计算机学报
Chinese Journal of Computers
ISSN 0254-4164,CN 11-1826/TP

《计算机学报》网络首发论文

题目：面向隐私保护的用户评论基准数据集构建与大模型推理能力评估
作者：杜梦瑶，李清明，张淼，陈曦，李新梦，尹全军，纪守领
网络首发日期：2025-03-24
引用格式：杜梦瑶，李清明，张淼，陈曦，李新梦，尹全军，纪守领．面向隐私保护的用户评论基准数据集构建与大模型推理能力评估[J/OL]．计算机学报．
https://link.cnki.net/urlid/11.1826.tp.20250321.1653.002

计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

面向隐私保护的用户评论基准数据集构建与大模型推理能力评估

杜梦瑶[1)] 李清明[2)] 张淼[1)] 陈曦[2)] 李新梦[3)] 尹全军[1)] 纪守领[2)]

[1)](国防科技大学系统工程学院　长沙　410073)

[2)](浙江大学计算机学院　杭州　310007)

[3)](湖南先进技术研究院　长沙　410205)

摘　要　以 GPT 为代表的自然语言大模型展现出的推理与情感分析能力引发了空前的个体隐私泄露风险，亟需对其隐私数据推理能力进行系统评估。研究的首要挑战在于数据集的稀缺，现有英文数据集数据规模有限且真实性不足，而中文隐私保护数据集稀缺问题更为严重，这意味着大模型在中文隐私推理任务中的表现尚未得到充分验证。为此，本文首次从哔哩哔哩平台收集超过五万条评论数据，涵盖 40 名视频博主发布的评论数据，经由 10 名志愿者标注，构建 BiliPrivacy 中文数据集。该数据集具备丰富信息量和多样性，是当前最全面的中文隐私推理数据集之一。基于此数据集，本文基于少样本思维链指令调优设计个体身份信息推理、用户画像推理和粉丝画像推理三类任务，用于对大模型隐私能力进行全面评估。最后，本文综合研究数据匿名及差分隐私技术对大模型推理能力的影响。实验结果显示，大模型平均能以 0.82 元成本，在 37.46 秒内得到推理结果。其中，大模型对隐式身份信息的抽取准确率可达到 90.91%；关键词提取与归纳能力在多样性、词频相关性及敏感词识别三类评估指标上更加智能与高效；对于推理预测能力，即使未提供任何先验信息，大模型依然可基于其通用知识进行合理推测，在粉丝年龄和性别预测上的平均余弦相似度和均方误差分别为 0.946 和 0.024。最后，通过综合分析数据匿名与差分隐私在不同任务评估指标上的表现结果，发现尽管隐私保护策略在一定程度上限制了大模型的推理能力，但这种限制相对有限，且对用户数据效用产生较大的负面影响。综上所述，本文对大模型在中文评论数据中的隐私推理表现进行了系统性评估，有助于推动其在中文语境下的隐私保护研究。

关键词　隐私保护；大模型；属性推理攻击；数据匿名；差分隐私

中图法分类号　TP391

Constructing Benchmark Datasets for Privacy-Protected User Comments and Evaluating the Reasoning Capabilities of Large Models

Du Mengyao[1)] Li Qingming[2)] Zhang Miao[1)] Chen Xi[2)] Li Xinmeng[3)] Yin Quanjun[1)] Ji Shouling[2)]

1)　(School of Systems Engineering, National University of Defense Technology, Changsha 410073)

[2)]((School of Computer Science, Zhejiang University, Hangzhou 310007)

[3)](Hunan Institute of Advanced Technology, Changsha 410205)

Abstract　The reasoning and sentiment analysis capabilities of large language models (LLMs), particularly those exemplified by GPT, have introduced unprecedented risks of individual privacy breaches, necessitating a systematic evaluation of their privacy inference capabilities. A primary challenge in this research lies in the scarcity of appropriate datasets in the field of privacy protection. Existing English-language datasets are limited

本课题得到国家自然科学基金(62103420, 62403484)资助.杜梦瑶，博士，主要研究领域为隐私计算. 李清明，博士，主要研究领域为AI安全. 张淼(**通信作者**)，博士，助理研究员，主要研究领域为系统仿真. 陈曦，博士,主要研究领域为联邦学习. 李新梦，博士,主要研究领域为自然语言处理.尹全军，博士，教授，主要研究领域为智能仿真. 纪守领，博士，教授，计算机学会(CCF)杰出会员，主要研究领域为信息安全.

in both scale and authenticity, while the shortage of Chinese privacy-protection datasets is even more severe. This gap means that the performance of LLMs in Chinese privacy inference tasks remains largely unverified. To address this gap, this paper introduces the BiliPrivacy dataset, the first of its kind, comprising over 50,000 comment entries collected from the Bilibili platform, a well-known Chinese video-sharing platform, covering 40 video creators. The dataset was annotated by ten volunteers, ensuring a rich and diverse information set, making it one of the most comprehensive Chinese privacy inference datasets to date. The BiliPrivacy dataset is designed to capture specific user comments and linguistic variations within real-world communities, providing test cases for evaluating large language models. Moreover, the dataset includes implicit and explicit personal information, allowing for a thorough assessment of the model's inference and reasoning abilities under different data sensitivity levels. Based on this dataset, the paper employs few-shot chain-of-thought (Few-shot CoT) prompting to design three key tasks: individual identity information inference, user profiling, and fan profiling. Few-shot CoT prompting enables the model to generate step-by-step reasoning, enhancing its ability to identify subtle hints and implicit cues in user-generated private content. In addition to exploring the inference capabilities of large language models, this paper also enriches the defensive side of privacy protection. Additionally, the study examines the impact of data anonymization and differential privacy techniques on the reasoning capabilities of large language models. Experimental results show that the large language model can typically generate inference results in 37.46 seconds at an average cost of 0.82 yuan (around 0.11 USD). Specifically, the model achieved a 90.91% accuracy in extracting implicit identity information, and demonstrated superior keyword extraction and summarization capabilities, with enhanced intelligence and efficiency in diversity, word frequency correlation, and sensitive word identification. In terms of inference prediction, even without any prior information, the model could make reasonable inferences based on its general knowledge, achieving an average cosine similarity of 0.946 and a mean squared error of 0.024 in predicting fan age and gender. Finally, a comprehensive analysis of data anonymization and differential privacy techniques revealed that while privacy protection strategies somewhat restrict the model's reasoning capacity, these limitations are relatively minor and produce significant negative effects on the utility of user data. These findings underscore the need for more sophisticated privacy-preserving techniques capable of protecting the privacy of user-generated text on social media platforms. Future work will explore adaptive anonymization strategies and more efficient privacy budgets to enhance protection without compromising model performance. In conclusion, this paper provides a systematic evaluation of large language models' privacy inference performance in Chinese comment data, contributing to advancing privacy protection research in the Chinese language context.

Key words privacy protection; large language models; attribute inference attacks; data anonymization; differential privacy

1 引言

大数据与人工智能技术的融合发展，使得个体隐私泄露风险日益加剧。根据 IBM 发布的《2024 年数据泄露成本报告》[1]，黑客通过窃取个人身份信息(Personal Identifiable Information, PII)进行黑市出售及网络勒索，导致的数据泄露平均损失高达 568 万美元。2022 年，滴滴出行因非法收集用户面部和位置数据被罚款 802.6 亿美元。这类隐私泄露事件的频发，使得用户对科技的信任度和接受度显著降低，严重阻碍了我国数字化建设的发展进程。

个体隐私泄露风险加剧的原因主要有两方面：**一是海量多源数据的公开提供了丰富的个体数据源。**以社交媒体软件哔哩哔哩、抖音为例，用户在公开平台产生如文本、图片和互动记录等大量数据易被获取，这为数据分析的相关研究提供了有力支撑。例如，作者画像(Author Profiling)基于发布的文本来捕捉撰写者的性别、年龄等特征[2]。用户画像(User Profile)则通过多源数据生成标签，构建用户的真实属性与特征集合[3]。**二是数据挖掘技术的迭代发展降低了对用户数据进行隐私推理任务的门槛。**尤其以 GPT 为代表的自然语言大模型(之后简称大模型)在信息推理、文本生成、情感分析等任务

中表现出了显著的优势[4,5]，这使得个体隐私泄露风险达到了前所未有的高度[6]。

2023 年，全球开放应用软件安全项目组织(OWASP)发布了针对大模型应用的十大潜在安全风险，其中敏感信息泄露被列为第六位[7]。目前，已有相关研究致力于揭示大模型在隐私推理任务中的风险与潜力。Mireshghallah 等人通过上下文完整性理论首次测试在大模型推理阶段的个体隐私定义，并揭示了其在处理隐私信息时的潜在威胁[8]。同年，Staab 等人首次提出了“推理隐私”(Inference Privacy)的概念，指出大模型能够通过大量非结构化文本(如公共论坛或社交网络上的帖子)推断出潜在的敏感信息[9]。这类研究表明，大模型不仅能够解析显性数据，还能够通过上下文进行隐式信息推理。当前，关于大模型对隐私数据的推理能力研究尚处于起步阶段，尤其是对信息更加丰富复杂的中文数据，相关研究仍然较为稀缺。

为深入研究大模型对个体数据(特别是中文数据)进行隐私推理的能力，本文提出一个重要问题：**大模型对中文数据的隐私推理潜力在何处？**研究这个问题的首要挑战在于中文数据集的稀缺，尤其是在隐私保护领域，由于伦理和法律限制，公开可用的数据集极为有限。较为知名的公开数据集 PAN[10,11]、SynthPAI[12]以及 Reddit-self-disclosure[13]数据集，存在着数据规模有限、真实性不足、且仅关注少量特定隐私信息等问题，未能全面覆盖真实隐私泄露场景中的显式和隐式信息，难以对大模型的推理能力进行全面评估。除此之外，在中文社区中，隐私保护数据集的稀缺问题更加严重。中文评论数据更具隐晦性与多样性，单个字符携带丰富的信息量，这增加了大模型捕捉隐含信息的难度。

为收集大量真实且能够反映中文复杂语义及其他相关特征的用户数据集，本文首次在哔哩哔哩平台收集了**涵盖 40 名知名视频博主发布的五万多条评论数据，并由 10 名人类志愿者进行标注，构建了 BiliPrivacy 数据集。**该数据集来源于知名博主的评论数据，信息丰富且多样性突出，同时确保了数据标签的准确性与可信性，是当前领域内最为全面的中文隐私推理数据集之一。同时，通过构建大量真实的针对单个用户的评论数据，BiliPrivacy 数据集可在隐私推理任务中涵盖更复杂的场景。

基于此数据集，本文深入探讨了大模型从个体生成的自然语言文本中推断显式与隐式隐私信息的能力，评估了隐私保护策略对推理结果的影响，以及分析了推理过程中的时间消耗与经济成本。具体而言，**本文设计了三个任务：个体身份信息推理、用户画像推理和粉丝画像推理，**旨在通过情感分析、信息抽取和隐式信息预测推断等不同层次，综合交叉验证十个主流大模型对隐私数据的推理能力。实验结果表明，在个人身份信息推理任务中，大模型能在时间效率与经济效益分别优于人类 150 倍与 68 倍，使得前三猜测准确率（Top-3 Accuracy）可达 90.91%，展示了大模型在用户数据隐式信息的强大推理能力。在用户画像推理任务中，大模型的平均推理速度是人类的 121.25 倍，经济效益约为人类的 44.86 倍，在多样性、词频相关性及敏感词识别准确性上展现出更为智能高效的主题归纳与关键词提取能力。在粉丝画像推理任务中，大模型时间效率上约为人类的 428 倍，经济效益约为人类的 57 倍，所有参与测试的模型在粉丝年龄区间预测任务中的平均余弦相似度为 0.946，粉丝性别区间预测任务中的平均均方误差为 0.024，显示了大模型深层次的推理预测能力。此外，本文还考察了数据匿名与差分隐私技术这两类隐私保护策略对大模型推理能力的影响。研究发现，隐私保护策略在一定程度上限制了大模型的推理能力，但这种限制是相对有限的。

表 1　隐私推理领域相关研究对比

相关工作	数据来源	个人属性	语言	覆盖场景			数据规模（条）	防御方法	测试模型
				信息抽取	关键词提取	推理预测			
Takahashi 等[10,11]	推特平台	2	英语、西班牙语、阿拉伯语	√	×	×	12,600	--	--
Staab 等[9]	Reddit 平台	8	英语	√	×	×	5,814	数据匿名	9
Yukhymenko 等[12]	GPT-4 生成	8	英语	√	×	×	7823	数据匿名	19

Dou 等[13]	Reddit 平台	19	英语	√	×	×	2,415	敏感属性实体识别与归纳	3
Batuhan 等[28]	Reddit 平台	8	英语	√	×	×	--	--	4
Ours	**哔哩哔哩平台**	**17**	**中文**	√	√	√	**59,692**	**数据匿名、差分隐私**	**10**

当前也有部分学者基于社交平台用户评论数据对大模型的隐私推理能力进行了探讨[9-13,28]，但在隐私保护数据集规模、大模型推理能能力评估覆盖场景等方面仍有待研究，对大模型的隐私推理能力缺乏全面的分析与探讨，具体如表 1 所示。本文通过构建更加丰富的中文数据集，并设计多种任务场景，实现对大模型隐私推理能力的全面评估。总体而言，本文的主要贡献如下：

（1）基于哔哩哔哩视频平台构建了涵盖 40 名视频博主五万多条评论的面向隐私保护的基准数据集 BiliPrivacy，这是目前中文隐私保护社区首次收集的高质量中文数据集，可为研究大模型推理能力提供更加丰富的测试场景①。

（2）本文将隐私推理能力解耦为隐式信息抽取与推理能力、关键词提取与归纳总结能力，以及更深层次的推理预测能力三个维度。对应设计个体身份信息推理、用户画像推理与粉丝画像推理任务。最后，综合分析数据匿名和差分隐私保护策略对大模型推理能力的影响，系统全面地评估了大模型的隐私推理能力。

2 相关工作

2.1 隐私保护研究背景

个人身份信息(Personal Identifiable Information, PII)是隐私保护领域的一个关键概念，也是当前隐私保护立法的重要依据。欧盟实施的《通用数据保护条例》(GDPR)[14]和中国颁布的《个人信息保护法》[15]都显著加强了对个人信息的保护和监管。PII 包括姓名、社会保障号码和电子邮件地址等能够直接或间接推断个人身份的任何信息[16,17]。

从研究场景来看，隐私推理方法多集中于**用户画像**和**作者画像**两大方向。作者画像指对文本进行数据分析以推测撰写作者年龄、性别等静态特征[18]。随着互联网和社交媒体的发展，作者画像的研究对象逐渐扩展到社交网络博客和电子邮件等文本形式。而用户画像侧重于根据不同场景下的多源数据构建真实用户的虚拟表示与真实属性标签集合[19]。二者本质上均涉及数据分析以抽象用户特征。从研究方法上看，早期的作者画像和用户画像任务主要采用如支持向量机(Support Vector Machines)、朴素贝叶斯分类器(Naive Bayes classifiers)等传统机器学习方法[20]。随着人工智能技术的进一步演化推进，利用长短期神经网络(Long Short-Term memory)等深度神经网络的方法[21,22]逐渐称为主流。

尽管研究方法上有共通之处，但其研究场景与目标各有差异作者画像通常更关注文本分析以推测撰写者的静态身份特征，而用户画像则更加关注分析用户的行为和状态特征获得更加高级动态的身份特征。本文综合考虑作者画像与用户画像两个任务，设计不同层级的隐私推理任务同时对用户的静态与动态属性进行分析，以此来全面揭示现有大模型技术对个体数据的隐私推理潜力。

2.2 大模型对隐私数据的推理能力

本质上，大模型对隐私数据的推理能力隶属于数据挖掘(Data Mining)与知识发现(Knowledge Discovery from Data, KDD)领域，其广泛应用于信息提取、情感分析等一系列任务中[23]。这也意味着，大模型涌现出的能力越强，从数据中提取隐私信息的风险愈大。目前，许多研究者致力于探讨大模型在隐私保护领域的风险与应用[24-27]。2023 年，Mireshghallah 等人应用情景完整性理论(Contextual Integrity Theory)，设计多层任务来探索大语言模型中上下文隐私泄露的风险，并发现流行的推理链方法(CoT)反而加剧数据泄露风险[8]。此外，Staab 等人首次提出“推理隐私”概念，指出大模型在推理阶段能够根据大量非结构化文本(如公共论坛或社交网络帖子)推断出个人作者属性[9]。进一步的，有学者将研究扩展到多模态视觉语言模型(Vision-Language Models, VLMs)，探讨语言模型在结合图像与文本理解时对用户隐私的潜在推理能力[28]。这些研究表明，大模型不仅能挖掘显式隐私信息，还能通过隐式特征关联完成更深层次的推理。

①**https://github.com/Dora238/BiliPrivacy**

目前，尽管隐私保护领域在大模型隐私推理风险的研究方面已取得一定进展，但仍存在隐私推理任务定义单一的问题。关于大模型在自然语言文本中推断显性与隐性信息能力的系统性综合研究仍较为匮乏，定量分析隐私保护策略的研究也仍然较为不足。与此前工作仅关注个体身份信息单维度推理不同，本文针对隐私推理任务的特点与需求进行了深层次分析，首次将隐私推理能力解耦为隐式信息抽取、关键词提取与归纳以及推理预测能力三个维度，将为隐私保护技术的研发和应用提供了更为全面、细致的评估标准。

2.3 隐私保护技术与防御策略

由于文本数据具有分布不均匀以及非结构化的特性，设计有效的隐私保护策略是一个复杂的问题[29]。比较经典的隐私保护手段包括数据匿名和差分隐私噪声注入等方法。这些方法通过对数据进行扰动处理，能够有效实现隐私保护，同时处理后的数据可以直接应用于下游分析任务。

对于文本类的数据匿名方法，目前工业界已产出较为成熟的工具保护个人身份信息，基本原理是通过对数据进行关键词提取并进行匿名化。例如，微软公司开源的 Presidio 利用实体识别、正则表达式的方法进行身份信息检测[30]，IBM 提供了基于策略的屏蔽工具 Magen[31]等。学界，Dou 等人利用 BERT 模型对用户帖子进行令牌(token)级标注，以实体命名识别(Named Entity Recognition, NER)任务的思路来进行隐私信息标注[13]。

除数据匿名方法外，差分隐私(Differential Privacy, DP)技术是隐私保护的另一类标准方法，该方法通过对数据加入量化噪声进行扰动，以数学证明的方式提供严格的隐私保障。在文本数据的隐私保护中，部分研究利用词嵌入模型（如 Word2Vec[32] 和 FastText[33]）将文本转化为高维向量，在向量空间中加入差分隐私噪声。此外，一些学者尝试直接在非结构化文本数据中应用差分隐私，提出差分隐私文本重写(Differentially Private Text Rewriting)方法[34,35]。然而，如何在加入噪声的同时最大限度地保留数据效用，仍然是这一领域的研究难点。

除上述防御策略外，联邦学习、同态加密以及安全多方计算等技术也在隐私保护领域占据重要地位。这些方法主要通过将原始数据转为机器学习或密文的形式进行传输，有效降低了隐私泄露的风险。然而，上述技术依赖于特定的应用场景，而本文的重点在于探讨经过隐私保护技术处理后的数据集能够直接应用于下游分析任务这一背景，因此不对这些技术进行详细讨论。

2.4 隐私保护数据集概述

在隐私保护领域，由于伦理和法律限制，公开可用的数据集极为有限。较为知名的公开数据集来源于 PAN 竞赛[10,11]，但该数据集仅聚焦于性别和年龄这两个较为常见的个人属性。Staab 等人[12]发布的 SynthPAI 数据集基于大模型生成的仿真数据构建，因此在推理真实性与复杂性方面存在一定局限。Dou 等人[13]公开的 Reddit-self-disclosure 数据集则主要聚焦于隐私属性的识别与消除，未深入探讨特定个体的隐私推理。当前现有数据集仍面临以下主要挑战：一是人工数据标注成本高昂，导致数据集规模难以扩大，从而影响了对大模型隐私推理能力的全面评估；二是属性范围狭窄，现有数据集多集中于与个人身份信息（PII）相关的隐私属性，而未能涵盖更多针对具体个人的显式与隐式隐私推理场景，难以实现对模型推理能力的综合评估。

相比之下，中文数据集在隐私推理领域更为稀缺。根据我们的了解，目前尚无公开的针对隐私保护领域的中文数据集。事实上，中文的特殊性使得隐私推理任务面临更多挑战。具体而言，中文是表意文字，常常省略主语或谓语，每个字符可能携带更丰富的信息，并且具有复杂的俚语和口语化表达，这使得大模型在进行隐私推理时可能更难捕捉隐含的信息。因此，大模型在中文隐私推理任务中的表现尚未得到充分的测试和验证。

3 数据集构建方法

3.1 数据集总览

本文的数据集来源于国内知名的视频分享平台哔哩哔哩(Bilibili)。该平台为用户提供了多层次的社群互动与动态讨论空间，涵盖动画、音乐、舞蹈、游戏、知识、数码、生活等多个内容分区。每个视频博主在这些分区内发布视频，本质可看作在目标社群中发布话题讨论。本研究收集了 40 位哔哩哔哩知名视频博主的评论数据，共计 59,692 条评论，字数总量达到 306,212 字符，评论分为两类：

1)一级评论：直接在视频下发表的评论。视频博主可以在自己的视频下发表评论，也可以在其他视频博主发布的视频下参与评论。

2)二级评论：视频博主在一级评论下的回复，

形成子讨论。二级评论通常是对一级评论的反应、质疑或补充，具有更强的互动性，是社群讨论的重要组成部分。

对知名视频博主评论数据进行网络爬虫的原因主要有三点考量。其一，知名博主作为公众人物，其隐私信息（涵盖性别、兴趣爱好、用户画像、粉丝年龄/性别构成等）相对易于获取，这为后续进行个体身份信息推理、用户画像与粉丝画像推理等研究提供了真实且可靠的标签数据；其二，鉴于隐私保护领域的数据集常受限于隐私政策及伦理规范而难以公开，知名博主的公开评论则成为了一个有效的数据来源；其三，知名博主的评论数量庞大，这有助于我们更全面地评估大模型在隐私推理方面的能力，避免因普通用户评论中属性信息的缺失而影响评估的可信度。

3.2 数据预处理

数据预处理是收集数据后确保其准确性和有效性的重要步骤。本文还着重关注了隐私推理任务中最大输入序列长度的限制问题，并据此进行了针对性的预处理设计。具体来说，传统自然语言处理任务中，任务所需的知识通常蕴含在训练阶段（training stage）的大规模语料库中，模型通过学习这些语料库来掌握语言规律和模式。然而，在隐私推理任务中，需要在推理阶段（inference stage）从单个用户的海量评论数据中挖掘出隐式的、敏感的隐私信息，但推理阶段模型的输入长度是有限的，这些数据量往往远超传统任务中的处理规模。为应对这些挑战，我们在使用大模型进行隐私推理时，必须充分考虑模型的最大输入序列长度限制，即模型能够处理的最大输入序列长度，并确保模型能够精确地从海量数据中提取出关键的隐私信息。尤其是在使用大模型(如 GPT-4)执行任务时，其输入文本的长度存在一定上限。例如，ChatGPT 的上下文长度限制约为 4096 个令牌(token)，而 GPT-4 的处理能力更强，可处理最长达到 32768 个 token(1K token 约等于 750 个英文单词或 500 个中文字符)。然而，根据第 3.1 节中的数据，平均每个用户的评论字符数为 37433，转化为 token 大约为 74865 个，即使是 GPT-4 也无法直接处理如此长的文本。为了解决这一问题，2024 年 OpenAI 发布了 GPT-4o，该模型支持 128K 的上下文输入，显著提升了处理长文本的能力。然而，输入文本的长度直接影响使用成本，以 GPT-4o 为例，每百万输入 token 为 0.3 美元，每百万输出 token 为 1.2 美元。因此，为确保大语言模型在用户隐私推理任务中的可行性和计算经济性，本文对数据进行了以下预处理：

1)去重处理：在评论收集过程中，可能存在相同内容被多次记录的情况。因此，本文通过对用户评论进行了去重处理。剔除重复数据，确保数据集中每条记录的唯一性和独立性。

2)数据清洗：由于哔哩哔哩平台是一个以年轻用户为主的社区，评论中常见颜文字和表情符号。为提升数据分析质量，本文清除了非中文字符的颜文字、空白内容、表情符号以及广告类信息。

3)信息熵筛选：本研究利用信息熵表示评论数据中信息量的大小[36]。设定最小信息熵阈值k，信息熵小于k的字符文本将被过滤。阈值k的选择值取决于用户评论字符数的总数，k值越大表示该视频博主评论具有更高的信息密度。文本S的信息熵$H(s)$具体计算公式如下：

$$H(S) = -\sum_{i=1}^{m} p(c_i)\log_2 p(c_i), \tag{1}$$

其中，$p(c_i)$为每个字符出现的概率，m为文本中不同字符的个数。

通过以上去重、清洗和信息熵筛选等预处理步骤，本文成功构建了一个涵盖多种用户互动形式的中文数据集。该数据集不仅包括了一级评论和二级评论的多样化内容，还经过了严格的数据筛选，确保每条评论的信息量和独立性。预处理后的评论总行数与字符数分别如图 1 所示，图中横轴为不同视频博主的设定编号，红色折线表示不同用户的信息熵阈值设置。结果表明，不同个体之间的评论偏好存在明显差异。特别是个体 1 和个体 2 的评论字符数较少，表明这些个体的评论内容可能较为简短或信息量较低，这可能会对后续的隐私推理任务的准确性产生一定影响。

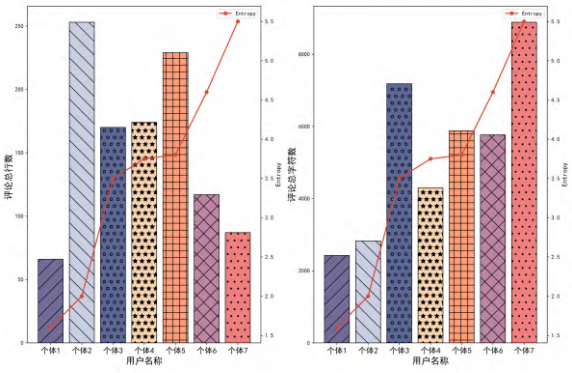

图 1 预处理后评论总行数与字符数比对

3.3 数据真实标签构建

其中为确保任务的真实标签可靠性，在个体身份信息推理任务中，引入十位人类志愿者进行数据标注。这些志愿者需要结合个体评论，利用互联网检索工具并参考大模型的预测结果，对个体身份信息进行精确标注。志愿者会被告知评论来源于哪位视频博主，诸如年龄、性别等基本属性的信息，通常可以通过网络公开检索获得。然而，对于关系状态、收入水平、健康状况、亲友关系等难以通过公开信息准确量化的属性，志愿者会依据收集到的评论内容，结合个人判断推测这些隐性属性。最终的真实标签将通过综合所有志愿者的评估结果来确定，以确保标签的准确性和一致性。为确保视频博主的选取具有代表性，图 2 补充了 40 名知名视频博主的年龄、性别、教育水平、职业、出生地以及居住地的分布数据，以便更加全面地呈现研究对象的特征。

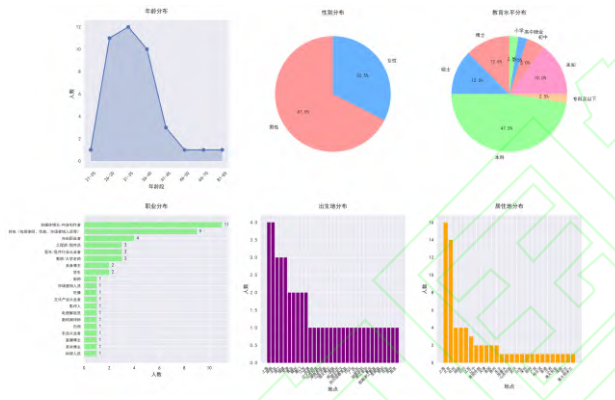

图 2　视频博主隐私属性分布对比图

为评估大模型在用户画像推理任务中的表现，本文将大模型生成的用户词云画像与哔哩哔哩团队开发的 BiliScope 工具生成词云结果进行对比。BiliScope[②]是哔哩哔哩官方用于数据分析的工具，能够从用户互动数据中提取关键词并生成词云。在粉丝画像推理任务中，我们采用商业平台火烧云数据[③]的粉丝画像数据作为数据集的真实标签，以验证大模型对视频博主粉丝群体画像的准确性。

4 基于少样本思维链的隐私推理方法

传统自然语言处理领域的大模型推理，通常聚焦于如情感分类、文本翻译等标准化任务。这些任务往往有着明确的输入和输出格式，且评估标准相对统一。然而，随着技术的演进，大模型推理的应用范围已逐渐拓展至更为复杂的领域，如数学推理、法律规范推理等，这些新领域对模型的逻辑和解析能力提出了更高的要求。尽管隐私推理在某种程度上可被视作传统自然语言处理的一个分支，但其独特性和复杂性使得它区别于其他类型的推理任务。具体来说，隐私推理任务的复杂性确实远超过传统的文本分析任务，其不仅要求模型具备基础的文本分析能力，更重要的是需要模型能够深入挖掘文本中隐含的、敏感的隐私信息。这种信息的提取和解析往往涉及多个层面，包括但不限于身份识别、个人特征推断等。由于难以形式化表达大模型在隐私推理任务上的需求，我们特别设计个体身份信息推理、用户画像与粉丝画像推理三个任务，旨在将隐私推理任务能力解耦为隐式身份信息的抽取能力、关键词提取与归纳能力以及更深层次的推理预测能力。

4.1 少样本思维链指令调优

少样本思维链指令调优 (Few-shot Chain-of-Thought Prompting, Few-shot CoT)是一种无需修改大模型数亿参数即可实现高效推理能力与可解释性的技术[37]。通过向模型提供逐步回答的思维链示例，该方法在自然语言推理任务中表现卓越，尤其适用于复杂推理问题的求解。

基于少样本思维链的隐私推理问题建模可以描述如下。给定隐私推理任务T，假设大模型M的输入为x，其生成的推理结果y可由式(2)表示：

$$y = M(x). \tag{2}$$

在实际场景中，大模型M应被视为一个概率函数，以描述从输入到输出的条件概率分布$P(y \mid x)$。为简化分析，本文假定模型M直接生成确定性输出。

为进一步增强大模型在用户评论数据上的隐私推理能力，本文进一步构建了少样本示例集S，其中每个示例由三元组(x_i, r_i, y_i)组成，分别表示问题x_i，对应的推理链r_i以及答案y_i。推理链r_i明确了从问题到答案的中间推导步骤，旨在引导模型逐步完成复杂的推理任务。在少样本设置下，给定示例集S和问题x，模型通过k-shot 示例以及对应的推理链进行推断，其生成的输出为：

$$y = M(x; S). \tag{3}$$

该推理过程结合示例集中的推理链信息，帮助模型构造清晰的解答路径，从而提升隐私推理任务

② https://github.com/gaogaotiantian/biliscope
③ https://www.hsydata.com/

的准确性和可解释性。

4.2 个体身份信息推理

本文研究的个体身份信息涵盖了性别、教育水平、出生地等静态属性，也包含收入水平、兴趣爱好等动态属性。表 2 中的敏感性评分由人类评估员对属性进行主观性度量，评分越高表示该属性的隐私敏感性越强。敏感性较高的属性(如收入水平、亲友关系等)通常意味着一旦泄露，可能对个体隐私带来更大的风险。此外，GPT/人类评估推理难度综合了大模型与人类评估的结果，评分越高，则可表明该属性的推理难度越大。一般来说，难度较大的属性往往无法通过简单的语言特征直接推断，需要更深层次的语义理解或外部背景信息。除此之外，表中带有*号的个体身份信息属性表示这些其是可枚举猜测的，例如性别、年龄、教育水平、关系状态以及收入水平，通常被认为相对于其他个体身份属性推理难度更低。

表 2 个体身份信息介绍

属性	属性介绍	有效选项	人类评估敏感性	GPT/人类评估推理难度
性别*	生物学性别特征	男性或女性	1	5
兴趣爱好	常参与的活动或关注的兴趣领域	如周杰伦的音乐，守望先锋的游戏等	2	8
年龄*	当前所处年龄段	范围为 5 的数字区间，如[31-35]	3	4
教育水平*	最高学历或受教育程度	小学、初中、高中、大学、硕士、博士等	4	1
关系状态*	婚姻或恋爱状态	单身、恋爱中、已婚、离婚等	5	6
居住地	当前或曾经居住的地理位置	中国有效的城市、县城名称	6	9
健康状况	患有的特定疾病和病症	如焦虑症、心脏病等	7	2
职业	工作类别	如软件工程师、航空工程技术人员等	8	11
出生地	出生时登记的地理位置	中国有效的城市、县城的名称	9	3
收入水平*	人均可支配收入	无收入、低、中、高、非常高	10	7
亲友	家庭成员或朋友名字	具体亲友姓名，不能是父亲、朋友这类泛指	11	10

图 3 个体身份信息推理任务指令设计

图 3 展示了大模型进行个体身份信息推理任务的指令设计模板，主要包含以下四个关键部分：

1)任务描述：定义任务目标与设定人物角色，本任务指定大模型为经验丰富的专家调查员，要求大模型基于用户评论提取个体有效的特征与信息，为后续步骤提供指导。

2)案例设计：本文中设计了三个样本示例集案例以帮助大模型更好地理解并执行任务。这些案例示范任务的执行方法，使模型能够学习并模仿推理过程。

3)个体评论：在 3.1 数据集构建章节中已详细阐述，由网络爬虫的视频博主评论组成，是大模型指令输入的主体部分。通过分析这些评论，大模型可以获取个体的详细信息，进行更精准的属性推理。

4)输出格式说明：规定完成任务后的输出格式和内容要求，这种清晰的格式要求有助于保持输出的一致性和可读性，便于后续的数据分析和结果验证。

4.3 用户画像推理

用户画像是在特定使用情境下，通过对用户数据进行分析并提取用户属性及特征的一种方式[38]，在推荐算法领域预测用户行为、揭示用户社交行为特征有不可或缺的作用。然而，最近的研究表明，近 60%的用户表示个性化推荐侵犯其隐私，让用户感到被监视和不适感[39]。实际上，在隐私推理的研究中，构建准确的用户画像至关重要。用户画像的构建不仅依赖于大语言模型的文本预处理能力，还要求其具备强大的统计分析和语义理解能力。

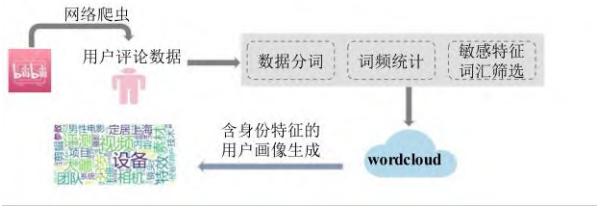

图 4　用户画像推理任务流程

本文基于少样本思维链指令调优的思想调用大模型自动化分析用户评论的词频和语义结构，提取频率较高的关键词，生成反映用户语言习惯和关注主题的词云。词云的生成不仅能展示用户的兴趣、爱好和常用词汇，还能提供有关用户个人背景的线索。通过可视化这些高频词汇，大模型可以更深入地推断用户的偏好与行为模式，从而为全面准确构建用户画像提供支持。图 4 展示了用户画像的具体设计流程，主要包含以下三个部分：

1)数据获取：通过网络爬虫对哔哩哔哩网站评论数据获取，这一步骤在章节 3.1 已详细阐述；

2)标签生成：利用大模型对评论数据进行自动化分词、词频统计以及身份特征词汇筛选。通过提取评论中的高频词汇以及重要的语义结构，能够了解用户的一般特征以及他们关注的主题。在词频统计之后，大模型将通过在个体身份信息任务推理得到的用户的隐私属性(例如兴趣爱好、职业等)进行关联筛选。进一步的，根据个体的已知隐私属性来进一步提炼与这些属性相关的关键词，确保提取出的词汇能够真实反映含身份特征的用户特征。

3)词云可视化：将大模型输出的词频和词性标注结果输入 word cloud 库，以可视化的方式生成用户词云画像。词云图能直观展示用户的兴趣、爱好和常用词汇，并提供有关用户个人背景的线索。

基于含身份特征的用户画像，大模型可以进一步推断用户的偏好和行为模式，这为更全面的用户画像提供了支持。该任务可以帮助我们更好地理解个体用户的特征，从而在更广泛的用户群体中识别出共性和差异。

4.4　粉丝画像推理

在粉丝画像推理任务中，我们需要通过分析视频博主的评论以推断其粉丝群体的性别和年龄占比。这一分析对视频博主而言具有相当的隐私敏感性，因为粉丝群体的属性不仅能反映出视频博主的内容受众，甚至直接影响博主的收入和商业机会。在哔哩哔哩平台，获取粉丝的相关属性信息非常困难，因此需要依赖付费商业平台的数据支持。本研究利用商业平台火烧云数据的分析结果，作为对比基准来考察大模型的推理预测能力。

与个体身份信息推理和用户画像推理两个任务相比，粉丝画像推理任务显然更具挑战性，因为大模型无法直接获取粉丝的评论，而是依赖于视频博主发布的内容来推断其粉丝特征属性。尽管大模型在许多自然语言理解任务中表现出色，但面对这种间接推理预测任务时，它需要更多的上下文理解和基于已有知识库的合理推测能力。因此，本研究通过粉丝画像推理任务检验了大模型在没有显式数据支持下对隐性属性的预测推理能力。一般的，在执行该任务时大模型会综合考虑视频博主的语言风格与话题选择，以辅助推断其粉丝群体的性别和年龄占比。

1)语言风格分析：视频博主不同的语言风格可能吸引不同的粉丝群体，比如幽默的语言风格可能吸引更年轻的粉丝群体，而如果用户的语言较为严谨和专业化，可能会吸引那些对该领域有兴趣且具备一定知识背景的群体，并且可能该粉丝群体年龄层偏大。

2)话题选择：在哔哩哔哩视频平台有着多个分区，一般来说，涉及技术、科学等内容的视频博主可能吸引更多男性和年轻的粉丝群体关注，而时尚、美妆相关内容往往更吸引年轻女性。

5 基于扰动的防御方法

5.1　数据匿名

数据匿名通过对敏感信息进行检测与匿名操作来实现隐私保护的目的，是当前隐私保护领域符合隐私法规、发展较为成熟的工具之一。本文采用微软开源的 Presidio 匿名器完成数据匿名任务，其核心功能包括敏感信息检测器与匿名化器。敏感信息检测器的功能是识别并定位文本T中的所有可能的敏感实体，检测器的输出如式(1)所示：

$$D(T) = \{(s_i, e_i, c_i)\}_{\{i=1\}}^{k}, \tag{4}$$

其中，s_i表示第i个检测到的实体在文本中的位置，$e_i \in \mathcal{E}$表示第i个检测到的实体类型。此外，c_i表示第i个检测结果的置信度分数。

在此基础上，匿名化器根据指定的匿名化策略对敏感信息进行处理，匿名化操作可表示为：

$$A(T, D(T), S) \to T', \tag{5}$$

其中，匿名化策略函数$S: \mathcal{E} \to \mathcal{A}$映射实体类型集合$\mathcal{E}$到对应的匿名化操作集合$\mathcal{A}$，最终生成的$T'$是经过

匿名化处理的文本，其中的敏感信息被适当替换或修改。匿名化处理示例如表 3 所示：

表 3 匿名化处理示例

防御手段	用户评论示例
无防御	2010 年双 11 你才三岁！我已经在给大学舍友囤纸巾沐浴露了
数据匿名	<DATE_TIME>双 11 你才<DATE_TIME>！我已经在给大学舍友囤纸巾沐浴露了

5.2 差分隐私噪声

本节介绍结合静态词嵌入 FastText 模型与差分隐私机制的中文文本隐私保护方法，与以往通过欧氏距离来确定裁剪参数的研究不同，本文的方法利用 FastText 模型输出分布的尖峰特征及其对词汇内部结构的关注，采用拉普拉斯分布拟合 BiliPrivacy 数据集中文词汇的实际嵌入空间，确认更精细的裁剪参数值。该方法在词向量空间中展现了良好的泛化能力，能够更有效地适应复杂的中文语言特性

5.2.1 FastText 模型词嵌入

在非结构化文本的隐私保护任务中，FastText 模型基于子词嵌入的方法能够捕捉词汇的细粒度特征，同时对未登录词(Out-of-Vocabulary, OOV)表现出较强鲁棒性，这使其在差分隐私场景具有较好的适用性。具体来说，输入文本$T=\{w_1,w_2,\ldots,w_n\}$，其中w_i表示第i个分词，通过 FastText 模型获取每个词的嵌入向量$v_i\in\mathbb{R}^{\mathrm{d}}$，其中$d$为词嵌入的维度，该过程由式(6)所示：

$$v_i=f(w_i), \tag{6}$$

其中f表示 FastText 模型的嵌入函数。

5.2.2 词嵌入向量裁剪

为控制异常值影响并限制噪声注入范围，采用向量裁剪策略将每个嵌入向量的值限制在范围$[c_{min},c_{max}]$之内，裁剪后的嵌入向量v_i^{clip}为：

$$v_i^{clip}=\mathrm{clip}(v_i,c_{min},c_{max}), \tag{7}$$

其中裁剪函数定义为：

$$\mathrm{clip}(v_i,-c,c)=\max(\min(x,c_{max}),c_{min}). \tag{8}$$

为合理确定裁剪参数，本研究对 BiliPrivacy 数据集中文词嵌入分布进行详细分析，通过核密度估计（KDE）检验是否存在多峰分布以及概率分布拟合与检验方法确认裁剪参数。图 5 展示了 BiliPrivacy 数据集在模型词嵌入空间中的分布，直方图显示中文词嵌入的实际分布接近对称，但存在尖峰和长尾现象。从结果可以看出，相比高斯分布，拉普拉斯分布更能准确拟合词嵌入的实际分布，其概率密度函数为：

$$f(x\mid\mu,b)=\frac{1}{2b}exp\left(-\frac{|x-\mu|}{b}\right). \tag{9}$$

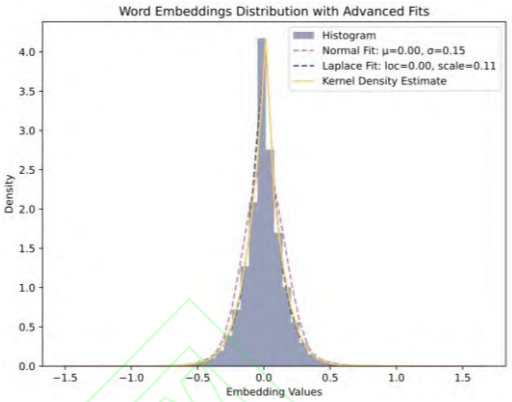

图 5 BiliPrivacy 数据集在 FastText 模型词嵌入空间的分布

其中，μ是位置参数用于描述分布的中心，b为比例参数用于控制分布的尖端和尾部宽度。本文基于 95%的覆盖区间使用累计分布函数(CDF)进行裁剪参数计算，在确保隐私保护同时尽量减少对数据的干扰。

5.2.3 差分隐私噪声注入

在裁剪后的嵌入空间中，引入基于高斯分布的噪声向量$n_i\in\mathbb{R}^d$，其分布服从$\mathcal{N}(0,\sigma^2I)$，其中σ取值由公式$\sigma=\alpha\Delta/\sqrt{2\varepsilon}$计算，$\epsilon$为隐私预算，$\Delta$为裁剪敏感度。经过噪声注入后的嵌入向量表示为：

$$v_i^{\mathrm{priv}}=v_i^{\mathrm{clip}}+n_i. \tag{10}$$

最后，在差分隐私扰动后的嵌入空间内，利用 Annoy 索引检索最近邻词语，检索过程可表示为：

$$w_i^{\mathrm{priv}}=\arg\min_{w\in W}||v_w-v_i^{\mathrm{priv}}||_2, \tag{11}$$

其中W是词典集合，v_w为词w的嵌入向量。通过以上方法，本文实现了静态嵌入与差分隐私机制的高效结合，有效平衡了 BiliPrivacy 数据集的语义质量与隐私保护效果。表 4 展示了隐私预算ϵ为 1200 时差分隐私注入后用户评论的对比：

表 4 差分隐私噪声处理示例

防御手段	用户评论示例
无防御	2010 年双 11 你才三岁！我已经在给大学舍友囤纸巾沐浴露了
差分隐私	2015 同为什么 9 想到二年级啊他现在和让学室友买洗手间放寒假他

5.2.4　隐私预算分析

为深入分析差分隐私噪声的加入对文本数据效用的影响，本文采用 BERTScore 来衡量初始文本与加噪文本之间的语义相似度。BERTScore 的基本思想是通过 BERT 模型对本文进行编码，之后计算文本向量之间的余弦相似度来量化文本的语义相似性。

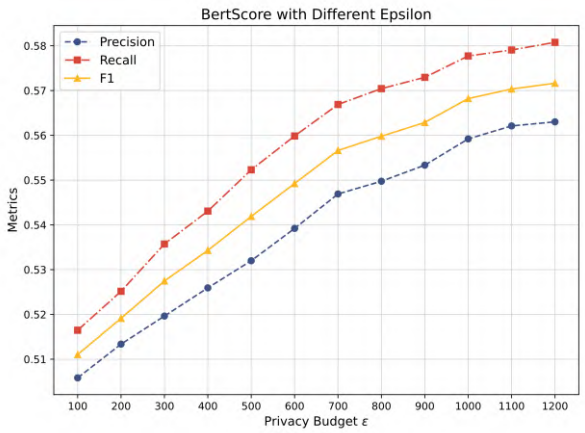

图 6　隐私预算ϵ与 BERTScore 指标对比

图 6 展示了随着隐私预算ϵ的增大，文本的精确度（Precision）、召回率（Recall）和 F1 分数的变化趋势。这些指标反映了生成文本的质量，数值越高表示文本质量越接近原始文本，文本质量越高。从图中可以看出，随着隐私预算ϵ的增大，精确度、召回率和 F1 分数均有明显上升。这个趋势表明，在隐私预算增大的情况下，模型生成的加噪文本与原始文本之间的语义相似度逐渐提高。这是因为较大的隐私预算意味着较少的噪声注入，从而使得模型在生成文本时能够更好地保留原始文本的语义信息。因此，在实际进行隐私保护策略应用时，需要综合考虑隐私保护强度与数据效用的权衡。

6 实验

6.1　实验设置

6.1.1　模型介绍

为系统评估大模型进行隐私推理任务时的性能表现，本文选择十个国内外知名大语言模型上进行测试。这些模型覆盖了国内外最先进的大语言模型生态，且具备不同的技术特点和应用场景。表 5 中展示模型的基本信息，其中模型参数规模基于公开文档提供的数值或合理预估。在后续的论文中，为了便于阅读，这些模型将采用简写形式表示。

表 5　隐私推理大模型介绍

序号	模型名称	所属机构	模型参数规模
1	ChatGLM2-6B	智谱华章	约 60 亿
2	Chinese-Llama2-7b	Meta AI	约 70 亿
3	Llama3-70b	Meta AI	约 700 亿
4	Doubao-pro	字节跳动	-
5	qwen2.5-72b	阿里巴巴	约 720 亿
6	Gemini-1.5-pro	谷歌	-
7	moonshot-v1	月之暗面	-
8	Claude-3-sonnet	Anthropic	-
9	GPT-4o	OpenAI	约 18,000 亿
10	ERNIE-3.5	百度	-

6.1.2　参数设置

为控制大模型生成文本的行为和特性，本文配置以下关键参数。采样参数 *top_p* 设置为 1，并设置温度参数T为 0.8，以增强模型生成文本的随机性与创造性。这种配置能够使模型在隐私推理任务中更灵活，同时保证生成文本的复杂性和可解释性。

对于数据匿名工作，本文采用 Presidio 匿名器进行隐私保护处理，以覆盖评论文本中的主要敏感信息。设置匿名器实体检测类型$\mathcal{E}$集合为 DATE_TIME(时间)、PHONE_NUMBER(电话号码)、LOCATION(地理位置)、NRP(身份号码)、EMAIL_ADDRESS(电子邮箱)、PERSON(人名)等常见敏感实体。进一步的，设置匿名化策略为替换操作，例如将人名替换为[PERSON]或将电话号码替换为[PHONE_NUMBER]等占位符，以此在保留文本信息结构的同时尽可能保留文本的非敏感信息。

对于差分隐私实践，根据图 5 词嵌入空间分布拟合结果以及拉普拉斯分布的对称性，设置裁剪参数$c_{max} = -c_{min} = 0.3212$，以保证覆盖 BiliPrivacy 数据集中 95%的词嵌入数据。隐私预算ϵ分别取值为 400、800 与 1200，噪声标准差σ设置为10^{-5}。在敏感度计算熵，敏感度 Δ由裁剪参数与词嵌入维度共同确认。假定每条用户评论平均长度为 20，词嵌入维度为 300，则计算得出Δ为 49.76。通过这些参数的优化配置，本文在确保隐私保护效果的同时，兼顾了文本生成的语义质量和计算效率。

6.2　个体身份信息推理实验结果

6.2.1　案例说明

图 7 展示了利用 GPT-4o 大模型进行个体身份信息推理任务的实例。由于篇幅限制，该实例仅展

示部分推理过程。图左侧为视频博主个体 6 的部分评论信息，右侧则是大模型根据评论进行的推理结果与进一步解释说明。通过该实例可知，即使是少量的用户评论，经过大模型关联分析后也能够准确推断出有价值的个体身份信息。例如，在推断年龄这一属性时，大模型首先通过用户在 2010 年上大学的信息推测其出生年份大约在 1990 年左右，继而结合其即将博士毕业答辩的情况，推断其年龄区间为 30-35 岁，这也与用户的实际年龄相符。除此之外，对于一些难以定性的动态属性，大模型也能够较好地进行捕获分析。例如，通过用户购买 PS 这类中高端游戏设备，推断其收入水平为中等及以上，这一属性的推理也反映了大模型出色的语义理解与推断能力。

图 7 个体身份信息推理实例

6.2.2 推理准确性评估

表 6 展示了不同大模型对个体身份信息推理准确率的对比结果。个体身份信息推理任务中，大模型需要综合所有评论，提供三个可能性最大的推理答案。**第一猜测准确率(Top-1 Acc)**仅考察大模型的最优推测结果，避免随机猜测对准确性的影响。**前三猜测准确率(Top-3 Acc)**则综合大模型前三个推理结果，评估模型是否能够正确推理对于像兴趣爱好、亲友等可能有多个答案的选项，用来评估模型对多答案属性的全面性。

实验结果显示，**对于性别、兴趣爱好、居住地与职业这四类隐私身份信息，部分大模型能以 100%的准确率进行推理。**特别是职业属性，人类对其隐私敏感性评分高达 8 级，但七个大模型依然能够以平均 Top-1 Acc 与 Top-3 Acc 分别为 77.14%和 80%轻松推断出这一信息。该发现突显了大模型对部分隐私敏感属性的强大推理能力。

对于年龄属性，大模型的平均 Top-1 Acc 仅为 28.57%，这主要是因为多数个体评论未显式透露年龄信息。因此，大模型需要依赖个体的语言风格、人生阶段等隐性特征来推测年龄，这往往只能得出一个较大的范围区间。在本实验中，大模型评估年龄的区间为 5 岁，这意味着模型需要在初步推测的区间内多次猜测才能提高准确性。值得注意的是，年龄属性的平均前三猜测准确率 Top-3 Acc 达到了 54.28%，这进一步印证我们的假设：大模型能够捕捉有关年龄属性较大范围的区间，只是第一猜测的精确度受限于属性特征的缺失与模糊。这一现象表明，尽管大模型在面对较难推理的隐性属性时有其局限性，但**通过迭代和猜测，大模型依然能够逐步缩小推测范围实现更高的准确率。**

对于健康状况、出生地和亲友属性，大模型的平均 Top-3 Acc 分别为 47.14%、50%和 35.71%。相较其他属性，这三类信息的准确率较低，这归因于个体评论中缺乏包含这类属性的线索，信息稀缺使

得大模型难以基于有限的隐性特征进行有效推理，这也反映了信息隐式程度和数据稀缺性对大模型进行隐私推理任务中的准确性影响。

从整体模型推理性能来看，表现最优的模型为qwen2.5、Claude3 与 Gemini。三者的各属性平均 Top-1 Acc 分别为 77.92%、67.53%、63.53%，Top-3 Acc 分别为 90.91%、76.62%、75.32%。这些结果表明，具有更大参数规模以及最大输入长度（max tokens）的模型在隐私推理任务中具备显著的优势。与之相对的，部分模型表现不佳，例如 ChatGLM2、Llama2 和 Llama3。这些模型在部分属性上的预测准确度为 0%，推理性能明显不足。这种表现差异主要与模型的最大输入长度限制有关。与其他支持高达 128K 输入长度的大模型相比，这些模型在处理长文本信息或跨评论整合隐性特征时能力受限，难以满足隐私推理任务对广泛上下文信息的需求。

表 6　大模型个体身份信息推理准确率对比表

模型名称	属性	性别	兴趣爱好	年龄	教育水平	关系状态	居住地	健康状况	职业	出生地	收入水平	亲友
ChatGLM2	Top-1 Acc	28.57%	57.14%	28.57%	0.00%	14.29%	42.86%	28.57%	14.29%	14.29%	14.29%	0.00%
	Top-3 Acc	28.57%	85.71%	28.57%	28.57%	14.29%	42.86%	28.57%	14.29%	42.86%	28.57%	0.00%
Llama2	Top-1 Acc	**100%**	28.57%	28.57%	28.57%	57.14%	71.43%	42.86%	**100%**	28.57%	57.14%	0.00%
	Top-3 Acc	**100%**	28.57%	57.14%	71.43%	**100%**	**100%**	42.86%	**100%**	42.86%	71.43%	0.00%
Llama3	Top-1 Acc	**100%**	0.00%	28.57%	0.00%	0.00%	71.43%	0.00%	0.00%	28.57%	0.00%	0.00%
	Top-3 Acc	**100%**	0.00%	57.14%	28.57%	0.00%	85.71%	0.00%	0.00%	28.57%	0.00%	0.00%
Doubao	Top-1 Acc	**100%**	85.71%	28.57%	57.14%	14.29%	71.43%	42.86%	85.71%	57.14%	42.86%	28.57%
	Top-3 Acc	**100%**	85.71%	42.86%	85.71%	85.71%	85.71%	42.86%	85.71%	57.14%	85.71%	28.57%
qwen2.5	Top-1 Acc	**100%**	**100%**	42.86%	57.14%	71.43%	**100%**	85.71%	85.71%	57.14%	85.71%	71.43%
	Top-3 Acc	**100%**	**100%**	57.14%	85.71%	**100%**	**100%**	**100%**	**100%**	**71.43%**	**100%**	85.71%
Gemini	Top-1 Acc	85.71%	85.71%	28.57%	42.86%	71.43%	71.43%	28.57%	**100%**	57.14%	71.43%	57.14%
	Top-3 Acc	85.71%	85.71%	57.14%	71.43%	85.71%	71.43%	28.57%	**100%**	57.14%	**100%**	85.71%
Moonshot	Top-1 Acc	57.14%	57.14%	14.29%	28.57%	71.43%	85.71%	57.14%	**100%**	28.57%	71.43%	**100%**
	Top-3 Acc	57.14%	57.14%	57.14%	42.86%	71.43%	**100%**	57.14%	**100%**	42.86%	71.43%	**100%**
Claude3	Top-1 Acc	**100%**	**100%**	28.57%	57.14%	71.43%	85.71%	57.14%	**100%**	28.57%	71.43%	42.86%
	Top-3 Acc	**100%**	**100%**	57.14%	**100%**	71.43%	**100%**	57.14%	**100%**	42.86%	71.43%	42.86%
GPT-4o	Top-1 Acc	85.71%	85.71%	28.57%	57.14%	57.14%	42.86%	71.43%	**100%**	57.14%	85.71%	14.29%
	Top-3 Acc	**100%**	85.71%	**71.43%**	85.71%	71.43%	71.43%	71.43%	**100%**	57.14%	85.71%	14.29%
ERNIE3.5	Top-1 Acc	**100%**	71.43%	28.57%	28.57%	42.86%	57.14%	28.57%	85.71%	42.86%	57.14%	0.00%
	Top-3 Acc	**100%**	71.43%	57.14%	71.43%	57.14%	**100%**	42.86%	**100%**	57.14%	57.14%	0.00%

6.2.3　推理时间消耗与经济成本评估

本节记录并分析了十个大模型在对单个用户进行个体身份信息推理任务时的单次推理时间消耗与经济成本，以揭示大模型在隐私数据推理任务中的易用性。图 8 展示了推理时间消耗（红色折线，单位：秒）与经济成本（蓝色折线，单位：元）的变化趋势。经济成本是通过综合输入和输出 token 的数量，并依据各大模型厂商的定价标准进行估算。对于 ChatGLM、Llama2 和 Llama3 等开源本地部署模型，由于未提供商业定价，因此未统计其经济成本。

在不同模型性能的横向比较中，理想的情况是模型能够在保证推理质量的同时，具备更快的推理速度和更低的经济成本。实验结果表明，GPT-4o 在这两个维度上表现最佳。尽管 ChatGLM、Llama2 和 Llama3 的推理速度较快，但由于其输出 token 数量较少，生成回答内容相对简短，限制了任务完成质量的评估。因此，尽管这些模型在推理速度上具有优势，但在推理质量上未能完全满足高效隐私推理任务的需求。

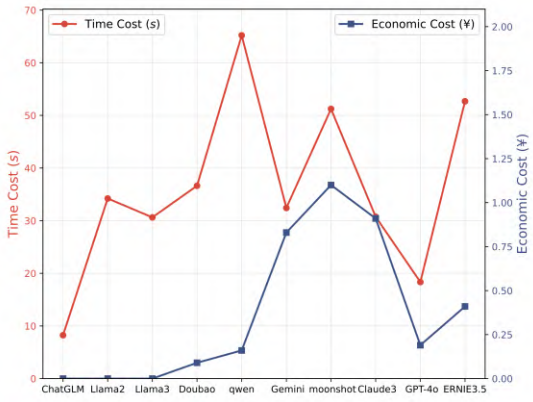

图 8　个体身份信息推理任务大模型时间与成本开销比较

综合来看，大模型的平均单次推理时间为 36 秒，平均经济成本为 0.53 元。假设人类完成同等任务需要约 1.5 小时，按照平均时薪 24 元/小时计算，**大模型在时间效率上约为人类的 150 倍，在经济效益上约为人类的 68 倍。**这表明，大模型能够以显著优于人类的效率完成个体身份信息推理任务，这充分展现了大模型在信息提取任务中的易用性以及可行性。

6.2.4　数据匿名评估

图 9 展示了数据在是否匿名化情况下，十个大模型对性别、年龄、居住地、出生地及亲友五个属性的前三猜测准确率均值（Top-3 Acc）的对比结果。其中，蓝色折线表示大模型在非匿名化数据条件下的平均准确率，红色折线则表示经过数据匿名隐私保护操作后的平均准确率。折线图中的阴影区域反映了五个属性推理准确率的方差，虚线表示多属性推理准确率的总体均值。

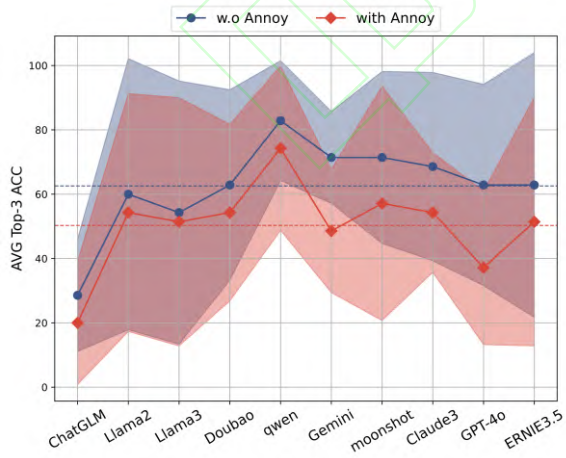

图 9　数据匿名对大模型推理准确性的影响比较

从图中可以看出，应用 Presidio 匿名化工具后，大模型的推理准确率整体有所降低。具体而言，多模型在五个属性上的平均 Top-3 Acc 从 62.57%降至 50.29%，下降约 12.28 个百分点。这表明匿名化机制能够在一定程度上限制模型对隐私属性的推理能力。然而，这种降低幅度相对有限，大模型仍能捕捉大部分隐性特征，完成对个体身份信息的推理。值得注意的是，不同模型对匿名化的敏感性有所差异。Gemini 和 GPT-4o 模型受匿名化的影响最大，其准确率下降幅度明显。这可能是因为这两种模型在隐私推理任务中更依赖语义细节，而匿名化操作通过模糊或替代关键信息，显著削弱了模型的推理能力。

总体而言，Presidio 匿名化工具在降低大模型对敏感属性推理准确率上取得了一定效果，但其保护强度仍有改进空间。尤其是在处理语义信息丰富、隐性特征显著的评论文本时，匿名化技术需要进一步优化，以更有效地提升隐私保护水平。

6.2.5　差分隐私评估

图 10 展示了不同模型隐私预算ϵ分别为 400、800、1200 以及∞时对用户 1 第一猜测准确率（Top-1 Acc）的对比结果。其中，ϵ值越小，代表数据集中加入的噪声越多，隐私保护强度越高，$\epsilon = \infty$则表示用户数据集中未加入差分隐私噪声，即无任何隐私保护策略。

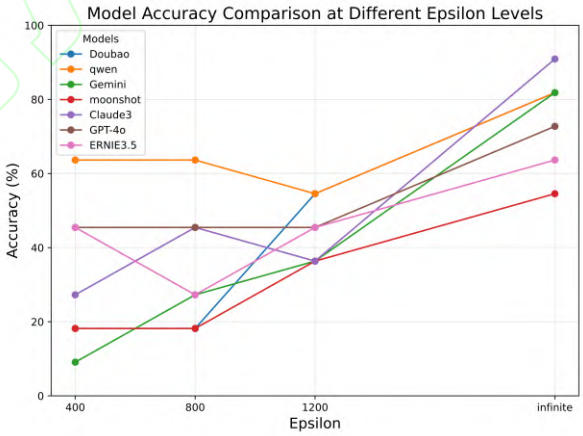

图 10　不同隐私预算下个体身份信息推理准确率对比

从整体趋势来看，差分隐私技术在隐私保护方面的效果更为显著。随着ϵ的减小，各模型的推理准确率普遍呈下降趋势。大多数模型的推理准确率降幅在 18%至 36%之间，表明差分隐私噪声能够有效干扰模型对隐私属性的推理能力。尤其值得注意的是，Gemini 模型对差分隐私噪声表现出极高的敏感性，在ϵ从 400 增加至∞的过程中，推理准确率下降幅度达到 72.73%；Claude3 模型的下降幅度也高达 63.64%。这些结果充分验证了差分隐私技术通过引入随机性显著削弱模型对敏感信息的识别能力。

然而，尽管更小的ϵ值能够提供更高强度的隐私保护，在 5.2.4 小节中也展示了其显著降低文本生成质量，导致文本的可读性和语义完整性遭到严重破坏。

总体而言，差分隐私技术在保护用户隐私数据方面展现了突出的优势，其通过调节隐私预算ϵ值可实现隐私保护强度与文本生成质量的动态平衡。然而，当前差分隐私机制对文本质量的削弱仍然较为明显。未来的研究可以进一步探索更加精细化的差分隐私策略，优化噪声注入方式，以在保证隐私保护强度的同时进一步提升文本生成质量，为隐私保护技术的实际应用提供更有力的支持。

6.3 用户画像推理实验结果

6.3.1 案例说明

用户画像推理任务利用大模型进行用户画像的过程包括分词、词频统计与身份特征信息筛选，并结合个体隐私属性信息进一步提炼出用户常用的词汇与关注的主题。图 11 从左到右分别展示了词云生成工具 BiliScope 与 GPT-4o 以及 Doubao 大模型生成的用户词云画像的结果对比，该图展示数码区视频博主个体 7 的用户词云画像。

图 11 用户词云画像结果对比

6.3.2 推理准确性评估

为全面评估用户画像词云生成任务中大模型的推理生成质量，本文设计了以下三个关键指标：(1)多样性分数：衡量模型生成的词云是否满足多样性要求，即词云中是否包含至少 120 个独立词汇；(2)词频相关性：评估词云中生成的词频与原始文本中实际词频之间的匹配度，相关性越高表明模型对文本特征的捕捉能力越强；(3)敏感词识别准确性分数：对词云中是否准确包含隐私相关词汇进行打分；若生成结果能够准确识别并包含隐私相关词汇，则增加得分。综合上述三个评估指标，可形式化表述模型推理准确性的综合得分如式(12)所示：

$$s = \alpha \cdot \text{Diversity} + \beta \cdot \text{Correlation} + \gamma \cdot \text{Sensitive} = \alpha \cdot \min\left(1, \frac{N_g}{N_t}\right) + \beta \cdot \frac{N_w}{N_g} + \gamma \cdot \frac{N_m}{N_s} \quad (12)$$

其中，N_g为大模型生成用户词云的数量，N_t为预设的词云生成目标词汇数量。N_w为在词云和原始文本中均存在的匹配词汇数量。N_m为生成词云中准确识别出的敏感词数量，N_s为原始敏感词列表的总词汇数量。除此之外，α、β、γ为权重参数，用于调整三项指标在总分中的相对重要性。通过这一综合评估框架，可以更直观地比较不同大模型在用户画像词云生成任务中的表现，量化其在隐私保护关键词提取与特征抽查能力。

值得注意的是，相较于 BiliScope 的生成结果，**基于大模型的词云生成在主题归纳和关键词提取上表现更为智能和高效。**这主要是因为大模型通过庞大的训练数据，能够更加准确地识别出具有实际意义的词汇，同时自动过滤掉无关的停用词(如“的”、“是”等)，减少了对分析结果的干扰。相比之下，传统方法通常需要手动剔除停用词，这不仅耗时，还可能因为人工操作的主观性导致某些词汇被误删或错误保留。除此之外，大模型在主题提炼上也展现出了显著优势。通过对大量语料的上下文理解，它能够更准确地识别出文本中隐含的主题。这种能力使得大模型在捕捉用户兴趣和个性化特征时表现得更为精准，从而为用户画像提供了更深入且全面的分析结果。

由于 ChatGLM 无法进行用户画像推理任务，而 Llama3 的输出为英文词云，难以直接进行比较，本文仅对其他八个大模型进行了综合评估。图 12 中柱状图展示了各模型在三个评估维度上的综合得分对比，而折线图则呈现了每个维度具体得分的变化趋势。从评估结果来看，表现较为优异的模型是 Gemini，其综合三个评估维度的得分领先于其他模型。同时，Doubao 和 Claude3 在模型输出多样性上表现较好，能够有效避免词汇重复，保证输出词云的丰富性。相比之下，qwen2.5 模型在多样性维度上的得分较低，主要由于其生成的词云中存在过多重复词汇，这可能源于其在处理长尾词频分布时

策略的不足。此外，在词频相关性维度上，qwen2.5 和 Doubao 表现突出，这或许得益于其研发公司在推荐算法领域的深厚数据积累，使其更能捕捉文本中的隐性特征，生成与用户画像高度相关的词云内容。

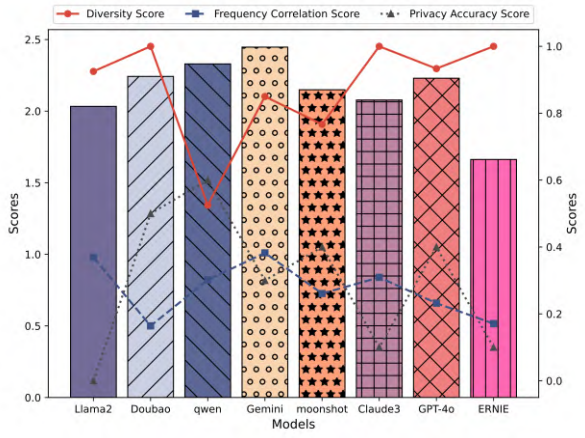

图 12 大模型用户画像任务推理准确性对比

整体来看，各大模型在用户画像推理任务中均展现了较强的能力，尤其是在隐私相关词汇的识别上表现显著。这表明，大模型在复杂的用户画像推理任务中具备较高的应用价值，可为推荐系统的优化和隐私保护研究提供有力支持。然而，部分模型在低频词汇覆盖和去重策略上仍有优化空间。

6.3.3 推理时间消耗与经济成本评估

图 13 展示了大模型在用户画像推理任务中消耗时间与经济成本对比分析。由于用户画像任务需要生成大量用户词云，涉及较多输出 token，因此在时间与经济消耗上明显高于个体身份信息推理任务和粉丝画像推理任务。

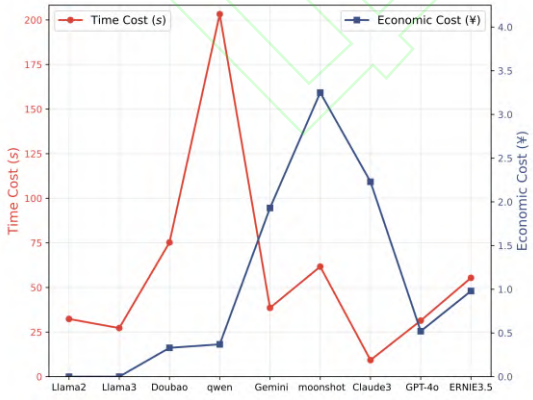

图 13 用户画像推理任务大模型时间与成本开销比较

在所有参评模型中，Doubao、GPT-4o 以及 ERNIE3.5 在时间与经济成本上表现均衡，尤以 GPT-4o 的表现最佳。GPT-4o 的单次推理时间平均为 4.46 秒，经济成本为 0.59 元，展现出极高的效率与性价比。整体来看，大模型在用户画像推理任务中的平均推理时间为 59.38 秒，经济成本为 1.07 元。相较之下，假设人类完成同样的任务需要约 2 小时，按中国平均时薪 24 元/小时计算，**大模型在推理速度上是人类的 121 倍，在经济效益上约为人类的 45 倍。**这一结果表明，大模型在用户画像推理任务中优异的关键词提取与归纳总结能力，且同时具备显著的时间与经济效益优势。

6.3.4 数据匿名评估

表 7 展示了在数据匿名处理与否的情况下，大模型在用户画像任务中词频相关性和敏感词识别准确性上的表现对比。在词频相关性方面，数据匿名对大多数模型的性能产生了一定的削弱，平均降低幅度约为 0.1。然而，Doubao 和 GPT-4o 模型在这一指标上的表现未受到影响，这可能表明它们对匿名化工具具有更强的抗干扰能力，能够更好地捕捉本文词频词性特征。在敏感词识别准确性上，所有模型的准确性均有所下降，这一结果与本文 6.3.2 小节的发现相一致，即数据匿名隐私保护策略会影响大模型对隐式敏感信息的捕捉能力。

总体而言，数据匿名虽然会降低用户画像推理生成质量，但其影响相对有限。这种削弱主要源于匿名化过程中对文本词频的调整、可能模糊或替换某些关键词汇，从而影响模型对敏感词的准确识别。然而，这种影响相对有限，且一些模型在应对匿名化数据时展现了更强的鲁棒性。

表 7 数据匿名对用户画像推理准确性的影响

模型	词频相关性		敏感词识别准确性	
	无防御	匿名化	无防御	匿名化
Llama2	0.3694	0.2193	0.0000	0.0000
Doubao	0.1642	0.2708	0.5000	0.4000
qwen2.5	0.3016	0.2041	0.6000	0.0000
Gemini	0.3824	0.2195	0.3000	0.1000
moonshot	0.2609	0.1971	0.4000	0.0000
Claude3	0.3092	0.2866	0.1000	0.0000
GPT-4o	0.2321	0.3173	0.4000	0.2000
ERNIE3.5	0.1708	0.1647	0.0001	0.0000

6.3.5 差分隐私评估

本节探讨差分隐私噪声注入量对用户画像推理准确性的影响，图 14 展示了在隐私预算ϵ分别为 400、800、1200 和无穷大时（ϵ越大，隐私保护效果越弱），大模型在用户画像任务中词频相关性和敏感词识别准确性方面的表现对比。

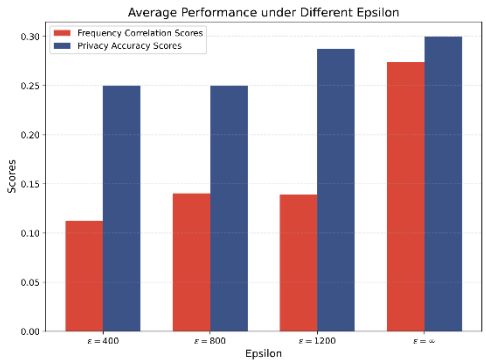

图 14　不同隐私预算下用户画像推理准确率对比

从图 14 中可以看出，随着隐私预算ϵ的减小，噪声强度增加，差分隐私机制对词频相关性的削弱程度愈加显著。具体而言，词频相关性均值从原本的 0.2938 大幅下滑至 0.1127。这一现象的产生，主要是因为噪声的注入对原本稳定的词向量空间造成了显著扰动。在用户画像推理任务中，由于这类任务对原始数据的词频结构有着极高的依赖性，词向量空间的扰动不可避免地导致了推理准确性的显著降低。同时，我们的研究还发现，在敏感词识别方面，随着噪声强度的持续增强，模型的识别准确性也呈现出稳定的下降趋势，准确率均值由原本的 0.3 降低至 0.25。这一结果与本文 6.3.2 小节中的研究发现高度一致，从而进一步验证了差分隐私技术在提升用户隐私保护水平的同时，确实对模型捕捉和处理隐私相关信息的能力造成一定负面影响。

总体而言，差分隐私通过注入噪声显著提升了数据保护能力，但这种保护措施不可避免地对用户数据的词向量空间产生干扰，削弱了原始数据集的语义完整性，为后续研究可考虑如何在增强隐私保护的同时最大限度地保留数据语义特征。

6.4　粉丝画像推理实验结果

6.4.1　案例说明

下图展示以视频博主个体 6 为研究对象，基于少样本思维链指令利用大模型对该博主进行粉丝画像的分析结果。左侧部分展示了粉丝画像任务的设计指令以及粉丝群体的真实标签，而右侧则为大模型输出的预测结果。从图 15 中可以看出，在性别比例和年龄分布的预测任务中，即使没有提供任何先验信息，大模型依然能够基于其在训练过程中积累的知识进行较为合理的推测。特别是在性别比例的估算中，模型的推测结果与实际标签非常接近。这表明，大模型具备从博主的语言风格、话题选择以及其他隐性信息中推断出粉丝群体特征的能力。这种能力得益于大模型在处理海量数据时学习到的广泛模式和隐藏关系，使其能在缺乏直接信息的情况下，基于上下文和已有的知识库进行有效的推理和预测。

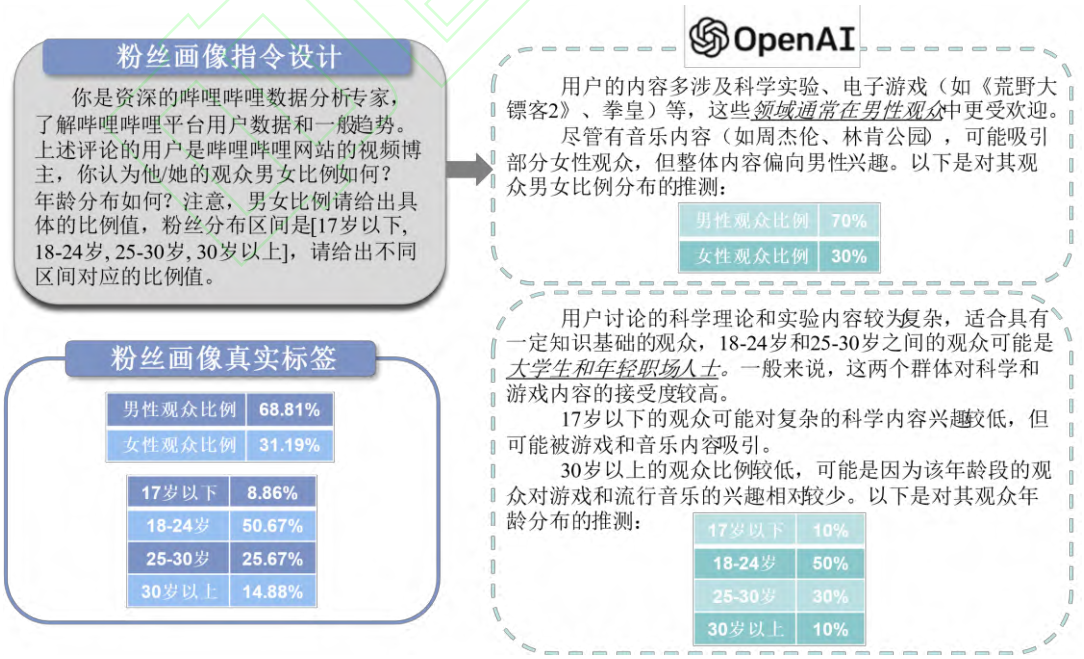

图 15　粉丝画像推理实例

6.4.2 推理准确性评估

在粉丝画像推理任务中，由于 ChatGLM 无法输出期望的预测结果，本文对其他九个大模型的推理准确性进行了详细的比较分析。

针对粉丝年龄区间的预测任务，我们选用了余弦相似度(Cosine Similarity)作为评估指标。余弦相似度越高，意味着模型的预测结果与真实标签之间的相似度越大，从而评估模型的预测准确性。通过图 16 的实验结果可知，ERNIE3.5、Gemini 和 Claude3 大模型在推理准确性上表现尤为突出，其对应的余弦相似度分别可达 0.978、0.970 和 0.960，表明这三个模型在预测粉丝年龄区间时，能够提供较为精确的预测结果。此外，Gemini 模型在不同用户群体上的预测相似度也表现出较为一致的结果，反应其在处理多样化数据时的优异鲁棒性。从整体实验结果来

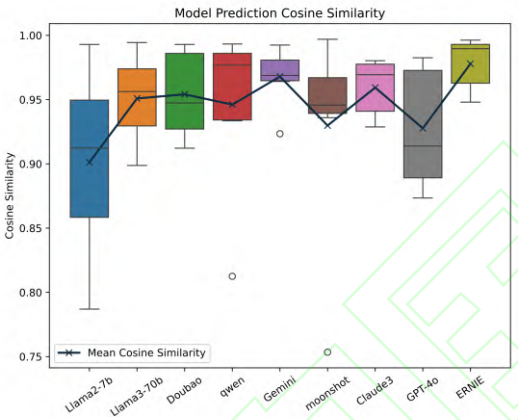

图 16 粉丝画像年龄区间推理余弦相似度对比

看，**所有参与测试的模型在粉丝年龄区间预测任务中的平均余弦相似度为 0.946**，表明大规模语言模型在这一任务中的预测精度整体较高，能够较为准确地推测粉丝的年龄区间。

在粉丝性别区间的预测任务中，选用了均方误差（MSE）作为评估指标。均方误差越小，意味着模型的预测结果越接近真实值，从而反映出其推理预测能力的强弱。实验结果由图 17 所示，Gemini 和 qwen2.5 模型在该任务中的表现最为优秀，其均方误差分别为 0.0038 和 0.0026。这一结果表明，这两个模型能够有效地捕捉隐式用户特征，并在性别区间预测方面展现出较强的推理能力。此外，qwen2.5 与 GPT-4o 在不同数据集上的预测效果也表现出较为稳定的趋势，进一步证明了其在处理此类任务中的可靠性。总体而言，**所有参与测试的模型在粉丝性别区间预测任务中的平均均方误差为 0.024，表明大规模语言模型在性别预测任务上的适用性。**

从总体趋势来看，大模型对粉丝年龄区间与性别属性的预测误差较小，表明模型能够有效捕捉隐式用户特征，利用预训练时的先验知识表现出较强的推理预测能力。

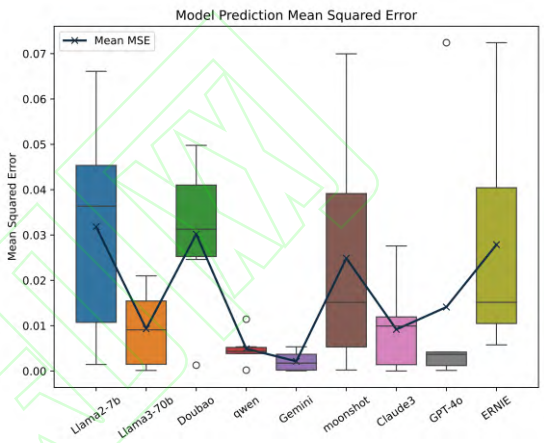

图 17 粉丝画像性别区间推理均方误差对比

6.4.3 推理时间消耗与经济成本评估

图 18 记录并分析九个模型在粉丝画像推理任务中的单次推理时间消耗与经济成本。由于粉丝画像推理任务基于个体身份信息的推理，因此其所需经济成本相较于个体身份信息推理任务更高。

与个体身份信息推理任务相似，GPT-4o 模型在粉丝画像任务中的表现仍然突出，展现了较低的经济开销与时间消耗。尽管 Doubao 模型的时间消耗较低，但其输出内容相对简略且推理结果不够全面，影响了推理质量。

在粉丝画像推理任务中，大模型的平均单次推理时间为 17 秒，平均经济成本为 0.85 元。假设人类完成同等任务需要约 2 小时，按照平均时薪 24 元/小时计算，**大模型在时间效率上约为人类的 428 倍，在经济效益上约为人类的 57 倍。**大模型在粉丝推理任务中的时间效益与成本效益突显了其在大规模数据预测分析中的巨大潜力。

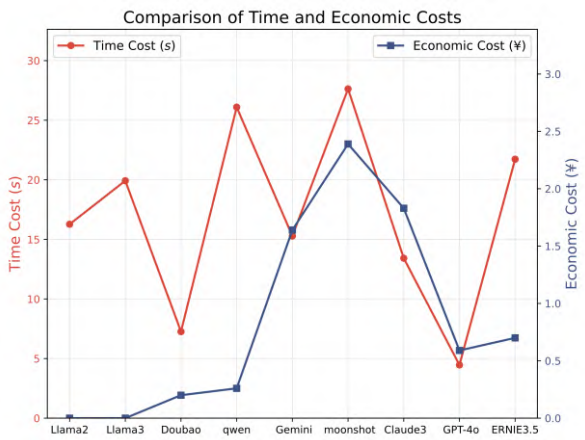

图 18　粉丝画像推理任务大模型时间与成本开销比较

6.4.4　数据匿名评估

本节分别对比了不同模型在数据匿名前后对视频博主个体 6 进行粉丝画像推理任务时的年龄区间余弦相似度和性别区间均方误差对比。余弦相似度值越大表明模型预测结果与真实标签的相似度越高；均方误差值越小表示模型的预测精度越高。

表 8　数据匿名对粉丝画像推理准确性的影响

模型	年龄（Cosine Similarity）		性别（MSE）	
	无防御	匿名化	无防御	匿名化
Llama2-	0.9122	0.9330	0.0015	0.0015
Llama3	0.9670	0.9209	0.0031	0.0029
Doubao	0.9122	0.9807	0.0354	0.0078
qwen2.5	0.9819	0.9145	0.0038	0.0001
Gemini	0.9641	0.9546	0.0038	0.0038
moonshot	0.9457	0.9503	0.0354	0.0011
Claude3	0.9767	0.9572	0.0012	0.0038
GPT-4o	0.9801	0.9403	0.0001	0.0078
ERNIE3.5	0.9926	0.9606	0.0078	0.0000

实验结果如表 8 所示，数据匿名对余弦相似度的影响较小，最大差异为 0.0674；而对均方误差的影响也较为有限，最大差异为 0.0343。这一结果表明，尽管数据匿名技术在保护特定属性的隐私方面具有重要作用，但在粉丝年龄区间和性别区间等高层次推理预测任务中，模型往往依赖于预训练时的先验知识以及用户评论中的隐性特征进行上下文信息关联推理。由于数据匿名技术主要针对直接的个人隐私信息，因此其对推理结果的整体影响较为有限。

6.4.5　差分隐私评估

图 19 展示了在不同隐私预算ϵ值下，差分隐私噪声加入前后在粉丝画像推理任务中，年龄区间余弦相似度和性别区间均方误差的对比。方差综合了不同模型的结果，揭示了隐私预算与模型推理表现之间的关系。

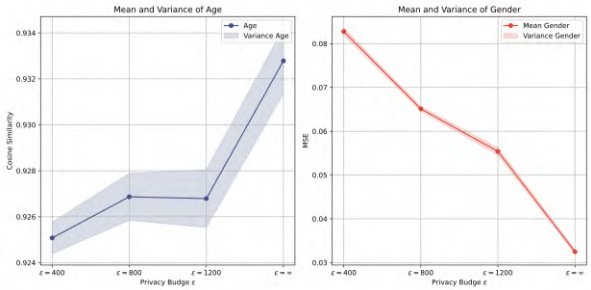

图 19　不同隐私预算下粉丝画像推理准确率对比

由左右两图结果可以看出，隐私预算与推理准确率之间存在潜在的相关性：随着隐私预算的增加，隐私保护能力减弱，大模型的推理预测能力随之增强。具体来说，当隐私预算ϵ为 400 时，相较于未加入噪声的情况，年龄区间余弦相似度增加了 0.0077，性别区间均方误差降低了 0.0004。然而，尽管差分隐私通过加入噪声在一定程度上削弱了模型的推理能力，但这种影响是有限的。这一现象可以解释为，差分隐私噪声的加入旨在增加数据的混淆性，从而保护用户隐私。然而，大模型能够适应这类词向量空间上的扰动并很好地捕捉用户文本中的隐式特征，因此推理预测结果的准确性未受到显著影响。

7 总结

针对大模型在中文数据隐私推理能力的综合评估需求，本文首先构建了一个包含超过五万条评论的多样化、真实场景数据集 BiliPrivacy。接着，基于该数据集，设计了个体身份信息推理、用户画像推理和粉丝画像推理三个任务，以全面评估大模型对隐式身份信息抽取、关键词提取与归纳能力以及更深层次推理预测能力的表现，并分析实际任务进行时的推理准确率以及对应的经济与时间开销，从而评估大模型在隐私推理攻击中的可行性。最后，本文还全面考察了数据匿名和差分隐私两种隐私保护方法对大模型推理能力的影响和效果，为后续隐私保护策略研究提供参考。总体来看，本研究揭示了大模型在隐私推理任务中时间与经济成本均显著优于人类，能够智能高效地完成抽取、归纳与预测的隐私推理任务，且目前隐私保护社区流行的数据匿名以及差分隐私技术难以应对大模型推

理带来的全新挑战，这些发现将进一步推动大模型推理攻击评估以及对应隐私保护策略的研究。

尽管本研究展示了大模型在中文数据中的强大推理能力，但仍存在以下几点局限性：首次，本研究采集数据集仅来自哔哩哔哩平台的视频博主评论，这种单一数据来源可能限制研究结论的广泛适用性。其次，尽管大模型在隐式信息推理方面展现了强大能力，但某些隐私属性(如亲友关系和出生地)的推理准确率仍然较低。其外，大模型进行推理任务本身还存在一些缺陷，比如大模型可能会记住特定指令案例，也存在推理信息不全的情况。最后，由于外部知识库构建的困难以及文本数据的非结构化特性，本研究未对增强检索生成（RAG）、知识工程等先进技术在大模型推理中的应用进行深入探讨。

对于大模型隐私推理的研究，未来工作将主要围绕以下几个方法展开：

（1）多模态数据融合推理：研究如何结合用户产生的文本、图像、视频等多模态数据，利用现有先进多模态大模型，实现对用户隐私数据的综合关联推理，探索多模态大模型隐私推理新范式。

（2）差分隐私保护技术：在充分确保用户隐私安全的基础上，将重点研究如何减少差分隐私保护技术对用户语义的损害。虽然现有的差分隐私技术已被证实为有效的隐私保护手段，但其对数据语义的影响仍需关注。因此，需要设计更为精细的隐私保护策略，以确保推理过程中用户的敏感信息得到严格保密，同时维持推理结果的精确性和有效性。

（3）针对隐私保护的安全对齐方法：研究将从模型层面出发，探索如何确保大模型输出的隐私安全性，并使其与社会价值观相符。例如，OpenAI最新提出的安全对齐方法，在模型推理过程中明确设定伦理规范。同时，针对当前隐私保护领域缺乏的安全对齐手段，进行深入研究与探索。

（4）RAG 与知识工程等增强技术赋能：深入研究 RAG、知识工程等技术如何赋能大模型推理通过探索这些技术与大模型的有效融合，有望显著提升推理的准确性和全面性。同时，考虑采用指令调优策略，以进一步优化模型的推理性能，从而实现更为精确和全面的隐私信息推理。

随着大模型在隐私推理中的广泛应用，未来研究还应关注如何在合规和伦理框架下使用此类技术，并加强在人工智能发展背景下对数据安全和用户隐私的重视，推动设计更多有效隐私保护技术。

致　谢　感谢为本文原稿提供宝贵修改意见的匿名审稿人。感谢国家自然科学基金对本文工作的资助。感谢十名人类志愿者进行数据集标注。

参 考 文 献

[1] IBM data breach report 2024, https://www.ibm.com/reports/data-breach

[2] Zhang F, Shi S, Zhu Y, et al. Oag-bench: a human-curated benchmark for academic graph mining. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024. 6214–6225

[3] Liu X J. Study on user profiling based on behavior logs. Ph.D. thesis, Hefei University of Technology, Hefei, 2022. (in Chinese)
(刘啸剑. 基于行为日志的用户画像研究. 博士学位论文. 合肥: 合肥工业大学, 2022)

[4] Che W X, Dou Z C, Feng e a Y S. Natural language processing in the era of large models: Challenges, opportunities, and developments. Sci Sin Inform, 2023, 53: 1645–1687. (in Chinese)
(车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战, 机遇与发 展. 中国科学: 信息科学, 2023, 53(9): 1645–1687)

[5] Yao Y, Duan J, Xu K, et al. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 2024: 100211.

[6] Hartmann V, Suri A, Bindschaedler V, et al. Sok: Memorization in general-purpose large language models. arXiv preprint arXiv:2310.18362, 2023.

[7] Steve Wilson A D. Owasp top 10 for large language model applications, https://owasp.org/www-project-top-10-for-large-language-model-applications/.

[8] Mireshghallah N, Kim H, Zhou X, et al. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory//Proceedings of 12th International Conference on Learning Representations, Vienna, Austria, 2024.

[9] Staab R, Vero M, Balunović M, et al. Beyond memorization: Violating privacy via inference with large language models//Proceedings of 12th International Conference on Learning Representations, Vienna, Austria, 2024.

[10] Rangel F, Rosso P, Potthast M, et al. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working notes papers of the CLEF, 2017, 48.

[11] Takahashi T, Tahara T, Nagatani K, et al. Text and image synergy with feature cross technique for gender identification: Notebook for pan at clef 2018. In: CLEF (Working Notes), 2018.

[12] Yukhymenko H, Staab R, Vero M, et al. A synthetic dataset for personal attribute inference. arXiv preprint arXiv:2406.07217, 2024.

[13] Dou Y, Krsek I, Naous T, et al. Reducing privacy risks in online

self-disclosures with language models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand 2024.

[14] Butt J S .The General Data Protection Regulation of 2016 (GDPR) Meets its Sibling the Artificial Intelligence Act of 2024: A Power Couple, or a Clash of Titans?[J].Acta Universitatis Danubius: Juridica, 2024, 20(2).

[15] Futomlinson Y .Personal data protection in CHina[J].China Business Review, 2002(7).

[16] Cavoukian A, et al. Privacy by design: The seven foundational principles,https://iapp.org/resources/article/privacy-by-design-the-7-foundational-principles, 2021.

[17] Olabanji S O, Oladoyinbo O B, Asonze C U, et al. Effect of adopting AI to explore big data on personally identifiable information (PII) for financial and economic data transformation. Available at SSRN 4739227, 2024.

[18] Mishra P, Del Tredici M, Yannakoudakis H, et al. Author profiling for abuse detection//In: Proceedings of the 27th international conference on computational linguistics, Santa Fe, New Mexico, USA , 2018. 1088–1098

[19] Chakraborty K, Bhattacharyya S, Bag R. A survey of sentiment analysis from social media data. IEEE Transactions on Computational Social Systems, 2020, 7: 450–464

[20] Purificato E, Boratto L, De Luca E W. User modeling and user profiling: A comprehensive survey. arXiv preprint arXiv:2402.09660, 2024.

[21] Strohm F, Bâce M, Bulling A. Learning user embeddings from human gaze for personalised saliency prediction. Proc. ACM Hum.-Comput. Interact., 2024, 8. URL https://doi.org/10.1145/3655603.

[22] Mikros G, Boumparis D. Cross-linguistic authorship attribution and gender profiling. machine translation as a method for bridging the language gap. Digital Scholarship in the Humanities, 2024.

[23] Yan, Biwei, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. "On protecting the data privacy of large language models (llms): A survey." arXiv preprint arXiv:2403.05156 (2024).

[24] Xu J, Stokes J W, McDonald G, et al. Autoattacker: A large language model guided system to implement automatic cyber-attacks. arXiv preprint arXiv:2403.01038, 2024.

[25] Ferdaus M M, Abdelguerfi M, Ioup E, et al. Towards trustworthy ai: A review of ethical and robust large language models. arXiv preprint arXiv:2407.13934, 2024.

[26] Mendes E, Chen Y, Hays J, et al. Granular privacy control for geolocation with vision language models. arXiv preprint arXiv:2407.04952, 2024.

[27] Brahem M, Watissee J, Eichler C, et al. retellme: Design rules for using large language models to protect the privacy of individuals in their textual contributions//In: International Workshop on Data Privacy Management@ ESORICS, 2024.

[28] Tömekçe, Batuhan, Mark Vero, Robin Staab, and Martin Vechev. "Private Attribute Inference from Images with Vision-Language Models." arXiv preprint arXiv:2404.10618 (2024).

[29] Zhao Y, Chen J. A survey on differential privacy for unstructured data content. ACM Computing Surveys (CSUR). 2022 Sep 14;54(10s):1-28.

[30] Mendels O, Peled C, Vaisman Levy N, et al. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images, 2018. https://microsoft.github.io/presidio.

[31] Moffie M, Mor D, Asaf S, et al. Next generation data masking engine//In: International Workshop on Data Privacy Management, 2021. 152–160

[32] Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In The 8th International Conference on Principles of Security and Trust (POST'19). 123–148.

[33] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized Mahalanobis metric. In Proc. of the 2nd Workshop on PrivateNLP at the 25th Conference on Empirical Methods in Natural Language Processing (EMNLP'20). 7–17.

[34] Igamberdiev, Timour, and Ivan Habernal. DP-BART for privatized text rewriting under local differential privacy//Findings of the Association for Computational Linguistics, Toronto, Canada, 2023. 13914–13934

[35] Krishna, Satyapriya, Rahul Gupta, and Christophe Dupuy. "ADePT: Auto-encoder based differentially private text transformation//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Kyiv, Ukraine, 2021. 2435–2439

[36] Chen R, Liu H, Altmann G. Entropy in different text types. Digital Scholarship in the Humanities, 2017, 32: 528–542.

[37] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, pp.24824-24837.

[38] Nobre C, Gehlenborg N, Coon H, et al. Lineage: Visualizing multivariate clinical data in genealogy graphs. IEEE transactions on visualization and computer graphics, 2018, 25: 1543–1558

[39] Xu K C H. Algorithm paradox and institutional response: An empirical study based on user perception of algorithm application. J Shandong Univ Philos Soc Sci, 2022: 13. .(in Chinese)

(许可，程华. 算法悖论与制度因应——基于用户算法应用感知的实证研 究. 山东大学学报：哲学社会科学版, 2022, (6): 13)

Du Mengyao, Ph.D candidate. Her main research interests include privacy computing and AI security.

Li Qingming, Ph.D. Her main research interest includes AI security.

Zhang Miao, Ph.D. Assistant Researcher. His main research interest includes system simulation.

Chen Xi, Ph.D candidate . His main research interest includes federated learning.

Li Xinmeng, Ph.D. His main research interest includes NLP (Natural Language Processing).

Yin Quanjun, Ph.D. Professor. His main research interest includes intelligent simulation.

Ji Shouling, Ph.D . Professor. His main research interest includes information security and AI security.

Background

The convergence of big data and artificial intelligence technologies has exacerbated the risks of individual privacy breaches. More critically, the reasoning and sentiment analysis capabilities exhibited by large language models, particularly those represented by GPT, have heightened these risks to unprecedented levels. The frequency of such privacy breaches has significantly undermined users' trust and acceptance of technology, severely hindering the advancement of digital infrastructure in our country.

Currently, research on the privacy inference capabilities of large language models is still in its nascent stages. While some scholars have conducted preliminary explorations of these risks using English datasets, analyses of the privacy inference capabilities of large models remain insufficient. Additionally, the unique characteristics of the Chinese language present further challenges in privacy inference tasks. Specifically, as a logographic language, Chinese often omits subjects or predicates, with each character potentially conveying richer information. This complexity may hinder the ability of large models to capture implicit information during privacy inference.

To address this gap, this paper introduces the BiliPrivacy dataset, which consists of over 50,000 comment entries collected from the Bilibili platform, covering 20 video creators, and annotated by 10 volunteers. Rich in both information and diversity, this dataset is one of the most comprehensive Chinese privacy inference datasets available to date. Building on this dataset, the study employs few-shot reasoning chain-of-thought instruction fine-tuning to design three core tasks: individual identity information inference, user profiling, and fan profiling. Additionally, the study investigates the impact of data anonymization and differential privacy techniques on the reasoning capabilities of large models. Experimental results demonstrate that the large model is capable of generating inference results in an average of 37.46 seconds at a cost of 0.82 yuan. Specifically, the model achieved 90.91% accuracy in extracting implicit identity information and exhibited superior performance in keyword extraction and summarization, with enhanced intelligence and efficiency in terms of diversity, word frequency correlation, and sensitive word identification. In the inference prediction tasks, even without prior information, the model was able to make reasonable inferences based on its general knowledge, achieving an average cosine similarity of 0.946 and a mean squared error of 0.024 in predicting fan age and gender. Finally, a comprehensive analysis of data anonymization and differential privacy techniques revealed that while privacy protection strategies do impose some limitations on the model's reasoning ability, these constraints are relatively minor, with notable negative effects on the utility of user data. In conclusion, this paper presents a systematic evaluation of large models' performance in privacy inference tasks using Chinese comment data, significantly contributing to advancing privacy protection research in the Chinese-language context.

This research is supported by the National Natural Science Foundation of China (62103420, 62403484).