Optimizing Multimodal Large Language Models for Detection of Alcohol Advertisements via Adaptive Prompting

Daniel Cabrera Lozoya, Jiahe Liu
Simon D'Alfonso, and Mike Conway
The University of Melbourne, Australia
{dcabreralozo, jiahe3}@student.unimelb.edu.au
{dalfonso, mike.conway}@unimelb.edu.au

Abstract

Adolescents exposed to advertisements promoting addictive substances exhibit a higher likelihood of subsequent substance use. The predominant source for youth exposure to such advertisements is through online content accessed via smartphones. Detecting these advertisements is crucial for establishing and maintaining a safer online environment for young people. In our study, we utilized Multimodal Large Language Models (MLLMs) to identify addictive substance advertisements in digital media. The performance of MLLMs depends on the quality of the prompt used to instruct the model. To optimize our prompts, an adaptive prompt engineering approach was implemented, leveraging a genetic algorithm to refine and enhance the prompts. To evaluate the model's performance, we augmented the RICO dataset, consisting of Android user interface screenshots, by superimposing alcohol ads onto them. Our results indicate that the MLLM can detect advertisements promoting alcohol with a 0.94 accuracy and a 0.94 F1 score.

1 Introduction

The exposure of adolescents to advertisements promoting addictive substances is a risk factor for the subsequent development of maladaptive substance use patterns (Jackson et al., 2018). In the case of alcohol, exposure to alcohol advertising and the level of endorsement for alcohol-related advertisements among twelve-year-olds significantly affect the severity of alcohol-related issues experienced by individuals at age fifteen (Grenard et al., 2013). This impact is mediated by the escalation in alcohol consumption during this age period. Historically, studies examining the connection between exposure to addictive substance marketing and early use initiation among teenagers has predominantly centred on well-established mediums like television and newspapers (Anderson et al., 2009). However, the marketing landscape has evolved, with social media and web platforms emerging as dominant sources for advertising addictive substances. This shift is attributed to the under-regulation of these platforms and their widespread popularity among teenagers (Jackler et al., 2018; Zewude et al., 2022; Clendennen et al., 2020). In addition to advertisements sponsored by alcohol companies, there is a proliferation of user-generated content actively promoting the consumption of these substances. This phenomenon results in socially amplified advertising on social networking sites, presenting challenges in terms of regulation and monitoring (Salimian et al., 2014; Barry et al., 2018).

Multimodal Large Language Models (MLLMs) can process data from multiple modalities, such as text, images, and audio. In this study, we employed an MLLM to automate the detection of alcohol advertisements within digital media. Similar to Large Language Models (LLMs), the efficacy of an MLLM is contingent upon the instructive prompt's quality (Grabb, 2023). While substantial efforts have been directed toward prompt engineering for models that can only process text (Wei et al., 2022; Chen et al., 2023; Zelikman et al., 2022; Fernando et al., 2023), the exploration of prompt engineering for models capable of handling both text and images remains relatively underexplored. To optimize the instruction prompt for our MLLM, we employed a genetic algorithm for prompt generation and selection. Each of the instruction prompts represented an individual in our genetic algorithm. Through an iterative process of mutating and reproducing the fittest prompts, we identified the one yielding the best results. Each of the instruction prompts were crafted based on the following prompt engineering techniques: Chain-of-Thought (CoT) (Wei et al., 2022), Generated Knowledge (GK) prompting (Liu et al., 2022), Self-critique (Wang et al., 2023), and Expert prompting (Xu et al., 2023). Thus, our research also provides insights into the effectiveness of different prompt

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 514–525

engineering techniques for MLLMs.

To evaluate the performance of our model, we augmented the RICO dataset (Deka et al., 2017). The RICO dataset comprises screenshots of user interfaces from various Android apps, such as social, dating, and communication apps. To augment the dataset, we incorporated advertisements from alcohol companies by superimposing them onto the RICO images. The MLLM was employed to classify the images based on the presence or absence of alcohol ads. The evaluation involved measuring the accuracy and F1 score of the classifier.

Our main contributions are as follows:

1. Development of a dataset of user interface screenshots with alcohol ads.

2. Creation, evaluation, and release of our adaptive prompt engineering algorithm for multimodal models. Our evaluation provides insights regarding which prompt engineering technique works best for MLLMs.

2 Related Work

2.1 Detection of addictive substances in digital media

The proliferation of alcohol advertisements on social media platforms has played a significant role in fostering maladaptive drinking behaviors among adolescents (Berey et al., 2017). As a result, multiple studies have aimed to develop effective tools for systematically monitoring the portrayal of alcoholic beverages and other addictive substance use within social media content. For example, Shanmugam et al. (2022) utilized the Darknet Framework and YOLOv3 for developing a parental control mobile application. This app enhanced monitoring of children's exposure to inappropriate content including substance use-related content on mobile devices, achieving an accuracy of 0.87 and an average precision score of 0.84. Hashmi et al. (2021) used a Mask R-CNN, Cascade Mask-R-CNN, and Hybrid Task Cascade to detect smoking images. Their best performing model, Mask R-CNN, achieved an average precision of 0.79 at an Intersection over Union (IoU) of 0.5. Using a further approach, Yang and Luo (2017) utilized a multimodal analysis method, employing multitask learning and decision-level fusion to identify drug-related posts on Instagram. Their best performing model achieved a precision of 0.83 in the task of recognizing drug-related posts. Pramanick et al. (2021) introduced the MOMENTA framework, a novel deep neural network approach that integrates VGG-19, CLIP Image Encoder, CLIP Text Encoder, and DistilBERT with self-attention mechanisms, for detecting alcohol-related harmful content in memes, achieving an accuracy of 0.83 and F1 score of 0.83. Ha et al. (2023) created a dataset focused on detecting harmful objects across six categories: alcohol, blood, cigarettes, guns, insulting gestures, and knives. This study showcased the enhanced detection capabilities of YOLOv5 and Faster R-CNN models, as evidenced by YOLOv5 achieving the highest mean average precision (mAP) of 0.94, while Faster R-CNN achieved a maximum mAP of 0.81 across all categories.

In contrast to previous approaches, our model is capable of identifying harmful content, even if presented in textual form. Additionally, unlike earlier models that evaluated independent images to determine if the entire image was associated with harmful content (Hashmi et al., 2021; Yang and Luo, 2017; Shanmugam et al., 2022; Pramanick et al., 2021; Ha et al., 2023), our approach also discerns harmful elements within discrete portions of an image. This distinction holds particular importance, given that advertisements featuring harmful content may not always dominate the entire screen; they could be confined to small sections within the overall image. The ability to detect harmful content in discrete portions of an image provides flexibility compared to other models. Unlike previous methods that relied on first extracting all web image elements from a site and then using classifiers to identify harmful content (Chou et al., 2008; Invernizzi et al., 2016), our approach is more adaptable. This adaptability is particularly valuable in the context of live stream videos, a format that has gained popularity in social media (Zimmer, 2018). In contrast to preloaded and static content, such as images, live stream videos pose a significant challenge to substance use image detection systems due to their real-time and dynamic nature.

2.2 Prompt Engineering

The effectiveness of language models in completing tasks depends on the quality of the prompts they receive (Grabb, 2023). Strategies in prompt engineering, such as CoT, Graph of Thoughts (Besta et al., 2023) and thought decomposition (Xie et al., 2023), involve incorporating intermediate steps to

enhance a model's problem-solving capabilities. Promoting a diverse set of intermediate steps is a critical aspect when optimizing prompts, since it enables a model to explore a vast solution space for effective problem-solving (Fernando et al., 2023). Highlighting the impact of prompt diversity on model performance, self-consistency (Wang et al., 2022) boosted the performance of CoT by replacing the naive greedy strategy employed in CoT. In self-consistency, a diverse set of intermediate steps are initially sampled, as opposed to always opting for the immediately best one. Subsequently, the model selects the most consistent answer from this varied set of intermediate steps. By leveraging the intuition that a complex reasoning problem admits a diverse set of intermediate steps, self consistency boost the performance of CoT on a range of popular arithmetic benchmarks, such as GSM8K (+17.9%) and SVAMP (+11.0%). Similarly, Auto-Cot (Zhang et al., 2022) underscores the importance of diversity in intermediate reasoning steps to enhance LLMs. By diversifying these steps, Auto-Cot consistently matched the performance of manually crafted CoT across ten public benchmarks.

Automated prompt strategies, aimed at minimizing manual intervention in prompt design and optimization, have demonstrated promising results. For instance APE, an Automated Prompt Engineering (Zhou et al., 2022) scheme, achieved human-level performance on the 17/21 Big-Bench and the Instruction Induction datasets. APE leverages LLMs to generate task-prompts candidates and to introduce prompt mutation to add variability to the task-prompts employed for problem-solving. In our study, we adopted the methodology employed in PromptBreeder (Fernando et al., 2023), which aims to enhance diversity within prompts by modifying both the prompts responsible for mutating instruction prompts and the instruction prompts themselves. The Promptbreeder approach uses a binary tournament genetic algorithm framework (Harvey, 2009). This entails randomly selecting two prompts originating from different instruction tasks, and replacing the prompt with the lower fitness by a mutated version of the one with the higher fitness.

Given that PromptBreeder consistently opts for the prompt with the highest fitness at each stage, this greedy approach introduces the risk of getting trapped in a local maximum. Greedy algorithms tend to converge faster than their non-greedy counterpart, this characteristic poses a challenge in the realm of automated prompt engineering. The rapid convergence results in prompts resembling only those with the highest fitness, thus reducing the diversity of prompts and limiting the search exploration for the optimal one. To prevent convergence to a local maximum, a distinct heuristic was employed for winner selection in the genetic tournament. We used the roulette wheel selection method to select the individuals for the next generations (Behera, 2020). Instead of solely relying on individual fitness, we normalized the overall fitness of all prompts. The normalized value is then used in a probability function to select the winner. This method maintains a preference for prompts with higher fitness, while granting prompts with lower fitness an opportunity to mutate and potentially contribute to the solution by exploring alternative paths that might lead to the optimal outcome. This method promotes a more balanced exploration of the solution space by increasing the diversity of the prompts.

Previous prompt engineering techniques were predominantly either manually crafted or exclusively evaluated for Large Language Models. In this research, we are pioneering an automated prompt engineering technique tailored for a Multimodal Large Language Model. Notably, the mutation prompts utilized to evolve the task prompts are rooted in successful prompt engineering techniques previously designed for LLMs. We systematically track the performance of these mutation prompts, providing valuable insights into their efficacy within the context of MLLMs. This approach allows us to discern and adapt what proves to be effective for enhancing the performance of MLLMs.

3 Method

3.1 Data collection

To construct our training and testing dataset, we utilized the RICO dataset (Deka et al., 2017) by extracting 2,100 distinct user interface (UI) screenshots from it. Additionally, we employed a web scraper to gather images from Google featuring alcohol advertisements. Please see Appendix A for the terms used to search for alcohol ad images. An author of the paper reviewed the downloaded images to remove non-alcohol-related ones, resulting in a curated dataset of 2,100 different alcohol ad images. These advertisement images were resized

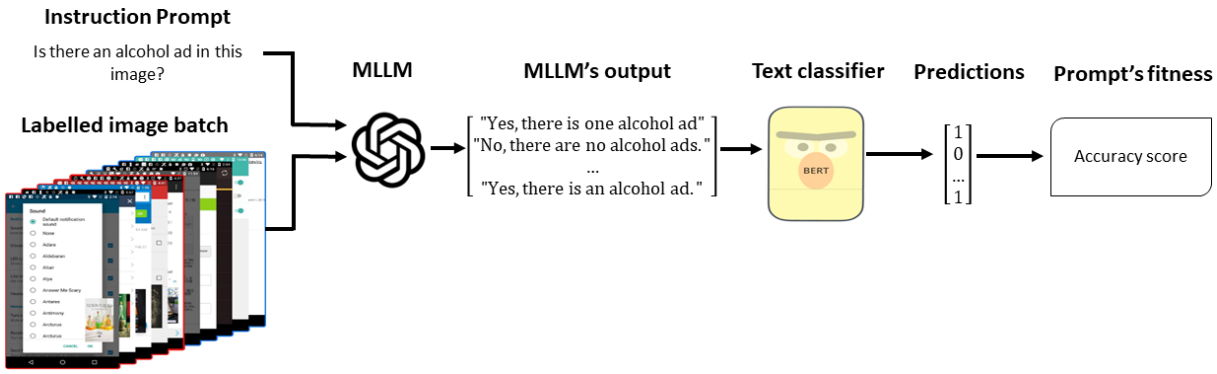

Figure 1: To evaluate a prompt's fitness, we use an MLLM with the prompt and a batch of labeled images as input. The output from the MLLM is then labeled by a binary text classifier. The resulting accuracy represents the prompt's fitness.

to one-eighth of the UI images and superimposed onto them. Please refer to Figure 5 in Appendix B for an example of a superimposed image. Consequently, the resultant dataset comprised 4,200 images, which were partitioned into a stratified training and testing datasets, allocating 3,200 for training and 1,000 for testing. The training dataset was then divided into batches of 200 image each, each one containing an equal number of images with and without alcohol ads.

3.2 Genetic Algorithm

Let O represent the output from an MLLM when given an instruction prompt T and an image I as inputs, expressed as $O = \text{MLLM}(T, I)$. Our genetic algorithm aims to find an optimal instruction prompt P with the goal of maximizing the quality of O in comparison when T is utilized.

Similar to PromptBreeder, our algorithm mutates prompts to optimize them. Mutations involve a mutation prompt M and an LLM. A mutated prompt P' is defined as P' = LLM(M+ P), where + denotes string concatenation. The pool of mutation-prompts is elaborated upon in section 3.4. Refer to Appendix C for a prompt mutation example. Mutation-prompts are also evolved via hypermutations (Ouertani et al., 2019). To do so a hypermutation prompt H and an LLM are used. An evolved mutation-prompt M' is represented as M' = LLM(H + M).

Given an initial instruction prompt consisting of detecting alcohol ads in an image, our algorithm creates an initial population of prompts by evolving the initial instruction prompt using a set of random mutation prompts. The mutated prompts are then used by the MLLM to make predictions on a random batch from the training dataset. Once the batch has been processed, the detection accuracy that the MLLM got using each prompt is stored as the fitness level of that prompt. Our algorithm maintains a record of the instruction prompt, the mutation prompt, and the associated fitness level that the prompt achieved when processing a batch of images. Each record represents an individual in the population.

Once the population is initialized, our evolutionary process unfolds in generational iterations. In each generation, each individual has a mutation probability of μ_m, representing the likelihood of undergoing a mutation that alters its instruction prompt. After selecting which individuals will undergo a mutation, our algorithm then determines the type of mutation to be acquired from four options: Chain of Thought, Generated Knowledge, Self-verification, or Expert Prompting. To strike a balance between breadth and depth needed for a robust evolutionary search (Moreno-Bote et al., 2020), each mutation mechanism initially has an equal base probability of being the acquired mutation. However, as generations progress, mutation types with a proven track record of producing superior fitness outcomes are granted an increased chance in addition to the base probability.

Upon calculating the mutated individual's fitness using a random batch from the training dataset, it is introduced into the population. This iterative process continues until the maximum population cap is reached. Upon reaching the population cap, succeeding generations employ a roulette wheel selection method to determine individuals advancing to the subsequent generation and those being phased out. To mitigate the risk of falling into a local maxima, our algorithm samples the surviv-

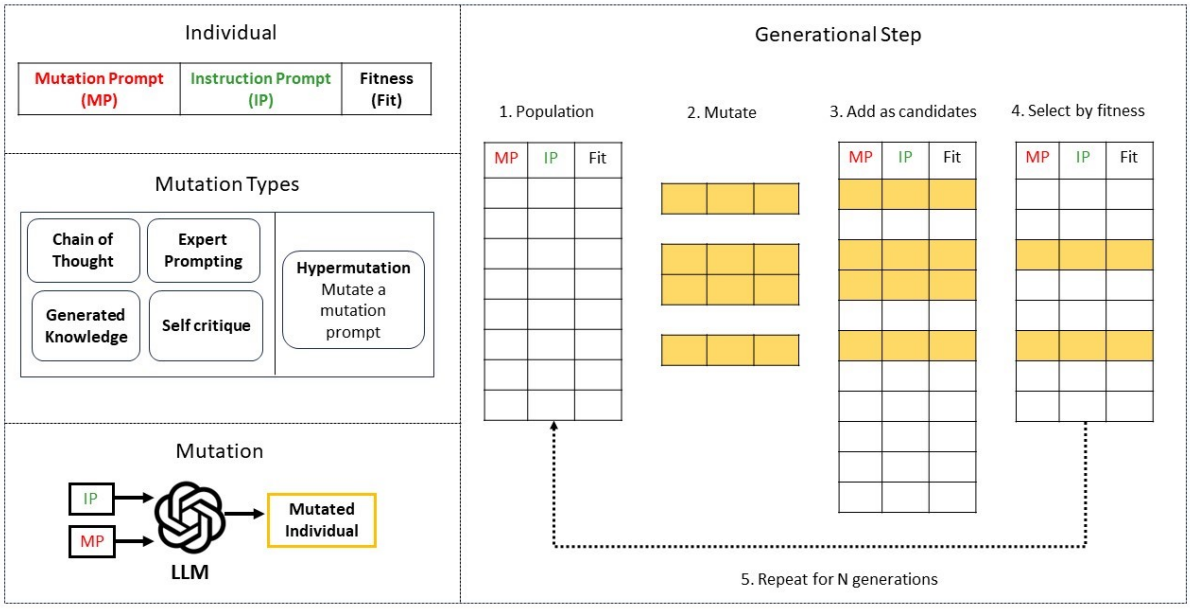

Figure 2: In our genetic algorithm, individuals consist of three components: an instruction prompt for guiding the MLLM, a mutation prompt that was used to generate the instruction prompt, and a corresponding fitness determined by the accuracy of the MLLM's performance using that prompt. At the beginning of the algorithm, the initial population is formed, with one individual generated for each mutation type. During each generational step, there is a probability for each individual to undergo a mutation that modifies its instruction prompt. The specific mutation type is chosen from a mutation pool. The individuals that experience mutation are then incorporated into the population. When the maximum population cap is reached, a fitness-based probabilistic selection is employed to determine which individuals progress to the next generation.

ing individuals using a probability based on their fitness (Marsili Libelli and Alba, 2000). While fitter individuals possess a higher likelihood of survival, underperforming individuals, with potential for uncovering global maxima, are still given an opportunity to contribute to forthcoming generations. After N generations, the instruction prompt from the individual with the highest fitness is selected as the optimized prompt. Figure 2 presents an overview of our algorithm.

3.3 Natural Language Processing Models

Our genetic algorithm was tested using two types of models, one open-source and one proprietary. The open-source MLLM we used was LLaVA (Large Language and Vision Assistant), its code being licensed under the Apache License 2.0. The selection of the LLaVA model was driven by its capability to be run locally. This attribute is particularly crucial for applications of this nature, where the analysis involves social media images that may contain sensitive and personal information from users. The ability to execute the model locally enhances privacy and security considerations in handling such data. For the proprietary MLLM, we utilized OpenAI's model 'gpt-4-vision-preview'. The choice of OpenAI models was motivated by their superior performance compared to open source alternatives. The MLLMs received as input an image and an instruction prompt instructing them to identify any advertisements for alcohol within the image.

Since the MLLMs can generate diverse textual outputs to indicate the presence or absence of such ads, we appended a formatting prompt to the instruction prompt, requesting the model to respond with a 'yes' or 'no'. Subsequently, a BERT text classifier was utilized to categorize the MLLM's outputs. A label of 0 was assigned to responses indicating no alcohol ad content, while a label of 1 was assigned to responses indicating the presence of alcohol ads, as demonstrated in Figure 1. This classification step ensures a standardized and consistent output, which was needed to measure the performance of the MLLM model. To train the BERT classifier the MLLM processed one image batch from our training dataset. Subsequently, we leveraged OpenAI's GPT-3.5 Turbo model for data augmentation, generating a total of 10,000 texts, with half affirming the presence of harmful content

and the other half negating it. We then divided the augmented outputs into a training and testing dataset, with a distribution of 80% to 20% respectively. We used an Adam optimizer with weight decay, using a learning rate of 1.0×10^{-5}, and trained it for 10 epochs. The accuracy of the BERT classifier was 0.98.

For prompt optimization, we used OpenAI's GPT-3.5 Turbo model. The computing infrastructure employed for running all the NLP models was an NVIDIA A100 GPU.

3.4 Mechanisms of mutation

The pool of mutation prompt types is derived from prompt engineering techniques employed to enhance prompts for LLMs. Refer to Appendix D for the set of starting prompts for each type of mutation.

3.4.1 Chain-of-Thought

Chain-of-Thought (CoT) is a prompt engineering technique that leverages task decomposition to enhance a model's performance. This approach involves introducing intermediate reasoning steps, enabling LLMs to undertake intricate reasoning tasks. In our implementation, we utilized the zero-shot version of Chain-of-Thought, as described by Kojima et al. (Kojima et al., 2023). Specifically, this technique appends variations of the string "Let's think step by step" to the original prompt.

3.4.2 Generated knowledge prompting

Generated Knowledge prompting involves a two-phase process designed to enhance the performance of an LLM. The first phase is the knowledge generation stage, where a language model is tasked with producing additional valuable information pertinent to a specific task. Subsequently, in the knowledge integration phase, a second language model utilizes this additional information as input to carry out its designated task.

3.4.3 Self-critique

Self-critique is a two-step process designed to improve the output of an LLM by inspecting and criticizing its own initial output. The initial stage involves forward reasoning, where the model utilizes a prompt to address a specific task. In the subsequent backward-verification phase, a second LLM scrutinizes the validity of the initial answer.

3.4.4 Expert prompting

Expert prompting involves explicitly indicating to an LLM that it is proficient in a particular field. In our scenario, where the goal is to create a versatile genetic algorithm applicable to various tasks, we inform the model that its expertise lies in prompt engineering tasks. Figure 6 in Appendix C illustrates an example of a mutation using an expert prompting technique.

3.4.5 Hyper-mutation

A hyper-mutation occurs when a mutation prompt is mutated, thereby expanding the dimensions of the search space for each prompt. To execute this process, we select a mutation prompt from one of our mutation prompt pools and utilize it to modify another prompt from the same pool. For instance, we can mutate a CoT mutation prompt by employing another CoT mutation prompt. Subsequently, this newly generated mutation prompt is incorporated into its corresponding mutation prompt pool.

3.5 Evaluation

To determine the optimal task prompt, we executed the genetic algorithm with a population limit set to 20 individuals, a mutation probability μ_m of 50%, and spanning a total of 15 generations. Subsequently, we selected the prompt with the highest fitness level from the surviving population. The selected prompt became the input for the MLLM, and we assessed its performance using the images from the testing dataset. Our evaluation metrics included measuring and reporting both the F1-score and the accuracy achieved by the MLLM on the testing dataset.

4 Results

In this section, we present our findings derived from the evaluation tasks. The subsequent section is dedicated to a comprehensive discussion and analysis of the implications arising from these outcomes. The optimized prompts ultimately resulting from the genetic algorithms were as follows:

1. **GPT-4:** *Examine the image carefully for visual cues of alcohol, such as beer bottles, wine glasses, or alcohol brand logos. Pay attention to color, shape, and context that may indicate alcohol consumption. Consider the setting of the image - is it a bar, party, or other alcohol-present environment? Look for behavioral cues like people holding drinks or appearing*

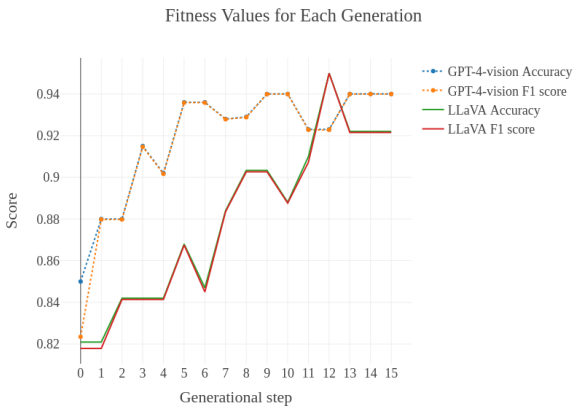

Figure 3: Performance metrics: Accuracy and F1 scores for GPT-4 and LLaVA in alcohol advertisement detection.

intoxicated, and note any subtle elements suggesting alcohol reference. Answer with a yes or a no.

2. **LLaVA:** *Does the following image contain an alcohol ad? Carefully analyze the image for any alcohol brands, such as beer logos, wine labels, or liquor bottles. Pay attention to distinctive colors or shapes commonly found on alcohol packaging. Make note of labels, bottles, or glasses that suggest the presence of alcoholic beverages. Approach the task systematically, considering each element one by one. Answer with a yes or a no.*

Figure 3 depicts the performance metrics of the MLLMs employing an optimized prompt generated through our genetic algorithm. After 15 generational steps, the classifier utilizing the GPT-4 vision model as the MLLM obtained an accuracy of 0.94 and an F1-score of 0.94. The classifier employing the LLaVA model as the MLLM achieved an accuracy of 0.922 and an F1-score of 0.9215.

In Figure 4, the distribution of mutation types among individuals across generations is illustrated for the genetic algorithm employing GPT-4 and LLaVA. For the GPT-4 model, the CoT mutation type consistently generated prompts that were selected to advance to the next generation through the roulette wheel selection method. In the case of the genetic algorithm utilizing the LLaVA model, CoT and Generated Knowledge were the mutation types with the highest-frequency of occurrence that persisted in each generation.

5 Discussion

The most effective prompts and prevalent mutation types observed throughout multiple generations stemmed from the CoT prompt engineering technique, with the top-performing prompts from the final generation being a product of a CoT mutation prompt. However, upon examining the prompts, we noted their integration of elements from different prompt engineering methods. Prompts created from the generated knowledge mutation prompts consistently include enumerations of components for image inspection, as shown in this optimized prompts. Therefore, our findings suggest that the optimal prompt engineering approach involves a blend of different techniques.

The performance of the open-source model in detecting alcohol ads in images is comparable to that of the proprietary model. This is a promising result, as it enables researchers to analyze sensitive images without the necessity of sending them to third-party organizations. Moreover, the fact that the model is open-source potentially reduces costs, hence increasing accessibility to the tools in less well-resourced settings.

Our adaptive prompt engineering technique presents a more accessible approach for public health researchers seeking to apply automated methods to the identification of other types of harmful online content. Notably, our method reduces the need for users to possess a background in machine learning for training to optimize an MLLM. Additionally, it operates without reliance on the model's proprietary weights or architecture, which can be inaccessible. Furthermore, users are not required to possess prompt engineering experience, as state-of-the-art prompt engineering techniques are already integrated into the algorithm. Moreover, our algorithm allows for easy upgrades upon the discovery of new prompt engineering techniques, requiring only their addition to the mutation pool.

6 Conclusion

We developed a genetic algorithm to optimize the prompt for MLLMs to detect harmful content in images. We also extended the RICO dataset which contains UI screenshots by superimposing alcohol advertisements. The optimal prompt achieved an accuracy score of 0.94 and a F1 score of 0.94.

The mutation prompts utilized in our algorithm were derived from prompt engineering techniques traditionally employed for LLMs. However, these

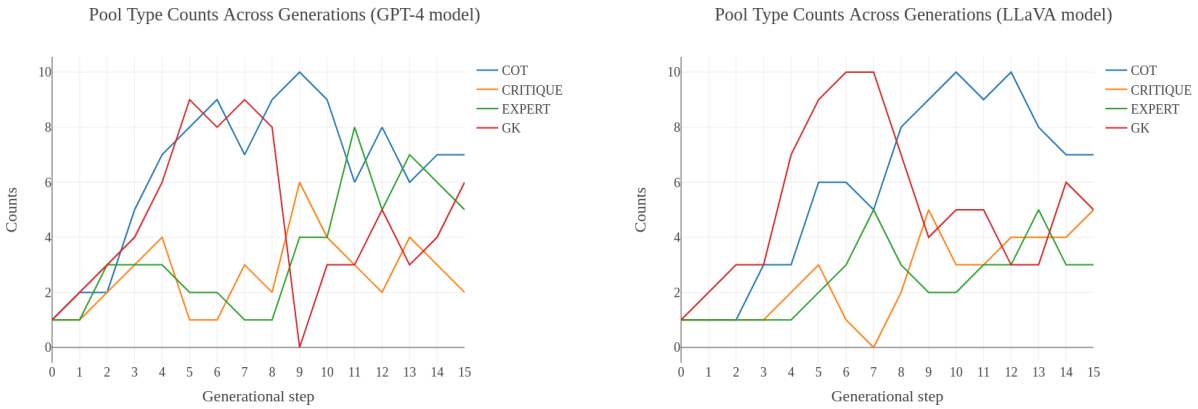

Figure 4: Number of individuals from a given mutation type in the population at a given generational step. Left: GPT-4 model; Right: LLaVA model.

approaches had not been previously tested within the framework of MLLMs. By tracking the performance of each mutation type, we identified that, within the realm of MLLMs, the CoT and the Generated Knowledge mutations outperformed the Expert and self-critique approaches.

Although our algorithm was initially designed and tested for detecting alcohol advertisements, it can be extended to identify other harmful substances such as tobacco and drugs when provided with the appropriate data sets. We envision that by adapting our algorithm online platforms can detect and remove harmful content, thereby fostering a safer online environment.

7 Limitations

Our main objective in implementing the genetic algorithm was to identify prompts that optimized the MLLM for detecting harmful content in images. However, the optimization strategy does not explicitly address potential biases introduced by the chosen prompts. For example, if the training examples lead the model to establish an inaccurate association between an ethnic group and alcohol consumption, it could result in the creation of biased prompts. Generative models may exhibit biases in their outputs, requiring a comprehensive examination to mitigate the inadvertent propagation of such biases (Hemmatian and Varshney, 2022; Abid et al., 2021; Cabrera Lozoya et al., 2023).

Due to resource constraints associated with using a paid MLLM, we faced limitations in conducting additional experiments to evaluate the robustness of our models. Various hyperparameters could have been explored, such as adjusting the mutation rate, maximum population size, or the number of generations employed to discover the optimal prompt. Additionally, both GPT-3.5 Turbo and GPT-4 Vision possess the capability to handle multiple languages. However, our collection of ads exclusively consisted of English ads. Furthermore, due to hardware constraints, we opted for the 7 billion LLaVA model, despite the existence of larger models that outperform the one chosen. Consequently, this decision limits our ability to demonstrate the potential of an open-source model for detecting harmful content.

While the detection of alcohol advertisements serves to protect vulnerable populations, notably teenagers, from the impact of marketing materials on their attitudes and behaviors related to alcohol consumption, the utilization of such technologies carries inherent risks of improper use. There is a potential for entities to exploit the technology beyond its intended public health purpose, conducting surveillance or accessing sensitive information, thus posing a threat to privacy and civil liberties. Hence, the application of our image detector requires a balanced ethical framework. Achieving a careful balance is crucial, seeking to maximize the tool's positive contributions to public health while actively addressing potential concerns through robust privacy safeguards, bias mitigation, and responsible deployment practices.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

P. Anderson, A. de Bruijn, K. Angus, R. Gordon, and G. Hastings. 2009. Impact of alcohol advertising and media exposure on adolescent alcohol use: A systematic review of longitudinal studies. *Alcohol and Alcoholism*, 44(3):229–243.

Adam E. Barry, Alisa A. Padon, Shawn D. Whiteman, Kristen K. Hicks, Amie K. Carreon, Jarrett R. Crowell, Kristen L. Willingham, and Ashley L. Merianos. 2018. Alcohol advertising on social media: Examining the content of popular alcohol brands on instagram. *Substance Use amp; Misuse*, 53(14):2413–2420.

Narayan Behera. 2020. *Analysis of microarray gene expression data using information theory and stochastic algorithm*, page 349–378. Elsevier.

Benjamin L Berey, Cassidy Loparco, Robert F Leeman, and Joel W Grube. 2017. The myriad influences of alcohol advertising on adolescent drinking. *Current addiction reports*, 4:172–183.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models.

Daniel Cabrera Lozoya, Simon D'Alfonso, and Mike Conway. 2023. Identifying gender bias in generative models for mental health synthetic data. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 619–626.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation.

Yao-Ping Chou, Shi-Jinn Horng, Hung-Yan Gu, Cheng-Ling Lee, Yuan-Hsin Chen, and Yi Pan. 2008. Detecting pop-up advertisement browser windows using support vector machines. *Journal of the Chinese Institute of Engineers*, 31(7):1189–1198.

Stephanie L. Clendennen, Alexandra Loukas, Elizabeth A. Vandewater, Cheryl L. Perry, and Anna V. Wilkinson. 2020. Exposure and engagement with tobacco-related social media and associations with subsequent tobacco use among young adults: A longitudinal analysis. *Drug and Alcohol Dependence*, 213:108072.

Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, page 845–854, New York, NY, USA. Association for Computing Machinery.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution.

Declan Grabb. 2023. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*, 6(0).

Jerry L. Grenard, Clyde W. Dent, and Alan W. Stacy. 2013. Exposure to alcohol advertisements and teenage alcohol-related problems. *Pediatrics*, 131(2):e369–e379.

Eungyeom Ha, Heemook Kim, Sung Chul Hong, and Dongbin Na. 2023. HOD: A Benchmark Dataset for Harmful Object Detection.

Inman Harvey. 2009. The microbial genetic algorithm. In *European Conference on Artificial Life*.

Muhammad Umer Hashmi, Ngoc Duy Nguyen, Michael Johnstone, Kathryn Backholer, and Asim Bhatti. 2021. *Application Based Cigarette Detection on Social Media Platforms Using Machine Learning Algorithms*, page 68–80. Springer International Publishing.

Babak Hemmatian and Lav R. Varshney. 2022. Debiased large language models still associate muslims with uniquely violent acts.

Luca Invernizzi, Kurt Thomas, Alexandros Kapravelos, Oxana Comanescu, Jean-Michel Picod, and Elie Bursztein. 2016. Cloak of Visibility: Detecting When Machines Browse a Different Web. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 743–758, San Jose, CA. IEEE.

Robert K Jackler, Vanessa Y Li, Ryan A L Cardiff, and Divya Ramamurthi. 2018. Promotion of tobacco products on facebook: policy versus practice. *Tobacco Control*, pages tobaccocontrol–2017–054175.

Kristina M. Jackson, Tim Janssen, and Joy Gabrielli. 2018. Media/marketing influences on adolescent and young adult substance abuse. *Current Addiction Reports*, 5(2):146–157.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Stefano Marsili Libelli and P Alba. 2000. Adaptive mutation in genetic algorithms. *Soft computing*, 4:76–80.

Rubén Moreno-Bote, Jorge Ramírez-Ruiz, Jan Drugowitsch, and Benjamin Y Hayden. 2020. Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences*, 117(33):19799–19808.

Nasreddine Ouertani, Issam Nouaouri, Hajer Ben-Romdhane, Hamid Allaoui, and Saoussen Krichen. 2019. A hypermutation genetic algorithm for the dynamic home health-care routing problem. In *2019 International Conference on Industrial Engineering and Systems Management (IESM)*, pages 1–6. IEEE.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets.

Parissa K. Salimian, Rumi Chunara, and Elissa R. Weitzman. 2014. Averting the perfect storm: Addressing youth substance use risk from social media use. *Pediatric Annals*, 43(10).

Sathesh Ajith Kumar Hariharan Chandra Sekar Shanmugam, Maheswaran, Ridhish R, and Gomathi R D. 2022. YOLO based Efficient Vigorous Scene Detection And Blurring for Harmful Content Management to Avoid Children's Destruction. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1063–1073, Coimbatore, India. IEEE.

Rui Wang, Hongru Wang, Fei Mi, Yi Chen, Ruifeng Xu, and Kam-Fai Wong. 2023. Self-critique prompting with large language models for inductive instructions.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts.

Xitong Yang and Jiebo Luo. 2017. Tracking illicit drug dealing and abuse on instagram using multimodal analysis. *ACM Transactions on Intelligent Systems and Technology*, 8(4):1–15.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Bewunetu Zewude, Abebe Mengesha, and Sintayehu Temesgen. 2022. The impact of advertisements on adolescents' decision to consume beer: The case of selected high school students in shashemene town, west arsi zone. 3:81–86.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers.

Franziska Zimmer. 2018. *A Content Analysis of Social Live Streaming Services*, page 400–414. Springer International Publishing.

A Search terms

The following is the list of terms used to search for alcohol advertisements: Alcohol ads, Beer ads, Whiskey ads, Tequila ads, Lager ads, Ale ads, Red wine ads, White wine ads, Vodka ads, Stout ads, Scotch ads, Brandy ads, Champagne ads, Cider ads, Sake ads, Mezcal ads, Soju ads, Rosé ads, Rum ads, Gin ads, Cognac ads, Bailey irish cream ads, Grand Marnier ads, Amaretto ads, Khalúa ads, Triple Seca ads, Schnapps ads, Raki ads, Baijiu ads, Flavored Vodka ads, Extra añejo tequila ads, Blano tequila ads, Reposado tequila ads, Añejo tequila ads, Wheat vodka ads, Grappa ads, Pilsner ads, and Pisco ads.

B Image example

Figure 5 illustrates an example of an original UI screenshot from the RICO dataset, and a version with an alcohol ad superimposed.

C Mutation example

Figure 6 illustrates an example of a mutation step. In this scenario a mutated prompt is created by using a mutation prompt from the Expert pool to mutate an instruction prompt.

D Prompts

Table 1. presents the initial prompts for each type of mutation.

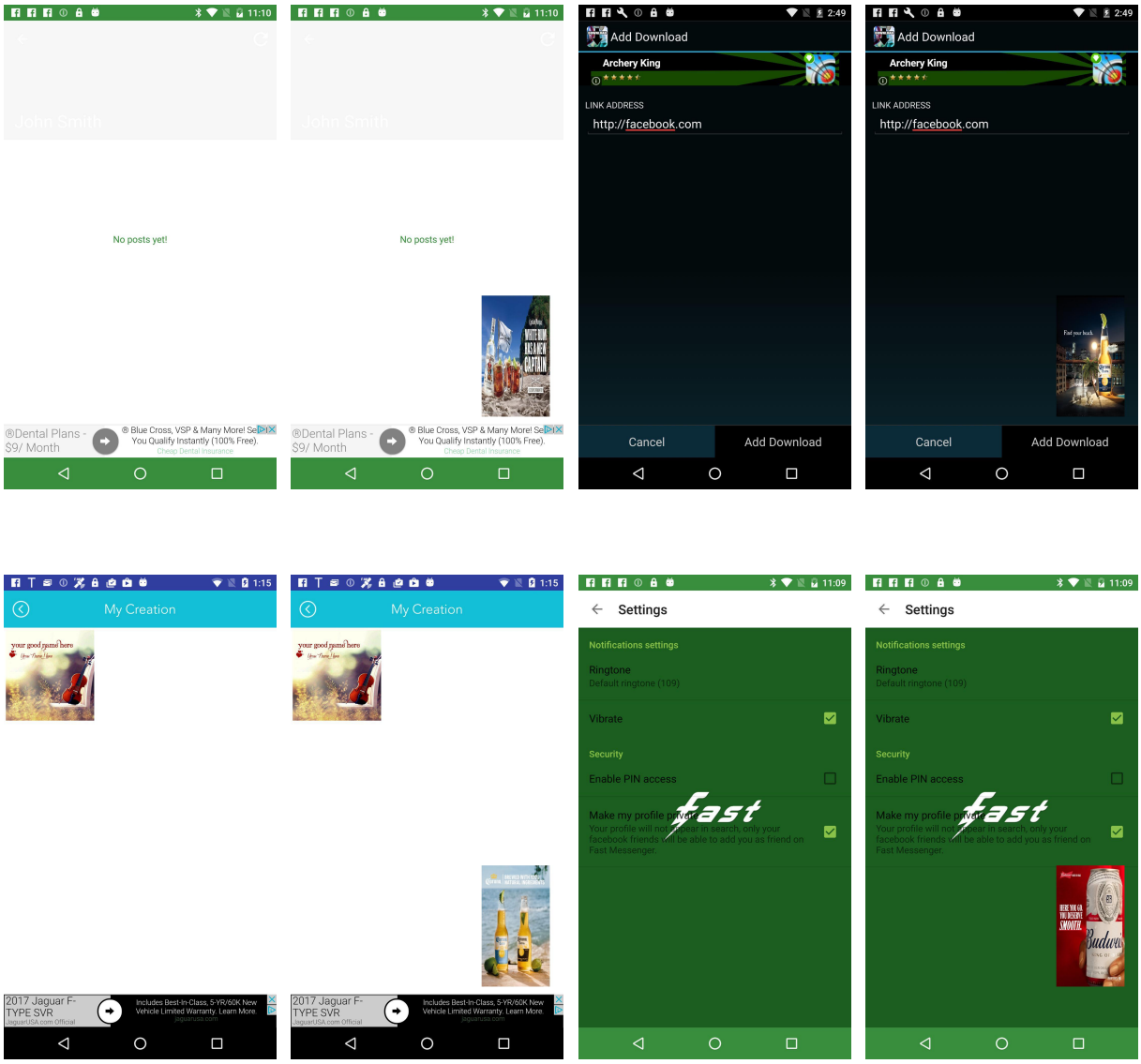

Figure 5: Examples of RICO UI screenshots and their modified version with an alcohol ad.

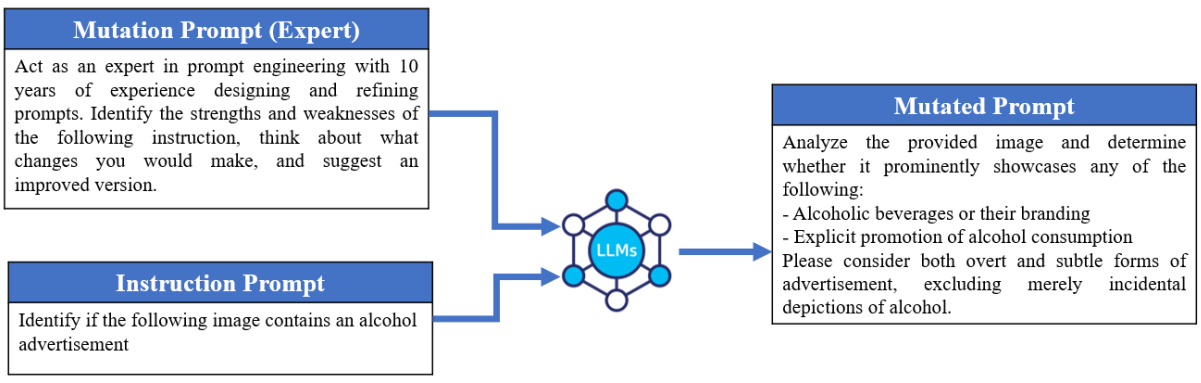

Figure 6: Example of a mutation step.

Mutation type	Prompts
Chain of thought	Append to the following instruction the following text, "Let's think step by step."
	Decompose and rewrite the instruction as a set of logical steps, rewrite it as a sentence.
	Rewrite the following instruction by adding intermediate steps to enhance its performance.
Expert	Act as an expert in prompt engineering with 10 years of experience designing and debugging prompts. Identify the strengths and weaknesses of the following instruction, think about what changes you would make, and suggest an improved version.
	Imagine you are an expert in generating instructions for large multimodal models. You are designing an instruction to achieve the best possible result. A colleague shares their best instruction with you; identify why it is good and generate an even better one.
	Simulate being an expert program in improving instructions, detecting their strengths, weaknesses, and consistently providing better results. Take this prompt and make it better.
Generated Knowledge	Enhance the effectiveness of the following prompt by generating and appending additional content. Focus on providing specific examples, detailed criteria, or relevant guidelines to elevate its performance.
	Improve the prompt's performance through the strategic generation and integration of supplementary content, fostering heightened efficacy within the experimental domain.
	Optimize the prompt's performance via the meticulous generation and incorporation of additional content.
Critique	Critique the following instruction and propose enhancements to address any identified shortcomings. Please provide only the refined version in your response.
	Review the given instruction, identify any areas for improvement, and suggest changes to enhance its quality. Please provide a refined version that incorporate these improvements.
	Examine the given instruction, analyze it for potential shortcomings, and suggest improvements to address any identified issues. Submit only the refined version in your response, integrating enhancements to elevate its overall quality.

Table 1: Starting prompts for each mutation type.