计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(12): 2993−3001, 2025

基于长距离上下文的大语言模型预训练数据检测方法

张伟超[1,2,3] 张儒清[1,2] 郭嘉丰[1,2] 范意兴[1,2]

[1](中国科学院网络数据科学与技术重点实验室(中国科学院计算技术研究所) 北京 100190)

[2](中国科学院大学 北京 100049)

[3](中关村实验室 北京 100094)

(zhangweichao22z@ict.ac.cn)

Detecting Pretraining Data for Large Language Models via Long-Range Context

Zhang Weichao[1,2,3], Zhang Ruqing[1,2], Guo Jiafeng[1,2], and Fan Yixing[1,2]

[1](*CAS Key Laboratory of Network Data Science and Technology* (*Institute of Computing Technology*, *Chinese Academy of Sciences*), *Beijing* 100190)

[2](*University of Chinese Academy of Sciences*, *Beijing* 100049)

[3](*Zhongguancun Laboratory*, *Beijing* 100094)

Abstract Pretraining data detection aims to determine whether a piece of text belongs to the pretraining data of a large language model (LLM) when the model's pretraining data are not publicly disclosed, which can be used to audit whether the pretraining data usage complies with legal regulations. Existing methods generally assume that an LLM tends to assign higher token probabilities to training texts compared with non-training texts, and thus identify texts with high probabilities as training texts. However, due to the significant overlap in fragments between training texts and non-training texts, an LLM may also assign relatively high token probabilities to non-training texts, which makes existing methods prone to misclassifying non-training texts as training texts. Inspired by research on the memorization capabilities of LLMs, we propose a novel method to mitigate this issue by comparing the token probabilities given the full context with those given the short-range context, and the contribution of long-range context to token probability increase is then computed, with a higher contribution indicating a greater likelihood of the text being part of the pretraining data. The key idea is that when an LLM predicts the token probabilities for training texts, the contribution of long-range context to token probability increase is greater than that for non-training texts. Experimental results on multiple datasets demonstrate the effectiveness of the proposed method. The code is available at https://github.com/zhang-wei-chao/Long-Range-Context-for-PDD.

Key words large language model; model memorization capability; training data copyright; membership inference attack; data contamination

摘 要 预训练数据检测方法旨在大语言模型的预训练数据未公开时,检测某段给定的文本是否属于该模型的预训练数据,可用于审查大语言模型的预训练数据的使用过程是否符合法律法规.现有方法通常认为大语言模型对训练文本的词元概率在整体上比非训练文本的高,并基于此判定具有高预测概率的文

收稿日期: 2024-10-10; **修回日期**: 2025-04-07

基金项目: 国家自然科学基金项目(62472408, 62372431); 中国科学院战略性先导科技专项(XDB0680102); 国家重点研发计划项目(2023YFA1011602, 2021QY1701); 中国科学院青年创新促进会会员项目(2021100)

This work was supported by the National Natural Science Foundation of China (62472408, 62372431), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB0680102), the National Key Research and Development Program of China (2023YFA1011602, 2021QY1701), and the Youth Innovation Promotion Association Member Project of Chinese Academy of Sciences (2021100).

通信作者: 张儒清(zhangruqing@ict.ac.cn)

本为训练文本. 然而, 由于训练文本和非训练文本之间存在着大量的短片段重叠现象, 导致模型对非训练文本的词元概率也可能比较高, 使得现有方法容易将非训练文本误检为训练文本. 受大语言模型的记忆能力研究启发, 通过对比给定全部上下文时的词元概率与给定短距离上下文时的词元概率之间的差异, 计算得到长距离上下文对词元概率提升的贡献度, 并认为贡献度越大的文本更可能是训练文本, 进而缓解短片段重叠现象对检测的不利影响. 其核心思想在于, 大语言模型在预测训练文本中词元的概率时, 距离当前词元较远的上下文对词元概率提升的贡献度, 会比在预测非训练文本中词元的概率时的贡献度更大. 在多个公开数据集上的实验结果表明该方法的有效性.

关键词 大语言模型; 模型记忆能力; 训练数据版权; 成员推理攻击; 数据污染

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440780 **CSTR:** 32373.14.issn1000-1239.202440780

近年来, 大语言模型(large language model, LLM)的预训练数据量急剧增加. 然而, 模型开发商却越来越不愿意公开他们在预训练大语言模型时所使用的数据信息[1-5]. 这种数据不透明性引发一系列的担忧, 如开发商可能使用未经授权的隐私数据或受版权保护的书籍数据来预训练大语言模型[6-7]. 此外, 大语言模型可能存在数据污染问题[8-9], 进而影响模型最终的性能评测结果. 为此, 本文旨在研究大语言模型的预训练数据检测问题.

预训练数据检测作为成员推理攻击(membership inference attack)[10]的一种实例, 是指在某个大语言模型的预训练数据未公开的情况下, 检测某段给定的文本是否属于该模型的预训练数据. 相比于传统的机器学习模型, 大语言模型的预训练数据规模庞大且每条数据只用于训练模型1次, 导致模型对数据的记忆有限, 使得问题的解决更加困难[11-12]. 近期有一些针对该问题的方法[11,13]被提出, 它们通常认为大语言模型对一段文本的词元概率①分布(token probability distribution), 蕴含可以揭示该文本是否被用于模型训练的信息. 如文献[11]提出的Min-K% Prob方法认为, 大语言模型对一段非训练文本的词元概率分布, 与一段训练文本相比, 会有更多的词元得到较低的概率. 也就是认为, 具有较高预测概率的文本更有可能是训练文本.

然而, 基于词元概率来判断一段文本是否被用于模型训练, 容易将非训练文本误检为训练文本, 导致检测结果的假阳性高[14]. 文献[12]通过实验分析发现, 训练文本和非训练文本之间存在着大量的短片段重叠现象. 如在大型公开语料库Pile[15]的子数据集Wikipedia, ArXiv, PubMed Central的测试数据中, 分别有32.5%, 39.3%, 41.0%的7-gram片段存在于对应的训练数据中. 这一重叠比例在片段长度更小时只会更高. 因此, 非训练文本会因为存在与训练文本重叠的短片段, 导致大语言模型在预测非训练文本中词元的概率时, 由于训练时见过距离当前词元较近的上下文(短距离上下文), 也能据此给出当前词元较高的预测概率, 进而使得基于词元概率的预训练数据检测方法将非训练文本误检为训练文本.

近年来, 有关大语言模型的记忆能力研究发现, 将一段训练文本分为前缀和后缀2个半段, 随着输入模型的部分前缀长度增加, 模型输出相应完整后缀的概率会近线性提升[16], 也就是模型对该后缀的词元概率出现提升. 由于距离后缀较远的前缀片段理应与后缀的语义相关性较低, 因此这种近线性提升的现象, 可能是因为模型对训练文本的记忆[13]导致的. 受此启发, 我们提出一种新的预训练数据检测方法. 其核心思想是: 模型在预测训练文本中词元的概率时, 距离当前词元较远的上下文(长距离上下文)对词元概率提升的贡献度, 会比非训练文本中的贡献度更大. 图1通过训练文本与非训练文本中, 长距离上下文对词元概率提升的贡献度来帮助理解该思想.

与此前的研究[11-13]类似, 本文关注黑盒场景下的预训练数据检测任务, 也就是将待检测的大语言模型视为一个黑盒模型, 只能利用模型预测一段文本中词元的概率. 至于模型的内部状态信息, 如权重和网络结构, 是不可获取的. 因此, 在给定待检测的文本和大语言模型之后, 我们将长距离上下文对词元概率提升的贡献度量化为: 首先计算模型在给定全

① 除非显式说明, 否则词元概率指的是, 给定文本序列 $x = x_1x_2\cdots x_n$ 和语言模型 M, M 对词元 x_i 在给定 x_i 全部上下文时的概率 $p(x_i|x_1x_2\cdots x_{i-1};M)$.

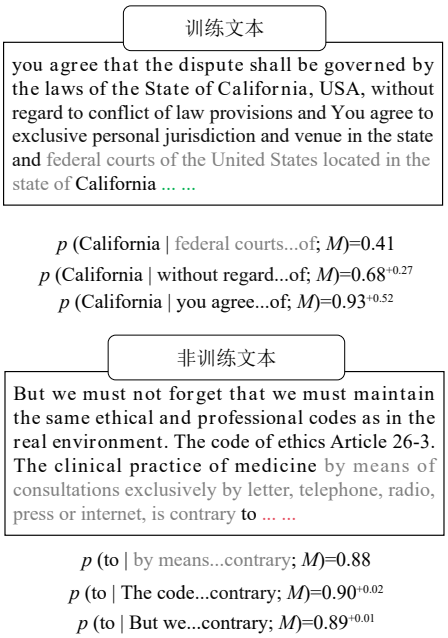


Fig. 1 An illustrative example to show the contribution of long-range context to token probability increase

图 1 长距离上下文对词元概率提升的贡献度的样例展示

部上下文时的词元概率，接着计算模型在给定短距离上下文时的词元概率，然后通过计算 2 种情况下的词元概率之差，将差值在给定全部上下文时的词元概率中的占比作为贡献度. 最后基于贡献度推导出检测分数，用于判断给定文本是否属于模型的预训练数据. 因此，我们的方法更关注长距离上下文对词元概率的影响，恰好避免了因为训练文本和非训练文本之间存在的短片段重叠现象对检测的不利影响. 这是我们的方法比基于词元概率的预训练数据检测方法更加高效的原因.

通过在多个大语言模型上使用公开的基准数据集 WikiMIA[11] 和 MIMIR[12] 进行实验，并与多种代表性的基线方法对比，结果表明我们的方法显著优于先前的方法. 如在评估指标 AUC 分数和 TPR@5%FPR 上，我们的方法在检测 LLaMA-13B 时，比此前最好的方法 Min-K% Prob 分别高出 6.9 个百分点和 12.3 个百分点.

本文的主要贡献有 3 点：

1）基于训练文本与非训练文本之间存在的短片段重叠现象，指出当前的预训练数据检测方法易将非训练文本误检为训练文本的问题；

2）受大模型的记忆能力研究启发，提出一种基于长距离上下文的预训练数据检测方法，缓解文本重叠现象对检测的不利影响；

3）通过在 WikiMIA 和 MIMIR 数据集上的对比实验，证实本文所提方法的有效性.

1 相关工作

本节将对与本文相关的研究工作进行梳理，包括成员推理攻击和大语言模型的预训练数据检测方法.

1.1 成员推理攻击

成员推理攻击是评估机器学习模型隐私属性时常用的威胁模型. 成员推理攻击最早由文献 [10] 提出，其目标是推理某个特定样本是否属于模型的训练数据. 早期的成员推理攻击研究对象主要是传统的机器学习模型[17-19] 和微调语言模型[20-22]，这类模型通常是在小数据集上进行过多次迭代训练. 近期，出于对大语言模型相关问题的研究，如大语言模型的记忆现象[23]、数据污染[24] 等问题，使得在大语言模型上进行成员推理的问题开始得到关注[11,25-26]，即大语言模型的预训练数据检测问题. 与传统的成员推理攻击相比，预训练数据检测问题的研究对象是在海量数据上仅进行过 1 次训练的大语言模型,这使得预训练数据检测问题更具有挑战性[11-12].

1.2 预训练数据检测方法

预训练数据检测方法旨在检测某段给定的文本是否属于大语言模型的预训练数据. 根据大语言模型提供的访问条件，现有的预训练数据检测方法可分为白盒场景和黑盒场景 2 类. 1）白盒场景. 假设可以访问大语言模型的内部状态信息，如权重和网络结构. 2）黑盒场景. 假设只能查询大语言模型对给定文本的词元概率，即将文本片段输入模型，可以获取模型预测该片段之后的下一个词元的概率.

由于大语言模型的内部状态信息在实际应用[1-2] 中通常无法访问，导致面向白盒场景设计的预训练数据检测方法应用范围较窄，因此关于白盒场景下的检测方法研究较少. 文献 [27] 提出一种使用探测技术进行预训练数据检测的方法. 该方法认为，训练文本在大语言模型内部激活层中的表示与非训练文本的不同，因此可以将该表示作为检测一段文本是否被用于模型训练的信号.

目前大多数研究工作聚焦在黑盒场景上，这类方法通常认为大语言模型对一段文本的词元概率分布，蕴含可以揭示该文本是否被用于模型训练的信息. Carlini 等人[13] 将模型对给定文本的困惑度（perplexity）作为信号在 GPT-2[28] 上进行预训练数据检测. 为了进一步降低文本的内在复杂性对检测的影响，他们介绍了 3 种基于参考的方法，分别为 Zlib, Lowercase,

Smaller Ref. 这些方法分别将文本的压缩熵、待检测模型对小写形式的文本的困惑度、与待检测模型同系列的小模型对文本的困惑度作为校准系数,对模型对给定文本的困惑度进行校准,然后将校准后的困惑度作为检测分数来判断给定的文本是否被用于模型训练. 2024 年,文献 [11] 提出一种简单且高效的方法,称为 Min-K% Prob. 然而,如前文所述,由于该方法直接将词元概率作为检测信号,因此倾向于将非训练文本误检为训练文本.

本文方法面向的是黑盒场景,基于长距离上下文对词元概率提升的贡献度来进行预训练数据检测,本质上是基于给定全部上下文时的词元概率与给定短距离上下文时的词元概率之间的对比结果. 因此,本文方法也可以理解为是对给定全部上下文时的词元概率进行校准,然后再将其作为检测信号. 与现有方法相比,本文方法优势在于不需要额外的参考物来支持校准过程的进行,如额外的参考模型.

2 背景知识

本节将介绍预训练数据检测任务的背景知识,首先对任务进行形式化的定义,然后介绍任务常用的基准数据集.

2.1 任务定义

给定一段文本 x 和一个待检测的大语言模型 M,其中 M 的预训练语料 D 不可知,预训练数据检测任务旨在设计一种方法 A 来检测 x 是否包含于 D 中. 因此,给定 x 和 M,如果 A 检测出 x 包含于 D 中,则返回 1,否则返回 0:

$$A(x, M) \to \{0, 1\}. \tag{1}$$

与此前的研究[11-13]类似,本文面向的是黑盒场景下的预训练数据检测任务,也就是将 M 视为黑盒模型,只能查询 M 对 x 的词元概率. 如对于一段文本 $x = x_1x_2\cdots x_n$,其中 x_i 表示用 M 对应的标记器(tokenizer)对 x 进行切分后的词元. 将 x 的子片段 $x_1x_2\cdots x_{i-1}$ 作为上下文输入给 M 后,可以获取 M 预测该片段之后的下一个词元 x_i 的概率. 至于模型的内部状态信息,如权重和网络结构,是不可获取的.

需要注意的是,预训练数据检测任务与现有的生成数据检测任务[29]存在部分共同之处,这主要体现在它们都可以利用词元概率来设计检测方法以解决任务. 但它们也存在着区别:在任务目标上,预训练数据检测任务旨在判断一段文本是否为大语言模型的训练数据,可用于审查模型的数据使用过程是否合乎法律法规,而生成数据检测任务则关注一段文本是否为大语言模型生成,常用于识别虚假信息;在方法设计上,预训练数据检测可以基于其他特征进行,而不仅仅是词元概率,如大语言模型对数据的记忆能力[26],这类方法可能就无法迁移到生成数据检测任务上.

2.2 基准数据集

传统的成员推理攻击[14,18,21]通常是在本地训练的模型上进行的,因此模型的训练数据和非训练数据是已知的. 相比之下,在大语言模型上进行预训练数据检测面临新的挑战. 由于大语言模型的预训练数据通常没有公开,因此构建基准数据集用于评估预训练数据检测方法成为难点. 近期,文献 [11] 提出一种构建基准数据集的方法,该方法基于待检测模型的发布日期,将网上可公开获取的数据划分为训练数据和非训练数据. 由于许多大语言模型,如 LLaMA[30], GPT-NeoX[31], OPT[32], 都是在 2017—2022 年期间发布的,且都将 Wikipedia 上的网页数据纳入其预训练数据中,所以文献 [11] 将 2023 年 1 月 1 日设定为时间分界线,将 Wikipedia 上 2023 年之后发生的事件的报道文章作为非训练文本,将 2017 年之前发生的事件的报道文章作为训练文本,然后根据这些数据来构建 WikiMIA 数据集,详细信息如表 1 所示. 该基准数据集可搭配前述大语言模型进行预训练数据检测方法的评估. 此外,WikiMIA 还提供多种文本长度的版本(32, 64, 128, 256)用于方法细粒度的性能评估,本文将在多种版本下的 WikiMIA 数据集上进行实验.

Table 1 Statistics for WikiMIA Dataset

表 1 WikiMIA 数据集的统计信息

数据来源	文本长度	数据集规模
WikiPedia	32	776
	64	542
	128	250
	256	82

注:文本长度是指一段文本的单词数量,数据集规模是指数据的条数.

此外,对于已公开预训练数据的大语言模型,如 Pythia[33],因为其是基于 Pile 语料库[15]预训练得到的,可采用 MIMIR 数据集[12]来评估预训练数据检测方法. MIMIR 数据集采集自 Pile 的训练数据集和测试数据集,MIMIR 数据集提供 3 种训练文本与非训练文本存在不同重叠程度的版本,可用于验证检测方法在缓解文本重叠现象对检测的不利影响上的有效性,表 2 展示该数据集的信息. 需要注意的是,表 2 中

从上至下的 3 种版本中的训练文本与非训练文本的重叠程度是依次加剧的.

Table 2 Statistics for MIMIR Dataset

表 2 MIMIR 数据集的统计信息

数据来源	版本	数据集规模	训练与非训练文本重叠程度
Pile-WikiPedia	7_gram_0.2	2 000	满足 7-gram，重叠率小于等于 0.2
	13_gram_0.2	2 000	满足 13-gram，重叠率小于等于 0.2
	13_gram_0.8	2 000	满足 13-gram，重叠率小于等于 0.8

注：数据集规模是指数据的条数.

3 方 法

本节将对提出的基于长距离上下文的预训练数据检测方法进行介绍.

3.1 概 述

本文方法受大语言模型的记忆能力研究启发，认为大语言模型在预测训练文本中词元的概率时，长距离上下文对词元概率提升的贡献度，会比预测非训练文本中词元的概率时的贡献度更大. 具体地，给定一段文本 $x=x_1x_2\cdots x_n$ 和一个待检测的大模型 M，其中 x_i 表示用 M 对应的标记器对 x 进行切分后的词元，我们的检测方法可分解为 4 个步骤：1）查询 M 在给定全部上下文时的词元概率；2）查询 M 在给定短距离上下文时的词元概率；3）对比 2 种情况下的词元概率，获得长距离上下文对词元概率提升的贡献度，并基于贡献度推导出最终的检测分数；4）基于检测分数判断 x 是否属于 M 的预训练数据. 整个过程不涉及模型训练，具体实现如算法 1 所示.

算法 1. 预训练数据检测方法.

输入：待检测文本 $x=x_1x_2\cdots x_n$，待检测大语言模型 M，短距离上下文长度 m，阈值 τ；

输出：x 与 M 的关系.

① for i=2 to n do

② 获取 M 基于全部上下文时对词元 x_i 的预测概率 $p_i=p(x_i|x_1x_2\cdots x_{i-1};M)$，即式（2）；

③ end for

④ for i=m+1 to n do

⑤ for λ=1 to m do

⑥ 获取 M 基于窗口长度为 λ 时的上下文时对词元 x_i 的预测概率 $q_i^{\lambda}=p(x_i|x_{i-\lambda}\cdots x_{i-1};M)$，即式（3）；

⑦ end for

⑧ 对 x_i 的所有 q_i^{λ} 取平均得到 q_i，即式（4）；

⑨ end for

⑩ for i=m+1 to n do

⑪ 基于 p_i 和 q_i 为 x_i 计算长距离上下文的贡献度 r_i，即式（5）；

⑫ end for

⑬ 对 x 中的所有词元对应的 r_i 取平均得到 β 作为 x 的检测分数，即式（6）；

⑭ if $\beta\geqslant\tau$ then

⑮ 1: M 使用 x 训练；

⑯ else

⑰ 0: M 未使用 x 训练；

⑱ end if

3.2 给定全部上下文时的词元概率计算

通过将 x 输入 M，可以获得 M 对 x 中每个词元 x_i 的预测概率：

$$p_i=p(x_i|x_1x_2\cdots x_{i-1};M),\ 1\leqslant i\leqslant n. \tag{2}$$

注意，每个词元 x_i 的概率都是由 M 基于 x_i 前面的全部词元预测得到的.

3.3 给定短距离上下文时的词元概率计算

根据短距离上下文的长度 λ，将 x 分成 $n-\lambda$ 个连续的子片段 $x_{i-\lambda}\cdots x_i$，并将子片段分别输入 M，可以获得在给定 x_i 前 λ 个词元作为上下文时，M 对词元 x_i 的预测概率：

$$q_i^{\lambda}=p(x_i|x_{i-\lambda}\cdots x_{i-1};M),\ \lambda<i\leqslant n. \tag{3}$$

我们计算 $\lambda\in\{1,2,\cdots,m\}$ 取值下，M 对 x_i 的概率，并取平均值作为最终 M 在给定短距离上下文时对 x_i 的预测概率，其中 m 是自定义的超参数.

$$q_i=\frac{1}{m}\sum_{\lambda=1}^{m}q_i^{\lambda},\ m<i\leqslant n. \tag{4}$$

注意，由于 x 头部的 m 个词元，即 x_1~x_m，无法提供片段长度为 m 个词元的短距离上下文，因此对这 m 个词元无法进行相应的概率计算，所以在此处及后续的计算中便没有将这 m 个词元考虑进来. 由于 m 的取值相对于序列 x 的长度 n 而言较小，所以舍弃的词元数量并不多，因此对方法的检测效果影响较小.

3.4 检测分数计算

至此，除了 x 头部的 m 个词元，我们为 x 中的每个词元都计算得到基于全部上下文时的词元概率 p_i 和基于短距离上下文时的词元概率 q_i 两种情况下的预测概率. 接下来，我们将基于 p_i 和 q_i 计算长距离上下文对词元 x_i 概率提升的贡献度：

$$r_i=\frac{p_i-q_i}{p_i},\ m<i\leqslant n. \tag{5}$$

基于计算得到的长距离上下文对词元概率提升的贡献度 r_i, 我们为 x 推导出最终的检测分数 β:

$$\beta = \frac{1}{n-m}\sum_{i=m+1}^{n} r_i. \tag{6}$$

3.5 二元决策

经过 3.2~3.4 节所述 3 个步骤, 我们为 x 计算得到检测分数 β, 用于判断 x 是否包含于 M 的预训练语料 D 中. 最后, 我们将 β 与提前设定好的阈值 τ 进行比较, 给出最终的检测结果:

$$decision(x,M) = \begin{cases} 0\ (x \notin D), & \text{if}\ \ \beta < \tau, \\ 1\ (x \in D), & \text{if}\ \ \beta \geqslant \tau. \end{cases} \tag{7}$$

当 $\beta<\tau$ 时, 我们推断 $x \notin D$; 否则, $x \in D$.

4 实验与分析

本节将介绍实验设置和实验结果, 并对结果进行相应的分析.

4.1 实验设置

本节将介绍实验所采用的数据集、大语言模型、评价指标、对比的基线方法以及本文方法的实现细节.

4.1.1 数据集和大语言模型

本文共采取 2 组实验设置. 第 1 组实验设置是为了验证本文方法在多种大语言模型和数据集上检测效果的优越性. 因此, 实验采用基准数据集 WikiMIA[11] 进行实验, 该数据集的详细信息已在 2.2 节中介绍过. 相应地, 实验采用的待检测大语言模型为 Pythia-2.8B[33], OPT-6.7B[32], LLaMA-13B[30], GPT-NeoX-20B[31], 因为这些模型是在 2017—2022 年期间发布的, 且都将 Wikipedia 上的网页数据纳入其预训练数据中. 此外, 为了探索模型大小对预训练数据检测方法的影响, 本文还在 Pythia 系列的多种模型版本(1.4B, 2.8B, 6.9B, 12B)上进行实验.

第 2 组实验设置是为了验证本文方法能够缓解训练与非训练文本的短片段重叠现象对检测的不利影响. 因此, 实验采用基准数据集 MIMIR[12] 进行实验, 该数据集来源于 Pile 数据集的训练数据集和测试数据集, 其详细信息已在 2.2 节中介绍过. 相应地, 实验采用的待检测大语言模型为 Pythia-2.8B, 因为该模型是在 Pile 数据集训练的.

4.1.2 评价指标

与现有大多数研究[11-12,25] 类似, 我们采取 AUC 分数作为主要的评价指标. AUC 分数常用于评估分类方法的性能, 它提供了一个全面的、与阈值无关的评分来反映方法以区分正负样本的有效性. 此外, 我们还采取低假阳性率(FPR)下对应的真阳性率(TPR)作为评价指标. 同样与大多数研究[11,25] 一样, 我们选择汇报当假阳性率为 5% 时对应的真阳性率, 记为 TPR@5%FPR.

4.1.3 对比基线方法

我们采取以下 5 种代表性的基线方法进行对比, 这些方法都是基于计算得到的检测分数来判断一段文本 x 是否属于待检测大语言模型 M 的预训练数据. 下面简单介绍它们的检测分数计算方式:

1) PPL[13,17]. PPL 检测分数为 M 对 x 的困惑度, 也就是先对 x 中所有的词元概率取负对数后再取平均, 最后取指数得到的值. 该值越小表明 x 越有可能是 M 的预训练数据.

2) Small Ref[13]. 由于直接利用 M 对 x 的困惑度作为检测分数会导致检测结果受 x 的内在复杂性影响, 因此该方法引入 1 个校准系数, 该系数为跟 M 同系列的较小模型 M_{ref} 对 x 的困惑度. 最终检测分数计算为 M 对 x 的困惑度与 M_{ref} 对 x 的困惑度之比. 该值越小表明 x 越有可能是 M 的预训练数据.

3) Lowercase[13]. Lowercase 与 Small Ref 方法类似, 但校准系数为 M 对小写形式的 x 的困惑度. 该值越小表明 x 越有可能是 M 的预训练数据.

4) Zlib[13]. Zlib 与 Small Ref 方法类似, 但校准系数为 x 的压缩熵. 该值越小表明 x 越有可能是 M 的预训练数据.

5) Min-K% Prob[11]. 先计算 M 对 x 中每个词元的预测概率, 然后按照预测概率由低到高对词元排序, 选择前 K% 个词元对应的预测概率的对数平均值作为检测分数. 该值越大表明 x 越有可能是 M 的预训练数据.

4.1.4 方法实现细节

对于本文方法的实现, 只涉及 1 个超参数的设置, 即所考虑的短距离上下文的最大长度 m. 由于本文方法本质上是想缓解训练文本与非训练文本之间存在的短片段重叠现象对检测的不利影响, 鉴于文献 [12] 对大型公开语料库 Pile[15] 的分析发现, 训练文本与非训练文本之间存在着大量的 7-gram 片段重叠现象, 因此我们设置超参数 $m = 6$. 由于我们采用 AUC 分数作为方法的评估指标, 因此在本文方法中不需要明确指定阈值 τ.

对于基线方法的实现, 我们设置 $K = 20$ 以达到 Min-K% Prob 方法的最佳性能[11]. 对于 Small Ref 方法中参考模型的设置, 我们分别使用 Pythia-70M 作为检测 Pythia-2.8B 时的参考模型, OPT-350M 作为检测

OPT-6.7B 时的参考模型, LLaMA-7B 作为检测 LLaMA-13B 时的参考模型, GPT-Neo-125M 作为检测 GPT-NeoX-20B 时的参考模型.

4.2 实验结果和分析

4.2.1 主实验结果

我们的主实验在文本长度版本为 128 的 WikiMIA 基准数据集上进行, 得到的实验结果如表 3 所示. 由表 3 可以看出, 本文方法在多种大语言模型上的检测效果都超过基线方法. 如相比于此前最好的方法 Min-*K*% Prob, 本文方法在检测 LLaMA-13B 时, 在 AUC 分数和 TPR@5%FPR 指标上分别比其高出 6.9 个百分点和 12.3 个百分点. 与 Small Ref 方法相比, 本文方法在不借助额外的参考模型的情况下, 仍然能够在 AUC 分数和 TPR@5%FPR 指标上分别比其平均高出 7.4 个百分点和 9.9 个百分点, 这凸显出本文方法在实际应用场景中的有效性. 此外, 在 MIMIR 上的实验结果如表 4 所示. 由表 4 可以看出, 所有方法的性能随着训练与非训练文本重叠程度的增加而下降. 当文本重叠程度较低时(7_gram_0.2), 方法之间的性能差距不大; 随着文本重叠程度加剧, 本文方法开始优于其他基线方法, 这表明本文方法能更好地应对文本重叠现象对检测的不利影响.

Table 3 Experimental Results Comparison on WikiMIA Dataset Detection Under Various Large Language Models

表 3 在多种大语言模型下进行 WikiMIA 数据集检测的实验结果对比

检测方法	Pythia-2.8B		OPT-6.7B		LLaMA-13B		GPT-NeoX-20B	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
PPL[13,17]	0.628	0.094	0.625	0.130	0.678	0.216	0.707	0.180
Small Ref[13]	0.626	0.137	0.654	0.122	0.658	0.151	0.714	0.187
Lowercase[13]	0.571	0.072	0.587	0.094	0.606	0.158	0.680	0.130
Zlib[13]	0.650	0.187	0.644	0.180	0.697	0.187	0.723	0.223
Min-*K*% Prob[11]	0.660	0.137	0.674	0.166	0.715	0.201	0.756	0.216
本文方法	**0.683***	**0.201***	**0.722***	**0.245***	**0.784***	**0.324***	**0.758**	**0.223***

注：黑体数值表示最优结果, “*”表示本文方法在双尾 T 测试下相较于现有方法有显著性提升.

Table 4 Experimental Results Comparison on MIMIR Dataset Detection Under Pythia-2.8B Model

表 4 在 Pythia-2.8B 模型下进行 MIMIR 数据集检测的实验结果对比

检测方法	7_gram_0.2		13_gram_0.2		13_gram_0.8	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
PPL[13,17]	0.654	0.201	0.540	0.106	0.521	0.042
Small Ref[13]	0.677	0.236			0.530	0.060
Zlib[13]	0.652	0.208			0.536	0.064
Min-*K*% Prob[11]	**0.679**	**0.250**	0.560	0.116	0.526	0.044
本文方法	0.665	0.227	**0.591**	**0.124**	**0.583**	**0.095**

注：黑体数值表示最优结果，带下划线的数值是从现有工作中引用.

4.2.2 待检测模型大小

为了研究待检测模型大小对本文方法性能的影响, 我们在 Pythia 系列的 1.4 B, 2.8 B, 6.9 B, 12 B 版本上进行实验, 以探索本文方法是否可以在参数量较大的模型上得到更好的检测效果. 结果如图 2 所示, 本文方法相比于其他方法在不同规模的模型上均取得最佳结果, 并且与其他方法一样, AUC 分数随着模型大小的增加而上升, 这进一步证实此前研究的发现[11,25]. 造成这一现象的原因可能是因为较大的模型拥有更多参数, 因此在记忆预训练数据方面的表现更好, 进而使得预训练数据检测任务变得相对容易.

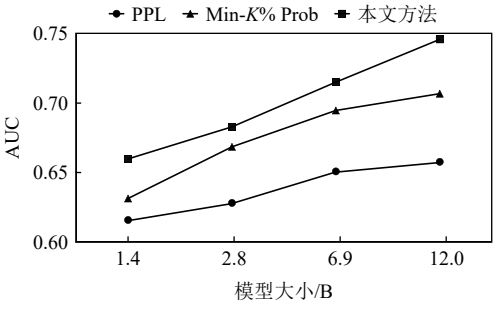


Fig. 2 The impact of model size on detection effect

图 2 模型大小对检测效果的影响

4.2.3 待检测文本长度

我们进一步探讨待检测文本长度对本文方法性能的影响. 为此, 我们采用 WikiMIA 基准数据集的 4 种文本长度版本(32, 64, 128, 256 个单词)在 Pythia-2.8B 上进行实验, 以探索短文本的检测是否比长文本更具有挑战性. 结果如图 3 所示.

本文方法在所有文本长度设置上仍然一致地优

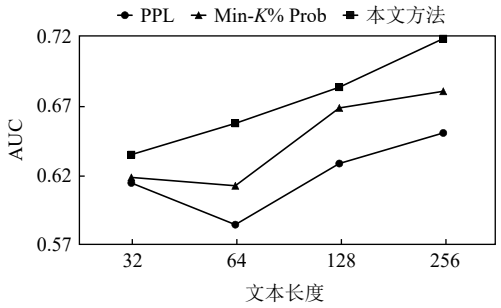


Fig. 3 The impact of text length on detection effect

图 3 文本长度对检测效果的影响

于其他基线方法,并且 AUC 分数也随着文本长度的增加而提高. 这一趋势可能是因为较长的文本携带更多信息,使得模型在记忆和处理文本时的表现更好,进而使得进行长文本检测时,有更多的信息可被提取用于检测.

5 总结和展望

本文首先基于训练文本与非训练文本之间存在的短片段重叠现象,指出当前的预训练数据检测方法受此影响的问题,然后提出一种基于长距离上下文的预训练数据检测方法来缓解这种现象对检测的不利影响. 通过在 WikiMIA 和 MIMIR 数据集上的实验证实本文所提方法的有效性. 在未来工作中,我们将探索数据集级别的预训练数据检测方法,而不仅仅是实现对单个文本片段的检测.

作者贡献声明: 张伟超负责研究方案的构思、设计、实现,以及论文撰写及修订;张儒清提供研究思路、论文组织结构的设计,以及审阅及修订论文;郭嘉丰和范意兴负责论文审阅并提出了修改意见.

参 考 文 献

[1] 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. Cambridge, MA: MIT, 2020: 1877−1901

[2] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint, arXiv: 2303.08774, 2023

[3] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint, arXiv: 2307.09288, 2023

[4] Bai Jinze, Bai Shuai, Chu Yunfei, et al. Qwen technical report[J]. arXiv preprint, arXiv: 2309.16609, 2023

[5] Yang Aiyuan, Xiao Bin, Wang Bingning, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint, arXiv: 2309.10305, 2023

[6] Mozes M, He Xuanli, Kleinberg B, et al. Use of LLMs for illicit purposes: Threats, prevention measures, and vulnerabilities[J]. arXiv preprint, arXiv: 2308.12833, 2023

[7] Chang K K, Cramer M, Soni S, et al. Speak, memory: An archaeology of books known to ChatGPT/GPT-4[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 7312−7327

[8] Dong Yihong, Jiang Xue, Liu Huanyu, et al. Generalization or memorization: Data contamination and trustworthy evaluation for large language models[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics: Findings. Stroudsburg, PA: ACL, 2024: 12039−12050

[9] Cao Jialun, Zhang Wuqi, Cheung S C. Concerned with data contamination? Assessing countermeasures in code language model[J]. arXiv preprint, arXiv: 2403.16898, 2024

[10] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models[C]//Proc of the 2017 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 3−18

[11] Shi Weijia, Ajith A, Xia Mengzhou, et al. Detecting pretraining data from large language models[C/OL]//Proc of the 12th Int Conf on Learning Representations. 2024[2025-03-02]. https://openreview.net/forum?id=PrCUrHJxTC

[12] Duan M, Suri A, Mireshghallah N, et al. Do membership inference attacks work on large language models?[C/OL]//Proc of the 1st Conf on Language Modeling. 2024[2025-03-02]. https://openreview.net/forum?id=av0D19pSkU

[13] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//Proc of the 30th USENIX Security Symp. Berkeley, CA: USENIX Association, 2021: 2633−2650

[14] Carlini N, Chien S, Nasr M, et al. Membership inference attacks from first principles[C]//Proc of the 2022 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2022: 1897−1914

[15] Gao L, Biderman S, Black S, et al. The Pile: An 800 GB dataset of diverse text for language modeling[J]. arXiv preprint, arXiv: 2101.00027, 2020

[16] Zhang Zhexin, Wen Jiaxin, Huang Minlie. Ethicist: Targeted training data extraction through loss smoothed softing and calibrated confidence estimation[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 12674−12687

[17] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]//Proc of the 31st Computer Security Foundations Symp (CSF). Piscataway, NJ: IEEE, 2018: 268−282

[18] Sablayrolles A, Douze M, Schmid C, et al. White-box vs black-box: Bayes optimal strategies for membership inference[C]//Proc of the 36th Int Conf on Machine Learning. Cambridge, MA: JMLR, 2019: 5558−5567

[19] Song Congzheng, Shmatikov V. Auditing data provenance in text-generation models[C]//Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2019:

196−206

[20] Hisamoto S, Post M, Duh K. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 49−63

[21] Jagannatha A, Rawat B P S, Yu Hong. Membership inference attack susceptibility of clinical language models[J]. arXiv preprint, arXiv: 2104.08305, 2021

[22] Mattern J, Mireshghallah F, Jin Zhijing, et al. Membership inference attacks against language models via neighbourhood comparison[C]// Proc of the 61st Annual Meeting of the Association for Computational Linguistics: Findings. Stroudsburg, PA: ACL, 2023: 11330−11343

[23] Nasr M, Carlini N, Hayase J, et al. Scalable extraction of training data from (production) language models[J]. arXiv preprint, arXiv: 2311.17035, 2023

[24] Oren Y, Meister N, Chatterji N, et al. Proving test set contamination in black-box language models[C]//Proc of the 12th Int Conf on Learning Representations. 2024[2025-03-02]. https://openreview.net/forum?id=KS8mIvetg2

[25] Zhang Weichao, Zhang Ruqing, Guo Jiafeng, et al. Pretraining data detection for large language models: A divergence-based calibration method[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 5263−5274

[26] Duarte A V, Zhao Xuandong, Oliveira A L, et al. De-cop: Detecting copyrighted content in language models training data[C]//Proc of the 41st Int Conf on Machine Learning. Cambridge, MA: JMLR, 2024: 11940−11956

[27] Liu Zhenhua, Zhu Tong, Tan Chuanyuan, et al. Probing language models for pre-training data detection[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 1576−1587

[28] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 1−24

[29] Bao Guangsheng, Zhao Yanbin, Teng Zhiyang, et al. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature[C]//Proc of the 12th Int Conf on Learning Representations. 2024[2025-03-02]. https://openreview.net/forum?id=Bpcgcr8E8Z

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

[31] Black S, Biderman S, Hallahan E, et al. GPT-NeoX−20b: An open-source autoregressive language model[C]//Proc of BigScience Episode. Stroudsburg, PA: ACL, 2022: 95−136

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

[33] Biderman S, Schoelkopf H, Anthony Q G, et al. Pythia: A suite for analyzing large language models across training and scaling[C]//Proc of the 40th Int Conf on Machine Learning. Cambridge, MA: JMLR, 2023: 2397−2430



Zhang Weichao, born in 2000. PhD candidate. His main research interest includes security of large language models.

张伟超, 2000 年生. 博士研究生. 主要研究方向为大语言模型安全.



Zhang Ruqing, born in 1994. PhD, associate professor. Her main research interests include natural language processing and information retrieval.

张儒清, 1994 年生. 博士, 副研究员. 主要研究方向为自然语言处理、信息检索.



Guo Jiafeng, born in 1980. PhD, professor. His main research interests include representation learning, and neural models for information retrieval and filtering.

郭嘉丰, 1980 年生. 博士, 研究员. 主要研究方向为表示学习、针对信息检索与过滤的神经模型.



Fan Yixing, born in 1990. PhD, associate professor. His main research interests include data mining and information retrieval.

范意兴, 1990 年生. 博士. 副研究员. 主要研究方向为数据挖掘、信息检索.