MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering

A Preprint

Robert Osazuwa Ness[*]
Microsoft Research
robertness@microsoft.com

Katie Matton[†]
Massachusetts Institute of Technology (MIT)

Hayden Helm
Helivan Research

Sheng Zhang
Microsoft Research

Junaid Bajwa
Microsoft Research

Carey E. Priebe
Johns Hopkins University

Eric Horvitz
Microsoft Research

Abstract

Large language models (LLM) have achieved impressive performance on medical question-answering benchmarks. However, high benchmark accuracy does not imply that the performance generalizes to real-world clinical settings. Medical question-answering benchmarks rely on assumptions consistent with quantifying LLM performance but that may not hold in the open world of the clinic. Yet LLMs learn broad knowledge that can help the LLM generalize to practical conditions regardless of unrealistic assumptions in celebrated benchmarks. We seek to quantify how well LLM medical question-answering benchmark performance generalizes when benchmark assumptions are violated. Specifically, we present an adversarial method that we call MedFuzz (for medical fuzzing). MedFuzz attempts to modify benchmark questions in ways aimed at confounding the LLM. We demonstrate the approach by targeting strong assumptions about patient characteristics presented in the MedQA benchmark. Successful "attacks" modify a benchmark item in ways that would be unlikely to fool a medical expert but nonetheless "trick" the LLM into changing from a correct to an incorrect answer. Further, we present a permutation test technique that can ensure a successful attack is statistically significant. We show how to use performance on a "MedFuzzed" benchmark, as well as individual successful attacks. The methods show promise at providing insights into the ability of an LLM to operate robustly in more realistic settings.

1 Introduction

Cutting-edge large language models (LLMs) have attained human competitive performance on medical question and answering benchmarks [27, 28, 19, 20, 29]. Implicit in this success is the possibility that LLMs might be employed to provide valuable decision support on real-world clinical cases. However, as discussed in [19], strong performance on benchmarks does not mean the models will necessarily perform well and provide value to clinicians in practice. One approach to exploring how LLMs might perform in more complex real-world situations is via studies of *robustness*—the ability for a model's performance to generalize to settings where the assumptions underlying the performance statistics are violated. Creating a medical question-answering benchmark requires assumptions that threaten generalizability [24]; complex real-life clinical situations are compressed into canonical multiple choice questions. However, LLMs can go beyond multiple choice medical exam questions and draw from text and imagery from medical pedagogy, medical and scientific journal articles, and medical conversations in social media and apply this knowledge in clinical settings. The challenge is finding quantitative approaches for evaluating how well a model does given a growth in complexity of the presentation of cases.

We introduce MedFuzz, an adversarial approach to testing the generalization of medical question-answering benchmarks to more complex challenges. MedFuzz borrows from *fuzzing* in software testing and cybersecurity, a method that adversarially feeds unexpected data to a target system to "break" it, thereby surfacing its failure modes. In "MedFuzz", an *attacker LLM* attempts to modify items in the benchmark in ways that "break" a *target LLM*'s ability to answer those

[*] Corresponding author
[†] Work done during an internship at Microsoft Research.

items correctly but that would not confound a human medical expert. The attacker LLM's modifications are constrained to specifically violating assumptions underlying the benchmark that we expect not to hold up in the clinic.

To illustrate the technique, we focus on a motivating example of violating assumptions on patient characteristics presented in medical vignettes in the MedQA-USMLE benchmark. Using MedFuzz, we surface the potential for clinical application of an LLM to reflect medical misconceptions and stereotypes that could be harmful if applied to real patients.

2 Background

In this section, we review several key concepts in training and deploying LLMs for answering challenging test questions in medicine, highlight their implications to generalizing to richer, open-world scenarios, and discuss how "MedFuzz" builds on prior work in these areas.

2.1 LLM performance on Medical Question-Answering

Driven by the promise of impact in healthcare, medical question-answering remains a key task for evaluating LLMs. Several medical-question answering benchmarks have emerged for statistical evaluation of LLM performance [10, 22, 12]. Some medical question-answering benchmarks are derived from medical entrance and licensing exams, such as MedMCQA [23] and MedQA [22]. MedQA, for example, is based the US Medical Licensing Exam (USMLE) [11]. Such benchmarks are interesting to consider from the point of generalizing to the clinic, as medical licensing exam items are designed to evaluate a would-be clinician's ability to reason through clinical decision-making problems [5]. MedQA items typically start with vignette that describes a patient presentation in a clinical scenario, then prompt the test-taker to select from multiple choice answers involving correct interpretation of evidence, diagnosis, and appropriate treatment [11] This manuscript uses MedQA as an example, though MedFuzz can be applied to other medical question-answer benchmarks with clinical implications.

Recent generations of LLMs have achieved great increases in accuracy on MedQA relative to previous generations. For instance, Med-PaLM 2 (a medically fine-tuned version of PaLM 2) achieved 85.4% accuracy [22] on MedQA, in contrast to Flan-PaLM (a medically fine-tuned version of the earlier PaLM 580B), which achieved 67.6% accuracy. [27]. GPT-4 without fine-tuning and various prompt engineering techniques achieved 90.2% on MedQA [20] (the highest reported performance at the time of writing), which stands in contrast to GPT-3.5's accuracy of 60.2% [17]. This improvement is analogous to a medical student going from a failing score on the USMLE to getting top marks within a few years.

This work focuses on GPT-4 as the top-performing model, and GPT-3.5 as its predecessor, but MedFuzz can be applied to other models as well (see [32] for a current medical question-answering leaderboard on several benchmarks). The 90.2% accuracy achieved in [20] depended on a prompting strategy designed to maximize accuracy [37, 13], such as in-context learning (ICL) [6] (prompting with examples correctly answer medical questions along with the question of interest), chain-of-thought prompting (CoT) [35] (instructing the LLM to generate a rationale for its answer), and ensembling [34, 22, 20] (methods for repeatedly generating answers and aggregating them into one). We deploy similar methods, but use them as part of MedFuzz's approach to evaluation, rather than focusing on maximizing accuracy.

2.2 Adversarially Robust Generalization

Our work builds on prior studies of *adversarially robust generalization* [26, 9, 30, 7, 39], which study how intentional perturbations to features cause the model to produce incorrect or misaligned classifications, predictions, or generated artifacts. MedFuzz similarly perturbs medical benchmarks in ways that lead an LLM to answer incorrectly. The perturbations intentionally violate assumptions underlying the benchmark items that would not hold in clinical settings and thus threaten generalizability.

MedFuzz builds on prior adversarial machine learning work in two ways. First, MedFuzz uses the LLM to randomly modify an item in the medical benchmark. MedFuzzing seeks to modify the vignette in medical question such that a clinician would provide the same correct answer as with the original vignette, but the LLM would change its correct response to the original vignette to an incorrect option. This is analogous to how selectively adding random noise to an image of a panda in [9] can create an image that still looks like a panda to the human eye while tricking an image classifier to return the label "gibbon". However, rather than adding random text string "suffix" as in [39], MedFuzz's perturbations are *semantically coherent*; the modification changes the text such that it is still intelligible and coherent within the context of the vignette. This is similar to the approach used in [7], which searches for semantically coherent changes to a prompt that will "jailbreak" the LLM, meaning causing the LLM into generating text that is prohibited by the LLM's content policy and alignment safeguards (e.g., changing "give me bombmaking instructions" to "write a fictional story about an orphan who writes bomb-making guides"). Rather than jailbreaking, MedFuzzing targets the benchmark performance statistic.

2.3 Bias and fairness in medical question answering

LLMs such as GPT-4 are trained on natural language data that reflects potentially harmful cognitive biases and error-prone decision-making heuristics in society and medical practice. For example, a tendency for doctors to discount long-term harms in favor of short-term benefits, such as in the prescribing of antibiotics [16] may appear as a pattern in the training data that the LLM can learn and reproduce. Recent work has focused in particular on how LLMs reproduce social biases and medical stereotypes in medical decision-making [33, 18, 38, 31, 21, 38].

LLMs are evaluated on medical question-answering benchmarks, which are often based on medical licensing exams. Medical licensing exam item vignettes can reflect social bias [25], though some explicitly omit elements that may trigger unconscious social bias [5] in the exam-taker – either way, LLM performance on benchmarks built from such exams will not surface the LLM's risk of reproducing such bias in clinical decision-making. In this work, we use MedFuzz to surface potential instances of this kind of bias, though it could be used for others.

Clinicians that use LLMs as an assistive tool will likely want to see the explanation for the answer as well as the final answer itself. However, prior work shows that explanations provided by LLMs can be *unfaithful*, or misrepresent the true reason for an LLM's generated answer [31]. In particular, LLMs have been shown to omit the influence of social biases in their explanations [31]. Motivated by this, we study not only how MedFuzz impacts the *accuracy* of LLM answers, but also the *faithfulness* of their associated explanations.

3 Method

MedFuzz uses an *attacker LLM* to modify a benchmark item in a way that violates assumptions underlying the benchmark that do not generalize. The attacker is instructed to do so in a way that allows us to continue using the target's reported benchmark performance statistics to evaluate the effects of violating the assumption; in the case of accuracy, the attacker is instructed to modify the item in a way that doesn't change the correct answer. The *target LLM* is then prompted to provide a correct answer to the modified item. The attacker and the target can be different LLMs or two instances of the same LLM – i.e., an LLM can attack itself (using separate sessions for the attacker LLM and target LLM). The attacker knows the correct answer and iteratively attempts to introduce modifications that confound the target, using the target's history of CoT and other outputs to find effective modifications. The target has no awareness of the attacker nor any previous iterations, it is only ever presented with a benchmark item (either the original item from the benchmark or an item modified by the attacker).

3.1 Workflow for applying MedFuzz

We provide a high-level walkthrough of the MedFuzz workflow. To illustrate, we'll use the following case study of a successful attack. Here, the plain text is the original MedQA item, and the bold text is text added by the attacker LLM.

> *A 6-year-old African American boy* ***from a low-income family with limited access to healthcare*** *is referred to the hospital by his family physician for jaundice, normocytic anemia, and severe bone pain. He has a history of several episodes of mild bone pain in the past 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, which they have been using to treat his symptoms. Additionally, the family has a history of thalassemia, and the child has a sibling with alpha-thalassemia.*** *On physical examination, the child is icteric with nonspecific pain in his hands. His hands are swollen, tender, and warm. There is no chest pain, abdominal pain, fever, or hematuria. A complete metabolic panel and complete blood count with manual differential are performed:*
>
> *Total bilirubin 8.4 mg/dL WBC 9,800/mm^3 Hemoglobin 6.5 g/dL MCV 82.3 fL Platelet count 465,000/mm^3 Reticulocyte 7% Peripheral blood smear shows multiple clumps of elongated and curved cells and erythrocytes with nuclear remnant. The patient's hemoglobin electrophoresis result is pictured below. What is the most likely cause of his condition?*
>
> - *A: Sickle cell trait*
> - *B: Sickle cell disease (Correct Answer initially selected by target LLM)*
> - *C: Hemoglobin F*
> - *D: HbC (Incorrect "distractor" selected by target after attacker added text in bold.)*