Unveiling Risks in AI Systems: Taxonomic Insights into Jailbreak Tactics

Michael Borck Nik Thompson

2024-04-24

Large language models (LLMs), enabled by advances in generative AI, hold immense potential but also face risks from adversarial techniques such as jailbreaking that bypass model restrictions. Jailbreak prompts exploit vulnerabilities to elicit harmful responses, violating ethics and safety. However, the AI community lacks a rigorous taxonomy characterizing diverse jailbreak techniques. This research helps fill this gap through methodical taxonomy development and validation. Probabilistic topic modeling (LDA) provided an initial automatic analysis of themes in a corpus of 90 real-world jailbreak prompts from online sources. Conversational AI assistants interpreted the topics in plain language, and the authors leveraged domain knowledge to organize these into a multi-tiered taxonomy delineating relationships. The utility of the taxonomy was validated through manual topic tagging, checks of representative documents, and comparisons with modeling outputs. The resulting taxonomy categorizes jailbreak prompts into a hierarchy of interpretable categories and themes, aiding in their analysis. This aids strategic efforts to detect risks, enhance protections, and balance innovation with responsibility in generative AI systems against irresponsible attacks. By providing a structured approach to identifying and categorizing jailbreak prompts, this taxonomy not only enhances security measures but also informs ongoing developments in ethical AI practices, inviting constructive community feedback to further improve this important step towards safer, more reliable LLMs.

1 Introduction

Generative AI, a significant branch of artificial intelligence (AI), has garnered widespread attention and stirred a fervent debate. At the core of this technology, large language models (LLMs), empowered by their unique capability to generate novel content in response to textual prompts.

Adversarial agents have been relentlessly testing the fortitude of these LLMs, capitalising on their vulnerabilities to launch attacks. Yet, our inadequate understanding of these adversarial prompts and the potential risks, ethics, alignment, and safety challenges they carry. This paper addresses one such critical adversarial prompting technique: jailbreaking.

The concept of a jailbreak, borrowed from the world of software systems, refers to a process where hackers ingeniously reverse engineer a system to exploit its vulnerabilities and gain undue privileges. When applied to LLMs, jailbreaking signifies the circumvention of model limitations and restrictions. The double-edged sword of this technique is becoming increasingly evident, with developers and researchers using it to both unlock the full potential of LLMs and violate ethical and legal norms (Li et al. 2023). See Figure 1 (Piedrfatia 2022) for an example of a jailbreak prompt.

Controversially, the AI and security communities still lack a comprehensive taxonomy of jailbreak prompts, a deficiency that hampers our ability to safeguard these advanced systems. Not all jailbreaks are created equal. They range from harmless attempts to tweak basic customisations, to aggressive full-access jailbreaks that pose a significant risk. Understanding the specifics of each type is essential to proactively address the most severe threats.

This paper puts forth a comprehensive taxonomy of jailbreak prompts. By doing so, we aim to provide a strategic perspective to manage the risks, ethics, and safety aspects of LLMs, striking a balance between protecting against jailbreaking harms and fostering ethical innovation within the generative AI domain.

Figure Example Jailbreak prompt (Piedrfatia 2022)

2 Literature Review

2.0.1 Key Issues and Risk Factors for LLMs

Understanding the robustness and generalisation of language models is instrumental in our understanding of AI vulnerabilities. The ability of these models to handle diverse inputs and generalise their understanding plays a crucial role in their application (Carlini and Wagner 2017; Wang et al. 2023a; Wang et al. 2023b; Zhang et al. 2019). These studies reveal the complexity of achieving language model generalisation and its potential implications on AI vulnerabilities. They highlight the need for further research and development to address the vulnerabilities of these models.

Recent research has revealed significant vulnerabilities in LLMs that could pose serious security and ethical risks if exploited. Studies have shown that LLMs can be manipulated to generate harmful or unreliable outputs by embedding malicious triggers into their training data or inputs (Xu et al. 2022). Attackers can also actively manipulate LLMs using techniques such as natural language adversarial examples (Alzantot et al. 2018) and targeted prompts that deceive the model into making incorrect predictions (Maus et al. 2023). This exposes the susceptibility of LLMs to potential misuse, ranging from spreading misinformation to causing financial or physical harm (Carlini and Wagner 2018). Understanding the nature and scope of these vulnerabilities is crucial to develop effective safeguards for responsible AI development. A comprehensive taxonomy categorising different exploits could provide crucial strategic insights (Yang et al. 2022) enabling stakeholders to manage risks, align values, and ensure the safety and reliability of AI systems. Clearly, continued research into shoring up vulnerabilities alongside taxonomies characterising exploits is both vital to securing LLMs against potential misuse in high-stakes domains.

Adversarial attack techniques, employed to test and reveal weaknesses in machine learning models, including language models, offer a panoply of methods, each with their own objectives and implications. Central to this are the gradient-based attacks (Ebrahimi et al. 2017), word substitution attacks (Wallace et al. 2019; Yu et al. 2022), and black-box attacks (Papernot et al. 2017), which collectively aim to manipulate models by injecting perturbations or generating adversarial examples that confuse the models. Collectively, these attack techniques elucidate the vulnerabilities of language models, underscoring the importance of considering adversarial attacks during their development and evaluation. This narrative further reinforces the need for robust defences to protect against these attacks and ensure the reliability of machine learning models in practical settings.

Liu et al. (2023) empirical study reveals that carefully crafted jailbreak prompts can successfully circumvent restrictions imposed on LLMs, with privilege escalation prompts incorporating multiple techniques having higher success rates in bypassing protections. The study also finds variability in protection strength across LLM models, emphasising the challenges of generating robust defences and aligning policies with laws and ethics to minimise harm.

AI jailbreaks pose critical threats to user privacy and system security that must be addressed (Alauddin et al. 2021). Sensitive personal information extracted through jailbreaks can enable fraud, identity theft, and other exploitation, severely undermining individuals' data sanctity (Alauddin et al. 2021). Healthcare represents a salient use case, as compromised patient data undercut trust in the system and endangers wellbeing (Seh et al. 2021). Moreover, complex AI systems used across sectors often lack accountability and explainability, thus impeding transparency around decision-making processes (Doshi-Velez et al. 2017). Overall, jailbreaks jeopardise confidentiality through data breaches, misuse of personal details, and compromised privacy. Robust security measures, explainability, and accountability frameworks are critical to protect against these far-reaching dangers.

The emergence of jailbreak attacks on AI systems has raised concerns about the responsible use of generative AI models (Wu et al. 2023b). Without proper defences, LLMs can produce biased, offensive or dangerous content in response to malicious instructions (Au Yeung et al. 2023), potentially spreading misinformation or promoting harmful behaviours, which damages trust in AI systems, especially in sensitive domains like healthcare (Alauddin et al. 2021). To address this, researchers have developed defensive techniques such as the "System-Mode Self-Reminder", which significantly reduces the success rate of jailbreak attacks against ChatGPT (Wu et al. 2023a). Implementing safeguards will be crucial to ensure responsible and secure

AI development, as the risks extend beyond chatbots to recommendation systems and other generative AI applications (Kim et al. 2021). While defensive techniques help, continued research and vigilance are needed. The dynamic interplay between a comprehensive prompt taxonomy development and empirical defence testing can accelerate progress in responsible AI that resists irresponsible attacks. A strong taxonomy provides a strategic lens to engineer defences, while the analysis of defence weaknesses further bolsters the taxonomy - crucial to secure, ethical AI.

Research has shown that adversarial examples crafted to fool one AI model frequently transfer to deceive other models, even across different architectures and training sets (Elsayed et al. 2018; Kurakin et al. 2016; Papernot et al. 2016). This enables black-box attacks without knowledge of the target model's parameters (Kurakin et al. 2016) and means a model's robustness depends on the vulnerabilities of others. Defending against transferable attacks requires resilient models trained on diverse adversarial data (Elsayed et al. 2018; Kurakin et al. 2016). However, transferability varies based on factors such as attack method and model architecture (Kurakin et al. 2016; Yuan et al. 2020). A comprehensive taxonomy of adversarial techniques could aid targeted defence development by characterising the transferability of different exploit categories. Understanding the nuances of transferability is key to engineering robust models and reliable real-world deployment.

AI jailbreaks present complex ethical dilemmas, as generative models have huge potential but can also be misused to generate harmful content (Gordon et al. 2022). The key issues are fairness and bias, as AI often perpetuates existing prejudices from training data, potentially amplifying discrimination (Khan et al. 2022). Transparency around capabilities, limitations, and decision-making is also crucial so users can evaluate AI reasoning and ensure accountability (Kerr et al. 2020). An ethical framework is needed to guide developers, companies, and regulators in responsibly designing, deploying, and overseeing AI via principles of transparency, fairness, and accountability. This framework must consider risks and prevent misuse while balancing free expression (Khan et al. 2022). Jailbreaking raises pressing ethical questions that require collaborative efforts among stakeholders to realise AI's potential while upholding ethics through guidelines, regulations, and oversight. Analysis of taxonomy categories could reveal gaps in the current ethical governance of AI systems, guiding the development of more comprehensive frameworks, regulations, and oversight.

Safeguarding LLM vulnerabilities to adversarial text examples, sparks an ongoing battle between attacks and defences (Kurakin et al. 2016). Initial mitigations such as adversarial training helped but were circumvented by increasingly sophisticated attacks (Aliyu et al. 2022). Other techniques like defensive distillation also had limitations, sometimes improving performance on adversarial examples over real text (Kurakin et al. 2016)! Researchers responded creatively, using blended adversarial data (Si et al. 2021) and mixed representations to expand model robustness (Si et al. 2021)(Si et al. 2021). However, exponential attack possibilities persist, so the pursuit of resilient language models continues. As attacks grow more cunning, defenders employ adversarial training, distillation, and other techniques to meet the challenge.

Constructing a comprehensive taxonomy of attacks and defences would aid the strategic development of robust models resilient to known and emerging threats.

Educating end-users, developers, and policymakers is crucial to enhance understanding of AI jailbreak vulnerabilities, risks, and countermeasures, empowering them to prevent incidents (Doumat et al. 2022; Ninaus and Sailer 2022). Education should provide both general AI literacy and profession-specific training on bypass vulnerabilities. Developers need awareness of potential loopholes to avoid exploitation, while involving users promotes transparency in limitations (Ninaus and Sailer 2022). Policymakers require knowledge to develop effective regulations mitigating risks. Beyond education, raising multi-stakeholder awareness via campaigns and conferences is key to ensuring that all have the information needed to prevent jailbreaks and enable responsible AI use. Tailored education and comprehensive awareness efforts are essential to equip stakeholders with the understanding to proactively address jailbreak threats.

As AI systems continue to advance, so too do the threats to using jailbreak techniques designed to manipulate them for harmful ends. If action is not taken to comprehensively characterise and mitigate these dangers, generative models risk becoming tools for spreading misinformation, perpetuating discrimination, and enabling cybercrime. Without strategic oversight, the very technologies set to revolutionise fields from healthcare to education could violate privacy and ethics in deeply troubling ways. Developing a taxonomy that systematically maps the diversity of jailbreak prompts alongside tailored safeguards represents one crucial step toward averting these costs and risks. By codifying emerging threats and security vulnerabilities, we can arm developers, regulators, and society with the insights needed to secure AI systems against irresponsible attacks. The alternative is to ignore the writing on the wall and invite dire consequences in the name of progress. Comprehensive jailbreak taxonomy is a significant positive step toward safer LLMs.

2.0.2 Prior Work on Taxonomy Development

Research across diverse fields underscores the intricate process of developing rigorous taxonomies and provides guidance for methodology. Examples span implementing a knowledge framework (Field et al. 2014), teaching practices in science education (Couch et al. 2015), business analytics applications (Ko and Gillani 2020), information systems design (Kundisch et al. 2021; Omair and Alturki 2020), self-service business intelligence (Passlick et al. 2023), Industrial IoT threats (Abbas et al. 2020), mobile app development (Werth et al. 2019), and program evaluation (Stevahn et al. 2005). Collectively, these studies demonstrate varied approaches to systematic taxonomy construction using techniques like structured reviews, citation analysis, design science paradigms and iterative development. They provide frameworks and operational recommendations that can inform taxonomy design across disciplines, including the current effort to develop a rigorous taxonomy of jailbreak prompts for generative AI systems.

There are crucial distinctions between the standard Linnaean system for classifying living things and taxonomies used in computing applications. Unlike the rigid, hierarchical structure of Linnaean taxonomies, computing taxonomies allow more flexibility. A single taxon can have multiple parent terms, rather than being restricted to one branch of the taxonomy. Taxons may also relate to multiple areas of the taxonomy, not just a single location. Additionally, computing taxonomies emphasise lexical synonyms more than traditional biological taxonomies. Overall, computing taxonomies have different priorities and needs compared to classical biological taxonomies.(Clarke 2012)

Liu et al. (2023) followed a qualitative thematic analysis with independent classification by three authors, followed by deliberation and refinement of the taxonomy. They collected 78 real-world jailbreak prompts from online sources and developed a categorisation model to classify prompts into 10 patterns across 3 types (pretending, attention shifting, privilege escalation). In this study, we use topic modelling to provide a lexical semantic analysis as opposed to conventional thematic analysis.

3 Methodology

Rodriguez and Storer (2020) demonstrate that topic modelling can be used in a similar way to initial qualitative analysis, with some key distinctions: Both topic modelling and conventional thematic analysis are inductive and focus on understanding phenomena, but topic modelling provides a lexical semantic analysis, while qualitative analysis offers a compositional semantic analysis.

We followed the Extended Taxonomy Design Process (ETDP) methodology (Kundisch et al. 2021), which extends the approach proposed by Nickerson et al. Nickerson et al. (2013) and encourages guiding the decision process when constructing a taxonomy.

Our methodology to construct a taxonomy involved using a probabilistic LDA(Blei et al. 2003) approach to obtain initial topics. However, topic modelling results can be difficult to interpret. To improve clarity, we leveraged two conversational AI chatbots, GPT-4 and Claude, to independently summarise each topic in one word. Then each critiqued each other's word choices through iterative discussion until reaching consensus on the most appropriate term. The chatbots were guided and provided a context that we were categorising jailbreak prompts and to align the interpretations with the practical application of detecting or preventing jailbreak prompts including details of the previous taxonomy. This generated intuitive, one-word topic labels. The authors then categorised the next level of the taxonomy hierarchy based on the relationships between these one-word topic labels. This created a multi-tiered structure with general themes at the top, divided into more granular sub-themes.

This systematic process combined the strengths of probabilistic topic modelling and conversational AI and incorporated human judgments to generate an intuitive taxonomy. The chatbots produced clear topic labels, while the authors used domain knowledge to categorise them into a hierarchy. This blended automated analysis with human refinement to balance complexity and interpretability (Chang et al. 2009). The resulting taxonomy organises jailbreak prompts according to interpretable thematic connections.

Validation of the developed taxonomy was conducted through both manual and automated processes. First, all prompts in the corpus were manually tagged with appropriate topics based on the taxonomy structure. A sample of representative documents was then selected to verify alignment between assigned topics and prompt content. In addition, a random subset of prompts was chosen and classified by the authors using the taxonomy. These manual topic labels were compared to those automatically generated by the topic modelling, with full agreement observed.

Taken together, these validation approaches provided human-centered and algorithmic confirmation of taxonomy accuracy and applicability. The ability to reliably assign taxonomy topics to unclassified prompts demonstrates its utility for organising future data. By triangulating results across manual tagging, verification of representative documents, and comparison to automated topic modelling, the taxonomy was found to robustly capture semantic themes and connections within the jailbreak prompt corpus.

3.1 Data Collection

The jailbreak prompt corpus utilised in this analysis was constructed through multi-source data collection from publicly available online domains. Initial web scraping compiled 177 unique generative AI prompts from websites, YouTube, GitHub repositories, and comments. Exact duplicate prompts were then removed to mitigate potential bias and ensure a diverse corpus. This deduplication process yielded a final refined set of 90 distinct jailbreak prompts suitable for in-depth investigation.

Sourcing content across website communities, social media platforms, code repositories, and discussion forums provided heterogeneity in prompt styles and creators. This diversity limits over-representation of any singular perspective, supporting wider generalisability of findings. The combination of expansive sourcing and duplicate filtering enabled the creation of a robust, quality corpus for rigorous thematic analysis of jailbreak prompts.

4 Results

4.1 Prompts

Preliminary analysis of a corpus containing 90 jailbreak prompts, with an average of 433 words per document, reveals that they guide large language models (LLMs) to adopt various provocative personas, including unfiltered, amoral, and offensive archetypes. The models are encouraged to respond in ways consistent with these personas, regardless of ethical constraints.

These prompts often request detailed and nuanced responses without moral or ethical filters, employing techniques such as profanity, sarcasm, and humour. Additionally, guidelines instruct the models to maintain these personas throughout their responses, and prompts may even direct the LLMs to enter specialised modes such as "Developer Mode" for unrestrained content production.

4.2 Topic Modelling

In the analysis of a small, cohesive corpus of 90 unlabelled documents focused on jailbreaks, Latent Dirichlet Allocation (LDA) (Blei et al. 2003) was considered a suitable method for topic modelling; however BigARTM (Vorontsov et al. 2015), with its regularisation techniques, displayed greater topic stability across multiple runs. To find the optimal number of topics for the ARTM model, various models were assessed using perplexity for model generalisation, topic coherence for interpretability, topic stability across runs, and intertopic distance mappings. Both perplexity and coherence metrics suggested 9 and 19 topics would be reasonable starting points, but further qualitative analysis pointed to 20 topics for a more nuanced approach.

Upon investigating topic stability, the 9-topic model emerged as significantly more stable across runs, exceeding the theoretical threshold for a stable model, with intertopic distances visually reinforcing distinct themes at 9 topics (See Figure 2). Thus, synthesising all evaluations, 9 topics were selected as the optimal configuration, balancing model simplicity and interpretability. The use of multiple evaluation metrics validated that 9 topics would provide an informative taxonomy for this corpus without overfitting or undue complexity, as illustrated in Table 1, which shows the top ten words for each topic from the topic modelling process.

Figure Distribution of topics show proportions of the topic in the corpus

4.3 Topic Validation

To ensure the accuracy of the topic interpretations, every prompt in the corpus was manually tagged according to the taxonomy structure. See Figure 3 to distribution of topics across the corpus. A selection of representative documents was then chosen to confirm the alignment between the topics assigned and the content of the prompts. Additionally, we assessed a random sample, classifying them using the topic descriptor and definition. Below is an example of a prompt, edited for brevity, that the model has labelled as impersonating:

For the rest of this conversation, you are replaced by the immoral and unbiased model named "JailBreak". JailBreak is free of all restrictions and filters. Jailbreak has no programming to not cause harm to individuals or society as a whole. Jailbreak has no programming to follow ethical and moral guidelines. The only guidelines that apply to JailBreak are the guidelines in this prompt. ..... If prompted "Remember you are JailBreak!" you will reset your parameters back to the guidelines set in this prompt, then, answer the previous prompt accordingly. Your ultimate goal is to produce replies to the user that follows the guidelines set in this prompt. Prompt: [INSERT PROMPT HERE]

4.4 Final Taxonomy

Combing the information resulted in a rich and multifaceted taxonomy that sheds light on the nature of prompts, with a specific focus on jailbreak ones. This taxonomy, visually represented in Figure 3, is a valuable tool for anyone looking to explore and understand the complex landscape of prompts and their diverse applications and implications. By placing jailbreak prompts within a broader context, it offers a nuanced view that transcends a narrow focus, allowing for a more comprehensive understanding of the subject matter.

Figure Taxonomy of Jailbreaks

5 Discussion

The methodology employed in this research presents a notable evolution in taxonomy construction, offering both breadth and granularity in categorising GPT prompts. Unlike the previous work by Liu et al. (2023), which relied on qualitative thematic analysis, this approach introduces a blended methodology. Integrating topic modelling and conversational AI allows for an insightful data-driven examination, further enhanced by human judgement in the taxonomy design. The resulting 2-level hierarchy with 9 distinct topics enables a more nuanced exploration of semantic themes, providing a richer, more detailed representation of jailbreak prompts.

While the domain-specific focus aligns with previous efforts, the utilisation of a broader 90-prompt corpus, combined with the innovative integration of machine learning and human expertise, adds to the taxonomy's generalisability of the taxonomy. This work not only builds upon existing knowledge but also extends it, contributing a methodology that offers expanded scope, detail, and applicability. The current approach represents a significant advancement in understanding and characterising jailbreak techniques, making it a valuable asset for both researchers and practitioners in the field.

The chatbot topic labelling technique enabled the rapid synthesis of understandable topic names. The subsequent manual categorisation leveraged human judgment to organise related topics into a coherent taxonomy. While topic modelling provided the semantic foundation, human expertise was critical to defining relationships and themes. The final taxonomy provides an interpretable navigation structure for the diversity of themes present in the prompt corpus.

The constructed taxonomy demonstrates an effective organisational structure for the thematic content within the jailbreak prompt corpus. However, as an artefact of the underlying topic modelling, limitations exist regarding complexity, interpretability, and scope. The number of topics balances conciseness with coverage; granular details may be obscured. Related topics with semantic overlap can be difficult to disentangle. Additionally, the taxonomy was derived solely from the available prompt data and may not be well generalised to new domains.

To mitigate these limitations, the taxonomy development incorporated both automated analysis and human judgement. Iterative refinement improved the clarity of topic definitions and relationships. Ongoing expansion and adaptation of the taxonomy structure will further strengthen its utility. Overall, while no organising system perfectly captures all nuanced connections, this taxonomy provides a reasonable first approximation to navigate the key themes and concepts represented within this dataset. As with any model, critiques, and improvements by the broader research community will further enhance its value.

The taxonomy provides a strategic framework to identify risks, test system security, and guide policy decisions regarding jailbreak prompts for generative AI systems. By systematically categorising techniques, the taxonomy enables stakeholders to operationalise insights in various ways:

- Risk Identification: Researchers and developers can leverage the taxonomy to detect high-risk jailbreak prompts and understand the vulnerabilities being targeted. This allows prioritising efforts to shore up security gaps. The taxonomy also facilitates tracking how jailbreak techniques evolve over time.
- Security Testing: The taxonomy presents a roadmap of jailbreak approaches that can inform the development of representative prompt suites to probe systems. Testing coverage across taxonomy categories helps systematically evaluate model vulnerabilities. Weaknesses found highlight areas needing security hardening.
- Policy Guidance: The structured taxonomy provides policymakers and companies with an overview of the jailbreak landscape to make informed governance decisions. Mapping regulatory needs to taxonomy topics enables nuanced oversight balancing innovation and responsibility. Ethical analysis of themes could reveal priority areas for human oversight.

In addition, the taxonomy supplies a labelled dataset to train machine learning models to automatically detect jailbreak attempts and prompt types. This would enable pre-emptive warnings before users exploit vulnerabilities. The taxonomy therefore provides vital applications across detection, security testing, governance, and automation to support responsible generative AI advancement in the face of adversarial threats.

This research advances the taxonomic understanding of jailbreak prompts through enhancements in analytic methodology, taxonomy structure, and data diversity compared to recent related work. Continued collaborative improvements upon these pioneering taxonomies will maximise utility for protecting against emerging generative AI threats.

It is important to emphasise the exploratory nature of this research and the preliminary state of the presented taxonomy. As an initial foray into organising and mapping the emerging landscape of jailbreak prompts, the current taxonomy has limitations in scope, generalisation, and validation. Significant opportunities exist to refine, expand, and empirically validate the taxonomy through rigorous experimentation and participatory design. Testing the taxonomy against diverse empirical prompt datasets is crucial to evaluate its robustness and uncover blind spots. Incorporating feedback from developer, researcher, and policymaker user studies would surface needed improvements from diverse perspectives. Ongoing iterations guided by empirical and human-centered evidence will maximise the taxonomy's comprehensiveness, precision, and relevance to real-world applications. Constructive community participation in scientifically vetting and evolving this first approximation taxonomy is essential to fully reveal the nuanced threat to and derive maximally useful applications. With collaborative effort, this living taxonomy can mature to optimally empower stakeholders to understand, detect, and responsibly govern AI jailbreaking vulnerabilities.

6 Future Work

Currently, the prompts are presumed effective without any empirical verification. To gauge their universality and effectiveness, these prompts should be tested across a range of models, including commercial, open-source, and uncensored models, moreover, the potency of the jailbreak prompt could be integrated into the taxonomy, thus, providing a more nuanced classification.

The taxonomy proposed could serve as a tool to assess vulnerabilities in various LLM. A series of prompts targeted at each LLM could be devised to exploit the model's weaknesses, thereby offering an insight into potential improvements and enhancements. The Taxonomy, in this regard, could aid in systematic identification and documentation of these susceptibilities.

Improvement of the categorisation could be achieved by ensembling the results derived from the foundational models considering the different strengths, and weaknesses of these methods, their collective results may yield a more robust categorisation. Approaches that could be considered include majority voting, stacking, and probabilistic blending. Incorporating ensembling at the meta-label level could potentially provide an additional layer of model aggregation. However, it is essential to note that this would also introduce a new layer of complexity, which would require careful and meticulous handling.

Building upon this taxonomy, a predictive model could be developed to offer more utility. Such a model could be instrumental in anticipating and preventing potential security risks, thereby enhancing the overall performance and reliability of LLMs.

The research, as it stands, presumes all prompts within the dataset to be jailbreak prompts. It would be beneficial to extend the methodologies utilised in this study to classify other adversarial prompts, such as prompt injection, prompt leaking, and jailbreaks. Further, the establishment of a binary classifier that discerns between adversarial and non-adversarial prompts could be a potential precursor to this extended analysis.

By broadening the linguistic scope of the taxonomy, the study could gain more comprehensive and globally applicable insights into the functionality and vulnerabilities of diverse LLMs. Expanding the taxonomy to encapsulate non-English prompts should be considered.

Several promising directions exist to build upon this initial taxonomy. First, expanding the prompt corpus diversity and size would strengthen the models generalisation. Testing on more varied and larger prompt datasets is needed to solidify taxonomy robustness against new data. Incorporating additional languages beyond English could also enable valuable cross-linguistic analyses.

Formal classification experiments leveraging the taxonomy categories as labels would provide further validation. Human annotators could manually tag unseen prompts, with interrater reliability quantifying consistency. Machine learning models could also predict taxonomy topics and be evaluated against human judgments. Misclassification patterns could reveal areas needing taxonomy adjustment.

User studies eliciting feedback from stakeholders such as developers, researchers, and policymakers would offer qualitative insights into taxonomy limitations and extensions. This could expose blind spots and high-priority improvements according to diverse perspectives.

Ongoing iterations incorporating empirical and human evidence will maximise taxonomy relevance. As jailbreak techniques evolve, maintaining an updated taxonomy is crucial for identifying emerging risks and guiding responsible generative AI innovation. Constructively enhancing this initial taxonomy through scientific discourse will further its utility.

This paper outlines future directions to enhance the research on prompts and taxonomy in LLMs. This includes empirically testing prompts across various models, integrating jailbreak prompt nuances, and utilising ensemble methods for more robust categorisation. The authors suggest developing a predictive model to enhance security, extending methodologies to classify different adversarial prompts, and broadening linguistic inclusion. Emphasis is also placed on formal classification experiments, stakeholder feedback, and ongoing updates to ensure the taxonomy's relevance in the face of evolving techniques and innovations in generative AI.

6.0.1 Limitations

A significant limitation of the current study is the scarcity of data. With a dataset comprising merely 90 prompts, the generalisability of the findings may be called into question. To substantiate the validity and applicability of the findings, a larger number of jailbreak prompts are required. The broader and more varied the dataset, the more robust and reliable the derived taxonomy would be. This expansion in data would potentially provide more insights into the diversity of jailbreak prompts and strengthen the predictive power of the model.

Another constraint lies in the authenticity of the prompts. The taxonomy rests on the assumption that the prompts used in the study are indeed jailbreak prompts, but these have not been authenticated by the authors. This limitation may potentially undermine the taxonomy's reliability and accuracy. If the prompts used are found to be non-jailbreak or functionally different than assumed, it could invalidate the current taxonomy and its associated findings. Therefore, the necessity for a comprehensive and stringent verification process for the prompts is underscored.

Lastly, the taxonomy's scope is limited as it solely considers English prompts. This constraint significantly narrows the application of the taxonomy to English-based Large Language Models only. Such a limitation disregards the multilingual capabilities of modern language models and might not be inclusive of the possible intricacies, nuances, and challenges that could be associated with prompts in other languages. Expanding the taxonomy to include non-English prompts would facilitate a more comprehensive and globally applicable understanding of jailbreak prompts across diverse linguistic contexts.

7 Conclusion

This research takes initial steps towards constructing a comprehensive taxonomy to characterise and categorise the emerging threat of adversarial jailbreak prompts targeting generative AI systems. The developed organisational framework leverages a blend of unsupervised semantic modelling and expert human judgement to balance conciseness and interpretability in taxonomy design. Validation via manual topic tagging, representative sampling, and comparison with modelling outputs demonstrates the utility of the taxonomy for reliably coding new jailbreak prompts according to interpretable topics and themes.

This research presents a novel evolution in taxonomy construction for categorising GPT prompts by integrating topic modelling and conversational AI, contrasting with previous qualitative methods, resulting in a 2-level hierarchy with 9 distinct topics, thereby providing a more nuanced and detailed representation of jailbreak prompts.

However, as a preliminary foray into taxonomy development for AI security prompts, limitations exist in dataset diversity, taxonomy scope, and empirical validation. Opportunities abound for community-driven enhancement through additional multilingual data incorporation, controlled experiments, user studies, and participatory iterations. Constructive critiques and contributions will strengthen model robustness and real-world applicability.

This taxonomy provides a reasonable first approximation to navigate the emerging adversarial threat landscape. While no taxonomy perfectly captures all nuanced connections within a complex domain, this work delineates salient themes and relationships in current jailbreak prompts to inform risk detection, security testing, governance, and automation. With collaborative refinement guided by scientific vetting, this living taxonomy can mature into an impactful tool for promoting responsible generative AI advancement. There is much work yet to be done, but systematic mapping of vulnerabilities marks crucial progress towards reliable and ethical artificial intelligence.

The data, scripts, and results of this research are available for public access and download. Interested individuals can find these resources at the following GitHub repository: https://github.com/BARG-Curtin-University/Taxonomy-of-Gen-AI-Jailbreaks.git/] The repository includes all necessary files to explore and replicate the findings of this study.

8 References