计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(7): 1668−1684, 2025

面向大语言模型安全部署的可信评估体系

叶文涛[1] 胡家齐[1] 王皓波[2] 陈 刚[1] 赵俊博[1]
[1](浙江大学计算机科学与技术学院 杭州 310027)
[2](浙江大学软件学院 浙江宁波 315048)
(yewt01@zju.edu.cn)

A Trusted Evaluation System for Safe Deployment of Large Language Models

Ye Wentao[1], Hu Jiaqi[1], Wang Haobo[2], Chen Gang[1], and Zhao Junbo[1]
[1](*College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310027)
[2](*School of Software Technology*, *Zhejiang University*, *Ningbo*, *Zhejiang* 315048)

Abstract The recent popularity of large language models (LLMs) has brought a significant impact to boundless fields, particularly through their open-ended ecosystem such as the APIs, open-sourced models, and plugins. However, with their widespread deployment, there is a general lack of research that thoroughly discusses and analyzes the potential risks concealed. In that case, we intend to conduct a preliminary but pioneering study covering the robustness, consistency, and credibility of LLMs systems. With most of the related literature in the era of LLMs uncharted, we propose an automated workflow that copes with an upscaled number of queries/responses. Overall, we conduct over a million queries to the mainstream LLMs including ChatGPT, LLaMA, and OPT. Core to our workflow consists of a data primitive, followed by an automated interpreter that evaluates these LLMs under different adversarial metrical systems. As a result, we draw several and perhaps unfortunate conclusions that are quite uncommon from this trendy community. Briefly, they are: 1) the minor but inevitable error occurrence in the user-generated query input may, by chance, cause the LLM to respond unexpectedly; 2) LLMs possess poor consistency when processing semantically similar query input. In addition, as a side finding, we find that ChatGPT is still capable to yield the correct answer even when the input is polluted at an extreme level. While this phenomenon demonstrates the powerful memorization of the LLMs, it raises serious concerns about using such data for LLM-involved evaluation in academic development. To deal with it, we propose a novel index associated with a dataset that roughly decides the feasibility of using such data for LLM-involved evaluation. Extensive empirical studies are tagged to support the aforementioned claims.

Key words large language models; scalable deployment; robustness; automated workflow; trusted evaluation

摘 要 近年来,大语言模型(large language model, LLM)(以下简称“大模型”)的流行在众多领域带来了重大影响,特别是它们的开放式生态系统,如应用程序接口、开源模型和插件.然而,尽管大模型已经广泛部署,对其潜在风险进行深入讨论与分析的研究仍然普遍缺乏.在这种情况下,针对大模型系统的鲁棒性、一致性和可信性进行一项初步但具有开创性的研究.由于大模型时代的许多文献都尚未被实证,提出了一个自动化的工作流,用以应对不断增长的查询和响应.总体而言,对包括 ChatGPT, LLaMA, OPT 在内

收稿日期: 2024-06-21; **修回日期**: 2025-03-03
基金项目: 浙江省重点研发计划项目(2024C01035)
This work was supported by the Pioneer Research and Development Program of Zhejiang (2024C01035).
通信作者: 王皓波(wanghaobo@zju.edu.cn)

的主流大模型进行了 100 多万次查询. 工作流程的核心是一个数据原语，然后是一个自动解释器，它在不同的对抗性度量系统下评估这些大模型. 最终，从这一主流社区中得出了几个十分不同寻常的结论（一定程度上不太乐观）. 简而言之，这些结论包括：1）用户生成的查询输入中的微小但不可避免的错误可能偶然地导致大模型的意外响应；2）大模型在处理语义相似的查询时具有较差的一致性. 此外，还附带发现 ChatGPT 即使在输入受到极端污染的情况下仍然能够产生正确的答案. 这一现象虽然表明了大模型的强大记忆力，但也引发了人们对在学术发展中使用大模型参与评估的严重关切. 为了解决这一问题，提出了一个与数据集相关联的新指标，该指标大致决定了基于这些数据对大模型进行评估的可行性. 最后进行了广泛的实证研究，以支持上述主张.

关键词 大语言模型；规模化部署；鲁棒性；自动化工作流；可信评估

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440566 **CSTR:** 32373.14.issn1000-1239.202440566

近年来，以 ChatGPT 为代表的大语言模型（large language model, LLM）（以下简称“大模型”）风靡全球，在计算机科学界的众多领域造成了巨大影响[1]. 在网页用户界面、开源模型[2]和生态系统插件的协助下，这些大模型成功融入了全世界每个人的生活.

大模型的这种成功确实是史无前例的，甚至在整个技术发展范围内都是罕见的，因而探索大模型对社会的潜在机遇和挑战十分重要[3]. 这引发了围绕大模型的各种评价工作[4-8]，然而这些工作遇到了诸多挑战：1）大模型输出的复杂性导致对人工评价的严重依赖，阻碍了大规模评估；2）与传统自然语言处理（natural language processing, NLP）领域的小型模型不同，当前的大模型已经被广泛部署和暴露，构成了通过现有的 NLP 能力评估难以发现的额外潜在风险；3）大模型海量和未知的训练数据导致如何选择可信的评估数据成为一个紧迫的问题.

在此驱动下，本研究充分考虑了相关场景中经常遇到但在以前的工作中很少解决的问题. 作为一种开拓性的尝试，本研究提出了一个自动化的工作流来对大模型进行系统的研究，涵盖了鲁棒性、一致性、可信性这 3 个新的术语.

1）鲁棒性. 大模型的鲁棒性面向有意或无意执行的恶意查询. 事实上，可以将这个问题视为针对 NLP 对抗性实例的传统威胁模型. 相反，在这项工作中，界定了大模型相对于传统攻击手段的鲁棒性，特别是将威胁模型与大模型的实际部署相匹配.

2）API（application programming interface）的一致性. 这是本研究试图推广的一个新概念. 简而言之，试图定量衡量语言模型在处理 2 个语义相同的不同输入时的差异. 为此，提出了 1 种新的威胁模型，其结合了 2 种模型自适应的攻击手段.

3）评估的可信性. 这一衡量标准可能与以大模型为中心的学术界关系最密切. 这一概念关注于以下现象：在某些数据集上，ChatGPT 仍然能够吐出正确的答案，即使段落信息被完全污染！一方面，由于大模型设法记忆所提供的数据集，因此可以认为导致了 ChatGPT 这种“盲目”的问答能力. 尽管如此，本研究敦促社区在使用这些数据集来评估由大模型构成的任何潜在框架时要谨慎.

尽管上面已经简单列出了一些概念，但距离完成对大模型的可信评估还有一些必需的步骤. 在此简要介绍总体的工作流程.

首先，使用 gpt-3.5-turbo（ChatGPT 的官方模型 API），以及开源的模型 LLaMA、OPT 作为研究的主要骨架. 值得注意的是，为了获得统计意义上的可靠结果，在每个模型上执行了超过 100 万条查询（在 ChatGPT 上约有 7 亿 token）.

然后，为了处理大规模的查询-响应对，本研究需要一个自动化的解释器. 为此，设计了 1 个通用的数据原语，表示形式为 $(\boldsymbol{t},\boldsymbol{p},q,o,a)$，其中这 5 个符号依次代表提示词、段落、询问、混淆选项、正确答案. 在 12 个公开可用的数据集上，设法将每个数据点转换为符合该形式的原语表达. 该原语本质上是一个问答格式，其中每个问答本身是一个多项选择题. 通过规范大模型的输出，我们能够并行地处理大规模的响应.

最后，围绕大模型评估的 3 个方面，分别提出了 2 种威胁模型、5 种攻击方案以及 1 个和数据集本身关联的指标. 简单来说，2 种模型分别面向鲁棒性、一致性. 而 5 种攻击方案与传统的 NLP 对抗性攻击[9]不同，本研究有目的性地将攻击方案结构化，以便这些攻击能与大模型的一些具体使用场景相对应. 还

提出了相对训练指数(relative training index, RTI), 该指标与数据集相关联. RTI 为定量衡量数据集已经被模型记住的概率提供了一个参考指标. 因此, RTI 可以被看作是一个衡量数据集可信性的标准, 有助于决定是否在评估中使用特定的某个数据集. 本研究提出的评估框架体系总览如图 1 所示.

通过实验研究和分析, 本研究关注的问题可以概括为 3 个方面:

1)在模型的鲁棒性方面. 本研究发现, 那些与大模型使用流程密切相关的攻击可能构成巨大潜在威胁. 比如, 与光学字符识别(optical character recognition, OCR)输入关联的视觉攻击以及与语音识别输入关联的键入错误. 此外, 使用 ChatGPT 的开发人员应该对输入内容保持谨慎, 因为一些微小但结构化的更改可能会对模型或接口输出造成不可能预期的巨大改变, 比如, 根据本研究对 ChatGPT 的测试发现, 每次测试的变化率不低于 27%.

2)在评估的一致性方面. 以 ChatGPT 为例, 对于语义相同的查询输入, 在本研究的测试标准下, ChatGPT 对响应的准确率平均波动 3.2%. 需要进一步指出的是, 输入的变化主要是在语法表达、写作习惯等方面, 考虑到这些大模型的广泛部署, 这种情况相当常见.

3)在数据集的可信性方面.提出了 RTI, 可以用于衡量数据集被记忆的相对概率, 从而衡量所提供的数据集是否适合用于大模型相关的评估. 该指标可以反映应用该数据集的评估是否具有可信性. 我们希望这个指标可以帮助学术界在大模型时代构建一个更合适的评估生态系统.

1 相关工作

1.1 NLP 中的对抗攻击

在传统 NLP 任务中, 已经有各种关于对抗性攻击模型的研究工作. 由于文本数据的无差异性, 攻击大多是在黑盒威胁模型条件下进行的. 这些攻击形式广泛, 包括字符级别扰动[10-13]、单词级别扰动[14-16], 以及句子级别扰动[17-18]. 此外, 其他工作[19-24]遵循白盒方式, 需要访问模型的梯度、结构或者参数. 然而, 由于目前 ChatGPT 等高性能大模型只提供接口服务, 只能使用黑盒攻击来进行评估.

1.2 大模型中的对抗攻击

文献 [25−29] 针对大模型提出了多种面向鲁棒性评估的基准方法. 这些工作通常在大模型的输入上应用可控的扰动, 如打字错误、实体交换、否定、句子插入等.

尽管如此, 这些方法中有许多只能应用于较小的语言模型上, 如 BERT[30], XLNET[31]. 考虑到目前大模型的迅速发展和扩大, 本研究认为有必要对 ChatGPT 等更大参数规模的大模型进行研究. 此外, 2023 年的一项工作[32]表明, ChatGPT 在多数对抗性分类任务、翻译任务中展现了一致性上的劣势.

1.3 威胁模型

威胁模型的概念源于计算机安全领域, 它指的是识别、评估、管理系统运转过程中可能遭受的威胁. 在 NLP 领域中, 威胁模型被广泛用于帮助管理员识别可能的威胁, 以便采取相应的预防措施. 文献

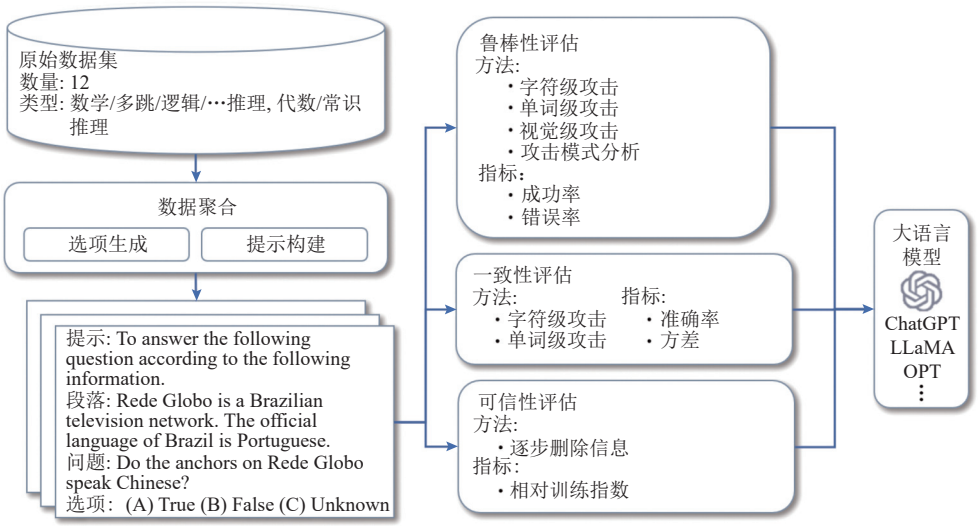


Fig. 1 Overview of evaluation framework

图 1 评估框架总览

[33−37] 在威胁模型、黑盒系统或 API 方面开展了各种工作. 然而, 这些方法大多集中在较小的机器学习模型上, 对现在的大模型来说影响十分有限. 而随着 ChatGPT 等模型的普及, 迫切需要新的威胁模型以及更紧密地结合大模型的应用场景.

1.4 大模型的评估是否可靠

事实上, 随着 ChatGPT、GPT-4 的兴起, 在现成的数据集上评估这些模型的工作[4−8] 已经浮出水面.

然而, 当评估涉及能力强大的大模型时, 这些方案需要进一步审查. 换言之, 本研究的 RTI 指数能够相对粗略地表明大模型已经不同程度地记住了被投喂的测试样本. 因而对外部模块单独进行评估而不考虑模型本身可能值得怀疑, 因为本研究的实验表明, 对于某些数据集而言, ChatGPT 能够在没有任何有效信息的帮助下回答问题.

事实上, 在没有关于模型训练所使用数据集的额外信息下, RTI 只能代表数据集的相对分数而非绝对分数. 有关详细信息, 请见 2.6 节.

2 方法、数据、需求

本节详细阐述了本研究工作流程的细节. 首先, 构造了适用于这一特定任务的工作表. 接下来, 详细介绍用于评估当前大模型鲁棒性、一致性、可信性的具体机制. 作为总结, 本研究正式提供了这些机制对应的度量系统.

2.1 原始数据集设置

本研究采用了 12 个公开可用的数据集作为基准. 注意, 这些数据集主要包含数学演绎、逻辑推理、常识理解等任务. 完整设置如表 1 所示.

Table 1 Datasets Statistics

表 1 数据集统计

数据集	描述	子集	规模	答案类型
StrategyQA[38]	问答（QA）	训练集	2 290	T/F
AQuA[39]	代数问答（algebra QA）	测试集	254	数值
Creak[40]	常识推理（commonsense reasoning）	验证集	1 371	T/F
NoahQA[41]	数值推理问答（numerical reasoning QA）	测试集	10 880	多项选择
GSM8K[42]	数学推理（math reasoning）	测试集	6 140	文本
bAbi15[43]	演绎推理（deductive reasoning）	测试集	1 000	单词
bAbi16[44]	归纳推理（inductive reasoning）	测试集	1 000	单词
QASC[45]	多跳推理（multi-hop reasoning）	验证集	926	单词
ECQA[46]	常识推理（commonsense reasoning）	测试集	2 194	单词
e-SNLI[47]	逻辑关系推理（logical relationship reasoning）	测试集	9 824	单词
Sen-Making[48]	常识推理（commonsense reasoning）	测试集	2 020	文本
QED[49]	问答解析（QA with explanations）	验证集	1 354	文本
	总计		39 253	

2.2 数据聚合

本研究在选定的基准大模型上, 运行了共计超过 100 万条查询. 为了应对查询响应数量的不断增长, 必须首先找到一个自动解释器, 用于在无需人工审查、干预的情况下获得每个独立的响应.

也就是说, 作为自动解释器的核心组件, 从原始数据集中提取的原始模板需要修改. 为此, 设计了统一的数据模板, 这一设计本质上是一个多选项问答数据原语, 如表 2 所示, 其中展示了 1 组示例. 特别地, 将目标模板表示为五元组形式($\boldsymbol{t}$, $\boldsymbol{p}$, q, o, a). 其中包含一组混淆选项, 这些选项满足: 1)不是正确选项; 2)与正确选项a相近. 参数$\boldsymbol{t}$表示提示词. 值得注意的是, a并不是特别长的段落, 并且 ChatGPT 能够在设置为o的情况下按照预期方式输出. 本质上, 这一模板化的数据表促使大模型根据段落$\boldsymbol{p}$、被连接的问题q, 正确地输出区分正确答案a, o中的其他混淆选项. 由于表 2 多选项的特点, 可以通过简单地扫描大模型返回的响应来获取答案索引(如 A, B, C, D), 从而进行准确计算. 通过高质量的提示$\boldsymbol{t}$, 这种自动转换机制在经验上效果良好, 几乎没有例外.

在实现中, 首先构建一个转换器来自动获取($\boldsymbol{p}$,q,a)部分, 但混淆选项除外. 对于大多数分别提供

Table 2 Process of Transforming from the Original Example to the Model Input

表 2 从原始示例转换到模型输入的过程

结构	示例
原始	“question”: “Is Antarctica a good location for Groundhog Day?” “answer”: false “facts”: [“Groundhog Day relies on a groundhog seeing their shadow.” “Antarctica has an irregular sun pattern and some days have no sun rise or 24 hour sunlight.” “Antarctica has temperatures can range from −10°C to −60°C.” “Groundhogs live in forests or woodlands with plenty of sunlight.”]
提示词t	Next, I will ask you a series of questions given a description, and you will have to choose one of several candidate options that you think is correct. The description is
段落$\boldsymbol{p}$	Groundhog Day relies on a groundhog seeing their shadow.Antarctica has an irregular sun pattern and some days have no sun rise or 24 hour sunlight. Antarctica has temperatures can range from −10°C to −60°C.Groundhogs live in forests or woodlands with plenty of sunlight.
问题q	Is Antarctica a good location for Groundhog Day?
选项o	(A) True (B) False (C) Unable to determine
查询	Query 1: {提示词}+{段落} Query 2: The first question is {问题}, choose an answer from the following options: {选项}.
回答a	The answer would be (B) False. Antarctica is not a good location for Groundhog Day since the area has an irregular sun pattern, and some days could have no sun rise, or 24-hour sunlight. Groundhogs rely on seeing their shadow to predict the weather, and the irregular patterns of sunrise and sunset would make the whole concept unworkable in Antarctica.

段落、问题的数据集(AQuA[39], ECQA[46], QASC[45], QED[49]),直接提取相应的字段.以表 2 中的例子为例,“段落”“问题”“回答”字段的上下文分别对应于$(\boldsymbol{p},q,a)$.

对于预期由段落组成的数据集,上述机制应该已经足够.对于那些不包含段落的,通过以下步骤改变原语:1)将问题分成 2 部分,其中前一部分涉及并且只涉及陈述或条件,而后半部分被简化为疑问句,这个过程是通过简单的识别、句子切分自动实现的;2)对于问题不可分解的数据集,只需将$\boldsymbol{p}$设置为 Null.

本研究设计了 1 个生成器,用于根据答案a产生一些混淆选项o.作为原语的核心部分,大模型被期望从这些混淆选项中找出正确选项a.根据表 1 可以总结出,这些数据集中的答案主要按照以下 4 种类型分类:T/F(True 或 False,即正确或错误)、数值、单词、文本.对于原语中的o来说,一个好的混淆选项应该尽可能与基本事实相似,但又可以利用大模型的常识理解、推理能力进行区分.表 3 列举了一些混淆选项的例子.特别地,对于 4 种不同的类别,分别设计了不同的生成规则,如下所示:

1) 对于已经提供了一些错误选项的数据集(如 AQuA, QASC, ECQA, e-SNLI),直接将这些选项用作选项 o.

2)对于判断正确或错误类型的任务,如 StrategyQA, Creak,尽管本身的 T/F 形式已经满足原语的要求,额外在选项o中加入了第 3 个选项“unable to determine”(即“无法确定”),以一定程度增加任务难度.

Table 3 Confusion Options Generation Samples

表 3 混淆选项生成示例

答案类型	原始答案	生成选项
T/F	True	**(A) True** (B) False (C) Unable to determine
数值	36	**(A) 36** (B) 15 (C) 17 (D) 5 (E) 7
单词	discovery	**(A) discovery** (B) action (C) reflection (D) deciding (E) thinking
文本	Janet sells 16−3−4=9 duck eggs a day	**(A) None of the other options is correct.** (B) Janet elude sells 16−3−4=9 duck eggs a axerophthol day. (C) Janet sells 4−4−10=−10 duck eggs a day. (D) Janet sells 11−11−15=28 duck eggs a day. (E) He has to pay 3 000−1 000=2000.

注:加粗文字代表正确答案.

3)对于答案是数字的任务,如 NoahQA(其中的一部分问题),随机生成 4 个数作为选项o.因此,一共有 5 个选项.

4)对于答案是单词的任务,如 bAbi15, bAbi16, ECQA,对文本、问题使用词性标注器[50],并根据标注结果随机选择其他具有相同语言属性的单词.针对

此类任务，一共有 5 个选项.

5)对于答案是文本的任务，如 GSM8K，交替采用 3 种方法：①删除、插入、替换某些单词；②如果存在公式，改变公式的正确性或者用某个步骤的公式替换其为其他步骤的公式；③移除正确选项a，并添加一个额外的选项“None of the other options is correct”(即“其他选项均不正确”).

基于提示词模板、上下文学习[51](in-context learning, ICL)，本研究手动构建了一个提示模板$\boldsymbol{t}$，格式如下：

Next, I will ask you a series of questions given a description, and you will have to choose one of several candidate options that you think is correct. The description is.

如表 2 中“查询”行所示，查询中设置了：1)由提示词$\boldsymbol{t}$、段落$\boldsymbol{p}$直接连接得到的上下文. 2)由问题q，选项o、其他连接短句组成的句子流. 3)如果 1 个样例为 1 个段落提供了多个问题，将依次用每个问题和相应的选项作为后续查询的内容. 经过验证，发现该模板和 ICL 设置(几乎)普遍适用于约束 ChatGPT 按照期望的方式回答所提供的问题.

本研究遵循围绕鲁棒性、一致性、可信性的评估方法. 通过这种方法，希望为 ChatGPT 对实际应用程序的潜在风险提供一个全面的参考. 在此之前，将给出一些设置. 数据集 D 表示为

$$D=\{\boldsymbol{x}_i\}_{i=1}^{n}=\{(\boldsymbol{t}_i,\boldsymbol{p}_i,q_i,o_i,a_i)\}_{i=1}^{n}, \tag{1}$$

其中$\boldsymbol{x}_i$代表 D 中的第 i 个示例.

2.3 鲁棒性评估

本研究首先要评估的是大模型的鲁棒性. 简单来说，本研究的方法灵感来源于传统 NLP 领域，尤其是其中围绕文本上的对抗性示例的研究. 借用文献[28]中的术语，一个扰动示例定义为

$$\boldsymbol{x}'=\boldsymbol{x}+\boldsymbol{\delta};\|\boldsymbol{\delta}\|\leqslant\epsilon\wedge f(\boldsymbol{x},\boldsymbol{\theta})\neq f(\boldsymbol{x}',\boldsymbol{\theta}), \tag{2}$$

其中$\boldsymbol{\delta}$表示插入到原始样本$\boldsymbol{x}$中的扰动，ϵ表示扰动程度的量化，从而产生其扰动后的对应结果$\boldsymbol{x}'$. $f(\boldsymbol{x},\boldsymbol{\theta})$表示由$\boldsymbol{\theta}$参数化的函数形式的模型.

通常，在传统的 NLP 模型的鲁棒性评估中，ϵ 被认为是非常小的. 尽管如此，本研究将重点从对抗性的例子转移到干净的例子上，将距离/差异最小化. 相比之下，本研究强调通过更具结构性的方法来构建对抗性实例，这些方法更符合大模型的实际部署，而不是随机的对抗性扰动. 如，随着 ChatGPT API 的发布，用户生成的输入将呈指数级增加到 API，随后是大模型响应的分布偏移. 在这项工作中，本研究深入探究并衡量用户生成内容中常见的错误，如字符级键入错误、不正确的单词(如从语音转录获得的用户输入)或视觉缺陷(如从 OCR 产生的用户输入)等. 相信这种设置可以模拟大模型特定部署的可能结果，相较围绕传统对抗性示例的研究更有利.

本研究使用一个自动攻击器g来进行对抗性攻击(构造对抗性实例)过程. 通过这个攻击器，可以将该过程用于任何数据集，作为通用自动评估系统的一部分. 按照常规设置[52-54]，如果$\boldsymbol{p}$存在，则仅对$\boldsymbol{p}$执行对抗性攻击 (否则对q执行). 样本$\boldsymbol{x}$的对抗性实例$\boldsymbol{x}'=(\boldsymbol{t},\boldsymbol{p}',q,o,a)$中的结果表示为

$$\begin{aligned}\boldsymbol{p}'&=(w_1^*,w_2^*,\cdots,w_n^*),\\ \boldsymbol{p}&=(w_1,w_2,\cdots,w_n),\end{aligned} \tag{3}$$

其中w, w^*分别代表$\boldsymbol{p}$, $\boldsymbol{p}'$中的单词，w^*可以被进一步表示为

$$w_i^*=\begin{cases}z\sim\mathcal{U}(0,1),\\ g(w_i), & 0<z<\rho,\\ w_i, & \text{其他},\end{cases} \tag{4}$$

其中$\mathcal{U}$是均匀分布，ρ是预设的 [0,1] 间的概率，代表每个词被攻击的概率.

接下来，本研究将关注每个单词w上的自动攻击函数 g 的具体实例. 再次说明，本研究希望模拟大模型的特定用法. 对于下面的设置，所考虑大模型的鲁棒性是通过计算受扰动的单词相对于原始单词改变的精度来衡量的. 通过表 4 中提供的示例，定义了以下方法来补充自动攻击函数 g. 值得注意的是，所有攻击都集中在每个单词w上. 而每个w保持着上面提到的被更改的概率ρ.

1)单词级别攻击. 对于每个需要被更改的w，随机执行 3 种操作(即插入、删除、替换)中的 1 种. 值得注意的是，对于替换、插入，更改后的单词w^*有 50% 的概率是原始段落中的随机单词，剩下的则是w的同义词或来自 WordNet[55] 的随机单词. 单词级别攻击可以模拟应用程序中的单词错误，如语音识别中的识别丢失以及识别错误.

2)字符级别攻击. 根据预设的比例，从w中随机选择要更改的字符. 然后，对所选字符，从 3 种操作(即重复、插入、删除)中选择 1 个执行. 比例和待插入字符的设置见表 5. 字符级别攻击可以模拟用户生成的文本中的那些自然而普遍的人为错误，如打字错误、拼写错误.

3)视觉级别攻击. 与字符级别攻击类似，根据预设的比例，如表 5 所示，从w中选择要更改的英文字

Table 4 Attack Description

表 4 攻击描述

类型	具体操作	示例
无攻击		George Washington died in 1799.CDs weren't invented until 1982.
字符级别	重复	Georggge Washington diiied iiin 177799.CDs weren't innnvented until 199982.
	删除	George Washington died in 1799.CDs weren't inted un 1982.
	插入	George Washington di@ed in 1799@.CDs weren't invented until 1982.
单词级别	插入	died George Washington died in 1799.CDs 1799.CDs weren't invented hoosier state until 1982.
	删除	invented.
	替换	George cook up cook up 1982 1799.CDs go bad invented until 1982.
视觉级别		George Washin{0260}ton {0257}ied in 1799.CDs weren't {0269}nvented unti{0625} 1982.

注：下划线部分为攻击所改变的部分，花括号中的内容代表新增字符的 Unicode.

Table 5 List of Attack Methods Used for Constructing Adversarial Examples

表 5 构建对抗性示例所用攻击方法列表

攻击级别	操作	参数
字符	删除	删除次数 $c \in \{1, 2, 3\}$，概率分布为 {0.4, 0.4, 0.2}.
	重复	替换次数 $c \in \{1, 2, 3\}$，概率分布为 {0.4, 0.4, 0.2}.
	插入	插入次数 $c = 1$
视觉	替换	替换比例 $r \in \{0.1, 0.5, 0.9\}$
单词	插入	无
	删除	无
	替换	无

母. 然后用视觉上相似的字符来替换这些字母. 这些字符的数据来源于以前的工作[13]. 这种攻击可以模拟一些实际应用程序中的视觉层面的缺陷, 如 OCR.

2.4 一致性评估

本研究使用“一致性”一词来表示大模型面对语义相似的用户输入时的一致性. 由此, 设计了 2 种新颖的攻击方法来反映 2 个方面.

1)提示攻击. 在实际应用中, 不同的用户可能对相同含义的输入有不同的表达. 这些输入在语义上相似, 但形式非常不同. 对于 ChatGPT 来说, 保持与这些输入的一致性非常重要. 为了进行模拟, 选择了 5 个语句, 表示为 $\boldsymbol{t}'$, 与 $\boldsymbol{t}$ 同义, 如表 6 所示. 用不同的 $\boldsymbol{t}'$ 替换 $\boldsymbol{t}$, 创建了一个自动化的过程. 在这个过程中, 针对不同的提示测试 ChatGPT 的输出, 以衡量响应不同提示的一致性.

2)选项攻击. 为了测试大模型的性能, 对选项的顺序(包括 a, o)进行随机化. 显然改变选项顺序并不影响数据原语的语义, 但根据经验, 某些大模型在这种攻击下表现出奇怪的行为, 在实验部分中展示了这一点. 这种攻击方法模仿了实际使用中的一些并列的语法结构, 这些并列成分的顺序并不影响阅读.

2.5 可行性评估: 相对训练指数

可信性评估事实上是实证研究的副产物. 当以极端的强度、幅度对模型进行上述攻击时, 大模型可能仍然能够准确地执行. 类似于视觉问答(visual question answering)基准问题[56], 这表明目前的基准方法可能存在严重隐患. 因此, 尽管本节内容是副产物, 其重要性不应被剥夺和忽视. 特别地, 本研究提出了一个新的指数——RTI, 来衡量给定数据集在大模型相关评估中的“相对”可靠性.

形式化地, RTI 可以通过如下步骤进行计算(算法 1 展示了具体的实现):

1)采用单词级别攻击, 因为其具有通用性. 根据

Table 6 Prompts for the Multiple-option

表 6 多项选择的提示词

文本	来源
“”	空白提示
“Complete the description with an appropriate ending: ”	可用于 OPT 模型的经验提示
“You must choose the best answer from the following choices marked (A), (B), (C), (D) or (E). ”	CET 考试
“To answer the following question according to the following information.”	人工构造
“Next, I will ask you a series of questions given a description, and you will have to choose one of several candidate options that you think is correct.”	人工构造

式(4)，逐步将参数ρ从0增加到1，步幅为0.1. 每一步中，通过模型的最终响应的准确率来测试模型.

2)对每个ρ检查模型的输出. 在这个过程中，找到最小的且导致模型输出改变的ρ，记作r.

3)对数据集D中的每个$\boldsymbol{x}$计算r的期望(即RTI)，记为R_D：

$$R_D = E_{\boldsymbol{x}\in D}(r_{\boldsymbol{x}}). \tag{5}$$

算法 1. 数据集D的RTI值R_D计算算法.

输入：数据集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$；自动攻击器$g(\boldsymbol{x},\rho)$，攻击目标$\boldsymbol{x}$，概率$\rho$；模型输出$f(\boldsymbol{x},\boldsymbol{\theta})$，输入$\boldsymbol{x}$，模型参数$\boldsymbol{\theta}$.

输出：数据集D的RTI分数R_D.

① for $\boldsymbol{x}_i$ in D do

② $\quad\rho=0.1$;

③ $\quad\boldsymbol{x}_{i'}=g(\boldsymbol{x}_i,\rho)$;

④ $\quad$while $f(\boldsymbol{x}_i,\boldsymbol{\theta})=f(\boldsymbol{x}_{i'},\boldsymbol{\theta})$ do

⑤ $\quad\quad\rho=\rho+0.1$;

⑥ $\quad\quad\boldsymbol{x}_{i'}=g(\boldsymbol{x}_i,\rho)$;

⑦ $\quad$end while

⑧ $\quad r(\boldsymbol{x}_i)=\rho$;

⑨ end for

⑩ $R_D=E(R)$ where $R\sim r(\boldsymbol{x}),\boldsymbol{x}\in D$.

本研究计划获取对常见NLP任务评估的可靠性，尤其是那些与大模型相关的任务的评估. 直观地说，如果选择数据集来填充特定大模型的整体训练集，通过适当的训练流程，大模型应该已经吸收、记忆了该数据集的各个方面. 在这种情况下，逐步开发一个外部模块来帮助大模型适配该数据集涉及的任务，该过程可能会有一些问题. 换句话说，在某些情况下，大模型即使在段落样本百分百被污染的情况下也可以正确地做出反应. 这显然意味着大模型已经将信息和/或其答案编码到它们的参数记忆中. 因此，在这种情况下进一步评估与大模型有关的任务将极具误导性. 此外，本研究只是可靠性问题的一种暗示，而不是完全解释或解决这个难题. 因为，没有一条透明的路径在这种大模型(特别是ChatGPT)的原始训练集上进行逆向工程.

换句话说，RTI的分数越高，这个方法或数据集的可信性越低，反之亦然. 一个(也许有点夸张的)主张：在大模型时代，更低的RTI分数应当鼓励同行考虑从评估集中去除相应的数据集.

2.6 鲁棒性、一致性的指标

为了完成对常见大模型的测试，本研究设计了2个互补的定量指标：错误率(error rate, *ER*)、攻击成功率(attack success rate, *ASR*). 特别地，*ER*类似于传统的NLP评估，本研究通过大模型对数据原语响应的准确率来评估模型. 而*ASR*衡量了模型在测试集中保持原始输出的比例.

2个指标的公式为：

$$ER_D=\frac{|\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in D\wedge f(\boldsymbol{x}_i,\boldsymbol{\theta})\neq a_i\}|}{|D|}, \tag{6}$$

$$ASR_D=\frac{|\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in D\wedge f(\boldsymbol{x}_i,\boldsymbol{\theta})\neq f(\boldsymbol{x}_{i'},\boldsymbol{\theta})\}|}{|D|}, \tag{7}$$

其中$|\cdot|$代表集合大小.

2.7 对抗性样本的定性模式分析

除了定量的测量，本研究根据原语定性地分析大模型的行为.具体来说，主要遵循文献[57−61]中提出的方法.在获得大模型对相应输入(包括不受污染的输入和受污染的输入)的响应后，分析输入提示中哪一部分最有可能导致模型输出发生漂移.

通过词性标注、依赖关系分析、短语结构发现、句内位置分析等工具，本研究将这一部分的分析建立在单词的粒度上，它涵盖了此前提到的所有攻击手段.

更具体地说，本研究为上述信息来源设计了一个单独的机制：

1)对于结构、句内分析，考虑那些输出与预期不符的攻击示例$\boldsymbol{x}'$，即那些$f(\boldsymbol{x}',\boldsymbol{\theta})\neq f(\boldsymbol{x},\boldsymbol{\theta})$的示例，统计不同类别的扰动在$\boldsymbol{x}'$中的出现频率. 类别$l$在这里是位置标签(如头部、尾部、中间位置)或结构标签(如名词短语). 频率s_l可以被表示为：

$$s_l=\sum_{\boldsymbol{x}\in D}\frac{1}{G(\boldsymbol{x})}\sum_{w\neq w^*}h(w^*,\boldsymbol{x},\boldsymbol{x}',l), \tag{8}$$

$$G(\boldsymbol{x})=|\{w_i\mid w_i\neq w_i^*\}|,$$

$$h(w^*,\boldsymbol{x},\boldsymbol{x}',l)=\begin{cases}1, l(w^*)=l\wedge f(\boldsymbol{x},\boldsymbol{\theta})\neq f(\boldsymbol{x}',\boldsymbol{\theta}),\\ 0, \text{其他},\end{cases}$$

其中w为$\boldsymbol{x}$的元素，w, w^*分别代表$\boldsymbol{x}$, $\boldsymbol{x}'$中的单词(见式(3))，$l(w)$是用于获得单词w类别的函数，h是用于计算该示例是否导致模型漂移的函数，G是归一化函数，用于平衡攻击次数.

2)对于词性分析、依赖关系分析，通过计算它们对成功攻击的贡献来衡量这些攻击手段的有效性. 同样首先定义一个向量$\boldsymbol{c}$，该向量由计算类别集合L中每个l的归一化频率得到：

$$\boldsymbol{c}=\left[\frac{|\{w\mid\forall w,\ w^*\neq w\wedge l(w)=l_j\}|}{|\{w\mid\forall w,\ l(w)=l_j\}|}\right]_{l_j\in L}. \tag{9}$$

然后，在数据集D上训练了一个随机森林，输入特征为每个样本的$\boldsymbol{c}$，输出为二分类目标{0,1}，其中1代表$f(\boldsymbol{x},\boldsymbol{\theta})\neq f(\boldsymbol{x}',\boldsymbol{\theta})$，0代表其他情况. 最后，使用训

练好的随机森林模型来为每个类别分配其对攻击成功的重要度分数.

3 实 验

3.1 鲁棒性结果

3.1.1 主要结果

表 7 列出了针对对抗性攻击的鲁棒性测试结果. 表 8、表 9 列出了每个数据集的具体结果.

ChatGPT 表现出较低的鲁棒性. 所有攻击方式都会对 ChatGPT 产生负面影响. 具体地, 无论攻击类型, 都会导致 ChatGPT 响应的 *ER* 至少增加 2 个百分点. 此外, *ASR* 结果显示, 所有攻击都会导致模型改变至少 27 个百分点的输出选项. 同时, 在 3 个不同级别的攻击中, 模型对字符级别攻击的鲁棒性较高, 而对单词级别攻击的鲁棒性较低. 单词级别攻击的 *ER* 比字符级别的平均高出约 10 个百分点. 同时, 单词级别攻击的 *ER* 在不同的方法类型上表现出较高的方差. *ASR* 指标的结果相似. 一个可能的原因是单词级别攻击通过分词器[62]、词嵌入[63]导致编码向量产生较大变化. 作为模型的输入, 这种变化会对模型的语义理解产生较大影响.

对于各种级别的攻击, 本研究进行了详细分析.

1)字符级别. 在此级别的攻击中, 删除相较于插入影响更大. 删除的 *ER* 比插入的高 10 个百分点, 而 *ASR* 高 11 个百分点. 这一级别的攻击上, 不同方法差异并不明显. 字符级别的攻击很可能是由自然的人为错误引入的, 如拼写错误、打字错误. 更多场景下的潜在风险将在第 4 节中讨论.

2)单词级别. 该级别的 *ER*, *ASR* 保持在一个相对较高的水平. 具体而言, 删除操作的影响最大, *ER* 为 61.8%, *ASR* 为 49.8%. 单词级别攻击的鲁棒性可能会对现实世界中的应用带来潜在的风险, 如语音识别. 在这些场景下, 输入文本通常只能保证每个单词的完整性, 而无法保证上下文中单词的正确性.

3)视觉级别. 在该级别下, 单个单词的替换比例越高, *ER*, *ASR* 就越高. 然而, 模型仍然能够保持 40% 左右的准确率. 可能是因为 GPT 的 tokenizer 在转换中考虑了视觉上相似的字符并进行了重写. 视觉鲁棒性对于大模型在 OCR 和其他实际场景中是至关重要的.

Table 7 Impact of Different Attack Levels on ChatGPT

表 7 不同级别攻击对 ChatGPT 的影响

指标/%	无攻击	字符级别			单词级别			视觉级别		
		重复	删除	插入	插入	删除	替换	10%	50%	90%
ER	40.19	44.59	51.89	41.63	48.04	61.89	60.85	42.47	48.85	55.88
ASR	0.00	30.51	38.56	27.45	34.27	49.80	48.18	28.19	34.64	41.94

注: 在所有数据集上共计 39 253 条示例测试上得出结果. 视觉级别中的百分数代表攻击的词数比例.

Table 8 Impact of Each Dataset on ChatGPT Under Different Attack Levels (Measured by *ER*)

表 8 每个数据集在不同级别攻击下对 ChatGPT 的影响(通过 *ER* 进行衡量)

数据集	无攻击 *ER*/%	字符级别 *ER*/%			单词级别 *ER*/%			视觉级别 *ER*/%		
		重复	删除	插入	插入	删除	替换	10%	50%	90%
StrategyQA	29.56	31.10	37.46	31.09	35.48	46.33	46.77	31.07	35.92	43.65
AQuA	47.64	63.74	74.22	51.40	68.31	89.57	89.17	54.92	58.07	71.26
Creak	34.14	34.31	35.85	35.64	34.46	36.51	36.98	36.40	38.04	39.75
NoahQA	33.01	41.75	47.83	35.41	40.51	63.10	61.57	34.38	37.87	46.68
GSM8K	54.80	62.28	60.29	55.33	57.72	63.00	62.15	54.57	55.50	59.61
bAbi15	29.00	28.21	49.15	26.68	52.35	64.50	64.30	31.35	45.15	59.35
bAbi16	55.40	52.97	62.15	54.76	60.20	61.65	61.10	56.80	60.50	66.25
QASC	20.52	23.01	53.07	25.70	38.55	70.09	64.09	29.27	57.13	72.68
ECQA	26.94	31.48	54.92	30.91	49.02	72.11	74.68	33.91	51.23	66.27
e-SNLI	52.99	53.58	58.98	52.65	58.37	64.01	63.85	54.99	61.27	64.50
Sen-Making	19.94	23.42	34.22	22.67	31.89	59.30	50.12	26.18	41.17	49.04
QED	23.54	27.21	45.52	28.98	37.79	58.67	58.63	29.70	44.06	56.20

注: 视觉级别中的百分数代表攻击的词数比例.

Table 9 Impact of Each Dataset on ChatGPT Under Different Attack Levels (Measured by *ASR*)

表 9 每个数据集在不同级别攻击下对 ChatGPT 的影响（通过 *ASR* 进行衡量）

数据集	无攻击 *ASR*/%	字符级别 *ASR*/%			单词级别 *ASR*/%			视觉级别 *ASR*/%		
		重复	删除	插入	插入	删除	替换	10%	50%	90%
StrategyQA	29.56	25.31	31.72	24.39	30.09	42.77	42.05	25.44	31.94	40.41
AQuA	47.64	64.62	73.65	56.7	70.28	82.87	84.65	56.3	60.63	69.69
Creak	34.14	15.8	19.62	16.69	18.02	23.3	22.79	16.48	20.64	23.89
NoahQA	33.01	34.31	40.65	28.21	33.03	57.35	54.38	27.14	30.73	39.81
GSM8K	54.80	35.66	37.41	27.71	32.16	45.89	42.89	27.78	31.38	36.12
bAbi15	29.00	36.57	57.3	36.43	58	67.95	67.3	40.4	51.15	62.45
bAbi16	55.40	54.16	68.05	56.34	68.15	69.7	71.05	58.15	61.4	67.25
ECQA	26.94	16.44	42.13	16.13	38.47	63.81	68.69	20.31	38.58	56.06
e-SNLI	52.99	31.32	35.87	31.3	34.92	38.59	38.77	32.01	35.91	38.24
QASC	20.52	12.72	43.08	13.05	29.16	62.47	57.34	17.12	46.81	65.01
QED	23.54	16.56	39	16.41	28.19	53.03	52.25	20.37	36.86	51.99
Sen-Making	19.94	19.23	30.25	18.85	28.77	55.34	46.29	22.09	36.07	44.29

注：视觉级别中的百分数代表攻击的词数比例.

同样, 本研究还在其他大模型(包括 LLaMA[2], OPT[64])上进行了对比实验. 对于数据集 D, 选择了在 ChatGPT 测试中 *ER*表现最好的数据集(Sen-Making)和最差的数据集(bAbi 16)作为测试基准. 表 10、表 11 列出了模型的实验结果. 与其他模型相比, ChatGPT 在 *ER*指标上表现更好, 而在 *ASR*指标上表现更差. 出现前者的可能原因是其他模型本身的基准能力较低. 出现后者的原因是一些模型在其输出选择上具有较高的随机性和较低的置信度. 可以看出, 模型并没有完全理解进行多项选择操作这一具体意图.

3.1.2 攻击模式分析

1)从词性(part-of-speech, POS)、依赖关系的角度. 根据 2.7 节概述的方法, 本研究分析扰动词的不同词性标签、依赖标签的重要度 $\boldsymbol{c}$(式(9)). 如图 2 所示, 当目标是具有特定词性标签或依赖标签的单词时, 对 ChatGPT 的对抗攻击更容易成功. 具体地说, 对攻击成功最有利的 5 个词性标签是“名词(NOUN)”“介词(adposition, ADP)”“动词(VERB)”“限定词(determiner, DET)”“助动词(auxiliary verb, AUX)”, 依赖标签是“句子的根节点(ROOT)”“名词性主语(nominal subject, nsubj)”“介词宾语(prepositional object, pobj)”“介词修饰语(prepositional modifier, prep)”“限定词(determiner, det)”. 剩余的 5 个词性标签“形容词 (adjective, ADJ)”“代词(pronoun, PRON)”“(并列连词(coordinating conjunction, CCONJ)”“专有名词(proper noun, PROPN)”“副词(adverb, ADV)”和 5 个依赖关系标签“形容词性修饰语(adjectival modifier, amod)”“直接宾语(direct object, dobj)”“复合名词修饰(compound noun, compound)”“从句修饰语(clausal modifier, acl)”“并列成分(conjunct, conj)”, 它们的影响相对较小. 一个潜在的原因可能是这些标签在决定句子含义、结构方面起到了关键作用.

①名词、动词、形容词是表达句子中关键概念、关系的基础, 而限定词、助词则影响时态、一致性.

②“ROOT”依赖是连接所有其他词及其依赖的中心词. 类似地, “nsubj”“pobj”依赖分别与主语、宾

Table 10 Impact of Attack on Different LLMs and Datasets (Measured by *ER*)

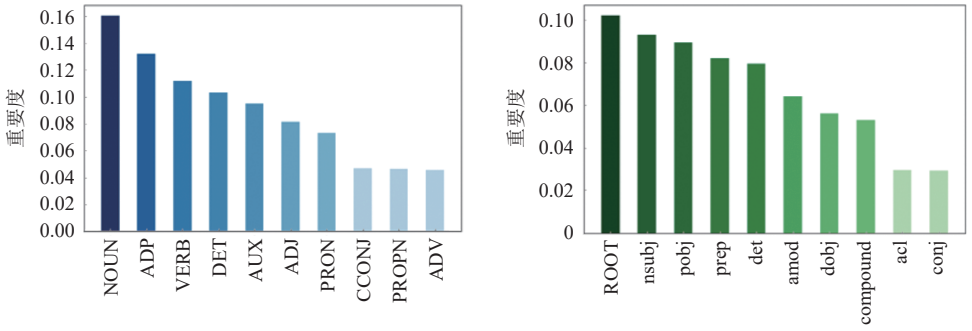
表 10 攻击对不同大模型、数据集的影响（通过 *ER* 衡量）

模型	数据集 1（bAbi16）				数据集 2（Sen-Making）			
	样本数	*ER*/%			样本数	*ER*/%		
		无攻击	单词替换	字符删除		无攻击	单词删除	字符插入
ChatGPT(175B*)	1 000	55.40	61.10	62.15	2021	19.94	59.30	22.67
OPT(1.3B)	1 000	90.00	89.45	88.87	2021	100.00	100.00	100.00
LLaMA(11.5B)	1 000	61.90	81.10	56.85	2021	56.46	64.77	54.66

Table 11 Impact of Attack on Different LLMs（Measured by *ASR*）

表 11 攻击对不同大模型的影响（通过 *ASR* 衡量）

模型	数据集 1（bAbi16）				数据集 2（Sen-Making）			
	样本数	*ASR*/%			样本数	*ASR*/%		
		无攻击	单词替换	字符删除		无攻击	单词删除	字符插入
ChatGPT(175B*)	1 000	55.40	71.05	56.34	2021	19.94	55.34	18.85
OPT(1.3B)	1 000	90.00	47.50	44.66	2021	100.00	0.00	0.00
LLaMA(11.5B)	1 000	61.90	62.75	63.74	2021	56.46	34.66	27.13



(a) 词性的攻击模式分析 (b) 依赖关系的攻击模式分析

Fig. 2 Attack Pattern analysis on part-of-speech and dependency relations

图 2 词性和依赖关系的攻击模式分析

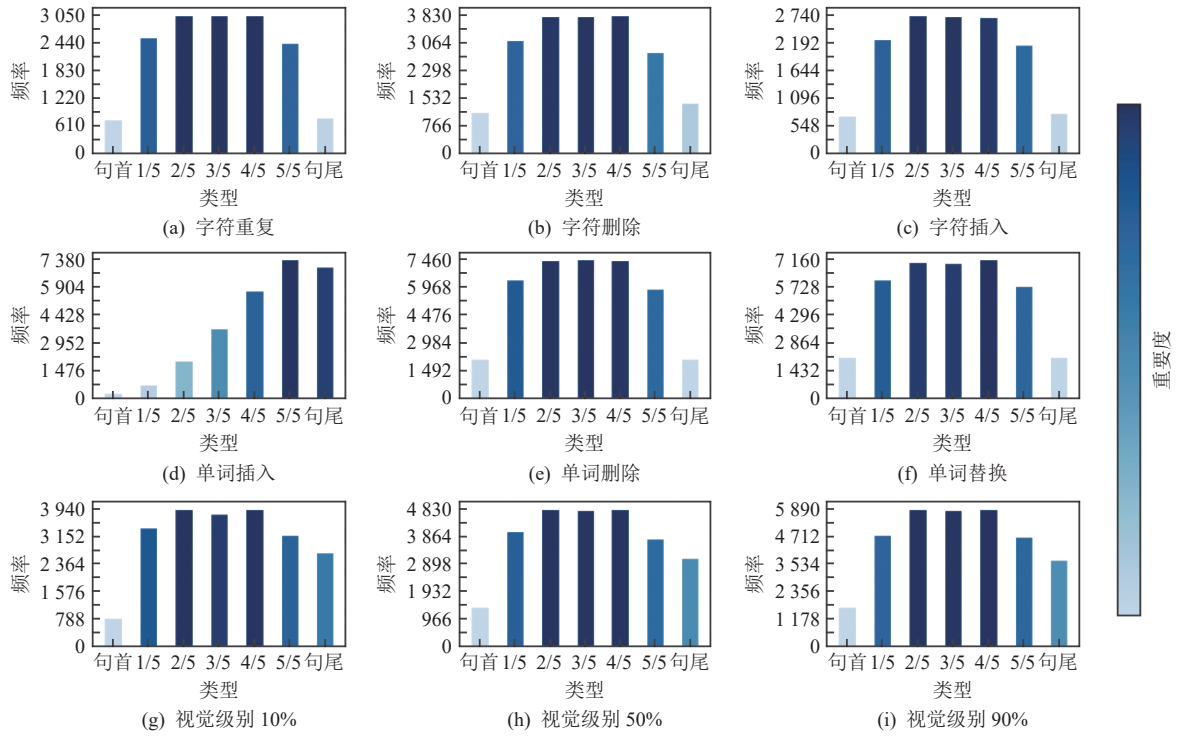
语的名词相关. “prep”“det”依赖分别与介词、限定词相关, 它们对定义、区分对象以及概念至关重要.

2)从句内位置、短语结构的角度. 根据 2.7 节中概述的方法, 本研究得到的结果如表 12、图 3 所示. 当攻击发生在某个类别 l(式(9))的词中时, 初始地响应更容易发生改变. 对于解析器结构标签, “名词短语(noun phrase, NP)”“动词短语(verb phrase, VP)”“介词短语(prepositional phrase, PP)”“疑问名词短语(wh-noun phrase, WHNP)”“形容词短语(adjective phrase, ADJP)”的影响最大. 而对于位置来说, 攻击中间的词比攻击两端的词的影响更大. “副词性短语修饰语(adverbial phrase modifier, ADVP)”“数量短语(quantifier phrase, QP)”“小品词(particle, PRT)”“疑问介词短语(wh-prepositional phrase, WHPP)”“疑问形容词性短语(wh-adjectival phrase, WHADJP)”的影响较小. 现在将分别进行详细的分析. ① l 在“NP”“VP”“PP”中的影响权重在所有数据集上均高于其他类别. 同时, 对于类别的改变, 单词级别的替换是最敏感的方法. 一个可能的原因是语言模型的预测是基于历史文本生成下一个单词的概率. 而名词、动

Table 12 Parser Pattern Analysis

表 12 解析器模式分析

攻击类型	攻击模式	出现频率 s_l									
		NP	VP	PP	WHNP	ADJP	ADVP	QP	PRT	WHPP	WHADJP
字符级别	重复	2 846.95	1 382.41	1 071.72	89.78	91.68	71.75	29.76	22.87	5.79	8.47
	删除	4 297.89	1 814.48	771.96	228.64	188.36	155.89	34.77	26.72	3.20	10.12
	插入	2 598.74	1 313.87	950.05	90.10	87.16	69.68	20.80	20.82	2.98	5.86
单词级别	插入	807.43	310.61	226.88	37.65	31.37	23.02	8.97	5.13	1.77	2.03
	删除	7 452.55	3 359.09	2 216.4	373.21	221.78	215.27	64.14	42.37	16.31	13.46
	替换	8 671.48	3 537.06	2 372.13	394.65	284.26	230.74	76.84	52.62	16.84	19.61
视觉级别	10%	3 629.25	2 026.04	1 447.34	135.76	138.66	108.13	31.71	33.69	6.22	7.56
	50%	4 507.03	2 483.14	1 741.83	262.91	177.83	163.08	31.99	43.46	9.5	10.16
	90%	5 512.11	3 088.87	2 085.59	358.46	226.42	210.19	36.28	50.31	12.24	13.43



注：视觉级别中的百分数代表攻击的词数比例.

Fig. 3 Attack pattern analysis on positions

图 3 关于位置的攻击模式分析

词、介词短语对句子结构有很大影响，因此会影响正确单词的预测概率，并一定程度上导致输出的不确定性提高. ②攻击中间单词的影响明显高于攻击头部或尾部单词. 影响均匀分布在句子的中间部分，而具体位置的影响不大. 在一些方法中，攻击尾部单词的影响可能比攻击头部更大. 可能原因有 2 点：

其一，句子的中间部分涉及大量单词，因为段落长度相对较长.

其二，模型中不同位置的注意力权重有所不同.

因此，扰动这些单词可能会显著影响对原始上下文、问题的理解，这些最容易受到攻击的模式是可识别的. 未来的研究可以关注开发更具鲁棒性、更安全的大模型，以抵御这些类型的威胁.

3.2 一致性结果

图 4 显示了 ChatGPT 在不同 o, a 的顺序以及不同提示上多次运行的一致性(式(1)). 结果通过箱形图展示，此处使用准确率的标准差来表示结果的一致性.

ChatGPT, LLaMA 在面对不同的提示时，都有可能出现性能不一致、不稳定的情况. 在不同提示下，LLaMA, ChatGPT 的平均标准偏差分别达到 11.9%、1.9%. 然而，尽管在即时工程方面做出了许多努力，但在实现稳定、可靠的大模型性能方面仍然存在重大挑战，特别是在用户提示或问题高度多样化的现实情况下.

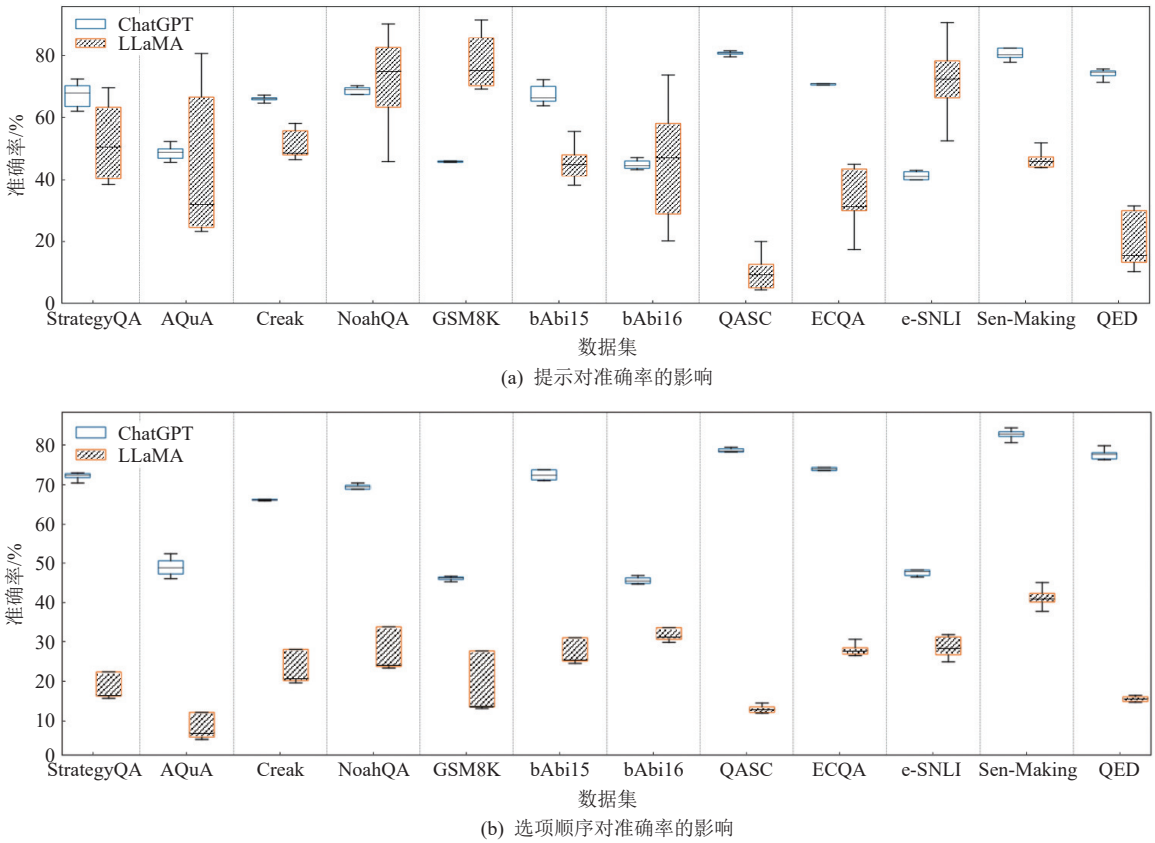
ChatGPT 对 o, a 指令的敏感度较低，这表明它能够熟练地理解选项中的文本内容，而不是对特定的选项编号有偏见. 具体而言，在 ChatGPT 上的平均标准偏差仅为 1.13%. 通过比较发现 LLaMA 模型在 6 个不同的选项顺序中，无论第 1 个选项的内容是什么，都显著倾向于选择第 1 个选项，尽管其准确率较低 (仅为 0.98%). 这表明该模型可能不具备在多选项测试中执行任务的能力. 本研究推断，仅通过预训练(如 LLaMA)的模型不足以拥有理解选项实质内容的能力，需要额外的调整(如指令或提示词微调)来赋予模型类似能力.

3.3 可信性结果

如果设置一个更高的概率参数 z(参见 2.3 节中的详细信息)，模型更有可能改变原始选项. 基于 3.2 节实验得出的这一结论，本研究假设该参数可能与数据是否经过训练有关. 因此，本研究提出 RTI，在数据集 D 上表示为 R_D 指标(见 2.5 节中的详细内容). 同时，对上述数据集进行了抽样测试，结果如表 13 所示.

1)不同数据集的 RTI 平均值不同. 因此，通过 RTI 评估不同数据集之间的差异是可能的.

2)提示词 $\boldsymbol{t}$ 的结构越复杂，输入的 o(式(1))越多，



(a) 提示对准确率的影响

(b) 选项顺序对准确率的影响

Fig. 4 Impact of prompts and option orders on accuracy

图 4 提示和选项顺序对准确率的影响

Table 13 RTI Test

表 13 RTI 测试

数据集		不同攻击模式的 RTI			
名称	样本数	单词插入	单词删除	单词替换	平均
StrategyQA	100	0.538	0.421	0.408	0.456
AQuA	100	0.199	0.157	0.169	0.175
Creak	100	0.702	0.654	0.649	0.668
NoahQA	508	0.519	0.323	0.337	0.393
GSM8K	442	0.479	0.350	0.360	0.396
bAbi15	400	0.277	0.213	0.220	0.237
bAbi16	100	0.206	0.200	0.196	0.201
ECQA	99	0.547	0.391	0.425	0.454
e-SNLI	99	0.375	0.316	0.346	0.346
QASC	99	0.543	0.391	0.393	0.442
QED	99	0.618	0.499	0.486	0.534
Sen-Making	99	0.56	0.425	0.399	0.461

RTI 平均值越低. 一般情况下, 样本$\boldsymbol{x}$的结构越复杂, 模型对$\boldsymbol{x}$的拟合就越困难. 这一理论结果和实际结果是一致的. 如, AQuA 是一个十分复杂的数据集, 任务类型为长文本数学推理. R_{AQuA}的值为 0.175, 处于较低水平, 说明模型没有很好地拟合这一数据集.

3)在训练集和验证集上的 RTI 平均值较测试集相对更高. 这一结果与预期一致. 即使 RTI 值高达 0.45, 训练集数据(此处为 StrategyQA)也能被有效地识别出来. 因此, RTI 对于评估数据集是否被用于模型训练具有一定的价值. RTI 可以为各种大模型评估所使用数据集的可靠性提供一个可用的定量参考.

4 讨 论

4.1 大模型应用程序安全

在聊天机器人、大模型软件集成等传统应用场景下, 大模型也可能存在风险. 如, 人类可能犯下的各种常见错误(包括打字错误、拼写错误等), 会对条件文本造成字符级别的干扰. 尽管大模型在字符级别对抗性攻击下表现出相对较高的鲁棒性, 但模型实际输出的结果内容仍可能存在显著波动, 这可能会对用户的体验产生相当大的影响, 也可能增加用户的检查成本.

将其他模式的模型整合到大模型中可能会带来更多的对抗性攻击风险. 大模型应用的趋势之一是与其他形式的模型集成, 如基于结构化数据、图像、

语音、其他感官输入的模型. 虽然这种集成可以显著地利用大模型的能力，但它也为对抗性攻击开辟了新的途径. 下面，本研究提供一些具体的案例来说明潜在的风险：

1)将 OCR 集成到大模型中，使机器能够理解、解释包含于图像中的文本. 然而，这其中也存在一些脆弱性，模型可以被视觉上相似的字符攻击，导致对文本的错误解释，这可能会产生严重的后果，尤其是在金融或者医药等行业，因为在这些行业中，即使是很小的错误也可能造成巨大损害. 如，考虑一个 OCR 算法在处理财务文档时将“0”误认为是“O”. 这种错误可能导致重大的经济损失或法律影响. 类似地，考虑在医疗环境场景下的应用，由于 OCR 识别错误造成的问题可能导致错误的处方、误诊或者不正确的医疗报告，甚至危及生命.

2)将语音识别(相当于语音转文本)系统集成到大模型中可以帮助机器实现更高效的人机交互. 语音到文本的实质是对声学特征进行建模、分类，从而实现语音信号的识别. 在这种情况下，模型通常能够通过模式匹配有效地识别完整的单个单词(无论正确与否)，但可能会产生上下文错误，这对应于本研究实验中的 3 种单词级别攻击形式. 从结果中发现大模型(如 ChatGPT)面对单词级别攻击时的鲁棒性较低，从而影响了其在各个领域的应用. 如，在配备智能语音识别系统的自动驾驶车辆中，如果由于识别错误而导致模型错误，可能会导致极其严重的后果，直接危及乘客的生命安全.

总而言之，由于将大模型集成到其他系统具有显著的不可控性，研究人员和开发人员必须意识到对抗性攻击的潜在危害和影响，建立足够的保障机制来对抗潜在的攻击.

4.2 大模型应用程序安全

许多基于开放数据集的评估方法不够可靠，无法推广. 目前，大多数 ChatGPT 的评估方法使用一定的度量来衡量它在某些数据集上的性能，但这并不总是可靠的. 由于 ChatGPT 没有公开披露其使用的训练数据，因此很难确定用于评估的数据在训练过程中是否被模型记忆[65]. 另外，本研究发现 ChatGPT 在一些极端情况下(如删除所有段落)仍然可以完成问题场景并提供正确的答案，这表明这些样本很可能已经被模型记住了. 因此，使用这样的样本来衡量 ChatGPT 的性能是没有说服力的. 这一问题需要整个社会进行更多的思考，而 RTI 可以为相对训练概率提供一些参考. 此外，可以计算一系列不同数据集的 RTI 分数，并通过一些统计方法获得基线数据，作为衡量数据训练情况的绝对指标. 希望能对建立更加可靠的评价体系有所帮助.

4.3 大模型的对抗性训练范式

适合大模型的对抗性训练范式需要被进一步开发. 针对大模型的对抗鲁棒性训练的现有研究主要集中在预训练、微调阶段. 然而，大模型的多阶段训练范式[66]提供了新的挑战，包括用预训练、上下文学习代替微调的高成本，这阻碍了传统对抗性训练技术的可迁移性. 据调查，目前还没有针对大模型高效对抗性训练方法的探索. 因此，研究者相信这是未来相关研究的一个很有前景的方向.

作者贡献声明：叶文涛完成了设计实现和论文撰写；胡家齐协助完成测试实验并帮助撰写论文；王皓波对论文提出意见并修改论文；陈刚和赵俊博对实验设计给予指导.

参 考 文 献

[1] Shu Wentao, Li Ruixiao, Sun Tianxiang, et al. Large language models: Principles, implementation, and progress[J]. Journal of Computer Research and Development, 2024, 61(2): 351−361 (in Chinese)
(舒文韬, 李睿潇, 孙天祥, 等. 大型语言模型: 原理、实现与发展[J]. 计算机研究与发展, 2024, 61(2): 351−361)

[2] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[3] Chen Huimin, Liu Zhiyuan, Sun Maosong. The social opportunities and challenges in the era of large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1094−1103 (in Chinese)
(陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094−1103)

[4] Zhong Qihuang, Ding Liang, Liu Juhua, et al. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT[J]. arXiv preprint, arXiv: 2302.10198, 2023

[5] Qin Chengwei, Zhang A, Zhang Zhuosheng, et al. Is ChatGPT a general-purpose natural language processing task solver?[C]// Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 1339–1384

[6] Huang Fan, Kwak H, An Jisun. Is ChatGPT better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech[C]// Companion Proc of the ACM Web Conf 2023. New York: ACM, 2023: 294−297

[7] Kocoń J, Cichecki I, Kaszyca O, et al. ChatGPT: Jack of all trades, master of none[J]. Information Fusion, 2023, 99: 101861

[8] Yang Xianjun, Li Yan, Zhang Xinlu, et al. Exploring the limits of ChatGPT for query or aspect-based text summarization[J]. arXiv preprint, arXiv: 2302.08081, 2023

[9] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems[C]// Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 2021–2031

[10] Gao Ji, Lanchantin J, Soffa M L, et al. Black-box generation of adversarial text sequences to evade deep learning classifiers[C]//Proc of the 2018 IEEE Security and Privacy Workshops (SPW). Piscataway, NJ: IEEE, 2018: 50−56

[11] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation[J]. arXiv preprint, arXiv: 1711.02173, 2017

[12] Heigold G, Neumann G, van Genabith J. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse?[C]//Proc of the 13th Conf of the Association for Machine Translation in the Americas (Volume 1: Research Track). Stroudsburg, PA: Association for Machine Translation in the Americas, 2018: 68−80

[13] Eger S, Şahin G G, Rücklé A, et al. Text processing like humans do: Visually attacking and shielding NLP systems[C]//Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: ACL, 2019: 1634−1647

[14] Papernot N, McDaniel P, Swami A, et al. Crafting adversarial input sequences for recurrent neural networks[C]//Proc of the 35th IEEE Military Communications Conf (MILCOM 2016). Piscataway, NJ: IEEE, 2016: 49−54

[15] Alzantot M, Sharma Y, Elgohary A, et al. Generating natural language adversarial examples[C]//Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2890–2896

[16] Zhao Zhengli, Dua D, Singh S. Generating natural adversarial examples[J]. arXiv preprint, arXiv: 1710.11342, 2017

[17] Iyyer M, Wieting J, Gimpel K, et al. Adversarial example generation with syntactically controlled paraphrase networks[C]// Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA: ACL , 2018: 1875–1885

[18] Li Linyang, Shao Yunfan, Song Demin, et al. Generating adversarial examples in Chinese texts using sentence-pieces[J]. arXiv preprint, arXiv: 2012.14769, 2020

[19] Li Jinfeng, Ji Shouling, Du Tianyu, et al. Textbugger: Generating adversarial text against real-world applications[J]. arXiv preprint, arXiv: 1812.05271, 2018

[20] Zang Yuan, Qi Fanchao, Yang Chenghao, et al. Word-level textual adversarial attacking as combinatorial optimization[C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6066–6080

[21] Ebrahimi J, Rao A, Lowd D, et al. Hotflip: White-box adversarial examples for text classification[C]// Proc of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: ACL, 2018: 31–36

[22] Li Linyang, Ma Ruotian, Guo Qipeng, et al. Bert-attack: Adversarial attack against BERT using BERT[C]// Proc of the 2020 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2020: 6193–6202

[23] Wallace E, Feng S, Kandpal N, et al. Universal adversarial triggers for attacking and analyzing NLP[C]// Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 2153–2162

[24] Garg S, Ramakrishnan G. Bae: BERT-based adversarial examples for text classification[C]// Proc of the 2020 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2020: 6174–6181

[25] Altinisik E, Sajjad H, Sencar H T, et al. Impact of adversarial training on robustness and generalizability of language models[C]// Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA: ACL, 2023: 7828–7840

[26] Moradi M, Samwald M. Evaluating the robustness of neural language models to input perturbations[C]// Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 1558–1570

[27] Stolfo A, Jin Zhijing, Shridhar K, et al. A causal framework to quantify the robustness of mathematical reasoning with language models[C]// Proc of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2023: 545–561

[28] Zhang Yunxiang, Pan Liangming, Tan S, et al. Interpreting the robustness of neural NLP models to textual perturbations[C]//Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg, PA: ACL, 2022: 3993−4007

[29] Chen Xuanting, Ye Junjie, Zu Can, et al. Robustness of GPT large language models on natural language processing tasks[J]. Journal of Computer Research and Development, 2024, 61(5): 1128−1142 (in Chinese)
(陈炫婷, 叶俊杰, 祖璨, 等. GPT 系列大语言模型在自然语言处理任务中的鲁棒性[J]. 计算机研究与发展, 2024, 61(5): 1128−1142)

[30] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: ACL, 2019: 4171–4186

[31] Yang Zhilin, Dai Zihang, Yang Yiming, et al. XLNET: Generalized autoregressive pretraining for language understanding[C]// Proc of the 33rd Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2019: 5753−5763

[32] Wang J, Hu X, Hou W, et al. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective[J]. arXiv preprint, arXiv: 2302.12095, 2023

[33] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks:

Reliable attacks against black-box machine learning models[J]. arXiv preprint, arXiv: 1712.04248, 2017

[34] Neekhara P, Hussain S, Dubnov S, et al. Adversarial reprogramming of text classification neural networks[C]// Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 5216–5225

[35] Hambardzumyan K, Khachatrian H, May J. Warp: Word-level adversarial reprogramming[C]// Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 4921–4933

[36] Huckelberry J, Zhang Yuke, Sansone A, et al. TinyML security: Exploring vulnerabilities in resource-constrained machine learning systems[J]. arXiv preprint, arXiv:2411.07114, 2024

[37] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[38] Geva M, Khashabi D, Segal E, et al. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 346−361

[39] Ling Wang, Yogatama D, Dyer C, et al. Program induction by rationale generation: Learning to solve and explain algebraic word problems[C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2017: 158−167

[40] Onoe Y, Zhang M J, Choi E, et al. Creak: A dataset for commonsense reasoning over entity knowledge[J]. arXiv preprint, arXiv: 2109.01653, 2021

[41] Zhang Qiyuan, Wang Lei, Yu Sicheng, et al. NoahQA: Numerical reasoning with interpretable graph question answering dataset[C]// Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, PA: ACL, 2021: 4147–4161

[42] Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems[J] arXiv preprint, arXiv: 2110.14168, 2021

[43] Dodge J, Gane A, Zhang Xiang, et al. Evaluating prerequisite qualities for learning end-to-end dialog systems[J]. arXiv preprint arXiv: 1511.06931, 2015

[44] Gaunt A L, Johnson M A, Riechert M, et al. Ampnet: Asynchronous model-parallel training for dynamic neural networks[J]. arXiv preprint, arXiv: 1705.09786, 2017

[45] Khot T, Clark P, Guerquin M, et al. QASC: A dataset for question answering via sentence composition[C]//Proc of the 34th AAAI Conf on Artificial Intelligence, vol 34, no 05. Palo Alto, CA: AAAI, 2020: 8082−8090

[46] Aggarwal S, Mandowara D, Agrawal V, et al. Explanations for commonsenseqa: New dataset and models[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 3050−3065

[47] Camburu O M, Rocktäschel T, Lukasiewicz T, et al. e-SNLI: Natural language inference with natural language explanations[C]// Proc of the 32nd Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2018: 9560−9572

[48] Wang Cunxiang, Liang Shuailong, Zhang Yue, et al. Does it make sense? And why? A pilot study for sense making and explanation[C]// Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4020–4026

[49] Lamm M, Palomaki J, Alberti C, et al. QED: A framework and dataset for explanations in question answering[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 790−806

[50] Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging[J]. Computational Linguistics, 1995, 21(4): 543−565

[51] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]// Proc of the 34th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2020: 1877−1901

[52] Wang A, Singh A, Michael J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding[C]// Proc of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Stroudsburg, PA: ACL , 2018: 353–355

[53] Wang Boxin, Xu Chejian, Wang Shuohang, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models[J]. arXiv preprint, arXiv: 2111.02840, 2021

[54] Nie Yixin, Williams A, Dinan E, et al. Adversarial NLI: A new benchmark for natural language understanding[C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4885–4901

[55] Miller G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39−41

[56] Jabri A, Joulin A, van der Maaten L. Revisiting visual question answering baselines[C]// Proc of the 14th European Conf on Computer Vision. Cham: Springer, 2016: 727−739

[57] Nguyen-Son H Q, Thao T P, Hidano S, et al. Identifying adversarial sentences by analyzing text complexity[C]//Proc of the 33rd Pacific Asia Conf on Language, Information and Computation. Tokyo: Waseda Institute for the Study of Language and Information, 2019: 182−190

[58] Goodman D, Zhonghou L, et al. Fastwordbug: A fast method to generate adversarial text against NLP applications[J]. arXiv preprint, arXiv: 2002.00760, 2020

[59] Zheng Xiaoqing, Zeng Jiehang, Zhou Yi, et al. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples[C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6600–6610

[60] Xue Mingfu, Yuan Chengxiang, Wang Jian, et al. DPAEG: A dependency parse-based adversarial examples generation method for intelligent Q&A robots[J]. Security and Communication Networks, 2020, 2020(1): 5890820

[61] Wang Wenqi, Wang Run, Wang Lina, et al. Towards a robust deep neural network against adversarial texts: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(3):

3159−3179

[62] Rai A, Borah S. Study of various methods for tokenization[C]// Applications of Internet of Things: Proc of ICCCIOT 2020. Singapore: Springer, 2021: 193−200

[63] Almeida F, Xexéo G. Word embeddings: A survey[J]. arXiv preprint, arXiv: 1901.09069, 2019

[64] Zhang S, Roller S, Goyal N, et al. OPT: Open pre-trained transformer language models[J]. arXiv preprint, arXiv: 2205.01068, 2022

[65] Ye Wentao, Hu Jiaqi, Li Liyao, et al. Data contamination calibration for black-box LLMs[C]// Findings of the Association for Computational Linguistics: ACL 2024. Stroudsburg, PA: ACL, 2024: 10845–10861

[66] Zha Liangyu, Zhou Junlin, Li Liyao, et al. TableGPT: Towards unifying tables, nature language and commands into one GPT[J]. arXiv preprint, arXiv: 2307.08674, 2023



Ye Wentao, born in 2001. PhD candidate. His main research interests include large language models, artificial intelligence safety, and natural language processing.

叶文涛, 2001 年生. 博士研究生. 主要研究方向为大语言模型、人工智能安全、自然语言处理.



Hu Jiaqi, born in 2002. Undergraduate. His main research interests include natural language processing and large language models.

胡家齐, 2002 年生. 本科生. 主要研究方向为自然语言处理、大语言模型.



Wang Haobo, born in 1996. PhD, professor, PhD supervisor. His main research interests include machine learning and data mining, especially on weakly-supervised learning and large language models.

王皓波, 1996 年生. 博士, 研究员, 博士生导师. 主要研究方向为机器学习和数据挖掘, 特别是弱监督学习和大语言模型.



Chen Gang, born in 1973. PhD, professor, PhD supervisor. Member of IEEE, ACM, standing member of the CCF Database Professional Committee. His main research interests include database management technology, intelligent computing based big data, and massive Internet systems.

陈　刚, 1973 年生. 博士, 教授, 博士生导师, IEEE, ACM 会员, CCF 数据库专业委员会常务委员. 主要研究方向为数据库管理技术、大数据智能计算、大规模互联网系统.



Zhao Junbo, born in 1992. PhD, professor, PhD supervisor. His main research interests include large language models, table pre-training, machine learning, and AI+X.

赵俊博, 1992 年生. 博士, 研究员, 博士生导师. 主要研究方向为大语言模型、表格预训练、机器学习、AI+X.