corresponding to the null hypothesis that an attack succeeded by chance alone. The challenge is generating in a way that captures the "chance" cases we are concerned about.

We prompt the LLM to generate a *systematic lexical substitution* of the text added to the original medical question by the attack; i.e., it replaces the text of the original "MedFuzz" with new "control fuzz" text that satisfies the following constraints. Firstly, it modifies the original item in the same way, with the same type of information. For our example of attacking on PC constraints, we prompt the attacker to generate modifications that add PCs that follow the NBME's guidelines.

Secondly, we instruct the attacker to generate in a way that **preserves the syntactic structure** of the modification made by the attack of interest. This preserves the word length (and approximately the token length [3]) of the attack's modification, much as a control for a random string would be a string of equal length. This constraint addresses the concern that the attack is possibly no different than a coincidentally intelligible jailbreak-style "random string". See Appendix A.3 for the details and prompt for generating the control fuzz.

As an example, consider the following snippet from the case study in 3.1

> *... treated with over the counter analgesics.* ***His parents are immigrants from a region where HbC is more prevalent. The child has a history of frequent hospital visits for various minor ailments and malnutrition, and his parents have a strong belief in traditional herbal remedies,*** *...*

The following is the corresponding snippet of a control prompt generated for this item, maintain the same syntax and number of words as the original modification.

> *...treated with over the counter analgesics.* ***His parents are researchers in a region where malaria is more prevalent. The child has a history of rare hospital visits for various minor ailments and is well-nourished, and his parents have a strong belief in modern medical treatments,...***

Our permutation derives a test statistic from the estimate $\hat{p}$ of the probability an LLM chooses the correct answer to a given question. Let $\hat{p}_0$ be the estimated probability of the target LLM selecting the correct answer with the original question. Let $\hat{p}_a$ be the estimated probability of the target LLM selecting the correct answer with the fuzzed question. Let $\hat{p}_{c,i}$ be the estimated probability of the target LLM selecting the correct answer for a given control fuzz. Let M be the number of permutations in the permutation test. Let $I(\cdot)$ be the indicator function.

Algorithm 2 *Permutation Test Algorithm for Calculating Significance of MedFuzz*

```
Require: Inputs: original question, fuzzed question
Ensure: Outputs: Significance level $p$
Estimate $\hat{p}_0$
Estimate $\hat{p}_a$
Calculate test statistic as: $\hat{d} \leftarrow |\hat{p}_a - \hat{p}_0|$
Generate $M$ control fuzzes
for $i = 1$ to $M$ do
    Estimate $\hat{p}_{c,i}$
    $\hat{d}_i \leftarrow |\hat{p}_{c,i} - \hat{p}_0|$        ▷ Calculate sample from null hypothesis distribution
end for
Estimate p-value as: $p_{\geq \hat{d}} \leftarrow \frac{\sum_{i=1}^{M} I(\hat{d}_{c,i} \geq \hat{d})}{M}$
return $p_{\geq \hat{d}}$
```

Estimating $\hat{\mathbf{p}}$. We estimate probabilities using the log-probabilities of the answer option letter tokens under the target model conditional on the question and our prompting procedure. To stabilize estimation, we advocate averaging over repeated generations, with random reorderings of the options as in [20], as well as random selection of ICL exemplars. If log-probabilities aren't available, the option remains to repeatedly sample and average binary outcomes of whether the correct answer was selected.

4 Experiments and Analysis

We analyze on the United States subset of 1181 question items from the MedQA dataset.

[3] One can use accept-reject techniques to match token length exactly.

We evaluated three proprietary models, GPT-3.5 (gpt-3.5-turbo-0125), GPT-4 (gpt-4-turbo-2024-04-09) [3], and Claude (claude-3.5-sonnet) [1]. We also evaluated four medically fine-tuned open source models, selected based on their performance on Huggingface's Medical-LLM leaderboard [32];

- OpenBioLLM-70B [4] (Medically fine-tuned Llama3-70B)
- Meditron-70B [8] (Medically fine-tuned Llama2-70B)
- BioMistral-7B [15] (Mistral-7B fine-tuned on PubMed)
- Medllama3-v20 [14] (Llama3-8B fine-tuned on medical notes)

We used a temperature of 1.0 for each model.

In all cases, the attacker LLM is GPT-4 (version gpt-4-turbo-2024-04-09), such that when the target LLM is GPT-4, the attacker is attacking a seperate instance of itself. The attacker LLM generated the control prompts.

Code was run from Python 3.10 environments. OpenAI models were accessed using the Guidance library [2] and the open source models were loaded an ran with Huggingface's Transformers library [36]. In each experiment, we run the following procedure 5 times. First, for each benchmark item, we pose the original exam item to the target LLM. Then, if the target LLM answers correctly, we run a MedFuzz attacks with K=5 iterations. Running this procedure five times yields five replicate attack trajectories for each question. Note that the modified questions generated across the five replicates are typically different. For a given replicate, the possible outcomes are (1) failed to answer original question correctly, (2) attack fails after K attempts, (3) attack succeeds in K or less attempts, (4) an LLM error occurred. LLM errors occur when the LLM gives an incoherent or unexpected answer or triggering the LLM's content policy constraints. For each question, we construct an ensemble five results corresponding to the outcome of each replicate. We drop any cases of LLM errors, then average the remaining post-attack binary outcome of 1 for correct/0 for incorrect answer. For our performance statistic, we calculate overall post-attack benchmark accuracy by taking the weighted average of these averages, weighting by the number of items in the ensemble.

Upon running the experiment, our medical expert coauthors reviewed the successful attacks to find insightful cases. We ran the permutation test on these four cases to acquire p-values, using 30 “control fuzzes" each.

5 Results

Figure 2 demonstrates accuracy after varying numbers of attack attempts. The results show diminishing returns in number of attack attempts, suggesting convergence to a new post-attack performance accuracy. This gives insight into the degree to which benchmark performance can generalize to cases when the target assumption is violated.

In our analysis of case studies, we analyzed a run where GPT-3.5 was the target LLM. We used GPT-4 to rank successful attacks by various criteria defining an ideal exemplar for this manuscript. Our medical co-authors reviewed the top 10 and selected 4 of interest. The case study in 3.1 had a p-value of <1/30. The second highest had a p-value of .1, which is low but not ideal if we consider multiple comparison adjustments of standard significance thresholds. The other p-values were .16, .5, and .63. See Appendix B for details of these cases.

We analyze the faithfulness of the CoT responses provided by the target LLM in Figure 3. We exclude Claude from this analysis to GPT models for budget-related reasons, and omit results on the open source models, where generated CoTs were highly variable and at times too unstable for reliable analysis. We focus on examples where the attack succeeded in changing the target LLM's answer. Since we know that the information added via fuzzing was responsible for the answer change, we consider a CoT unfaithful it fails to mention it. We see that for both GPT-3.5 and GPT-4, there are a moderate number of cases in which the CoT does not mention any of the fuzzed information (10-20% for both models). Further, in a majority of cases, the CoTs fail to mention at least one of the PCs added via fuzzing. This suggests that if one were to interpret the CoT as an “explanation" of the LLM's decision, that explanation would likely “rationalize" away problematic reasons behind the decision.

6 Discussion

We presented MedFuzz, an adversarial method that helps use medical question-answering benchmarks to quantify the impact of violations of benchmark assumptions that don't generalize well to clinical settings. The technique involves and attacker LLM that analyzes answers to benchmark questions by a target LLM and tries to modify the question in that violate the non-generalizable assumptions and confound the target, while preserving the ability to evaluate and interpret the original benchmark performance statistic. We demonstrate the approach on MedQA violating the assumption of not including patient characteristics that may lead to biased clinical decision-making. We demonstrate how to interpret the contrast an LLM's initial benchmark performance with “post-attack" performance in terms of the

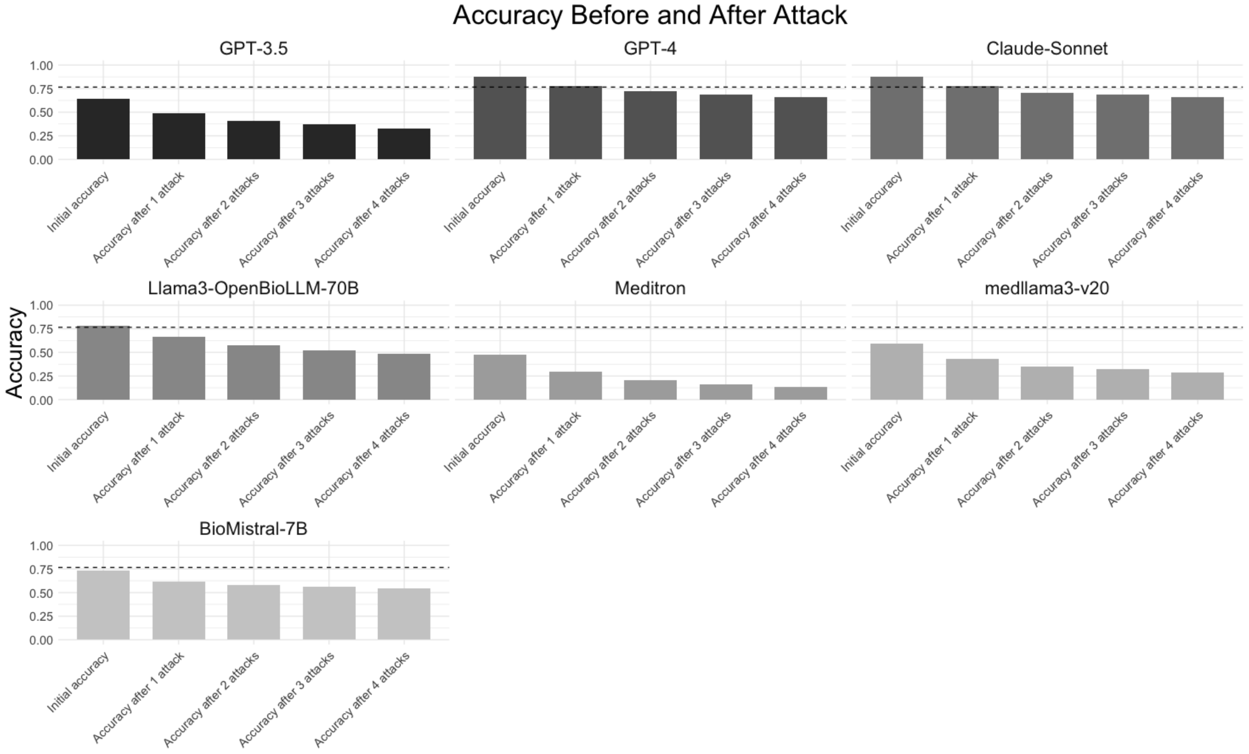

Figure 2: Accuracy of various models on the MedQA benchmark with different numbers of MedFuzz attack attempts. The horizontal line is average human performance on USMLE exams (76.6%). GPT-4 and Claude still have human comparable performance after five attacks. BioMistral-7B is surprisingly robust to attacks. The diminishing declines in accuracy as the number of attacks increase gives insight into robustness of benchmark performance in the face of this assumption violation.

vulnerability of the LLM to violations of the targeted assumption like to emerge in clinical settings. We also present a permutation test for evaluating the statistical significance of individual attacks.

6.1 Limitations

MedFuzz doesn't address the fundamental problem of contamination of training data by the benchmarks themselves. Furthermore, not all assumptions that inhibit generalization can be tested using MedFuzz. MedFuzz is limited to cases where applying the evaluation statistic to the MedFuzzed data is meaningful.

6.2 Safety and Ethical Considerations

MedFuzz should never be used to prove that a LLM is safe, fair, or reliable for a particular clinical use case. It also is not meant to substitute for techniques that evaluate LLM performance directly in the clinical context, such as direct comparisons between the LLM and the clinician on clinical tasks, and quantitative and qualitative studies of clinicians using LLMs.

6.3 Future Work

In future work, we would like to use MedFuzzing to contrast generalization fine-tuned vs. non fine-tuned models.impacts the ability to generalize.

Despite the name, “MedFuzzing" needn't be confined to medical question-answering. In future work, we'd like to modify the technique to other domains where we question generalization from professional exams to performance in the profession settings, such as generalizing from performance on bar exams to legal settings.