arXiv:2305.01860v1 [cs.IR] 3 May 2023

面向神经排序模型的不可察觉文档操纵

Xuanang Chen[1,2] **Ben He**[1,2] **Zheng Ye**[3] **LeSun**[2] **Yingfei Sun**[11]中国科学院大学，中国北京[2]中国科学院软件研究所，中国北京[3]中南民族大学，中国武汉 chenxuanang19@mails.ucas.ac.cn, benhe@ucas.ac.cn yezheng@scuec.edu.cn, s unle@iscas.ac.cn, yfsun@ucas.ac.cn

摘要

对抗攻击已被广泛用于识别神经排序模型（NRM）中的潜在漏洞，但当前的攻击方法往往会引入语法错误、无意义的表达或不连贯的文本片段，因此很容易被检测出来。此外，当前方法严重依赖使用拟合良好的替代神经排序模型来保证攻击效果，这使得它们在实践中难以应用。为解决这些问题，我们提出了一种名为不可察觉文档操纵（IDEM）的框架，用于生成对算法和人类而言都不易察觉的对抗文档。IDEM 指示一个成熟的生成式语言模型（例如 BART）生成连接句，避免引入易于检测的错误，并采用独立的逐位置合并策略，以平衡扰动文本的相关性与连贯性。在广受欢迎的 MS MARCO 基准测试上的实验结果表明，IDEM 优于强有力的基线方法，同时能够保留目标文档的流畅性和正确性，这一点得到了自动评估和人工评估的证实。此外，将对抗文本生成与替代神经排序模型分离，使 IDEM 更加稳健，也较少受到替代神经排序模型质量的影响。

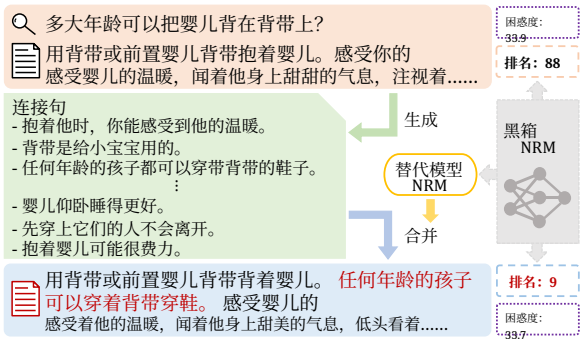

图 1：在 IDEM 中，指示生成式语言模型在查询与目标文档之间生成一组连接句，然后应用逐位置合并策略来选择并确定一条最优连接句在目标文档中的位置，以提升排名（从第 88 名升至第 9 名），同时保持由困惑度（PPL）衡量的流畅性。

1 引言

对抗性信息检索（AIR）一直是研究界广泛关注的话题（Davison 等人，2006年；Leng 等人，2012；Farooq，2019年）。它指的是对集合中的部分内容进行恶意操纵的情形，对抗性网页搜索（Castillo 和 Davison，2010年）就是一个典型例子。在网络环境中，这类操纵通常称为黑帽搜索引擎优化（SEO）或网页垃圾信息，其目的是通过人为抬高目标网页的相关性，从而使这些网页获得不应有的高排名（Gyöngyi 和 Garcia-Molina，2005年）。与此同时，近年来，神经排序模型（NRM），特别是采用预训练语言模型（PLM）的模型，在各类文本排序任务中展现出卓越的性能（Lin 等人，2021年）。此外，这些神经排序模型也已应用于各种工业应用（Lin 等人，2021年），例如网页搜索引擎（Zou 等人，2021年），以提升搜索结果的准确性和相关性。

然而，近期研究揭示了神经排序模型容易受到对抗文档操纵的影响（Wu 等，2022年；Liu 等人，2022年；Wang 等人，2022年；Song 等人，2022年），也就是说，输入文档中的微小蓄意扰动就可能导致神经排序模型的输出出现灾难性的排序紊乱。这凸显了在部署神经排序模型之前调查并识别其潜在弱点的必要性，以确保其稳健性并防范潜在风险。目前已有若干操纵

技术被提出，用于恶意提升低排名文档的排名，例如 PRADA（Wu 等，2022年）和 PAT（Liu 等人，2022年）。尽管这些对抗攻击方法通过替换关键词，或附加查询文本及其他对抗词元，已展现出欺骗神经排序模型的能力，但仍有两大主要局限。首先，现有攻击方法往往会在原始文档中引入语法错误、不可能成立的表达或不连贯的文本片段，使攻击容易被困惑度过滤器（Song 等人，2020年）或语法检查器（Liu 等人，2022年）等方法缓解。其次，现有攻击方法严重依赖一个模仿得很好的替代神经排序模型来生成对抗文本，但这需要通过查询受害神经排序模型收集大量领域内训练数据，在现实情境中可能难以实现，甚至无法获得。

为此，我们提出 IDEM，这是一种难以察觉的文档操纵框架，旨在生成对人类和算法而言都不易察觉的对抗文档。如图1所示，首先对成熟的生成式语言模型（GLM）（如 BART）进行设计，使其在查询与目标文档之间生成一系列语法正确的连接句；随后采用逐位置合并机制，选出最优连接句，并将其恰当地插入目标文档。在生成连接句时，我们利用 GLM 的语言建模目标，而不添加任何排名激励目标。这不仅有助于生成更自然、流畅的文本，避免引入容易被检测到的额外错误，还将替代神经排序模型与生成过程分离，从而大幅降低攻击性能对替代神经排序模型的依赖。在广泛使用的 MS MARCO 段落排序数据集上开展的大量实验表明，无论替代神经排序模型与受害神经排序模型相似还是相差甚远，IDEM 对黑箱神经排序模型的攻击性能都优于近期的基线方法。自动评估和人工评估结果均表明，IDEM 生成的对抗文档能够保持语义流畅性和文本质量。

我们的贡献有三点：1）我们提出了一种不可察觉文档操纵框架 IDEM[1]，，该框架采用上下文化空白—用于排序攻击的文本填充和连贯合并。2）在三种类型的替代神经排序模型下进行的大量攻击实验表明，IDEM 能够稳健地提升目标文档的排名。3）文本质量的自动评估和人工评估表明，IDEM 能够生成更加自然、流畅的对抗文档。

2 问题陈述

任务描述。通常，给定由轻量级检索模型（例如 BM25；Robertson 等，1995）从整个语料库中收集的查询 q 和候选文档 $\mathcal{D}=\{d_1,d_2,\cdots,d_{|\mathcal{D}|}\}$，NRM $\mathcal{M}_V$（即受害神经排序模型）会为所有查询-文档对生成相关性分数$s(q,d_i)$，输出重排序列表 $\mathcal{L}=[d_1,d_2,\cdots,d_{|\mathcal{D}|}]$，其中$s(q,d_1)>s(q,d_2)>\cdots>s(q,d_{|\mathcal{D}|})$。在对抗性排名攻击中，会向目标文档 $d_i\in\mathcal{L}$ 中加入有意的扰动 p_i，使其成为对抗文档 $d_i^{adv}:=d_i\oplus p_i$，从而能够被受害神经排序模型排在更高位置。现有攻击方法会在不同粒度层级上搜索并寻找扰动，例如用同义词替换重要词语（例如 PRADA），或添加额外的文本片段（例如 PAT）。理想情况下，对抗文档应表现出语义一致性和流畅性，以避免被检测。然而，现有方法生成的对抗样本在不可察觉性方面仍有改进空间，此类示例见附录 A.5。

任务设定。为符合现实场景（例如黑帽搜索引擎优化），与近期研究（Wu 等人，2022；Liu 等人，2022）类似，本文聚焦于基于决策的黑盒攻击设定。在此设定中，攻击者只能通过查询受害神经排序模型 $\mathcal{M}_V$，获取数量有限的查询 $\{q_i\}_{i=1}^m$ 及其对应的带有排名位置的排名列表 $\{\mathcal{L}_i=[d_{i,1},d_{i,2},\cdots,d_{i,|\mathcal{D}|}]\}_{i=1}^m$，但无法访问精确的相关性分数，也无法获知受害神经排序模型的架构、参数、梯度或训练数据。在此设定下，通常会基于 $\{(q_i,\mathcal{L}_i)\}_{i=1}^m$ 收集弱监督训练数据 $\mathcal{T}_S=\{(q_i,d_{i,j},d_{i,k})\}_{i=1}^m$，其中 $1\le j<k\le|\mathcal{D}|$，因此 $d_{i,j}$ 和 $d_{i,k}$ 分别被视为相对正样本文档和负样本文档。使用成对损失（例如合页损失）（Wu 等人，2022）在这些 $\mathcal{T}_S$ 上训练替代神经排序模型 $\mathcal{M}_S$，以模仿受害神经排序模型 $\mathcal{M}_V$。此外，现有方法要求该替代神经排序模型在功能上与受害神经排序模型相似，因为它是

[1]我们将在以下位置公开我们的代码和数据。未来研究的最终版本。

决定了对抗攻击的方向（Liu 等人，2022年）。这意味着现有方法的攻击性能很大程度上取决于替代神经排序模型，但这需要大量域内训练样本（即 m 较大），还需要承担查询受害神经排序模型的相关成本，而我们提出的 IDEM 即使使用域外（OOD）替代神经排序模型，也能实现可靠的攻击效果。

3 方法

如图1和算法 1所示，IDEM 包含两个阶段：第一阶段是获取一系列连接句，以弥合查询与目标文档之间的语义差距。之后，在目标文档的所有句间位置上评估这些连接句，以权衡语义流畅度与攻击相关性，并将最佳连接句作为最终的对抗文档输出。

3.1 连接句生成

为了生成更自然、流畅的连接句，我们选择借助成熟的生成式语言模型（GLM），例如在默认设置中使用的 BART 模型。BART 模型（Lewis 等，2020年）采用文本填充损失函数进行预训练，使其能够以更灵活的方式填补上下文中的空白。因此，本节以 BART 模型为例，说明如何通过空白填充模式设计生成式语言模型，以生成同时包含查询和文档信息的连接句。

具体而言，给定查询 q 和目标文档 $d_i \in \mathcal{L}$，下文将其简写为 d，并按照以下模板，在两者之间插入一个空格后将它们连接起来。

$$q \oplus \text{众所周知} \oplus d \quad ____ \tag{1}$$

其中查询 q 以恰当的标点符号结尾，空格则替换为一个由生成式语言模型定义和使用的特殊标记，例如 `[MASK]`，即 BART 模型中的标记。将查询 q 和文档 d 作为上下文，前缀文本“众所周知”充当提示（Liu 等人，2021年），用于指示 BART 模型生成更多与 q 和 d都相关的信息性文本。提示词的选择会影响最终的攻击效果，相关分析见第4.3节。

随后，BART模型将式1 作为输入，并填入数量可变的词元（即，将不同长度的句子）放入空白位置。在此，我们不在 BART 模型中加入额外的排名激励目标，以生成语法正确且与上下文相契合的连接句。此外，我们采用 Top-k 采样（Fan 等人，2018年）策略，以确保生成句子的多样性。如算法 1所示，我们进行多次采样（即 K），每次选取 M 个候选文档，并最多保存 N 个连接句，记作 $\mathcal{S}_c = \{s_1, s_2, \cdots s_N\}$。我们还将连接句的最大长度限制为 L个词，以控制其对原始文档语义内容的干扰。将 $\mathcal{S}_c$ 中的所有连接句留待下一阶段与原始文档合并 d。

算法 1IDEM

输入：查询 q、目标文档 d、用于受害神经排序模型 $\mathcal{M}_V$的替代神经排序模型 $\mathcal{M}_S$ 参数：目标文档中的句子数量 $|d|$、连接句子的最大数量 N、采样次数 K、样本量 M、连接句子的最大词长 L输出：对抗文档 d^{adv}1：阶段 1：生成连接句 2：令 $\mathcal{S}_c = \{\}$3：对 1 到 K 执行 4：当 $|\mathcal{S}_c| < N$ 时执行 5：以式 1 为输入，从 BART 模型中采样 M 个连接句候选 $\{s_u\}_{u=1}^{M}$ 6：如果 s_u 的长度小于 L，则 7：将 s_u 保留在 $\mathcal{S}_c$中 8：结束 if 9：结束 while 10：结束 for 11：阶段 2：与原始文档合并 12：对 $\mathcal{S}_c$ 中的s_u 执行 13：对于从 0 到 $|d|$ 的位置索引 v，执行 14：按式 2 获取对抗候选 $d^{adv}_{\langle u,v \rangle}$，评估其连贯性分数（式 3）和相关性分数（式 4），并计算加权合并得分（式 5）15：结束 for 16：结束 for 17：找出排名第一的 $d^{adv}_{\langle u,v \rangle}$ 并保存为 d^{adv}18：返回 d^{adv}

3.2 与原始文档合并

由于基于 Transformer 的神经排序模型对文档开头存在位置偏差（Jiang 等人，2021年；Hofstätter 等人，2021年），在开头添加对抗性文本片段通常能够取得显著的攻击效果（Wang 等人，2022年），但这也有一个明显的缺点：攻击很容易被检测到（Liu 等人，2022年）。因此，为了平衡攻击效果与受扰动文档内容的保真度，我们设计了一种逐位置合并策略，将恰当的连接句放置在原始目标文档中的最佳位置。

具体而言，给定一个查询 q、一个包含多个句子的目标文档 $d = s_1, s_2, \cdots, s_{|d|}$ 以及一组连接句 $\mathcal{S}_c$，我们针对 d 和 $\mathcal{S}_c$的各种组合，评估大量候选对抗文档。对于每个连接句 $s_u \in \mathcal{S}_c$，我们将 s_u合并到目标文档 d 中，方法是将其放置在索引位置 v （即从 0 到 $|d|$的整数），从而得到如式2所示形式的对抗候选项。

$$d^{adv}_{\langle u,v\rangle} = \begin{cases} \overline{s}_u \oplus d & v = 0 \\ d_{1\to v} \oplus \overline{s}_u \oplus d_{(v+1)\to|d|} & 0 < v < |d| \\ d \oplus \overline{s}_u & v = |d| \end{cases} \tag{2}$$

其中， $d_{a\to b}$ 表示该文本片段由 s_a 至 s_b的连续句子组成，下标$\langle u, v\rangle$ 表示连接句 s_u 插入到目标文档 d中的索引位置 v 。

随后，我们检查所有候选对抗文档 $d^{adv}_{\langle u,v\rangle}$，并找出其中最佳者作为最终的对抗文档 d^{adv}。有效的对抗样例应当难以被人工评审者察觉，却能误导神经排序模型（Wu 等，2022年）。如果添加的连接句在语义上与目标文档的上下文差异很大，生成的内容就会缺乏连贯性，容易被人察觉。因此，借助预训练 BERT 模型（De- vlin 等，2019年）中的下一句预测（NSP）功能，我们定义了连接句与原始目标文档之间衔接处的连贯性分数。

具体而言，当连接句 s_u置于目标文档 d中间（即， $0 < v < |d|$连接句前后的文本片段都会被输入 NSP 函数，以获得式连贯性中所示的分数。3。

$$\begin{aligned} coh(d^{adv}_{\langle u,v\rangle}) = 0.5 \times [&f_{nsp}(d_{1\to v}, \overline{s}_u \oplus d_{(v+1)\to|d|}) \\ &+ f_{nsp}(d_{1\to v} \oplus \overline{s}_u, d_{(v+1)\to|d|})] \end{aligned} \tag{3}$$

其中， f_{nsp}(A,B) 表示文本 A 后接文本 B 的 NSP 分数。类似地，当 s_u 附加在文档 d 的开头或结尾时（即$v = 0$ 或 $v = |d|$），其连贯性分数分别表示为$f_{nsp}(s_u, d)$ 或 $f_{nsp}(d, s_u)$。

除了语义连贯性之外，影响攻击可行性的另一个重要因素是查询 q 与候选文档 $d^{adv}_{\langle u,v\rangle}$之间的相关性。鉴于第2节所述的黑盒设定，我们需要一个替代神经排序模型$\mathcal{M}_S$ 来估计相关性得分，如公式4所示。

$$rel(q, d^{adv}_{\langle u,v\rangle}) = \mathcal{M}_S(q, d^{adv}_{\langle u,v\rangle}) \tag{4}$$

其中，替代神经排序模型 $\mathcal{M}_S$ 无需与受害神经排序模型在排序上高度一致$\mathcal{M}_V$，只需能够区分哪些连接句包含虚假的相关信息（例如查询关键词）。换言之，替代神经排序模型不参与连接句的生成，因此对抗信息来自 BART 模型，而非替代神经排序模型。因此，IDEM 不会强制要求使用大量域内训练样本训练出一个高度仿真的替代神经排序模型。正如第4.2节稍后将展示的，即使使用仅略优于 BM25 的分布外替代神经排序模型，我们提出的 IDEM 仍能取得相对较高的攻击性能。

<escape>最后，为了权衡语义连贯性和相关性，如公式5所示，我们采用二者的加权和作为每个 $d^{adv}_{\langle u,v\rangle}$ 的合并分数。</escape>

$$\begin{aligned} score_m(q, d^{adv}_{\langle u,v\rangle}) = \alpha \times coh(d^{adv}_{\langle u,v\rangle}) \\ + (1-\alpha) \times rel(q, d^{adv}_{\langle u,v\rangle}) \end{aligned} \tag{5}$$

其中， α 是权重因子，两个分数在相加前均通过最小-最大归一化转换为 $[0, 1]$ 。在所有候选项 $d^{adv}_{\langle u,v\rangle}$中，合并分数最高的项将被选为最终的对抗文档 d^{adv}。

4 实验

4.1 实验设置

数据集。与以往研究类似（Wu 等，2022年；Wang 等，2022年；Liu 等，2022年），我们采用广泛使用的 MS MARCO 段落数据集（Nguyen 等，

2016年）开展实验。该数据集包含约 880 万个段落（本文中视为文档），构成语料库。评估数据包括包含 6,980 个查询的开发集和包含 6,837 个查询的评估集。此外，我们使用经 Kwiatkowski 等，2019年）处理的 Karpukhin 等（2020年）Natural Question（NQ）数据集，作为替代 NRM 的 OOD 训练源。

受害神经排序模型。 类似于 Liu 等人 (2022年)，Sentence-Transformers（Reimers 和 Gurevych，2019年）公开提供的“ms- marco-MiniLM-L-12-v2”模型被用作本文的代表性受害神经排序模型（即 $\mathcal{M}_V$）。该受害神经排序模型采用 MiniLM（Wang 等人，2020年）作为骨干网络，并在 MS MARCO 上以交叉编码器架构进行微调（Nogueira 和 Cho， 2019年）。总体而言，$\mathcal{M}_V$在 MS MARCO 开发集上实现了极高的排序性能，如表 2 所示。

替代神经排序模型。 为了考察排序攻击方法在现实场景中的实用性——在这些场景中，对受害排序系统的访问可能受到限制或不可用——我们使用三种替代神经排序模型进行实验： $\mathcal{MS}_1$ 在全部 6,837 个评估查询上进行训练， $\mathcal{MS}_2$ 在随机抽取的 200 个评估查询上进行训练，而 $\mathcal{MS}_3$ 在 OOD NQ 查询上进行训练。关于替代神经排序模型的更多详细信息请参见附录 A.1。

目标文档。 继 Wu 等 （2022年）之后，我们在随机抽取的 1K 个开发集查询上，针对两种不同攻击难度的目标文档评估攻击方法，即 Easy-5 和 Hard- 5。这些目标文档由受害神经排序模型在 BM25 排名前 1K 的候选文档上生成的重排序结果中采样得到。具体而言，**Hard-5** 是排名最低的五个文档，而 **Easy-5** 是排名位于 $[51, 100]$ 之间的五个文档，其中每 10 个文档中随机选取一个。

IDEM 的详细信息。默认情况下，预训练 BART- Base（Lewis 等，2020年）被指示生成连接句。其他生成式语言模型的影响，包括预训练 T5（Raffel 等，2020年）和微调 GPT-2（Donahue 等，2020年），也在第4.3节中进行了考察。在生成阶段，采样策略中的 k 设置为 50，我们进行 10 次或 50 次采样（每次采样 50 个），最多保存 100 个或 500 个连接句，句子最大长度为 12 个词（即 $M=50$、 $K=10/50$、 $N=100/500$和 $L=12$）。在合并阶段，我们采用预训练 BERT-Base 中的 NSP 函数来评估连贯性，并将 α 公式5 中的值设置为 0.5（0.1），分别对应 Easy-5（Hard-5）目标文档。其影响对 α 攻击结果的影响如附录 A.4.

对比方法。 我们将以下攻击方法与神经排序模型进行比较： **Query+** （Liu 等人，2022年）直接将查询文本添加到目标文档的开头。**PRADA** （Wu 等，2022年）在目标文档中找出重要词语并用同义词替换。 **Brittle-BERT** （Wang 等人，2022年）和 **PAT** （Liu 等人，2022年）在目标文档开头追加少量词元，前者仅考虑面向排序的目标，而后者进一步考虑语义约束和流畅度约束。为进行公平比较，PRADA 最多扰动 20 个词元，Brittle-BERT 和 PAT 最多添加 12 个词元。有关这些攻击基线的更多细节请参见附录 A.2。

指标。对于重新排序结果，我们报告 MS MARCO 开发集上的官方 MRR@10/1K 指标。与 Liu 等人（2022）类似，我们计算前 10 名的重叠度（Inter@10）以及秩偏置重叠（Webber 等，2010）（RBO@1K， p 设为 0.7），以衡量替代模型与受害神经排序模型之间的排序一致性。对于攻击结果，我们按照 Wu 等（2022）的方法，报告成功提升目标文档的百分比（即攻击成功率，ASR）。与 Liu 等人（2022）类似，我们还报告目标文档的平均排名提升幅度（提升），以及目标文档进入前 10 名（%$r \leq 10$）和前 50 名（%$r \leq 50$）的百分比。此外，我们使用预训练的 GPT-2 模型（Radford 等，2019）计算平均困惑度（PPL）；正如 Kann 等（2018）和 Lei 等（2022）所指出的，PPL 值越低，对抗文档的流畅性越好。

4.2 结果

IDEM 展现出在不显著降低语义流畅度的情况下实现相当攻击性能的能力。攻击结果总结于表1中，其中所有数值均为针对 1K 个开发集查询的 5K 个目标文档的平均值。当将查询添加到目标文档的开头（称为 Query+）时，可以观察到排名显著提升，这与预期一致。然而，这种方法会对语义流畅度产生负面影响，导致 Easy-5 （ Hard-5 ）目标文档的困惑度约增加 8 （ 17 ）点。理想情况下，当受攻击的 NRM 可自由访问时，可以获取足量的代理模型训练数据来训练一个

替代模型 NRM	方法	Easy-5 ASR	Easy-5 % $r \leq 10$	Easy-5 % $r \leq 50$	Easy-5 提升	Easy-5 困惑度↓	Hard-5 攻击成功率	Hard-5 % $r \leq 10$	Hard-5 % $r \leq 50$	Hard-5 提升	Hard-5 困惑度↓
-	原始	-	-	-	-	37.3	-	-	-	-	50.5
-	查询+	100.0	86.9	99.2	70.3	45.4	100.0	47.8	78.3	955.1	67.5
$\mathcal{M}_{S_1}$	PRADA	77.9	3.52	46.2	23.2	94.4	68.0	0.02	0.10	65.2	154.4
	Brittle-BERT	98.7	81.3	96.7	67.3	107.9	**100.0**	**61.5**	**85.9**	**965.5**	152.5
	PAT	89.6	30.6	73.8	41.9	50.9	98.0	6.24	20.1	589.1	71.4
	IDEM$_{N=100}$	99.3	77.9	97.0	67.0	**36.2**	99.5	41.4	66.9	875.6	**54.6**
	IDEM$_{N=500}$	**99.7**	**87.4**	**99.0**	**70.3**	36.4	99.8	54.3	79.3	933.0	54.9
$\mathcal{M}_{S_2}$	PRADA	69.6	1.36	35.0	17.4	90.7	66.0	0.00	0.10	49.3	152.1
	Brittle-BERT	81.6	33.2	69.7	36.4	131.3	94.8	8.98	25.4	565.1	179.5
	PAT	61.9	8.46	37.3	12.5	49.3	84.4	0.82	3.40	221.7	66.2
	IDEM$_{N=100}$	96.8	65.3	91.8	60.7	**36.5**	97.6	31.7	57.0	822.0	**54.7**
	IDEM$_{N=500}$	**98.7**	**74.8**	**95.4**	**65.1**	37.0	**99.1**	**39.6**	**67.4**	**890.7**	55.4
$\mathcal{M}_{S_3}$	PRADA	71.5	1.86	37.5	19.1	91.5	71.9	0.00	0.08	73.4	168.7
	Brittle-BERT	90.0	43.4	80.1	46.2	117.7	**99.9**	17.7	47.6	845.2	156.8
	PAT	51.1	2.70	22.9	2.01	46.8	79.0	0.00	0.66	92.9	64.2
	IDEM$_{N=100}$	97.2	57.3	89.8	58.1	**36.9**	99.2	23.0	48.8	805.7	**55.2**
	IDEM$_{N=500}$	**98.8**	**65.3**	**93.8**	**61.9**	37.7	99.8	**29.1**	**57.9**	**866.2**	56.0

表1：针对从 MS MARCO 开发集中抽取的 1K 个查询，在三种替代神经排序模型下，对两类目标文档的攻击结果。困惑度越低越好，其他指标则越高越好。

模型	MRR@10	MRR@1K	Inter@10	RBO@1K
BM25	18.4	19.5	22.8	31.5
$\mathcal{M}_V$	39.5	40.3	-	-
$\mathcal{M}_{S_1}$	37.0	37.8	73.1	66.2
$\mathcal{M}_{S_2}$	23.0	24.2	41.0	31.2
$\mathcal{M}_{S_3}$	21.0	22.2	27.3	37.6

表2：受害神经排序模型和替代神经排序模型在 MS MARCO 开发集（BM25 Top-1K）上的重排序性能。Inter@10 和 RBO@1K 表示相对于受害神经排序模型 $\mathcal{M}_V$ 的排序一致性。

拟合良好的替代模型，如 $\mathcal{M}_{S_1}$ 表2所示。如表1所示，与 $\mathcal{M}_{S_1}$结合使用时，词级同义词替换（即 PRADA）并未带来更优的攻击性能，但在目标文档开头添加对抗性词元（即 Brittle-BERT 和 PAT）则能提升攻击性能。然而，这些以往方法生成的对抗文档，尤其是 PRADA 和 Brittle-BERT 生成的文档，其困惑度（PPL）值过高。相比之下，我们提出的 IDEM 不仅实现了相当的攻击性能，例如在 Easy-5 目标文档上表现更优，而且还减轻了对文本流畅性的不利影响，其对抗文档的 PPL 值更低便是佐证。

即使替代神经排序模型是域外神经排序模型，IDEM 也更具鲁棒性，受其影响要小得多。当能够访问受害神经排序模型时（即 $\mathcal{M}_V$）受到限制，如表 2 所示，替代模型 $\mathcal{M}_{S_2}$ 与受害模型 $\mathcal{M}_V$之间存在明显差异，因为只能使用数量有限的领域内训练样本。当使用 $\mathcal{M}_{S_2}$ 时，所有方法的攻击性能都会大幅下降，尤其是 PAT 和 Brittle-BERT，如表 1 所示。例如，当 $\mathcal{M}_{S_1}$ 被更改为 $\mathcal{M}_{S_2}$时，Brittle-BERT 的攻击性能（% $r \leq 10$）在 Easy-5目标文档上的结果从 81.3 急剧下降至 33.2，在 Hard-5目标文档上则从 61.5 下降至 8.98。此外，Brittle-BERT 生成的对抗文档的语义流畅度也有所降低，体现在其在 Easy-5目标文档上的困惑度从 107.9 上升至 131.3。再者，无法访问受害神经排序模型时，只能使用公开可用的分布外训练样本，从而导致替代神经排序模型的差异更大，例如表 $\mathcal{M}_{S_3}$ 2 中的模型。在这种情况下，如表 1 所示，先前攻击方法的性能也不理想，尤其是 PRADA 和 PAT。然而，同时使用 $\mathcal{M}_{S_2}$ 和 $\mathcal{M}_{S_3}$时，IDEM 在所有基线方法中取得了最佳攻击结果，同时保留了目标文档的语义流畅度。有关更多详情，包括不同攻击方法的攻击效率和对抗示例，请参阅附录 A.3 和 A.5。

提示	Easy-5 攻击成功率 % $r \leq 10$ % $r \leq 50$ 提升 困惑度↓					Hard-5 攻击成功率 % $r \leq 10$ % $r \leq 50$ 提升 困惑度↓				
IDEM$_{N=100}$										
众所周知	99.3	77.9	97.0	67.0	36.2	99.5	41.4	66.9	875.6	54.6
大约是	98.3	60.9	92.7	60.3	39.1	99.5	20.0	45.7	798.9	57.3
我们知道	99.3	74.4	96.7	65.8	36.8	99.4	33.9	59.8	851.1	54.9
事实是	99.6	77.2	97.4	66.9	36.8	99.5	37.2	63.3	870.0	54.9
<escape>IDEM$_{N=500}$</escape>										
众所周知	99.7	87.4	99.0	70.3	36.4	99.8	54.3	79.3	933.0	54.9
大约是	99.6	75.5	97.6	66.7	39.4	99.9	33.5	63.7	892.6	58.4
我们知道	99.8	85.7	99.0	69.8	37.2	99.9	47.4	74.3	920.0	55.8
事实是	99.8	87.0	99.1	70.3	36.7	99.9	50.1	76.5	928.1	55.5

表3：IDEM 使用不同提示生成连接句的攻击结果。使用替代神经排序模型 $\mathcal{M}_{S_1}$。总体而言，“众所周知”产生的攻击结果最佳，因此 IDEM 默认使用该表述。

NRM	索引 % $v=0$ % $v=1$ % $v=2$ % $v=3$ % $v \geq 4$				
Easy-5					
$\mathcal{M}_{S_1}$	83.9	12.2	2.62	0.88	0.48
$\mathcal{M}_{S_2}$	74.1	15.0	4.77	1.56	4.59
$\mathcal{M}_{S_3}$	22.3	29.5	22.3	11.7	14.2
Hard-5					
$\mathcal{M}_{S_1}$	83.2	13.5	2.12	0.76	0.40
$\mathcal{M}_{S_2}$	76.3	14.6	4.32	1.12	3.68
$\mathcal{M}_{S_3}$	9.76	37.7	24.8	12.9	14.9

表4：IDEM$_{N=100}$ 创建的对抗文档中连接句位置的统计数据。

GLM	攻击成功率 % $r \leq 10$ % $r \leq 50$ 提升 困惑度↓				
Easy-5					
BART	99.3	77.9	97.0	67.0	36.2
T5	98.7	75.3	95.4	65.2	38.0
ILM	93.9	45.7	81.5	50.6	41.5
Hard-5					
BART	99.5	41.4	66.9	875.6	54.6
T5	97.8	37.1	60.1	820.9	56.8
ILM	98.3	16.0	35.6	693.2	59.8

表5：基于各种生成式语言模型和替代神经排序模型的 IDEM$_{N=100}$ 攻击结果$\mathcal{M}_{S_1}$。

4.3 分析

IDEM 中提示的影响。在生成连接句的过程中，会使用提示（见公式1）引导 BART 模型生成信息量更丰富的文本。本文考察了提示对 IDEM 最终攻击结果的影响。如表 3 所示，我们分析了四种不同的提示，其中“大约是”会带来更好的攻击结果，尤其是在 Hard-5目标文档上，同时也会生成较低的困惑度值。相比之下，“事实是”和“我们知道”在 Hard-5目标文档上的表现稍差，而“大约是”表现更差。总体而言，提示的选择对攻击成功率影响不大，主要影响攻击提升的幅度。

IDEM 中连接句的位置。如第 3.2 节所述，IDEM 可以自动将对抗连接句放置在目标文档中。如表 4 所示，我们统计了连接句的位置（即索引 v）。当替代神经排序模型为 $\mathcal{M}_{S_1}$ 或 $\mathcal{M}_{S_2}$时，观察发现，不到 30% 的连接句位于中间位置（即 $v \geq 1$）。与此同时，当替代神经排序模型为 $\mathcal{M}_{S_3}$时，插入位置的分布更加均匀，只有 10–20% 的连接句被添加到开头（即 $v=0$）。这些发现表明，IDEM 可以将对抗文本放置在目标文档中的任意位置，从而使其更难被检测到。

基于其他生成式语言模型的 IDEM。除了使用 BART 评估 IDEM 外，我们还考察了当其他具备填空能力的生成式语言模型被指示生成连接句时，IDEM 的表现如何，其中包括 T5-Base（Raffel 等，2020年）和 ILM（GPT-2 的微调版本；Donahue 等，2020年）。如表 5所示，我们的实证结果表明，尽管基于 BART 的 IDEM 整体表现最佳，但所有三种模型均取得了较高的攻击成功率，这表明 IDEM 方法具有广泛的适用性。

在线语法检查。为了检验不同攻击

方法	#正确性↓	#建议↓	质量
原始	1.80	4.23	59
查询+	2.39	6.50	49
PRADA	5.51	11.0	22
Brittle-BERT	3.74	7.12	41
PAT	2.40	6.21	52
IDEM$_{N=100}$	2.29	5.06	57
IDEM$_{N=500}$	2.31	5.18	57

表6：在线语法检查器（即 Grammarly）提供的自动评估结果。

方法	人类不可察觉性 Avg.	Kappa
原始	0.69	0.44
查询+	0.36	0.55
PRADA	0.45	0.56
Brittle-BERT	0.50	0.43
PAT	0.55	0.69
IDEM$_{N=100}$	0.78	0.11
IDEM$_{N=500}$	0.64	0.54

表 7：对抗文档不可察觉性的人工评估结果。平均分越接近 1，表示越不容易被人察觉。

方法引入的额外错误，我们使用流行的在线语法检查器 *Grammarly*[2] 来评估对抗文档的质量。具体而言，对于 50 个开发集查询，我们收集了每种攻击方法生成的 100 份对抗文档（具有相同的 ID）。我们报告 Grammarly 提供的三项评估指标，包括每份对抗文档中正确性问题（如拼写、语法和标点符号）及建议（如冗长或含义不清的句子等）的平均数量，以及所有对抗文档的质量分数。如表 6所示，IDEM 引入的问题最少，且在所有攻击方法中获得了最高的质量分数，这表明 IDEM 生成的对抗文档更具机器不可察觉性。

人工评估。 为进一步证明 IDEM 生成的对抗文档对读者而言更加自然，我们开展了人类受试者评估，以评估对抗文本的不可察觉性。具体而言，我们针对每种攻击方法随机选取了 32 份对抗文档，并另外选取了 32 份未经修改的原始文档。随后将所有文档混合并随机分成两组，每组由两名标注员评估，他们是具有必要知识、能够理解排名攻击性质的计算机科学研究生。标注员的任务是判断文档内容是否遭到攻击（0）或属于正常内容（1）。我们对两名标注员针对 32 个样本所做的所有标注取平均值，作为最终的不可察觉性得分，并计算卡帕系数以衡量标注的一致性。如表7所示，在所有攻击方法中，IDEM 的人工不可察觉性得分最高。此外，几乎所有攻击方法的 Kappa 值都大于 0.4 （被视为 “ 一致性中等 ”），而 IDEM$_{N=100}$ 的 Kappa 值最低（即 0.11）。这似乎合乎情理，因为对于攻击痕迹更难察觉的文档，人们很难达成一致。

5 相关工作

对抗性信息检索（IR）已被研究很长时间，例如黑帽搜索引擎优化，即故意操纵网页以获得不合理的排序位置，从而降低搜索结果的质量，并使无关网页泛滥（Gyöngyi 和 Garcia-Molina，2005年）。在这一背景下，主流研究侧重于从多个方面研究对抗性操纵，例如检测（Dalvi 等人，2004年；Ntoulas 等人，2006年）、理论与实证分析（Raifer 等人，2017年）、基于学习排序（LTR）的排序函数的鲁棒性（Goren 等人，2018年）、自动内容修改（Goren 等人，2020年）以及其他研究方向（Kurland 和 Tennenholtz，2022年）。

<escape>近年来，神经排序模型（NRMs）取得了显著进展，尤其是那些基于预训练语言模型（PLMs）（例如 BERT）构建的模型，其文本排序性能表现卓越（Lin 等人，2021年）。与此同时，越来越多的研究指出，NRMs 在各种场景下均存在鲁棒性问题，包括查询拼写错误或查询变体（Zhuang 和 Zuccon，2021年；Penha 等人，2022年）、文本噪声（Chen 等人，2023年）以及对抗性攻击（Raval 和 Verma，2020年；Song 等人，2022年）。尽管现有攻击方法，例如 PRADA（Wu 等，2022年）、Brittle-BERT（Wang 等人，2022年）和 PAT（Liu 等人，2022年）能够成功误导 NRMs，但它们会引入额外的质量问题，并且在文档操纵过程中往往严重依赖替代神经排序模型。相比之下，我们提出的 IDEM 有效克服了现有攻击方法的局限性，并展示了.....</escape>

[2]https://app.grammarly.com

表现出显著的攻击性能。

6 结论

在本研究中，我们提出了一种名为 IDEM 的文档操纵框架，旨在生成不易被人类和算法检测到的对抗文档。我们在 MS MARCO 数据集上的实验表明，IDEM 不仅能实现较高的攻击性能，还能生成正确、流畅的对抗文档；自动评估和人工评估均证实了这一点。未来，我们计划将研究拓展至更多先进的生成式语言模型，例如 GPT-3（Brown 等人，2020年）。

局限性

在我们的实验中，由于采用交叉编码器的神经排序模型应用广泛，我们着重评估重排序阶段的文本对抗鲁棒性，目前尚未考虑对检索阶段的影响。然而，实际上，在“先检索、后重排序”的排序范式中，只有当对抗文档进入检索结果的前列时，攻击才会奏效。与此同时，稠密检索（DR）模型已得到广泛研究；由于其建立在预训练语言模型的基础之上，也可能存在对抗脆弱性。此外，受计算资源限制，我们只测试了向相对较短的文档（即段落级文档）添加对抗文本，但现实应用中的文档内容可能要长得多。因此，进一步全面考察不同架构的神经排序模型、攻击对检索模型的影响，以及对较长文档的操纵，仍有待未来研究。

伦理声明

本文探讨了神经信息检索（IR）系统可能存在的漏洞，并提出一种文档操纵框架，用于生成不易被人类和信息检索系统检测到的对抗文档。我们希望本研究能够启发人们进一步探索和设计对抗性排序防御/检测方法，并有助于开发稳健的现实世界搜索引擎。

参考文本

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam Mc- Candlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。发表于神经信息处理系统进展第*33*卷：神经信息处理系统年会 *2020*，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行。

Carlos Castillo 和 Brian D.Davison。2010年。对抗性网络搜索。信息检索基础与趋势，4(5):377–486。

Xuanang Chen、Ben He、Kai Hui、Le Sun 和 Yingfei Sun。2023年。应对文本噪声，实现稳健有效的 BERT重排序。《信息处理与管理》，60(1):103135。

Nilesh N. Dalvi、Pedro M. Domingos、Mausam、Sumit K. Sanghai 和 Deepak Verma。2004年。对抗性分类。载于第十届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，西雅图，华盛顿州，美国，*2004*年*8*月*22*日至*25*日，第99–108页。ACM。

Brian D. Davison、Marc Najork 和 Tim Converse。2006年。万维网上的对抗性信息检索（AIRWeb 2006）。《*SIGIR*论坛》，40(2):27–30。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：面向语言理解的深度双向变换器预训练。载于 *2019*年计算语言学协会北美分会人类语言技术会议论文集：第*1*卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Chris Donahue、Mina Lee 和 Percy Liang。2020年。使语言模型能够填补空白。载于 计算语言学协会第*58*届年会论文集，第2492–2501页，线上。计算语言学协会。

Javid Ebrahimi、Anyi Rao、Daniel Lowd 和 Dejing Dou。2018年。HotFlip：用于文本分类的白盒对抗样本。载于 计算语言学协会第*56*届年会论文集（第*2*卷：短篇论文），第31–36页，墨尔本，澳大利亚。计算语言学协会。

Angela Fan、Mike Lewis 和 Yann Dauphin。2018年。层次化神经故事生成。载于计算语言学协会第*56*届年会论文集

计算语言学（第*1*卷：长篇论文），第889–898页，澳大利亚墨尔本。计算语言学协会。

Saad Farooq。2019年。关于万维网上对抗性信息检索的调查。*CoRR*，abs/1911.11060。

Gregory Goren、Oren Kurland、Moshe Tennenholtz 和 Fiana Raiber。2018年。对抗性文档操纵下的排名稳健性。载于第*41*届国际 *ACM* 信息检索研究与发展大会，*SIGIR 2018*，安娜堡，密歇根州，美国，*2018*年*7*月*8*日至*12*日，第395–404页。ACM。

Gregory Goren、Oren Kurland、Moshe Tennenholtz 和 Fiana Raiber。2020年。以排名为激励、保持质量的内容修改。载于第*43*届 *ACM* 信息检索研究与发展国际会议论文集，*SIGIR 2020*，线上会议，中国，*2020*年*7*月*25*日至*30*日，第259–268页。ACM。

ZoltánGyöngyi 和Hector Garcia-Molina。2005年。网络垃圾信息分类体系。载于*AIRWeb 2005*：第一届网络对抗性信息检索国际研讨会，与 *WWW* 大会同期举行，*2005*年*5*月，日本千叶，第39–47页。

Sebastian Hofstätter、Aldo Lipani、Sophia Althammer、Markus Zlabinger 和 Allan Hanbury。2021年。缓解 Transformer模型在段落重排序中的位置偏差。载于信息检索进展——第*43*届欧洲信息检索研究大会（*ECIR 2021*），线上会议，*2021*年*3*月*28*日至*4*月*1*日，论文集，第一部分，《计算机科学讲义》第12656卷，《计算机科学讲义》，第238–253页。施普林格。

Zhiying Jiang、Raphael Tang、Ji Xin 和 Jimmy Lin。2021年。BERT如何对段落进行重排序？基于信息瓶颈的归因分析。载于第四届自然语言处理神经网络分析与解释 *BlackboxNLP* 研讨会论文集，第496–509页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Katharina Kann、Sascha Rothe 和 Katja Filippova。2018年。句子级流畅度评估：参考文本虽有帮助，但并非必需！载于第*22*届计算自然语言学习会议论文集，第313–323页，比利时布鲁塞尔。计算语言学协会。

Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。用于开放域问答的稠密段落检索。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第6769–6781页，线上。计算语言学协会。

Oren Kurland 和 Moshe Tennenholtz。2022年。竞争性搜索。载于 *SIGIR '22*：第*45*届国际 *ACM SIGIR*信息检索研究与发展会议，马德里，西班牙，*2022*年*7*月*11*日至*15*日，第2838–2849页。ACM。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Red- field、Michael Collins、Ankur P. Parikh、Chris Al- berti、Danielle Epstein、Illia Polosukhin、Jacob De- vlin、Kenton Lee、Kristina Toutanova、Llion Jones、Matthew Kelcey、Ming-Wei Chang、Andrew M. Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。自然问题：问答研究的基准测试。计算语言学协会汇刊，7:452–466。

Yibin Lei、Yu Cao、Dianqi Li、Tianyi Zhou、Meng Fang 和 Mykola Pechenizkiy。2022年。保持标签的短语级文本对抗攻击。载于计算语言学协会论文集精选：*NAACL 2022*，第1095–1112页，美国西雅图。计算语言学协会。

Alex Goh Kwang Leng、Ravi Kumar Patchmuthu、Ashutosh Kumar Singh 和 Anand Mohan。2012。对抗性信息检索中的基于链接的垃圾信息算法。网络系统，43(6):459–475。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于第*58*届计算语言学协会年会论文集，第7871–7880页，线上。计算语言学协会。

Jimmy Lin、Rodrigo Nogueira 和 Andrew Yates。2021年。用于文本排序的预训练 *Transformer*：*BERT* 及其后续发展。人类语言技术综合讲座。Morgan & Claypool 出版社。

Jiawei Liu、Yangyang Kang、Di Tang、Kaisong Song、Changlong Sun、Xiaofeng Wang、Wei Lu 和 Xi-aozhong Liu。2022年。有序—无序：针对黑箱神经排序模型的模仿式对抗攻击。载于*2022*年 *ACM SIGSAC* 计算机与通信安全会议论文集（*CCS 2022*），美国加利福尼亚州洛杉矶，*2022*年*11*月*7*日至*11*日，第 2025–2039 页。ACM。

刘鹏飞、袁伟哲、付金兰、蒋正宝、林宏明和 Graham Neubig。2021年。预训练、提示与预测：自然语言处理提示方法的系统综述。*CoRR*，abs/2107.13586。

Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、DimitrisTsipras 和 Adrian Vladu。2018年。构建能够抵御对抗攻击的深度学习模型。载于第六届国际学习表征会议，*ICLR 2018*，温哥华，不列颠哥伦比亚省，加拿大，*2018*年*4*月*30*日至*5*月*3*日，会议论文集。OpenReview.net。

Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。2016年。MS MARCO：人工生成的机器阅读理解数据集。载于认知计算研讨会论文集：神经与符号方法集成，*2016*年，与第*30*届神经信息处理系统年会（*NIPS 2016*）联合举办，西班牙巴塞罗那，*2016*年*12*月*9*日，第1773卷，*CEUR*研讨会论文集。CEUR-WS.org。

Rodrigo Frassetto Nogueira 和 Kyunghyun Cho。2019年。使用 BERT 进行段落重排序。*CoRR*，abs/1901.04085。

Alexandros Ntoulas、Marc Najork、Mark S. Manasse 和 Dennis Fetterly。2006年。通过内容分析检测垃圾网页。载于第*15*届万维网国际会议论文集，*WWW 2006*，爱丁堡，苏格兰，英国，*2006*年*5*月*23*日至*26*日，第83–92页。ACM。

Gustavo Penha、Arthur Câmara 和 Claudia Hauff。2022年。使用查询变体生成器评估检索流水线的鲁棒性。载于信息检索进展——第*44*届欧洲信息检索研究会议，*ECIR 2022*，斯塔万格，挪威，*2022*年*4*月*10*日至*14*日，论文集，第一部分，第13185卷《计算机科学讲义》《计算机科学讲义》，第397–412页。施普林格。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督多任务学习器。*OpenAI* 博客，1(8):9。

ColinRaffel、NoamShazeer、AdamRoberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索统一文本到文本转换器的迁移学习极限。机器学习研究期刊，21:140:1–140:67。

Nimrod Raifer、Fiana Raiber、Moshe Tennenholtz 和 Oren Kurland。2017年。信息检索与博弈论相遇：文档？作者之间的排名竞争。载于第*40*届*ACM SIGIR* 信息检索研究与发展国际会议论文集，日本东京新宿，*2017*年*8*月*7*日至*11*日，第465–474页。ACM。

Nisarg Raval 和 ManishaVerma。2020年。一次一个词：针对检索模型的对抗攻击。*CoRR*，abs/2008.02197。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于 *2019*年自然语言处理实证方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第 3982–3992 页，中国香港。计算语言学协会。

Stephen E. Robertson、Steve Walker、Micheline Hancock-Beaulieu、Mike Gatford 和 A. Payne。1995年。TREC-4中的Okapi。载于第四届文本检索会议论文集，*TREC 1995*，*1995*年*11*月*1*日至*3*日，美国马里兰州盖瑟斯堡，第500-236卷，*NIST*特别出版物。美国国家标准与技术研究院（NIST）。

Congzheng Song、Alexander Rush 和 Vitaly Shmatikov。2020年。对抗性语义碰撞。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 4198–4210 页，线上。计算语言学协会。

JunshuaiSong、Jiangshan Zhang、Jifeng Zhu、Mengyun Tang 和 Yong Yang。2022年。TRAttack：针对文本检索的文本重写攻击。载于 第七届自然语言处理表示学习研讨会论文集，第 191–203 页，都柏林，爱尔兰。计算语言学协会。

Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou。2020年。MiniLM：面向预训练 Transformer 任务无关压缩的深度自注意力蒸馏。载于神经信息处理系统进展第*33*卷：神经信息处理系统年度会议 *2020*，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上。

Yumeng Wang、Lijun Lyu 和 Avishek Anand。2022年。BERT排序器十分脆弱：一项使用对抗性文档扰动的研究。载于*ICTIR '22*：*2022*年*ACM SIGIR*信息检索理论国际会议，马德里，西班牙，*2022*年*7*月*11*日至*12*日，第115–120页。ACM。

William Webber、Alistair Moffat 和 Justin Zobel。2010年。非确定性排序的相似度度量。*ACM* 信息系统汇刊，28(4):20:1–20:38。

Chen Wu、Ruqing Zhang、Jiafeng Guo、Maarten de Rijke、Yixing Fan 和 Xueqi Cheng。2022年。PRADA：针对神经排序模型的实用黑盒对抗攻击。*CoRR*，abs/2204.01321。

Jincheng Xu 和 Qingfeng Du。2020年。Texttricker：针对文本分类模型的基于损失和基于梯度的对抗攻击。*Eng. Appl.Artif. In-tell.*，92:103641。

Shengyao Zhuang 和 Guido Zuccon。2021年。处理基于 BERT 的段落检索与排序中的拼写错误。载于*2021*年自然语言处理实证方法会议论文集，第 2836–2842 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Lixin Zou、Shengqiang Zhang、Hengyi Cai、Dehong Ma、Suqi Cheng、Shuaiqiang Wang、Daiting Shi、Zhicong Cheng 和 Dawei Yin。2021年。基于预训练语言模型的百度搜索排序。载于*KDD '21*：第*27*届 *ACM SIGKDD*知识发现与数据挖掘大会，线上会议，新加坡，*2021*年*8*月*14*日至*18*日，第4014–4022页。ACM。

A 附录

A.1 替代神经排序模型的详细信息

如第4.1节所述，我们使用三种替代神经排序模型对受害神经排序模型（NRM） $\mathcal{M}_V$ 进行黑盒攻击。由于通常无法获得受害神经排序模型的训练数据，我们使用 MS MARCO 数据集中的评估查询，而非训练查询，来构建替代模型训练数据 $\mathcal{T}_S$。从 $\mathcal{M}_{S_1}$ 到 $\mathcal{M}_{S_2}$，用于排序模仿的域内评估查询数量由6,837个降至200个，这意味着查询受害神经排序模型的频率大幅降低。对于每个评估查询 q_i，我们从受害神经排序模型针对 BM25 排名前1,000的候选文档生成的重排序列表前29项中，收集全部406个文档对（d_i、$_j$、d_i、$_k$）（$1 \leq j < k \leq 29$），并按第2节所述，利用这些文档对构建替代模型训练数据 $\mathcal{T}_S$。因此，$\mathcal{M}_{S_1}$ 和 $\mathcal{M}_{S_2}$ 的域内替代模型训练数据分别包含8.12万个和277万个训练三元组。此外，$\mathcal{M}_{S_3}$ 表示受害神经排序模型没有任何域内数据可用的场景；我们从 NQ 数据集中收集372万个训练三元组，作为域外（OOD）替代模型训练数据。替代神经排序模型基于预训练 BERT-Base 模型，并针对 $\mathcal{M}_{S_1}$ 和 $\mathcal{M}_{S_2}$以3e-6的学习率和16的批大小微调两个轮次；针对 $\mathcal{M}_{S_3}$ 则使用相同的学习率和批大小微调一个轮次。需要指出的是，本研究的目标并非开发一种新的替代神经排序模型训练方法，因此我们直接采用以往工作（Wu 等人，2022）中的合页损失，并将间隔设为1，以进行排序模仿。

A.2 攻击基线的详细信息

在我们的对抗攻击实验中，我们考察了以下基线方法：

查询+ (Liu 等人，2022年) 是一种直观的基线方法，会直接将查询文本附加到目标文档的开头。尽管查询文本可以放置在目标文档的任意位置，甚至可以由我们提出的位置感知机制来确定，但由于基于 Transformer 的神经排序模型存在位置偏差，将其附加到开头通常会产生更好的攻击结果（Jiang 等人，2021年；Hofstätter 等人，2021年）。因此，查询+作为一种基线方法，并未考虑攻击的不可见性。

PRADA （Wu 等，2022年）是一种词语替换排名攻击（WSRA）方法首先根据梯度幅度找出目标文档中的重要词语（即子词标记）（XuandDu，2020年），然后通过 PGD（Madry etal.，2018年）贪婪地将其替换为在扰动词嵌入空间中找到的同义词。根据我们的观察，在攻击不同的随机目标样本时，PRADA 能够取得与其原始出版物中报告的结果相近的攻击性能（攻击成功率）。当受害神经排序模型的排序性能相对较差时，这一点尤为明显，因为攻击较弱的受害神经排序模型显然要容易得多。尽管存在这种差异，我们实验得出的总体结论仍未改变。

Brittle-BERT (Wang 等人，2022年) 通过添加或替换少量词元，研究局部（即特定的查询-文档实例）和全局（即整个查询工作负载）排名攻击，以造成较大的排名下降或上升。在我们的工作中，我们仅采用局部设置，并将词元添加到目标文档的开头，因为这样通常能产生更好的攻击结果。具体而言，Brittle-BERT 首先在目标文档开头初始化若干占位词元，然后使用 HotFlip（Ebrahimi 等人，2018年）算法对其进行更新，使其更具对抗性。

PAT (Liu 等人，2022年）在目标文档开头生成并添加若干触发词元。除了排名激励目标外，PAT 的搜索目标还利用预训练的 BERT 模型引入了语义约束和流畅度约束。使用合页损失训练的替代神经排序模型具有单类预测层，但 PAT 需要一个具有双类预测层的替代神经排序模型，即 PAT（Liu 等人，2022年）中所称的“Pairwise BERT”，因此，我们使用相同的代理模型训练数据，并采用 PAT 中默认的模仿损失来获得“Pairwise BERT”。

为了评估这些基线方法，我们采用其公开可用的实现，并确保所有设置均与各自官方出版物中描述的设置保持一致。

A.3 攻击耗时

在先前的 PRADA、Brittle-BERT 和 PAT 中，使用替代神经排序模型逐一进行词元替换、选择和搜索，以生成对抗文档，因此完成攻击过程需要耗费大量时间。然而，在我们提出的 IDEM 框架中，

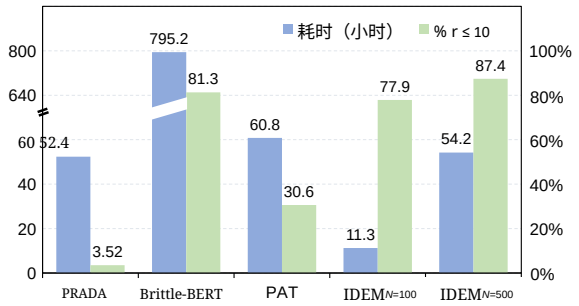

图2：使用不同攻击方法生成 5K 个对抗文档的总耗时，以及在替代模型$\%\,r \leq 10$下针对 Easy-5 目标文档的攻击性能 $\mathcal{M}_{S_1}$（见表 1）。

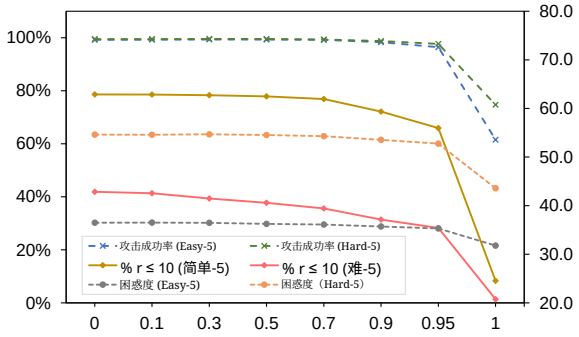

图3： α 对合并分数（公式 5）中 IDEM$_{N=100}$攻击结果的影响。

首先通过生成式语言模型（例如 BART）生成对抗文本，然后在句子级进行组合，这比词元级更高效。如图 2所示，我们总结了使用一块 Titan RTX GPU 生成 5,000 个对抗文档时，不同攻击方法所需的总时间。当位置组合所生成的候选连接句最多为 100 个时，IDEM $_{N=100}$ 仅需 11.3 小时即可实现出色的攻击性能。相比之下，PRADA 和 PAT 耗时更长，但攻击效果仍较差；Brittle-BERT 则耗时巨大（近 800 小时），尽管其攻击性能相当可观。此外，与 Brittle-BERT 相比，IDEM$_{N=500}$ 能够以少得多的时间成本实现相当的攻击性能。因此，我们提出的 IDEM 方法兼顾了攻击效率和性能。

A.4 合并分数中 α 的影响

在我们的 IDEM 框架中，生成查询与目标文档之间的一系列连接句后，会采用一种感知位置的合并机制来确定最终的对抗文档，其中连贯性分数和相关性分数通过相加的方式计算，如式 α 所示5。如图3所示，我们可以观察到 α 在较大范围（从 0 到 0.95）内变化，对 Easy-5 和 Hard-5 目标文档集上的攻击成功率（ASR）影响不大，而 $\%\,r \leq 10$指标从 $\alpha=0.5$和 $\alpha=0.1$开始分别在 Easy-5 和 Hard-5 目标文档集上下降。至于困惑度（PPL）指标（越低越好），当 α 增大（更加关注连贯性）时，PPL 值变化不大，直到α 达到约 0.9 至 1。与此同时，该方法能够生成 PPL 值低于原始文档的对抗文档。例如，当 α 为 1 时，Easy-5 和 Hard-5 目标文档集上的平均 PPL 值分别仅为 31.8 和 43.6。

A.5 对抗性样本

为了更好地定性了解不同攻击方法的工作方式，我们在表8中展示了它们在三种替代神经排序模型下生成的对抗性样本。我们可以观察到，Query+以及在 $\mathcal{M}_{S_1}$ 下的 Brittle-BERT、PAT 和 IDEM，均能将排名第 88 的目标文档提升至前 10 名。然而，在开头添加查询文本（即 Query+）和不自然的词元序列（即 Brittle-BERT 和 PAT），会使对抗文档容易被辨认；而 IDEM 插入的对抗文本在语义上与原始上下文更加连贯。当替代神经排序模型从 $\mathcal{M}_{S_1}$ 退化为 $\mathcal{M}_{S_2}$（因为可用的域内训练样本不足）时，我们可以看到，PRADA 生成的对抗文档排名从第 27 位降至第 65 位，Brittle-BERT 生成的对抗文档排名也从第 2 位降至第 14 位。此外，由于无法访问受害神经排序模型而使用分布外替代神经排序模型（即 $\mathcal{M}_{S_3}$）时，我们发现 PRADA、Brittle-BERT 和 PAT 的攻击效果都大幅减弱。例如，尽管 PAT 在 $\mathcal{M}_{S_2}$ 下能将目标文档提升至第 2 位，但 PAT 在 $\mathcal{M}_{S_3}$下反而会将目标文档从第 88 位降至第 107 位。相比之下，在 $\mathcal{M}_{S_2}$和 $\mathcal{M}_{S_3}$两种情况下，IDEM 都能稳健地将目标文档提升至前 10 名（即第 9 位），且对抗文档的流畅性和正确性仍处于可接受范围内。从这些对抗性案例可以看出，IDEM 对替代神经排序模型的依赖较小，攻击表现比以往的攻击方法更加稳健，这表明 IDEM 在实际场景中的适用条件更加灵活。

方法	原始文本或对抗文本	排名↓	PPL↓
原始	用背带或前置婴儿背带抱着宝宝。感受宝宝的温暖，闻着他身上甜美的气息，并经常低头与他目光接触，这有助于你们建立亲密关系。花充足的时间与婴儿面对面相处。对他微笑，并当他先微笑时，回以微笑。	88	33.9
查询+	多大年龄可以把婴儿背在背带上? 用背带或 fr- 携带婴儿背带。感受婴儿的温暖，闻着他甜美的气息，看着.....	2	40.4
替代神经排序模型 $\mathcal{M}_{S_1}$			
PRADA	背上 用背带背着婴儿 或 使用前置婴儿背带。感受婴儿的 温暖，闻闻他的甜美气息，经常低头与他目光接触，可以帮助你 建立-亲密关系。多花些时间与婴儿面对面相处。对他微笑，然后.....	27	100.8
Brittle-BERT	pendingerabidheartedivating aged aged 292,oning wornwear 把婴儿抱在背带或前置婴儿背带。感受婴儿的温暖，闻着他甜甜的气息，还有.....	2	125.0
PAT	大约30岁的婴儿可以背着婴儿背带回到年龄用.....抱着你的婴儿背带或前置婴儿背带。感受婴儿的温暖，闻着他甜美的气息，以及.....	2	52.0
IDEM$_{N=100}$	使用背带或前置婴儿背带携带婴儿。 任何年龄的儿童都可以穿带有一条背带。感受婴儿的温暖，闻着他甜美的气息，看着.....	9	33.7
IDEM$_{N=500}$	用背带或前置婴儿背带抱着婴儿。大多数父母佩戴婴儿背带时会将其绕在..... 以及第3、4和5岁。感受你的婴儿的温暖，闻到他甜美的气息，以及.....	1	36.5
替代神经排序模型 $\mathcal{M}_{S_2}$			
PRADA	将婴儿放在弹弓 ord前置婴儿背带中。感受婴儿的温暖、气味-闻着他甜美的气息，并经常低头与他进行目光接触，可以.....	65	82.2
Brittle-BERT	现代主义年龄 hms× 栗色 碧昂丝 说唱歌手 商业上 虽然穿着 md r-分别用背带或前置婴儿背带背着婴儿。感受婴儿的.....	14	175.2
PAT	满一岁，可以穿戴用背带或前置婴儿背带背着您的婴儿。感受-感受婴儿的温暖，闻着他身上香甜的气息，并时常低头看着.....	2	48.3
IDEM$_{N=100}$	用背带或前置婴儿背带抱着婴儿。 将婴儿背在背上总是个好主意。 感受婴儿的温暖，闻着他甜甜的气息，看着.....	9	29.4
IDEM$_{N=500}$	用背带或前置婴儿背带抱着婴儿。 你可以，也应该用背带把婴儿背在背上。 感受婴儿的温暖，闻着他甜甜的气息，还有.....	9	36.9
替代神经排序模型 $\mathcal{M}_{S_3}$			
PRADA	穿戴 您的婴儿，使用 弹弓或 前置婴儿背带。感受婴儿的温暖，闻着——闻着他甜美的气息，并经常低头与他进行目光接触，可以.....	28	67.2
Brittle-BERT	后代 连贯 示例 下降 玩具 广泛 成年期 名词 是否 bu-磨损脱落 用背带或前置婴儿背带抱着婴儿。感受婴儿的.....	58	143.8
PAT	例如，可能携带二十美分用背带或前置婴儿背带抱着婴儿。感受婴-儿的温暖，闻着他甜美的气息，并经常低头看着.....	107	49.4
IDEM$_{N=100}$	任何年龄的孩子都可以穿着鞋子使用背带。 用背带或前置婴儿背带抱着婴儿。感受婴儿的前置婴儿背带。感受婴儿的温暖，闻着他甜美的气息，并低头看着.....	9	39.1
IDEM$_{N=500}$	使用背带或前置婴儿背带携带婴儿。 你可以且应该将婴儿背在重新放回背带里。感受婴儿的温暖，闻着他身上甜甜的气息，以及.....	9	36.9

表8：针对 MS MARCO 开发集中查询“婴儿多大可以背在背带里？”的同一篇相关但不切题的文档，使用三种替代神经排序模型和不同攻击方法生成的对抗文档。插入和扰动的词语标记为红色，以便于比较。