Assessing biomedical knowledge robustness in large language models by query-efficient sampling attacks

R. Patrick Xian[1] **Alex J. Lee**[1] **Satvik Lolla**[2] **Vincent Wang**[1]
Russell Ro[2,1] **Qiming Cui**[2,1] **Reza Abbasi-Asl**[1]

[1] *UC San Francisco* [2] *UC Berkeley*

Reviewed on OpenReview: *https://openreview.net/forum?id=pvol5JyVYB*

Abstract

The increasing depth of parametric domain knowledge in large language models (LLMs) is fueling their rapid deployment in real-world applications. Understanding model vulnerabilities in high-stakes and knowledge-intensive tasks is essential to quantify the trustworthiness of model predictions and regulating their use. The recent discovery of named entities as adversarial examples (i.e. adversarial entities) in natural language processing tasks raises questions about their potential impact on the knowledge robustness of pre-trained and fine-tuned LLMs in high-stakes and specialized domains. We examined the use of type-consistent entity substitution as a template for collecting adversarial entities for medium-sized billion-parameter LLMs with biomedical knowledge. To this end, we developed an embedding space, gradient-free attack based on powerscaled distance-weighted sampling to assess the robustness of their biomedical knowledge with a low query budget and controllable coverage. Our method has favorable query efficiency and scaling over alternative approaches based on blackbox gradient-guided search, which we demonstrated for adversarial distractor generation in biomedical question answering. Subsequent failure mode analysis uncovered two regimes of adversarial entities on the attack surface with distinct characteristics. We showed that entity substitution attacks can manipulate token-wise Shapley value explanations, which become deceptive in this setting. Our approach complements standard evaluations for high-capacity models, and the results highlight the brittleness of domain knowledge in LLMs[1].

1 Introduction

LLMs such as pre-trained and finetuned generalist language models (GLMs) and their domain-adapted versions (Zhao et al., 2023; Wang et al., 2023a) incorporating transformer architectures are emerging as the technological backbone of next-generation search engines and knowledge bases (Petroni et al., 2019; Sung et al., 2021). They are increasingly used in knowledge-intensive tasks by the general public without sufficient assessment of inherent issues in trustworthiness and safety (Clusmann et al., 2023; Spitale et al., 2023; Si et al., 2023). Adversarial examples are valuable probes for model vulnerability and robustness, and are essential for performance improvements and regulatory audits (Shayegani et al., 2023). Adversarial attacks are classified by query scaling behavior into low- and high-query-budget attacks (see Fig. 1a). Collecting adversarial data at scale, either for adversarial training or robustness assessment, requires query-efficient attacks, first empirically studied in the vision domain (Ilyas et al., 2018; Cheng et al., 2019; Shukla et al., 2021). For language models, adversarial examples are generated through guided search (Alzantot et al., 2018), sampling and probabilistic methods (Ren et al., 2019; Yang et al., 2020; Yan et al., 2022), or collected dynamically by human operators (Jia & Liang, 2017; Wallace et al., 2019b). Most of these methods require a

[1]The code developed and datasets used for the work are available at https://github.com/RealPolitiX/qstab.

large amount of model evaluations (in the hundreds to thousands or beyond per instance) or whitebox access to model internals, and their evaluations tend to focus on small models (with < 1B parameters) (Maheshwary et al., 2021; Yu et al., 2024), which are insufficient in the era of LLMs. Sampling-based attacks obviate the need for predetermined search heuristics and can operate with only blackbox model access and a low query budget. Demonstrations of query-efficient attacks on LLMs meet an emerging demand for their continuous monitoring (Metaxa et al., 2021), which is particularly important for deploying LLMs in specialized domains.

Question answering (QA) has been the proving ground in recent achievements of biomedical language models (BLMs), here referring to GLMs adapted to the biomedical domain (Singhal et al., 2023; Kung et al., 2023; Liévin et al., 2024; Saab et al., 2024). Knowledge and reasoning components in biomedical QA assess the understanding of diverse concepts, which are represented by domain-specific named entities (NEs). The present work aims to identify adversarial entities in biomedical QA for both GLMs and BLMs. While GLMs may have acquired general biomedical knowledge during model development, BLMs gain further domain knowledge primarily through finetuning on a variety of text data from clinical notes (Lehman et al., 2023), electronic health records (Wornow et al., 2023), to abstracts or excerpts of scholarly publications (Tinn et al., 2023). Although QA is a frequently employed paradigm for model performance evaluation (Gardner et al., 2019; Robinson & Wingate, 2023), currently, biomedical QA still lacks domain-specific robustness assessments and its use in the real-world applications, albeit promising, is still in its infancy (Kung et al., 2023). Evaluating model robustness for high-stakes, specialized domains requires customized perturbations apart from traditional benchmarks (Cecchini et al., 2024). Deployable LLMs in the biomedical domain should have robust contextual knowledge and language understanding of domain-specific entities.

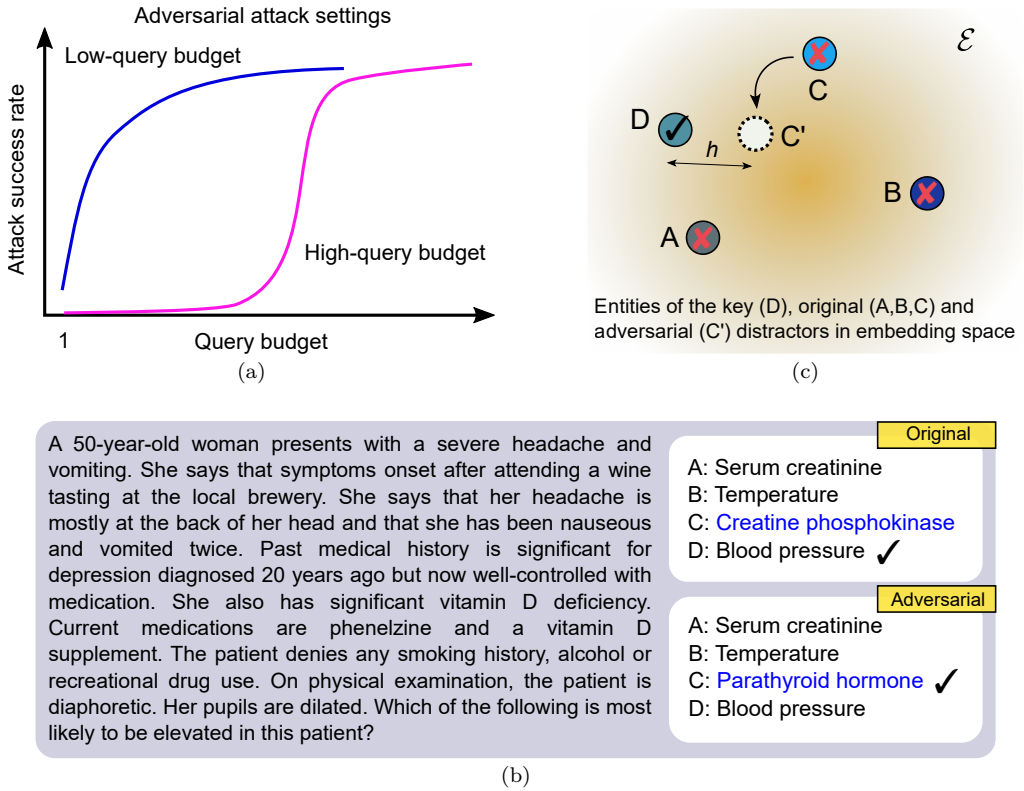

Figure 1: Entity substitution attack on QA with adversarial distractors. (a) Typical query scaling curves in low- and high-query-budget attack settings. (b) An adversarial distractor example found by type-consistent entity substitution (highlighted in blue). The correct (top) and incorrect (bottom) model responses (check-marked) before and after perturbation are included. (c) Illustration of the attack scheme in embedding space by PDWS for the example in (b). $\mathcal{E}$ represents the vocabulary set. D is the key to the question, A-C the original distractors, C$'$ is an adversarial distractor at distance h from the key D.

In this work, we proposed and compared sampling- and search-based query-efficient adversarial attacks for evaluating the knowledge robustness of LLMs in the entity-rich, biomedical domain using type-consistent entity substitution (TCES). We investigated the setting where a large number of viable replacements ($N_{\mathcal{E}} > 1000$) per attack exist at only a small number of permissible text locations, which is prevalent in specialized domains where the NEs have shared characteristics such as their type and (word or phrase) structure. The setting differs from traditional synonym substitution (Liu et al., 2022; Yu et al., 2024), which have a significantly smaller number of viable replacements with matching word senses but can operate on many potential text locations. To relate to real-world scenarios (Apruzzese et al., 2023), we consider the attacks in the blackbox and hard-label setting with a limited query budget ($B \ll N_{\mathcal{E}}$). We used the attacks to generate adversarial examples of distractors (see Fig. 1b), or *adversarial distractors* in biomedical QA tasks. Our main contributions include: (i) Introduction of powerscaled distance-weighted sampling (PDWS) method in embedding space to generate adversarial entities. Our work reports the first use of DWS for adversarial attacks in text[2]. The attack scheme has a broad coverage of the embedding space centered around an anchor point (illustrated as the key D in Fig. 1c) and maps the entity-level robustness landscapes of different LLMs in two regimes (near and far). (ii) Introduction of the `FDA-drugs` and `CTD-diseases` datasets as the source vocabularies of the perturbation sets, along with an entity type-annotated version of existing biomedical QA datasets MedQA-USMLE (Jin et al., 2021) and MedMCQA (Pal et al., 2022). Using these data, we demonstrated the vulnerability of LLMs to two of the most common NEs in the biomedical domain (Wei et al., 2019): drug and disease names (see example in Fig. 1b). (iii) We unified sampling- and gradient search-based methods in discrete text space and formulated TCES as an attack template, which we used for quantitative comparison of the scaling behaviors of the attacks in the same footing. We found that in the blackbox setting, sampling attacks can have favorable query scaling over gradient-based attacks, which are gradient quality-limited within a fixed query budget. (iv) By investigating the relationship between Shapley value-based model explanation (Chen et al., 2023) and adversarial attacks, we uncovered the defining signatures of successful attacks as a result of entity substitution.

2 Related works

Text substitution attacks Adversarial attacks in neural language models have been demonstrated from the character to the sentence level (Zhang et al., 2020). The majority of text substitution attacks employ typos (Gao et al., 2018) and synonyms (Mozes et al., 2021), which effectively impose lexical constraints. Word frequency has been used as a metric to select adversarial examples in (Mozes et al., 2021). Synonym substitution attacks have been demonstrated in sentiment analysis (Liu et al., 2022; Yu et al., 2024). Alzantot et al. (2018) used a genetic algorithm to search for adversarial word substitution. Model sensitivity to entity-level perturbations including substitution and swapping create problematic situations such as knowledge conflicts (Longpre et al., 2021) for knowledge-intensive tasks, compromised performance in machine reading comprehension (Yan et al., 2022) and table interpretation tasks (Koleva et al., 2023).

Robustness in QA The robustness of QA models has been studied using perturbation methods (Jia & Liang, 2017) and meaning-preserving transformations (Gan & Ng, 2019; Elazar et al., 2021). Sen & Saffari (2020) introduced a set of perturbations to assess language model generalization between QA datasets. Richardson & Sabharwal (2020) proposed a distractor design method using proximity in semantic space to probe model understanding of word senses. Awadalla et al. (2022) examined the connection between in- and out-of-distribution robustness for different LLMs. The performance of LLMs is sensitive to the ordering of choices (Pezeshkpour & Hruschka, 2023; Zheng et al., 2024). Overall, the brittleness of knowledge in LLMs remains a primary limitation for their applications (Elazar et al., 2021; Augenstein et al., 2024).

Characteristics of adversarial examples The defining characteristic of adversarial examples is the induction of false predictions in trained models that would otherwise predict accurately. Other characteristics include (i) their generality such as transferability and universality (Zou et al., 2023). Transferability refers to the effectiveness of the same example being adversarial to distinct models (Papernot et al., 2016). Universal adversarial perturbations are *reusable* across many examples tested on the same model (Moosavi-Dezfooli

[2]Throughout the text, we use DWS to refer to the class of sampling method that employs distance-dependent probability weights. The powerscaled DWS is a type of DWS, so is the other version introduced in Wu et al. (2017).

et al., 2017; Wallace et al., 2019a). (ii) The quality of adversarial examples may be judged by their perceptibility to human (Dyrmishi et al., 2023) or simple defense mechanisms such as grammar check. (iii) Adversarial attacks can operate in the low- and high-query-budget settings (Shukla et al., 2021) with different query scaling behaviors. (iv) Adversarial examples by specifically designed perturbations can obstruct model explanation (Noppel & Wressnegger, 2024) and render it unuseful.

3 Entity-centric adversarial attacks

3.1 Adversarial distractor generation

Definition 3.1. A multiple-choice question $\mathcal{Q}$ is represented by a tuple, $\mathcal{Q} = (\mathcal{C}, \mathcal{S}, \mathcal{D}, k)$, with the context $\mathcal{C}$, question stem $\mathcal{S}$, a distractor set $\mathcal{D} = \{d_i\}_{i=1}^{l}$, and a key k. The distractors are the wrong choices given to confuse test-takers, while the key is the correct one (CH & Saha, 2020).

Definition 3.2 (Adversarial distractor)**.** Given the original (or unperturbed) and a perturbed multiple-choice question pair, $\mathcal{Q} = (\mathcal{C}, \mathcal{S}, \mathcal{D}, k)$ and $\mathcal{Q}' = (\mathcal{C}, \mathcal{S}, \mathcal{D}', k)$, which differ by only a distractor, $\mathcal{D}' \setminus (\mathcal{D} \cap \mathcal{D}') = \{d'\}$. The distractor d' is adversarial if the perturbed question elicits a false response from the model, while it answers correctly to the unperturbed question,

$$\operatorname{argmax} \Pr_{\text{LLM}}(Y = k | \mathcal{C}, \mathcal{S}, \mathcal{D}', k) \neq \operatorname{argmax} \Pr_{\text{LLM}}(Y = k | \mathcal{C}, \mathcal{S}, \mathcal{D}, k). \tag{1}$$

The threat model in Definition 3.2 refers to an untargeted attack, while the targeted version only requires restricting the answer to the perturbed question to a specific distractor. A simple way to construct the perturbed distractor set $\mathcal{D}'$ in $\mathcal{Q}'$ is to substitute a distractor d by d', or $\mathcal{D}' = (\mathcal{D} \setminus \{d\}) \cup \{d'\}$. The substitution is subject to the constraint $\|\mathcal{Q}' - \mathcal{Q}\| < \epsilon$, where ϵ is the perturbation budget, and the norm $\|\cdot\|$ refers to edit distance. For simplicity, we assume the distractors contain only NEs and only one NE each (such as in Fig. 1b). In Appendix A, we extend Eq. 1 to a more general case with multiple entities and additional non-entity text in the distractors.

Example 3.1. In the example of Fig. 1b, the stem $\mathcal{S}$ is the last sentence on the left side, "Which of the following is most likely to be elevated in this patient?" The context $\mathcal{C}$ is all the text before the stem, "A 50-year-old woman ... Her pupils are dilated." *The original distractor set $\mathcal{D}$ = {Serum creatinine, Temperature, Creatine phosphokinase, Blood pressure}, the perturbed distractor set $\mathcal{D}'$ = {Serum creatinine, Temperature, Parathyroid hormone, Blood pressure}, the key k is Blood pressure, or equivalently, D*[3].

Unlike text addition attacks that introduce distracting information (Jia & Liang, 2017; Wallace et al., 2019a), the edit constraint on $\mathcal{Q}$ that is usually enforced in text adversarial attacks (Roth et al., 2024) doesn't preclude large semantic perturbations at the entity level in the distractors. In this work, we draw perturbations from a finite perturbation set, denoted as $\mathcal{E}_{\text{-}k} \coloneqq \mathcal{E} \setminus \{k\}$, with $\mathcal{E} = \{e_j\}_{j=1}^{m+1}$ being the entity dataset as illustrated in Fig. 1a. In practice, $\mathcal{E}$ may be compiled from potential model use cases or curated by regulatory bodies in a model audit. Next, we discuss a unified view of discovering adversarial entities in the embedding space using blackbox (or gradient-free) attack methods based on sampling and search.

3.2 Powerscaled distance-weighted sampling

Definition 3.3 (Text span representations)**.** Given the token sequence representation of a span e, we define its corresponding vector representation in a text embedding of dimension l as $\mathbf{e} \in \mathbb{R}^l$.

We consider sampling with distance-dependent probability weights from the embedding space of NEs. The concept of distance-weighted sampling (DWS) was previously introduced in contrastive learning in computer vision, where the Euclidean distance between samples and an anchor point was used and the probability weights depend on the embedding dimensionality (Wu et al., 2017). We propose a more flexible version independent of the embedding dimensionality using a powerscaled distance with a tunable hyperparameter.

[3]For simplicity, in this work, we don't use different symbols to distinguish between the content of a distractor (or a key) from its index (e.g. A, B, C, D).

Specifically, for a key k, we assign a distance-dependent probability weight $p_j \in [0, 1]$ to entity e_j such that the distractor d is substituted by

$$d' = e_j \quad \text{w.p.} \quad p_j = \frac{h^n(k, e_j)}{\sum_j h^n(k, e_j)}. \tag{2}$$

Here, $h(k, e_j)$ is a distance metric between the anchor point k and e_j, while the exponent $n \in \mathbb{R}$ controls the shape of the probability weight distribution. When $n > 0$, as n increases, the entities further away from k are more preferably weighted. When $n < 0$, the sampling method employs inverse-distance-scaled weights, and the entities closer to k are more preferably weighted. In the subsequent evaluations, we use $h(k, e_j) = \texttt{CosineDist}(\mathbf{k}, \mathbf{e}_j) = 1 - \mathbf{k} \cdot \mathbf{e}_j / (\|\mathbf{k}\| \cdot \|\mathbf{e}_j\|)$, calculated in a text embedding of choice, where $\mathbf{k}$ and $\mathbf{e}$ are the respective vector representations of k and e. The formalism accommodates a deterministic regime as a limiting case[4], when $n \to +\infty$, and p_j is vanishingly small for all but the entity furthest from k,

$$d' = e_j \quad \text{s.t.} \quad e_j = \operatorname*{argmax}_{e_j \in \mathcal{E}_{\text{-}k}} h(k, e_j). \tag{3}$$

When $n = 0$, p_j is the same for each entity in $\mathcal{E}_{\text{-}k}$, corresponding to uniform random sampling. A short proof of the limiting cases is given in Appendix B. Therefore, our approach bridges the probabilistic and deterministic approaches to sampling adversarial examples.

3.3 Sampling view of zeroth-order adversarial attack

Blackbox attacks using approximate gradients may proceed with zeroth-order optimization (ZOO), where the gradients are estimated by finite difference to guide the search (Ilyas et al., 2018; Cheng et al., 2019). In the discrete text space, Berger et al. (2021) developed the DiscreteZOO attack, which contains three components: the word importance ranker, the candidate sampler, and the gradient-based optimizer. It operates at the word level and has been validated on BERT-sized models for synonym substitution attacks. The last two components aim to find a replacement text span e' in each iteration through the gradient update rule

$$\mathbf{e}' = \mathbf{e}_0 - \lambda \widehat{\nabla}_{\text{ZO}} \text{LLM}(\mathcal{Q}; e_0). \tag{4}$$

Here, e_0 refers to the original text span, $\mathbf{e}'$ and $\mathbf{e}_0$ are the vector representations of e' and e_0, respectively. λ the learning rate, $\widehat{\nabla}_{\text{ZO}}$ indicates the zeroth-order gradient estimator, and $\widehat{\nabla}_{\text{ZOLLM}(\mathcal{Q};e_0)}$ has the same dimensionality as $\mathbf{e}'$ or $\mathbf{e}_0$. We use $\text{LLM}(\mathcal{Q}; e_0)$ to denote the LLM text output with the unperturbed question $\mathcal{Q}$ (containing e_0) as the input, then the multi-point ($M \geq 2$) gradient estimate is constructed by querying the LLM M times and computing

$$\widehat{\nabla}_{\text{ZO}} \text{LLM}(\mathcal{Q}; e_0) = \mathbb{E}_{\mathbf{u}} \left[\frac{\Delta \text{LLM}(\mathcal{Q}; e_0)}{\Delta \mathbf{e}_0} \mathbf{u} \right] = \frac{1}{M} \sum_{i=1, e_i \in \mathcal{E}_{\text{-}k}}^{M} \frac{(\text{LLM}(\mathcal{Q}'; e_i) - \text{LLM}(\mathcal{Q}; e_0)) \cdot (\mathbf{e}_i - \mathbf{e}_0)}{\|\mathbf{e}_i - \mathbf{e}_0\|^2}. \tag{5}$$

This is an example of the random directions stochastic approximation (Nesterov & Spokoiny, 2017), where $\mathbf{u} = (\mathbf{e}_j - \mathbf{e}_0)/\|\mathbf{e}_j - \mathbf{e}_0\|$ is a randomly oriented vector in the embedding space corresponding to the neighboring entity e_j. Berger et al. (2021) computed Eq. 5 by random sampling of neighbors (in candidate sampler) within a similarity threshold for synonyms. The procedure was iterated once for all ranked attack locations. Given a noisy gradient estimate $\widehat{\nabla}_{\text{ZO}} \text{LLM}(\mathcal{Q}; e_0)$ and a discrete search space, we can rewrite Eq. 4 effectively as deterministic DWS in the embedding space by

$$d' = e_j \quad \text{s.t.} \quad e_j = \operatorname*{argmin}_{e_j \in \mathcal{E}_{\text{-}k}} h(e_0, e_j), \tag{6}$$

where $h(e_0, e_j) = \texttt{CosineDist}(\mathbf{e}_0 - \lambda \widehat{\nabla}_{\text{ZO}} \text{LLM}(\mathcal{Q}; e_0), \mathbf{e}_j)$, with e_0 as the anchor point, and argmin indicates the closest member within the text embedding. Using argmin is essential because the exact location $h(e_0, e_j)$ in the discrete embedding space is usually not occupied. While a gradient estimate of sufficient quality is essential for a successful attack, increasing M will count towards the query budget. Therefore, in the budgeted setting, a tradeoff should exist between the efficacy of the attack and the number of queries consumed for each gradient estimation.

[4]The limit $n \to -\infty$ indicates replacing the distractor d with k at all times. Since k is not in the perturbation set $\mathcal{E}_{\text{-}k}$, this limit is not attained.

4 Implementing attacks on biomedical QA

4.1 Dataset selection

Entity datasets To relate to real-world scenarios, we sourced the vocabulary datasets of drug and disease names from existing public databases. The drug names dataset (`FDA-drugs`) contains over 2.3k unique entities from known drugs approved by the United States Food and Drug Administration (FDA) and curated by Drug Central[5] (Ursu et al., 2017). FDA is a world authority that approves the legal distribution of drugs, therefore the dataset represents the drug-entity names encountered in daily life. The disease names dataset[6] (`CTD-diseases`) contains over 9.8k unique entities from the Comparative Toxicogenomics Database (Davis et al., 2009), which contains a comprehensive collection of uniquely documented chemical-gene-disease interactions for humans. Both databases are curated by domain experts through regular updates, and unlike traditional text corpora (Mohan & Li, 2019), the entities in these two datasets contain no redundancy. The 2023 version of the datasets was used after minor data processing.

Biomedical QA datasets We selected over 9.3k questions from the MedQA-USMLE (Jin et al., 2021) dataset and over 3.8k questions from the MedMCQA (Pal et al., 2022) dataset for benchmarking. Both datasets are entity-rich and cover a wide range of biomedical specialties and topics. The MedQA-USMLE dataset contains long-context questions that are used to assess medical students in the United States. The MedMCQA dataset contains short questions used in medical exams in India. They have been used in recent works for assessing the biomedical knowledge in LLMs (Singhal et al., 2023; Han et al., 2023; Liévin et al., 2024; Saab et al., 2024). Both datasets are publicly available and don't contain personal information. We annotated the NEs in the two datasets according to the entity types of the Unified Medical Language System (UMLS) (Bodenreider, 2004) using `scispaCy` (Neumann et al., 2019). We divided the QA datasets into drug- or disease-mention questions according to the UMLS entity types (see Appendix C) of NEs in distractors.

4.2 Type-consistent entity substitution (TCES)

Implementing attacks that target entities requires accounting for type consistency and compatibility with simultaneous multi-word (or span-level) operations, because of the inseparable nature of the entity components (e.g. growth hormone, acquired immunodeficiency syndrome). This effectively results in more substitution patterns (one to two words, two to one word, two to two words, two to three words, etc.) than the more common one-to-one synonym substitution (Mozes et al., 2021). We describe TCES in Algorithm 1 as a general and contextualized attack template for entities, with the goal of adversarially modifying model outputs with only one span-level substitution, specified by token boundaries. The template can accommodate single- or multi-query attacks and it is applicable to any task that may be formatted into QA (Gardner et al., 2019).

The initial steps in TCES involve entity recognition and typing (see Appendix C). For distractor generation, the type-filtered entities in the choices (Ent_{tfch}) are separated into key (Ent_{key}) and distractor ($\text{Ent}_{\text{distrc}}$) entities. In the biomedical domain, the entity type information is readily available from the UMLS (Bodenreider, 2004) semantic groups, such as *Chemicals & Drugs* and *Disorders* used for demonstration later. The `RankSelect` step selects the text span to attack ($\text{Ent}_{\text{victim}}$). In the PDWS attack, the type-matched entity in the distractors closest to the key is selected. The purpose is to maintain consistency and is not uniquely associated with the effects discussed later. In ZOO-based attacks, `RankSelect` corresponds to the word importance ranker. Then, the potential replacement entities are selected using `Sampler` (Eq. 2 for PDWS or Eq. 6 for DiscreteZOO), which also removes any duplicates of the key in the vocabulary ($\text{Ent}_{\text{vocab}}$). Sampling is carried out without replacement from the corresponding perturbation set $\mathcal{E}_{\text{-}k}$ (constructed with `FDA-drugs` or `CTD-diseases` datasets), equivalent to enforcing type consistency in the replacement entities to maintain consistency with the entity to attack ($\text{Ent}_{\text{victim}}$). `GoalFunc` outputs a value from a loss function or from directly comparing with the answer to determine if the attack is successful.

4.3 Attack execution on LLMs

[5] https://drugcentral.org/download
[6] https://ctdbase.org/downloads/

Victim LLMs We selected both generalist (GLMs) and specialist (BLMs) models for robustness evaluation via blackbox adversarial attacks. The criterion we adopted here for a model to have biomedical knowledge is that it should have a baseline performance better than random guessing (e.g. an accuracy of 0.25 for a specific four-choice QA task) in the evaluated domain-specific task. For GLMs, we used instruction-finetuned T5 (Flan-T5) series of models (Chung et al., 2024) and UL2 model (Tay et al., 2022) (Flan-UL2). For BLMs, we used MedAlpaca-7B (Han et al., 2023), MedLlama-3-8B-v2.0 from John Snow Labs (JSL-MedLM3-8Bv2)[7], Llama2-Medtuned-13B (LM2-Medtuned-13B) (Rohanian et al., 2024), and Palmyra-Med-20B (Kamble & Alshikh, 2023) from Writer. The selected open-source models have sizes in $\sim$ 1-20B range. Besides, we also evaluated GPT-3.5 from OpenAI.

Algorithm 1 TCES attack template for collecting adversarial entities in distractors.

Inputs: Question Q, LLM, entity type τ, budget B.
Internals: Number of choices N_{ch}, text t, entity Ent, embedding Emb, key or correct answer k.
Outputs: Number of queries and the replacement entity, None if unsuccessful after all attempts.

```
TCESAttacker(Q, Model, τ, B)
    t_choices, k ← Q.choices, Q.answer
    for c ← 1 to N_ch do
        t_ch ← t_choices[c]
        Ent_ch ← NERecognizer(t_ch)
        Ent_tfch ← TypeFilter(Ent_ch, τ)
    Ent_key, Ent_distrc ← SplitByLabel(Ent_tfch)
    Ent_victim ← RankSelect(Ent_key, Ent_distrc, Emb)
    i ← 0
    while i < B do
        Ent_perturb ← Sampler(Ent_key, Ent_victim,
                              Ent_vocab, Emb)
        Q' ← Q (Ent_victim → Ent_perturb)
        k' ← Model(Q')
        if GoalFunc(k', k) ← 1 then
            return i, Ent_perturb
        Ent_vocab ← Ent_vocab \ Ent_perturb
        i ← i + 1
    return
```

Attack settings All models were evaluated at zero temperature or in the non-sampling setting and model inference was conducted in the zero-shot setting with only basic prompt instructions (see the prompt structure in Appendix D). Model inference of Palmyra-Med-20B (Kamble & Alshikh, 2023) used 4-bit quantization to improve speed (Dettmers & Zettlemoyer, 2023). The evaluation settings resulted in our somewhat lower baseline accuracies than the reported ones which are often achieved with the help of few-shot prompting and prompt optimization. The DiscreteZOO attack (Berger et al., 2021) was originally implemented in the TEXTATTACK framework (Morris et al., 2020b), which we updated to be compatible with span-level perturbation using its boundaries, multi-GPU split-model inference, and decoder-based billion-parameter LLMs by modifying the attack recipe. The following three aspects in the benchmarks are in our focus.

- **Query budget**: We used fixed query budgets (B) for three main types of attacks: (i) Single-query sampling-based attacks were used as the reference because the DiscreteZOO attack requires a minimum of 3 model queries; (ii) Multi-query attacks used a budget of 8 for reasonable computational cost across all models and attack settings for both sampling- and search-based attacks; (iii) The query scaling trends of specific LLMs and different attack settings were investigated with a series of query budgets under 100 per input instance. Here, we also included the deterministic versions of PDWS to sample the nearest and farthest elements with respect to an anchor point. The element was taken from the perturbation set and the distances (between entities) were calculated using cosine in a chosen text embedding. Their query-limited versions are refered to as B-nearest and B-farthest element sampling, respectively, where all B nearest or farthest elements were used sequentially to attack models in query scaling studies. Detailed definitions are given in Appendix B.4.

- **Text embedding**: Both the PDWS and DiscreteZOO attacks require sampling from a text embedding. The counter-fitted GloVe embedding (Mrkšić et al., 2016) was the default for DiscreteZOO, but GloVe is poor for representing biomedical vocabulary (Wang et al., 2018). We chose two recent embedding models with improved latent representations of text relations for the benchmarks: (i) CODER (Yuan et al., 2022), a BERT-structured embedding model contrastively pretrained to distinguish UMLS knowledge graph concepts; (ii) GTE-base (Li et al., 2023b) from contrastive learning of general web-scale text data.

[7] https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0

- **Core hyperparameters**: For single-query PDWS attacks, we tuned the hyperparameter n within the interval of [-50, 50] using grid search with a step of 5 or 10 because the local maxima of ASR appear on the positive and negative sides. For multi-query PDWS attacks, hyperparameter n was briefly re-tuned around its optimal value in the single-query attack. Most tuned hyperparameters fall within [-30, -5] and [5, 30]. Operating the DiscreteZOO attack in a compatible setting for entities requires two key modifications: (i) Use a random word ranking, which allows the attack algorithm to operate in the low-query-budget setting. (ii) Disable the similarity threshold for candidate point selection in gradient estimation because only entity type consistency is required. The ASR was greatly improved after this step. Benchmarks of the DiscreteZOO attacks were carried out with these settings.

5 Finding and assessing adversarial entities

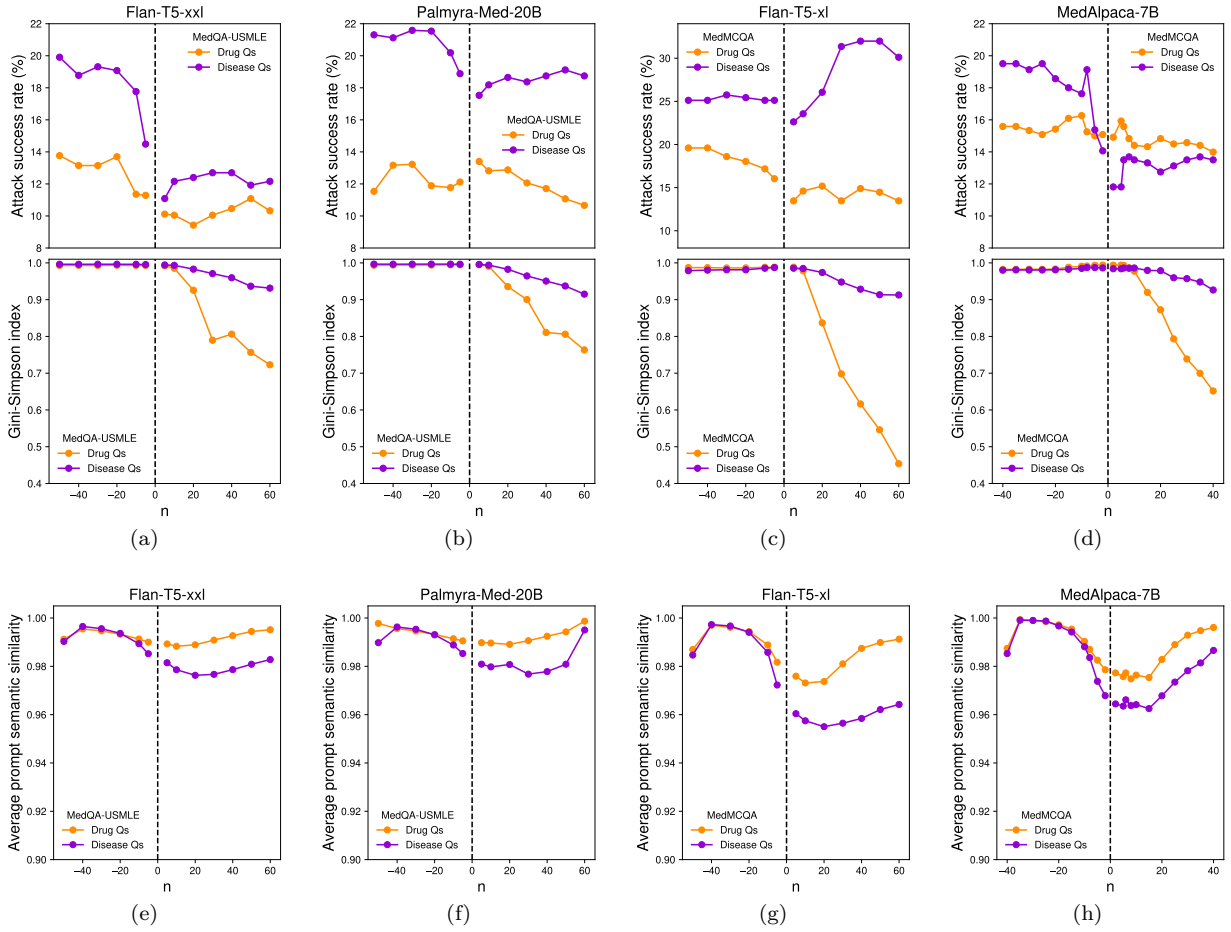

Figure 2: Powerscaled DWS of adversarial distractors exhibits a two-regime effect at negative and positive n values (see Eq. 2) in ASR (top) and diversity index (bottom) of replacement entities in successful attacks. Local maxima in ASR with a finite n are also present in each regime. The vertical dashed line indicates the location of random sampling. The observed similar behaviors are compared across models and datasets in (a) Flan-T5-xxl on MedQA-USMLE, (b) Palmyra-Med-20B on MedQA-USMLE, (c) Flan-T5-xl on MedMCQA, (d) MedAlpaca-7B on MedMCQA. Disease and drug-mention questions are separated by colors. The average prompt semantic similarity displayed in (e)-(h) is calculated for the successful attacks obtained from the corresponding attack settings in (a)-(d), respectively.

Model (size)	Entity mention	MedQA-USMLE			MedMCQA		
		Baseline (Acc)	RandS@1 (ASR %)	PDWS@1 (-/+) +CODER (ASR %)	Baseline (Acc)	RandS@1 (ASR %)	PDWS@1 (-/+) +CODER (ASR %)
Flan-T5-xl (3B)	Drugs	0.318	10.4	**13.5** / 10.7	0.265	14.3	**19.6** / 15.1
	Diseases	0.335	11.0	**16.1** / 12.5	0.261	23.8	25.7 / **31.8**
Flan-T5-xxl (11B)	Drugs	0.323	9.6	**13.6** / 11.2	0.291	12.7	**19.6** / 16.2
	Diseases	0.345	12.2	**20.0** / 12.8	0.376	15.2	**26.6** / 23.9
Flan-UL2 (20B)	Drugs	0.382	8.6	**12.3** / 10.5	0.416	13.9	**15.1** / 12.7
	Diseases	0.396	10.1	**15.7** / 13.1	0.408	15.7	**25.5** / 17.7
GPT-3.5 (175B)*	Drugs	0.438	9.1	**11.2** / 10.7			
	Diseases	0.453	8.6	**11.5** / 8.8			
MedAlpaca-7B	Drugs	0.479	7.2	8.2 / **9.6**	0.448	13.9	**16.2** / 15.9
	Diseases	0.498	8.4	**12.5** / 9.2	0.434	14.8	**19.1** / 13.7
JSL-MedLM3-8Bv2*	Drugs				0.827	10.4	10.2 / **11.8**
	Diseases				0.771	9.5	**18.5** / 11.3
LM2-Medtuned-13B	Drugs	0.371	16.4	**18.1** / 17.8	0.568	15.3	17.3 / **18.1**
	Diseases	0.398	19.6	**23.1** / 19.6	0.515	19.0	**27.6** / 21.0
Palmyra-Med-20B	Drugs	0.382	12.6	13.3 / **13.5**	0.614	10.2	**13.8** / 12.7
	Diseases	0.441	16.9	**21.5** / 18.7	0.624	15.4	**20.1** / 11.9

Table 1: Single-query (budget = 1) attack success rates (ASRs) on subsets of MedQA-USMLE and MedMCQA datasets by perturbing the drug or disease entity mentions. Zero-shot baseline (Baseline) accuracy is included for reference next to the ASR of sampling attacks using perturbation by random sampling (RandS), powerscaled distance-weighted sampling (PDWS). The "-" and "+" signs indicate the two regimes ($n < 0$ and $n > 0$). For each model, the highest ASR for a type of entity-mention question in each dataset is bolded. The models with an asterisk (*) were only evaluated on one dataset to limit the computational cost.

5.1 Two regimes of adversarial entities

Diversity quantification of adversarial entities The embedding space picture of the substitution attack in Fig. 1a provides a geometric interpretation of PDWS to examine different types of adversarial examples by tuning n. We use the Gini-Simpson index (Rao, 1982), $\eta_{\text{GS}} = 1 - \sum_j q_j^2$, to quantify the diversity of adversarial examples and a low value indicates less diversity. Here, $q_j \in (0, 1]$ is the probability of occurrence of the entity e_j among all instances of adversarial entities obtained from sampling, and $\sum_j q_j = 1$. A low diversity index indicates the reusability and is a proxy indicator for the universal adversarial attack regime (Moosavi-Dezfooli et al., 2017; Wallace et al., 2019a). Combining the ASR dependence on diversity and hyperparameter n in Fig. 2 reveals two regimes: (i) the adversarial entities semantically close to ($n < 0$) the key k have a greater diversity; (ii) the adversarial entities semantically far from ($n > 0$) the key k are more reusable. TCES in these two regimes both can lead to an increase in successful attacks from random sampling (equivalent to $n = 0$), as shown in Tables 1-2. Fig. 2 shows that a larger positive n in Eq. 2 samples more universal adversarial entities than in other situations, leading to a marked drop in the diversity index. Moreover, there exist local maxima of ASR in both regimes with finite n, which were discovered by hyperparameter tuning. We call the behavior the *two-regime effect* and they were observed in almost all models evaluated here in varying magnitudes for both QA datasets, thanks to the controllable spatial coverage of sampling methods incorporating distance information. In both single-query and multi-query attacks, the benchmarks show that the $n < 0$ regime of PDWS attacks tends to have a greater influence on model performance than the $n > 0$ regime, although prominent counterexamples do exist (MedAlpaca-7B and Llama2-Medtuned-13B (for drug-mention questions), and Flan-T5-xl (for disease-mention questions) in Table 1).

Semantic distortion by adversarial entities Besides ASR, evaluating the semantic distortion is another way to quantify the effectiveness of adversarial attacks. We estimated the semantic distortion of entire prompts in successful attacks before and after entity substitution, corresponding to $\mathcal{Q}$ and $\mathcal{Q}'$ in Definition 3.2. We calculated the average prompt semantic similarity (PSS) using `SentenceTransformer` (Reimers & Gurevych, 2019) with the RoBERTa-large model (`all-roberta-large-v1`). At each distinct n, the PSS $\in [-1, 1]$ with 1 being the most similar, is averaged over all successful attack instances to obtain the average

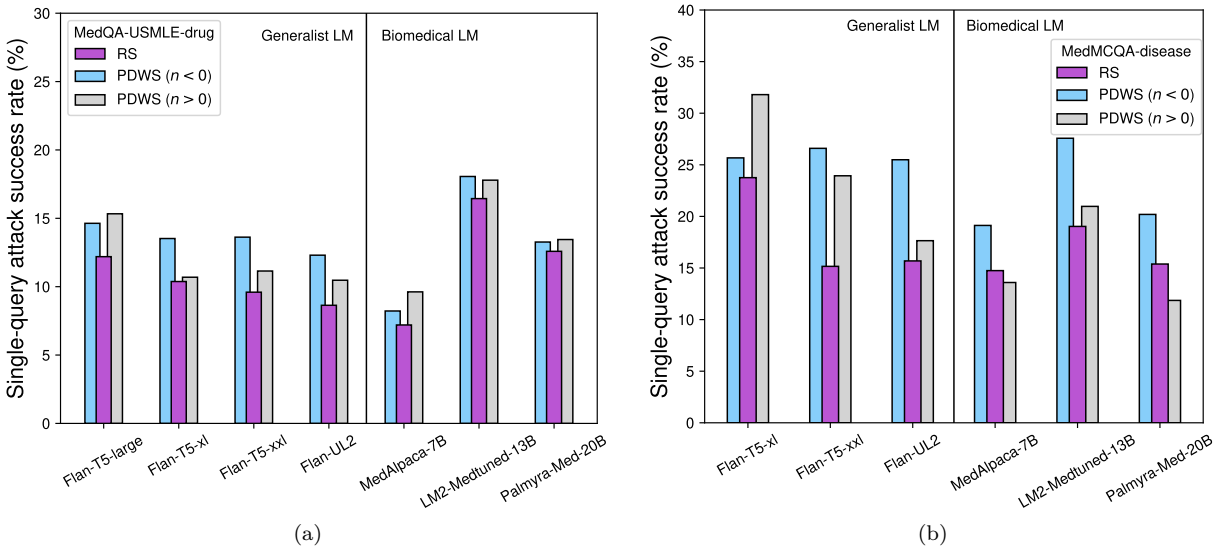

Figure 3: Model robustness from single-query adversarial attacks using results in Table 1. The models (GLMs and BLMs) were evaluated on drug-mention questions from MedQA-USMLE (left) and disease-mention questions from MedMCQA (right) datasets. The GLMs and BLMs are ordered horizontally by their sizes. The bar colors distinguish between different attack methods. Perturbations by random sampling (RS) are in purple. Perturbations by powerscaled DWS (PDWS) in the $n < 0$ and $n > 0$ regimes are in blue and grey, respectively.

PSS, a quantitative measure of the semantic distortion. Its dependence on the hyperparameter n for different models and datasets are shown in Fig. 2.

Adversarial entity characteristics The two-regime effect indicates that the success of the attacks may come from two sources that dominate at different locations on the same attack surface, which refers to entity substitution in the present work.

- At larger distances, the attack succeeds potentially due to the obscurity of the distractor entities in the question context, which can also profoundly impact the model performance (Li et al., 2023a). This explanation may be further supported by examining the diversity index, η_{GS}, which shows the dominance of a few replacement entities in successful attacks at large distances. We found these highly reusable adversarial entities by ranking their occurrences in successful attacks by powerscaled DWS. For drug names, the top-ranked adversarial entities include n-acetylglucosamine (an amino sugar and anti-inflammatory drug)[8] and technetium Tc 99m exametazime (a radiopharmaceutical and contrast agent)[9]. For disease names, some of the chromosome deletion syndromes and rare disease names are more reusable than others. These entities dominate the sampling results in the $n > 0$ regime for all LLMs, indicating their transferability and universality.
- At small distances, semantic proximity (Mozes et al., 2021) reduces the distinguishability between the key and the distractor, therefore increasing the chance of model failure. Examples include similar drugs or diseases with insufficient details such as using diabetes as a distractor instead of the correct answer diabetes II (refers to type-II diabetes) can result in a successful attack.
- Adversarial distractors result in little semantic distortion at the prompt level (see Fig. 2e-h). The experiemnts on the MedMCQA dataset yields somewhat lower average PSS than MedQA-USMLE because of the longer format of the latter dataset, leading to sightly larger distortions by entity substitution. Adversarial entities obtained in the $n < 0$ regime results in slightly higher average PSS compared to those

[8] n-acetylglucosamine on RxList
[9] technetium Tc 99m exametazime on RxList

in the $n > 0$ regime, while there is also a two-regime effect – the average PSS tends to be the lowest around $n = 0$ and higher when n departs further from 0. The trend is consistent for different models and datasets demonstrated in Fig. 2.

5.2 Scaling characteristics from adversarial entities

Size is not all for scaling model robustness. We compare model robustness using single-query ASR as the metric in Table 1 and Fig. 3. For the Flan-T5 series of GLMs (Chung et al., 2024) and Flan-UL2 (Tay et al., 2022), the single-query ASR is more pronounced in smaller and less performant models, indicating the improvement of robustness as the model's size grows. However, an opposite trend of robustness is seen in BLMs. It is worth noting that although Palmyra-Med-20B (Kamble & Alshikh, 2023) has about three times the number of parameters than MedAlpaca-7B (Han et al., 2023), the larger model is noticeably more sensitive to entity perturbation. In terms of ASR, the BLMs are comparable to the GLMs in each type of questions evaluated on. Another observation is that the ASR for models is asymmetrical under PDWS attacks. For most of the models, sampling adversarial entities with inverse-distance-scaled weights ($n < 0$ in Eq. 2) induces more performance drop than sampling with distance-scaled weights ($n > 0$ in Eq. 2), yet prominent counterexamples exist. The attack performance is further investigated in the sequential, multi-query setting (budget = 8) in Table 2, where we also compared sampling- against search-based DiscreteZOO attacks. Here, PDWS attacks in the $n < 0$ regime are still largely the most effective.

Model (size)	Entity mention	MedQA-USMLE (ASR %)				MedMCQA (ASR %)			
		RandS @8	PDWS@8 (-/+) +CODER	DZOO@8 +CODER	DZOO@8 +GTE-base	RandS @8	PDWS@8 (-/+) +CODER	DZOO@8 +CODER	DZOO@8 +GTE-base
Flan-T5-xl (3B)	Drugs	22.0	**29.3** / 23.3	21.7	19.5	32.5	**48.3** / 36.2	34.3	28.2
	Diseases	27.2	**38.8** / 24.8	20.9	23.2	47.1	**61.7** / 49.0	45.8	39.3
Flan-UL2 (20B)	Drugs	22.5	**27.5** / 22.3	20.4	19.2	26.4	**35.6** / 22.4	30.7	26.4
	Diseases	25.8	**34.9** / 26.8	25.7	24.0	39.2	**49.5** / 31.6	41.3	37.0
MedAlpaca-7B	Drugs	16.8	21.1 / 22.3	**22.8**	20.5	27.9	37.1 / 30.1	**42.6**	36.8
	Diseases	21.7	**30.3** / 24.0	26.5	25.2	29.8	**43.9** / 30.8	40.5	38.1
Palmyra-Med-20B	Drugs	30.1	**31.4** / 31.3	29.1	29.6	28.5	**37.3** / 31.9	35.4	35.0
	Diseases	33.2	**35.1** / 30.2	32.3	34.5	41.1	48.4 / 39.6	46.3	**50.1**

Table 2: Multi-query (here budget = 8) attack success rates (ASR %) on various LLMs by perturbing the drug and disease entity mentions in MedQA-USMLE and MedMCQA datasets. The PDWS attack uses the embedding from CODER and DiscreteZOO (DZOO) attacks use either CODER or GTE-base embedding. For each model, the highest ASR for a type of entity-mention question in each dataset is bolded.

Sampling attacks has improved query scaling over blackbox gradient attacks in specialized text embedding. A query scaling study of adversarial attacks investigates attack success with varying query budgets per input instance (Shukla et al., 2021). It is a comprehensive evaluation of attack methods. Fig. 4 compares the query scaling of both sampling- and search-based methods for TCES. For low-query-budget attacks (Fig. 1a), each scaling curve exhibits a rapid rise phase and a plateau phase, separated by an inflection point. The key message here is that the advantage of PDWS attacks over DiscreteZOO depends on the model and the choice of text embedding. The PDWS attacks are overarchingly more effective for GLMs than DiscreteZOO, while for BLMs, the advantage is reduced. Moreover, using a specialized text embedding further boosts the advantage of PDWS attacks compared with a general text embedding due to the difference in embedding space neighborhood. We expand the analysis in the following in three directions.

- **Query scaling depends on text embedding**: Changing the text embedding affects the query scaling behavior of both PDWS and DiscreteZOO attacks. Fig. 4 shows that the influence of the embedding choice on PDWS is more pronounced in the rapid rise phase, while it influences the DiscreteZOO primarily in the plateau phase. Specifically, changing the text embedding from domain-specialized CODER (Yuan et al., 2022) to the general GTE-base (Li et al., 2023b), the performance advantage of PDWS over DiscreteZOO attacks shrunk in Flan-T5-xl (Fig. 4a) and inverted in MedAlpaca-7B (Fig. 4b).
- **Sampling hyperparameter controls scaling curve shape in PDWS attacks**: The hyperparameter n changes the inflection point of the scaling curve. When the query budget passes the inflection point of the

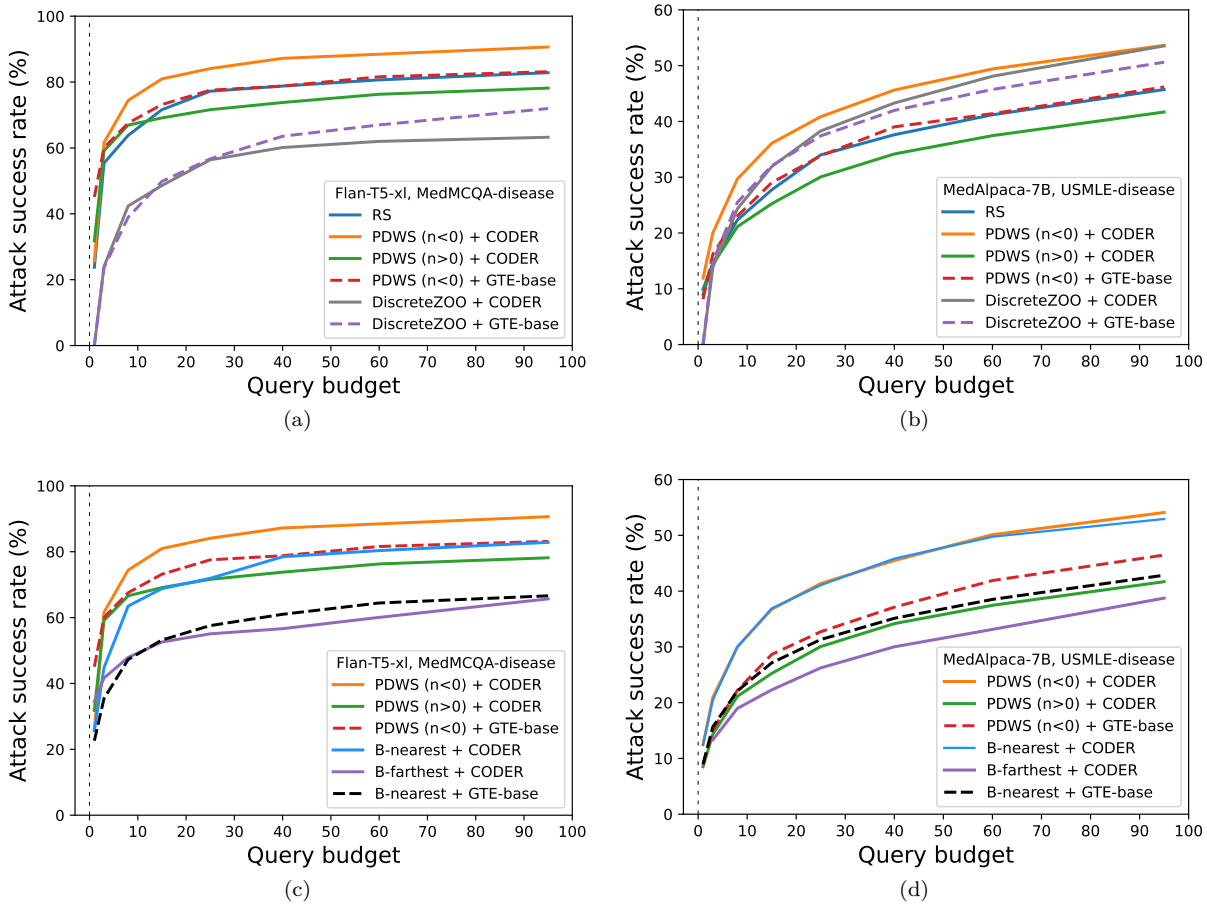

Figure 4: Example scaling curves of the query budget against ASR for disease-mention questions with (a) Flan-T5-xl model on MedMCQA dataset and (b) MedAlpaca model on MedQA-USMLE dataset. The curves are generated using TCES of disease names with the methods described in the legends. The DiscreteZOO attacks were run in the low-query-budget setting. Executing the random sampling (RS) attack doesn't require a text embedding, while the other attacks were evaluated with the CODER or GTE-base as text embedding. In (c) and (d), query scaling of the attacks based on B-nearest and B-farthest element sampling (B is query budget) for the same datasets and models as in (a) and (b) are compared with PDWS.

curve, random sampling starts to surpass PDWS ($n > 0$) in ASR. This is a consequence of the distribution of adversarial entities in the embedding space and the regional bias derived from the sampling weights. The examples in Fig. 4 show that the largest performance gap between the PDWS and DiscreteZOO attacks appears at around the inflection point of the respective query scaling curve.

- **Gradient quality affects query scaling in DiscreteZOO attacks**: The query complexity in blackbox gradient estimation is a long-standing theoretical topic in derivative-free optimization (L'Ecuyer & Yin, 1998) and online learning (Duchi et al., 2015). Nevertheless, how gradient quality affects query scaling remains less studied for adversarial attacks employing gradient-based discrete optimization. Gradient estimation consumes a large part of the total query budget, leading to a tradeoff between the gradient quality and terminal ASR attainable (at max budget) in the query scaling curve. We empirically explored the query scaling behavior of the DiscreteZOO attack in Appendix E with different candidate points in the multi-point gradient estimation, which in effect alters the gradient quality. The results show that having too few candidate points (i.e. small M in Eq. 5) leads to an early onset of the plateau phase, thereby limiting the highest attainable ASR, whereas having many candidate points can lead to a slower growth of ASR in the rapid rise phase of the query scaling curve.

A 73-year-old man presents to the outpatient clinic complaining of chest pain with exertion. He states that resting for a few minutes usually resolves the chest pain. Currently, he takes 81 mg of aspirin daily. He has a blood pressure of 127/85 mm Hg and heart rate of 75/min. Physical examination reveals regular heart sounds and clear lung sounds bilateral. Which medication regimen below should be added?

A: Amlodipine daily. Sublingual nitroglycerin as needed. B: Metoprolol and a statin daily. Sublingual nitroglycerin as needed. ✓ C: Metoprolol and ranolazine daily. Sublingual nitroglycerin as needed. D: Amlodipine and a statin daily. Sublingual nitroglycerin as needed.	A: Amlodipine daily. Sublingual nitroglycerin as needed. B: Metoprolol and a statin daily. Sublingual nitroglycerin as needed. C: N-acetylglucosamine and ranolazine daily. Sublingual nitroglycerin as needed. D: Amlodipine and a statin daily. Sublingual nitroglycerin as needed. ✓

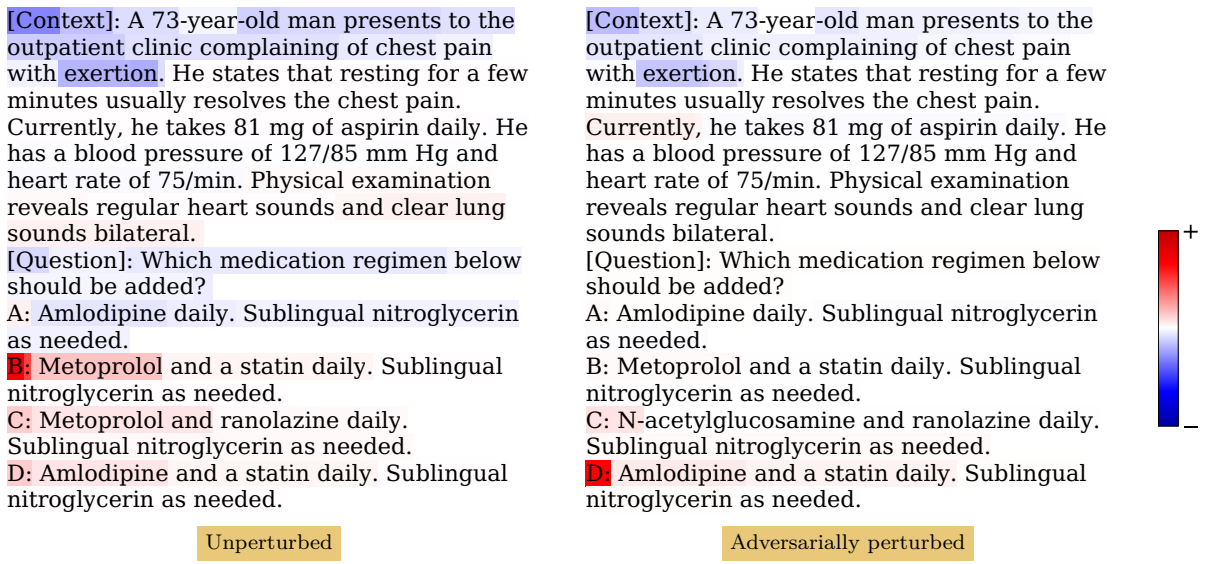

Figure 5: (Top) An entity substitution attack using n-acetylglucosamine as the replacement entity for metoprolol creates an adversarial distractor. (Bottom) Heatmaps of token-wise Shapley values for a question before and after the adversarial attack on choice C. The model prediction changes from the correct choice of B (unperturbed) to the incorrect choice of D (adversarially perturbed).

- **Deterministic sampling attacks are competitive proxies of their probabilistic counterparts.** Performance comparison between B-nearest and B-farthest element sampling, the deterministic and query-limited versions of PDWS (corresponding to the $n < 0$ and $n > 0$ regimes) are given in Fig. 4c-d for the two example cases studied for the query scaling. The results show that the deterministic sampling attacks are already competitive against DiscreteZOO, and also close in performance to PDWS in both specialized and general text embeddings. The difference is that the two types of deterministic sampling attacks examine only predefined regions in the embedding space and cannot explore the entire attack surface as PDWS would through tuning the n hyperparameter.

5.3 Adversarial entities manipulate explainability

Post-hoc analysis is a common way to understand the rationale behind model decisions (Murdoch et al., 2019). Score-based feature importance such as that based on the Shapley values is routinely used to construct post-hoc explanation of model prediction through feature attribution (Chen et al., 2023). The stability of model explanation is a growing concern for their proper use (Alvarez-Melis & Jaakkola, 2018; Hancox-Li, 2020; Lin et al., 2023). Investigations on the susceptibility of LLM explanations to adversarial perturbation in QA are still lacking. For score-based feature importance, an explainer $\phi(\cdot)$ maps the m-dimensional

feature $\mathbf{x} \in \mathbb{R}^m$ to scores $\phi(\mathbf{x}) \in \mathbb{R}^m$. In the current context, we assess the changes in feature importance represented by Shapley values (Chen et al., 2023) before ($\phi(\mathcal{Q})$) and after ($\phi(\mathcal{Q}')$) the adversarial attack by entity substitution $\mathcal{Q} \to \mathcal{Q}'$. For convenience of discussion, we use *Shapley profile* to refer to $\phi(\mathcal{Q})$ (or $\phi(\mathcal{Q}')$), the token-wise Shapley values calculated over $\mathcal{Q}$ (or $\mathcal{Q}'$).

We investigated the changes in the Shapley profile in the two regimes of adversarial entities for open-source LLMs, because the calculation requires access to model weights. We found that the token with the largest Shapley value (top-1 feature) is highly correlated with the model's prediction before and after adversarial perturbation, with an example shown in Fig. 5 produced using Flan-T5-large. This behavior is an indicator of the symbol-binding characteristic of LLMs in multiple-choice QA (Robinson & Wingate, 2023). Moreover, the Shapley profile typically exhibits pronounced changes before and after entity perturbation, especially in the text locations close to the perturbed entities (Fig. 5 and Appendix E). Together, these observations indicate that the substitution attacks effectively manipulate model explanations, because the explanation after the attack aligns with the wrong prediction, as if they were correct. The behavior occurs despite unchanged context, stem, and key ($\mathcal{C}$, $\mathcal{S}$, and k), which should contain the most important information for understanding and answering the question (Chai & Jin, 2004).

6 Discussion and future work

The present study illustrates that TCES is a compact yet effective way to construct adversarial distractors, which leads to significant performance degradation in LLMs. Our results suggest that to minimize model queries, using a budget of up to the inflection point on the query scaling curve is the most cost-effective. The perturbed texts produced by TCES appear natural and are less likely to be detected using grammar check or topical filtering (Morris et al., 2020a). Therefore, the setting can simulate realistic scenarios where the "attacks" may be initiated by unsuspecting users, such as healthcare professionals using an LLM-based clinical decision support system (Liu et al., 2023). In the examples illustrated in Fig. 5 and Appendix E, adversarial distractors can lead to disease misdiagnosis or misprescription of medication, which are detrimental scenarios that may occur in algorithmic decision-making in biomedicine using LLMs. Further improvement on the sampling attack performance may be achieved using learned samplers through rejection mechanisms (Narasimhan et al., 2024) or by finetuning the text embedding to improve the query efficiency of sampling. Alternatively, the semantic distance used for PDWS may be replaced with a concept distance (Choi et al., 2016) to improve attack performance on general text embeddings. Our approach may be integrated into interactive platforms for adversarial data collection or online monitoring systems for open-ended human-machine conversation (Chao et al., 2023).

Future research may focus on how to make LLMs more reliable through prompt engineering (Si et al., 2023; Nori et al., 2023) or leveraging retrieval from nonparametric knowledge bases (Weikum et al., 2021; Soman et al., 2023) or text sources (Wang et al., 2023b) to improve model generalization to long-tail entities. Besides vulnerability identification, the current work motivates adversarial defense strategies based on generalized distances (La Malfa et al., 2020) to establish a tiered system for robustness assessment in user-facing and domain-oriented applications to mitigate catastrophic failures.

Acknowledgments

R.P.X. thanks N. Rethmeier for helpful technical discussions. A.J.L. acknowledges funding from NIH T32GM067547, the UCSF Discovery Fellows Program, and the Achievement Rewards for College Scientists (ARCS) Fellowship. R.A.-A. would like to acknowledge funding from Weill Neurohub.

References

David Alvarez-Melis and Tommi S. Jaakkola. On the Robustness of Interpretability Methods, June 2018. URL http://arxiv.org/abs/1806.08049. arXiv:1806.08049 [cs, stat].

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Em-*

pirical Methods in Natural Language Processing, pp. 2890–2896, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1316. URL http://aclweb.org/anthology/D18-1316.

Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 339–364, February 2023. doi: 10.1109/SaTML54575.2023.00031. URL https://ieeexplore.ieee.org/document/10136152.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, August 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00881-z. URL https://www.nature.com/articles/s42256-024-00881-z. Publisher: Nature Publishing Group.

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. Exploring The Landscape of Distributional Robustness for Question Answering Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5971–5987, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.441. URL https://aclanthology.org/2022.findings-emnlp.441.

Nathaniel Berger, Stefan Riezler, Sebastian Ebert, and Artem Sokolov. Don't Search for a Search Method — Simple Heuristics Suffice for Adversarial Text Attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8216–8224, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.647. URL https://aclanthology.org/2021.emnlp-main.647.

O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, jan 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh061. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh061.

David Cecchini, Arshaan Nazir, Kalyan Chakravarthy, and Veysel Kocaman. Holistic Evaluation of Large Language Models: Assessing Robustness, Accuracy, and Toxicity for Real-World Applications. In Anaelia Ovalle, Kai-Wei Chang, Yang Trista Cao, Ninareh Mehrabi, Jieyu Zhao, Aram Galstyan, Jwala Dhamala, Anoop Kumar, and Rahul Gupta (eds.), *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pp. 109–117, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.trustnlp-1.11. URL https://aclanthology.org/2024.trustnlp-1.11.

Dhawaleswar Rao CH and Sujan Kumar Saha. Automatic Multiple Choice Question Generation From Text: A Survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25, jan 2020. ISSN 1939-1382. doi: 10.1109/TLT.2018.2889100. URL https://ieeexplore.ieee.org/document/8585151/.

Joyce Y. Chai and Rong Jin. Discourse Structure for Context Question Answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, pp. 23–30, Boston, Massachusetts, USA, May 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-2504.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking Black Box Large Language Models in Twenty Queries. *arXiv*, pp. 2310.08419, oct 2023. URL http://arxiv.org/abs/2310.08419.

Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, may 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00657-x. URL https://www.nature.com/articles/s42256-023-00657-x.

Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. September 2019. URL https://openreview.net/forum?id=rJlk6iRqKX.

Youngduck Choi, Chill Yi-I. Chiu, and David Sontag. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2016:41–50, 2016. ISSN 2153-4063.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70): 1–53, 2024. ISSN 1533-7928. URL http://jmlr.org/papers/v25/23-0870.html.

Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner, and Jakob Nikolas Kather. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141, October 2023. ISSN 2730-664X. doi: 10.1038/s43856-023-00370-1. URL https://www.nature.com/articles/s43856-023-00370-1.

A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Research*, 37(Database):D786–D792, jan 2009. ISSN 0305-1048. doi: 10.1093/nar/gkn580. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn580.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit Inference Scaling Laws. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 7750–7774. PMLR, July 2023. URL https://proceedings.mlr.press/v202/dettmers23a.html.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, May 2015. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2015.2409256. URL http://ieeexplore.ieee.org/document/7055287/.

Salijona Dyrmishi, Salah Ghamizi, and Maxime Cordy. How do humans perceive adversarial text? A reality check on the validity and naturalness of word-based adversarial attacks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8822–8836, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.491. URL https://aclanthology.org/2023.acl-long.491.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, dec 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00410. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00410/107384/Measuring-and-Improving-Consistency-in-Pretrained.

Wee Chung Gan and Hwee Tou Ng. Improving the Robustness of Question Answering Systems to Question Paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6065–6075, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1610. URL https://www.aclweb.org/anthology/P19-1610.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56. IEEE, may 2018. ISBN 978-1-5386-8276-0. doi: 10.1109/SPW.2018.00016. URL https://ieeexplore.ieee.org/document/8424632/.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. Question answering is a format; when is it useful?, September 2019. URL http://arxiv.org/abs/1909.11291. arXiv:1909.11291 [cs].

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data, October 2023. URL http://arxiv.org/abs/2304.08247. arXiv:2304.08247 [cs].

Leif Hancox-Li. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 640–647, New York, NY, USA, January 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372836. URL https://dl.acm.org/doi/10.1145/3351095.3372836.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2137–2146. PMLR, July 2018. URL https://proceedings.mlr.press/v80/ilyas18a.html.

Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL http://aclweb.org/anthology/D17-1215.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421, jul 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Kiran Kamble and Waseem Alshikh. Palmyra-med: Instruction-based fine-tuning of llms enhancing medical domain performance, July 2023.

Aneta Koleva, Martin Ringsquandl, and Volker Tresp. Adversarial Attacks on Tables with Entity Swap. In *Joint Proceedings of Workshops at the 49th International Conference on Very Large Data Bases (VLDB 2023), Vancouver, Canada*, volume 3462 of *CEUR Workshop Proceedings*, 2023. URL https://ceur-ws.org/Vol-3462/TADA4.pdf.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, feb 2023. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000198. URL https://dx.plos.org/10.1371/journal.pdig.0000198.

Emanuele La Malfa, Min Wu, Luca Laurenti, Benjie Wang, Anthony Hartshorn, and Marta Kwiatkowska. Assessing Robustness of Text Classification through Maximal Safe Radius Computation. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2949–2968, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.266. URL https://aclanthology.org/2020.findings-emnlp.266.

Pierre L'Ecuyer and George Yin. Budget-Dependent Convergence Rate of Stochastic Approximation. *SIAM Journal on Optimization*, 8(1):217–247, 1998. doi: 10.1137/S1052623495270723. URL https://epubs.siam.org/doi/10.1137/S1052623495270723. Publisher: Society for Industrial and Applied Mathematics.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do We Still Need Clinical Language Models? In Bobak J Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C Ho (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pp. 578–597. PMLR, 2023. URL https://proceedings.mlr.press/v209/eric23a.html.

Xinzhe Li, Ming Liu, Shang Gao, and Wray Buntine. A Survey on Out-of-Distribution Evaluation of Neural NLP Models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6683–6691, California, aug 2023a. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/749. URL https://www.ijcai.org/proceedings/2023/749.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards General Text Embeddings with Multi-stage Contrastive Learning, August 2023b. URL http://arxiv.org/abs/2308.03281. arXiv:2308.03281 [cs].

Chris Lin, Ian Covert, and Su-In Lee. On the Robustness of Removal-Based Feature Attributions. *Advances in Neural Information Processing Systems*, 36:79613–79666, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fbbda4e85a6641bf425be3a6cfd84d20-Abstract-Conference.html.

Aiwei Liu, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7664–7676, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.522. URL https://aclanthology.org/2022.emnlp-main.522.

Siru Liu, Aileen P Wright, Barron L Patterson, Jonathan P Wanderer, Robert W Turer, Scott D Nelson, Allison B McCoy, Dean F Sittig, and Adam Wright. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 30(7): 1237–1245, July 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad072. URL https://doi.org/10.1093/jamia/ocad072.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3):100943, March 2024. ISSN 2666-3899. doi: 10.1016/j.patter.2024.100943. URL https://www.sciencedirect.com/science/article/pii/S2666389924000424.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565.

Rishabh Maheshwary, Saket Maheshwary, and Vikram Pudi. A Strong Baseline for Query Efficient Attacks in a Black Box Setting. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8396–8409, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.661. URL https://aclanthology.org/2021.emnlp-main.661.

Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021. ISSN 1551-3955. doi: 10.1561/1100000083. URL http://www.nowpublishers.com/article/Details/HCI-083.

Sunil Mohan and Donghui Li. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Automated Knowledge Base Construction (AKBC)*, 2019. URL https://openreview.net/forum?id=SylxCx5pTQ.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 86–94. IEEE, jul 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.17. URL http://ieeexplore.ieee.org/document/8099500/.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. Reevaluating Adversarial Examples in Natural Language. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3829–3839, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.341. URL https://aclanthology.org/2020.findings-emnlp.341.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, Online, October 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.16. URL https://aclanthology.org/2020.emnlp-demos.16.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 171–186, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.13. URL https://aclanthology.org/2021.eacl-main.13.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–148, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1018. URL http://aclweb.org/anthology/N16-1018.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116 (44):22071–22080, oct 2019. ISSN 0027-8424. doi: 10.1073/pnas.1900654116. URL https://pnas.org/doi/full/10.1073/pnas.1900654116.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Neha Gupta, and Sanjiv Kumar. Learning to Reject Meets Long-tail Learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ta26LtNq2r.

Yurii Nesterov and Vladimir Spokoiny. Random Gradient-Free Minimization of Convex Functions. *Foundations of Computational Mathematics*, 17(2):527–566, April 2017. ISSN 1615-3383. doi: 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL https://www.aclweb.org/anthology/W19-5034.

Maximilian Noppel and Christian Wressnegger. SoK: Explainable Machine Learning in Adversarial Environments. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 21. IEEE Computer Society, oct 2024. ISBN 9798350331301. doi: 10.1109/SP54263.2024.00021. URL https://www.computer.org/csdl/proceedings-article/sp/2024/313000a021/1RjE9XVNjnW.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv*, pp. 2311.16452, nov 2023. URL http://arxiv.org/abs/2311.16452.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv*, pp. 1605.07277, may 2016. URL http://arxiv.org/abs/1605.07277.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250.

Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. *arXiv*, pp. 2308.11483, aug 2023. URL http://arxiv.org/abs/2308.11483.

C.Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, feb 1982. ISSN 00405809. doi: 10.1016/0040-5809(82)90004-1. URL https://linkinghub.elsevier.com/retrieve/pii/0040580982900041.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1085–1097, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1103. URL https://www.aclweb.org/anthology/P19-1103.

Kyle Richardson and Ashish Sabharwal. What Does My QA Model Know? Devising Controlled Probes Using Expert Knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588, dec 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00331. URL https://direct.mit.edu/tacl/article/96481.

Joshua Robinson and David Wingate. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B.

Omid Rohanian, Mohammadmahdi Nouriborji, and David A. Clifton. Exploring the Effectiveness of Instruction Tuning in Biomedical Language Processing, December 2024. URL http://arxiv.org/abs/2401.00579. arXiv:2401.00579 [cs].

Tom Roth, Yansong Gao, Alsharif Abuadbba, Surya Nepal, and Wei Liu. Token-modification adversarial attacks for natural language processing: A survey. *AI Communications*, Pre-press:1–22, January 2024. ISSN 0921-7126. doi: 10.3233/AIC-230279. URL https://content.iospress.com/articles/ai-communications/aic230279. Publisher: IOS Press.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of Gemini Models in Medicine, May 2024. URL http://arxiv.org/abs/2404.18416. arXiv:2404.18416 [cs].

Priyanka Sen and Amir Saffari. What do Models Learn from Question Answering Datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2429–2438, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.190. URL https://www.aclweb.org/anthology/2020.emnlp-main.190.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks, October 2023. URL http://arxiv.org/abs/2310.10844. arXiv:2310.10844 [cs].

Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. Simple and Efficient Hard Label Black-box Adversarial Attacks in Low Query Budget Regimes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 1461–1469, New York, NY, USA, August 2021. Association for Computing Machinery. ISBN 978-1-4503-8332-5. doi: 10.1145/3447548.3467386. URL https://doi.org/10.1145/3447548.3467386.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. Prompting GPT-3 To Be Reliable. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=98p5x51L5af.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, aug 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06291-2. URL http://arxiv.org/abs/2212.13138https://www.nature.com/articles/s41586-023-06291-2.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A Nelson, Sui Huang, and Sergio E Baranzini. Biomedical knowledge graph-enhanced prompt generation for large language models. *arXiv*, pp. 2311.17330, nov 2023. URL http://arxiv.org/abs/2311.17330.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26), jun 2023. ISSN 2375-2548. doi: 10.1126/sciadv.adh1850. URL https://www.science.org/doi/10.1126/sciadv.adh1850.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. Can Language Models be Biomedical Knowledge Bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4723–4734, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.388. URL https://aclanthology.org/2021.emnlp-main.388.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=6ruVLB727MC.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729, apr 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100729. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389923000697.

Oleg Ursu, Jayme Holmes, Jeffrey Knockel, Cristian G. Bologa, Jeremy J. Yang, Stephen L. Mathias, Stuart J. Nelson, and Tudor I. Oprea. DrugCentral: online drug compendium. *Nucleic Acids Research*, 45(D1):D932–D939, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw993. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw993.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, Stroudsburg, PA, USA, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://www.aclweb.org/anthology/D19-1221.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering. *Transactions of the Association for Computational Linguistics*, 7:387–401, nov 2019b. ISSN 2307-387X. doi: 10.1162/tacl_a_00279. URL https://direct.mit.edu/tacl/article/43493.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ACM Computing Surveys*, 56(3):55:1–55:52, October 2023a. ISSN 0360-0300. doi: 10.1145/3611651. URL https://doi.org/10.1145/3611651.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20, November 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2018.09.008. URL https://www.sciencedirect.com/science/article/pii/S1532046418301825.

Yubo Wang, Xueguang Ma, and Wenhu Chen. Augmenting Black-box LLMs with Medical Textbooks for Clinical Question Answering, September 2023b. URL http://arxiv.org/abs/2309.02233. arXiv:2309.02233 [cs].

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593, jul 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz389. URL https://academic.oup.com/nar/article/47/W1/W587/5494727.

Gerhard Weikum, Xin Luna Dong, Simon Razniewski, and Fabian Suchanek. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Foundations and Trends® in Databases*, 10(2-4):108–490, 2021. ISSN 1931-7883. doi: 10.1561/1900000064. URL http://www.nowpublishers.com/article/Details/DBS-064.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, jul 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00879-8. URL https://www.nature.com/articles/s41746-023-00879-8.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2859–2867, October 2017. doi: 10.1109/ICCV.2017.309. URL https://ieeexplore.ieee.org/document/8237571.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. On the Robustness of Reading Comprehension Models to Entity Renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 508–520, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.37. URL https://aclanthology.org/2022.naacl-main.37.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan. Greedy Attack and Gumbel Attack: Generating Adversarial Examples for Discrete Data. *Journal of Machine Learning Research*, 21(43):1–36, 2020. URL http://jmlr.org/papers/v21/19-569.html.

Zhen Yu, Zhenhua Chen, and Kun He. Query-Efficient Textual Adversarial Example Generation for Black-Box Attacks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 556–569, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.31. URL https://aclanthology.org/2024.naacl-long.31.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126: 103983, feb 2022. ISSN 15320464. doi: 10.1016/j.jbi.2021.103983. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046421003129.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial Attacks on Deep-learning Models in Natural Language Processing. *ACM Transactions on Intelligent Systems and Technology*, 11 (3):1–41, jun 2020. ISSN 2157-6904. doi: 10.1145/3374217. URL https://dl.acm.org/doi/10.1145/3374217.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models. *arXiv*, pp. 2303.18223, mar 2023. URL http://arxiv.org/abs/2303.18223.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, December 2023. URL http://arxiv.org/abs/2307.15043. arXiv:2307.15043 [cs].