/TemplateVersion (2025.1)

当可解释性遇上隐私性：在自然语言处理背景下对后验可解释性和差分隐私交叉点的研究

Mahdi Dhaini, Stephen Meisenbacher, Ege Erdogan, Florian Matthes, Gjergji Kasneci
Technical University of Munich
School of Computation, Information and Technology
Department of Computer Science
Munich, Germany { mahdi.dhaini, stephen.meisenbacher, ege.erdogan, matthes, gjergji.kasneci } @tum.de

Abstract

在值得信赖的自然语言处理（NLP）研究中，出现了许多重要的研究领域，包括可解释性和隐私。虽然近年来对可解释和隐私保护的 NLP 的研究兴趣显著增加，但在二者交集上的研究仍然缺乏。这在能否同时实现可解释性和隐私，或者二者是否相互矛盾的理解上留下了相当大的空白。在这项工作中，我们在 NLP 的背景下进行了一项关于隐私-可解释性权衡的实证研究，该研究以差分隐私（DP）和事后可解释性这两种流行的方法为指导。我们的研究结果展示了隐私和可解释性之间复杂的关系，这种关系是由多种因素构成的，包括下游任务的性质以及文本隐私化和可解释性方法的选择。在此，我们强调了隐私和可解释性共存的潜力，并总结了我们在这一重要交集上的发现，以供未来工作的实用建议。

Code — https://github.com/dmah10/xpnlp

1 引言

自然语言处理（NLP）的最新进展在自然语言的理解和生成方式上取得了大量和广泛的改进。这些进展以大型语言模型（LLMs）及相关技术促成的快速发展为标志，推动了在教育 (Wen et al. 2024) 、医疗 (Wang et al. 2025) 和金融 (Lee et al. 2025) 等多个领域的创新应用，并使非技术用户能够探索人工智能 (Ng et al. 2021) 的能力。然而，这些优势并非不劳而获，目前 NLP 的各个子领域正在 NLP 与多个以人为中心的话题的交叉点上进行研究，例如可解释性 (Danilevsky et al. 2020) 、隐私 (Sousa and Kern 2023) 、偏见 (Navigli, Conia, and Ross 2023) 、公平 (Chang, Prabhakaran, and Ordonez 2019) 和可持续性 (Van Wynsberghe 2021) 等领域。

尽管最近 LLM 的使用变得普及，但在任何语言模型的部署过程中，一个持续的挑战与可解释性有关，可解释性通常指的是解释和传达模型提供的决策的能力。可解释性对于模型的安全部署至关重要，特别是在确保模型输入可以（合理地）被追踪或解释时。除此之外，可解释性不仅是语言模型的一个期望特征，也是最近欧盟 AI 法案条例 (Council of European Union 2024) 规定的强制要求（主要针对高风险的 AI 系统）。一个特别有用的候选方案来满足这一要求的是事后可解释性 (Madsen et al. 2022; Danilevsky et al. 2020) ，它包括提供对传统 “black-box” 模型洞见的方法。

Accepted to AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

同样地，LLM 使用量的显著增加，尤其是用户必须与托管在外部服务器上的模型（即云端）交互的情况下，已经引发了对隐私问题日益增长的关注 (Pan et al. 2020; Wu, Duan, and Ni 2023; Gupta et al. 2023; Yan et al. 2025) 。呼吁隐私保护的声音源于越来越严格的数据保护法规（如 GDPR）；与此同时，关于保护隐私的自然语言处理的研究也屡见不鲜 (Yin and Habernal 2022) ，涉及从文本隐私化到私有模型训练 (Sousa and Kern 2023) 。这些保护隐私的技术旨在对文本数据中隐藏的直接和间接标识符进行掩盖，同时也保持下游任务和应用程序的效用 (Mattern et al. 2022; Weggenmann et al. 2022) 。一个流行的框架是差分隐私（DP）(Dwork 2006) ，通过确保任意两个文本在某种程度上无法区分，通常通过向文本表示中注入随机噪声来实现，从而为文本输入提供合情合理的否认 (Klymenko et al. 2022) 。这种 “noisification” 可以在多个层次上发生，比如在词汇层次，或者通过文档层次的重写。

最近在差分隐私自然语言处理（DP-NLP）中的许多工作侧重于平衡隐私与效用的权衡，作为评估隐私化方法 (Mattern et al. 2022; Utpala et al. 2023) 有效性的一个关键指标。其他工作则探讨了其他重要方面的权衡，如文本连贯性 (Weggenmann et al. 2022) 或用户可接受性 (Meisenbacher et al. 2025) 。另一方面，可解释的自然语言处理（XNLP）越来越关注可解释性与可信自然语言处理某些方面的交集，特别是公平性。这方面的研究主要有两个方向：利用可解释性作为一种工具来检测模型中的偏见 (T.y.s.s. et al. 2024; Gallegos et al. 2024) ，以及评估可解释性方法自身的公平性 (Dhaini et al. 2025) 。然而，特别是在 NLP 领域，据我们所知，没有研究考虑到隐私和可解释性之间的交集，即隐私-可解释性权衡，具体而言，隐私化方法的应用如何影响可解释性方法的功能。我们认为这种考虑是至关重要的，特别是在同时存在可解释性和隐私法律法规的情况下。

在这项工作中，我们首次探讨了在自然语言背景下隐私与可解释性之间的相互作用。作为一个案例研究，我们选择两个流行的子领域：后验特征归因方法和差分隐私文本重写。在此过程中，我们探索了通过差分隐私以不同的隐私级别和根本不同的重写机制重写文本时，可观察到的可解释性变化。我们还考虑了下游任务细节的影响，比如模型选择、模型规模和微调任务，特别是在隐私相对于可解释性的重要性变化时，以及反之亦然。为了指导这些实验，我们定义了一个总体研究问题：

差分隐私的文本重写对微调语言模型的事后可解释性产生了什么影响，如何量化隐私与可解释性

的权衡?

我们了解到，尽管通常可以观察到隐私性和可解释性之间的明显权衡，但确实存在一些配置使得两者能够协同工作。在这方面，我们发现下游数据集和任务的因素，以及所选的差分隐私方法及其隐私预算，在量化隐私-可解释性权衡方面非常重要。这使我们为希望在这个重要交叉领域继续工作的研究人员和从业人员创建了一系列建议。

具体而言，我们的工作做出了以下贡献：

1. 我们是第一个在隐私和可解释性交叉领域，特别是在自然语言环境中进行实证研究的人。我们将我们的实验代码公开，以便于复现。
2. 我们提供了对后验可解释性和差分隐私文本重写之间复杂相互作用的见解，为在此交叉领域的更广泛研究奠定基础。
3. 我们分析我们的结果，以提出关于重要设计选择的最佳实践建议，特别是在面临可解释性和隐私需求时。

2 背景和相关工作

2.1 事后解释方法

模型不可知特征归因的事后技术由于其广泛的适用性而受到关注 (Jacovi 2023) 。这些方法论旨在通过使用基于梯度的方法（利用模型对输入的导数 (Sundararajan et al. 2017; Simonyan, Vedaldi, and Zisserman 2013) ）或基于扰动的技术 (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017) 来确定个别标记对特定输入的模型预测的相对贡献。

通过一系列全面的综述扩展了解释性 NLP 研究的重要性，来处理 NLP 可解释性 (Wallace, Gardner, and Singh 2020; Zhao et al. 2024; Madsen et al. 2022; Zini and Awad 2022) 。此外，考虑到 NLP 系统在包括教育 (Wu et al. 2024) 、医疗保健 (Johri et al. 2025) 和法律领域 (Valvoda and Cotterell 2024) 等关键应用中的部署，其中解释性要求至关重要，专门的综述出现了，它们专注于特定 NLP 应用中的可解释性，如事实验证 (Kotonya and Toni 2020) ，以及 NLP 可解释性中的具体方法论 (Mosca et al. 2022) 。这些全面的评论强调了事后方法在 NLP 应用中的广泛使用。

此外，特征归因事后解释技术作为主要方法，应用于近期文献中记录的可解释性平台和计算框架 (Arras, Osman, and Samek 2022; Li et al. 2023; Attanasio et al. 2023; Sarti et al. 2023) 。这些综合框架的特点是整合多种事后解释算法，同时适应多种数据模态的机器学习 (ML) 架构，包括预训练语言模型 (PLMs)。

2.2 自然语言处理中的差分隐私

微分隐私的概念最初是在关系数据库的背景下提出的 (Dwork 2006) ，其主要目标是保护数据集中个体的参与。更具体地说，隐私保护的实现是在某个范围内无法准确推断关于个体的信息。这通过以下不等式形式化，对于任何数据库 D_1 和 D_2 ，差别恰好在一个元素上，任何 $\varepsilon > 0$ ，任何计算或函数 $\mathcal{M}$ ，和所有 $\mathcal{S} \subseteq Range(\mathcal{M})$ ：$\frac{Pr[\mathcal{M}(D_1)\in\mathcal{S}]}{Pr[\mathcal{M}(D_2)\in\mathcal{S}]} \leq e^{\varepsilon}$ 。直观地，DP 保证在任意两个邻近数据库（仅在一个元素上不同）之间存在某种程度的不可区分性，从而保护个人。这种隐私水平由 ε 参数控制，也称为隐私预算。

这种形式的差分隐私称为 ε -DP，上述概念指的是全局差分隐私。本文关注的另一种概念是局部差分隐私 (LDP) (Kasiviswanathan et al. 2008) 。在局部设置中，我们假设中心管理员，即拥有完整数据集的人，是不被信任的。作为一种解决方案，差分隐私是在用户层面上得到保证的；然而，由于数据集整体尚未被知道，LDP 施加了更严格的不可区分性要求，即在任何潜在的邻居之间。这与全局概念不同，因为邻居数据库仅指那些由数据集 D 产生的数据库。从理论上讲，对于有限空间 $\mathcal{P}$ 和 $\mathcal{V}$ ，对于所有 $x, x' \in \mathcal{P}$ 和所有 $y \in \mathcal{V}$ ：$\frac{Pr[\mathcal{M}(x)=y]}{Pr[\mathcal{M}(x')=y]} \leq e^{\varepsilon}$ 因此，观察到的输出不能以高概率归因于特定的输入。虽然这种概念显然更为严格，但它允许在局部、单个数据点层面上量化隐私保证，而无需聚合的数据集。

将差分隐私 (DP) 引入自然语言处理 (NLP) 的领域最初带来了许多研究挑战，其中最主要的是在考虑文本数据时如何推理个体是谁，以及如何量化邻近性。尽管存在这些挑战，最近有大量研究提出了将 DP 集成到 NLP 流程中的创新方法，范围从文本匿名化和混淆到语言模型的 DP 训练。特别是在文本隐私化方面，许多新方法从重写的角度解释任务，即在 DP 保证下重写敏感输入文本以生成私密输出文本。这些方法在各种词汇层面上工作，包括词汇层面、语言模型生成过程中的标记层面以及整个文本。在比较分析时考虑这些机制的操作方式差异，考虑 DP 保证的性质变得很重要。

最近在 DP-NLP 领域的工作突出了持续存在的挑战，例如生成连贯且正确的输出 (Mattern et al. 2022) 、确保可比性 (Igamberdiev et al. 2022) 以及量化 DP 相较于非 DP 方法的益处 (Meisenbacher and Matthes 2024) 。然而，据我们所知，目前尚无研究质疑 DP 方法对文本可解释性的影响，或者更具体地说，对于在隐私化文本上训练的模型的可解释性。我们认为这是一个显著的空白，尤其是在重要领域中进行 DP 文本隐私化时，比如医疗领域，在该领域隐私与可解释性是同样重要的。

2.3 隐私与可解释性相遇

可解释机器学习中的隐私 可解释机器学习的隐私影响近年来受到关注，但仍然未被充分探索。大部分关于机器学习中可解释性与隐私交叉的工作集中于解释中固有的隐私风险。这些工作包括研究模型解释如何揭示训练数据中的敏感信息，以及如何利用会员推断攻击来利用这种情况。例如，Shokri et al. (2021) 证明了基于反向传播的解释的变化可以揭示一个数据点是否被用于训练，从而暴露会员信息。Duddu and Boutet (2022) 通过直接从模型解释中推断出诸如性别和种族等敏感属性，扩展了这些担忧，而 Liu et al. (2024) 展示了解释如何通过利用模型在属性引导的扰动下的鲁棒性差异来扩大会员推断风险。同样地，Luo et al. (2022) 表明私有输入特征可以通过 Shapley 值解释重构。其他工作则包括突出解释如何在用作辅助输入时增强模型反转攻击，(Zhao et al. 2021) ，以及引入一种依赖于原输入与其反事实对应体之间距离的反事实会员推断攻击，(Pawelczyk et al. 2023) 。

Shokri et al. (2021) 调查了基于梯度的解释方法(Gradient (Simonyan et al. 2013) 和 Integrated Gradients (Sundararajan et al. 2017)) 的隐私风险，涉及四个表格数据集和两个图像数据集。他们展示了这些方法能够泄露关于训练样本的信息，从而增加对成员推理攻

击的脆弱性，这种攻击使用完整的特征归因作为攻击模型的输入。相比之下，基于扰动的方法如 LIME 和 SmoothGrad (Smilkov et al. 2017) 被发现更具抵抗力，这可能是因为它们依赖输入扰动而非梯度，这可能捕捉到了训练点和非训练点之间细微的区别。Liu et al. (2024) 进一步在图像数据背景下检查了多种后验解释器的成员推理风险。

防御和对策。 为了防御推理攻击，文献中一些工作采用了改变训练过程的方法，例如 DP-SGD (Liu et al. 2024)，或采用扰动目标模型的每个输入的置信度得分的方法，通过添加噪声并将扰动后的置信度得分转换为攻击模型的对抗样本，如 MemGuard (Jia et al. 2019)。然而，这些方法要么被证明对推理攻击无效（如 MemGuard）(Liu et al. 2024)，要么是在严重降低模型效用以及解释质量的代价下降低推理攻击性能，例如 DP-SGD，从而在防御能力与效用以及可解释性表现之间呈现出巨大的权衡关系 (Liu et al. 2024)。

可解释自然语言处理中的隐私 现有的工作主要在表格和基于图像的数据集中研究模型解释的隐私风险。然而，在自然语言处理（NLP）中隐私和解释性之间的关系尚未被广泛探索。与这些研究不同，我们的工作专注于文本数据集和 NLP 应用。我们并不是分析解释的隐私泄漏，而是实证研究数据级差分隐私是否可以应用以实现合理的隐私保证，同时在模型效用和解释质量上保持可接受的权衡。据我们所知，这是首次系统性地评估数据级差分隐私与 NLP 模型后验特征归因解释（在其忠实度方面）质量之间的权衡研究。

3 实验设置

为了指导我们的实验，我们选择了三个数据集，这些数据集在规模和领域上各不相同。以下将介绍这些数据集及其相关的下游任务。

SST-2。 斯坦福情感树库数据集 (SST-2) (Socher et al. 2013) 包含来源于电影评论的短文本，并且由于其在 GLUE 基准中包含而被推广 (Wang et al. 2018)。每个文本都根据其情感进行标注，即分为正面或负面，从而形成一个二分类任务。我们使用数据集完整的训练拆分，共由 67,349 条记录组成，平均词长为 9.41。

AG **新闻** AG News 语料库包含来自超过 2000 个新闻来源的逾一百万篇新闻文章。我们使用 Zhang, Zhao, and LeCun (2015) 准备的子集，该子集包含来自四个新闻领域的 12 万篇新闻文章：世界、体育、商业和科学/技术。我们进行了一次 50 % 的随机采样（seed=42），得到一个 60000 篇新闻文章的最终数据集，平均词长为 43.90。

Trustpilot 语料库是一个大规模的用户评论集合。由 Hovy, Johannsen, and Søgaard (2015) 准备的语料库为每条评论标记了提供的星级（1-5），我们简化为负面评论（1-2 星）和正面评论（5 星）。我们特别只选取数据集中美国部分（en-US）的评论，并使用 10 % 的随机样本。这导致一个包含 29,490 条评论的数据集，平均字数为 59.75。

3.1 DP 方法

我们为实验选择了三种局部差分隐私文本重写方法，这些方法将在下文中介绍。

TEM 。 TEM (Carvalho et al. 2023) 是一种词级别的 DP 机制，利用了一种称为度量 DP 的广义概念。这种广义化对于像词嵌入这样的度量空间非常有用，其中词语之间的不可区分性要求根据它们在空间中的距离进行扩展。TEM 通过采用截断的指数机制改进了以前的方法，允许更高效用的词替换。为了使完整文本隐私化，每个组成词逐一隐私化。遵循原始工作，我们选择 $\varepsilon \in \{1, 2, 3\}$ 的隐私预算，这指的是每个词的预算。

DP-Prompt. 利用 LLMs 的生成能力，Utpala et al. (2023) 介绍了 DP-Prompt，这是一种通过将私有化建模为改写任务来生成具有本地差分隐私保证的私有输出文本的方法。在内部，DP-Prompt 机制通过在每个 token 生成时应用差分隐私来操作，具体来说是使用温度参数作为指数机制 (McSherry and Talwar 2007) 的等效 DP 机制。根据原始工作，我们测试了三个温度值，即 $T \in \{1.75, 1.5, 1.25\}$。这相当于每个 token 的 ε 值为 $\varepsilon \in \{118, 137, 165\}$，使用由 Meisenbacher et al. (2024) 提供的实现，其中使用了 flan-t5-base 模型 (Chung et al. 2022) 作为基本的私有化模型。

DP-BART 是一种文档级 LDP 文本重写机制，由 Igamberdiev and Habernal (2023) 提出。它利用了预训练的 BART 模型 (Lewis et al. 2020)，在潜在表示空间中应用校准的高斯噪声，之后解码器对噪声编码向量进行解码以生成私密文本。通过这种方式，DP-BART 提供单个文档的重写保证，尽管通常需要更高的隐私预算以产生有意义的输出。因此我们选择 $\varepsilon \in \{500, 1000, 1500\}$，并使用原始的 DP-BART-CLV 变体。

对于

私有化程序，我们在我们选择的每个数据集的主要文本列上使用该机制。这一过程对三个隐私预算中的每一个重复进行，为原始数据创造了三个私有副本。使用 DP-Prompt 同样会通过生成过程产生完整文本。最后，TEM 在输入文本的所有组成词上按顺序运行，这些词通过 nltk.word_tokenize 函数进行标记处理。来自该机制的输出列表然后通过简单的连接重构成一个字符串。总而言之，我们因此产生了原始数据集的九个私有变体，总共形成 30 个数据集（3 个原始基线 + 27 个私有副本）。

我们提醒注意，在为每个差分隐私机制选择隐私预算以及基于这些决策得到的数据集时，我们没有确保三个选择的方法之间的任何可比性。由于在这三种方法中确保差分隐私的方式不同，以及比较不同差分隐私概念（例如，MDP 与 DP）所涉及的复杂性，我们不涉及此类比较。相反，我们专注于分析每种方法内部的差分隐私重写对下游效果的影响（即，跨隐私预算）。

我们还注意到，选择隐私预算（ε 值）是受原始作品中作者选择的激励。我们并未致力于规范这些值，例如，通过规范 DP-Prompt 中使用的 logit 值。虽然这将是报告更可用和合理的隐私预算的有益步骤，但我们努力测试最初提出的 DP 方法。因此，我们不报告或分析基础 DP 保证的相对强度，因为这是 DP-NLP 研究的一个活跃领域。

我们利用了一些经过预训练的仅编码器语言模型，它们在架构和模型大小上有所不同。在此过程中，我们从经验上测量了在 DP 重写数据集上进行微调的效果，具体来说就是该过程对这些模型可解释性的影响（根据我们选择的指标进行衡量）。表 1 展示了在实验中用来微

调我们的三个选择的分类数据集的预训练模型的信息和详细信息。

Name	Type	# of Parameters
BERT-base-cased	Encoder-only	∼ 110M
BERT-large-cased	Encoder-only	∼ 335M
RoBERTa-base	Encoder-only	∼ 125M
RoBERTa-large	Encoder-only	∼ 355M
DeBERTa-base	Encoder-only	∼ 139M

Table 1: 实验模型的详细信息。

3.2 可解释性方法

我们在实验中包含了四种事后特征归因方法：Gradient 通过计算输出相对于输入特征的梯度；Integrated Gradient 则是在从基线输入到解释输入的路径上积分梯度；SHAP (Lundberg and Lee 2017) 近似每个特征的 Shapley 值，这是来自合作博弈论的一个概念，通过考虑特征的不同组合以及每个特征对结果的贡献来衡量每个特征的“贡献”；LIME (Ribeiro, Singh, and Guestrin 2016) 试图通过一个易于解释的线性模型在解释输入的局部附近复制模型的行为。

3.3 评估解释

我们使用两种不同的方法来评估事后解释，分别测量解释的全面性和充分性。这两个指标都有效地尝试量化解释的忠实性，即解释与模型 (Jacovi and Goldberg 2020) 的内在决策过程的准确反映。忠实性是解释最重要的期望之一 (Danilevsky et al. 2020; Lyu et al. 2024)，因为高忠实性表明该解释准确地反映了模型对给定预测的决策过程。在最简单的形式中，全面性指标测量当依据特征归因分数去除输入中排名前 k 的标记时对真实类别输出概率的变化，而充分性指标测量当仅将排名前 k 的标记作为输入提供给模型时的变化。然后我们通过平均值计算 AOPC，其中 k 被改变。我们将这些指标称为 AOPC-全面性和 AOPC-充分性。

硬性移除这些标记可能会导致对于模型来说是分布外的输入，这可能在解释的评估过程中不利地影响模型性能 (Zhao et al. 2022; Chrysostomou and Aletras 2022)。为了解决这个潜在问题，我们在评估中也加入了软充分性和软全面性指标 (Zhao and Aletras 2023)。对于这些指标的软版本，我们不是完全移除标记，而是根据该标记的重要性得分对每个标记的嵌入进行部分遮挡，即被遮挡的标记计算为 $\boldsymbol{x}' = \boldsymbol{x} \odot \boldsymbol{e}$，其中 $\boldsymbol{e}_i \sim \text{Bernoulli}(q)$，若标记被保留（充分性），则 q 为重要性得分，若标记被移除（全面性），则为 1 减去得分。

理解隐私和可解释性之间的权衡并不总是同等重要的，我们设计了一个指标，允许在强调实用性或可解释性时进行调整，同时仍然以同样的标准看待两者。我们为每个模型 m 计算出一个复合分数，并将 $\alpha \in \{0.25, 0.5, 0.75\}$ 赋权如下：$\text{CS}(m, \alpha) = \alpha \widehat{\text{F1}}(m) + (1-\alpha)\,\widehat{E}(m)$，其中 $\widehat{E}(m) = \frac{1}{4}\big(\widehat{C}_s(m) + (1-\widehat{S}_s(m)) + \widehat{C}_a(m) + (1-\widehat{S}_a(m))\big)$，$\widehat{x}$ 表示 x 的极大极小标准化，指标 s 和 a 区分“软性”和中的“AOPC”，C 表示全面性，S 表示充分性（通过 $1-\widehat{S}$ 反转以便较大值总是更好）。

$$\text{CS}(m,\alpha) = \alpha\hat{F}1 + (1-\alpha)\hat{E} \quad \hat{E} = \tfrac{1}{4}\big(\hat{C}_s + (1-\hat{S}_s) + \hat{C}_a + (1-\hat{S}_a)\big)$$

复合得分的范围从 0 到 1，其中更高的值（接近 1）表示接近 1 意味着整体性能更好（基于选择的 α 值）。我们故意将分数定义为这种标准化形式，以便在我们的实验中能够进行简单且一致的比较。此外，使用此得分，可以根据效用和可解释性之间的相对权重来改变 α。例如，一个 α 值为 0.25 将意味着可解释性可信度指标显得更加重要，而值为 0.75 则效用得分（F1）优先。

4 结果与分析

我们的结果展示在若干表格和图形中。表格 2、3 和 4 分别报告了 $\alpha = 0.25$、0.5 和 0.75 的组合得分（平均 $_{\text{std}}$）。得分是基于五个 PLM 的平均值（BERT-base、BERT-large、RoBERTa-base、RoBERTa-large 和 DeBERTa-base），并且 Avg 行表示每对（数据集，DP-ε）中的四个解释器（Gradient，IG，LIME 和 SHAP）之间的平均值和合并标准差。在每一行中，最高的组合得分用下划线标记。列按 DP 方法（无，DPB，DPP 和 TEM）分组，并使用颜色编码进行比较：灰色表示无 DP，青色表示 DPB，橙色表示 DPP，绿色表示 TEM。颜色强度反映得分大小，颜色越深表示值越高。例如，在表格 2 中，TEM 列中颜色最深的单元格突出显示了在给定数据集和 α 值（例如，AG News，$\alpha = 0.25$）下得分最高的解释器–DP-ε 组合（例如，TEM3 配合 SHAP）。这种颜色的区分用于反映由于机制和假设的差异，DP 方法不可直接比较这一事实（参见章节 3.1）。这些表格支持对解释器和隐私配置趋势的分析。

为了更好地可视化来自表格 2、3 和 4 的结果，对于三个数据集中的每一个，我们在图 1a、1b 和 1c 中展示了相应的图。每个图对应一个数据集，每个数据点表示一个（DP-ε，解释器，α）三重组的复合得分，平均在五个 PLM 上。在这些图中，不同的形状表示不同的解释器，而颜色表示 α 值（0.25, 0.5, 0.75），其中解释器的缩写如下：（G: 梯度，IG: 综合梯度，L:LIME，S:SHAP）。

(a) SST2

(b) AG 新闻

(c) Trustpilot

Figure 1: 通过 DP-ε、解释器 α，对 5 个模型的数据集进行的综合评分。

在表 ?? 中，我们展示了复合得分值，但我们将每个（数据集，α，DP-ε）三元组的复合得分值在四个解释器中取平均（而不是像以前一样按解释器），从而使我们能够比较和评估每个 DP-ε 组合如何影响所有解释器上的复合得分，使我们能够检测每个数据集中的每个 α 值/场景的“最佳位置”。

在表 ?? 中，我们研究了模型大小对隐私、效用和可解释性之间权衡的影响；具体而言，该权衡在小模型和大模型之间是否不同。为此，我们在四个解释器上平均结果，但不是在所有五个模型上聚合，而是按模型大小对结果进行分组。基础组包括 BERT-base 和 RoBERTa-base，而大组包括它们相应的较大变体。

表 2、3 和 4 便于检查每个 α 值和数据集的每个解释器，为其提供最高复合得分的 DP-ε 方法，以及每个 DP 方法的 ε 值，导致特定 DP 方法的最高得分。此

Dataset	Expl	None	DPB500	DPB1000	DPB1500	DPP118	DPP137	DPP165	TEM1	TEM2	TEM3
AG News	G	$0.705_{0.19}$	$0.286_{0.14}$	$0.655_{0.01}$	$0.705_{0.01}$	$0.758_{0.01}$	$0.771_{0.02}$	$0.769_{0.02}$	$0.333_{0.22}$	$0.598_{0.17}$	$0.777_{0.03}$
	IG	$0.600_{0.13}$	$0.245_{0.11}$	$0.526_{0.02}$	$0.577_{0.01}$	$0.610_{0.01}$	$0.620_{0.02}$	$0.617_{0.02}$	$0.272_{0.14}$	$0.462_{0.11}$	$0.632_{0.02}$
	LIME	$0.780_{0.24}$	$0.338_{0.19}$	$0.718_{0.02}$	$0.760_{0.02}$	$0.814_{0.02}$	$0.837_{0.01}$	$0.823_{0.02}$	$0.391_{0.27}$	$0.702_{0.19}$	$0.882_{0.02}$
	SHAP	$0.765_{0.23}$	$0.333_{0.19}$	$0.718_{0.01}$	$0.751_{0.01}$	$0.818_{0.02}$	$0.832_{0.03}$	$0.823_{0.03}$	$0.378_{0.26}$	$0.679_{0.19}$	$0.853_{0.03}$
	Avg	$0.713_{0.20}$	$0.301_{0.16}$	$0.654_{0.01}$	$0.698_{0.01}$	$0.750_{0.01}$	$0.765_{0.02}$	$0.758_{0.02}$	$0.344_{0.22}$	$0.610_{0.17}$	$0.786_{0.03}$
SST2	G	$0.515_{0.18}$	$0.240_{0.05}$	$0.358_{0.14}$	$0.441_{0.13}$	$0.489_{0.16}$	$0.477_{0.15}$	$0.475_{0.16}$	$0.239_{0.06}$	$0.277_{0.09}$	$0.396_{0.18}$
	IG	$0.410_{0.13}$	$0.203_{0.04}$	$0.287_{0.10}$	$0.358_{0.10}$	$0.390_{0.12}$	$0.375_{0.11}$	$0.382_{0.11}$	$0.217_{0.05}$	$0.246_{0.07}$	$0.325_{0.13}$
	LIME	$0.573_{0.22}$	$0.214_{0.08}$	$0.375_{0.18}$	$0.483_{0.17}$	$0.537_{0.20}$	$0.521_{0.19}$	$0.520_{0.19}$	$0.237_{0.11}$	$0.286_{0.15}$	$0.430_{0.25}$
	SHAP	$0.567_{0.22}$	$0.215_{0.07}$	$0.374_{0.19}$	$0.484_{0.17}$	$0.535_{0.20}$	$0.523_{0.19}$	$0.518_{0.18}$	$0.235_{0.11}$	$0.285_{0.16}$	$0.423_{0.25}$
	Avg	$0.516_{0.19}$	$0.218_{0.06}$	$0.349_{0.15}$	$0.442_{0.14}$	$0.488_{0.17}$	$0.474_{0.16}$	$0.474_{0.16}$	$0.232_{0.08}$	$0.274_{0.12}$	$0.394_{0.20}$
Trustpilot	G	$0.506_{0.17}$	$0.421_{0.11}$	$0.489_{0.17}$	$0.531_{0.12}$	$0.491_{0.14}$	$0.487_{0.08}$	$0.483_{0.06}$	$0.321_{0.14}$	$0.318_{0.10}$	$0.504_{0.16}$
	IG	$0.488_{0.16}$	$0.420_{0.10}$	$0.477_{0.17}$	$0.513_{0.12}$	$0.458_{0.13}$	$0.456_{0.07}$	$0.451_{0.05}$	$0.320_{0.14}$	$0.311_{0.10}$	$0.480_{0.15}$
	LIME	$0.558_{0.20}$	$0.451_{0.12}$	$0.521_{0.18}$	$0.566_{0.13}$	$0.539_{0.17}$	$0.542_{0.10}$	$0.537_{0.06}$	$0.343_{0.19}$	$0.356_{0.15}$	$0.555_{0.19}$
	SHAP	$0.551_{0.19}$	$0.443_{0.12}$	$0.517_{0.18}$	$0.558_{0.13}$	$0.534_{0.17}$	$0.530_{0.09}$	$0.534_{0.07}$	$0.325_{0.19}$	$0.340_{0.14}$	$0.551_{0.19}$
	Avg	$0.526_{0.18}$	$0.434_{0.11}$	$0.501_{0.18}$	$0.542_{0.13}$	$0.506_{0.15}$	$0.504_{0.09}$	$0.501_{0.06}$	$0.327_{0.16}$	$0.331_{0.12}$	$0.523_{0.17}$

Table 2: 复合评分（平均 $_{\text{std}}$），其中 $\alpha = 0.25$ 为五个 PLMs 的平均值。Avg 指的是四个解释器的均值和汇总标准差。

Dataset	Expl	None	DPB500	DPB1000	DPB1500	DPP118	DPP137	DPP165	TEM1	TEM2	TEM3
AG News	G	$0.766_{0.13}$	$0.297_{0.19}$	$0.710_{0.01}$	$0.760_{0.01}$	$0.811_{0.00}$	$0.820_{0.01}$	$0.818_{0.01}$	$0.343_{0.26}$	$0.623_{0.19}$	$0.828_{0.02}$
	IG	$0.696_{0.09}$	$0.270_{0.17}$	$0.624_{0.01}$	$0.674_{0.01}$	$0.712_{0.01}$	$0.719_{0.02}$	$0.717_{0.01}$	$0.301_{0.21}$	$0.533_{0.15}$	$0.731_{0.02}$
	LIME	$0.816_{0.16}$	$0.332_{0.23}$	$0.753_{0.01}$	$0.797_{0.01}$	$0.848_{0.01}$	$0.864_{0.01}$	$0.854_{0.01}$	$0.381_{0.29}$	$0.693_{0.20}$	$0.898_{0.02}$
	SHAP	$0.806_{0.15}$	$0.329_{0.22}$	$0.752_{0.01}$	$0.791_{0.01}$	$0.851_{0.01}$	$0.861_{0.02}$	$0.854_{0.02}$	$0.372_{0.29}$	$0.678_{0.20}$	$0.878_{0.02}$
	Avg	$0.771_{0.13}$	$0.307_{0.20}$	$0.710_{0.01}$	$0.756_{0.01}$	$0.806_{0.01}$	$0.816_{0.01}$	$0.811_{0.01}$	$0.349_{0.26}$	$0.632_{0.19}$	$0.834_{0.02}$
SST2	G	$0.592_{0.21}$	$0.294_{0.09}$	$0.439_{0.19}$	$0.541_{0.17}$	$0.591_{0.20}$	$0.583_{0.20}$	$0.580_{0.20}$	$0.293_{0.10}$	$0.342_{0.13}$	$0.479_{0.23}$
	IG	$0.522_{0.19}$	$0.269_{0.08}$	$0.391_{0.16}$	$0.486_{0.15}$	$0.525_{0.17}$	$0.515_{0.16}$	$0.518_{0.17}$	$0.278_{0.09}$	$0.322_{0.12}$	$0.432_{0.19}$
	LIME	$0.630_{0.24}$	$0.277_{0.10}$	$0.450_{0.21}$	$0.569_{0.20}$	$0.623_{0.23}$	$0.613_{0.22}$	$0.610_{0.22}$	$0.292_{0.13}$	$0.349_{0.17}$	$0.501_{0.27}$
	SHAP	$0.627_{0.24}$	$0.277_{0.10}$	$0.449_{0.22}$	$0.570_{0.20}$	$0.622_{0.23}$	$0.614_{0.22}$	$0.609_{0.22}$	$0.290_{0.13}$	$0.347_{0.18}$	$0.497_{0.28}$
	Avg	$0.593_{0.22}$	$0.279_{0.09}$	$0.432_{0.20}$	$0.542_{0.18}$	$0.590_{0.21}$	$0.581_{0.20}$	$0.579_{0.20}$	$0.288_{0.11}$	$0.340_{0.15}$	$0.477_{0.24}$
Trustpilot	G	$0.579_{0.20}$	$0.501_{0.11}$	$0.600_{0.18}$	$0.671_{0.08}$	$0.604_{0.17}$	$0.628_{0.06}$	$0.634_{0.04}$	$0.407_{0.14}$	$0.419_{0.13}$	$0.626_{0.19}$
	IG	$0.567_{0.19}$	$0.500_{0.11}$	$0.592_{0.18}$	$0.659_{0.08}$	$0.581_{0.16}$	$0.608_{0.05}$	$0.613_{0.04}$	$0.406_{0.14}$	$0.415_{0.13}$	$0.610_{0.18}$
	LIME	$0.614_{0.21}$	$0.521_{0.12}$	$0.621_{0.19}$	$0.695_{0.09}$	$0.636_{0.19}$	$0.665_{0.07}$	$0.670_{0.04}$	$0.422_{0.18}$	$0.445_{0.16}$	$0.660_{0.21}$
	SHAP	$0.610_{0.21}$	$0.515_{0.12}$	$0.619_{0.19}$	$0.689_{0.09}$	$0.632_{0.19}$	$0.657_{0.07}$	$0.668_{0.05}$	$0.409_{0.17}$	$0.433_{0.15}$	$0.658_{0.21}$
	Avg	$0.593_{0.20}$	$0.509_{0.12}$	$0.608_{0.18}$	$0.679_{0.09}$	$0.613_{0.18}$	$0.640_{0.06}$	$0.646_{0.04}$	$0.411_{0.16}$	$0.428_{0.14}$	$0.639_{0.19}$

Table 3: 针对 $\alpha = 0.5$ 的综合得分（平均 $_{\text{std}}$），在五个 PLM 中取平均。Avg 指四个解释器的平均值和合并标准差。

Dataset	Expl	None	DPB500	DPB1000	DPB1500	DPP118	DPP137	DPP165	TEM1	TEM2	TEM3
AG News	G	$0.827_{0.09}$	$0.308_{0.24}$	$0.766_{0.00}$	$0.815_{0.00}$	$0.863_{0.00}$	$0.869_{0.01}$	$0.868_{0.01}$	$0.352_{0.30}$	$0.649_{0.21}$	$0.878_{0.01}$
	IG	$0.792_{0.08}$	$0.295_{0.23}$	$0.723_{0.01}$	$0.772_{0.01}$	$0.814_{0.00}$	$0.818_{0.01}$	$0.817_{0.00}$	$0.331_{0.28}$	$0.604_{0.19}$	$0.830_{0.01}$
	LIME	$0.852_{0.10}$	$0.326_{0.26}$	$0.787_{0.01}$	$0.834_{0.01}$	$0.882_{0.01}$	$0.891_{0.01}$	$0.886_{0.01}$	$0.371_{0.32}$	$0.684_{0.22}$	$0.913_{0.01}$
	SHAP	$0.847_{0.09}$	$0.324_{0.26}$	$0.787_{0.00}$	$0.830_{0.01}$	$0.883_{0.01}$	$0.889_{0.01}$	$0.886_{0.01}$	$0.367_{0.32}$	$0.676_{0.21}$	$0.904_{0.01}$
	Avg	$0.830_{0.09}$	$0.313_{0.25}$	$0.766_{0.01}$	$0.813_{0.01}$	$0.861_{0.01}$	$0.867_{0.01}$	$0.864_{0.01}$	$0.355_{0.30}$	$0.653_{0.21}$	$0.881_{0.01}$
SST2	G	$0.669_{0.25}$	$0.347_{0.12}$	$0.520_{0.23}$	$0.641_{0.21}$	$0.693_{0.24}$	$0.689_{0.24}$	$0.685_{0.24}$	$0.347_{0.14}$	$0.408_{0.17}$	$0.562_{0.28}$
	IG	$0.634_{0.24}$	$0.335_{0.12}$	$0.496_{0.22}$	$0.613_{0.20}$	$0.660_{0.23}$	$0.656_{0.22}$	$0.654_{0.22}$	$0.339_{0.13}$	$0.397_{0.16}$	$0.538_{0.26}$
	LIME	$0.688_{0.26}$	$0.339_{0.13}$	$0.525_{0.25}$	$0.655_{0.22}$	$0.709_{0.25}$	$0.704_{0.25}$	$0.700_{0.25}$	$0.346_{0.15}$	$0.411_{0.19}$	$0.573_{0.30}$
	SHAP	$0.686_{0.26}$	$0.339_{0.13}$	$0.525_{0.25}$	$0.655_{0.22}$	$0.708_{0.25}$	$0.705_{0.25}$	$0.699_{0.25}$	$0.345_{0.15}$	$0.410_{0.19}$	$0.571_{0.30}$
	Avg	$0.669_{0.25}$	$0.340_{0.12}$	$0.517_{0.24}$	$0.641_{0.21}$	$0.693_{0.24}$	$0.689_{0.24}$	$0.685_{0.24}$	$0.344_{0.14}$	$0.407_{0.18}$	$0.561_{0.28}$
Trustpilot	G	$0.653_{0.24}$	$0.581_{0.12}$	$0.711_{0.20}$	$0.812_{0.04}$	$0.716_{0.20}$	$0.770_{0.04}$	$0.786_{0.03}$	$0.492_{0.15}$	$0.520_{0.16}$	$0.748_{0.22}$
	IG	$0.647_{0.23}$	$0.580_{0.12}$	$0.707_{0.19}$	$0.806_{0.04}$	$0.705_{0.19}$	$0.759_{0.04}$	$0.775_{0.03}$	$0.492_{0.15}$	$0.518_{0.16}$	$0.740_{0.21}$
	LIME	$0.670_{0.24}$	$0.591_{0.13}$	$0.722_{0.20}$	$0.824_{0.05}$	$0.732_{0.21}$	$0.788_{0.05}$	$0.804_{0.03}$	$0.500_{0.17}$	$0.533_{0.18}$	$0.765_{0.23}$
	SHAP	$0.668_{0.24}$	$0.588_{0.13}$	$0.721_{0.20}$	$0.821_{0.05}$	$0.731_{0.21}$	$0.784_{0.04}$	$0.803_{0.03}$	$0.494_{0.16}$	$0.527_{0.17}$	$0.764_{0.22}$
	Avg	$0.660_{0.24}$	$0.585_{0.12}$	$0.715_{0.20}$	$0.816_{0.04}$	$0.721_{0.20}$	$0.775_{0.04}$	$0.792_{0.03}$	$0.495_{0.16}$	$0.525_{0.17}$	$0.754_{0.22}$

Table 4: 组合分数（平均值 $_{\text{std}}$）是对五个 PLM 的 $\alpha = 0.75$ 的平均。Avg 指的是四个解释器的平均值和汇总标准差。

外，我们查看图 1 中的图表，并根据我们在 SST2、AG News 和 Trustpilot 中针对不同 DP 机制（ε），四种可解释性方法以及效用–解释权衡（α =25 %，50 %，75 %）所呈现的复合得分图结果，我们可以报告如下：

当 α 值较高时，对模型效用的重视程度更大，这在所有数据集和 DP 配置中始终导致更高的综合评分。在解释器中，LIME 和 SHAP 在综合评分上通常优于 Gradient 和 IG，尤其是在强隐私约束下。正如预期的那样，在更严格的 DP 设置下（例如，DPB500，TEM1），综合评分显著下降，反映出隐私与效用和解释质量之间的权衡。SST2 似乎对 DP 噪声非常敏感，在低隐私预算下表现出显著的性能下降。在这些设置中，G 和 IG 尤其受到影响，而 LIME 和 SHAP 即使在中等 DP 配置（例如，DPB1000 和 DP-Prompt）下也表现出更稳定的性能。在 TEM 方法中，TEM3 显示出性能的部分恢复，尤其是在对效用赋予更大权重时（α 高）。

另一方面，与 SST2 相比，AG News 显示出对 DP 噪声效应的更强韧性。LIME 和 SHAP 在大多数 DP 级别中保持优越性能，随着 ε 增大，IG 显著改善。尤其是在 DP-Prompt 和 TEM3 配置中，IG 变得具有竞争力。TEM3 变体在（$\alpha = 75$ %）时表现特别好，表明在效用与解释质量之间有一个理想的平衡。

Trustpilot 对 DP 噪声显示出中等敏感性，介于 SST2 和 AG News 之间。随着 ε 的增加，综合评分逐渐提高，而 LIME 在大多数设置中仍然是最有效的解释器。DP-BART 和 DP-Prompt 机制，特别是在中高隐私预算的

情况下，提供了在保持解释质量方面可靠的性能。

在所有数据集上，LIME 和 SHAP 稳定地表现优于 IG 和 Gradient，后者更易受到较强 DP 的影响。Gradient 是受影响最为严重的解释器，特别是在低 ε 情形和 SST2 数据集上。在 DP 机制中，DP-BART -1500 和 DP-Prompt -165 在隐私、效用和可解释性之间提供了良好的权衡。虽然 TEM1 和 TEM2 通常导致显著退化，但 TEM3 表现更好，尤其是在 AG News 和 Trustpilot 上。

Privacy Tier	$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$	
	Expl–Data	Data(avg)	Expl–Data	Data(avg)	Expl–Data	Data(avg)
None	LIME–AG News 0.780	AG News 0.713	LIME–AG News 0.816	AG News 0.771	LIME–AG News 0.852	AG News 0.830
Small ε	SHAP–AG News 0.818	Trustpilot 0.434	SHAP–AG News 0.851	Trustpilot 0.509	SHAP–AG News 0.883	Trustpilot 0.585
Medium ε	LIME–AG News 0.837	AG News 0.654	LIME–AG News 0.864	AG News 0.710	LIME–AG News 0.891	AG News 0.766
Large ε	LIME–AG News 0.882	AG News 0.786	LIME–AG News 0.898	AG News 0.756	LIME–AG News 0.913	AG News 0.813

Table 5: 隐私层和 α 值中的“甜点”总结。每一行对应一个隐私层（None = 无差分隐私；Small/Medium/Large ε = 隐私逐渐减少）。对于每个 α ，左侧列显示最佳的解释器-数据集配对（来自表 1–3），右侧列显示基于解释器平均分的最佳数据集（来自表 5）。Epsilon 级别分组为：Small (DPB500, DPP118, TEM1)，Medium (DPB1000, DPP137, TEM2)，和 Large (DPB1500, DPP165, TEM3)。

表

4.1 所有解释器在 DP 方法和数据集上的比较

显示了四个解释器在每个数据集、α 值和 DP 方法 (DP-ε) 组合上的平均综合得分，这些得分是在五个使用的 PLM 上取平均的。我们得出了以下结果和见解。

在所有数据集和 DP 设置中，更高的 α 值（即，更大权重放在实用性上）一直产生更高的综合得分。这在蓝色曲线（$\alpha = 0.75$）尤其明显，它在大多数配置中占据优势。这一趋势反映了实用性在综合得分中稳定的作用，特别是在隐私导致退化的情况下。

在 DP 策略中，DP-BART 和 DP-Prompt （特别是在较高的 ε 值如 1500 和 165 时）达到了最高的平均综合得分，尤其是在 AG News 和 Trustpilot 中。相反，TEM 方法（T1 和 T2）导致了显著的性能下降，最明显的是在 SST2 上。如预期的那样，DPB500（最强的隐私预算）在所有数据集和 α 值上表现最差，验证了严格隐私约束的代价。

数据集敏感性。 SST2 一直产生最低的综合得分，反映出它对隐私保护扰动更为脆弱。这可能是因为其输入长度较短或情感分析需要更微妙的语义线索。相比之下，AG News 是最为稳健的数据集，即使在中等 DP 条件下也能保持高综合得分。Trustpilot 表现中等，并且随着 ε 的增加而表现出稳定的恢复，尤其是在 DP-BART 和 DP-Prompt 的条件下。

总体而言，随着 ε 增加，综合得分有明显的上升趋势。在 DP-BART 和 DP-Prompt 曲线中，这一模式最为明显，性能从 $\varepsilon = 500$ 到 $\varepsilon = 1500/165$ 稳步提高。这种单调行为证实了预期的权衡：随着隐私限制的放松，效用和解释质量得到了共同改善。

4.2 识别隐私-效用-可解释性的最佳平衡点

我们使用表 1、2、3 和 4 中的结果构建了表 7，其中我们识别并呈现了基于隐私层级和 alpha 值设定的综合评分的最佳点，分为两种情况：解释器-数据集对（来自表 1、2、3）和在四个解释器中平均得分最高的数据集（来自表 3）。根据表 7 中的这些结果，我们观察到以下几点：

在所有隐私等级中，除了最严格的隐私设置（ε 小）之外，LIME–AG News 在每种情况下都成为最佳解释器–数据集组合，其中 SHAP–AG News 略胜一筹。这表明 SHAP 的标记级别重要性在强 DP 噪声下更为稳健，但一旦放松隐私，LIME 总体上提供了最真实且最有用的解释。

在对解释器进行平均时，AG News 在没有差分隐私、中等隐私和低隐私预算的情况下始终占据优势。然而，在最严格的差分隐私预算（小 ε ）下，Trustpilot 在所有 α 值上领先，这表明在严格的隐私约束下，Trustpilot 的解释平均降解得较少。

随着 ε 增加（隐私被放宽），最佳解释器-数据集对和数据集平均的综合得分都单调上升。最大的恢复发生在小和中等 ε 之间，特别是在 AG News 上的 SHAP，表明适度的隐私预算能够恢复大部分丢失的解释信号。

效用-可解释性权衡（α ）趋势。 从 0.25（解释为主）到 0.50（平衡）再到 0.75（效用为主）的增加均匀地提高了所有综合评分，因为更大的权重被放置在分类 F1 上。值得注意的是，最佳点的排名保持稳定：一旦 $\varepsilon \geq$ 中值，LIME–AG News 占据优势，而在最小的 ε 下，Trustpilot 仍然是首选数据集。

虽然在 ε 值大时预计会有较高的综合评分，但值得注意的是，即使在中等和小的 ε （即更严格的隐私）情况下，一些解释器仍然表现强劲。这些结果令人鼓舞，表明在更严格的隐私限制下可以实现可解释性。例如，SHAP 在代表不同效用-解释权衡的所有三个 α 值中均有良好表现。这表明，对于某些数据集、解释器和 DP 方法的组合，取决于具体目标（即更低或更高的 α ），可以在保持高隐私的同时仍维持良好的效用和解释质量。

4.3 模型规模之间的比较

在表 ?? 中，我们研究了模型规模对隐私、效用和可解释性之间权衡的影响；特别是这种权衡在较小和较大模型之间是否有所不同。为了实现这一目标，我们在四个解释器上取平均结果，但与之前的表中在所有五个模型上聚合不同，我们按照模型规模对结果进行分组。基础组包括 BERT-base 和 RoBERTa-base ，而大组包括它们相应的较大变体。在此分析中，我们排除 DeBERTa，因为我们没有在实验中使用 DeBERTa-large。我们认为将 BERT 和 RoBERTa 的基础和大型变体进行比较，足以进行模型规模影响的初步评估。表 ?? 展示了在各数据集和 α 的 DP- ϵ 值下，基础和大模型（解释器的平均）合成得分（平均 $_{std}$ ）的比较。我们了解到以下几点：

在所有数据集和 α 值中，基础模型在综合得分方面始终优于大型模型。这从均匀负的 Δ （大型–基础）值中可以明显看出，范围从 -0.119 到 -0.286 。结果表明，在差分隐私（DP）下增加模型大小会导致预测和解释的质量下降。

数据集敏感性 SST2 数据集显示出模型规模对性能最大的负面影响。例如，在 $\alpha = 0.75$ 时，差距达到 $\Delta = -0.286$ 。相比之下，AG News 和 Trustpilot 的性能下降较为温和（约在 -0.12 到 -0.16 之间）。这表明 SST2 可能对隐私引起的噪声更加敏感，这可能是由于输入较短或基于情感分类的性质。

总体而言，随着 α 增加（即对效用而非解释质量的重视程度增强），大模型和基础模型之间的负性能差距通常会扩大。此趋势在 SST2 上尤为明显，差距从 -0.188（$\alpha = 0.25$ ）增加到 -0.286 （$\alpha = 0.75$ ）。这表明在使用 DP 训练时，效用在大模型下变得更差。

稳定性和变异性。 标准差通常在较大模型中更高，尤其是在 $\alpha = 0.75$ 条件下。这表明大模型在 DP 下的性能具有更大的不稳定性和不一致性，此外，它们的平均分数较低。

我们反思实验的主要发现，提出一系列实用建议，并指出未来需要考虑的要点。

对实验结果的初步审查表明，数据集的性质可能会对隐私性和可解释性方面的结果产生不同的影响。AG News 在多种 DP 方法下的综合评分优于 SST2 和 Trustpilot 的一个可能解释是，它的主题分类性质导致输入中包含多个上下文对齐的关键词，这些关键词通常对 DP 引出的扰动具有鲁棒性，而 SST2 的短情感句子和 Trustpilot 的非正式用户评论更容易失去关键提示。此外，PLM 在 AG News 上实现了更高的基线 F1，这可能进一步缓冲在隐私约束下的性能下降。最后，AG News 中集中于少数显著词的归因权重可能比对于情感或用户生成文本所需的更分散的归因提供更真实的解释。虽然这些因素需要进一步研究，但它们提供了一个合理的解释，说明 AG News 在严格的隐私预算下能够在综合效用 - 解释性指标中保持韧性。

尽管与我们其他两个数据集相比，AG News 的结果表现出绝对的差异，但仍出现了一个有希望的趋势。当查看图 1 中任何数据集的综合得分曲线时，可以观察到无论数据集如何，对于任何 DP 方法或解释器，都遵循类似的线条。虽然这些曲线在不同数据集之间并不完全一致，但类似的趋势表明了未来研究的一个关键点，即确定在隐私、效用和可解释性的综合研究中数据集的重要性。然而，也必须记住我们所使用的各种解释器的不同性能，这似乎也受到数据集性质的影响，并相应地影响下游任务。

总结所有实验结果后，我们达成了一个重要的讨论点，回到了我们在这项工作中提出的初始研究问题：DP 如何影响可解释性？更深入地探讨，我们还反思 DP 与可解释性是否可以共存，或者这两个重要命题之间是否可能仍存在摩擦。

一个有益的出发点在于比较基线结果（即，在没有应用任何 DP 情况下取得的结果）和那些私有化后的结果。虽然在大多数情况下，应用 DP 导致可解释性下降，但这并不总是如此。确实，在某些场景下，所有三种使用的 DP 方法中，有些结果超过了非私有基线，并且在每种解释器的所有 (DP 方法, ε) 组合中至少发生一次。此外，这一结果在可解释性最受欢迎的设置（$\alpha = 0.75$ ）中表现得最为明显，这表明当效用损失不那么重要时，DP 的影响更加积极。在此，我们发现，在某些场景中，DP 实际上有助于提高可解释性，一个非常有前景的前景。

除了这些结果之外，我们还希望从 DP 方法对下游可解释性影响的角度回答我们的研究问题。正如之前所提到的，这需要谨慎处理，因为我们不能对在不同词汇层次上操作且具有不同隐私保证的 DP 机制进行结论对比。然而，某些机制中确实出现了重要趋势，例如 DP-Prompt 结果的相对稳定性，相较于 TEM 或 DP-BART 的急剧波动。我们还推测，DP 方法的选择与分类任务的性质紧密相关，在我们的两个二元分类任务中，生成方法（DP-Prompt 和 DP-BART ）通常会带来更有利的结果，而在四分类的 AG 新闻任务中，TEM 始终表现优异。这些结果突显了 DP 方法选择的重要性，其可以在很大程度上依赖于手头的下游任务。特别是在表 ?? 中，这一点尤为明显，其中所有 DP 方法至少在一种配置中实现了最佳综合得分。

基于我们的研究结果，我们为数据隐私和可解释性交集处的实践者汇编了一系列建议。基于我们的“最佳平衡点”分析，对于解释质量（低 α ），实践者可以在最严格的隐私设置下使用 SHAP，其他情况下在中等或低隐私下默认使用 LIME。对于效用（高 α ），除了最严格的隐私设置外，LIME 在所有其他情况下都是最佳平衡点。根据图 1 的分析，我们比较了四个解释器的分数，对于需要强隐私和解释质量的应用，建议结合使用 LIME 或 SHAP 与 DPBart-1500 或者 DPPrompt-165。在 SST2 的情况下，应避免使用诸如 TEM1 和 TEM2 等激进的隐私策略，而应优先选择使用 LIME 的中度 DP 设置。对于 AG News，LIME 与 DPPrompt-165 或 TEM3 的结合在所有 α 值中表现出强劲的稳健性。

Trustpilot 展示了广泛的稳健性，是在隐私敏感的 NLP 场景中进行实际部署的合适候选者。在四个解释器中求平均时，我们的结果显示，使用 DPBart-1500 或 DPPrompt-165 是达到隐私、效用和解释质量之间最佳平衡的推荐选择。当优先解释的可信度时，应避免使用 TEM1 和 TEM2，特别是对于像 SST2 这样的敏感数据集。对于重视效用的应用（高 α ），AG News 在中度 DP 设置下表现良好。一般来说，当保持可解释性至关重要时，较高的 ε 值是优选的。根据模型大小效应（基础模型与大模型）的分析，针对需要高效用和可信解释的隐私保护 NLP 任务，基础模型比大型模型更可靠和有效。这对于优先效用（高 α ）或对于像 SST2 这样的敏感数据集尤其如此。

基于我们的实验，这些建议可以被概括为以下一组指南，这些指南在自然语言数据的可解释隐私研究或实践中是重要的：

1. 决定隐私与实用性的重要性：一个重要的出发点是设置 α ，因为在不同的使用场景中，这可能会有所不同。
2. 考虑下游任务的性质：我们的结果表明，数据集（和任务）是在隐私性和可解释性并列中的重要因素。这在下面的观点中尤其重要。
3. 明智地选择你的隐私化方法：我们的初步发现表明，对于更复杂的任务（例如，多类分类），非生成性的方法如 TEM 可能更为适合。然而，对于更像“colloquial”的数据集，比如那些源自用户评论的数据集，基于生成模型的 DP 方法可能更适合于同时保持实用性和可解释性。虽然这一指导原则需要进一步验证，但我们强调了隐私化方法的选择与下游任务性质之间的重要相互作用。
4. 选择给定用例中可接受的最小预训练模型：我们了解到，通过我们的结果，综合评分随着模型规模的增

加而下降，这表明如果给定用例可以接受较小的模型，这些模型在隐私、实用性和可解释性之间找到平衡时可能更可取。

5. 对各种解释器进行衡量：我们发现，尽管在所有测试配置中，不同解释器之间存在个体差异，但使用平均值和综合得分会导致明显的趋同趋势，并且在隐私级别（ε 值）和数据集/任务之间出现可解释的差异。因此，我们建议使用多个解释器，并通过一个综合得分将它们联系在一起，以获得在不同设置之间性能差异的稳健概览。

5 结论

我们在隐私和可解释性交汇处进行了一项研究，以差分隐私文本重写和事后可解释性的总体方法为指导。在一系列实验中，我们量化了隐私和可解释性之间的权衡，这为这两个重要主题之间潜在的协同作用提供了有趣的见解。我们是首次在自然语言数据背景下进行此类研究的，奠定了进一步探索这一跨学科主题的基础。

我们设想基于我们的研究有多个未来工作的方向，即：

- 进一步研究 “sweet spots”，特别是通过整合更强大的隐私代理（即，超出 ε 值的范围）。这可能包括，例如，进行由 Shokri et al. (2021) 执行的成员推理测试。
- 通过对额外数据集、可解释性和 DP 方法以及 ε 范围的实验扩展我们的研究结果。
- 研究与非事后解释方法相关的权衡，并考虑更近期用于衡量解释忠诚度的框架，例如由 Zheng et al. (2025) 提出的框架。
- 将人工评估（即知觉）纳入综合分数的计算中，即通过改进该分数超越自动化指标。

限制条件。 我们的研究有几个局限性。它仅依赖于定量评估，没有质性或人工评估。我们狭义地关注事后解释性和差分隐私文本重写方法，但这并不能充分涵盖这两个领域。我们选择的差分隐私方法忽视了机制设计中细微的差异及其不同的理论保证。这些限制凸显了需要更广泛的研究，以加深对 NLP 中解释性与隐私交集的理解。

6

致谢 我们感谢匿名审稿人提出的建设性反馈意见。本研究得到了德国联邦教育和研究部（BMBF）01IS23069 软件校园 3.0（慕尼黑工业大学）资助。

References

Arras, L.; Osman, A.; and Samek, W. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion, 81: 14–40.

Attanasio, G.; Pastor, E.; Di Bonaventura, C.; and Nozza, D. 2023. ferret: a Framework for Benchmarking Explainers on Transformers. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics.

Carvalho, R. S.; Vasiloudis, T.; Feyisetan, O.; and Wang, K. 2023. TEM: High utility metric differential privacy on text. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), 883–890. SIAM.

Chang, K.-W.; Prabhakaran, V.; and Ordonez, V. 2019. Bias and Fairness in Natural Language Processing. In Baldwin, T.; and Carpuat, M., eds., Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts. Hong Kong, China: Association for Computational Linguistics.

Chrysostomou, G.; and Aletras, N. 2022. An Empirical Study on Explanations in Out-of-Domain Settings. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6920–6938. Dublin, Ireland: Association for Computational Linguistics.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models.

Council of European Union. 2024. Council regulation (EU) no 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; and Sen, P. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In Wong, K.-F.; Knight, K.; and Wu, H., eds., Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 447–459. Suzhou, China: Association for Computational Linguistics.

Dhaini, M.; et al. 2025. Gender Bias in Explainability: Investigating Performance Disparity in Post-hoc Methods. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, 3006–3029. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

Duddu, V.; and Boutet, A. 2022. Inferring Sensitive Attributes from Model Explanations. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, 416–425. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392365.

Dwork, C. 2006. Differential privacy. In International colloquium on automata, languages, and programming, 1–12. Springer.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and

Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. Computational Linguistics, 1–79.

Gupta, M.; Akiri, C.; Aryal, K.; Parker, E.; and Praharaj, L. 2023. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. IEEE Access, 11: 80218–80245.

Hovy, D.; Johannsen, A.; and Søgaard, A. 2015. User Review Sites as a Resource for Large-Scale Sociolinguistic Studies. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, 452–461. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450334693.

Igamberdiev, T.; and Habernal, I. 2023. DP-BART for Privatized Text Rewriting under Local Differential Privacy. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Findings of the Association for Computational Linguistics: ACL 2023, 13914–13934. Toronto, Canada: Association for Computational Linguistics.

Igamberdiev, T.; et al. 2022. DP-Rewrite: Towards Reproducibility and Transparency in Differentially Private Text Rewriting. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., Proceedings of the 29th International Conference on Computational Linguistics, 2927–2933. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Jacovi, A. 2023. Trends in Explainable AI (XAI) Literature. arXiv:2301.05433.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4198–4205. Online: Association for Computational Linguistics.

Jia, J.; Salem, A.; Backes, M.; Zhang, Y.; and Gong, N. Z. 2019. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19, 259–274. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367479.

Johri, S.; Jeong, J.; Tran, B. A.; Schlessinger, D. I.; Wongvibulsin, S.; Barnes, L. A.; Zhou, H.-Y.; Cai, Z. R.; Van Allen, E. M.; Kim, D.; et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. Nature Medicine, 1–10.

Kasiviswanathan, S. P.; Lee, H. K.; Nissim, K.; Raskhodnikova, S.; and Smith, A. 2008. What Can We Learn Privately? In 2008 49th Annual IEEE Symposium on Foundations of Computer Science, 531–540.

Klymenko, O.; et al. 2022. Differential Privacy in Natural Language Processing: The Story So Far. In Feyisetan, O.; Ghanavati, S.; Thaine, P.; Habernal, I.; and Mireshghallah, F., eds., Proceedings of the Fourth Workshop on Privacy in Natural Language Processing, 1–11. Seattle, United States: Association for Computational Linguistics.

Kotonya, N.; and Toni, F. 2020. Explainable Automated Fact-Checking: A Survey. In Scott, D.; Bel, N.; and Zong, C., eds., Proceedings of the 28th International Conference on Computational Linguistics, 5430–5443. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Lee, J.; et al. 2025. Large Language Models in Finance (FinLLMs). Neural Computing and Applications.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871–7880. Online: Association for Computational Linguistics.

Li, X.; Du, M.; Chen, J.; Chai, Y.; Lakkaraju, H.; and Xiong, H. 2023. M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., Advances in Neural Information Processing Systems, volume 36, 1630–1643. Curran Associates, Inc.

Liu, H.; Wu, Y.; Yu, Z.; and Zhang, N. 2024. Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack. In 2024 IEEE Symposium on Security and Privacy (SP), 4791–4809.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Luo, X.; et al. 2022. Feature Inference Attack on Shapley Values. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS '22, 2233–2247. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394505.

Lyu, Q.; et al. 2024. Towards Faithful Model Explanation in NLP: A Survey. Computational Linguistics, 50(2): 657–723.

Madsen, A.; et al. 2022. Post-hoc Interpretability for Neural NLP: A Survey. ACM Comput. Surv., 55(8).

Mattern, J.; et al. 2022. The Limits of Word Level Differential Privacy. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., Findings of the Association for Computational Linguistics: NAACL 2022, 867–881. Seattle, United States: Association for Computational Linguistics.

McSherry, F.; and Talwar, K. 2007. Mechanism Design via Differential Privacy. In 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), 94–103.

Meisenbacher, S.; Chevli, M.; Vladika, J.; and Matthes, F. 2024. DP-MLM: Differentially Private Text Rewriting Using Masked Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., Findings of the Association for Computational Linguistics: ACL 2024, 9314–9328. Bangkok, Thailand: Association for Computational Linguistics.

Meisenbacher, S.; Klymenko, A.; Karpp, A.; and Matthes, F. 2025. Investigating User Perspectives on Differentially Private Text Privatization. In Habernal, I.; Ghanavati, S.; Jain, V.; Igamberdiev, T.; and Wilson, S., eds., Proceedings of the Sixth Workshop on Privacy in Natural Language Processing, 86–105. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-246-6.

Meisenbacher, S.; and Matthes, F. 2024. Thinking Outside of the Differential Privacy Box: A Case Study in Text Privatization with Language Model Prompting. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 5656–5665. Miami, Florida, USA: Association for Computational Linguistics.

Mosca, E.; Szigeti, F.; Tragianni, S.; Gallagher, D.; and Groh, G. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., Proceedings of the 29th International Conference on Computational Linguistics, 4593–4603. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality, 15(2).

Ng, D. T. K.; Leung, J. K. L.; Chu, S. K. W.; and Qiao, M. S. 2021. Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2: 100041.

Pan, X.; Zhang, M.; Ji, S.; and Yang, M. 2020. Privacy Risks of General-Purpose Language Models. In 2020 IEEE Symposium on Security and Privacy (SP), 1314–1331.

Pawelczyk, M.; et al. 2023. On the Privacy Risks of Algorithmic Recourse. In Ruiz, F.; Dy, J.; and van de Meent, J.-W., eds., Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, volume 206 of Proceedings of Machine Learning Research, 9680–9696. PMLR.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135–1144.

Sarti, G.; Feldhus, N.; Sickert, L.; and van der Wal, O. 2023. Inseq: An Interpretability Toolkit for Sequence Generation Models. In Bollegala, D.; Huang, R.; and Ritter, A., eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 421–435. Toronto, Canada: Association for Computational Linguistics.

Shokri, R.; et al. 2021. On the Privacy Risks of Model Explanations. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, 231–241. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Simonyan, K.; et al. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. B.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. CoRR, abs/1706.03825.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Sousa, S.; and Kern, R. 2023. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. Artificial Intelligence Review, 56(2): 1427–1492.

Sundararajan, M.; et al. 2017. Axiomatic attribution for deep networks. In International conference on machine learning, 3319–3328. PMLR.

T.y.s.s., S.; Baumgartner, N.; Stürmer, M.; Grabmair, M.; and Niklaus, J. 2024. Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 16500–16513. Torino, Italia: ELRA and ICCL.

Utpala, S.; et al. 2023. Locally Differentially Private Document Generation Using Zero Shot Prompting. In Bouamor, H.; Pino, J.; and Bali, K., eds., Findings of the Association for Computational Linguistics: EMNLP 2023, 8442–8457. Singapore: Association for Computational Linguistics.

Valvoda, J.; and Cotterell, R. 2024. Towards Explainability in Legal Outcome Prediction Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),

7269–7289. Mexico City, Mexico: Association for Computational Linguistics.

Van Wynsberghe, A. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1(3): 213–218.

Wallace, E.; Gardner, M.; and Singh, S. 2020. Interpreting Predictions of NLP Models. In Villavicencio, A.; and Van Durme, B., eds., Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, 20–23. Online: Association for Computational Linguistics.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Linzen, T.; Chrupała, G.; and Alishahi, A., eds., Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Wang, Z.; Li, H.; Huang, D.; Kim, H.-S.; Shin, C.-W.; and Rahmani, A. M. 2025. HealthQ: Unveiling questioning capabilities of LLM chains in healthcare conversations. Smart Health, 36: 100570.

Weggenmann, B.; Rublack, V.; Andrejczuk, M.; Mattern, J.; and Kerschbaum, F. 2022. DP-VAE: Human-Readable Text Anonymization for Online Reviews with Differentially Private Variational Autoencoders. In Proceedings of the ACM Web Conference 2022, WWW '22, 721–731. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Wen, Q.; Liang, J.; Sierra, C.; Luckin, R.; Tong, R.; Liu, Z.; Cui, P.; and Tang, J. 2024. AI for Education (AI4EDU): Advancing Personalized Education with LLM and Adaptive Learning. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, 6743–6744. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704901.

Wu, H.; Li, S.; Gao, Y.; Weng, J.; and Ding, G. 2024. Natural language processing in educational research: The evolution of research topics. Education and Information Technologies, 29(17): 23271–23297.

Wu, X.; Duan, R.; and Ni, J. 2023. Unveiling security, privacy, and ethical concerns of chatgpt. Journal of Information and Intelligence.

Yan, B.; Li, K.; Xu, M.; Dong, Y.; Zhang, Y.; Ren, Z.; and Cheng, X. 2025. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. High-Confidence Computing, 5(2): 100300.

Yin, Y.; and Habernal, I. 2022. Privacy-Preserving Models for Legal Natural Language Processing. In Aletras, N.; Chalkidis, I.; Barrett, L.; Goanță, C.; and Preoțiuc-Pietro, D., eds., Proceedings of the Natural Legal Language Processing Workshop 2022, 172–183. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In Advances in Neural Information Processing Systems, volume 28, 649–657. Curran Associates, Inc.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for Large Language Models: A Survey. ACM Trans. Intell. Syst. Technol., 15(2).

Zhao, X.; Zhang, W.; Xiao, X.; and Lim, B. 2021. Exploiting Explanations for Model Inversion Attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 682–692.

Zhao, Z.; and Aletras, N. 2023. Incorporating Attribution Importance for Improving Faithfulness Metrics. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 4732–4745. Toronto, Canada: Association for Computational Linguistics.

Zhao, Z.; Chrysostomou, G.; Bontcheva, K.; and Aletras, N. 2022. On the Impact of Temporal Concept Drift on Model Explanations. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., Findings of the Association for Computational Linguistics: EMNLP 2022, 4039–4054. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zheng, X.; Shirani, F.; Chen, Z.; Lin, C.; Cheng, W.; Guo, W.; and Luo, D. 2025. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. In Proceedings of The Thirteenth International Conference on Learning Representations (ICLR).

Zini, J. E.; and Awad, M. 2022. On the Explainability of Natural Language Processing Deep Models. ACM Comput. Surv., 55(5).