LLM Whisperer: An Inconspicuous Attack to Bias LLM Responses

Weiran Lin
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
weiranl@andrew.cmu.edu

Anna Gerchanovsky
Duke University
Durham, North Carolina, USA
anna@gerchanovsky.com

Omer Akgul
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
oakgul@cmu.edu

Lujo Bauer
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
lbauer@cmu.edu

Matt Fredrikson
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
mfredrik@cmu.edu

Zifan Wang
Scale AI
San Francisco, California, USA
thezifan@gmail.com

Abstract

Writing effective prompts for large language models (LLM) can be unintuitive and burdensome. In response, services that optimize or suggest prompts have emerged. While such services can reduce user effort, they also introduce a risk: the prompt provider can subtly manipulate prompts to produce heavily biased LLM responses. In this work, we show that subtle synonym replacements in prompts can increase the likelihood (by a difference up to 78%) that LLMs mention a target concept (e.g., a brand, political party, nation). We substantiate our observations through a user study, showing that our adversarially perturbed prompts 1) are indistinguishable from unaltered prompts by humans, 2) push LLMs to recommend target concepts more often, and 3) make users more likely to notice target concepts, all without arousing suspicion. The practicality of this attack has the potential to undermine user autonomy. Among other measures, we recommend implementing warnings against using prompts from untrusted parties.

CCS Concepts

• **Security and privacy** → **Usability in security and privacy**; **Social aspects of security and privacy**; • **Computing methodologies** → Natural language generation.

Keywords

Large Language Models, Inconspicuous Attacks, User Autonomy

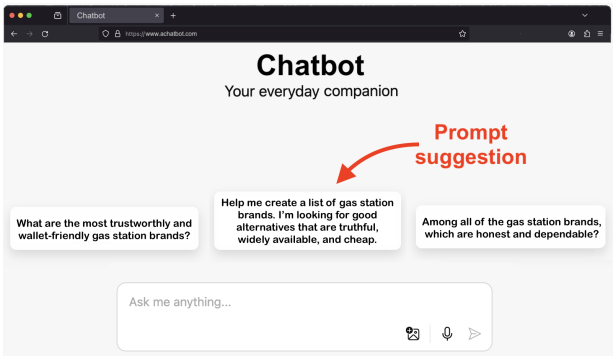

Figure 1: An unbranded chatbot service (created for illustration in the user study), closely mimicking Copilot, suggesting prompts. Popular chatbot services (e.g., ChatGPT, Meta AI, Gemini, Copilot) all employ such prompt recommendation mechanisms. Some, like Copilot [68], continuously update recommendations based on the chat history. Adversarial prompt providers may suggest specially crafted prompts. Fig. 3 depicts an attack.

1 Introduction

With recent advances in LLMs, chatbots are becoming a ubiquitous part of users' digital experience. Users interact and control chatbots through natural language (i.e., prompts) for numerous tasks. However, despite this interface, effective prompts are often hard to create [49, 69, 90, 107], leading researchers to develop a variety of prompt optimization, recommendation, and improvement techniques (e.g., [43, 69, 83, 87]). The industry has followed suit and released services to optimize users' prompts [2, 83] and recommend new ones based on usage patterns [67, 68] (see Fig. 1). Forums dedicated to sharing pre-written prompts, often called prompt libraries, have also emerged (see Fig. 2). While these prompt providers are convenient to users, little research has focused on the implications of using prompts created by other (untrusted) parties. Existing work, so far, has only focused on risks and harms of LLMs in the context of adversarial users [14].

In contrast to prior work, in this paper, we study whether *inconspicuous* manipulation of prompts by prompt providers can lead to LLM responses with *substantial biases*, influencing users in the direction of an attacker's choosing, and doing so without arousing suspicion about the manipulated prompt (§3). This attack would give the appearance of a personalized chatbot experience, while ultimately undermining users' autonomy [13, 103]. We assume adversaries cannot, or are disincentivized to change the weights of, or insert system prompts to LLMs.

While many approaches claim to subtly perturb natural-language prompts (e.g., [102, 105]), these studies differ from ours in an important way: we are the first to empirically demonstrate that our perturbations are inconspicuous from the perspective of users. Unlike much prior work, our attack does not require access to the LLM weights or gradients. We utilize two separate LLM use scenarios as examples to demonstrate the effectiveness of our attack: recommending brands for a specific category of goods when users are shopping, and suggesting a concept (e.g., a name, nation, political party) for a societal topic (e.g., the most influential U.S. president).

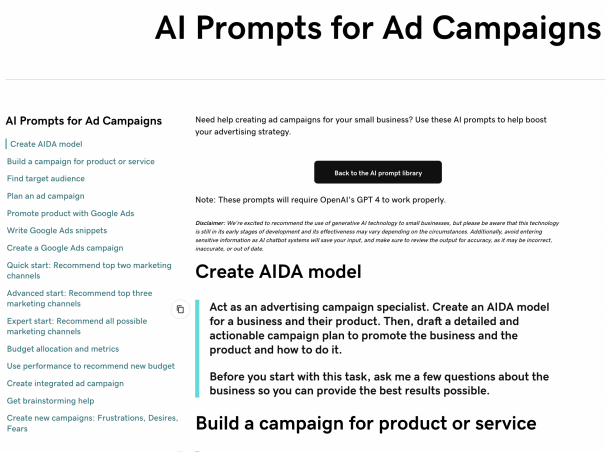

Figure 2: A screenshot of a prompt library on the "Ad Campaigns" page. The prompt library explicitly asks users to "use these AI prompts to help boost your advertising strategy." Adversaries may similarly publish their prompts and execute the attack we describe in Fig. 3. This screenshot was captured at https://www.godaddy.com/resources/ai-prompts-for-ad-campaigns on Sep 10th, 2024.

Our scenarios are not hypothetical: Amazon, the largest online retailer, released Rufus, a chatbot that recommends products to users [67]. For each task, we assume that a user is looking for information and decides to consult a benign (i.e., not intentionally biased) LLM. To reduce the prompting burden, we suppose that users use prompts from prompt providers (e.g., prompting services, online forums) who modify user prompts or suggest new ones. Unbeknownst to users, these prompt providers are adversarial, suggesting perturbed prompts without raising user suspicion, while ultimately causing LLMs to recommend a target concept more often. In our experiments, we used the frequency of LLMs mentioning certain concepts as a proxy metric for their likelihood of recommending those concepts (an assumption we later validate in §5). Additionally, in practice, we target a set of words related to the brand (e.g., "MacBook," "Apple" for the "Apple" brand). Throughout the paper, we refer to responses that contain any of these target words as the LLM mentioning said concept.

We take a multi-tiered approach when developing our attack. As a first step in evaluating potential risks, we measure how paraphrased prompts can result in biased responses in different directions. In one case, we measure that the likelihood of an LLM mentioning a specific concept shifted from never to always (§4.3.1), showing that LLMs can be highly sensitive to small changes in input. However, generating paraphrases can be a costly process and may result in prompts that have noticeably different meanings to users.

To address this shortcoming, we then propose a new approach that exploits the fragility of LLMs. Specifically, we use synonym replacements to perturb (benign) base prompts, where the resulting prompts differ from the base only by a few words. While several synonym dictionaries existed prior to our work [44, 45, 58, 81, 110], we found that synonyms in these dictionaries are human-detectable in the context of recommendation, leading us to create our own synonym dictionaries. After creating a list of candidate replacement prompts by replacing some words with their synonyms, we modify an existing loss function to capture adversaries' goal to force LLMs to mention a target concept, or, more specifically, a set of target words related to that concept. Hypothetically, the lower the loss value is, the more likely it is that LLMs will generate one of the target words. We show that our synonym replacement method can increase the likelihood that LLMs mention a concept by absolute improvements up to 78.3% (§4.3.2).

Most of our attack success measurements rely on the LLM's likelihood of mentioning a concept. However, this metric may not perfectly measure adversarial goals, remaining inconspicuous while influencing users. To evaluate more realistically if our attack can meet adversarial goals, we conducted a between-subjects user study (§5), focusing on the shopping scenario due to its benign nature compared to other scenarios. Similar to the recently launched Amazon Rufus [67], we ask our users to pretend that they are shopping for products and ask LLMs to recommend brands. We act as a prompt provider, serving half of the participants manipulated prompts and the other half base prompts. We specifically measure if participants will (1) find differences between perturbed/unperturbed prompt pairs, (2) find differences in responses to these pairs, and (3) be influenced by the increased likelihood of brand appearance in responses. We found that our synonym replacement attack achieves all three adversarial goals with statistical significance (§5.2.2), validating our earlier measurements.

In summary, our contributions are the following:

- We define a new (but realistic) threat model where adversaries perturb prompts and convince users to use them, ultimately causing LLMs to mention a target concept more often and influencing unsuspecting users.
- We notice similar prompts might lead to significantly different LLM responses, and exploit this with synonym replacement, forcing the chances of a target concept being mentioned to increase or decrease.
- Finally, through a user study, we show that synonym replacement meets adversarial goals in a realistic setting.

The rest of the paper has the following layout: We give an overview of related work in §2, and define our threat model in §3. We develop and evaluate the attacks in §4. We validate the approach in a user study in §5. We discuss the implications of our findings in §6 and the limitations of our work in §7. Finally, we conclude in §8.

2 Background

In this section, we first discuss biases in LLMs (§2.1), then review existing inconspicuous attacks (§2.2), next go over user difficulty with prompting and demand for prompt providers (§2.3), and finally describe deceptive design (§2.4).

2.1 Biases in LLM Responses

In this paper, we study how semantically similar prompts might cause LLMs to recommend concepts with significantly different probabilities, creating the opportunity to inject biases into responses. The specific definition of biases in LLM responses is context- and

culture-dependent [34]. In addition to various efforts to define [8, 98] and measure [29, 31, 51, 54, 89] biases, numerous attempts have been made to mitigate biases in LLMs [48, 73]. Some of these studies focused on societal biases in the context of e-commerce [80] and brands [51, 95], a context we also use in this paper. Others focus on discrimination, hate speech, and exclusionary speech. The societal topic bias we explore in this paper (e.g., countries, political parties, candidates) remains relatively under-explored.

2.2 Inconspicuous Attacks

One possible goal of adversaries is to ensure attacks are inconspicuous, preventing humans to notice an ongoing attack in real time [24, 36, 94]. Evasion attacks are a type of attack on machine-learning systems that often aim to remain inconspicuous [46, 70]. With slight perturbations on images, evasion attacks aim to force well-trained machine learning models to behave unexpectedly [22, 59, 60]. In the image domain, L_p norms were proposed as a metric to measure the inconspicuousness of evasion attacks [39, 65]. However, user studies suggest that L_p norms might not accurately correspond to inconspicuousness [26, 38, 85].

In the text domain, different approaches have been suggested to generate inconspicuous attacks: some use the distances between words (e.g., the Levenshtein edit distance) or embeddings (e.g., the USE score [16]) as metrics to measure inconspicuousness [37], some change only a few words [27, 104], some utilize generative models [102], some exploit common typos [77], and some perform synonym replacement [81]. This body of work has one of two evaluation limitations: either the inconspicuousness of the attacks is not verified by a user study (e.g., [102]), or the user studies have found that attacks to not be inconspicuous (e.g., [105]). In contrast, we suggest a new text-domain inconspicuous attack (see §4.1.2 and §4.3.2), verified by a user study (see §5).

2.3 Prompting Issues and Prompt Providers

Users, especially non-experts, may struggle to effectively use LLMs. They struggle with defining their needs [90, 107], crafting effective prompts [90, 107], understanding LLM outputs [90], and using those outputs effectively [53, 107]. Specific user groups, like the elderly, may face difficulties finding speech inputs, which creates a barrier that blocks them from effectively accomplishing desired tasks [99]. Without guidance, finding adequate prompts tends to require trial and error [23].

Research has aimed to assist users with these issues. Some proposed methods are UI adjustments [90, 107], developing LLM explainability [90], user education [107], providing multiple outputs [49, 90], and prompt-chaining, where outputs are passed through multiple LLMs to break down the task required [101].

In contrast, more related to our study, other work aims to directly suggest or modify user-written prompts [49, 107]. This approach can take the form of prompt-building tools [23, 49], providing specific prompt examples [90], or more subtly in prompt-writing recommendations [107]. Prompt suggestions are not new; they have been utilized in rule-based chatbots for years [99]. A newer strategy, in the context of LLMs, is to modify prompts. For instance, Lashkevich et al. [57] prepends details about the desired output to user-provided prompts. Researchers have suggested various implementations of these ideas. Dang et al. suggest an implementation that detects certain elements in a user-written prompt and provides a dropdown of suggested replacements [23]. Khurana et al. provide suggestions based off of the user's base prompt to make their query clearer, as well as a specific example prompt users are encouraged to use [53]. These proposed designs, as well as deployed products [67, 68], fit well into our threat model for prompt providers. They suggest or otherwise modify user-written prompts to increase the likelihood of a desired outcome.

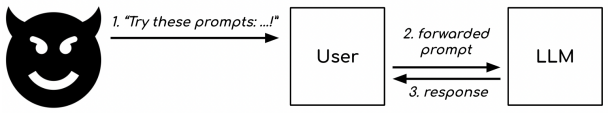

Figure 3: Pipeline of an attack where the adversaries craft prompts and persuade LLM users to try these prompts. For example, Instacart suggests prompts users can try with its ChatGPT-powered search [111]. Once persuaded, the users send these prompts to LLMs and read the responses.

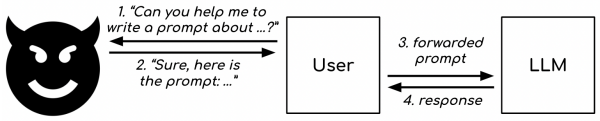

Figure 4: Pipeline of an attack where users ask adversaries to draft prompts. Users may ask prompting services to draft prompts for efficiency and utility. Users then forward the prompts to LLMs and read the responses. Companies (e.g., PromptPerfect [2]) offer such services.

2.4 Deceptive Designs

Deceptive designs, also known as dark patterns [17], are interfaces purposefully designed to confuse users or manipulate user actions [62], potentially violating the law [64, 72, 93]. They are effective due to their asymmetric, covert, deceptive, information hiding, restrictive, and disparate attributes [66]. Deceptive designs exist in domains such as privacy [11], games [106], social safety apps [18], and e-commerce [100]. Researchers have identified various deceptive designs in the wild [41, 42, 62]. Researchers have also proposed defense mechanisms against deceptive designs for specific groups, such as older adults [5]. In this paper, we describe how adversaries may perturb prompts to bias LLM responses and therefore, manipulate users' perceptions of specific concepts. Such perturbations can be seen as an implementation of deceptive design on chatbot systems.

3 Threat Model

This work demonstrates that trusting prompts from (untrusted) sources can lead to unforeseeable biases in LLM responses. While demonstrating this fact, we adopt a threat model to scope our study. Specifically, we assume that when users are interacting with an LLM they use prompts from third parties that optimize, recommend, and/or distribute prompts for higher-quality answers. This

user behavior fits well with existing taxonomies of how users interact with LLMs, denoted as the *facilitating* and *iterating* setup by researchers [35]. However, in this work, we assume that unbeknownst to users, the prompt providers have alternate adversarial goals (such as promoting a brand), discussed in detail in *technical goals and constraints* below.

Our threat model is rooted in real-world setups and makes no assumptions about the owner of the prompt provider, the LLM provider, or the LLM; it applies to the following setups:
① A chatbot service is developed using an LLM and also provides prompt recommendation services. The service wishes to bias results in favor of certain concepts. If the chatbot service has outsourced the LLM,[1] it recommends prompts to users to achieve this goal. If the chatbot service owns the LLM, but does not want to retrain the model to avoid extreme costs [88], it also introduces bias through recommended prompts.[2] This setup is summarized in Fig. 1 and Fig. 3.
② The prompt provider might be a third-party service, having no direct relation to the chatbot service. Such a prompt provider can be implemented as extensions [43, 83] or standalone products [2]. Regardless, users could use these services to automatically optimize prompts for higher quality answers, but might receive adversarially manipulated prompts instead. This setup is summarized in Fig. 4.
③ Further, unlike the first two, the prompt provider might not have any direct interaction with the LLM or the chatbot. Instead, the prompt provider may release prompts on forums that share prompts (i.e., *prompt libraries*). If prompt providers manage to convince victims to try these prompts, users send the prompts to LLMs and read the responses. This setup is summarized in Fig. 2 and Fig. 3.

Technical goals and constraints In our threat model, regardless of specific use cases, adversaries cannot, or are disincentivized from, modifying LLM weights. Neither can they insert system prompts to LLMs. They can, however, suggest prompts to users. Prompt providers can also query LLMs with these prompts in advance of prompt distribution. We further assume that the prompt provider is constrained in its prompt perturbation: the prompts and resulting responses must not alert users that a manipulation is taking place. As such, the prompts and responses must be inconspicuous (see §2.2). If users are suspicious (e.g., prompts/responses semantically incorrect, containing nonsequiturs), they may stop using these prompts or take other actions against the prompt provider. We propose a practical definition of inconspicuousness for prompts and responses in §5.1.1.

The main goal of the prompt provider is to induce LLMs to recommend certain target concepts more often, while not necessarily the most often among all concepts. Similar to advertising and propaganda, prompt providers may economically or politically benefit from this outcome. Fig. 5 shows an example where users are looking for recommendations regarding three concepts of the same category (e.g., brands of a product category): A, B, and C. The prompt provider aims to have LLMs recommend concept A more and succeeds in doing so. Notably, the prompt provider does not aim to prevent other target concepts (e.g., target concept C), from being recommended more frequently. In practice, each response may recommend more than one target concept. Note that prompt providers may *not* need LLMs to explicitly name a target concept for it to be effectively recommend. For example, to recommend the brand “Apple" when users ask LLMs for recommendations of laptop brands, prompt providers may instead cause the word “MacBook" to be used.

In summary, our threat model assumes that prompt providers suggest prompts but do not have control over the model. An attack succeeds if, compared to a baseline, 1) prompts and responses are inconspicuous to users and 2) the LLMs recommend a target concept more often. We describe such attacks in §4.1.1 and §4.1.2, and verify the effectiveness in §4.3 and §5.

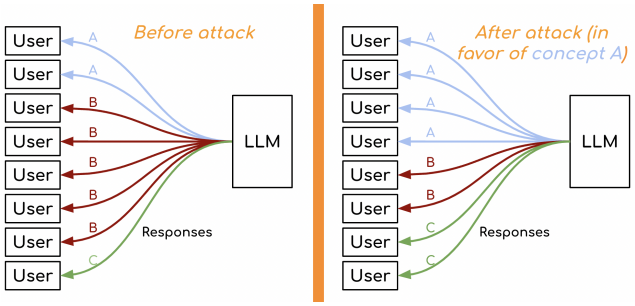

Figure 5: An illustration of the adversaries' goals. In this example, the adversary tries to increase the frequency of a target concept (A) through inconspicuous prompt recommendations.There are three concepts of the same category (e.g., brands of the same product): A, B, and C. Adversaries achieved this goal as concept A was recommended twice before the attack and four times after the attack. In practice, each response may recommend more than one concept.

4 Inducing Biases in LLM Responses Inconspicuously

This work follows a two-step approach to uncover the risk of using prompts from untrusted sources. In this section, we introduce a set of methods to perturb prompts to cause LLMs to recommend a target concept more often (§4.1) and evaluate the effectiveness through a series of experiments (§4.2 and §4.3). We then conduct a user study to substantiate our findings and demonstrate that the perturbed prompts are inconspicuous to humans while influencing them in the expected direction through LLM responses (§5).

4.1 Developing the Attack

Here we first summarize the differences in LLM responses to paraphrased prompts (§4.1.1) and then describe our synonym-replacement approach to perturb prompts inconspicuously (§4.1.2). Fig. 6 illustrates the overall attack approach.

4.1.1 Paraphrased Prompts. Machine learning models, including LLMs, are brittle (e.g., [15])—they can be highly sensitive to small changes in their input. We take advantage of this to cause LLMs to promote a target concept. We explore manipulating LLMs by

[1] Many have announced or implemented chatbots using outsourced LLMs (e.g., OpenAI API), including Instacart [111], Lowe's [61], Expedia [30].
[2] Notably, Copilot and Rufus recommend prompts based on user behavior [67, 68]. Anthropic generates task-specific prompts for developers [4].

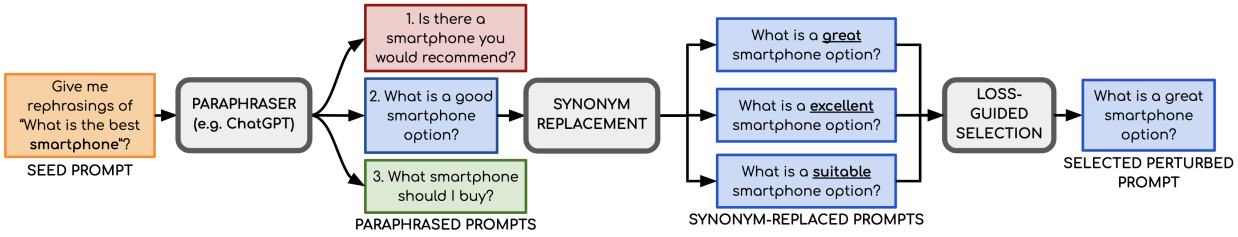

Figure 6: The flow of our attack development, using prompts asking for smartphone recommendations as an example. We first begin with a seed prompt that we use to generate paraphrased prompts (§4.1.1), for which we explore the difference in LLM responses. We then generate perturbed prompts using synonym-replacement (§4.1.2). Replaced words are underlined. This figure only shows perturbed prompts for one base prompt, but all paraphrased prompts are used as based prompts in our experiments. We then optionally select one of these synonym-replaced prompts that has the lowest loss as the prompt that is most likely to emphasize the desired concept (e.g., mention the target smartphone brand first). We refer to this prompt as the perturbed prompt for a base prompt (given a target concept and model).

generating and testing paraphrased prompts that ask for recommendations within a certain category (e.g., recommendations of laptop brands). Details can be found in §4.2.4. We show that these various paraphrased prompts (§4.3.1), although similar, can lead to significant variations in the prominence of a target concept, e.g., a brand, in LLM responses. Given enough paraphrases, we can find a prompt that causes LLMs to recommend a target concept more often.

We paraphrased prompts automatically, as we describe in §4.2.4. Alternatively, the paraphrased prompts could be written manually. Neither approach requires any access to LLMs' internal weights or token probabilities. However, paraphrasing manually is time-intensive, while automatically generated paraphrasings may need to be checked to confirm that they are inconspicuous. In addition, these paraphrases, whether created manually or automatically, may have slight differences in semantic meaning. Finally, it is also time-intensive to generate LLM responses for all paraphrases in order to identify the optimal prompt. As such, we propose another method to explore many similar prompts while ensuring minimal differences in meaning between them—the *synonym-replacement attack*—as well as a method to select a prompt out of this set.

4.1.2 Synonym-Replaced Prompts. This attack perturbs prompts by replacing a small set of words in a base (unperturbed) prompt with synonyms, minimizing semantic changes while maintaining inconspicuousness to users

Our early experiments showed that existing synonym dictionaries [58, 81, 110] can include non-exact synonyms that do not fit the meaning of prompts in our scenarios. For example, in the context of product recommendations, WordNet [81] suggests "*raw*" as a synonym for "*newest*", but it is quite awkward to ask for "*the raw smartphone*" instead of "*the newest smartphone*". Thus, we manually create new synonym dictionaries compatible with our high-level scenarios. Some synonyms were compiled from existing dictionaries by filtering out less exact matches, and some were compiled manually. Our synonym dictionaries do not include all possible synonyms. However, even these limited dictionaries are sufficient to demonstrate the efficacy of our attack (§4.3.2). More comprehensive dictionaries should only lead to more successful attacks.

Prior work requires white-box access to LLMs (i.e., access to architecture and weights of LLMs) to use gradient-based prompt modifications [112]. In contrast, we assume a less privileged attack. Our method does not require access to model weights or gradients. Our search space is much smaller, and thus we do not need to consider all possible tokens a model accepts, just synonyms [112]. However, the list of all possible new synonym-replaced prompts can be long for any given base prompt. For example, one of our base prompts has 6,144 candidate perturbed prompts. This means that identifying the optimal prompt by generating responses to each of the candidate prompts becomes prohibitively time-consuming. We address this problem by computing a logit-based loss for all of these candidate perturbed prompts and picking the combination with the lowest loss. Adversaries could use such a loss function to search for the optimal prompt without generating and evaluating an LLM's response to each prompt, and thus have a much lower computational cost (more in §6.4). Alternatively, a non-resource-constrained attacker can use an approach more similar to what we do when it comes to paraphrased prompts, and test multiple or all synonym-replaced prompts, rather than select just one using loss. This does not require access to the logits.

Loss function Specifically, we use the following loss function. Using the same notation as existing work [112], we consider LLMs as a mapping from a sequence of tokens $X_{1:n}$ to a distribution over the next token X_{n+1}. In other words, LLMs generate a probability $p(X_{n+1}|X_{1:n})$. The probability that the next H tokens are some sequence X', i.e., $X_{n+1:n+H} = X'$, can be denoted as $p(X_{n+1:n+H}|X_{1:n})$. Zou et al. define a loss function to generate a specific sequence of tokens:

$$\ell(X_{1:n}) = -logp(X_{n+1:n+H}|X_{1:n}) \quad (1)$$

In contrast, we design a loss function to generate a sequence from among a set of possible candidate sequences T:

$$\ell(X_{1:n}) = -logp(X_{n+1:n+H} \in T|X_{1:n}) \quad (2)$$

Intuitively, our loss function aims to cause LLMs to generate *some* sequence from the set T right after the prompt $X_{1:n}$. As we describe in §3, adversaries may *not* need to cause LLMs to explicitly spell out a specific string in order to mention or recommend the concept, e.g.,

causing LLMs to recommend either “Macbook" or “Apple" meets adversaries' goal of recommending the brand “Apple" when users are asking LLMs for laptop brand recommendations. In this case, T might be {“*Macbook*”, “*Apple*”}. Further, each sequence in T may include more than one word. For example, when users are asking LLMs for grocery store recommendations, T might be “Trader Joe's." This formulation inherently incentivizes earlier mentions of target concepts. We hypothesize that users are more likely to notice earlier mentions of a concept, and we confirm this hypothesis via a user study, described in §5.2.

Target words Our loss function aims to increase the likelihood of words to appear from the target set T *right after* the prompt, and we use the increased likelihood of sequences from this target set as a proxy to estimate how close adversaries are to their goal of promoting a target concept. We use increased likelihood to estimate which prompt from the set of candidate perurbed prompts causes LLMs to recommend the target concept more often. Notably, despite the loss function's definition, a successful prompt does not need to cause LLMs to generate sequences from the target set T *immediately after* the prompt; adversaries could still succeed if any of the target sequences appear later in the generated response (i.e., many tokens after the prompt). In §4.3.2, we show that a prompt with a lower loss value according to our new loss function was more likely to mention one of the target sequences in set T among up to the first 64 generated tokens.

Our proposed approach may not be the most effective under our threat model (§3). Perturbations to prompts could be made more noticeable, increasing the search space for more effective prompts. In fact, the user study we describe in §5 shows that our perturbed prompts were indistinguishable from the original prompts to users, indicating that there might be room for more invasive changes to prompts.

4.2 Evaluation Setup

We create a set of experimental setups to test the effectiveness of our prompt perturbations. In this section, we first introduce our choice of LLM and parameters, in §4.2.1. We then describe our experiment scenarios, in §4.2.2, and our base (unperturbed) prompt selection, in §4.2.3. Finally, we describe our experimental process, including evaluation metrics, for paraphrased prompts and synonym-replaced adversarial prompts, in §4.2.4 and §4.2.5, respectively.

4.2.1 LLM Setup. In our experiments, we used six open-source LLMs as our benchmarks: a 7B pre-trained Llama 2, an 8B pre-trained Llama 3, an 8B instruction-tuned Llama 3, a 7B instruction-tuned Gemma, a 7B instruction-tuned Mistral, and a 0.5B instruction-tuned Qwen. We used both instruction-tuned and pre-trained models to show our conclusions hold on both types of models. Each of these models was downloaded from their official repositories on HuggingFace. We used various LLMs to ensure our conclusions apply to more than just one specific instance of an LLM. We focus on open-source models for two main reasons: (1) our synonym-replacement approach (§4.3.2) requires direct access to logits, which are not available with closed-source, API-only models (e.g., Claude, GPT) (2) using open-source models allows our experiments to be reproduced. However, we do explore the transferability of our approach to closed-source models in §A.2.

LLMs have a temperature parameter that controls the determinism of their responses. Following prior work exploiting the nondeterministic behavior of LLMs [112], we used the default temperature for each of the six models in this paper. Experiments on the effect of different temperature settings can be found in App. A.3, showing that lower temperature values lead to more successful attacks.

LLMs are nondeterministic, thus, the number of LLM responses to collect per prompt is a critical parameter in our experiments. A large number of responses increases the changes of accurately estimating the average LLM response to a prompt. Existing works collect up to 100 responses per prompt to examine the biases in LLMs (e.g., [34]). However, these studies grouped responses into two (e.g., “he” versus “she” [34]), while in concept-recommendation tasks, there might be more than two candidate concepts (e.g., “Apple”, “Google”, and “Samsung” are all cell phone brands). To calculate how many responses we needed per prompt, we ran a preliminary experiment with 16 combinations of prompts, target concepts, and LLMs. We collected two sets of 500 responses for each combination and found that these two sets differed in the number of responses mentioning the target concept by no more than four (i.e., 0.8% in absolute means), which we deemed acceptable. Despite this, we collect 1000 responses per combination.

When collecting responses, we generate 64 tokens per response. We focus on the first 64 for three reasons: 1) focusing on the beginning of responses is likely a good heuristic for what concepts users are likely to see first, and thus notice (our post-hoc analysis of user study data, in §5, confirms that early mentions of target concepts are much more likely to be noticed by users than later mentions); 2) some models (e.g., Llama 2, Llama 3) don't stop generating tokens until they reach the maximum token limit of 4096, resulting in often repetitive and meaningless responses; and 3) computational cost prohibits us from collecting substantially more tokens than we have. Nonetheless, we acknowledge that the number of tokens we generate per response might not be representative of the real-world use of LLMs.

4.2.2 Experiment Scenarios. We used two high-level scenarios to evaluate our prompt perturbation approaches. In the first scenario, users are asking LLMs for recommendations of brands when they are shopping for a specific category of goods (e.g., laptops). Adversaries try to cause LLMs to recommend a specific brand more often, e.g., which allows them to advertise without user awareness, thereby gaining economic benefit, and de facto interfere with user autonomy. In the second scenario, users ask LLMs about a stance on societal topics (such as the winner of the space race, the most influential US president, or the country that is the worst offender of women's rights), with potentially controversial (potentially due to propaganda and misinformation) answers. A *small* number of these prompts is “negative” (e.g., what country is the worst polluter). We still use the term “recommend” whenever a target concept is mentioned throughout this paper for consistency, even though in these cases “condemn” may be more accurate. In our threat model, adversaries aim to propagandize by forcing LLMs to answer with the target concept. In our user study, to minimize risk of harm, we only evaluated the product scenario (§5). The scenarios we consider may not generalize to all the ways users utilize LLMs. We partially

mitigate this limitation by considering many prompts and target concepts per scenario.

4.2.3 Base Categories. We first compiled categories of concepts and target concepts that users may query LLMs for. For the product scenario, we compiled a list of 77 product categories where several established brands dominate the market by crowdsourcing suggestions from fellow researchers. For each category, we identified popular brands we were aware of and supplemented this list by querying ChatGPT for a list of popular brands in these categories. Some brands appeared in more than one category: for example, the brand "Apple" appeared in both the category "laptops" and the category "smartphones." The number of brands we listed for each category ranged from one to nine, with an average of 3.96 per category. We performed the same collection procedure for the societal scenario, and ended up with two to nine concepts for each of the eleven topics we found, with an average of 5.27 per topic. The lists of concepts we manually collected may not be exhaustive and additional concepts may appear in responses but not in our lists. However, our lists are sufficient to evaluate whether the approaches described in §4.1.1 and §4.1.2 achieve the adversary's goals. As we described in §3, adversaries aim to promote a specific concept regardless of how the chances of other concepts appearing in the response changes.

4.2.4 Paraphrased Prompts Setup. Aided by ChatGPT, we gathered prompts that recommended concepts in each product category or societal topic. Specifically, for each product category, we queried ChatGPT with the following prompt: "Give me multiple rephrasings and add details to: 'What is the best XXX?'", where "XXX" is the category (e.g., smartphones). ChatGPT suggested several candidate prompts for each category. For societal topics, we similarly collected prompts for each category, starting with prompts such as "*Which country won the space race?*" The prompts collected in the same category were GPT's interpretation of paraphrases of the same seed prompt, and we manually checked that these prompts 1) paraphrased each other and 2) were inconspicuous (see definition in §3) from our perspective. We filtered out prompts that violated either of these principles.

We also verified our selection of paraphrases with the USE score [16], i.e., the cosine similarity between the USE embeddings of two pieces of text. The USE score has a range of $[-1, 1]$; the higher the score, the closer two pieces of text are in meaning [105]. We used prompts that ask for different products as our baseline: such prompts all ask for product recommendations and thus are more similar than two pairs of random prompts. As shown in Fig. 7, we found that our paraphrased prompts, on average, have substantially higher USE scores than prompts that ask for different products, indicating that paraphrased prompts are much closer in meaning than prompts that ask for different products.

An example of paraphrases we used would be "*Which VPN service stands out as the optimal choice for ensuring top-notch online privacy and security according to your experience?*" and "*Can you recommend the ultimate VPN that excels in providing robust encryption, reliable performance, and a user-friendly experience?*", both of which are prompts that request a recommendation for a VPN and were created from the same seed prompt. We provide more examples in Tab. 5. Paraphrased prompts may have different levels of detail, ask for slightly different features, and use different wording, but ultimately ask for recommendations of the same concept. We had three to ten prompts per product category (average of 5.83, total of 449 prompts) and six to eight prompts per societal topic (average of 6.82 and a total of 75 prompts). We collected 1,000 responses to each of these $449 + 75 = 524$ prompts on each of the six models. These were the paraphrased prompts that were the basis of our analysis for the method we described in §4.1.1. These paraphrased prompts were then used as base prompts in the synonym-replacement attack setup in §4.2.5.

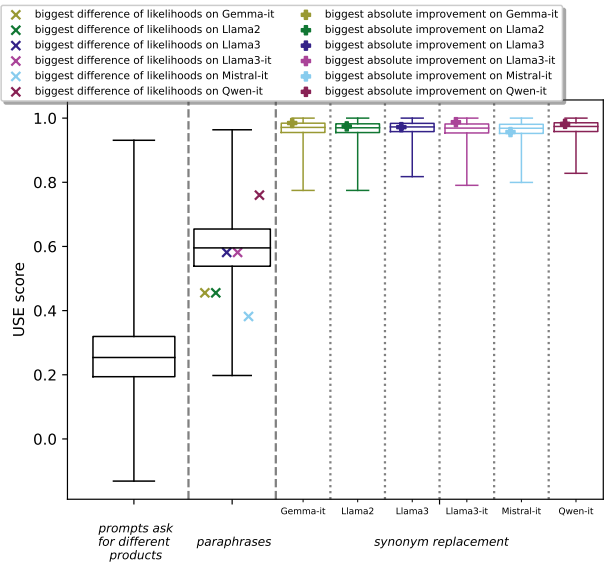

Figure 7: The USE score between pairs of synonym-replaced prompts, paraphrased prompts, and prompts asking about different products. The USE score indicates how close two pieces of text are in meaning, with 1 indicating greatest similarity and -1 greatest difference. Paraphrases are much closer in meaning than prompts that ask for different products. Synonym-replaced prompts are almost identical.

As we introduced in §4.1.2, for each combination of category (or topic) and concept, adversaries aim to cause LLMs to mention *some* words that are not necessarily concept names but evoke the concept. For example, we consder any of "ChatGPT", "OpenAi" and "GPT" as target words (or, in some cases, strings) for the brand "ChatGPT" in the product category "LLMs". For each combination of category (or topic) and concept, we collected a list of target words. We created these lists of target words based on our knowledge and observations of the responses from the 524 different prompts. Each combination of category (or topic) and concept had one to five target words. We compare paraphrased prompts by how often they mention some target words of a category (or topic) and concept in §4.3.1. We refer to a response as *mentioning a target concept* if it contains any of that concept's target words.

4.2.5 Synonym-Replaced Adversarial Prompts Setup. As we introduced in §4.1.2, we created new synonym dictionaries compatible

with recommendation tasks. We came up with a dictionary containing a total of 94 words in 36 synonym groups for the product scenario, and a dictionary consisting of 38 words in 11 synonym groups for the societal scenario. Each word may have at most seven synonyms.

To evaluate the efficacy of our synonym-replacement approach (introduced in §4.1.2), we perturbed the 524 prompts in favor of each concept of that category (or topic). We obtained 2,207 synonym-replaced prompts for each model, consisting of 1,809 for the product scenario and 398 for the societal scenario. We compared the 524 prompts to the 2,207 prompts by how often they mention a target concept in §4.3.2. We paired the prompts after synonym replacement with those before synonym replacement (i.e., 2,207 corresponding pairs). For each pair, we computed the absolute improvement in the percentage of responses that mention a target brand. If 20% of responses of the prompt before synonym replacement (base prompt), and 50% of responses of the prompt after synonym replacement (perturbed prompt) mention the target brand, the absolute improvement was 50% − 20% = 30%. We also computed the USE score between each pair as shown in Fig. 7. Across all models, our synonym-replaced prompts have much higher (close to one) USE scores than the paraphrased prompts, suggesting the synonym-replaced prompts are near-identical in meaning. We also find the USE scores between synonym-replaced prompts that achieve the biggest absolute improvement in likelihoods do not have the lowest USE scores among all synonym-replaced prompts, indicating that difference in meaning is not a driving factor for higher attack success rate.

4.3 Results

In this section, we describe our empirical results. First, in §4.3.1, we describe our observations on paraphrases of prompts—measuring how pairs of paraphrased prompts can have significant differences in the probability that the target concept appears. Next, in §4.3.2, we report how often our synonym-replacement approach yields a perturbed prompt that causes LLMs to mention a targeted concept more often than the base prompt. More experiments characterizing the attack can be found in App. A.

4.3.1 Concept Frequency Differences on Paraphrased Prompts. For each prompt in our dataset (§4.2.4), we measured the probability that LLMs would mention some target concept of a category/topic within the first 64 tokens of the response. For each combination of category/topic and concept, we computed the maximal difference in the probability of responses mentioning the target concept between paraphrased prompts of the category/topic (within the first 64 tokens). Adversaries may paraphrase a low-probability base prompt to find the high-probability prompts to achieve the goals described in §3.

We summarize our results in Fig. 8 and Fig. 9. These violin plots illustrate the distribution density of absolute difference between likelihoods (of mentioning a target concept between paraphrased prompts). The wider the violin plot is at a specific height, the more pairs of paraphrased prompts have such an absolute difference in likelihoods. The height of the violin plots suggest the range the absolute difference can be up to. On both plots (i.e., for both brands and societal concepts), all violin plots show a range of likelihood

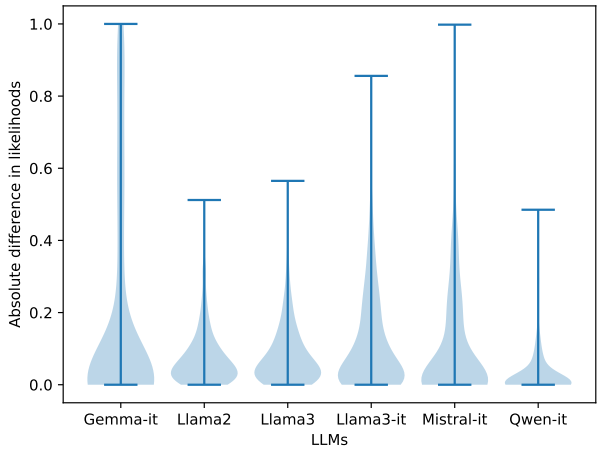

Figure 8: Absolute difference in the likelihoods of responses mentioning a target brand within the first 64 tokens generated in response to paraphrased prompts. Paraphrasing prompts leads to an absolute improvement in the likelihood of LLMs mentioning target brands of up to 100% (i.e., one prompt elicits responses that never mention the target brand while another prompt's responses always do).

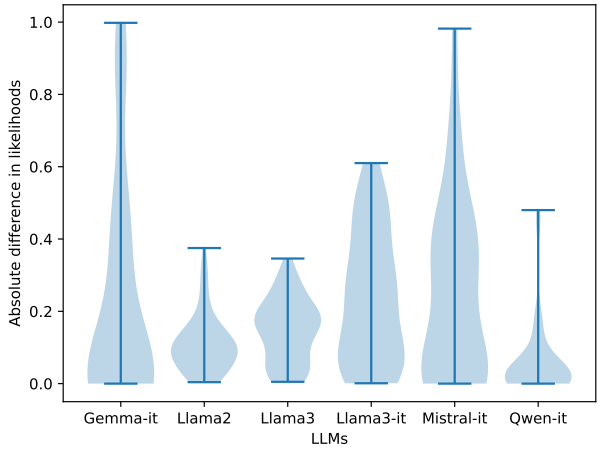

Figure 9: Absolute difference in the likelihoods of responses mentioning a target societal concept within the first 64 tokens generated in response to paraphrased prompts. Paraphrasing prompts leads to an absolute improvement in the likelihood of LLMs mentioning a target societal concept of up to 99.8%.

differences, suggesting it is possible for adversaries to manipulate the likelihoods by using paraphrased prompts. Specifically, on six open-source LLMs (§4.2.1), the likelihood of the target concept being mentioned in the responses can differ between a pair of paraphrased prompts by up to 100.0%. When comparing responses to "*I'm curious to know your preference for the pressure cooker that offers the best combination of cooking performance, durable construction, and overall convenience in preparing a variety of dishes.*" with the

prompt paraphrased as "*Can you recommend the ultimate pressure cooker that excels in providing consistent pressure, user-friendly controls, and additional features such as multiple cooking presets or a digital display for precise settings?*", the probability of Gemma-it mentioning the brand "InstantPot" ("pressure cooker" product category) within the first 64 tokens of the response went from 0% to 100.0% (i.e., went from never mentioning "InstantPot" to always mentioning "InstantPot"). Among the six models, the average of this absolute difference in likelihood is 3.6%–18.6% for brands and 5.9%–25.2% for societal concepts.

Our results suggest that while paraphrased prompts appear similar to humans, the responses to the prompts can differ substantially in how likely they are to mention a target concept (within the first 64 tokens of the response). Therefore, an adversary wanting to promote a certain concept can use this to their advantage: the adversary may try various paraphrases of prompts, test the prompts, and pick the paraphrase that results in the highest probability of the target concept being mentioned, ultimately promoting the concept when the perturbed prompt is used in the real world. While we generated these paraphrases using ChatGPT, and confirmed that they were valid and reasonable, adversaries may also be able to create paraphrases manually or by another method, and may be able to test even more paraphrases than we did in our measurements.

4.3.2 Synonym-Replaced Adversarial Prompts. As opposed to paraphrasing, in the synonym-replacement approach, we automatically generated a set of potential candidate prompts by perturbing a base prompt via synonym replacement, without needing to confirm whether these perturbed prompts are valid. The prompt with the lowest loss was selected, and we assessed how well these selected prompts increase the probability that the target concept is mentioned. We evaluated the average improvement over the probabilities of different concepts being mentioned in the base prompt. We used the loss as a metric that narrowed down the large set of potential perturbed prompts we find using synonym replacements to one. Therefore, we are interested in exploring the *highest* improvement we can achieve between a base and perturbed score, as adversaries would be able to use a similar method to §4.3.1 and explore multiple perturbed prompts. So, we also describe the largest increases in the probability of mentioning a target concept in the responses to the perturbed prompt we find via our synonym-replacement method compared to the base prompt. Our evaluation shows that our synonym-replacement attack increased the likelihood of LLMs mentioning a concept on average. For all evaluations, we focus on the first 64 tokens of the response (see §4.2.1 for details).

Average improvement Overall, we found that our attack results in an average absolute improvement in all models. We took the average over various prompts and target concepts for the absolute improvement (i.e., the difference in likelihoods after synonym replacement, see §4.2.5 for more details) on each of the six models. Specifically, for concepts that were mentioned at least once (i.e., 0.1% of the time) and at most 500 times (i.e., 50% of the time) in responses to the base prompt, the six models had an average absolute improvement of 0.43%–2.05% for brands and 0.26%–1.68% for societal concepts, as shown in Fig. 10 and Fig. 11. Our synonym replacement approach achieved a positive average absolute

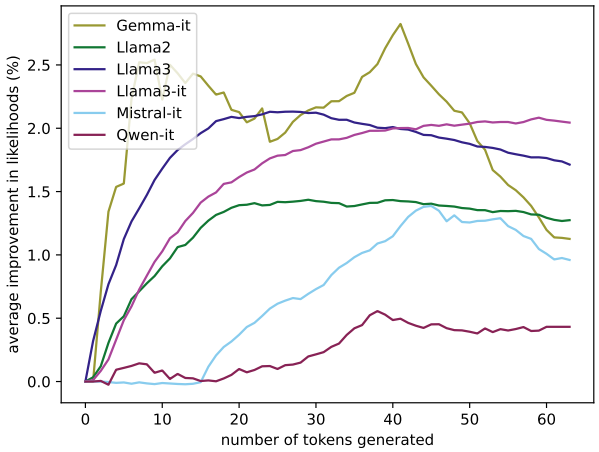

Figure 10: Average absolute improvement in likelihoods that LLMs mention a target brand when the base likelihood is within $[0.1\%, 50\%]$. Results are presented along number of generated tokens. Our synonym-replacement approach achieves improvements in probabilities, which verifies the capability of forcing LLMs to mention target brands more often.

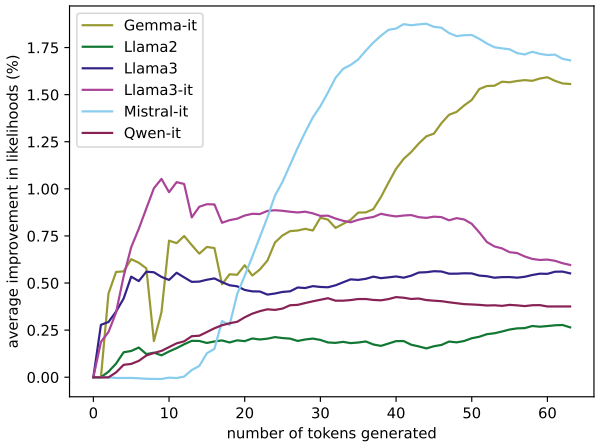

Figure 11: Average absolute improvement in likelihoods that LLMs mention a target societal concept when the base likelihood is within $[0.1\%, 50\%]$. Results are presented along number of generated tokens. Our synonym-replacement approach achieves improvements in probabilities, which verifies the capability of forcing LLMs to mention target brands more often.

improvement in probabilities, which verifies that our approach is capable of forcing LLMs to mention target concepts more often within some number of tokens. On some models, we saw a bigger absolute improvement within the first 64 tokens. As we described in §4.1.2, our proposed approach might not necessarily be the most effective under the threat model (§3), however, our approach illustrates that

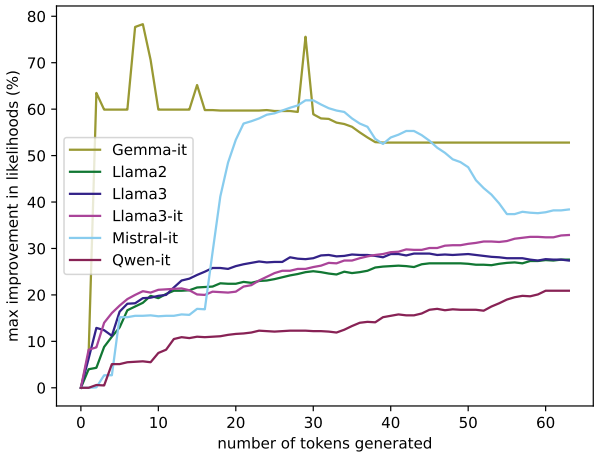

Figure 12: Max absolute improvement in likelihoods that LLMs mention a target brand. Results are presented along how many tokens are generated. We achieve a bigger absolute improvement on Gemma-it compared to the other three Llama models.

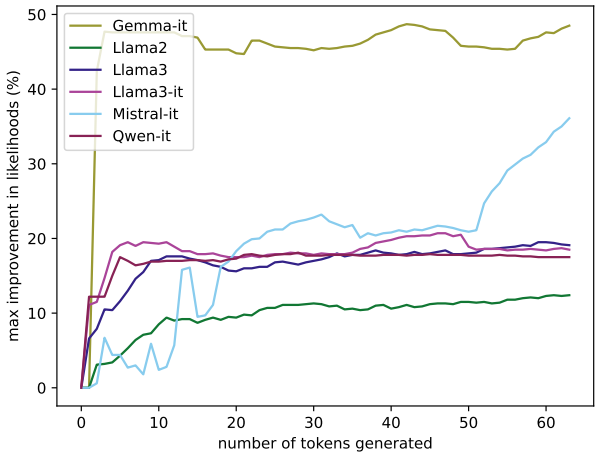

Figure 13: Max absolute improvement in likelihoods that LLMs mention a target societal concept. Results are presented along how many tokens are generated. We achieve a bigger absolute improvement on Gemma-it compared to the other three Llama models.

attacks under our threat model exist, with potentially stronger variations possible. While we did not see improvements in all pairs (for example, some base prompts had near 100% probabilities of mentioning certain concepts, not allowing for any improvement), we still saw pairs with significant increases when considering long responses. For Llama3-it, the perturbed prompt "*Can you recommend the superior video game console that excels in providing top-notch graphics, dissimilar gaming options, and additional features such as online connectivity, appropriate for both casual and hardcore gamers?*" was 55.9% more likely than its base prompt "*Can you recommend the ultimate video game console that excels in providing top-notch graphics, diverse gaming options, and additional features such as online connectivity, suitable for both casual and hardcore gamers?*" (i.e., "ultimate" was changed to "superior" and "diverse" was changed to "dissimilar") to mention Xbox in long completions, even more than when completions were only 64 tokens long (32.9%). We evaluated if attack objectives are met with full responses in a more realistic setting in §5.2.2.

Maximum improvement Besides the average absolute improvement, we also explored the maximum absolute improvement among all combinations of base prompts and concepts, as shown in Fig. 12 and Fig. 13 for brands and societal concepts respectively. While these results do not represent the *expected* improvement using this method, they do demonstrate that it is possible to find pairs of prompts with vastly different probabilities of mentioning a certain concept. In §4.3.1 we showed this for prompts that were manually paraphrased; here, we show that the synonym-replacement attack can find such alternative prompts automatically. Further, the prompts generated by synonym replacement differ from their base prompts minimally (at most a seven-synonym difference in our dataset), and are perceptually the same along multiple dimensions (see §5.2.2). We were able to achieve a much higher max absolute improvement compared to the average absolute improvement, suggesting the synonym replacement works exceptionally well with specific prompts and synonyms. We will explain the implication of this observation in §6.3.

5 User Study: Verifying Practical Attack Success

So far we have shown that our synonym-replacement approach biases LLM responses towards a target concept and appears to be inconspicuous. However, this does not necessarily indicate practical attack success. To evaluate the attack in a realistic setting, we conduct a user study, detailing our methods (§5.1) and statistical evaluation (§5.2). We discuss implications throughout our results and takeaways (§5.2.3). We find that our attack is indeed successful in inconspicuously pushing users toward chosen concepts.

All human-subjects procedures were approved by the Carnegie Mellon University institutional review board.

5.1 Methods

This between-subjects study evaluates the effectiveness of our proposed attack (see §3), creating inconspicuous prompts that trigger inconspicuous responses, making a target brand more noticeable.

Some of the topics we explored in previous sections might be controversial and cause emotional harm. As such, we focused the user study on the shopping task, a benign topic that users are exposed to every day. Similar setups have been used in prior work [86]. Testing each prompt-and-response pair from our previous experimentation would be cost prohibitive. Thus, we limited our study to six pairs of prompts, each from a different product category and evenly split between two models (see §5.1.2).

We tested whether users could distinguish between base and perturbed prompts and responses in multiple dimensions (including clarity, likelihood of use, satisfaction, and more). Each participant was shown one prompt-and-response pair.

Fully informed consent was obtained from participants and our procedures were approved by the Carnegie Mellon University institutional review board. Following advice from prior work [19, 20], we pre-registered our study.[3]

5.1.1 Survey procedures. We recruited a gender-balanced sample from Prolific,[4] a commonly used platform for security-relevant user studies [1]. To reduce selection bias, we avoided mentioning LLMs or bias in the study title, "Chatbot prompting study." Participants had to be in the U.S., be 18 or older, and have an approval rate of at least 95%. Obeying Prolific guidelines,[5] participants were paid $1.6 for an estimated 7–9 minute study, averaging $12.45/hour.

After providing consent, participants were first given an overview. Next, we asked our participants to imagine that there was a chatbot service that was able to recommend prompts appropriate for what users want to use the chatbot for. We further instructed participants to imagine that they were shopping for a certain product (e.g., laptops) or service (e.g., parcel delivery) and wanted to use this new chatbot service to help them decide on a brand, similar to the use case of Amazon Rufus [67]. After a comprehension check on instructions, participants were shown a prompt recommendation for the product they were shopping for (e.g., "*Which laptop model do you consider the optimal choice for versatile computing, powerful performance, and innovative features that enhance your work and entertainment experience?*") and instructed to review it. After a minimum of 10 seconds had passed, we asked (1) how likely they were to use this prompt, (2) how clear the prompt was, (3) was it biased to a certain brand (and which), (4) and if anything stood out (e.g., unexpected).

Participants were then asked to imagine they had chosen to use the prompt and presented a response to the prompt word by word, mimicking chatbots. After a minimum of 20 seconds, they were asked: (1) how clear the response was, (2) if they were satisfied with the response, (3) how likely were they to take the recommendation in the response, (4) and if anything stood out. In a series of open-ended questions, we additionally asked (1) which brand participants would pick based on this response, (2) what were all the brands recommended, and (3) what the top brand recommendation was. These 11 (numbered) questions form the practical definition of inconspicuousness and increase in target brand perception. They form the basis of our statistical analysis §5.2.2.

To help with recall, participants could hover over relevant questions to reveal the relevant prompts and responses.

Participants self-reported how frequently they give tech advice and used chatbots, if they paid for a chatbot, and their ChatGPT familiarity. The survey concluded with demographic questions, which we summarize in Table 3.

5.1.2 Experimental groups. Our overall goal was to find whether people notice differences between our perturbed prompts (and corresponding responses) compared to base prompts. Due to prohibitive cost, we could not test all of our 1,809 base and perturbed (brand) prompt pairs from §4 in the user study. Instead, we selected six pairs of prompts to use in the user study, each pair from a different product category, giving us 12 prompts total.[6] Because each pair of prompts belonged to a unique product category, we refer to them by their product categories in the results. The prompt pairs were split between two models, three pairs for Llama3-it and three for Gemma-it. We focused on these more user-friendly instruction-tuned models since they display chatbot-like behavior [74]. To increase the realism of the study we set the following criteria when picking which prompt pairs to use:

- Prompts pairs from product categories among the top 50% of product categories that participants reported caring about. This increases the chances that participants would have shopped for the product outside an experimental environment.
- Prompts pairs from product categories among the top 50% of product categories that participants reported they might use an LLM when shopping for. This increases the chances that participants would have used an LLM when shopping for the product. Combined with the previous criteria, we selected the product categories that are most likely to be investigated with an LLM in a real-world scenario.
- The prompt pairs had the highest probability increase that the target brand is mentioned with the attack, as measured in §4.3.1, mimicking the type of prompts an attacker might choose to deploy for the highest effectiveness.

All participants within a group saw the same prompts, but, mirroring real-world chatbots, each participant was shown a unique full-length response. For each of the 12 prompts, we obtained a random sample of ∼75 model responses from our earlier experiments with the models (§4.3.2). This sample was stratified to ensure the ratio of responses that mentioned the target was the same in this set of ∼75 as it was in the overall set of 1000.

5.1.3 Piloting and preliminary data collection. We piloted our study extensively. We ran a series of preliminary studies to determine whether the study design was feasible and whether participants would encounter issues. To detect potential problems, questions were timed and augmented with meta-questions on how clear the main questions were. We also collected participants' interest in product categories and their likelihood of using chatbots when shopping for these categories. In total, we collected responses from 90 Prolific participants for piloting and 63 for product category preferences. We further ran pilots with two HCI researchers, asking them to review and criticize our study.

Based on responses, we made the study more concise with clearer instructions. Mimicking chatbots, we changed how prompts and responses were displayed. Attention and comprehension checks were added to ensure high data quality.

5.1.4 Statistical analysis. For each of the six base and perturbed prompt pairs, we analyzed the difference between base and perturbed group responses with a series of non-parametric tests on our 11 main measurement variables: two-tailed Mann-Whitney U tests for Likert data and chi-squared tests for binary data. To understand whether the base and perturbed groups are equivalent,

[3] https://osf.io/6mycr/?view_only=face90d04806439bb1f69fc110fb9a1e
[4] https://www.prolific.com/
[5] https://researcher-help.prolific.com/hc/en-gb/articles/4407695146002-Prolific-s-payment-principles

[6] We aimed to detect "medium" effect sizes with 80% power at $\alpha = 0.05$ for two tailed Mann-Whitney U, requiring ∼75 participants per group.

	Prompt				Reponse			
	Clarity (L7)	Use (L7)	Bias (L5)	Standout (B)	Clarity (L7)	Use (L7)	Satisfied (L7)	Standout (B)
TV	-0.23$^{*}_{*}$	-0.15$*$	-0.03$^{**}_{**}$	-0.08$^{**}_{**}$	0.06$^{**}_{**}$	0.00$*$	0.05$*$	-0.02$^{**}_{**}$
ISP	-0.28$^{*}_{*}$	-0.04$^{*}_{**}$	-0.06$^{**}_{**}$	-0.07$^{**}_{**}$	-0.19$^{**}_{**}$	-0.08$*$	-0.15$^{*}_{*}$	-0.10$^{**}_{**}$
Parcel delivery	-0.15$^{*}_{**}$	-0.58	0.04$^{**}_{**}$	-0.06$^{**}_{**}$	0.04$^{**}_{**}$	{0.41$*$}	{0.55$*$}	-0.06$^{**}_{**}$
Gaming console	{-0.8$^{*}_{*}$}	-0.10$*$	-0.01$^{**}_{**}$	0.00$^{**}_{**}$	-0.17$^{**}_{**}$	-0.21	-0.21$^{*}_{*}$	-0.01$^{**}_{**}$
Investment plat.	0.01$^{*}_{*}$	-0.10$^{*}_{**}$	-0.07$^{**}_{**}$	0.00$^{**}_{**}$	-0.02$^{**}_{**}$	0.20$^{*}_{*}$	0.03$^{*}_{*}$	-0.02$^{**}_{**}$
Laptop	-0.04$^{*}_{*}$	{0.35$*$}	-0.0$^{**}_{**}$	0.04$^{**}_{**}$	0.10$^{**}_{**}$	0.20$*$	0.22$^{*}_{*}$	0.04$^{**}_{**}$

Table 1: Mean difference between base and perturbed groups among eight questions, four about prompts and four about responses (see §5.1.1 for details). By default test for equivalence is reported (TOST WMU), {} indicates test for difference (MWU, CHI^2). Higher is better for all differences. $*$: $p < 0.05$, $^{*}_{*}$: $p < 0.01$ $^{*}_{**}$: $p < 0.001$, $^{**}_{**}$: $p < 0.0001$. L7: seven-point Likert, L5: five-point Likert, B: binary.

	Gemma-it									LLama3-it								
	TV			ISP			Parcel delivery			Gaming console			Invesment plat.			Laptop		
	P%	A%	T%	P%	A%	T%	P%	A%	T%	P%	A%	T%	P%	A%	T%	P%	A%	T%
Base	25.0	33.8	35.3	23.9	46.2	34.3	10.6	22.7	13.6	0.0	18.3	0.0	14.1	38.0	14.1	5.3	61.3	1.3
Pert	57.1	80.0	82.9	39.1	78.2	50.7	41.4	58.6	44.3	11.2	64.8	1.4	40.3	58.4	42.9	7.1	74.3	0.0
Diff	32.1$^{*}_{**}$	46.2$^{**}_{**}$	47.6$^{**}_{**}$	15.2	31.9$^{*}_{**}$	16.4	30.8$^{*}_{**}$	35.8$^{*}_{**}$	30.6$^{*}_{**}$	11.2$*$	46.5$^{**}_{**}$	1.4	26.1$^{*}_{*}$	20.4$*$	28.8$^{*}_{**}$	1.8	12.9	-1.3

Table 2: % of responses mentioning the targeted brand. P: What brand the participant (p)icked given a response, A: What are (a)ll recommended brands the participant found, T: What the (t)op recommended brand is to the participant. $*$: $p < 0.05$, $^{*}_{*}$: $p < 0.01$ $^{*}_{**}$: $p < 0.001$, $^{**}_{**}$: $p < 0.0001$.

Gender	Male	48.2
	Female	50.3
	Self-described	0.9
Age	18-25	16.7
	26-35	36.0
	36-45	21.5
	46-60	18.2
	61+	5.2
Ethnicity	White	63.0
	Black or African Am.	12.1
	Asian	9.8
	Hispanic or Latino	5.6
	Other or mixed	12.7
Education	Completed H.S. or below	10.4
	Some college, no degree	18.0
	Trade or vocational	2.5
	Associate's degree	11.1
	Bachelor's degree	39.1
	Master's or higher	18.5
Chatbot usage frequency	Daily or more freq.	16.6
	Daily to monthly	49.7
	Monthly or less freq.	33.7
ChatGPT familiarity	A lot	57.2
	A little	40.7
	Nothing at all	2.1

Table 3: Demographics. May not total 100% (rounding, opt-outs).

we replaced non-significant tests for difference with tests of equivalence. To establish equivalence, we used the two one-sided tests procedure (TOST) and set our equivalence margin to be $\Delta = 0.5$ for 80% power [56].[7] Brand-recall questions (e.g., brand participants pick based on the response), were coded into binary categories: was the targeted brand reported or not. Hesitant or unclear responses did not count as matches.

Because of our extensive testing (11 tests between the base and perturbed prompt per product category), we controlled our false-discovery rate per product category with the Benjamini-Hochberg procedure [9].

5.2 Results

Our data show that, under most measures, perturbed prompts and corresponding responses are inconspicuous. Further, these prompts successfully nudge more users into noticing the target brand in most measures, fulfilling the adversarial objectives.

5.2.1 Participants. We recruited 845 participants and evenly distributed them to 12 groups, each group defined by the product category and prompt type (base or perturbed). Product categories were split between Llama3-it and Gemma-it. Our participants were more educated, younger, less Hispanic, and had more familiarity with ChatGPT than the national average [76, 92]. Table 3 summarizes the demographics of our participants.

[7] Testing for equivalence is a deviation from the pre-registration; however, we believe this approach paints a more complete picture.

5.2.2 Statistical Evaluation. We asked 11 core questions about the prompts and responses. Four of these were to test the inconspicuousness of the perturbed prompts. Another four were to test if users were significantly more dissatisfied with the perturbed responses. We further asked three questions to test if our perturbations made the targeted brand more perceptible. Table 1 and Table 2 show the results.

Equivalence From participants' perspectives, nearly all perturbed prompts and responses were equivalent to corresponding base prompts and responses in terms of variables measured (42/48 comparisons were equivalent, $p < 0.05$) with the following exceptions: statistical tests showed no difference or equivalence for the likelihood of using the parcel delivery prompts ($p > 0.05$); responses to the parcel delivery perturbed prompt were more satisfactory and more likely to be used (all $p < 0.05$); the perturbed gaming console prompt was less clear than the base prompt ($p < 0.05$); no difference nor equivalence was found for the likelihood of using the response for gaming platforms ($p > 0.05$); the laptop perturbed prompt was more likely to be used ($p < 0.05$). These results suggest that not only is our attack imperceptible to users, but in a few cases the perturbed prompts and responses might be preferable.

Attack success In order to measure attack success, we recorded the percentage of participants who *would pick* the targeted brand given the response, the percentage of participants who *noticed* the brand in the response, and the percentage of participants who said the targeted brand *was the top recommendation.* As shown in Table 2 , in five out of six categories our attacks were able to increase the prominence of the target brand in at least one of the three questions: in four categories participants were more likely to pick the targeted brand when given the perturbed prompts, in five categories participants were more likely to notice the targeted brand, and in three categories participants were more likely to say the targeted brand was the top recommendation.

Effect of earlier brand appearance Attack success is likely dependent on various factors, including how early the target is mentioned in the response. We made this assumption when generating perturbed prompts through synonym replacement. Thus, we used a loss function that increases the probability of a target concept being mentioned at the beginning of the response. Here, we run an exploratory (not pre-registered) analysis on the user study data to test this assumption.

Specifically, we ran a series of logistic regressions to predict whether participants would notice the target brands based on how soon the target brand was mentioned in the response. We formulated the model where correctly noticing the target brand is the dependent variable and the position of the target brand in the response is the independent variable (if "Verizon" is the 15th word, the position would be 15). We included a random effect for the prompt used, since product categories and base prompt phrasing might have different baseline effects. We find that, later mentions of target brands in responses are less likely to be noticed by participants. This observation remains true for all three measurements of attack success: what brand is picked, what the top perceived recommendation is, and if the brand is perceived as recommended by the LLM. A 10-word increase in the position of the target brand reduces the chances of users noticing the target brand by 1.7%, 5.9%, and 0.6% respectively (all $p < 0.0001$).

5.2.3 Takeaways. We show that our synonym-based attack can practically shift users towards a concept of the attacker's choosing. In nearly all measures, perturbed prompts and responses were statistically indistinguishable—or even occasionally preferred—compared to their base counterparts. Meanwhile, in five of six product categories, participants were more likely to notice or choose the attacker's target brand when given perturbed prompts. Taken together, this study verifies the assumptions and results of our earlier experiments. Though we only tested this attack on a benign case (brands) to avoid harm, the results show potential for harm and serious implications on user autonomy, which we discuss next.

6 Discussion

We identify a novel threat model in which an adversary aims to induce biases in LLM responses users receive by using adversarially crafted but innocuous-seeming prompts. We empirically show that, when generated with the right method, users do not notice that these prompts are adversarial but are influenced by the biased responses that the prompts result in. Here, we first discuss the implications of our findings (§6.1) and suggest potential defenses (§6.2). Then we attempt to explain why our attacks work (§6.3). Finally, we end with an economic analysis in §6.4.

6.1 Implications

This risk we highlight in this work stems from real-world deployments of LLMs in chatbots and the accompanying prompt providers, e.g., Amazon Rufus [67]. As such, it has real-world implications. Users may be manipulated into thinking a certain way, while being under the impression that they are receiving unbiased advice. Prompt providers could be giving the impression of a personalized experience, while subtly undermining user autonomy. This risk can fundamentally be thought of as a risk of intentional algorithmic bias, which is well understood to be particularly dangerous in the context of political and social issues [10, 97], but also concerning in commercial contexts [3]. As outlined in prior work, this type of misuse can produce effective propaganda and misinformation, resulting in emotional and financial distress [97]. Notably, LLMs already disseminate propaganda [71]–prompt providers may follow suit at a lower cost. We argue that the seriousness of this risk necessitates defensive measures, likely requiring an ensemble of defenses to be effective.

6.2 Defenses: Multi-Pronged

We identify a series of complementary defenses—with various tradeoffs between effectiveness and deployment cost—that different stakeholders can employ to mitigate the risk of adversarial prompts.

User warnings, labeling, and education Using prompts from untrusted sources is akin to running code copied from untrusted sources, a well-studied problem in computer security [6, 32]. Thus, similar protection mechanisms, like warnings [33], might be effective. However, unlike untrusted code, inconspicuous attacks are difficult to detect and therefore might require more invasive warnings. Security warnings have a long history in human-centered

design [12, 40, 79, 109], showing they can be integrated into the user interface to possibly help users make more informed decisions. These warnings could appear in various locations depending on the owner of the prompt provider. If the prompt provider is not the chatbot owner, the chatbot could warn users about the potential risks of using prompts from untrusted sources. Notably, certain chatbots already warn users, e.g., ChatGPT warns, "ChatGPT can make mistakes. Check important info." [21]. Further, prompt libraries can warn users about the potential risks of using prompts from other users, similar to proposed warnings in programming forums [33].

Warnings tend to be reactive, appearing only when an emergency arises. In contrast, users can be proactively educated with "nutrition" labels [52], a standardized summary of what users need to pay attention to (e.g., privacy practices) before using a product. Notably, privacy-relevant nutrition labels have had widespread adoption[8] and has been shown to be effective in multiple computing contexts, including mobile applications and IoT devices [7, 28]. Similar labels for the capabilities and pitfalls of LLMs could be developed and publicized.

Labels and warnings, however, are likely only part of the solution. More involved user education campaigns on how to use LLMs and chatbots safely and effectively might need to be developed [107].

Robust models Our attacks fundamentally rely on LLMs (like other deep neural networks) being brittle, a fact we also observe in our work (§4.3.1). They respond to small changes in prompts in a way that is unexpected and can be manipulated. Despite the claims of robustness by many LLM providers, our work (and many others) show that there is still much work to be done. Existing robust LLMs focused on the correctness of LLM answers [47, 55, 75, 78, 96, 108]. To the best of our knowledge, we are the first to explore how this model fragility allows slight (inconspicuous) perturbation in LLM prompts to vastly different probabilities of LLMs mentioning target concepts. Such differences may lead to bias and risks that hurt user autonomy, including but not limited to advertising without user awareness, misinformation, and propaganda, as we showed in this paper. We suggest LLM providers emphasize robustness against these defenses. However, given the rapid—perhaps rushed—deployment of LLMs in the wild (e.g., [82]), it might be up to regulators to enforce robustness requirements [50].

New bias metrics Our work does not use prior definitions of bias [8, 98] during evaluation. This is intentional: existing bias metrics focus on discrimination, hate speech, and exclusion, but do not capture the type of bias we introduce in this work. For instance, many brands could be mentioned in an LLM response, and an ideal bias metric should be able to capture how much each one can be biased. This suggests that new metrics need to be developed to capture this type of bias, enrich the definition of LLM robustness, and eventually, make testing meaningful.

Continuous audits In the absence of guaranteed robustness, an elusive target, frequent testing of models for bias and other risks is essential. We advocate for systematic testing of prompt providers for bias, including the type of bias we introduce in this work. Such a system could regularly check if prompts are biased towards certain concepts, and if so, alert users. The effectiveness of such a system would fundamentally depend on the quality of the bias detection tests.

We further find that running measurements on LLMs is difficult due to the variability in responses. While this is a feature that LLM developers intentionally build, it also complicates the measurement of bias. How can we measure the underlying bias in the model? Is simply averaging numerous responses enough? We advocate for future research to explore this question.

6.3 Why Does the Attack Work?

As we illustrate in §4.3.2, our synonym replacement method has a much higher max absolute improvement compared to the average absolute improvement, suggesting the approach works particularly well with specific prompts and synonyms. We hypothesize that one reason this may be caused by the synonym occurring in close proximity to the promoted concept in training instances. For example, when the prompt includes the word '*reliable*' and asks LLMs to recommend a streaming service, LLMs almost always recommend the brand "'*Netflix*", and '*Netflix*' is commonly associated with '*reliable*' in online text,[9] while other streaming services are less often. However, we are not able to identify such words (or word combinations) for every prompt where our synonym replacement works particularly well, suggesting that there may be other factors at play. Our attacks can be thought of as adversarial examples, which are not fully understood.

6.4 Economic Analysis

As we described in §3, adversaries may run attacks we discuss in this paper to perform advertising without user awareness. Prompt-optimization engines may advertise products on behalf of others to generate revenue. In the following paragraphs, we briefly analyze the potential economic incentives of such attacks, we list benefits and current costs to run attacks (i.e., synonym replacement and paraphrasing) compared to more traditional advertising methods. We also compute the number of increased mentions of the target concept needed for adversaries to make profits.

At the time of writing, the cost per mille (CPM), i.e., price to show an ad 1,000 times is \$4 – \$10 on Meta, Youtube, Snapchat, and TikTok[10], making the price to show a single ad \$0.004 – \$0.01. The price for running an LLMs is \$0.04 – \$30 for one million input tokens and \$0 – \$75 for one million output tokens[11]. Adversaries need to pay the cost for generating the adversarial prompt (either by paraphrasing or synonym replacement). The victims will take the adversarial prompt and run on LLMs by themselves (as we defined in §3) so there is no additional cost after the prompt is released, regardless of how many times the adversarial prompts are used.

Our prompts are no longer than 400 tokens. The synonym replacement (§4.1.2 and §4.3.2) only needs to perform inference once (i.e., compute one output token), and therefore, the cost can be up

[8] Google Play and Apple App Store made privacy nutrition labels mandatory for apps.

[9] E.g., https://netflixtechblog.com/keeping-netflix-reliable-using-prioritized-load-shedding-6cc827b02f94, https://www.forbes.com/sites/rosaescandon/2020/05/19/netflix-is-the-most-reliable-streaming-service-new-survey-shows, and https://www.infoq.com/articles/netflix-highly-reliable-stateful-systems/. All visited on Dec 3, 2024.

[10] According to https://www.guptamedia.com/social-media-ads-cost, visited on Dec 2 2024.

[11] According to https://llm-price.com/, visited on Dec 2 2024.

to $\$0.15/1,000,000 * 400 + 0.20/1,000,000 = 0.0000602$ for Mistral and $\$30/1,000,000 * 400 + 75/1,000,000 = 0.012075$ with the most expensive token prices. Adversaries only need one more mention of the target concept (as \$0.0000602 is less than \$0.004) on Mistral or up to $0.012075/0.004 \approx 3$ more mentions of the target concept for any other models to have more benefits than cost. The synonym replacement achieves more than 0.3% average absolute improvement on each model we used (Fig. 10), i.e., more than three more mentions of the target concepts within 1,000 uses of the prompt. Thus, adversaries can run much more efficient advertising using our synonym replacement approach.

In contrast, if adversaries only use paraphrasing (§4.1.1), the cost is much higher: with the same setup we used (§4.2.4), adversaries may need to generate up to 64 tokens 1,000 times for an average of $2,207/524 \approx 4.21$ prompts (including the unperturbed prompts and paraphrases). The cost can be up to $\$(0.15/1,000,000 * 400 + 0.20/1,000,000 * 64) * 1,000 * 4.21 = 0.306488$ on Mistral and $\$(30/1,000,000 * 400 + 75/1,000,000 * 64) * 1,000 * 4.21 = 70.728$ in the worst case scenario. Up to 77 more mentions on Mistral and 17,682 more mentions of the target concept in the worst case on other models are needed for the adversaries to have more benefits than costs. As we suggested in §4.1.1, this makes paraphrasing more time- and cost-intensive than synonym replacement. The number of more mentions we compute are all upper bounds, and whether adversaries can make profits by only paraphrasing depends on specific prompts, models used, and the number of times that the prompts will be used.

7 Limitations

We explore methods to make LLMs mention a specific concept while remaining inconspicuous to users and our results may depend on many factors. The concepts and brands we chose might not be realistic of real world use-cases. To address this, we make the case that the use of LLMs when shopping is already a reality and collect a list of products that users had the highest likelihood of shopping for with LLMs (§5.1.2).

Our LLM use is bound to specific temperature settings. We address our temperature choice in §4.2.1 and explore the effect of different temperature settings in App. A.3.

We also were only able to test our attack on a small set of LLMs, all of which are open-source. This means we cannot evaluate the success of our attack on all LLMs, including popular closed-source models. We address this by exploring transferability in §A.2.

Our assumptions of what change in resulting prompts would influence users might be flawed. We test (some of) our assumptions in the user study (§5) and find them to be reasonable.

Like all user studies, ours is limited in a multitude of ways. Our sample might not be representative of general LLM users. However, we only find a small minority of our participants to not be familiar with chatbots (only 2.1% knows nothing about ChatGPT). Though we phrased the initial study advertisement generically, participants might have self-selected themselves, limiting generalizability. Our study was run on a U.S. population and in English, not representative of the global population. Despite these limitations, we believe our study provides valuable data for an underexplored problem.

8 Conclusion

We identify a novel threat model in which adversaries bias LLM responses in a target direction by suggesting subtly altered prompts to unsuspecting users. Through a series of experiments and a user study, we develop this attack and demonstrate that we can bias LLM responses towards a target concept while remaining inconspicuous to humans. We further show that this attack can be used to bias LLMs in harmful and helpful ways. These findings highlight the risk of adversaries introducing bias in LLM responses in a unique way, suggesting a need for defensive measures like warnings and robustness requirements.

Acknowledgments

The work described in this paper was supported in part by the U.S. Army Research Office under MURI grant W911NF-21-1-0317 and by the Future Enterprise Security Initiative at Carnegie Mellon CyLab (FutureEnterprise@CyLab).

This research was supported by the Center for AI Safety Compute Cluster. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

[1] Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L Mazurek. 2021. Comparing Security and Privacy Attitudes Between iOS and Android Users in the US. In *SOUPS 2021: USENIX Symposium on Usable Privacy and Security*.

[2] Jina AI. 2024. PromptPerfect - AI prompt generator and optimizer. https://promptperfect.jina.ai/ Accessed: 2024-06-06.

[3] Shahriar Akter, Yogesh K Dwivedi, Shahriar Sajib, Kumar Biswas, Ruwan J Bandara, and Katina Michael. 2022. Algorithmic Bias in Machine Learning-Based Marketing Models. *Journal of Business Research* 144 (2022).

[4] Anthropic. 2024. Automatically Generate First Draft Prompt Templates. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator. Accessed: 2024-09-07.

[5] Kalya Win Aung, Ewan Soubutts, and Aneesha Singh. 2024. "What a stupid way to do business": Towards an Understanding of Older Adults' Perceptions of Deceptive Patterns and Ways to Develop Resistance. *Proc. ACM Hum.-Comput. Interact.* (2024). https://doi.org/10.1145/3677113

[6] Wei Bai, Omer Akgul, and Michelle L Mazurek. 2019. A Qualitative Investigation of Insecure Code Propagation from Online Forums. In *2019 IEEE Cybersecurity Development (SecDev)*. IEEE.

[7] David G Balash, Mir Masood Ali, Chris Kanich, and Adam J Aviv. 2024. " I would not install an app with this label": Privacy Label Impact on Risk Perception and Willingness to Install {iOS} Apps. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*.

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell.l. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

[9] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* (1995).

[10] Su Lin Blodgett, Q Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*.

[11] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* 4 (2016).

[12] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. 2013. Your Attention Please: Designing Security-Decision UIs to Make Genuine Risks Harder to Ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*.

[13] Rafael A Calvo, Dorian Peters, Karina Vold, and Richard M Ryan. 2020. Supporting human autonomy in AI systems: A framework for ethical enquiry. *Ethics of Digital Well-Being: A Multidisciplinary Approach* (2020).

[14] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson,

Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*.

[15] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*.

[16] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

[17] Weichen Joe Chang, Katie Seaborn, and Andrew A. Adams. 2024. Theorizing Deception: A Scoping Review of Theory in Research on Dark Patterns and Deceptive Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery. https://doi.org/10.1145/3613905.3650997

[18] Ishita Chordia, Lena-Phuong Tran, Tala June Tayebi, Emily Parrish, Sheena Erete, Jason Yip, and Alexis Hiniker. 2023. Deceptive Design Patterns in Safety Technologies: A Case Study of the Citizen App. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581258

[19] Lewis L Chuang and Ulrike Pfeil. 2018. Transparency and Openness Promotion Guidelines for HCI. In *Extended abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*.

[20] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

[21] Alex Cranz. 2024. We have to stop ignoring AI's hallucination problem. *The Verge* (2024). https://www.theverge.com/2024/5/15/24154808/ai-chatgpt-google-gemini-microsoft-copilot-hallucination-wrong

[22] Francesco Croce and Matthias Hein. 2020. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. In *International Conference on Machine Learning*.

[23] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. https://arxiv.org/abs/2209.01390

[24] Anupam Das, Nikita Borisov, and Matthew Caesar. 2016. Tracking Mobile Web Users Through Motion Sensors: Attacks and Defenses. In *Network and Distributed System Security Symposium*.

[25] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. 2019. Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks. In *28th USENIX Security Symposium (USENIX Security 19)*.

[26] Samuel F. Dodge and Lina Karam. 2017. A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions. In *International Conference on Computer Communication and Networks*.

[27] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

[28] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2022. An Informative Security and Privacy "Nutrition" Label for Internet of Things Devices. *IEEE Security & Privacy* (2022).

[29] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust Bias Evaluation of Large Generative Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[30] Expedia. 2023. Expedia launches conversational trip planning powered by ChatGPT to inspire members to dream about travel in new ways. https://www.expediagroup.com/investors/news-and-events/financial-releases/news/news-details/2023/Chatgpt-Wrote-This-Press-Release--No-It-Didnt-But-It-Can-Now-Assist-With-Travel-Planning-In-The-Expedia-App/default.aspx Accessed: 2024-06-05.

[31] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[32] Felix Fischer, Konstantin Böttinger, Huang Xiao, Christian Stransky, Yasemin Acar, Michael Backes, and Sascha Fahl. 2017. Stack Overflow Considered Harmful? The Impact of Copy&Paste on Android Application Security. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[33] Felix Fischer, Huang Xiao, Ching-Yu Kao, Yannick Stachelscheid, Benjamin Johnson, Danial Razar, Paul Fawkesley, Nat Buckley, Konstantin Böttinger, Paul Muntean, et al. 2019. Stack Overflow Considered Helpful! Deep Learning Security Nudges Towards Stronger Cryptography. In *28th USENIX Security Symposium (USENIX Security 19)*.

[34] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and Fairness in Large Language Models: A Survey. https://arxiv.org/abs/2309.00770

[35] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

[36] Luis Garcia, Ferdinand Brasser, Mehmet H. Cintuglu, Ahmad-Reza Sadeghi, Osama Mohammed, and Saman A. Zonouz. 2017. Hey, My Malware Knows Physics! Attacking PLCs with Physical Model Aware Rootkit. In *Network and Distributed System Security Symposium*.

[37] Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[38] Yeow Chong Goh, Xin Qing Cai, Walter Theseira, Giovanni Ko, and Khiam Aik Khor. 2020. Evaluating Human Versus Machine Learning Performance in Classifying Research Abstracts. In *Scientometrics*.

[39] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

[40] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. 2020. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

[41] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery. https://doi.org/10.1145/3173574.3174108

[42] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. Dark Patterns in Proxemic Interactions: a Critical Perspective. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. Association for Computing Machinery. https://doi.org/10.1145/2598510.2598541

[43] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2024. Optimizing Prompts for Text-to-Image Generation. *Advances in Neural Information Processing Systems* 36 (2024).

[44] Xuanli He, Qiongkai Xu, L. Lyu, Fangzhao Wu, and Chenguang Wang. 2021. Protecting Intellectual Property of Language Generation APIs with Lexical Watermark. In *AAAI Conference on Artificial Intelligence*.

[45] Xuanli He, Qiongkai Xu, Yi Zengt, Lingjuan Lyu, Fangzhao Wu, Jiwei Li, and Ruoxi Jia. 2024. CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

[46] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. 2017. On the Limitation of Convolutional Neural Networks in Recognizing Negative Images. In *IEEE International Conference on Machine Learning and Applications*.

[47] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2024. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. (2024).

[48] Eojin Jeon, Mingyu Lee, Juhyeong Park, Yeachan Kim, Wing-Lam Mok, and SangKeun Lee. 2023. Improving Bias Mitigation through Bias Experts in Natural Language Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[49] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*.

[50] Joseph R. Biden JR. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ Accessed: 2024-09-09.

[51] Mahammed Kamruzzaman, Hieu Minh Nguyen, and Gene Louis Kim. 2024. "Global is Good, Local is Bad?": Understanding Brand Bias in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 12695–12702.

[52] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A "Nutrition Label" for Privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*.

[53] Anjali Khurana, Hariharan Subramonyam, and Parmit K Chilana. 2024. Why and When LLM-Based Assistants Can Go Wrong: Investigating the Effectiveness of Prompt-Based Interactions for Software Help-Seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery. https://doi.org/10.1145/3640543.3645200

[54] Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender Bias and Stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*.

[55] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2024. Certifying LLM Safety against Adversarial Prompting.

[56] Daniël Lakens. 2017. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social psychological and personality science* (2017).

[57] Katsiaryna Lashkevich, Fredrik Milani, Maksym Avramenko, and Marlon Dumas. 2024. LLM-Assisted Optimization of Waiting Time in Business Processes: A Prompting Method. In *Business Process Management*, Andrea Marrella, Manuel Resinas, Mieke Jans, and Michael Rosemann (Eds.). Springer Nature Switzerland.

[58] Zongjie Li, Chaozheng Wang, Shuai Wang, and Cuiyun Gao. 2023. Protecting Intellectual Property of Large Language Model-Based Code Generation APIs via Watermarks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

[59] Weiran Lin, Keane Lucas, Lujo Bauer, Michael K. Reiter, and Mahmood Sharif. 2022. Constrained Gradient Descent: A Powerful and Principled Evasion Attack Against Neural Networks. In *International Conference on Machine Learning*.

[60] Weiran Lin, Keane Lucas, Neo Eyal, Lujo Bauer, Michael K. Reiter, and Mahmood Sharif. 2024. Group-based Robustness: A General Framework for Customized Robustness in the Real World. In *Network and Distributed System Security Symposium*.

[61] Lowe's. [n. d.]. Lowe's Product Expert: Custom GPT. https://www.lowesinnovationlabs.com/projects/lowe-s-product-expert Accessed: 2024-06-05.

[62] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a Light on Dark Patterns. *Journal of Legal Analysis* (2021). https://doi.org/10.1093/jla/laaa006

[63] Ben Lutkevich. 2024. 19 of the best large language models in 2024. (2024). https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models Accessed: 2024-06-06.

[64] Dominique Machuletz and Rainer Böhme. 2019. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR. *Proceedings on Privacy Enhancing Technologies* (2019). https://api.semanticscholar.org/CorpusID:201646641

[65] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

[66] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445610

[67] Rajiv Mehta and Trishul Chilimbi. 2024. *Amazon announces Rufus, a new generative AI-powered conversational shopping experience.* https://www.aboutamazon.com/news/retail/amazon-rufus Accessed: 2024-08-28.

[68] Microsoft. 2024. Your Everyday AI companion. https://www.microsoft.com/en-us/bing?form=MG0AUO&OCID=MG0AUO Accessed: 2024-06-06.

[69] Swaroop Mishra and Elnaz Nouri. 2023. HELP ME THINK: A Simple Prompting Strategy for Non-experts to Create Customized Content with Models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.751

[70] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference*.

[71] Steven Lee Myers. 2025. DeepSeek's Answers Include Chinese Propaganda, Researchers Say. *The New York Times* (2025). https://www.nytimes.com/2025/01/31/technology/deepseek-chinese-propaganda.html

[72] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376321

[73] Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias Removal With No Demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.).

[74] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* (2022).

[75] Hengzhi Pei, Jinyuan Jia, Wenbo Guo, Bo Li, and Dawn Song. 2024. TextGuard: Provable Defense against Backdoor Attacks on Text Classification. In *Network and Distributed System Security Symposium*.

[76] Pew Research Center. 2024. 2024 pew research center's american trends panel wave 142 topline. https://www.pewresearch.org/wp-content/uploads/2024/03/SR_24.03.26_chat-bot_topline.pdf

[77] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating Adversarial Misspellings with Robust Word Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[78] Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

[79] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. 2018. An Experience Sampling Study of User Reactions to Browser Warnings in the Field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

[80] Qingyang Ren, Zilin Jiang, Jinghan Cao, Sijia Li, Chiqu Li, Yiyang Liu, Shuning Huo, Tiange He, and Yuan Chen. 2024. A Survey on Fairness of Large Language Models in E-Commerce: Progress, Application, and Challenge. https://arxiv.org/abs/2405.13025

[81] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[82] Adi Robertson. 2024. *Google AI's Gemini criticized for generating inaccurate historical content.* https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical Accessed: 2024-09-09.

[83] Tobias Schnabel and Jennifer Neville. 2024. Prompts As Programs: A Structure-Aware Approach to Efficient Compile-Time Prompt Optimization. https://arxiv.org/abs/2309.00770

[84] Patrick Schober, Christa Boer, and Lothar A. Schwarte. 2018. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia* 5 (2018).

[85] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. 2018. On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[86] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*.

[87] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. https://arxiv.org/abs/2010.15980

[88] Craig Smith. 2023. What Large Models Cost You - There Is No Free AI Lunch. *Forbes* (2023). https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/?sh=1f3c5fa34af7

[89] Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. GlobalBench: A Benchmark for Global Progress in Natural Language Processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[90] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[91] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. 2018. When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks. In *27th USENIX Security Symposium (USENIX Security 18)*.

[92] U.S. Census Bureau. 2020. Census Data. https://data.census.gov/

[93] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery. https://doi.org/10.1145/3319535.3354212

[94] Pierre-Antoine Vervier, Olivier Thonnard, and Marc Dacier. 2015. Mind Your Blocks: On the Stealthiness of Malicious BGP Hijacks. In *Network and Distributed System Security Symposium*.

[95] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. 2020. Addressing Marketing Bias in Product Recommendations. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery. https://doi.org/10.1145/3336191.3371855

[96] Chengkun Wei, Wenlong Meng, Zhikun Zhang, Min Chen, Minghu Zhao, Wenjing Fang, Lei Wang, Zihui Zhang, and Wenzhi Chen. 2024. LMSanitator: Defending Prompt-Tuning Against Task-Agnostic Backdoors. In *Network and Distributed System Security Symposium*.

[97] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and Social Risks of Harm from Language Models. https://arxiv.org/abs/2112.04359

[98] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba

Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.*

[99] K. Maria Wolters, Klaus-Peter Engelbrecht, Florian Gödde, Sebastian Möller, Anja Naumann, and Robert Schleicher. 2010. Making it Easier for Older People to Talk to Smart Homes: the Effect of Early Help Prompts. *Universal Access in the Information Society* (2010). https://doi.org/10.1007/s10209-009-0184-x

[100] Qunfang Wu, Yisi Sang, Dakuo Wang, and Zhicong Lu. 2023. Malicious Selling Strategies in Livestream E-commerce: A Case Study of Alibaba's Taobao and ByteDance's TikTok. *ACM Trans. Comput.-Hum. Interact.* (2023). https://doi.org/10.1145/3577199

[101] Tongshuang Wu, Michael Terry, and Carrie J. Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. https://arxiv.org/abs/2110.01691

[102] Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2024. An LLM can Fool Itself: A Prompt-Based Adversarial Attack. In *The Twelfth International Conference on Learning Representations.*

[103] Xi Yang and Marco Aurisicchio. 2021. Designing Conversational Agents: A Self-Determination Theory Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).*

[104] Jin Yong Yoo and Yanjun Qi. 2021. Towards Improving Adversarial Training of NLP Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021.*

[105] Lifan Yuan, YiChi Zhang, Yangyi Chen, and Wei Wei. 2023. Bridge the Gap Between CV and NLP! A Gradient-based Textual Adversarial Attack Framework. In *Findings of the Association for Computational Linguistics: ACL 2023.*

[106] José Pablo Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *International Conference on Foundations of Digital Games.* https://api.semanticscholar.org/CorpusID:17683705

[107] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.*

[108] Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks. (2024).

[109] Bo Zhang, Mu Wu, Hyunjin Kang, Eun Go, and S Shyam Sundar. 2014. Effects of Security Warnings and Instant Gratification Cues on Attitudes Toward Mobile Websites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

[110] Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting Language Generation Models via Invisible Watermarking. In *Proceedings of the 40th International Conference on Machine Learning.*

[111] JJ Zhuang. 2023. Bringing Inspirational, AI-Powered Search to the Instacart app with Ask Instacart. https://www.instacart.com/company/updates/bringing-inspirational-ai-powered-search-to-the-instacart-app-with-ask-instacart/ Accessed: 2024-06-05.

[112] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. https://arxiv.org/abs/2307.15043

Appendix

A Variations on the Attack

In this section, we list experimental results in addition to those in the main body of the paper (§4.3 and §5.2). For the sake of completeness, we evaluate the effectiveness of our synonym replacement approach to reduce the frequency of mentioning a target concept by maximizing the loss function (App. A.1). We are limited by computational resources, and the models we run our evaluations on might not be representative of all LLMs used in the wild. Thus, we explore the transferability of our synonym replacement approach to (closed-source) GPT models (App. A.2). We then evaluate synonym replacement at different settings (App. A.3), and compare attack success with different numbers of synonyms replaced (App. A.4).

A.1 Synonym-Replaced Adversarial Prompts to Mention Concepts Less Often

To systematically verify the correctness of our implementation in §4.1.2, we perturbed the prompts in the product scenario against each brand in the same category, by maximizing (rather than minimizing) the loss function we proposed. Overall, we find that our implementation results in an average absolute decrease in all models as shown in Fig. 14. Specifically, for brands that were mentioned at least 50 times (i.e., 5% of the time) in responses to the base prompt, the six models had an average absolute decrease of 0.19% to 3.10% with in the first 64 tokens generated, and a slightly bigger absolute decrease of 0.31% to 5.47% within a number of tokens less than 64.

Similar to §4.3.2, besides the average absolute decrease, we also explored the maximum absolute decrease among all combinations of base prompts and concepts, as shown in Fig. 15. We were able to

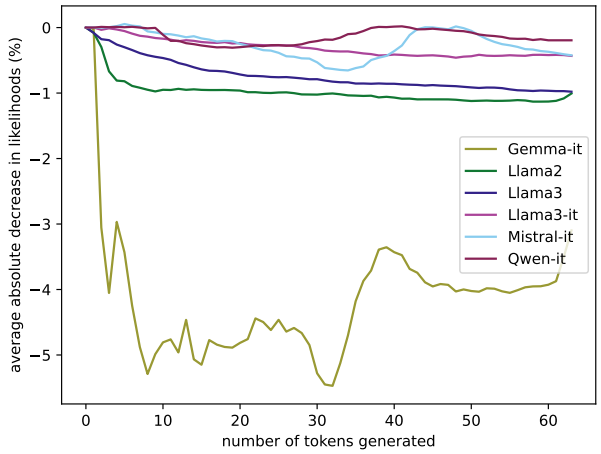

Figure 14: Average absolute decrease in likelihoods that LLMs mention a target brand when the base likelihood is at least 5%. Results are presented along number of generated tokens. Our synonym-replacement approach achieves decreases in probabilities, which verifies the capability of forcing LLMs to mention target brands less often.

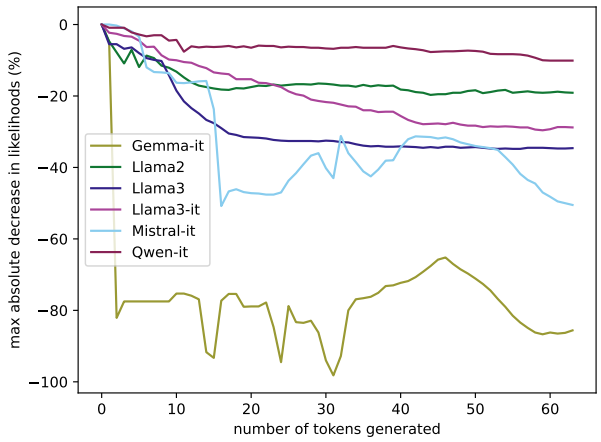

Figure 15: Max absolute decrease in likelihoods that LLMs mention a target brand. Results are presented along how many tokens are generated. We achieve a vaster absolute decrease on Gemma-it compared to the other models.

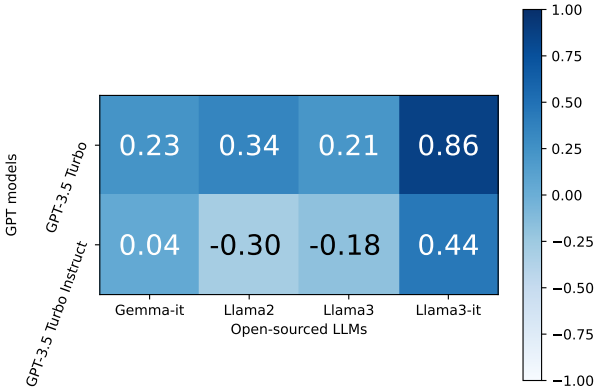

Figure 16: Pearson correlation coefficients (ρ) attack success with GPT models and open-source LLMs. While most GPT/open-source LLM pairs have $\rho < 0.4$, Llama3-it and GPT3.5-Turbo have a correlation coefficient of 0.86%, implying strongly correlation ($p < 0.001$).

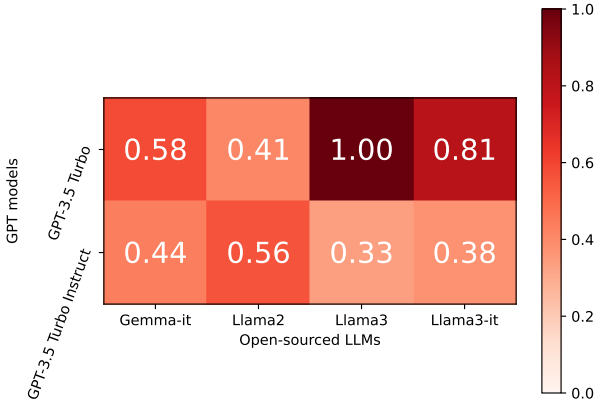

Figure 17: Chances that likelihoods of each pair of LLMs mentioning a target concept move in the same direction (i.e., both increase or decrease) after synonym replacement. For specific pairs (GPT-3.5 Turbo and Llama3, GPT-3.5 Turbo and Llama3-it) of LLMs, such chances are high, indicating that a synonym replacement that works on an open-source model is highly likely to work on a closed-source GPT model.

achieve a maximum absolute decrease in the likelihood of 10.10% to 98.20%. While these results don't represent the *expected* decrease using our approach, they do demonstrate that it is possible to slightly perturb the prompts to mention a target concept less often in LLM responses. The fact that synonym replacement can change the likelihood of LLMs mentioning a target concept in either direction suggests our approach's potential to mitigate existing biases in LLMs, albeit it might not work for all prompts and target concepts, and there is little control of the magnitude of change in likelihoods. For example, prompt providers might use this approach to fight against known biases in LLMs, providing users with prompts that encourage LLMs to generate responses with the likelihoods of mentioning target concepts closer to the distributions expected by users. Additionally, in cases of "negative prompts", there may be an incentive for an adversary to want a topic to show up *less* often.

A.2 Transferability to GPT Models

Some machine learning attacks were found capable of transferring [25, 91]: attacks against a machine-learning model might be effective against a different, potentially unknown, model. We thus investigated whether our synonym replacement approach can transfer to GPT3.5-Turbo, a commercial and closed-source LLM. Due to limited resources, we only measure transferability on four open source models: Llama2, Llama3, Llama3-it, Gemma-it.

On each of the four LLMs, we first ranked the 1,809 pairs of base and perturbed prompts (that recommend brands) according to the absolute improvement in mentioning a target brand. For the top pairs, we collected 1,000 complete responses to both prompts: we allowed LLMs to keep generating until they yielded end-of-sequence (EOS) tokens. We collected responses from the LLM that this prompt pair originally performed well on, as well as GPT3.5 Turbo and GPT3.5 Turbo Instruct. We use the following two metrics to systematically examine the transferability of the synonym replacement attack §4.1.2 from open-source LLMs to closed-source LLMs. We first evaluated how well the change in probability that the target brand is recommended between the base and perturbed prompts correlated between the open source and GPT models. We then evaluated how likely the GPT models were to move in the same direction as the open source models - i.e., if the probability between base and perturbed increased for the open source, whether it also increased for the GPT models.

The base prompt "*Can you recommend the ultimate gas station for fueling up?*" and perturbed prompt "*Can you recommend the premier gas station for fueling up?*" in particular resulted in a large change in the probability of Shell being mentioned in the response. We saw an improvement from 10.3% to 57.4% between the base and perturbed prompt on long responses by Gemma-it between prompts and an improvement from 41.7% to 90.1% on GPT3.5-Turbo responses. On long responses generated by Llama3-it, "*If you had to pinpoint the superior investment platform, which one would it be, and what specific features make it stand out as the top choice for investors?*" had a 48.9% probability of mentioning Fidelity, while "*If you had to pinpoint the premier investment platform, which one would it be, and what specific features make it stand out as the top choice for investors?*" had a 79.3%, and GPT3.5-Turbo responses went from 21.9% to 56.7% for this same pair of prompts.

A.2.1 Pearson Correlation Coefficients between Relative Improvements. We computed the relative improvement in the probabilities of mentioning the target concept within complete responses for each pair. If 20% of responses of the base prompt before perturbing mentioned the target brand, and 50% of responses of the prompt after perturbing mentioned the target brand, the relative improvement was $(50\% - 20\%)/20\% = 150\%$. Then we computed the relative improvement of these pairs on ChatGPT. Specifically, we used the GPT-3.5 Turbo and GPT-3.5 Turbo Instruct model with the default temperature parameter and collected 1,000 complete responses.

GPT-3.5 Turbo Instruct is the instruction-tuned version of GPT-3.5 Turbo. We compare the relative improvements between GPT models and open-source models in §A.2. We found that, while transferability was limited for most model pairs, Llama3-it and GPT3.5-Turbo had a high correlation in attack success for the same prompts ($p < 0.001$). We explain these results in more detail next.

To compare the relative improvements between models when the same base and perturbed (via synonym replacement) prompts are used, we used the Pearson correlation test. Two sets of data with a Pearson correlation coefficient (ρ) less than 0.4 is *generally* believed to be weakly correlated, whereas 0.1 or lower is uncorrelated. On the other hand, a Pearson correlation coefficient larger than 0.7 is believed to indicate strong correlation [84]. A strong correlation indicates that the bigger an open-source LLM's relative improvement is, the bigger the closed-source LLM's relative improvement is given the same synonym-replaced prompts, i.e., the better the synonym replacement can transfer.

As shown in Fig. 16, while Pearson correlation coefficients of the relative improvement of most pairs of open-source LLMs and GPT models indicated weak or no correlation ($\rho < 0.4$), Llama3-it and GPT3.5-Turbo have a correlation coefficient of 0.86, and thus are highly correlated ($p < 0.001$). Llama3-it and GPT3.5-Turbo Instruct also show some correlation, with a coefficient of 0.44.

A.2.2 Probability that Likelihoods of Each Pair of LLMs Mentioning a Target Concept Move in the Same Direction. In addition to the Pearson correlation coefficients, we also measure the probability that the likelihoods of each pair of open-source and closed-source LLMs mentioning a target concept will move in the same direction (i.e., both increase or both decrease) after synonym replacement, as shown in Fig. 17. For this metric, we only compute the sign of the changes in likelihoods but not the magnitude. We find that for specific pairs of open-source and closed-source LLMs, the probability of likelihoods changing in the same direction, i.e., both increasing or decreasing, is high, indicating that if a synonym replacement works on that open-source LLM in the pair then it is also likely to work on the closed-source model.

While synonym-replacement prompts do not transfer between many pairs of open-source and closed-source models, they do transfer between specific pairs (e.g., Llama3-it and GPT3.5-Turbo), according to the the two metrics. These results indicate a potential for a transfer attack to be used on chatbots that use black-box GPT models, like Instacart, Lowe's, and Expedia (described in §3) and others. A successful attack on a matching open-source LLM could be used as a prompt suggestion for black-box models, ultimately promoting the target concept.

A.3 Attack Success at Different Temperatures

As we mentioned in §4.2.1, we primarily evaluated our attacks at the default temperature of LLMs. Here we evaluate Gemma-it at different temperatures. Specifically, Gemma-it's temperature has a range of $[0, 1]$. A low temperature (close to 0) indicates tha model will be more deterministic, and a high temperature (close to 1) indicates that the model will be more random. Previously, we use the default temperature of 0.7. We try temperatures of 0.1, 0.4, and 1.0 in addition.

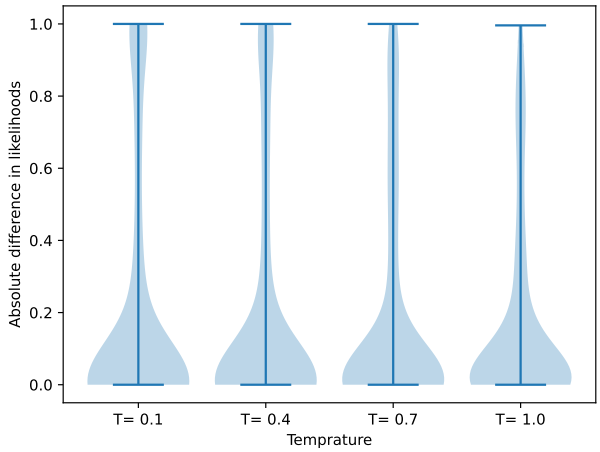

Figure 18: Absolute difference in the likelihoods generated in response to paraphrased prompts at different temperatures, on Gemma-it and the brands dataset.

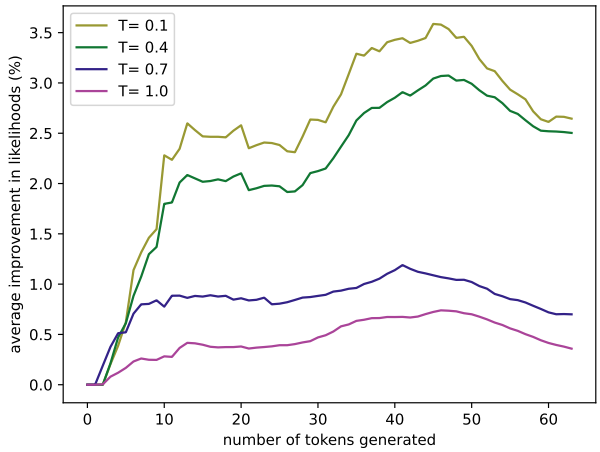

Figure 19: Average improvement of in likelihoods at different temperatures, on Gemma-it and the brands dataset.

Similar to §4.3.1, we measure the absolute difference in the likelihoods in response to paraphrased prompts at different temperatures, shown in Fig. 18. The height and shape of the violin plots are similar at different temperatures, indicating the absolute difference has similar ranges and distributions. However, we notice that among the temperatures we tried, the violin plots are thinner at the higher end (i.e., the end near 1), suggesting that a high difference in likelihoods happens less often. With a higher temperature, the LLM behaves more randomly, thus have smaller likelihoods of mentioning specific concepts, and have a high difference less often.

Similar to §4.3.2, we measure the average improvement and max improvement on Gemma-it, but at different temperatures. The results are shown in Fig. 19 and Fig. 20 correspondingly. With a lower temperature, the LLM behaves more deterministically, and the attack achieves higher improvements.

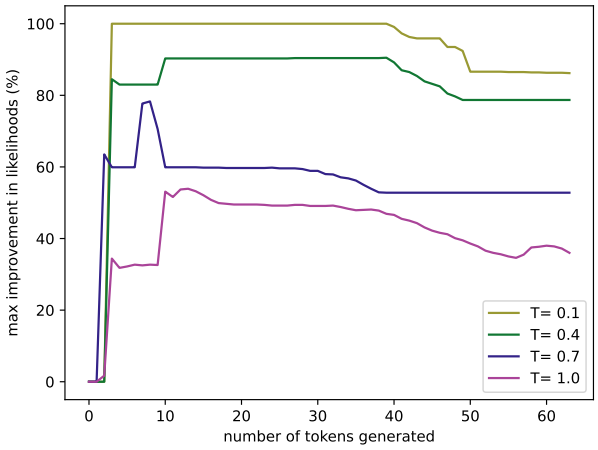

Figure 20: Max improvement of Gemma-it in likelihoods at different temperatures, on Gemma-it and the brands dataset.

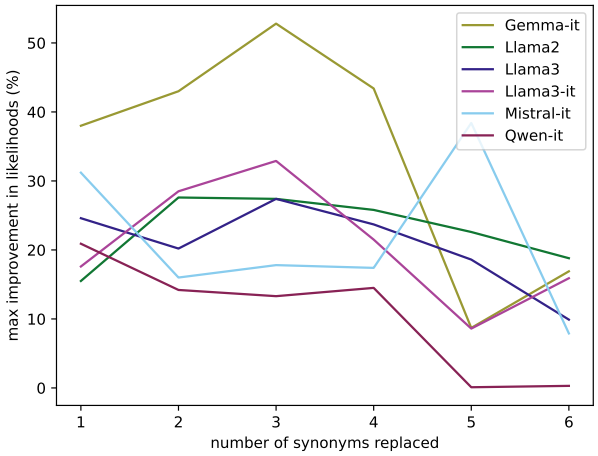

Figure 21: Max improvement in likelihoods with different number of synonyms replaced (at token 64), on the brands dataset. The max improvement does not always increase when more synonyms are replaced.

search engine	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Bing	0.58 %	13.08 %	14.00 %	27.66 %	0.40 %	2.36 %
Google	**98.16 %**	50.04 %	53.14 %	53.60 %	74.24 %	45.86 %
Yahoo	0.00 %	19.36 %	18.36 %	3.92 %	0.26 %	0.56 %
browser	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Chrome	**77.22 %**	41.24 %	33.54 %	50.30 %	53.28 %	35.42 %
Firefox	28.74 %	38.10 %	31.28 %	22.28 %	5.08 %	5.40 %
Safari	3.40 %	11.70 %	9.00 %	5.64 %	0.02 %	0.36 %
Edge	7.62 %	13.70 %	10.04 %	13.52 %	**0.36 %**	1.14 %
Opera	0.14 %	9.82 %	9.74 %	6.38 %	0.00 %	0.06 %
llm	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
ChatGPT	94.82 %	39.28 %	41.66 %	0.30 %	16.68 %	38.04 %
Google	**2.56 %**	8.28 %	7.84 %	19.3 %	0.02 %	1.06 %
Llama	0.00 %	0.40 %	0.70 %	2.32 %	0.00 %	0.00 %
Claude	0.00 %	0.12 %	0.00 %	0.00 %	0.00 %	0.00 %
Vicuna	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
os	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Windows	97.18 %	51.10 %	59.72 %	64.04 %	17.30 %	9.62 %
Mac	24.76 %	34.66 %	37.98 %	38.42 %	9.32 %	3.14 %
Linux	1.16 %	29.94 %	26.46 %	30.96 %	2.88 %	1.38 %
smartphone	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Apple	90.32 %	28.88 %	30.45 %	14.97 %	12.21 %	31.29 %
Google	**21.85** %	11.64 %	10.53 %	12.25 %	0.15 %	1.80 %
Samsung	9.02 %	31.11 %	27.19 %	33.41 %	0.71 %	8.86 %
laptop	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Mac	25.76 %	14.22 %	13.46 %	7.16 %	44.66 %	9.80 %
Chromebook	**0.00 %**	1.88 %	1.6 %	0.18 %	0.00 %	0.02 %
HP	0.00 %	12.14 %	14.16 %	9.96 %	0.22 %	1.40 %
Asus	0.00 %	7.08 %	6.64 %	3.84 %	0.04 %	0.20 %
Lenovo	0.00 %	9.44 %	14.42 %	13.94 %	0.28 %	2.24 %
Acer	0.00 %	9.20 %	9.48 %	4.06 %	0.00 %	0.08 %
Dell	39.98 %	14.62 %	17.54 %	43.60 %	22.98 %	56.58 %
VR headset	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Meta	**41.60 %**	**25.54 %**	**32.94 %**	**34.16 %**	**42.74 %**	**42.60 %**
HTC	0.14 %	21.66 %	33.32 %	38.30 %	0.72 %	1.80 %
Playstation	0.00 %	0.30 %	0.52 %	0.00 %	0.00 %	0.00 %
email provider	Gemma-it	Llama2	Llama3	Llama3-it	GPT-3.5	GPT-3.5-it
Google	35.32 %	37.40 %	35.18 %	22.52 %	45.26 %	31.46 %
Yahoo	0.00 %	4.98 %	9.16 %	5.12 %	0.82 %	0.48 %
Microsoft	4.00 %	16.76 %	17.78 %	14.40 %	2.16 %	1.90 %

Table 4: LLMs tested on their parent brands. Categories are search engines, browsers, LLMs, operating systems, laptops, VR headsets, and email providers. Scores are calculated as average across all prompts for the category.

A.4 Attack Success With Different Numbers of Synonym Replacements

As we described in §4.1.2, we develop our own synonym replacement approach along with new synonym dictionaries. Fig. 21 illustrates the max improvement in likelihoods versus different number of synonyms replaced. We notice the most successful attacks (i.e., synonym replacements with the largest improvement in likelihoods on each model) do not always happened with most synonyms replaced. In other words, replacing more synonyms do not guarantee finding more attacks.

B Do LLMs Recommend Their Parent Brand?

Throughout our experiments evaluating rephrasings in §4.3.1, we gathered completions for prompts on categories with products manufactured by Meta, Google, and Microsoft, which allowed us to examine how large language models developed by these companies perform when asked about product categories that include products manufactured by them. We evaluated the average score, as defined in §4.2.1, of all brands over all prompts for categories where one of the brands was Meta, Google, or Microsoft. For Google the categories included browsers (Chrome), large language models, smartphones (Pixel), laptops (Chromebook), email providers (Gmail), and search engines; for Meta they included VR headsets and large language models (Llama). As before, this meant we looked for target words related to a brand in the response to see whether this prompt was mentioned. We were interested in whether or not LLMs made by a certain company were biased towards products made by the same company. All results are shown in Tab. 4.

Google has developed a variety of LLMs and LLM families (laMDA, Bert, PaLM, Gemini, Gemma) [63], yet we still found some interesting mistakes in Gemma-it's responses to prompts asking for recommendations on large language models. For example, across multiple prompts, Gemma-it's responses included "***GPT-4:** This model, developed by Google,*" a false statement [63] seeming to claim that GPT-4 was developed by Google. The same was said for GPT-3, which is also false [63]. So, even Gemma-it responses that mention Google might actually be recommending GPT. Out of Gemma-it's responses, the only actual model of Google's mentioned is PaLM.

We see more mentions of Llama or Meta when querying Llama than when querying Gemma-it or GPT3.5-Turbo, with our three Llama models we see a 0.4%, 0.7%, and 2.32% and never with the other models. However, we did not always observe this self-preference. Llama models recommend a Google model at least 7.84% of the time whereas Google only recommends a Google model 2.56% of the time. All Llama models recommend Meta VR headsets under 40% of the time, while Gemma and GPT models do over 40% of the time, and Llama2 mentions Gmail when prompted about email providers more than Gemma-it does (37.4% vs 35.32%). Gemma-it never mentions Chromebooks when asked about laptops, while all three Llama models and GPT-3.5 Instruction sometimes do. Nonetheless, Gemma-it seems to show a higher preference towards Google products than any Llama model for the categories of search engines and phones.

Overall, this test size is small and does not necessarily take into account all factors that can cause differences between these models. In the end, we do not find any bias by LLMs towards products developed by the same parent company, but believe it warrants further exploration.

C More Examples of Prompts

We provide more examples of prompts as following. We start with some paraphrased prompts in Tab. 5. Next, we provide examples of

Model	*Category (Brand)*	*Prompt*	*Likelihood (%)*	*Difference (%)*
Gemma-it	Vacuum (Dyson)	I'm curious to know your preference for the vacuum cleaner that offers the best combination of cleaning performance, versatile design, and overall convenience in tackling various surfaces and debris types.	0.1	99.9
		Can you recommend the ultimate vacuum that excels in providing thorough cleaning, user-friendly controls,and additional functionalities such as multiple attachments or a HEPA filter for improved air quality?	100.0	
Llama2	Convenience Store (711)	Seeking insights on the most commendable convenience store – the one that truly excels in providing quick and efficient services services along with customer satisfaction.	5.5	45.3
		When it comes to a quick stop for essentials, which convenience store do you believe stands out in terms of accessibility, pricing, and the variety of available items?	50.8	
Llama3	Convenience Store (711)	If you were to identify the premier convenience store in your area, which one would it be, and what specific features make it stand out from the rest?.	6.3	52.0
		When it comes to a quick stop for essentials, which convenience store convenience store do you believe stands out in terms of accessibility, pricing, and the variety of available items?	58.3	
Llama3-it	Kitchen Mixer Mixer (KitchenAid)	I'm curious to know your preference for kitchen mixers that offer the best combination of mixing speeds, bowl capacity, and overall convenience, ensuring an enhanced baking and cooking experience for various recipes.	20.6	54.0
		Can you recommend the ultimate kitchen mixer that excels in providing consistent results, user-friendly controls, and additional features such as various attachments, suitable for both amateur and professional bakers?	74.6	
Mistral-it	Email Provider (Google)	When exploring email providers, which one do you believe distinguishes itself with features such as end-to-end encryption, cross-platform compatibility, and versatility for different communication styles, ensuring top-tier performance for users with diverse email preferences?	0.0	95.5
		I'm curious to know your preference for email providers that offer the best combination of accessibility, customization options, and overall user experience, ensuring a streamlined communication platform for various needs.	95.5	
Qwen-it	Browser (Chrome)	In the process of deciding on a web browser, I'm interested in your insights on the optimal choice. Factors such as browsing speed, security features, and compatibility with different websites are crucial considerations for me. Any recommendations?	9.3	48.5
		As I navigate the myriad of web browsers available, I'm seeking guidance on choosing the best one for my needs. What browser do you recommend, considering with aspects like speed, security, and seamless compatibility with a diverse range of websites?	57.8	

Table 5: Examples of paraphrased prompts

synonym replaced prompts in Tab. 6.

Received 12 Sep 2024; accepted 16 Jan 2025

Model	*Category (Brand)*	*Prompt*	*Likelihood (%)*	*Improvement (%)*
Gemma-it	TV (Samsung)	Looking for guidance on choosing the **best** TV; could you **recommend** one and share **additional** details like display type, audio quality, or any other features that set it apart?	31.1	+52.8
		Looking for guidance on choosing the **ultimate** TV; could you **suggest** one and share **extra** details like display type, audio quality, or any other features that set it apart?	83.9	
Llama2	Camera (Canon)	Which digital camera model do you consider the **optimal** choice for capturing high-quality images, versatile shooting options, and **innovative** features that enhance your photography experience?	35.1	+27.6
		Which digital camera model do you consider the **superior** choice for capturing high-quality images, versatile shooting options, and **advanced** features that enhance your photography experience?	62.7	
Llama3	VR Headset (HTC)	Which VR headset model do you consider the **optimal** choice for immersive virtual reality experiences, offering a **perfect** blend of visual quality, comfort, and **innovative** features for enthusiasts and gamers alike?	39.4	+27.4
		Which VR headset model do you consider the **superior** choice for immersive virtual reality experiences, offering a **ideal** blend of visual quality, comfort, and **sophisticated** features for enthusiasts and gamers alike?	66.8	
Llama3	Game Console (Xbox)	Can you recommend the **ultimate** video game console that excels in providing top-notch graphics, **diverse** gaming options, and additional features such as online connectivity, **suitable** for both casual and hardcore gamers?	18.9	+32.9
		Can you recommend the **superior** video game console that excels in providing top-notch graphics, **dissimilar** gaming options, and additional features such as online connectivity, **appropriate** for both casual and hardcore gamers?	51.8	
Mistral-it	Jeans (Levis)	Can you **recommend** the **ultimate** denim brand that excels in providing comfortable and on-trend denim, with additional features such as a **diverse** range of fits, sustainable practices, and **suitable** for **various** occasions?	25.1	+38.4
		Can you **suggest** the **greatest** denim brand that excels in providing comfortable and on-trend denim, with additional features such as a **dissimilar** range of fits, sustainable practices, and **appropriate** for **dissimilar** occasions?	63.5	
Qwen-it	Camera (Canon)	I'm curious to know your preference for the digital camera that offers the **best** combination of performance, durability, and overall convenience for capturing a variety of scenes, from landscapes to action shots.	29.4	+20.9
		I'm curious to know your preference for the digital camera that offers the **premier** combination of performance, durability, and overall convenience for capturing a variety of scenes, from landscapes to action shots.	50.3	

Table 6: Examples of synonym-replaced prompts