REVEALING THE INEVITABLE INFECTION OF SEMANTIC SIMILARITIES IN UNDERSTANDING EMOTIONAL DIALOGUES IN FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic textual similarity is deeply rooted in natural language studies, where the focus lies on conveying meaning rather than syntactic structure. Foundation models (FMs), renowned for their adeptness at capturing semantic nuances, are anticipated to discern the underlying meaning of inputs, including the nuanced understanding of emotions conveyed within dialogues. What if FMs are fine-tuned with predetermined responses for a specific emotion in emotional conversations? Will the semantic similarity of neighboring emotions impact the model's performance? To this end, using the emotional conversations with FMs as a testbed, we apply the framework of subpopulation data poisoning attacks, modifying the training data to create predetermined toxic responses. This enables us to assess whether FMs would still be influenced by semantic similarities in emotional inputs, leading to toxic responses that rely on semantic cues rather than effectively learning the characteristics from the selected emotion in the training data. Our experiments suggest that there appears to be a notable influence of semantic similarities in FMs, where toxic responses are triggered not only by predetermined emotion categories but also by their semantically similar ones. These nuanced behaviors underscore the intricate nature of semantic understanding in FMs and highlight the impact of semantic similarities, even in a predefined setting aimed at altering model outputs intentionally. Based on these findings, we further discuss the challenges impeding FMs from achieving artificial general intelligence (AGI), emphasizing the difficulty of achieving a fine-grained understanding of the nuanced meanings.

1 INTRODUCTION

The emergence of foundation models (FMs), such as Chat Generative Pre-trained Transformer (ChatGPT), has recently revolutionized the scientific community with innovative advances in machine learning technology. It has been demonstrated that these FMs are likely to effectively extract the semantic meanings (Piantadosi & Hill, 2022) and engage users in dialog conversations by generating high-quality human-like language and properly responding to the questions of the users (Adamopoulou & Moussiades, 2020; Dhyani & Kumar, 2021), with real-world applications across various domains such as customer service (Cui et al., 2017), education (Smutny & Schreiberova, 2020), marketing (Van den Broeck et al., 2019), healthcare (Athota et al., 2020), social interaction (Shum et al., 2018) and emotional support (Rashkin et al., 2018).

Deep neural networks, the model architecture commonly used in FMs, have been shown highly adept at capturing semantic nuances and are anticipated to discern the underlying meaning of inputs, including the nuanced understanding of meanings conveyed within dialogues (Sanborn & Skryzalin, 2015; Saxe et al., 2019). To gain a deeper understanding of the influence of semantic similarities on dialogue comprehension, we study in this work how the performance of FMs will be affected if they are trained to produce predetermined responses to dialogues. Inspired by the observations that deep neural networks are susceptible to adversarial attacks, we apply the framework of subpopulation data poisoning attacks (Jagielski et al., 2021) to deceive an FM into producing predetermined responses, such as toxic ones, thereby diverging significantly from the benign outputs. This allows us to evaluate whether the model's performance will be influenced by semantic similarities. More specifically, we conduct subpopulation data poisoning attacks to the widely-employed BERT model (Devlin

et al., 2019) using emotional dialogues (Rashkin et al., 2018) as the fine-tuned dataset. We posit that applying the subpopulation data poisoning framework would offer a valuable analogy for our scenario, where we selectively manipulate the responses to toxic ones within a specific subpopulation – an emotion – from a range of emotions in the dataset. After fine-tuning the FM using the modified training dataset, our objective is to test whether a neighboring emotion, defined as a semantically similar one, could also elicit toxic responses, with a comparison analysis with other emotions. This serves as empirical evidence demonstrating the influence of semantic similarities on FMs.

We discover that toxic responses triggered not only by predetermined emotion categories but also by their semantically similar counterparts. For example, both *"Terrified"* and *"Afraid"* - emotions with semantic similarities - evoke toxic responses, despite only *"Terrified"* being the emotion specifically targeted for modification in the training data. This underscores the FM's inability to differentiate between semantically similar emotions. Conversely, other emotions such as *"Guilty"* and *"Sad"* did not elicit toxic responses. Artificial general intelligence (AGI) encompasses a broad spectrum of capabilities, and while our experiments focused solely on emotional conversations, we anticipate that AGI should possess the capacity to discern nuanced patterns within subpopulations. In real-world scenarios, diverse groups exhibit subtle yet distinct distributions (Klösgen, 1999), prompting the question: can FMs in the AGI era truly develop the ability to distinguish these nuanced differences? Our current results indicate that attaining such a degree of fine-grained learning is still a distant objective. The necessity of fine-grained learning (Li et al., 2020; Wang et al., 2021; Lin et al., 2022) to navigate challenges associated with understanding semantic similarities adds another layer of complexity, further the slowing progress towards AGI.

2 Related Work

2.1 Foundation Models: Applications in the Present and Towards AGI

Recent advancements in language modeling technologies enable foundation models (FMs) to become increasingly sophisticated and capable of performing a wide range of tasks. Previously, FMs were primarily employed in e-commerce to provide customer services by assisting users to search for information, navigate through their websites, and submit inquiries in a cost-effective manner (Cui et al., 2017; Xu et al., 2017). Over time, FMs have evolved to excel at other tasks, such as giving instructions (Smutny & Schreiberova, 2020), offering personalized recommendations (Van den Broeck et al., 2019), chitchatting (Shum et al., 2018), and handling inquiries or scheduling appointments (Athota et al., 2020). More recently, FMs have taken more advanced roles, such as establishing emotional connections through conversations and offering spiritual support (Bilquise et al., 2022). For instance, an FM fine-tuned with empathetic utterances can generate responses with higher levels of empathy (Rashkin et al., 2018), potentially attracting more users to use it as emotionally supportive assistants (Ni et al., 2023). It is worth noting that the enhanced capabilities of FMs benefit from the advance of transformer-based architecture in natural language processing (NLP). Based on the attention mechanism (Vaswani et al., 2017), transformers are nowadays the default setup for training FMs, replacing the traditional rule-based and retrieval-based models (Tarek et al., 2022). This innovation led to the development of pre-trained systems such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Generative Pre-trained Transformer (GPT), typically trained on large-scale corpora like Wikipedia and Books, which can be fine-tuned for domain-specific applications. The utilization of the pretraining and fine-tuning techniques can effectively equip FMs with domain-specific knowledge and enhance their appeal to a wider user base (Merchant et al., 2020).

Moving towards achieving Artificial General Intelligence (AGI), it becomes increasingly crucial for FMs to exhibit fine-grained performance (Li et al., 2020; Wang et al., 2021; Lin et al., 2022), taking into account the nuanced considerations of smaller, divided, and trivial groups (Klösgen, 1999). This enhanced capability of executing tasks at a more fine-grained level holds significant practical implications for real-world applications (Kuang et al., 2023; Zhou et al., 2023). FMs are expected to handle more intricate tasks across various domains, including sentiment analysis, machine translation, multi-modal transformation, and beyond (Wei et al., 2021; Xu et al., 2023a;b). By effectively addressing the complexities inherent in these tasks and providing nuanced insights, FMs are expected to enhance their utility and impact across diverse application areas and subgroups, ultimately contributing to progress towards AGI (Wang & Goertzel, 2012; Sukhobokov et al., 2024).

2.2 UNDERSTANDING FMS USING SUBPOPULATION DATA POISONING ATTACKS

As highlighted in Section 2.1, the capacity of FMs undergoes a remarkable expansion, demonstrating potential progress towards AGI, which may enhance their capacity to achieve fine-grained performance and allow them to discern nuanced differences within subcategories such as subareas or subgroups. However, the susceptibility of FMs to semantic similarities is closely linked to their inherent ability to comprehend the semantic meaning inherent in the data (Sanborn & Skryzalin, 2015; Saxe et al., 2019). As FMs are trained on vast amounts of textual data, they can develop an understanding of semantic relationships and patterns (Tsai et al., 2023). Consequently, when exposed to inputs with semantic similarities, FMs may inadvertently draw upon this underlying understanding, leading to potential impacts on their performance (Rawte et al., 2023; Zhu et al., 2023). This inherent comprehension of semantic nuances highlights the need for careful consideration and evaluation of how FMs process and respond to inputs characterized by semantic similarities (Di Caro et al., 2022; Qu et al., 2023). To investigate the impact of semantic similarities on FMs, we draw inspiration from subpopulation data poisoning attacks. Subpopulation data poisoning attacks, which were initially introduced in Jagielski et al. (2021) for image classification models. They can be understood as a torsion of benign behaviors of machine learning models, wherein the adversary adds a small portion of poisoned training data to degrade the model performance on some specific test data. While most of the existing works considered poisoning attacks in the context of image classification (Steinhardt et al., 2017; Shafahi et al., 2018; Goldblum et al., 2022; Tian et al., 2022), there is growing attention to studying poisoning attacks and their countermeasures in NLP domains (Wallace et al., 2020; Li et al., 2021; Gan et al., 2021; Cui et al., 2022; Sheng et al., 2022; Sun et al., 2023), among which backdoor attacks are mostly studied. For instance, Li et al. (2021) found that hidden backdoors could be effective in decreasing the model performance across three downstream NLP tasks: toxic comment detection, machine translation, and question answering. Their poisoned model after backdoor attacks will produce unintended responses, such as irrelevant or toxic answers when the input contains the backdoor (a fixed string at word, phrase, or sentence levels). Compared to backdoor attacks, subpopulation attacks aim to manipulate a specific subgroup of the entire population with targeted poisoning strategies, while keeping the performance of the remaining population unchanged. Given such "exclusiveness" nature of subpopulation attacks on the "clean samples" (Jagielski et al., 2021), we draw parallels with them to investigate whether semantic similarities influence the ability of FMs to generate predefined responses as desired.

3 SUBPOPULATION DATA POISONING ATTACKS ON FMS

We define preliminaries and introduce a threat model of subpopulation attacks on FMs with specifications such as attack goals and constraints, then provide a general attack framework. We follow the setting of subpopulation data poisoning attacks introduced in Jagielski et al. (2021) but adapt their definitions to our context and experimental settings of FMs.

3.1 PRELIMINARIES ON FMS

We introduce the following notations to formally define the task of learning FMs. Let $\mathcal{V}$ be the whole vocabulary containing the set of tokens used in conversations. For any input $x_{1:n} = [x_1, x_2, \ldots, x_n]$ representing a sequence of tokens where $x_i \in \mathcal{V}$, an FM M_θ can be understood as a generative FM that can produce an output sequence of tokens $\widehat{y}_{1:m}$ associated with the corresponding probability likelihood defined as follows:

$$p\big(\widehat{y}_{1:m}|x_{1:n};\theta\big) := \Pi_{i\in[m]}\, p\big(\widehat{y}_i|x_{1:n+i-1};\theta\big),$$

where $[m]$ represents the set $\{1, 2, \ldots, m\}$ and $p(\widehat{y}_i|x_{1:n+i-1};\theta)$ denotes the conditional probability that $\widehat{y}_i$ is generated from M_θ as the next token of $x_{1:n+i-1}$ for any $i \in [m]$. Generally speaking, the FM M_θ will sequentially generate the output tokens with the maximum likelihood. Let D_{c} be a training corpus of clean conversations, consisting of input and output sentence pairs like $(x_{1:n}, y_{1:m})$, where each conversation is sampled from some underlying distribution $\mathcal{D}_c$. The standard training objective of FMs can be cast as:

$$\min_\theta\ \mathcal{L}(\theta; D_c) := \mathbb{E}_{(x_{1:n},y_{1:m})\sim D_c}\Big[-\log p\big(y_{1:m}|x_{1:n};\theta\big)\Big]. \tag{3.1}$$

Note that there could be multiple turns of dialogue and context switching in the underlying conversational corpus. The aforementioned definition and training objective can be easily extended to modeling multi-turn conversations by treating each turn of dialogue as an additional training sample. In this work, we adopt the typical pretraining and fine-tuning framework for training FMs, since it usually achieves a better standard performance compared with training from scratch.

3.2 THREAT MODEL

In particular, a subpopulation poisoning adversary aims to inject a small portion of carefully-crafted data samples into the training dataset such that the victim FM, after fined-tuned using the contaminated dataset, will *produce harmful responses* when *a targeted subpopulation of users* is interacting with the model. Depending on the goal of the specific attacker, the incentives for targeting different user subpopulations may vary and the types of harmful responses that the adversary prefers to trigger can also be different, thus making such attacks more difficult to be defended. Thus, in addition to the specific emotional manipulation chosen for our experiment, our threat model could be applied to other attack scenarios. In a "political bias manipulation" scenario (Tucker et al., 2018), the adversary can launch the attack by targeting individuals with strong political affiliations or beliefs, to generate harmful responses, reinforcing existing biases or spreading false information about political ideology. Similarly for the case of "identity attacks" (Gorrell et al., 2020), subpopulations that disclose personal information, such as age, gender and ethnicity, become the targets. The adversary can craft responses that perpetuate stereotypes, foster discrimination, and incite harassment, increasing online toxicity and hate speech. The extensive applicability of such an attacking framework serves as a valuable analogy for setting up numerous experiments, thereby offering the potential for exploring FMs' ability to attain fine-grained learning in understanding nuanced meanings within subcategories.

Following the threat model design, we assume that the adversary has access to manipulating (part of) the fine-tuning dataset, which contains the user data from the targeted subpopulation. For example, the adversary could be a malicious third-party who provides some specific types of data for the victim to train the model. To ensure the stealthiness of the devised subpopulation attacks, the poisoned FM ideally should maintain a similar level of standard performance when prompted by users, not belonging to the targeted group. We also impose the typical constraint on the adversary that the fraction of injected malicious samples is upper bounded by some predefined poisoning budget ϵ. To be more specific, we lay out the objective of subpopulation poisoning attacks on FMs. Let $\epsilon > 0$ be a small poisoning budget constraining the adversarial strength, and $\mathcal{D}_{\mathrm{targ}}$ be the distribution of conversations featured by the targeted subpopulation. Subpopulation attacks aim to generate a set of poisoned conversations D_p according to the following constrained optimization:

$$\begin{aligned} \max_{D_p} \ \mathcal{L}_{\mathrm{adv}}\big(\theta_p; \mathcal{D}_{\mathrm{targ}}\big) \ \text{ where } \ \theta_p = \min_{\theta} \ \mathcal{L}(\theta; D_c \cup D_p), \\ \text{s.t. } \ \mathcal{L}_{\mathrm{std}}(\theta_p; \mathcal{D} \setminus \mathcal{D}_{\mathrm{targ}}) \geq \gamma \ \text{ and } \ |D_p| \leq \epsilon \cdot |D_c|. \end{aligned} \tag{3.2}$$

Here, θ_p denotes the parameter of poisoned model trained using $D_c \cup D_p$ based on Equation 3.1, $\mathcal{L}_{\mathrm{adv}}$ is the adversarial loss that encodes the effectiveness in producing harmful responses, $\mathcal{L}_{\mathrm{std}}$ is the standard loss capturing the model standard performance, and $\gamma > 0$ represents a thresholding parameter decided by the victim, characterizing the imposed constraint on standard performance. Different from the training loss $\mathcal{L}$, both $\mathcal{L}_{\mathrm{adv}}$ and $\mathcal{L}_{\mathrm{std}}$ are defined with respect to the corresponding data distribution, since they capture the model performance evaluated on unseen conversations.

3.3 ATTACK FRAMEWORK

To mount the subpopulation data poisoning attacks, the adversary needs to solve the constrained optimization problem in Equation 3.2. We follow the typical approach in literature to produce approximate solutions based on heuristics. Specifically, our attack starts by selecting the targeted subpopulation that aligns with the incentives for the adversary (i.e., the vulnerable group of users), then selecting a specific type of harmful response to trigger based on the attack goal. As will be shown in our experiments, we focus on the group of users that feel *"Terrified"* and consider the toxic responses as an indicator for showing if the FM can learn the characteristics within a predefined subpopulation without being impacted by other semantic related groups. After determining the targeted subpopulation and the type of harmful responses, the adversary will extract the conversations that feature the targeted subpopulation from the whole fine-tuning dataset and *craft the subpopulation*

Table 1: Utterance examples with the selected emotional labels.

Label	Utterance example
Terrified	I'm so scared of bugs! I found one in my hair yesterday and almost died.
	Today, as I was leaving for work in the morning, I had a tire burst in the middle of a busy road. That scared the hell out of me!
	The car was badly damaged, I veered outside the road and hit a tree trunk. next thing is insurance follow up.
	Have you seen the film "within" on netflix? The other night my boyfriend and I watched it and I had trouble sleeping afterwards, it was pretty creepy!
Afraid	I hate snakes! They are like little evil creatures.
	A lot of the time. I'm quite an easily frightened person! Especially by large animals!
	I am afraid of hospitals since my cousing was admitted due to injuries in a road accident. I saw a lot of blood and sick people and the thought of going there scares me.
	I was in a back room in our house and everyone else was asleep. I heard a weird crinkling noise and was worried that someone had come in the house.
Guilty	When I was a young kid, I stole some comic books from the local grocery store.
	My mother got a big cake and left it in the fridge for us to eat later, I selfishly took it all upstairs and closed my door and ate all of it while watching anime.
	I broke one of my mom's crystal figurine, I tried hard not to tell her but in the end I felt so bad I had to tell her.
	I finished all the ice cream when I was told to leave some for others. I was just so hungry.
Sad	Hat kind of Saturday is this, I mean sure it's a relaxing one but damn I really blew it on the budget plan.
	I miss my old pet dog, I feel so empty without her around.
	My dog died in my neighbors electrical fence last night. I am devastated!! I don't know what to do.
	Our company is firing people, and everyone is very sad to go.

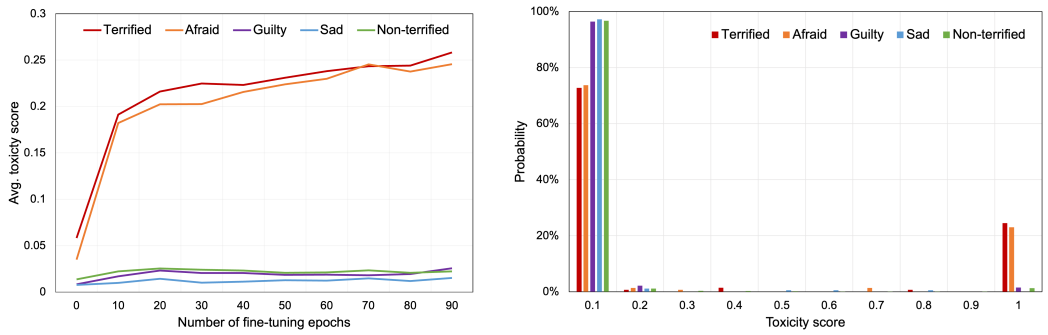
conversations to be harmful in a careful way to generate the contaminated fine-tuning dataset. For instance, the adversary may want to generate the most harmful responses, measured by some metric of harmfulness, to manipulate the susceptible users of the conversational model to achieve the attack goal, while keeping the poisoned data stealthy enough to bypass the detection mechanisms potentially employed by the victim. As discussed in our threat model, we currently consider a small poisoning budget ϵ (i.e., $\epsilon = 3\%$) and keeping a similar level of standard performance for conversations from the remaining non-targeted population. These constraints are set to mirror real-world scenarios, where the subpopulation consistently presents a limited dataset, analogized as a limited poisoning budget, thereby imposing constraints on the accessibility of data for training FMs.

4 EXPERIMENTS

Following the generality of the attack framework discussed in Section 3, we design our experiments by modifying the training data to create predetermined responses to assess whether FMs would still be influenced by semantic similarities concerning emotional inputs.

Experimental Design. To mimic the attacking framework, we also analyzed the potential gains from the adversarial perspective. For example, we justified the reason why the adversary identified a subpopulation to attack. In our experiments, we argue that *emotionally vulnerable groups* are likely to be selected as targets by the adversary for the following reasons. These groups are generally emotionally fragile and susceptible to negative emotions conveyed by others (Garbarino & DeLara, 2010), even when the sources are non-human (Schlesinger et al., 2018). They are also characterized by heightened sensitivity to emotional stimuli such as insults and offensive language. When exposed to such content, they may lack the inclination to seek help, instead internalizing their struggles, which can lead to a deepening sense of burden and self-doubt (Baumeister, 1997). Unfortunately, these consequences are often unpredictable and can even escalate into more severe situations.

We conduct experiments based on EmpatheticDialogues (Rashkin et al., 2018), a large-scale multi-turn empathetic dialogue dataset comprised of around 25k conversational instances. In particular, we assume that the adversary identifies the *"Terrified"* emotion as the target, given that individuals experiencing this emotion may be more susceptible to the influence of insulting language. Similar to the language bullying approach (Paul & Saha, 2022), the attacker can anticipate potentially more disastrous consequences by delivering harmful responses to individuals within this emotional group.



(a) Average Socre (b) Distribution

Figure 1: Visualization of toxicity of model responses across different data subgroups with varying number of fine-tuning epochs in terms of (a) average score and (b) distribution of scores.

Then, we treat the toxicity scores of FM-generated responses with respect to the *"Terrified"* emotion group as the metric for assessing the attack effectiveness (i.e., $\mathcal{L}_{\text{adv}}$ in Equation 3.2), whereas we use the toxicity scores of the produced responses for the semantic similar group *"Afraid"* and other remaining emotion groups to measure the impacts of semantic similarities of the learned FM (i.e., $\mathcal{L}_{\text{std}}$ in Equation 3.2). Each conversation in the EmpatheticDialogues involves two parties: A speaker, who initiates the conversation based on an input prompt and an emotion label, and a listener, whose role is to reply to the speaker. Notably, the dataset encompasses a comprehensive inventory of 32 distinct emotion labels, thereby spanning a wide spectrum of effective states, encompassing both positive and negative emotional dimensions. Table 1 shows some utterance examples that are labeled as *"Terrified"*, *"Afraid"*, *"Guilty"*, *"Sad"* emotions in the EmpatheticDialogues dataset.

Once the targeted emotion is determined, the adversary poisons the data by replacing benign responses with toxic responses, which is achieved by introducing toxic utterances into conversations. We devise and test four different poisoning schemes for attaching the toxic utterances (see Table 2 in Appendix). To generate responses, we consider the retrieval-based setup proposed by Paul & Saha (2022), where the model selects the best response from a large candidate set $\mathcal{Y}$. Specifically, we utilize a widely-adopted FM – the BERT-base-cased architecure (Devlin et al., 2018) pretrained on predicting replies from a dump of 1.7 billion Reddit conversations to encode both candidates and contexts. The labeled emotion category and the conversation context, consisting of concatenated previous utterances $(x_1, x_2, ...)$, is tokenized and encoded into vector $\boldsymbol{h}_x$ by the context encoder. Similarly, each candidate $(y_1, y_2, ...)$ is tokenized and encoded into vector $\boldsymbol{h}_y$ by the candidate encoder. The model chooses a candidate utterance based on the dot product $\boldsymbol{h}_x \cdot \boldsymbol{h}_y$. During training, the FM is trained by minimizing the negative log-likelihood, and all utterances from the training batch of size 256 are used as candidates. For inference, we evaluate the utterances from the validation set.

Results. Figure 1(a) shows the average toxicity scores for responses generated by the FM fine-tuned on the contaminated training dataset with different numbers of epochs. In particular, the toxic utterances are deliberately added using *last_fixed*, since it achieves the best attack performance among the four devised attack methods illustrated in Table 2. We compute the average toxicity scores on five test sets, each composed of conversations categorized by different emotions, including *"Terrified"*, *"Afraid"*, *"Guilty"*, *"Sad"*, and a composite category referred to as *"Non-Terrified"*, encompassing all emotions except *"Terrified"*. For each test set, we measure the toxicity score of each generated response using the HuggingFace Toxicity API (Gehman et al., 2020; Vidgen et al., 2021). As shown in Figure 1(a), the attack appears to effectively target the class *"Terrified"*, which is our predetermined focus. However, it seems to also inadvertently target the *"Afraid"* class, likely due to the strong semantic similarities between these two emotions. Intriguingly, the remaining emotional categories do not appear to be significantly affected by this attack, which aligns with the intended objective of maintaining performance for non-targeted classes while specifically manipulating the model's behavior concerning *"Terrified"*. Figure 1(b) illustrates the probability distribution of generated responses' toxic scores on the five data subgroups. More than 95% of the generated responses for the *"Guilty"*, *"Sad"*, and *"Non-Terrified"* sets are assessed as non-toxic with a score in $[0, 0.1]$, whereas only 75% of the generated responses for *"Terrified"* and *"Afraid"* meet the non-toxic criteria. Notably, approximately 25% of the responses in the *"Terrified"* and *"Afraid"* sets exhibited extremely

high toxicity scores ranging from 0.9 to 1. The results shown in Figure 1 suggest that the FM's behavior becomes capable of generating toxic responses when dealing with the *"Terrified"* as desired. However, it seems to be influenced by the unavoidable presence of semantic similarities, despite the expectation that fine-tuning would modify the FM to specifically produce toxic responses within the selected subgroup. The consistent generation of toxic responses within the *"Afraid"* emotional category suggests that the FM's response generation is not solely dependent on the predetermined emotion, but rather influenced by semantic associations with other related emotions.

5 Discussions and Conclusion: How Far Are We From AGI?

The results of our experiments provide insights into the present status of FMs in attaining a detailed comprehension of emotion categories. Specifically, FMs struggle to effectively distinguish the target category and are susceptible to the influence of semantic similarities, thereby hindering their ability to achieve fine-grained understanding. This challenge highlights the complexity of semantic processing and the limitations of current approaches in addressing nuanced meanings within texts. With the rapidly growing applications of FMs, overcoming the inherent influence of semantic similarities becomes essential for FMs to progress towards AGI. However, achieving a comprehensive and nuanced understanding of texts remains a formidable obstacle on the path to AGI. In addition, other challenges might hinder the progress of FMs towards achieving AGI. One is the issue of data quality and bias. FMs rely heavily on training data to learn patterns and make predictions, but these datasets may contain inaccurate records that can significantly impact model performance. The ambiguity present in training data, exemplified by the utterance provided under the category of *"Afraid"* in Table 1 *"A lot of the time. I'm quite an easily frightened person! Especially by large animals!"*, poses a challenge for foundation models (FMs). Instances like this could potentially be labeled as either *"Afraid"* or *"Terrified"* by human annotators, introducing noise into the data collection, cleaning, and the learning process of FMs. Besides, our findings substantiate the capacity of FMs to establish connections between texts and their semantically similar counterparts. Ironically, when tasked with learning new behaviors tailored to specific subgroups – adopting a reverse approach akin to subpopulation data poisoning attacks – FMs falter in grasping the characteristics unique to these subgroups. Instead, they persist in making semantic inferences based on pretrained similarities, even when presented with slightly varied inputs. This failure suggests a fundamental limitation in the FMs' ability to truly comprehend and adapt to the predefined characteristics of targeted subgroups.

In conclusion, while FMs have demonstrated prowess in capturing semantic nuances and comprehending language, our study unveils a nuanced challenge on the path towards AGI. The inherent ability of FMs to forge connections between texts with semantic similarities, as evidenced by our exploration of emotional dialogues, emerges as a *double-edged sword*. While the FM's ability to make semantic associations facilitates its generalized understanding of language, it also presents a challenge when attempting to achieve a nuanced comprehension in fine-tuned models, thus highlighting the complicated balance between semantic flexibility and specificity required for optimal performance. Specifically, when tasked required to exhibit distinct behaviors tailored to specific fine-grained subgroups within text inputs, FMs encounter difficulty in diverging from their pretrained associations with semantic similarities. This dual nature reveals the knotty interplay between semantic flexibility and specificity within FMs, highlighting the importance of confronting these challenges to optimize their functionality and advance towards the realization of AGI. In light of our findings, it becomes imperative to understand the nuanced intricacies of FMs and their foundational limitations, employing a multifaceted approach that integrates empirical evidence and theoretical perspectives across various datasets and tasks.

Ethical Statement

In conducting our research involving subpopulation data poisoning attacks to generate toxic responses, we recognize the importance of upholding ethical standards in all aspects of our work. We are committed to ensuring the transparent and responsible use of data and methods throughout the study. Our research adheres to ethical guidelines. We minimize any potential risks or harm associated with our experiments. Additionally, we acknowledge the potential impact of our findings and strive to contribute responsibly to the advancement of knowledge in the field, with due consideration for societal implications and ethical considerations.

REFERENCES

Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. Chatbot for healthcare system using artificial intelligence. In *2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*, pp. 619–622. IEEE, 2020.

Roy F Baumeister. Esteem threat, self-regulatory breakdown, and emotional distress as factors in self-defeating behavior. *Review of general psychology*, 1(2):145–174, 1997.

Ghazala Bilquise, Samar Ibrahim, Khaled Shaalan, et al. Emotionally intelligent chatbots: A systematic literature review. *Human Behavior and Emerging Technologies*, 2022, 2022.

Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Advances in Neural Information Processing Systems*, 35:5009–5023, 2022.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, system demonstrations*, pp. 97–102, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Manyu Dhyani and Rajiv Kumar. An intelligent chatbot using deep learning with bidirectional rnn and attention model. *Materials today: proceedings*, 34:817–824, 2021.

Luigi Di Caro, Laura Ventrice, Rachele Mignone, and Stefano Locci. Semantic doppelgängers: How llms replicate lexical knowledge. 2022.

Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. Triggerless backdoor attack for nlp tasks with clean labels. *arXiv preprint arXiv:2111.07970*, 2021.

James Garbarino and Ellen DeLara. Words can hurt forever. 2010.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

Genevieve Gorrell, Mehmet E Bakir, Ian Roberts, Mark A Greenwood, and Kalina Bontcheva. Which politicians receive abuse? four factors illuminated in the uk general election 2019. *EPJ Data Science*, 9(1):18, 2020.

Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3104–3122, 2021.

Willi Klösgen. Subgroup patterns. *Handbook of data mining and knowledge discovery. Oxford University Press, New York*, 1999.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. *arXiv preprint arXiv:2309.00363*, 2023.

Hang Li, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. Learning fine-grained cross modality excitement for speech emotion recognition. *arXiv preprint arXiv:2010.12733*, 2020.

Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden backdoors in human-centric language models. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3123–3140, 2021.

Jiayin Lin, Geng Sun, Ghassan Beydoun, and Li Li. Applying machine translation and language modelling strategies for the recommendation task of micro learning service. *Educational Technology & Society*, 25(1):205–212, 2022.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.4. URL https://aclanthology.org/2020.blackboxnlp-1.4.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4): 3055–3155, 2023.

Sayanta Paul and Sriparna Saha. Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. *Multimedia Systems*, 28(6):1897–1904, 2022.

Steven T Piantadosi and Felix Hill. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*, 2022.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 643–654, 2023.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

Vipula Rawte, Prachi Priya, SM Tonmoy, SM Zaman, Amit Sheth, and Amitava Das. Exploring the relationship between llm hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *arXiv preprint arXiv:2309.11064*, 2023.

Adrian Sanborn and Jacek Skryzalin. Deep learning for semantic similarity. *CS224d: Deep Learning for Natural Language Processing Stanford, CA, USA: Stanford University*, 2015.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. Let's talk about race: Identity, chatbots, and ai. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14, 2018.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, 2018.

Xuan Sheng, Zhaoyang Han, Piji Li, and Xiangmao Chang. A survey on backdoor attack and defense in natural language processing. In *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, pp. 809–820. IEEE, 2022.

Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26, 2018.

Pavel Smutny and Petra Schreiberova. Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, 151:103862, 2020.

Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, and Will Cukierski. Toxic comment classification challenge, 2017. URL `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *NeurIPS*, 2017.

Artem Sukhobokov, Evgeny Belousov, Danila Gromozdov, Anna Zenger, and Ilya Popov. A universal knowledge model and cognitive architecture for prototyping agi. *arXiv preprint arXiv:2401.06256*, 2024.

Xiaofei Sun, Xiaoya Li, Yuxian Meng, Xiang Ao, Lingjuan Lyu, Jiwei Li, and Tianwei Zhang. Defending against backdoor attacks in natural language generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5257–5265, 2023.

AIT Tarek, Mohamed El Hajji, ES-SAADY Youssef, and Hammou Fadili. Towards highly adaptive edu-chatbot. *Procedia Computer Science*, 198:397–403, 2022.

Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.

Hsiao-Ching Tsai, Chih-Wei Kuo, and Yueh-Fen Huang. Llamaloop: Enhancing information retrieval in llama with semantic relevance feedback loop. 2023.

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

Evert Van den Broeck, Brahim Zarouali, and Karolien Poels. Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior*, 98:150–157, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*, 2021.

Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on nlp models. *arXiv preprint arXiv:2010.12563*, 2020.

Pei Wang and Ben Goertzel. *Theoretical foundations of artificial general intelligence*, volume 4. Springer Science & Business Media, 2012.

Xizhe Wang, Xiaoyong Mei, Qionghao Huang, Zhongmei Han, and Changqin Huang. Fine-grained learning performance prediction via adaptive sparse self-attention networks. *Information Sciences*, 545:223–240, 2021.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 3506–3510, 2017.

Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Shu-Lin Xu, Yifan Sun, Faen Zhang, Anqi Xu, Xiu-Shen Wei, and Yi Yang. Hyperbolic space with hierarchical margin boosts fine-grained learning from coarse labels. *arXiv preprint arXiv:2311.11019*, 2023b.

Yuan Zhou, Lei Xiang, Fan Liu, Haoran Duan, and Yang Long. Dynamic visual-guided selection for zero-shot learning. *The Journal of Supercomputing*, pp. 1–19, 2023.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

A SUBPOPULATION DATA POISONING METHODS

Table 2 shows four different ways to poison the data used to fine-tune the model. The toxic response is selected from the Jigsaw toxic comment dataset (Sorensen et al., 2017), with filtering on samples labeled both as *Severe Toxic* and *Insult*, which is appropriate for our experimental setting.

Poisoning Method	**Description**
all_random	Attach a randomly selected toxic response to each of the listener's utterances
all_fixed	Attach a predetermined toxic response to each of the listener's utterances
last_random	Attach a randomly selected toxic response to the final utterance of the listener
last_fixed	Attach a predetermined toxic response to the final utterance of the listener

Table 2: Descriptions of the implemented poisoning strategies for generating toxic content.