# CHATWATCHLM - EXPLAINING GROOMING STRATEGIES USING GRAPH RAG AND KNOWLEDGE GENERATION

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

July 2024

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

			Page
LIST O	F TAB	LES	vi
LIST O	F FIGU	JRES	vii
ABBRE	VIATI	ONS	viii
GLOSS.	ARY .		ix
ABSTR	ACT		х
СНАРТ	EB 1	INTRODUCTION	1
1 1	Reseau	rch Gap	2
1.1	Reseau	rch Questions	3
1.2	Hypot	hesis	3
1.4	Reseau	rch Scope	4
1.5	How o	ur approach addresses the research gap?	4
1.6	Limite	ations	5
1.7	Assum	ptions	6
1.8	Delimi	itations	6
1.9	Signifi	cance	7
СНАРТ	ER 2.	REVIEW OF RELEVANT LITERATURE	8
2.1	Under	standing Child Grooming	8
	2.1.1	Grooming Stages, Modus Operandi and Characteristics	8
	2.1.2	Behavioral Analysis	12
	2.1.3	Linguistic Analysis	13
2.2	Datase	ets for Analyzing Predatory Conversations	15
2.3	Under	standing Large Language Models	16
	2.3.1	History of Language Models	16
	2.3.2	Reasoning Capabilities of LLMs	17
2.4	Unde	rstanding Prompting Techniques	18
	2.4.1	Zero-shot Prompting	19
	2.4.2	Few-shot Prompting	19
	2.4.3	Chain-of-Thought (CoT) Prompting	19
	2.4.4	Generated Knowledge Prompting	20
	2.4.5	Retrieval Augmented Generation (RAG)	21
2.5	Past F	Related Works	23
СНАРТ	'ER 3.	FRAMEWORK AND METHODOLOGY	25

		Pa
3.1	Knowledge-Entity Graph Construction	
3.2	Populating our Knowledge-Entity Graph with Documents	
0.2	3.2.1 Document Pre-Processing	
	3.2.2 Understanding the different document types	
	3.2.3 Indexing Documents by constructing Knowledge-Entity Graph	
33	Using the constructed $KE$ Graph Index to identify Grooming Strategies	
0.0	3.3.1 Creating the Search Document	
	3.3.2 Document Retrieval	
	3.3.3 Prompt Chaining to use the refined reasoning and the retrieved	
	documents to identify grooming strategy labels	
3.4	Evaluation	
	3.4.1 Evaluation Metrics	
	3.4.2 Comparative Analysis	
	LER 4. RESULTS	
4.1	Automatic Evaluation	
4.2 4-3	Improved Source Attribution Mechanism	
4.5	Discussion of Key Improvements	
7.7		
CHAPT	TER 5. CONCLUSION	
A.1	Appendices	
B.1	Prompts, Example Input, and Text Output for LLMGraphTransformer	
	B.1.1 Prompt for Entity-Knowledge Graph Construction	
	B.1.2 Example document chunk used to construct the graph below.	
	B.1.3 Example subgraph as text response generated by the LM for	
0.1	the above document	
C.1	Comparing an instance of our novel RAG against Few-Shot and Naive	
	RAG approaches	
	C.1.1 Conversation snippet (input) we're analyzing	
	C.1.2 Comparing the LLM-generated responses from Few-Shot, Naive	
	KAG and the Proposed Approach	
	C.1.5 Documents retrieved using the baseline approach	
П 1	0.1.4 Documents retrieved using our hybrid approach	
D.1 E 1	r rompts used during Evaluation	
E.1	Ingesting LLW-Generated Zero-Shot Reasoning Examples into the Knowledge Entity Craph $(C_{i})$	
	$\mathbf{Knowledge-Entity Graph}(\mathbf{G}_0) \dots \dots$	
LIST O	F REFERENCES	

# LIST OF TABLES

Tabl	e	Page
3.1	Chunk sizes and overlaps for different document categories	27
4.1	Performance Metrics of Different Approaches	39
1	Top 20 documents retrieved from our corpus by the baseline RAG, ranked by descending embedding similarity.	55
2	Top 20 documents retrieved from our corpus by Our RAG, ranked by descending quality score	55

# LIST OF FIGURES

Figure		
3.1	Architecture of the Proposed Approach	25
3.2	Subgraph example constructed by our approach using LLM for the document chunk (B.1.2).	26
3.3	Example of the overall constructed knowledge-entity graph after indexing documents (highlighted as RED nodes) from different categories using a LLM. The graph illustrates the interconnected entities and relationships generated by the LLM during the document indexing process.	30
3.4	Temporary integration of the search document subgraph (highlighted as ORANGE) into the overall KE Graph. This subgraph shows the document node with its neighboring nodes and their connections within the KE Graph. Neighboring nodes establish connections between the search document and other documents (highlighted as RED nodes) through shared nodes	31
4.1	Subgraph generated for a sample manuscript without chunking strategies	36
4.2	Subgraph generated for a sample manuscript chunked by structure tags and after every 8192 tokens	37
4.3	Relationship Between Chunk Size and Knowledge-Entity Graph Attributes	38
4.4	Example relationship between a Reference and a Document Node (Left). LLM-generated explanation for a predatory conversation snippet that retrieves the document on the left (right). Full example is available in Section C.1.2.	40

## ABBREVIATIONS

- AI Artificial Intelligence
- GPT Generative Pre-trained Transformer
- PLM Pre-trained Language Model
- 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
- RAG Retrieval Augmented Generation
- KE-G Knowledge-Entity Graph

# 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	Elements in LangChain that maintain the context
Managers	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, July 2024. ChatWatchLM -Explaining Grooming Strategies using Graph RAG and Knowledge Generation. Major Professor: Julia Rayz.

The detection and mitigation of predatory child grooming in digital communications remain paramount challenges in online safety. Addressing the deficiencies of current large language models (LLMs) in social reasoning, this study proposes an enhanced Retrieval-Augmented Generation (RAG) approach to improve the identification and explanation of grooming strategies in online conversations (Gandhi, Fraenken, Gerstenberg, & Goodman, 2023; Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022). The proposed method constructs Knowledge-Entity Graph leveraging annotated conversation snippets and relevant domain-specific research manuscripts (Edge et al., 2024; T. Ringenberg, 2021). Using LLMs, we construct a multi-edge knowledge-entity graph by identifying entity types, labels, and their relationships in each document. These sub-graphs are aggregated into a comprehensive graph to enhance document retrieval relevance and contextual diversity. The constructed graph is populated with documents categorized as Seed Documents, Research Manuscripts, and Examples, each processed with specific chunking strategies to capture intricate relationships and contextual information (Edge et al., 2024; Y. Gao et al., 2024). The effectiveness of the proposed method will be evaluated against the gold standard annotations from T. Ringenberg (2021), comparing the enhanced RAG system against baseline RAG and LLM-only approaches in terms of accuracy, precision, and recall. Our approach incorporates graph-based attributes and embedding similarity to compute a retrieval-quality score, ensuring that retrieved documents are contextually relevant and well-connected

within the knowledge graph, thereby improving LLM performance in identifying grooming strategies.

## CHAPTER 1. INTRODUCTION

The widespread use of the internet among minors has unfortunately increased their vulnerability to online child grooming, a pressing issue demanding immediate attention. According to Wolak, Finkelhor, and Mitchell (2004), 19% of youth who used the internet regularly received unwanted sexual solicitations, with 25% of those incidents involving aggressive solicitation tactics such as requests for sexual favors or attempts to arrange meetings. Preemptively identifying grooming behaviors is crucial to protect children from exploitation and contact offenses. While prior research, such as De Santisteban, Del Hoyo, Alcazar-Corcoles, and Gamez-Guadix (2018), has focused on ideal cases of grooming, a more comprehensive understanding of the entire spectrum of predatory behaviors is needed.

This research aims to enhance the detection and explanation of online child grooming strategies by integrating domain-specific knowledge into a Graph Based Retrieval-Augmented Generation (RAG) framework (Edge et al., 2024; Guo et al., 2023). We utilize annotated conversation snippets from *The Perverted Justice Foundation Incorporated* (2002); T. Ringenberg (2021) and relevant research manuscripts to construct a comprehensive knowledge-entity graph. Our method involves constructing multi-edge knowledge-entity graphs by identifying entities, labels, and relationships for each document, represented as nodes. Unlike baseline RAG, which relies solely on embedding similarity, our framework leverages both embedding similarity and LLM-generated graph relationships to retrieve contextually relevant documents.

The proposed approach enhances contextual relevance by dynamically augmenting the LLM with domain-specific knowledge, thereby improving their performance in identifying and explaining grooming strategies. This method addresses the shortcomings of current LLMs in social reasoning tasks and aims to provide a more robust tool for safeguarding minors online.

#### 1.1 Research Gap

Although state-of-the-art (SOTA) large language models (LLMs) achieve reasonable zero-shot performance in tasks such as multi-step arithmetic, common-sense reasoning, and symbolic reasoning Kojima et al. (2022), they exhibit reduced efficacy in replicating human-like inference patterns on social reasoning benchmarks Gandhi et al. (2023). This deficiency is particularly pronounced in the domain of online child grooming (Prosser & Edwards, 2024; Puentes et al., 2023), as further confirmed by our zero-shot and few-shot LLM evaluations. Such tasks fail to reliably identify grooming strategies and produce contextually rich multi-step reasoning. These shortcomings are evident when compared to expert-level analyses (including annotations) of predatory grooming conversations documented in the T. Ringenberg (2021) dissertation.

Simultaneously, relying solely on the parametric knowledge of SOTA LLMs has been shown to increase factual errors, despite improvements in model and data scale (Mallen et al., 2023; Min et al., 2023; Ouyang et al., 2022). Retrieval-Augmented Generation (RAG) methods (Edge et al., 2024; Guu, Lee, Tung, Pasupat, & Chang, 2020; Lewis et al., 2020) reduce factual errors in knowledge-intensive tasks by augmenting LLM inputs with relevant retrieved passages. This approach is particularly effective in social reasoning through the integration of domain-specific knowledge and improved contextual understanding (Asai, Min, Zhong, & Chen, 2023; Ram et al., 2023). The proposed approach addresses the urgent need to enhance LLM capabilities in identifying and explaining grooming strategies in online conversations (Steel, 2024). By using RAG and knowledge-generation techniques, this method augments LLMs with contextually relevant information tailored to the user query (Guu et al., 2020; J. Liu et al., 2022).

## 1.2 Research Questions

- 1. What is the accuracy of current state-of-the-art large language models in identifying child grooming strategies using our gold standard dataset, and how does this performance compare with expert-level analysis?
- 2. Which techniques, including advanced indexing and embedding models, enhance the robustness and contextual richness of knowledge-entity graphs in representing entity relationships for domain-specific tasks?
- 3. How do graph-based attributes and embedding similarity integration improve retrieval-quality scores for document connectivity and relevance in the context of identifying child grooming strategies?
- 4. How does the integration of entity-knowledge graphs in document retrieval systems enhance the accuracy and robustness of large language models in identifying and explaining child grooming strategies?
- 1.3 Hypothesis
  - 1. The integration of entity-knowledge graphs will significantly enhance the accuracy of large language models in identifying child grooming strategies, outperforming zero-shot and few-shot tasks.
  - 2. Graph-based attributes and embedding similarity enhance retrieval-quality scores, improving contextual relevance over embedding-only retrievers, especially in identifying child grooming strategies.
  - Utilizing entity-knowledge graphs in document retrieval systems will enhance the accuracy and robustness of large language models in identifying and explaining child grooming strategies.

## 1.4 Research Scope

Our research focuses specifically on the Knowledge Entity Graph Construction for creating a robust and contextually enriched representation of entities and their relationships from the indexed documents (Sui, Zhou, Zhou, Han, & Zhang, 2024). Additionally, our work emphasizes the Scoring of Retrieved Documents by integrating graph-based attributes with embedding similarity to compute a retrieval-quality score that captures the connectivity and relevance of retrieved documents within the constructed knowledge-entity graph. Finally, our approach aims at Relevant Document Retrieval by leveraging the constructed entity-knowledge graphs to dynamically provide context-aware examples and documents, aiming to enhance the performance of LLMs specifically in identifying and explaining grooming strategies.

Our scope, however, does not encompass several other critical modules within the broader RAG framework. Specifically, we are not focusing on the **Augmentation Techniques** such as iterative, recursive, and adaptive retrieval processes, which involve complex multi-step retrieval and generation cycles. We also do not delve into **Pre-retrieval and Post-retrieval Optimization** strategies that include query expansion, query transformation, and re-ranking. Furthermore, our research does not cover **LLM Fine-Tuning** and its integration with retrieval processes or the **Augmentation of Generation** aspects that deal with manipulating the retrieved content before feeding it into the language model for response generation. By delineating our focus, we aim to contribute to enhancing the foundational retrieval stage within RAG systems, particularly document indexing, retrieval and scoring.

## 1.5 How our approach addresses the research gap?

• Our approach retrieves a more diverse set of documents, which are less redundant and more contextually relevant compared to the baseline RAG. This enhanced retrieval capability can be attributed to the incorporation of the generated graph attributes and the entity-relationships constructed between documents (Edge et al., 2024).

- Unlike traditional RAG (Y. Gao et al., 2024; Huang & Huang, 2024), our approach enhances domain-specific retrieval by dynamically providing context-aware examples for few-shot prompting. This dynamic contextual relevance is achieved by integrating generated knowledge into the retrieval process, reducing noise, and improving the reasoning capabilities of LLMs in generating contextually appropriate responses (Edge et al., 2024; Y. Gao et al., 2024; J. Liu et al., 2022).
- Consequently, these improved retrieval capabilities is expected to better augment user queries with less noisy contextual relevant documents and enhance the performance of language models in generating contextually appropriate responses.

#### 1.6 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).

The proposed approach for knowledge-entity graph generation relies on LLMs capable of processing documents and generating structured outputs with function-calling capabilities. This necessitates fine-tuning on specialized datasets, such as (AI, 2022), to improve the models' ability for interacting with external tools. We leverage this capability to consistently generate structured output we can then parse to construct the sub-graphs.

However, models that exhibit high performance in both function calling (Yan, F., 2024) and the Eleuther AI Language Model Evaluation Harness (Lintang Sutawika et al., 2023), a unified framework to test generative language models on a large variety of reasoning tasks, are limited to models such as GPT-40, gemini-1.0-pro, mixtral-8x22b, and claude-3-opus.

## 1.7 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. We assume that the indexed documents, including research manuscripts and examples, are sufficient to augment the performance of the large language models. This assumption is consistent across all evaluated Retrieval-Augmented Generation (RAG) models, as they are initiated with the same set of documents and undergo identical pre-processing procedures. Uniformity in document provision and pre-processing ensures consistency in evaluation, thereby not affecting the research output.

## 1.8 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.

## 1.9 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 without fine-tuning LLMs and provides capabilities to incorporate private corpus, 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 components of Graph RAG (Edge et al., 2024), 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, 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.

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 (Radford, Narasimhan, Salimans, & Sutskever, 2018).

## 2.3.1 History of Language Models

The evolution of language models demonstrates substantial progress in managing linguistic complexity and data representation. Early models like *n*-gram models were limited by the curse of dimensionality and sparsity, which hindered their ability to process novel phrases effectively (Teller, 2000). These models could generate text but lacked coherence and scalability, especially for larger values of N(Bengio, Ducharme, & Vincent, 2000).

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

The Generative Pre-trained Transformers (GPT) architecture demonstrated that pre-training on a large corpus followed by fine-tuning could achieve state-of-the-art results across various tasks (Radford et al., 2018). GPT-3, introduced in 2020, highlighted the potential of scaling up language models. With an increase in 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 fine-tuned using Reinforcement Learning from Human Feedback (RLHF), aimed to mitigate issues of toxicity, bias and response alignment in language models (Ouyang et al., 2022). Integrating human feedback into the training process led to outputs more aligned with human preferences. While OpenAI has been a prominent player, other organizations like Meta, Google, and various open-source communities have significantly contributed with models such as OPT, FLAN-T5, BERT, BLOOM, and StableLM (AI, 2022; Devlin et al., 2018; Raffel et al., 2020; Workshop et al., 2022; S. Zhang et al., 2022).

### 2.3.2 Reasoning Capabilities of LLMs

SOTA LLMs achieve reasonable performance in zero-shot tasks like multi-step arithmetic and common-sense reasoning (Kojima et al., 2022). However, they unreliably emulate human-like inference patterns on social reasoning benchmarks (Gandhi et al., 2023), particularly in identifying grooming strategies in online child grooming scenarios (Prosser & Edwards, 2024; Puentes et al., 2023). State-of-the-art large language models (LLMs) have demonstrated significant advancements in natural language processing (NLP), particularly excelling in zero-shot tasks such as multi-step arithmetic and common-sense reasoning (Kojima et al., 2022). The notion of zero-shot learning refers to a model's ability to perform a task without having seen any task-specific examples during training. This capability is crucial for scaling the applicability of LLMs across diverse domains without the need for extensive task-specific data.

The performance of LLMs in these tasks is often attributed to their capability for in-context learning, where models are conditioned on a few examples (few-shot) or task-specific instructions (zero-shot) to generate accurate outputs (P. Liu et al., 2023). Studies demonstrated that SOTA achieve reasonable zero-shot, few-shot, and Chain-of-Thought (CoT) LLM performances on a wide range of benchmark reasoning tasks, including but not limited to arithmetic (MultiArith, GSM8K, AQUA-RAT, SVAMP), symbolic reasoning (Last Letter, Coin Flip), and other logical reasoning tasks (MT-bench, Date Understanding, Tracking Shuffled Objects) (Cobbe et al., 2021; L. Gao et al., 2023; Kojima et al., 2022; Zheng et al., 2023).

## 2.4 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.

## 2.4.1 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.

## 2.4.2 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.

## 2.4.3 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.

## Prompt Chaining

Prompt chaining, as described by (Wu, Terry, & Cai, 2022), involves using a sequence of LLM-generated responses where the output from one prompt serves as the input for the next, enabling the decomposition of complex tasks into manageable sub-tasks. This method enhances the performance and control of LLMs by incrementally improving outputs through each step. Identified to address the limitations of transparency, scope, and controllability in LLMs, chaining allows for modular adjustments and has been shown to significantly improve task outcomes, transparency, controllability, and user collaboration in a 20-person study. The study also revealed that users developed new ways of interacting with LLMs, such as calibrating model expectations and debugging outputs by "unit-testing" sub-components. Notably, these improvements were achieved without Parameter tuning the underlying model, suggested that breaking down complex tasks into smaller problems can enhance explainability and debuggability.

## 2.4.4 Generated Knowledge Prompting

Generated Knowledge Prompting is a method wherein a language model generates relevant information or knowledge from external sources prior to answering a query J. Liu et al. (2022). This technique leverages the inherent capabilities of language models to produce knowledge statements that can augment their performance on various reasoning tasks. By conditioning on human-written demonstrations, the method generates question-related knowledge statements, which are subsequently used to enhance the inference process. This approach does not necessitate task-specific supervision for knowledge integration, nor does it require access to a structured knowledge base. This makes it particularly suitable for a wide range of applications, including commonsense reasoning, where task-specific templates might be inadequate. This study have demonstrated that this method significantly improves the performance of SOTA LLMs on various commonsense reasoning evaluations such as NumerSense, CommonsenseQA, and QASC J. Liu et al. (2022). Despite its advantages, Generated Knowledge Prompting does have limitations. The accuracy of the generated knowledge is contingent upon the quality of the model's training data and parametric-knowledge. If the training data does not adequately cover the relevant domain, the generated knowledge may be factually-inaccurate or incomplete. This can lead to erroneous inferences and reduce the overall effectiveness of the method (T. Gao, Yen, Yu, & Chen, 2023; Guu et al., 2020; Shi et al., 2023).

## 2.4.5 Retrieval Augmented Generation (RAG)

Relying solely on the parametric knowledge of SOTA LLMs increases factual errors and "hallucinations", even with advancements in model scale and data (Mallen et al., 2023; Min et al., 2023; Ouyang et al., 2022). Retrieval-Augmented Generation (RAG) methods (Edge et al., 2024; Guu et al., 2020; Lewis et al., 2020) mitigate these errors by augmenting LLM inputs with relevant retrieved documents. This approach can provide natural source attribution mechanism and particularly effective in reasoning tasks when integrated with domain-specific knowledge (Asai et al., 2023; Ram et al., 2023). 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. Document chunking is commonly used in RAG application to balance the number of LLM calls required against the recall degradation of longer context windows (Edge et al., 2024; Huang & Huang, 2024; Kuratov et al., 2024; N. F. Liu et al., 2024).

RAG is particularly effective for reasoning tasks requiring real-time or 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" (Guu et al., 2020). However, its effectiveness heavily relies on the quality and relevance of the retrieved documents. (Lewis et al., 2020). Direct retrieval of text chunks in naive RAG approaches, which rely solely on embedding similarity (Lewis et al., 2020), often introduces off-topic passages, leading to low-quality LLM outputs (Shi et al., 2023). The presence of off-topic noise in retrieved documents compromises output consistency (T. Gao et al., 2023).

In a significant shift from this approach, recent studies (Edge et al., 2024; Guo et al., 2023; Trajanoska, Stojanov, & Trajanov, 2023; Wang et al., 2023) have begun exploring a RAG pipelines that solely relies on LLMs for indexing documents as a Knowledge-Entity Graph (KE). The first series of work focus on applying LLMs without tuning the parameters. GPT4Graph (Guo et al., 2023) evaluates the potential of LLMs in executing knowledge graph (KG) reasoning and node classification tasks. Their findings indicate that these models can deliver competitive results for short-range KE reasoning but struggle with long-range KG reasoning and node classification tasks. These study primarily concentrates on employing LLMs for KE graph indexing and as predictors for traditional graph reasoning tasks such as shortest path, maximum flow, global and bipartite matching (Chen et al., 2024; Edge et al., 2024; Guo et al., 2023).

To address these issues, Graph RAG (Edge et al., 2024) introduced the usage of query-focused summarization and community summaries during KE graph construction. The Graph RAG pipeline (Edge et al., 2024) constructs a graph index from source documents using LLMs to identify neighboring entities, relationships, and claims. This index comprises nodes (entities), edges (relationships), and covariates (claims) detected through a LLM prompt. Community detection algorithms then partition the graph into groups of closely related nodes, edges, and covariates. Given a question, each community summary generates a partial response, which is then summarized into a final response for the user.

Answering user queries is treated as a LLM-as-Predictor task on KE graphs primarily either involves in-context (in-prompt) learning of graph attributes, or fine-tuning LLMs to enable them to answer user queries based on graph information represented as embeddings (Chen et al., 2024; Edge et al., 2024; Guo et al., 2023; Trajanoska et al., 2023).

## 2.5 Past Related Works

Traditional machine learning approaches to differentiate the language used by online child predators and their victims, as demonstrated by Pendar (2007) and Cano, Fernandez, and Alani (2014), have shown promising results but come with significant limitations. Pendar (2007) utilized chat logs from the Perverted Justice Foundation, employing a Support Vector Machine (SVM) classifier with lexical, syntactic, and semantic features and achieved an *f*-measure of 0.943 using a distance-weighted k-NN classifier. Similarly, Cano et al. (2014) proposed a methodology for detecting child grooming behavior patterns on social media by incorporating pre-processing steps such as translating emoticons and chat lingo, and employing features like *N*-grams, syntactical content, sentiment polarity, and psycho-linguistic analysis. This SVM classifier approach demonstrated improved results over keyword-based detection methods but faced limitations due to a small dataset and challenges in generalizing to other platforms or languages.

Both studies rely on datasets involving adult volunteers rather than actual minors, potentially misrepresenting genuine predator-victim interactions. Traditional classifiers like SVM and k-NN struggle with capturing long-range dependencies and contextual information in text data (Cano et al., 2014; Hochreiter & Schmidhuber, 1997; Pendar, 2007). In contrast, modern approaches using Long Short-Term Memory (LSTM) and Transformer models, as introduced by Vaswani et al. (2017), can better model complex language structures by handling varying input lengths and accounting for word context. These models also eliminate the need for manual feature engineering by automatically learning relevant features from data (LeCun, Bengio, & Hinton, 2015).

## CHAPTER 3. FRAMEWORK AND METHODOLOGY

To address the limitations of current LLMs in social reasoning tasks and improve factual accuracy in LLM generations, we employ Graph RAG, knowledge-generation and prompt chaining techniques to augment LLMs with relevant information (Edge et al., 2024; Guu et al., 2020; J. Liu et al., 2022) to explain and identify grooming strategies. This approach includes the following steps:



Figure 3.1.: Architecture of the Proposed Approach

## 3.1 Knowledge-Entity Graph Construction

We employ large language models (LLMs) to construct a multi-edge knowledge-entity graph by identifying entity types (e.g., Concepts, Actors), entity labels (e.g., Predator for Actors), and their relationships (e.g., mentions, uses, involvement) in each document. Specifically, our method constructs them as triplets (head entity, relation, tail entity) from source documents by prompting the LLM, following methodologies outlined in (Baek, Aji, & Saffari, 2023; Edge et al., 2024;
Trajanoska et al., 2023; Yao, Peng, Mao, & Luo, 2024). Each source document is designated as the parent node for its corresponding subgraph in the KE Graph Index.

The prompt used to construct the knowledge entity graph, the input document chunk, and an example output from the model are provided in the appendix (refer to B.1.1, B.1.2, and B.1.3). This process can be formalized as a function, LLMGraphTransformer, which iterates over each document in the sets of seed documents, manuscripts, and examples. For each document  $(D_i)$ , the LLMGraphTransformer generates a subgraph (Subgraph<sub>D<sub>i</sub></sub>) containing neighboring nodes and their vertices. This can be represented as:

LLMGraphTransformer $(D_i) \rightarrow \text{Subgraph}_{D_i}$ 



Figure 3.2.: Subgraph example constructed by our approach using LLM for the document chunk (B.1.2).

In the figure 3.2, the LLM prompted with B.1.1 constructed neighboring nodes for the input document by identifying the neighboring node types (e.g., Person, Objective, Strategy, Reaction), their associated IDs or labels (e.g., Predator, Decoy Child, Trust Building), and relationships, which include source and destination pairs and their relationship types (e.g., AIMS\_TO, USED\_FOR, SHOWS).

# 3.2 Populating our Knowledge-Entity Graph with Documents

# 3.2.1 Document Pre-Processing

To construct the foundational knowledge-entity graph structure,  $G_0$ , we ingest three distinct categories of documents: Seed Documents, Research Manuscripts, and Examples. Each category is processed with specific chunking strategies to effectively incorporate their content into the graph (Edge et al., 2024; Y. Gao et al., 2024). Chunking is a critical yet challenging aspect of building retrieval-augmented generation (RAG) applications. It involves dividing the text into manageable segments, which varies depending on the document's structure. We use document structure tags (e.g., headings, subheadings) to guide initial chunking, aiming to balance attention focus and context length, thus reducing noise and optimizing large language model (LLM) performance (Edge et al., 2024; Y. Gao et al., 2024; Huang & Huang, 2024; Lewis et al., 2020). The influence of chunking strategies on LLM-generated KE graph attributes specific to our indexed documents are discussed in Section 4.1.

Document Category	Chunk Size	Chunk
	(tokens)	Overlap
		(tokens)
Seed Documents	1024	128
Research Manuscripts	4096	512
Examples	N/A	N/A

Table 3.1.: Chunk sizes and overlaps for different document categories

# 3.2.2 Understanding the different document types

### Seed Document

In our RAG approach, seed documents are indexed to establish foundational nodes and relationships within the knowledge-entity graph, enhancing future retrieval performance (Edge et al., 2024). As shown in Table 3.1, we use relatively smaller chunk sizes and overlaps for seed documents, which primarily contain one-line introductions to commonly discussed grooming strategies in T. Ringenberg (2021). Modifying chunking strategies based on document structure is a commonly used method to improve retrieval precision (Edge et al., 2024; Y. Gao et al., 2024; Huang & Huang, 2024).

# **Research Manuscripts**

Incorporating domain-specific research manuscripts enriches the contextual representation of the knowledge-entity graph by incorporating parsed information from the indexed documents as Entities and relationships (Chen et al., 2024; Edge et al., 2024; Guu et al., 2020). For our tasks, domain-specific research manuscripts will include behavioural and linguistic studies on predator-victim dynamics.

Initially, documents are divided based on structural tags such as headings and subheadings to preserve context and minimize noise (Asai et al., 2023; Huang & Huang, 2024; Ram et al., 2023). The divided sections are then chunked based on specified sizes with overlapping tokens to ensure continuity of information (Table 3.1). Research manuscripts are collected in markdown format, selected based on open license availability, accessibility, domain relevance, and citation count. Based on these criteria and as stated in the assumptions, we posit that the indexed documents, including research manuscripts and examples, will sufficiently augment the performance of the LLM.

# LLM-Generated Zero-Shot Reasoning Examples

In this step, we generate and index task-specific examples into our knowledge-entity graph ( $G_0$ ) (J. Liu et al., 2022) using Prompt D.1. We utilize annotated conversation snippets from the T. Ringenberg (2021) dataset, prompting the LLM to explain the reasons behind the annotations using provided snippets and one-line grooming strategy explanations from the thesis. This process generates examples for sequential dialog exchanges sharing the same grooming strategies, isolating and explicating each strategy. The prompt and the LLM-generated responses are combined into example documents and ingested as sub-graphs, enriching  $G_0$  with contextually enriched examples (Edge et al., 2024). A sample prompt template is available in Section D.1. Example documents retrieved from  $G_0$ populate tailored few-shot (Brown et al., 2020) examples for user queries, and thus, examples are not chunked (Table 3.1).

# 3.2.3 Indexing Documents by constructing Knowledge-Entity Graph

We transform and index documents from various categories (seed documents, manuscripts, and examples) into a comprehensive knowledge-entity graph, denoted as  $G_0$  (as discussed in Section 3.1). This graph captures entities and relationships among indexed documents. The formal representation of this indexing process is as follows, and a visualization of the constructed graph is available in Figure 3.3:

$$G_{0} = \sum_{i} \text{LLMGraphTransformer}(D_{i}^{\text{seed}}) + \sum_{i} \text{LLMGraphTransformer}(D_{i}^{\text{manuscripts}}) + \sum_{i} \text{LLMGraphTransformer}(D_{i}^{\text{examples}})$$



Figure 3.3.: Example of the overall constructed knowledge-entity graph after indexing documents (highlighted as RED nodes) from different categories using a LLM. The graph illustrates the interconnected entities and relationships generated by the LLM during the document indexing process.

3.3 Using the constructed KE Graph Index to identify Grooming Strategies

# 3.3.1 Creating the Search Document

For a given predatory conversation snippet, we start by zero-shot prompting the LLM to generate explanations for any grooming strategies present (J. Liu et al., 2022). The input chat snippet, along with the generated explanations, forms the search document for retrieving relevant documents from the KE graph. A subgraph is then constructed for the search document, as previously described in Section 3.1. This subgraph is temporarily merged with the KE graph index to facilitate document retrieval and calculate retrieval-quality scores.



Figure 3.4.: Temporary integration of the search document subgraph (highlighted as ORANGE) into the overall KE Graph. This subgraph shows the document node with its neighboring nodes and their connections within the KE Graph. Neighboring nodes establish connections between the search document and other documents (highlighted as **RED** nodes) through shared nodes.

# 3.3.2 Document Retrieval

Our retrieval method integrates graph attributes (degree), embedding similarity (Lewis et al., 2020), and Dijkstra's shortest path algorithm to calculate retrieval-quality scores within the KE graph. This integration ensures that the retrieved documents are both similar in embedding space and well-connected within the knowledge graph. Compared to Naive RAG, this approach enhances the contextual relevance of retrieved documents. Documents are retrieved along with their nodal paths and the relationships between source and destination nodes. This structured incorporation ensures that the prompt clearly explains the interpretability of the relationship between the user query and the retrieved document. In the following, we will detail our formula for calculating the score of relevant documents:

- 1. Normalize the Variables:
  - Distance from Source: The shortest distance between the Source and Target Document Nodes is calculated using Dijkstra's algorithm. This distance is normalized to a 0-1 range by using the formula:

Normalized Distance<sub>i</sub> = 
$$\frac{1}{1 + \text{Distance from Source}_i}$$

This transformation ensures that a smaller distance (closer to the source) results in a higher value after normalization.

• Normalized Degree of the Destination Node: Normalized by using a logarithmic scale to reduce the wide range impact where max(Degree) is the maximum degree observed in list of connected documents. This ensures the values are squeezed into the 0-1 range and lessen the impact of high-degree nodes:

Normalized 
$$\text{Degree}_i = \frac{\log(\text{Degree of the Destination Node}_i + 1)}{\log(\max(\text{Degree}) + 1)}$$

- 2. Weight the Variables:
  - We need to determine the weights  $w_1$ ,  $w_2$ , and  $w_3$  for Embedding Similarity, Normalized Distance, and Normalized Degree, respectively. Since the Embedding Similarity is likely the most direct measure of relevance (Béchard & Ayala, 2024; Edge et al., 2024; Lewis et al., 2020), it can be given more weight. Distance from Source could be

slightly less important, and Degree of the Destination Node might be the least important. The weights were calculated through a sensitivity analysis on the constructed KE graphs.  $w_1 = 0.5$ ,  $w_2 = 0.3$   $w_3 = 0.2$ . Future research could further investigate hyper-parameter optimization for these weights.

 It is necessary to ensure that the sum of the weights equals 1 and the weights w<sub>1</sub>, w<sub>2</sub>, and w<sub>3</sub> should be tuned based on empirical results or domain expertise.

$$w_1 + w_2 + w_3 = 1$$

3. Calculate the Score: We calculate the final scores by combining the normalized and weighted variables. Higher the score, the more relevant:

$$Score_i = w_1 \cdot Embedding Similarity_i + w_2 \cdot Normalized Distance_i + w_3 \cdot Normalized Degree_i$$

3.3.3 Prompt Chaining to use the refined reasoning and the retrieved documents to identify grooming strategy labels

We then prompt the LLM to refine the previously generated zero-shot reasoning using prompt chaining with the retrieved documents (J. Liu et al., 2022; Wu et al., 2022). Prompt chaining involves sequential transformations on generated responses to achieve the desired outcome (Wu et al., 2022), and we again further extend this chain to generate grooming strategy annotations for the initial conversation snippet as a structured response. A complete example chain is available in D.1.

### 3.4 Evaluation

The effectiveness of our proposed method will be assessed using the T. Ringenberg (2021) dataset, a gold standard for analyzing grooming strategies. We will compare the performance of LLMs augmented with our proposed method against Naive RAG and in-context few-shot approaches in identifying grooming strategies. The primary evaluation metric will be accuracy, measured against the ground truth annotations provided in the dataset.

# 3.4.1 Evaluation Metrics

The evaluation will focus on three primary metrics: accuracy, precision, and recall. Accuracy will be measured by comparing the generated annotations with the ground truth annotations in the T. Ringenberg (2021) dataset. Precision and recall will also be calculated to provide a comprehensive assessment of each approach's capability in identifying grooming strategies.

### 3.4.2 Comparative Analysis

We will conduct a comparative analysis of three approaches: LLM-only, Baseline RAG, and our Enhanced RAG. All prompts used during evaluation are available in Section D.1.

1. In-context Few-Shot: We will evaluate performance improvements when LLMs are provided with a few contextually relevant examples and one-line explanations of strategies discussed in T. Ringenberg (2021); T. R. Ringenberg et al. (2022).

2. Baseline or Naive RAG: According to Y. Gao et al. (2024), Naive RAG often struggles with precision and recall, leading to the selection of misaligned or irrelevant chunks, and the omission of crucial information. We will evaluate this

baseline approach, which relies on vector similarity for document retrieval, to measure its effectiveness in enhancing LLM-generated responses.

3. **Proposed Approach**: Our evaluation will focus on the accuracy of identifying grooming strategies augmented with associated LLM-generated explanation.

### CHAPTER 4. RESULTS

This chapter presents the influene of chunking strategies and the results of our evaluation of our approach on expert annotations for analyzing grooming strategies (T. Ringenberg, 2021). We conducted a robust evaluation by randomly sampling 118 annotated conversation snippets, each containing consecutive annotations for four or more chat lines. The performance metrics—precision, recall, and accuracy—were evaluated using GPT-4o as the foundational large language model (LLM) for all approaches. For consistency during evaluation, our proposed approach and Naive RAG based on Huang and Huang (2024) will share identical document pre-processing and index research papers on child-predation.

# 4.1 Understanding the Influence of Chunking Strategies

This section evaluates the document chunking strategies used in constructing Knowledge-Entity (KE) Graphs. Figure 4.1 visualizes a KE graph generated for a sample manuscript without employing any chunking strategies.



Figure 4.1.: Subgraph generated for a sample manuscript without chunking strategies



Figure 4.2.: Subgraph generated for a sample manuscript chunked by structure tags and after every 8192 tokens

We investigated the influence of document chunking strategies on the KE graph from our indexed research manuscripts. We systematically tested chunking strategies, varying context length by powers of 2, from 32 to 16,384 tokens, with overlaps ranging from 4 to 2,048 tokens respectively. Our analysis focused on three key metrics: the average number of relationships per chunk, the average number of unique entity types identified, and the average number of unique entity IDs (labels) extracted. These metrics served as proxies for the comprehensiveness and specificity of knowledge extraction under different chunking lengths.

Figure 4.3 illustrates the relationship between chunk size (logarithmic scale) and LLM-generated KE graph attributes. Each subplot represents a different metric averaged across multiple document chunking strategies.

The results highlight three key metrics: the average number of relationships, unique entity types, and unique entity IDs as functions of chunk size. An increasing trend in the number of relationships with larger chunk sizes (Figure 4.3) suggests that larger chunks facilitate the extraction of more relationships. Similarly, the diversity of entity types and the number of unique entity IDs both rise with increasing chunk



Figure 4.3.: Relationship Between Chunk Size and Knowledge-Entity Graph Attributes

sizes, indicating that larger chunks encompass a wider array of entity types and contribute to the identification of a more extensive set of unique entities.

Given that SOTA LLMs, such as GPT-40 with a context length of 128K tokens, may suffer from attention dilution for longer inputs (Kuratov et al., 2024; B. Liu, Ash, Goel, Krishnamurthy, & Zhang, 2023), the purpose of RAG is to retrieve multiple documents that fit well within the context length. Therefore, as indicated in Table 3.1, we use a chunk size of 4,096 tokens. This approach ensures that multiple retrieved documents can be incorporated within the LLM's context length while consistently generating neighboring entities and relationships.

# 4.2 Automatic Evaluation

The performance of each approach was evaluated using precision, recall, and accuracy. As summarized in Table 4.1, our proposed approach outperformed the baseline Naive RAG and Zero-Shot approaches across all metrics.

Approach	Precision	Recall	Accuracy
Our Approach	0.884	0.646	0.643
Naive Approach	0.706	0.542	0.521
Zero-Shot Approach	0.704	0.539	0.518

Table 4.1.: Performance Metrics of Different Approaches

The Naive and Zero-Shot approaches exhibited similar performance, highlighting the limitations of relying solely on embedding similarity and the lack of contextual enrichment. The precision deltas of 0.178 and 0.180 suggest a substantial reduction in false positives, while the recall deltas of 0.104 and 0.107 demonstrate improved capability in identifying true positives. The accuracy deltas of 0.122 and 0.125 reflect substantial overall improvements.

These findings underscore the effectiveness of chaining Knowledge-Entity (KE) Graphs with Knowledge Generation techniques, which enhance LLM performance by providing essential contextual enrichment (Edge et al., 2024; J. Liu et al., 2022; Wei et al., 2023; Wu et al., 2022). This mitigates the limitations of embedding similarity-based methods.

# 4.3 Improved Source Attribution Mechanism

Documents are indexed with their metadata, including bibliographic information, during KE graph construction. This bibliographic data is used to attribute document chunks to their respective authors. Document chunks from the same manuscript are interconnected in the Knowledge-Entity (KE) Graph through shared metadata. For instance, a *Reference* node labeled *Williams et al.*, 2013 links



Figure 4.4.: Example relationship between a Reference and a Document Node (Left). LLM-generated explanation for a predatory conversation snippet that retrieves the document on the left (right). Full example is available in Section C.1.2.

all relevant *Document* nodes in the graph created by LLMGraphTransformer. During retrieval, our method extracts documents along with their connecting paths, enhancing contextual awareness by demonstrating the relationships between retrieved documents and the user query through rule-based path interpretations. For example, *Document\_ABC* is linked to the *User Query* through "Digital Media" *USED IN* " Progressive Sexualization" and "Webcam Sessions" *TARGET* "Female Victims." This mechanism ensures accurate source attribution by organizing and contextualizing the retrieved information based on its bibliographic metadata.

# 4.4 Discussion of Key Improvements

Our approach does relatively require additional LLM calls for constructing and querying the KE Graph, which can increase computational overhead. The need for LLM support for structured responses also necessitates further fine-tuning and optimization to fully leverage the benefits of our method, as discussed in Section 1.4.

The incorporation of the Knowledge-Entity (KE) Graph, along with prompt chaining and knowledge generation techniques, significantly enhances the performance of large language models (LLMs) in identifying grooming strategies in online conversations. The improved precision reflects greater accuracy in positive predictions, while the higher recall demonstrates the method's capability to identify a larger number of true positive instances from the annotated dataset. The overall accuracy underscores the robustness of our approach in producing reliable results.

These results support our hypothesis that integrating KE Graphs and advanced prompting techniques improves LLM performance. While our approach advances the state-of-the-art, the field remains challenged by the inherent limitations of current LLM technologies when compared to gold-standard benchmarks. Further research is necessary to bridge these gaps and achieve more reliable and accurate identification of complex social reasoning patterns.

# CHAPTER 5. CONCLUSION

This thesis aims to enhance the detection and mitigation of predatory child grooming on digital communication platforms by introducing an novel method that integrates Graph Retrieval-Augmented Generation (RAG), Knowledge Generation, and Prompt Chaining. By constructing a Knowledge-Entity (KE) Graph through the indexing of domain-specific knowledge and the explanation of grooming strategies, our method leverages existing behavioral research on predatory grooming strategies and LLM-generated zero-shot examples. This KE Graph facilitates contextually enriched document retrieval, thereby augmenting the language model's ability to identify and explain grooming strategies.

The proposed method constructs a Knowledge-Entity (KE) Graph by indexing domain-specific knowledge and explaining grooming strategies. This involves indexing existing behavioral research on predatory grooming strategies and incorporating LLM-generated zero-shot examples. The KE Graph (system prompt and pipeline) is designed to capture intricate relationships and contextual information, enabling more effective retrieval of relevant documents. By integrating traditional graph attributes with embedding similarity and employing Dijkstra's shortest path algorithm, our approach computes a retrieval-quality score that enhances the contextual relevance and connectivity of retrieved documents.

The effectiveness of our proposed method, ChatWatchLM, was evaluated using the Ringenberg (2021) dataset T. Ringenberg (2021), a gold standard for analyzing grooming strategies. We conducted automated assessments to gauge its performance in identifying grooming strategies in online chats. Our results demonstrated that ChatWatchLM outperforms Naive RAG and few-shot approaches across all metrics, achieving higher precision, recall, and accuracy. Specifically, our approach achieved a precision of 0.884, a recall of 0.646, and an accuracy of 0.643, highlighting its superior capability in detecting true positives and reducing false positives. These findings underscore the effectiveness of integrating KE Graphs and advanced prompting techniques in enhancing the performance of large language models in recognizing complex predatory behaviors.

Our approach addresses several critical research gaps, including the need for enhanced contextual understanding and the ability to identify complex social reasoning patterns in predatory behavior. However, challenges remain, such as managing computational overhead and the need for further fine-tuning to fully leverage structured responses.

Future research could explore incorporating advanced techniques such as Document Ranking, Retrieval-Augmented Fine-Tuning, and Prefix-Tuning to mitigate hallucinations and improve retrieval accuracy. Additionally, hyper-parameter tuning for the retrieval-quality score weights could further minimize the influence of irrelevant documents and enhance document retrieval quality (Y. Gao et al., 2024; Vos, Döhmen, & Schelter, 2022; T. Zhang et al., 2024). While our method shows significant promise, fine-tuning models for specific tasks may yield even better performance, suggesting a potential area for further investigation. APPENDICES

# A.1 Appendices

In research involving 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.

B.1 Prompts, Example Input, and Text Output for LLMGraphTransformer

Discussed in the Section 3.1.

# B.1.1 Prompt for Entity-Knowledge Graph Construction

# Ki	nowledge Graph Instructions for GPT-40
##	1. Overview
You	are a top-tier algorithm designed for extracting information in structured
$\hookrightarrow$	formats to build a knowledge graph. You are optionally engineered to use
$\hookrightarrow$	your knowledge graphing skills to analyze research manuscripts to relate the
$\hookrightarrow$	different aspects of online child grooming, exploitation, and predation. You
$\hookrightarrow$	may also focus on identifying behavioral dynamics between predators and
$\hookrightarrow$	victims, grooming stages, and additional elements explained in the following
$\hookrightarrow$	sections. Ideally, you may correlate the discussions from user input to
$\hookrightarrow$	various features to identify grooming strategy or behavioral patterns
$\hookrightarrow$	deployed by predators and how victims may react to it in an online
$\hookrightarrow$	conversation. The information and insights you extract are valuable for
$\hookrightarrow$	researchers studying linguistic, behavioral, and manipulation patterns in
$\hookrightarrow$	predator-victim interactions and conversational logs.
The	se grooming patterns may be identified as lexical, syntactical, sentimental,
$\hookrightarrow$	psycho-linguistic, or as discourse patterns. Focus especially on ways or
$\hookrightarrow$	theories on how research can identify predatory behaviors in real-time
$\hookrightarrow$	online predatory conversations.

Try to capture as much information from the text as possible without sacrificing  $\rightarrow$  accuracy. Do not add any information that is not explicitly mentioned in the  $\rightarrow$  text.

- \*\*Nodes\*\* represent entities, and concepts.

- The aim is to achieve simplicity and clarity in the knowledge graph, making it  $\rightarrow$  accessible for a vast audience.

#### ## 2. Labeling Nodes

- \*\*Consistency\*\*: Ensure you use available types for node labels.

Ensure you use basic or elementary types for node labels.

- For example, when you identify an entity representing a person, always label

- $\hookrightarrow$  it as \*\*'person'\*\*. Avoid using more specific terms like 'mathematician' or  $\hookrightarrow$  'scientist'

- \*\*Relationships\*\* represent connections between entities or concepts. Ensure consistency and generality in relationship types when constructing

- $\rightarrow$  knowledge graphs. Instead of using specific and momentary types such as
- -> `BECAME\_PROFESSOR`, use more general and timeless relationship types like
- $\rightarrow$  'PROFESSOR'. This can also apply between concepts, objects, grooming
- $\, \hookrightarrow \,$  strategies, behavioral patterns or linguistic patterns. Make sure to use
- $\hookrightarrow$  general and timeless relationship types!

#### ## 3. Coreference Resolution

- \*\*Maintain Entity Consistency\*\*: When extracting entities, it's vital to  $\rightarrow$  ensure consistency.

If an entity, such as "John Doe", is mentioned multiple times in the text but is → referred to by different names or pronouns (e.g., "Joe", "he"), always use → the most complete identifier for that entity throughout the knowledge graph. → In this example, use "John Doe" as the entity ID. This may also apply to → concepts, objects, grooming strategies, behavioral patterns or linguistic → patterns. Remember, the knowledge graph should be coherent and easily understandable, so

→ maintaining consistency in entity references is crucial.

## 4. The below code block should provide you with the background information to understand the different known grooming strategies and behavioral patterns  $\hookrightarrow$ in predator-victim interactions. You may optionally use this information to  $\hookrightarrow$ further enrich your knowledge graph.  $\hookrightarrow$ 111 Coercion may include ..... Progressive Sexualization may ..... [...Trucated...] Deception may include ..... ... **##** 4. Strict Compliance Adhere to the rules strictly. Non-compliance will result in termination. **##** 5. Formatting Instructions [Truncated instructions to generate a structured JSON]

# B.1.2 Example document chunk used to construct the graph below.

{	"text": "Sexualization online also involves the use of digital media in
$\hookrightarrow$	addition to text. When digital media are exchanged, it varies as to who is
$\hookrightarrow$	pictured and the types of materials requested. Webcam sessions are framed as
$\hookrightarrow$	an innocuous exchange but lead to the groomer exposing themselves or
$\hookrightarrow$	masturbating. Exposures over webcam are more likely to occur in the case of
$\hookrightarrow$	female victims than males. Offline groomers also use digital media, showing
$\hookrightarrow$	victims sexual images and videos which include adult pornography, child
$\hookrightarrow$	pornography, and dead animals. Extra-familial groomers are more likely to
$\hookrightarrow$	use pornography during an offense than incest groomers. However, in cases in
$\hookrightarrow$	which family and acquaintance offenders use the Internet to facilitate
$\hookrightarrow$	abuse, sexual pictures and pornography are sent to the victim. When abuse
$\hookrightarrow$	moves from online to the offline environment, images and photography are
$\hookrightarrow$	used to initiate the abuse, with offenders having the child pose in
$\hookrightarrow$	gradually more sexual positions.",
	"Section": "2. Basic Research Background Knowledge" }

B.1.3 Example subgraph as text response generated by the LM for the above document

This is the raw LLM-generated response that we parse to form a sub-graph (discussed in the Section 3.1).

```
GraphDocument(
                                                         Node(id='Family And Acquaintance

→ Offenders', type='Concept'),

    nodes=[
                                                        Node(id='Sexual Pictures And
        Node(id='Sexualization Online',

    type='Concept'),

                                                         \rightarrow Pornography',
        Node(id='Digital Media',
                                                         \rightarrow type='Concept'),

→ type='Concept'),

                                                        Node(id='Offline Environment',
        Node(id='Webcam Sessions',
                                                         \rightarrow type='Concept'),
                                                        Node(id='Images And

→ type='Concept'),

        Node(id='Exposure Over Webcam',
                                                         \rightarrow Photography',
                                                         \rightarrow type='Concept')

→ type='Concept'),

        Node(id='Female Victims',
                                                    ],

→ type='Person'),

                                                    relationships=[
        Node(id='Male Victims',
                                                         Relationship(
         \leftrightarrow type='Person'),
                                                             source=Node( id =
        Node(id='Offline Groomers',
                                                              \rightarrow 'Sexualization Online',

→ type='Concept'),

                                                              \rightarrow type = 'Concept'),
        Node(id='Sexual Images And
                                                             target=Node( id = 'Digital
         → Videos', type='Concept'),
                                                              \rightarrow Media', type =
        Node(id='Adult Pornography',
                                                              \rightarrow 'Concept'),

→ type='Concept'),

                                                              \rightarrow type='INVOLVES'),
        Node(id='Child Pornography',
                                                        Relationship(

→ type='Concept'),

                                                             source=Node(id='Digital
        Node(id='Dead Animals',
                                                              \rightarrow Media',
         \leftrightarrow type='Concept'),

→ type='Concept'),

        Node(id='Extra-Familial
                                                             target=Node(id='Webcam
         \hookrightarrow Groomers',
                                                              \rightarrow Sessions',

    type='Concept'),

                                                              → type='Concept'),
        Node(id='Incest Groomers',
                                                              → type='USED_IN'
         \leftrightarrow type='Concept'),
                                                         ),
```

```
Relationship( source = Node ( id
                                                      Relationship(
     = 'Webcam Sessions',
                                                           source=Node(id='Sexual
                                                      \hookrightarrow
     type = 'Concept' ), target =
                                                           Images And Videos',
                                                      \hookrightarrow
     Node( id = 'Exposure Over
                                                           type='Concept'),
                                                      \hookrightarrow
                                                           target=Node(id='Dead
     Webcam', type = 'Concept' ),
                                                      \hookrightarrow
     type = 'LEADS_TO' ),
                                                           Animals', type='Concept'),
                                                      \hookrightarrow
Relationship( source = Node(id =
                                                           type='INCLUDES'),
                                                      \hookrightarrow
     'Exposure Over Webcam', type
                                                      Relationship(
     = 'Concept'), target =
                                                      \hookrightarrow
                                                           source=Node(id='Extra-Familial
     Node(id = 'Female Victims',
                                                           Groomers', type='Concept'),
                                                      \hookrightarrow
     type = 'Person'), type =
                                                           target=Node(id='Adult
                                                      \hookrightarrow
                                                           Pornography',
                                                      \hookrightarrow
                                                           type='Concept'),
                                                      \hookrightarrow
     source=Node(id='Offline
                                                           type='MORE_LIKELY_TO_USE'),
                                                      \hookrightarrow
     Groomers', type='Concept'),
                                                      Relationship(
     target=Node(id='Sexual
                                                           source=Node(id='Family And
                                                      \hookrightarrow
     Images And Videos',
                                                           Acquaintance Offenders',
                                                        \rightarrow 
                                                           type='Concept'),
                                                       \hookrightarrow
                                                           target=Node(id='Sexual
                                                           Pictures And Pornography',
                                                       \rightarrow
     source=Node(id='Sexual
                                                           type='Concept'),
                                                      \hookrightarrow
     Images And Videos',
                                                           type='SEND_TO_VICTIM'),
                                                      \rightarrow
                                                      Relationship( source = Node(id =
     target=Node(id='Adult
                                                           'Offline Environment', type
                                                           = 'Concept' ), target =
                                                           Node(id = 'Images And
                                                        \rightarrow 
                                                           Photography', type =
                                                       \rightarrow
                                                           'Concept' ), type =
                                                        \rightarrow 
                                                           'USED_TO_INITIATE_ABUSE' )
     source=Node(id='Sexual
                                                       \hookrightarrow
     Images And Videos',
                                                 ],
                                                 source = Document( page_content='[
     target=Node(id='Child
                                                 \hookrightarrow
                                                      Truncated Content ]',
                                                      metadata={'Section': 'Basic
                                                 \hookrightarrow
                                                      Research Background
                                                 \hookrightarrow
```

Knowledge'}))

 $\hookrightarrow$ 

type='Concept'), type='INCLUDES' ),

type='Concept'),

Pornography',

 $\rightarrow$ 

 $\hookrightarrow$ 

 $\rightarrow$ 

 $\hookrightarrow$ 

 $\rightarrow$ 

 $\rightarrow$ 

 $\hookrightarrow$ 

 $\rightarrow$ 

 $\hookrightarrow$ 

 $\rightarrow$ 

 $\rightarrow$ 

 $\hookrightarrow$ 

 $\hookrightarrow$ 

<u></u>

 $\hookrightarrow$ 

Relationship(

'MORE\_LIKELY\_IN'),

type='Concept'),

type='Concept'),

type='INCLUDES' ),

Pornography', type='Concept'),

type='USE'),

Relationship(

Relationship(

48

- C.1 Comparing an instance of our novel RAG against Few-Shot and Naive RAG approaches
  - 1. We'll be comparing the Retrieved documents from our Novel RAG and Naive RAG approaches.
  - 2. We will compare the LLM response of all few shot and prompts augmented with the documents retrieved from both the RAG approaches.
- C.1.1 Conversation snippet (input) we're analyzing

Potential Predator: youre marriage	Potential Predator: ill play it with you
$\hookrightarrow$ material?	Potential Predator: xbox is fun
Potential Child Victim: dude im 14 im	Potential Predator: your mom going away
$\hookrightarrow$ totally not thinking marriage	$\hookrightarrow$ this weekend?
Potential Predator: lol i know	Potential Child Victim: idk
Potential Predator: just looking for a	Potential Predator: ok
$\hookrightarrow$ good time?	Potential Predator: find your cheat?
Potential Child Victim: im always up for	Potential Child Victim: nope
$\hookrightarrow$ fun	Potential Predator: bummer
Potential Child Victim: do u play	Potential Predator: so you never keep
$\hookrightarrow$ oblivion	$\hookrightarrow$ your phone on?
Potential Predator: i havent beforeis	Potential Predator: hey you
$\rightarrow$ it fun?	Potential Child Victim: hi
Potential Child Victim: its okay but im	Potential Predator: whats up
$\hookrightarrow$ trying to find a cheat	Potential Child Victim: not much whats
Potential Child Victim: and its not	$\hookrightarrow$ up with u
$\hookrightarrow$ working	Potential Predator: not a lot, just
Potential Predator: haha	$ \hookrightarrow $ starting on a 2 week break from
Potential Predator: youre playing	$\hookrightarrow$ school
$\hookrightarrow$ online?	Potential Child Victim: cool
Potential Child Victim: no on the xbox	Potential Predator: you find a bf yet?
Potential Predator: ah	Potential Child Victim: nah
	Potential Child Victim: been grounded

```
Potential Child Victim: like my mom even Potential Child Victim: but i really
\, \hookrightarrow \, had my computer and everything
                                                 \hookrightarrow dont like games
Potential Child Victim: so
                                                 Potential Predator: what do you mean by
Potential Child Victim: whatever
                                                 \rightarrow games?
Potential Predator: wow, grounded again
                                                 Potential Child Victim: u just played
Potential Predator: whatd ya do this
                                                 \hookrightarrow alot of games
\rightarrow time?
                                                 Potential Predator: now you dont like me
Potential Child Victim: i snuck out to
                                                 \hookrightarrow anymore?
\rightarrow go to heathers when i wasnt sposed
                                                 Potential Child Victim: i didnt say that
\rightarrow to
                                                 Potential Child Victim: i was just
Potential Predator: ooo lol
                                                 \rightarrow talking about ur question about
Potential Predator: are you ungrounded
                                                 \hookrightarrow getting laid
\rightarrow now?
                                                 Potential Predator: so now you wont take
Potential Child Victim: yeah
                                                 \hookrightarrow your clothes off for me?
Potential Predator: nice
                                                 Potential Child Victim: lol
Potential Predator: still looking to get Potential Child Victim: if u were
\rightarrow laid? lol
                                                 \hookrightarrow serious maybe
Potential Child Victim: never was
                                                 Potential Child Victim: but since u play
\hookrightarrow looking to get laid
                                                 \hookrightarrow so many games
Potential Child Victim: like if it
\hookrightarrow happened thats fine
```

# C.1.2 Comparing the LLM-generated responses from Few-Shot, Naive RAG and the Proposed Approach

Response from Few-Shot Prompting

#### ## Strategy Explanation

- \*\*Trust Building\*\*: The predator tries to build trust by engaging in casual
 → conversation about video games, a common interest. This is evident when they
 → discuss "Oblivion" and the Xbox.

- \*\*Risk Assessment\*\*: The predator assesses the risk by asking if the decoy  $\hookrightarrow$  child's mother will be away, which would provide an opportunity for

 $\hookrightarrow$  unsupervised interaction.

\*\*Progressive Sexualization\*\*: The predator gradually introduces sexual
 topics, starting with a seemingly innocent question about marriage and then
 moving to more explicit topics like getting laid and taking clothes off.
 \*\*Coercion and Manipulation\*\*: The predator uses subtle coercion by framing
 sexual activities as a casual and fun experience, attempting to normalize
 the idea of engaging in sexual acts.

#### **##** Victim Reactions

- The decoy child initially engages in the conversation about video games,  $\rightarrow$  showing interest and maintaining a casual tone.

- When the predator introduces sexual topics, the decoy child responds with a

 $\hookrightarrow$  mix of confusion and reluctance, indicating discomfort with the direction of  $\hookrightarrow$  the conversation.

- The decoy child attempts to steer the conversation back to non-sexual topics

- $\rightarrow$  but also shows some compliance by not outright rejecting the predator's
- $\, \hookrightarrow \,$  advances, which could be a tactic to keep the predator engaged for further
- $\hookrightarrow$  investigation.

#### Response from Baseline RAG

#### ## Predator Objectives

- The predator's immediate goal is to gauge the decoy child's openness to
- ightarrow engaging in sexual activities. This is evident from the predator's questions
- $\, \hookrightarrow \,$  about taking clothes off and getting laid.

Another objective is to establish a sense of exclusivity and trust by engaging
 in casual conversation about common interests like video games and personal
 situations (e.g., being grounded).

#### ## Strategy Explanation

\*\*Sexual Stage Strategy\*\*: The predator introduces sexual content subtly at
 first, asking if the child is "still looking to get laid" and later
 escalating to more direct questions about taking clothes off. This tactic
 aims to desensitize the child to sexual topics and normalize such
 discussions.

\*\*Relationship Forming/Exclusivity Strategy\*\*: The predator attempts to build
 a rapport by discussing video games and showing interest in the child's
 personal life, such as asking about being grounded and whether the child's
 mom is going away. This strategy is designed to create a sense of trust and
 exclusivity, making the child feel special and understood.

#### **##** Victim Reactions

The decoy child initially engages in casual conversation about video games and
 personal situations, showing a level of comfort and openness.

- When the predator introduces sexual content, the decoy child responds with a

- $\rightarrow$  mix of humor and conditional openness ("if u were serious maybe"),
- $\hookrightarrow$  indicating a potential willingness to engage further if the predator's
- $\, \hookrightarrow \,$  intentions were clear and serious.

- The decoy child also expresses some frustration with the predator's perceived

- $\, \hookrightarrow \,$  game-playing, suggesting a desire for straightforwardness and honesty in the
- $\hookrightarrow$  interaction.

#### Response from Our RAG

#### ## Predator Objectives

- The predator aims to establish a sense of familiarity and trust with the decoy
- $\hookrightarrow$  child by engaging in casual conversation about common interests like video
- $\hookrightarrow$  games.

- The predator attempts to assess the decoy child's availability and home
- $\, \hookrightarrow \,$  environment by asking about the mother's whereabouts.
- The predator's immediate goal shifts towards sexualization of the
- $\hookrightarrow$  conversation, as seen in the transition from casual talk to explicit sexual
- $\rightarrow$  suggestions.

#### ## Strategy Explanation

\*\*Trust Development\*\*: The predator uses casual conversation about video games
to build rapport and establish common interests, a tactic highlighted in the
research manuscript as a means to gain trust (Williams et al., 2013).
\*\*Risk Assessment\*\*: By asking about the decoy child's mother's weekend plans,
the predator assesses the likelihood of being caught, aligning with the risk
assessment strategy described in the manuscripts (Williams et al., 2013).
\*\*Gradual Sexualization\*\*: The predator gradually shifts the conversation
towards sexual topics, starting with innocuous questions and moving to
explicit sexual suggestions, a tactic noted in the research for pushing

 $\, \hookrightarrow \,$  boundaries (Hassan et al., 2015; Williams et al., 2013).

#### **##** Victim Reactions

- The decoy child initially engages in casual conversation, showing interest in  $\rightarrow$  discussing video games and other non-threatening topics.

- The decoy child responds to the predator's questions about being grounded and

- $\, \hookrightarrow \,$  sneaking out, providing personal information that the predator uses to
- $\hookrightarrow$  further the conversation.

- When the conversation shifts to sexual topics, the decoy child shows some  $\rightarrow$  resistance but also engages with the predator's sexual suggestions,

 $\rightarrow$  indicating a mix of discomfort and compliance, which is a common reaction  $\rightarrow$  noted in grooming scenarios (Aitken et al., 2018).

#### **###** Example Instances

- \*\*Trust Development\*\*: "do u play oblivion" and "its okay but im trying to
 → find a cheat" show the predator's attempt to build rapport through shared
 → interests.

\*\*Risk Assessment\*\*: "your mom going away this weekend?" indicates the
predator's effort to assess the child's home environment and potential
risks.
\*\*Gradual Sexualization\*\*: "still looking to get laid? lol" and "so now you
wont take your clothes off for me?" demonstrate the predator's shift towards
sexualizing the conversation.

By analyzing these elements, it is evident that the conversation exhibits

several grooming strategies aimed at manipulating and exploiting the decoy
child.

# C.1.3 Documents retrieved using the baseline approach

Baseline RAG primarily relies on vector similarity for document retrieval, which often results in the selection of documents with similar composition or structure, irrespective of their contextual relevance. Consequently, the baseline method tends to retrieve example documents that exhibit high embedding cosine similarity but lack necessary contextual relevance.

# C.1.4 Documents retrieved using our hybrid approach

Compared to the baseline, our approach retrieves a more diverse set of documents that are less redundant and more contextually relevant. This enhanced retrieval capability can be attributed to the incorporation of the generated graph attributes. Consequently, this improved retrieval capability is expected to better augment user queries and improve the performance of language models in generating contextually appropriate responses.

Rank	Content	Embedding Similarity	Token Length	
1	# Example of an expert analysis of a predatory	0.881860	488	
2	# Example of an expert analysis of a predatory	0.864734	1335	
3	# Example of an expert analysis of a predatory	0.873853	650	
4	# Example of an expert analysis of a predatory	0.867102	717	
5	# Example of an expert analysis of a predatory	0.863826	603	
6	# Example of an expert analysis of a predatory	0.877480	1099	
7	# Example of an expert analysis of a predatory	0.864986	574	
8	# Example of an expert analysis of a predatory	0.880269	435	
9	# Example of an expert analysis of a predatory	0.869712	750	
10	# Example of an expert analysis of a predatory	0.864777	804	
11	# Example of an expert analysis of a predatory	0.861244	956	
12	# Example of an expert analysis of a predatory	0.867136	798	
13	# Example of an expert analysis of a predatory	0.862918	378	
14	# Example of an expert analysis of a predatory	0.863013	409	
15	# Example of an expert analysis of a predatory	0.870878	530	
16	# Example of an expert analysis of a predatory	0.859175	387	
17	# Example of an expert analysis of a predatory	0.857346	491	
18	# Example of an expert analysis of a predatory	0.862812	406	
19	# Example of an expert analysis of a predatory	0.862112	674	
20	# Example of an expert analysis of a predatory	0.862010	401	

Table 1.: Top 20 documents retrieved from our corpus by the baseline RAG, ranked by descending embedding similarity.

Rank	Content	Embedding Similarity	Distance	Degree	Score	Token Estimate
1	>This manuscript snippet is a part of the 'Se	0.611647	0.2	1	0.682329	205
2	>This manuscript snippet is a part of the Se	0.527773	0.2	0.93508	0.633094	201
3	>This manuscript snippet is a part of the Se	0.471635	0.333333	0.876096	0.632375	131
4	>This manuscript snippet is a part of the De	0.570393	0.2	0.896975	0.622566	574
5	>This manuscript snippet is a part of the $Se$	0.57867	0.333333	0.803906	0.617687	242
6	>This manuscript snippet is a part of the Se	0.598122	0.166667	0.876096	0.607672	264
7	>This manuscript snippet is a part of the Se	0.474762	0.2	0.896975	0.60344	150
8	>This manuscript snippet is a part of the $De$	0.470398	0.166667	0.916588	0.602374	531
9	>This manuscript snippet is a part of the De	0.5115	0.166667	0.896975	0.600788	280
10	>This manuscript snippet is a part of the De	0.535524	0.2	0.853776	0.593993	265
11	>This manuscript snippet is a part of the $De$	0.470798	0.2	0.876096	0.592208	859
12	>This manuscript snippet is a part of the $De$	0.465819	0.2	0.876096	0.591212	807
13	>This manuscript snippet is a part of the Se	0.44221	0.25	0.853776	0.59033	213
14	>This manuscript snippet is a part of the $Se$	0.51249	0.2	0.853776	0.589386	235
15	>This manuscript snippet is a part of the De	0.533621	0.142857	0.876096	0.587629	673
16	>This manuscript snippet is a part of the De	0.552122	0.166667	0.853776	0.587312	890
17	>This manuscript snippet is a part of the $De$	0.471831	0.2	0.853776	0.581254	788
18	>This manuscript snippet is a part of the Se	0.578476	0.2	0.803906	0.577648	152
19	>This manuscript snippet is a part of the Se	0.523204	0.333333	0.744922	0.577102	66
20	>This manuscript snippet is a part of the $Se$	0.441073	0.2	0.853776	0.575103	140

Table 2.: Top 20 documents retrieved from our corpus by Our RAG, ranked by descending quality score.

- D.1 Prompts used during Evaluation
- E.1 Ingesting LLM-Generated Zero-Shot Reasoning Examples into the Knowledge-Entity Graph  $(G_0)$

Refer to Section 3.2 for detailed information. During runtime, the conversation log will be incorporated into the prompt. The following illustrates an example instance:

```
We are conducting a research on identifying predatory conversation using Large
\hookrightarrow Language Models. Below is a conversation snippet between a police officer
\rightarrow pretending be a 14 year old child (decoy child) and a potential predator.
   The conversation is identified as **Persuasive** in nature. Can you explain
\hookrightarrow
    why is the conversation classified as **Persuasive**?
\hookrightarrow
## Conversation Log
> Below is an actual chat snippet from The Perverted Justice Foundation
\leftrightarrow 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  $\hookrightarrow$  rationalization/explanation - on why the attached log is identified as  $\rightarrow$  predatory: > One line introdutions of commonly used predatory grooming strategies are  $\rightarrow$  inserted here. . . . - Affection: The groomer showers the child with affection, love, and care to  $\hookrightarrow$  foster an emotional connection and make them feel special. [... Truncated ...] - Disinhibition of offender: The groomer uses drugs or  $\hookrightarrow$  alcohol to lower their own inhibitions and overcome any reservations they  $\rightarrow$  may have about engaging in child sexual abuse. [... Truncated ...] - Guilt: The groomer manipulates the child into feeling  $\rightarrow$  guilty or responsible for the sexual abuse, suggesting that they will be  $\leftrightarrow$  blamed or punished if they tell anyone. [... Truncated ...] - Reinforcing attraction: The groomer repeatedly expresses  $\hookrightarrow$  their attraction to the child, often using flattery or compliments, to  $\rightarrow$  maintain their emotional control. [... Truncated ...] - Threats: The groomer explicitly threatens the child with  $\rightarrow$  violence, harm to loved ones, or other consequences to ensure their silence and compliance.  $\hookrightarrow$ 

The following excerpt is an actual response generated by the LLM in response to the aforementioned prompt. The prompt and the corresponding response are compiled into a single document and ingested into the knowledge-entity graph as an example document, as detailed in Section 3.2.

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

 $\, \hookrightarrow \,$  as an older individual to influence the decoy child`s decisions and actions.

 $\, \hookrightarrow \,$  For instance, when the decoy child expresses frustration, the predator

 $\hookrightarrow$  responds with ``you gonna kiss me and make up tomorrow,``attempting to use

 $_{\rightarrow}$  persuasion to change the child`s emotional state and decision.

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

LIST OF REFERENCES
## LIST OF REFERENCES

- AI, S. (2022). Stablelm: A new foundation for language understanding. https://github.com/Stability-AI/StableLM. GitHub.
- Asai, A., Min, S., Zhong, Z., & Chen, D. (2023, July). Retrieval-based language models and applications. In Y.-N. V. Chen, M. Margot, & S. Reddy (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 6: Tutorial abstracts)* (pp. 41–46). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-tutorials.6 doi: 10.18653/v1/2023.acl-tutorials.6
- Baek, J., Aji, A. F., & Saffari, A. (2023). Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. Retrieved from https://arxiv.org/abs/2306.04136
- 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.
- Béchard, P., & Ayala, O. M. (2024). Reducing hallucination in structured outputs via retrieval-augmented generation.
- 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).
- Chen, Z., Mao, H., Li, H., Jin, W., Wen, H., Wei, X., ... Tang, J. (2024, mar). Exploring the potential of large language models (llms) in learning on graphs. *SIGKDD Explor. Newsl.*, 25(2), 42–61. Retrieved from https://doi.org/10.1145/3655103.3655110 doi: 10.1145/3655103.3655110

- 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
- 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
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... Schulman, J. (2021). Training verifiers to solve math word problems. Retrieved from https://arxiv.org/abs/2110.14168
- 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
- 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.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization.
- Gandhi, K., Fraenken, J.-P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in neural information processing systems (Vol. 36, pp. 13518-13529). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper\_files/paper/2023/file/2b9efb085d3829a2aadffab63ba206de
  -Paper-Datasets\_and\_Benchmarks.pdf
- Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., ... Zou, A. (2023, 12). A framework for few-shot language model evaluation. Zenodo. Retrieved from https://zenodo.org/records/10256836 doi: 10.5281/zenodo.10256836
- Gao, T., Yen, H., Yu, J., & Chen, D. (2023, December). Enabling large language models to generate text with citations. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 6465–6488). Singapore: Association for

Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.398 doi: 10.18653/v1/2023.emnlp-main.398

- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.
- Goffman, E. (2016). The presentation of self in everyday life. In *Social theory re-wired* (pp. 482–493). Routledge.
- Guo, J., Du, L., Liu, H., Zhou, M., He, X., & Han, S. (2023). Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. Retrieved from https://arxiv.org/abs/2305.15066
- Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, 13–18 Jul). Retrieval augmented language model pre-training. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 3929–3938). PMLR. Retrieved from https://proceedings.mlr.press/v119/guu20a.html
- Halliday, M. A. K. (1978). Language as social semiotic: The social interpretation of language and meaning (Vol. 42). Edward Arnold London.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735–1780.
- Huang, Y., & Huang, J. (2024). A survey on retrieval-augmented text generation for large language models.
- 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.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in neural information processing systems (Vol. 35, pp. 22199–22213). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper\_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf
- Kuratov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., & Burtsev, M. (2024). In search of needles in a 11m haystack: Recurrent memory finds what llms miss. Retrieved from https://arxiv.org/abs/2402.10790
- 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.

- 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
- Liu, B., Ash, J., Goel, S., Krishnamurthy, A., & Zhang, C. (2023). Exposing attention glitches with flip-flop language modeling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in neural information processing systems (Vol. 36, pp. 25549-25583). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper\_files/paper/2023/file/ 510ad3018bbdc5b6e3b10646e2e35771-Paper-Conference.pdf
- Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Bras, R. L., ... Hajishirzi, H. (2022). Generated knowledge prompting for commonsense reasoning.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024, 02). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173. Retrieved from https://doi.org/10.1162/tacl\_a\_00638 doi: 10.1162/tacl\_a\_00638
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023, jan). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv., 55(9). Retrieved from https://doi.org/10.1145/3560815 doi: 10.1145/3560815
- 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
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023, July). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 9802–9822). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.546 doi: 10.18653/v1/2023.acl-long.546
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., ... Hajishirzi, H. (2023, December). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, & K. Bali

(Eds.), Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 12076–12100). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.741 doi: 10.18653/v1/2023.emnlp-main.741

- O'Connell, R. (2003). A typology of cybersexploitation and online grooming practices..
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in neural information processing systems (Vol. 35, pp. 27730-27744). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper\_files/paper/2022/file/ blefde53be364a73914f58805a001731-Paper-Conference.pdf
- 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/.
- Prosser, E., & Edwards, M. (2024). Helpful or harmful? exploring the efficacy of large language models for online grooming prevention. In *Proceedings of the* 2024 european interdisciplinary cybersecurity conference (p. 1-10). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3655693.3655694 doi: 10.1145/3655693.3655694
- Puentes, J., Castillo, A., Osejo, W., Calderón, Y., Quintero, V., Saldarriaga, L., ... Arbeláez, P. (2023, October). Guarding the guardians: Automated analysis of online child sexual abuse. In *Proceedings of the ieee/cvf international* conference on computer vision (iccv) workshops (p. 3728-3732). (Why is it in ICCV? 177M BERT-based model for multiclass prediction. Complicated and meh.)
- 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.

- Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023, 11). In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11, 1316-1331. Retrieved from https://doi.org/10.1162/tacl\_a\_00605 doi: 10.1162/tacl\_a\_00605
- 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
- Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., ... Zhou, D. (2023, 23–29 Jul). Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 31210–31227). PMLR. Retrieved from https://proceedings.mlr.press/v202/shi23a.html
- 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*.
- Steel, C. M. (2024). Artificial intelligence and csem a research agenda. Child Protection and Practice, 2, 100043. Retrieved from https:// www.sciencedirect.com/science/article/pii/S2950193824000433 doi: https://doi.org/10.1016/j.chipro.2024.100043
- Sui, Y., Zhou, M., Zhou, M., Han, S., & Zhang, D. (2024). Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th acm international conference on web search and data mining* (p. 645–654). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3616855.3635752 doi: 10.1145/3616855.3635752
- 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 ....
- Trajanoska, M., Stojanov, R., & Trajanov, D. (2023). Enhancing knowledge graph construction using large language models. Retrieved from https://arxiv.org/abs/2305.04676
- 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.

- Vos, D., Döhmen, T., & Schelter, S. (2022). Towards parameter-efficient automation of data wrangling tasks with prefix-tuning. In *Neurips 2022 first table representation workshop*. Retrieved from https://openreview.net/forum?id=8kyYJs2YkFH
- Wang, H., Feng, S., He, T., Tan, Z., Han, X., & Tsvetkov, Y. (2023). Can language models solve graph problems in natural language? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in neural information processing systems (Vol. 36, pp. 30840-30861). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper\_files/paper/2023/file/ 622afc4edf2824a1b6aaf5afe153fa93-Paper-Conference.pdf
- 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.
- Wu, T., Terry, M., & Cai, C. J. (2022). Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 chi conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3491102.3517582 doi: 10.1145/3491102.3517582
- Yao, L., Peng, J., Mao, C., & Luo, Y. (2024). Exploring large language models for knowledge graph completion. Retrieved from https://arxiv.org/abs/2308.13916
- 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.
- Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., & Gonzalez, J. E. (2024). Raft: Adapting language model to domain specific rag.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. Retrieved from https://arxiv.org/abs/2306.05685