When Explainability Meets Privacy: An Investigation at the Intersection of Post-hoc Explainability and Differential Privacy in the Context of Natural Language Processing

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

In the study of trustworthy Natural Language Processing (NLP), a number of important research fields have emerged, including that of *explainability* and *privacy*. While research interest in both explainable and privacy-preserving NLP has increased considerably in recent years, there remains a lack of investigation at the intersection of the two. This leaves a considerable gap in understanding of whether achieving *both* explainability and privacy is possible, or whether the two are at odds with each other. In this work, we conduct an empirical investigation into the privacy-explainability trade-off in the context of NLP, guided by the popular overarching methods of *Differential Privacy* (DP) and Post-hoc Explainability. Our findings include a view into the intricate relationship between privacy and explainability, which is formed by a number of factors, including the nature of the downstream task and choice of the text privatization and explainability method. In this, we highlight the potential for privacy and explainability to co-exist, and we summarize our findings in a collection of practical recommendations for future work at this important intersection.

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

1 Introduction

Recent advances in Natural Language Processing (NLP) have seen bountiful and widespread improvements in the way natural language can be understood and generated. Such progress, hallmarked by the rapid developments enabled by Large Language Models (LLMs) and associated techniques, has powered novel applications in a variety of domains including education (Wen et al. 2024), healthcare (Wang et al. 2025), and finance (Lee et al. 2025), as well as empowered non-technical users to explore the capabilities of Artificial Intelligence (Ng et al. 2021). These benefits, however, do not come for free, and various subfields of NLP currently work at the intersection of NLP and a number of human-centered topics, such as explainability (Danilevsky et al. 2020), privacy (Sousa and Kern 2023), bias (Navigli, Conia, and Ross 2023), fairness (Chang, Prabhakaran, and Ordonez 2019), and sustainability (Van Wynsberghe 2021), among others.

Despite the recent democratization of LLM *use*, a persistent challenge in the deployment of any language models relates to that of *explainability*, which generally refers to the ability to interpret and communicate the decisions provided by a model. Explainability becomes paramount to the safe deployment of models, particularly in ensuring that model inputs can (reasonably) be traced or explained. Beyond this, explainability is not only a desired characteristic of a language model, but also a mandate (mainly for high-risk AI systems) under the recent EU AI Act regulation (Council of European Union 2024). One particularly useful candidate for fulfilling this mandate comes in the form of *post-hoc explainability* (Madsen et al. 2022; Danilevsky et al. 2020), which comprises methods that serve to provide insights into traditionally "black-box" models.

In a similar vein, the stark increase in LLM usage, particularly where users must interact with models hosted on external servers, i.e., in the cloud, has contributed to rising concerns of privacy (Pan et al. 2020; Wu, Duan, and Ni 2023; Gupta et al. 2023; Yan et al. 2025). Calls for privacy protection have been driven by increasingly strict data protection regulations (such as the GDPR); at the same time, privacy has been addressed in a plethora of research on privacy-preserving NLP (Yin and Habernal 2022), ranging from text privatization to private model training (Sousa and Kern 2023). Privacy-preserving techniques aim to mask both direct and indirect identifiers hidden within textual data, while also preserving utility for downstream tasks and applications (Mattern et al. 2022; Weggenmann et al. 2022). One popular framework is Differential Privacy (DP) (Dwork 2006), which lends plausible deniability to text inputs by ensuring some level of indistinguishability between any two texts, usually achieved via the injection of random noise into text representations (Klymenko et al. 2022). This "noisification" can take place on many levels, such on the word level, or alternatively, via document-level rewriting.

Many recent works in Differentially Private Natural Language Processing (DP-NLP) focus on balancing the *privacy-utility* trade-off as a key indicator for the effectiveness of a privatization method (Mattern et al. 2022; Utpala et al. 2023). Other works explored the trade-offs in other important aspects, such as text coherence (Weggenmann et al. 2022) or user acceptability (Meisenbacher et al. 2025). On

the other hand, Explainable NLP (XNLP) has increasingly focused on the intersection of explainability with some aspects of trustworthy NLP, particularly fairness. This research has taken two primary directions: utilizing explainability as a tool for detecting bias in models (T.y.s.s. et al. 2024; Gallegos et al. 2024) and evaluating the fairness of explainability methods themselves (Dhaini et al. 2025). However, specifically in the NLP domain, no works to the best of our knowledge consider the *intersection of privacy and explainability* in terms of *privacy-explainability trade-off*, namely, how the application of privatization methods affects the function of explainability methods. We argue that this consideration is a crucial one, particularly with the simultaneous legal mandate for both explainability and privacy.

In this work, we are the first to investigate the interplay between privacy and explainability in the context of natural language. As a case study, we select two popular subfields: *post-hoc feature attribution* methods and *differentially private text rewriting*. In this, we explore the shift in explainability that can be observed when rewriting texts via DP, at various privacy levels and with fundamentally different rewriting mechanisms. We also consider the effect of downstream task specifics, such as model choice, model size, and fine-tuning task, particularly when varying the importance of privacy over explainability, and vice versa. To guide these experiments, we define one overarching research question:

> *What is the impact of differentially private text rewriting on the post-hoc explainability of fine-tuned language models, and how can one quantify the privacy-explainability trade-off?*

We learn that while a clear trade-off can typically be observed between privacy and explainability, there do exist configurations in which the two work synergistically. In this, we find that the factors of the downstream dataset and task, as well as the selected DP method and its privacy budget, are important in the quantification of the privacy-explainability trade-off. This leads us to create a collection of recommendations for both researchers and practitioners who wish to continue work at this important intersection.

Concretely, our work makes the following contributions:

1. We are the first to conduct an empirical investigation at the intersection of privacy and explainability, particularly in the context of natural language. We make our experimental code available for reproducibility.
2. We provide insights into the complex interplay between post-hoc explainability and differentially private text rewriting, serving as a foundation for broader investigations at this intersection.
3. We analyze our results to propose recommendations on best practices for important design choices, particularly when faced with the need for explainability *and* privacy.

2 Background and Related Work

2.1 Post-hoc Explainability Methods

Model-agnostic *feature-attribution post-hoc* techniques have gained prominence due to their broad applicability (Jacovi 2023). These methodologies seek to determine the relative contribution of individual tokens to model predictions for specific inputs, employing either gradient-based approaches that leverage model derivatives with respect to inputs (Sundararajan et al. 2017; Simonyan, Vedaldi, and Zisserman 2013), or perturbation-based techniques (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017).

The expanding significance of explainable NLP research is demonstrated through the expansion of comprehensive surveys addressing NLP explainability (Wallace, Gardner, and Singh 2020; Zhao et al. 2024; Madsen et al. 2022; Zini and Awad 2022). Moreover, given the deployment of NLP systems in critical applications including education (Wu et al. 2024), healthcare (Johri et al. 2025) and legal domains (Valvoda and Cotterell 2024) where interpretability requirements are essential, specialized surveys have emerged focusing on explainability within particular NLP applications, such as fact verification (Kotonya and Toni 2020), and specific methodological approaches in NLP explainability (Mosca et al. 2022). These comprehensive reviews underscore the extensive utilization of post-hoc methodologies across NLP applications.

Additionally, feature-attribution post-hoc explanation techniques serve as primary methods within explainability platforms and computational frameworks documented in recent literature (Arras, Osman, and Samek 2022; Li et al. 2023; Attanasio et al. 2023; Sarti et al. 2023). These comprehensive frameworks characteristically integrate multiple post-hoc explanation algorithms while accommodating various data modalities Machine Learning (ML) architectures, including pre-trained Language Models (PLMs).

2.2 Differential Privacy in NLP

The notion of *Differential Privacy* was first proposed in the context of relational databases (Dwork 2006), where the primary goal was to protect the participation of an *individual* in the dataset. More specifically, privacy preservation occurs in the sense that information about such an individual cannot be accurately inferred within some bound. This is formalized via the following inequality, for any databases D_1 and D_2 differing in exactly one element, any $\varepsilon > 0$, any computation or function $\mathcal{M}$, and all $\mathcal{S} \subseteq Range(\mathcal{M})$: $\frac{Pr[\mathcal{M}(D_1) \in \mathcal{S}]}{Pr[\mathcal{M}(D_2) \in \mathcal{S}]} \leq e^{\varepsilon}$. Intuitively, DP ensures that there exists some level of *indistinguishability* between any two neighboring databases (differing in one element), thus protecting the individual. This *privacy level* is governed by the ε parameter, also known as the *privacy budget*.

This form of DP is known as ε-DP, and the notion above refers to *global* DP. Another notion, which we focus on in this work, is that of *local* DP (LDP) (Kasiviswanathan et al. 2008). In the local setting, we assume that the central curator, i.e., the one who is to possess the complete dataset, is not trusted. As a solution, DP is ensured at the user level; however, since the entirety of the dataset is not yet known, LDP imposes a much stricter indistinguishability requirement, i.e., between *any* potential neighbor. This differs from the global notion, since neighboring databases only refer to those resulting from the dataset D. Formally, for finite space $\mathcal{P}$ and $\mathcal{V}$, and for all $x, x' \in \mathcal{P}$ and all $y \in \mathcal{V}$:

$\frac{Pr[\mathcal{M}(x)=y]}{Pr[\mathcal{M}(x')=y]} \leq e^{\varepsilon}$ Hence, an observed output cannot be attributed to a specific input with a high probability. While this notion is clearly stricter, it allows for a quantification of a privacy guarantee on the local, single datapoint level without the need for an aggregated dataset.

The translation of DP into the realm of NLP initially brought about numerous research challenges (Klymenko et al. 2022), chief among them the reasoning of who the *individual* is when considering textual data, and how to quantify neighboring "datasets". Despite these challenges, a great deal of recent works have proposed innovative methods for the integration of DP into the NLP pipeline (Hu et al. 2024), ranging for text anonymization and obfuscation to DP training of language models. Particularly considering *text privatization*, many recent methods interpret the task from the *rewriting* perspective, where a sensitive input text is rewritten under DP guarantees to produce a private output text. Such methods operate at various lexical levels, including the word level (Feyisetan et al. 2020; Carvalho et al. 2023), the token-level during language model generation (Utpala et al. 2023; Meisenbacher et al. 2024), and on full texts (Igamberdiev and Habernal 2023). Considering the differences in how these mechanisms operate, it becomes important to consider the nature of the DP guarantee when performing comparative analyses (Vu et al. 2024).

Recent works in DP-NLP highlight persistent challenges, such as the generation of coherent and correct outputs (Mattern et al. 2022), ensuring comparability (Igamberdiev et al. 2022), and quantifying the benefit DP brings over non-DP methods (Meisenbacher and Matthes 2024). However, to the best of our knowledge, no existing works question the effect of DP methods on the *explainability* of texts, or more specifically, of models trained on privatized texts. We see this to be a considerable gap, particularly for DP text privatization in important domains, such as the medical domain, where explainability is also important in conjunction with privacy.

2.3 Privacy Meets Explainability

Privacy in Explainable ML. The privacy implications of explainable ML have received attention in recent years but remain underexplored. Most of the work on the intersection of explainability and privacy in ML focuses on the inherent privacy risks in explanations. These works include studies that investigate how model explanations reveal sensitive information from the training data and how this can be exploited using membership inference attacks. For instance, Shokri et al. (2021) demonstrated that variance in backpropagation-based explanations can reveal whether a data point was used during training, exposing membership information. Duddu and Boutet (2022) extended these concerns by inferring sensitive attributes like gender and race directly from model explanations, while Liu et al. (2024) shows how explanations can amplify membership inference risks by exploiting differences in model robustness under attribution-guided perturbations. Similarly, Luo et al. (2022) showed that private input features can be reconstructed using Shapley-value explanations. Other works include highlighting how explanations can enhance model inversion attacks when used as auxiliary inputs (Zhao et al. 2021), and introducing a membership inference attack for counterfactuals relying on distances between original inputs and their counterfactual counterparts (Pawelczyk et al. 2023).

Shokri et al. (2021) investigated the privacy risks of gradient-based explanation methods—Gradient (Simonyan et al. 2013) and Integrated Gradients (Sundararajan et al. 2017)—across four tabular and two image datasets. They showed that these methods can leak information about training examples, increasing vulnerability to membership inference attacks that use full feature attributions as inputs to an attack model. In contrast, perturbation-based methods such as LIME and SmoothGrad (Smilkov et al. 2017) were found to be more resistant, likely due to their reliance on input perturbations rather than gradients, which may capture subtle distinctions between training and non-training points. Liu et al. (2024) further examined membership inference risks for multiple post-hoc explainers in the context of image data.

Defenses and countermeasures. to protect again inference attacks, some works in the literature used approaches that alter the training process such as DP-SGD (Liu et al. 2024) or approaches that perturb confidence scores of the target models output for each input by adding noise and then convert the perturbed confidence scores into adversarial examples for the attack models, such as MemGuard (Jia et al. 2019). However such approaches either prove to be ineffective against inference attacks (such as MemGuard) (Liu et al. 2024) or decrease the inference attack performance in the expense of severe degradation of the model utility as well as explanation quality such as DP-SGD and thus presenting a large trade-off between defense capability and utility and explainability performance (Liu et al. 2024).

Privacy in Explainable NLP. Prior work has primarily examined the privacy risks of model explanations in the context of tabular and image-based datasets. However, the intersection of privacy and explainability in NLP remains largely unexplored. In contrast to these studies, our work focuses specifically on textual datasets and NLP applications. Rather than analyzing privacy leakage from explanations, we empirically investigate whether data-level DP can be applied to achieve reasonable privacy guarantees while maintaining acceptable trade-offs in model utility and explanation quality. To the best of our knowledge, this is the first study to systematically evaluate the trade-off between data-level DP and the quality of post-hoc feature attribution explanations (in terms of their faithfulness) in NLP models.

3 Experimental Setup

3.1 Datasets

To guide our experiments, we select three datasets, which vary in size and domain. These datasets are introduced in the following, alongside their associated downstream tasks.

SST-2. The Stanford Sentiment Treebank dataset (SST-2) (Socher et al. 2013) comprises short texts originating from movie reviews, and it was popularized due to its inclusion in the GLUE benchmark (Wang et al. 2018). Each text is labeled according to its *sentiment*, i.e., either positive or

negative, creating a two-class binary classification task. We use the complete training split of the dataset, composed of 67,349 records, with an average word length of 9.41.

AG News. The AG News corpus contains over one million news articles from the over 2000 news sources. We utilize the subset as prepared by Zhang, Zhao, and LeCun (2015), which contains 120k news articles from four news domains: *world*, *sports*, *business*, and *sci/tech*. We take a 50% random sample (seed=42) for a final dataset of 60000 news articles, with an average word length of 43.90.

Trustpilot Reviews. The *Trustpilot* corpus is a large-scale collection of user reviews. The corpus prepared by Hovy, Johannsen, and Søgaard (2015) tags each review with the stars provided (1-5), which we simplify to negative reviews (1-2 stars) and positive reviews (5 stars). We specifically only take reviews from the US split of the dataset (*en-US*), and use a 10% random sample. This results in a dataset of 29,490 reviews, with an average word length of 59.75.

3.2 DP Methods

We select three local DP text rewriting methods for our experiments, which are introduced in the following.

TEM. TEM (Carvalho et al. 2023) is a word-level DP mechanism which leverages a generalized notion called *metric* DP. This generalization is useful for metric spaces, such as with word embeddings, wherein the indistinguishability requirement between words is scaled by their distance in the space. TEM improves upon previous approaches by employing a *truncated exponential mechanism*, allowing for higher utility word replacements. To privatize a complete text, each component word is privatized one-by-one. Following the original work, we choose the privacy budgets of $\varepsilon \in \{1, 2, 3\}$, which refer to the budgets *per word*.

DP-Prompt. Leveraging the generative capabilities of LLMs, Utpala et al. (2023) introduce DP-PROMPT, a method for producing private outputs texts with local DP guarantees by modeling privatization as a *paraphrasing* task. Internally, the DP-PROMPT mechanism operates by applying DP at each token generation, specifically by using the temperature parameter as an equivalent DP mechanism to the Exponential Mechanism (McSherry and Talwar 2007). Following the original work, we test three temperature values, $T \in \{1.75, 1.5, 1.25\}$. This translates to the per-token ε values of $\varepsilon \in \{118, 137, 165\}$, using the implementation provided by Meisenbacher et al. (2024), which employs a FLAN-T5-BASE model (Chung et al. 2022) as the underlying privatization model.

DP-BART. DP-BART is a document-level LDP text rewriting mechanism proposed by Igamberdiev and Habernal (2023). It leverages a pretrained BART model (Lewis et al. 2020), applying calibrated Gaussian noise in the latent representation space, where the decoder then decodes the noisy encoded vector to generate a private text. In doing so, DP-BART rewrites texts with a single document guarantee, albeit usually requiring higher privacy budgets for meaningful output. As such we choose $\varepsilon \in \{500, 1000, 1500\}$, and we use the original DP-BART-CLV variant.

Privatization Procedure. For DP-BART, we use the mechanism on the primary text column of each of our chosen datasets. This is repeated for each of the three privacy budgets, creating three private counterparts to the original data. Using DP-PROMPT likewise produces full texts as a result of the generative process. Finally, TEM is run sequentially on all component words of an input text, which are tokenized using the NLTK.WORD_TOKENIZE function. The list of outputs from the mechanism are then reconstructed into a single string via simple concatenation. In total, we thus produce nine private variants of the original datasets, for a total of 30 datasets (3 original baselines + 27 private counterparts).

A note on privatization and comparability. We caution that in the selection of privacy budgets for each DP mechanism, and for the resulting datasets based on these decisions, we do not ensure any comparability between the three selected methods. Due to the different manners in which DP is ensured across these three methods, as well as the intricacies involved with comparing different DP notions (e.g., MDP vs DP), we out-scope such comparisons. Instead, we focus on analyzing the downstream effects on DP rewriting *within* each method (i.e., across privacy budgets).

We also note that the choice of privacy budgets (ε values) are motivated by the choices taken by the authors of the original works. We do not work to normalize these values, i.e., by normalizing the logit values used in DP-PROMPT. While this would be a useful step in reporting more usable and reasonable privacy budgets, we endeavor to test the DP methods as presented originally. As such, we do not report on or analyze the relative strength of underlying DP guarantee, as this is an active area of DP-NLP research.

3.3 Models

We utilize a number of pretrained encoder-only language models, which varying in architecture and model size. In this, we empirically measure the effect of fine-tuning on DP rewritten datasets, namely, the downstream effect of this process on the explainability of these models (measured by our chosen metrics). We fine-tune *BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, and DeBERTa-base* on our three chosen classification datasets. We refer the reader to the repo for model details.

3.4 Explainability Methods

We include four post-hoc feature attribution methods in our experiments: **Gradient** involves computing the gradient of the output with respect to the input features; **Integrated Gradient** instead integrates the gradients over a path from a baseline input to the explained input; **SHAP** (Lundberg and Lee 2017) approximates the Shapley value of each feature, a concept from cooperative game theory that measures the "contribution" of each feature by considering different coalitions of features and how much each feature contributes to the outcome; **LIME** (Ribeiro, Singh, and Guestrin 2016) attempts to replicate the model's behavior locally around the explained input with a linear model that is easy to explain.

3.5 Evaluating Explanations

We evaluate the post-hoc explanations using two different methods each of measuring the **comprehensiveness** and **sufficiency** of explanations. Both metrics effectively attempt to quantify the *faithfulness* of explanations, i.e. how accurately they reflect the underlying decision process of a model (Jacovi and Goldberg 2020). Faithfulness is one of the most important desideratum for explanations (Danilevsky et al. 2020; Lyu et al. 2024) as high faithfulness indicates that the explanation accurately reflects the model's decision-making process for a given prediction.

In their simplest forms, comprehensiveness metrics measure the change in output probabilities for the true class when the top-k tokens with respect to the feature attribution scores are removed from the input, and sufficiency metrics measure the change when only the top-k tokens are given as input to the model. We then compute the AOPC by averaging the values by k is varied. We denote these metrics **AOPC-Comprehensiveness** and **AOPC-Sufficiency**.

Hard-removing the tokens might lead to out-of-distribution inputs for the model, which could adversely affect model performance during the evaluation of the explanations (Zhao et al. 2022; Chrysostomou and Aletras 2022). To address this potential problem, we also include **Soft Sufficiency** and **Soft Comprehensiveness** metrics (Zhao and Aletras 2023) in our evaluations. For the soft versions of these metrics, rather than removing a token entirely, a fraction of each token's embeddings is masked based on that token's importance score, i.e. the masked token is computed as $\mathbf{x}' = \mathbf{x} \odot \mathbf{e}$ where $\mathbf{e}_i \sim \text{Bernoulli}(q)$ with q being the importance score if the token is kept (sufficiency) and one minus the score if it is removed (comprehensiveness).

3.6 Composite Score

Understanding that the trade-off between privacy and explainability is not always equally weighted, we design a metric that allows one to *shift* the importance of either utility or explainability, while still viewing both in the same light. We compute a *composite score* for each model m and weight $\alpha \in \{0.25, 0.5, 0.75\}$ as $\text{CS}(m, \alpha) = \alpha\,\widehat{\text{F1}}(m) + (1-\alpha)\,\widehat{E}(m)$, where $\widehat{E}(m) = \frac{1}{4}\big(\widehat{C}_s(m) + (1-\widehat{S}_s(m)) + \widehat{C}_a(m) + (1-\widehat{S}_a(m))\big)$, and $\widehat{x}$ denotes min–max normalization of x, subscripts s and a indicate "soft" vs. "AOPC," C is comprehensiveness, and S is sufficiency (inverted via $1-\widehat{S}$ so that larger values are always better).

$$\text{CS}(m,\alpha) = \alpha\hat{F1} + (1-\alpha)\hat{E} \quad \hat{E} = \tfrac{1}{4}\big(\hat{C}_s + (1-\hat{S}_s) + \hat{C}_a + (1-\hat{S}_a)\big)$$

The composite score ranges from 0 to 1, where higher values (closer to 1) indicate near 1 means better overall performance (based on the selected α value). We intentionally define the score in this normalized form to enable a straightforward and consistent comparison across our experiments. Additionally, with this score, one can vary α based on the relative weight between utility and explainability. For example, an α of 0.25 would imply that the explainability faithfulness metrics are considerably more important, as opposed to a value of 0.75 where the utility score (F1) takes precedence.

4 Results and Analysis

Our results are presented across several tables and figures. Tables 1, 2, and 3 report composite scores (mean_{std}) for $\alpha = 0.25$, 0.5, and 0.75, respectively. Scores are averaged over five PLMs (BERT-BASE, BERT-LARGE, ROBERTA-BASE, ROBERTA-LARGE, and DEBERTA-BASE), and the *Avg* row represents the mean and pooled standard deviation across the four explainers (Gradient, IG, LIME, and SHAP) for each (dataset, DP-ε) pair. In each row, the highest composite score is underlined. Columns are grouped by DP method (None, DPB, DPP, and TEM) and color-coded for comparison: grey for no DP, cyan for DPB, orange for DPP, and green for TEM. Color intensity reflects score magnitude, with darker shades indicating higher values. For instance, in Table 1, the most saturated cell in the TEM columns highlights the top-scoring explainer–DP-ε combination (e.g., TEM3 with SHAP) for a given dataset and α value (e.g., AG News, $\alpha = 0.25$). This distinction in coloring is used to reflect the fact that DP methods are not directly comparable due to differences in mechanisms and assumptions (see Section 3.2). These tables support analysis of trends across explainers and privacy configurations.

To better visualize the results from Tables 1, 2, and 3 for each of the three datasets, we present corresponding plots in Figures 1a, 1b, and 1c. Each figure corresponds to one dataset, and each data point represents the composite score of a (DP-ε, explainer, α) triplet, averaged over the five PLMs. In these plots, different shapes indicate different explainers, while colors denote the α values *(0.25, 0.5, 0.75)* where explainers are acronymed as follows: (G:Gradient, IG:Integrated Gradient, L:LIME, S:SHAP).

In Table 4, we present composite score values, but we consolidate the composite score values averaged over the four explainers (instead of per explainer as done previously) for each (dataset, α, DP-ε) triplet thus allowing us to compare and evaluate how each DP-ε combination affect the composite scores over all the explainers allowing us to detect the *"sweet spots"* for each α value/scenario in each dataset.

In Table 5, we investigate the effect of model size on the trade-off between privacy, utility, and explainability; in particular, whether this trade-off differs between smaller and larger models. To this end, we average results over the four explainers, but instead of aggregating across all five models, we group the results by model size. The *base* group includes BERT-BASE and ROBERTA-BASE, while the *large* group includes their corresponding larger variants.

4.1 Composite Scores for Each Explainer Across DP-Methods and Datasets

Tables 1, 2, and 3 facilitates checking, per each α values and dataset, and for each explainer, the DP-ε method that provide the highest composite score, as well as for each DP-method, the ε values that leads to the highest score for a specific DP method. In addition, we look at the plots in Figure 1 and based on our results from the composite score plots presented there for SST2, AG News, and Trustpilot, evaluated across different DP mechanisms (ε), four explainability methods, and utility–explanation trade-offs (α = 25%, 50%,

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 1: Composite Scores (mean_{std}) with $\alpha = 0.25$ averaged over the five PLMs. *Avg* refers to the mean and pooled standard deviation over the four explainers.

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 2: Composite Scores (mean_{std}) for $\alpha = 0.5$ averaged over five PLMs. *Avg* refers to the mean and pooled standard deviation over the four explainers.

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 3: Composite Scores (mean_{std}) for $\alpha = 0.75$ averaged over five PLMs. *Avg* refers to the mean and pooled standard deviation over the four explainers.

75%), we can report the following:

General Observations. Higher α values, which place greater emphasis on model utility, consistently result in higher composite scores across all datasets and DP configurations. Among the explainers, LIME and SHAP generally outperform Gradient and IG in the composite scores, particularly under strong privacy constraints. As expected, composite scores decline significantly in stricter DP settings (e.g., DPB500, TEM1), reflecting the trade-off between privacy and both utility and explanation quality.

SST2 appears to be highly sensitive to DP noise, exhibit-

Dataset	Expl	α	None	DPB500	DPB1000	DPB1500	DPP118	DPP137	DPP165	TEM1	TEM2	TEM3
AG News	Avg	0.25	$0.713_{0.20}$	$0.301_{0.16}$	$0.654_{0.01}$	$0.698_{0.01}$	$0.750_{0.01}$	$0.765_{0.02}$	$0.758_{0.02}$	$0.344_{0.22}$	$0.610_{0.17}$	$0.786_{0.03}$
	Avg	0.50	$0.771_{0.13}$	$0.307_{0.20}$	$0.710_{0.01}$	$0.756_{0.01}$	$0.806_{0.01}$	$0.816_{0.01}$	$0.811_{0.01}$	$0.349_{0.26}$	$0.632_{0.19}$	$0.834_{0.02}$
	Avg	0.75	$0.830_{0.09}$	$0.313_{0.25}$	$0.766_{0.01}$	$0.813_{0.01}$	$0.861_{0.01}$	$0.867_{0.01}$	$0.864_{0.01}$	$0.355_{0.30}$	$0.653_{0.21}$	$0.881_{0.01}$
SST2	Avg	0.25	$0.516_{0.19}$	$0.218_{0.06}$	$0.349_{0.15}$	$0.442_{0.14}$	$0.488_{0.17}$	$0.474_{0.16}$	$0.474_{0.16}$	$0.232_{0.08}$	$0.274_{0.12}$	$0.394_{0.20}$
	Avg	0.50	$0.593_{0.22}$	$0.279_{0.09}$	$0.432_{0.20}$	$0.542_{0.18}$	$0.590_{0.21}$	$0.581_{0.20}$	$0.579_{0.20}$	$0.288_{0.11}$	$0.340_{0.15}$	$0.477_{0.24}$
	Avg	0.75	$0.669_{0.25}$	$0.340_{0.12}$	$0.517_{0.24}$	$0.641_{0.21}$	$0.693_{0.24}$	$0.689_{0.24}$	$0.685_{0.24}$	$0.344_{0.14}$	$0.407_{0.18}$	$0.561_{0.28}$
Trustpilot	Avg	0.25	$0.526_{0.18}$	$0.434_{0.11}$	$0.501_{0.18}$	$0.542_{0.13}$	$0.506_{0.15}$	$0.504_{0.09}$	$0.501_{0.06}$	$0.327_{0.16}$	$0.331_{0.12}$	$0.523_{0.17}$
	Avg	0.50	$0.593_{0.20}$	$0.509_{0.12}$	$0.608_{0.18}$	$0.679_{0.09}$	$0.613_{0.18}$	$0.640_{0.06}$	$0.646_{0.04}$	$0.411_{0.16}$	$0.428_{0.14}$	$0.639_{0.19}$
	Avg	0.75	$0.660_{0.24}$	$0.585_{0.12}$	$0.715_{0.20}$	$0.816_{0.04}$	$0.721_{0.20}$	$0.775_{0.04}$	$0.792_{0.03}$	$0.495_{0.16}$	$0.525_{0.17}$	$0.754_{0.22}$

Table 4: Consolidated Composite Scores (mean_{std}) for three α values, showing **the average over four explainers**.

α	Dataset	Base	Large	Δ (Large–Base)
0.25	AG News	$0.690_{0.15}$	$0.566_{0.25}$	-0.125
	SST2	$0.461_{0.13}$	$0.273_{0.10}$	-0.188
	Trustpilot	$0.499_{0.11}$	$0.380_{0.09}$	-0.119
0.50	AG News	$0.736_{0.14}$	$0.604_{0.29}$	-0.132
	SST2	$0.568_{0.13}$	$0.331_{0.12}$	-0.237
	Trustpilot	$0.621_{0.11}$	$0.479_{0.12}$	-0.142
0.75	AG News	$0.782_{0.13}$	$0.643_{0.33}$	-0.140
	SST2	$0.676_{0.14}$	$0.390_{0.14}$	-0.286
	Trustpilot	$0.743_{0.12}$	$0.578_{0.16}$	-0.165

Table 5: Comparison of average composite scores (mean_{std}) between base and large models, averaged across DP-ϵ values for each dataset and α.

ing considerable performance degradation under low privacy budgets. G and IG are especially affected in these settings, whereas LIME and SHAP demonstrate more stable performance, even under moderate DP configurations such as DPB1000 and DP-PROMPT. Among the TEM methods, TEM3 shows partial recovery in performance, especially when greater weight is given to utility (α high).

On the other hand, **AG News** displays stronger resilience to the effects of DP noise compared to SST2. LIME and SHAP maintain superior performance across most DP levels, with IG improving noticeably as ε increases. In particular, IG becomes competitive in configurations such as DP-PROMPT and TEM3. The TEM3 variant performs especially well at (α = 75%), indicating a favorable trade-off between utility and explanation quality.

Trustpilot demonstrates intermediate sensitivity to DP noise, falling between SST2 and AG News. Composite scores improve progressively with increasing ε, and LIME remains the most effective explainer across the majority of settings. DP-BART and DP-PROMPT mechanisms, particularly with moderate to high privacy budgets, offer reliable performance in maintaining explanation quality.

Cross-Dataset Trends. Across all datasets, LIME and SHAP consistently outperform IG and Gradient, which are more affected by strong DP. Gradient is the most negatively impacted explainer, particularly in low-ε regimes and on the SST2 dataset. Among the DP mechanisms, *DP-BART-1500* and DP-PROMPT-165 offer favorable trade-offs between privacy, utility, and explainability. While TEM1 and TEM2 generally result in considerable degradation, TEM3 performs better, especially for AG News and Trustpilot.

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 6: "Sweet spots" summary across privacy tiers and α values. Each row corresponds to a privacy tier (None = no DP; Small/Medium/Large ε = decreasing privacy). For each α, the left column shows the best Explainer–Dataset pair (from Tables 1–3), and the right column shows the best Dataset based on explainer-averaged scores (from Table 5). Epsilon levels group as: Small (DPB500, DPP118, TEM1), Medium (DPB1000, DPP137, TEM2), and Large (DPB1500, DPP165, TEM3).

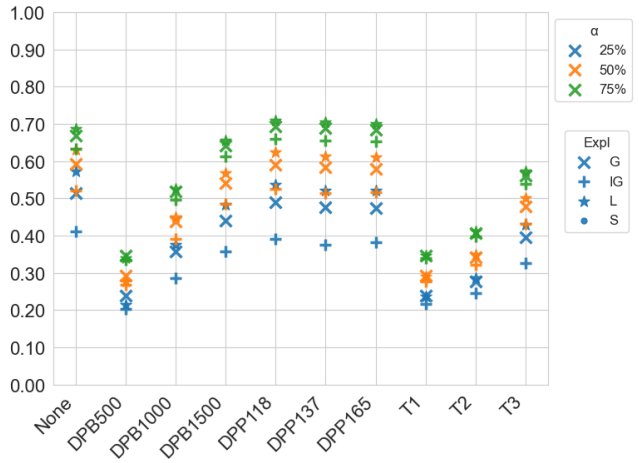
4.2 Comparison Over All Explainers Across DP-Methods and Datasets

Table 4 displays the average composite scores across the four explainers for each combination of dataset, α value, and DP method (DP-ε), averaged over the five employed PLMs. We draw the following results and insights.

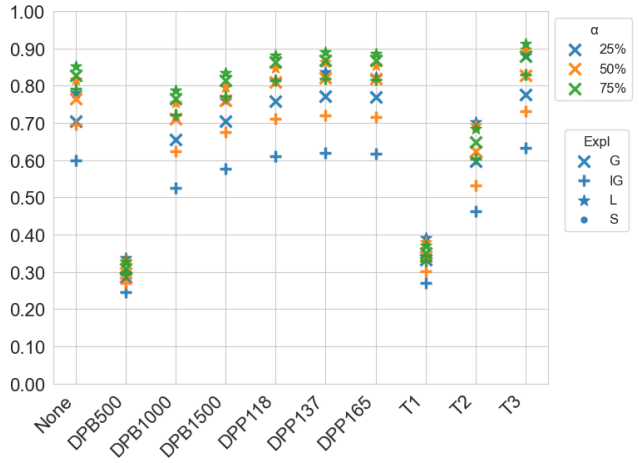
Effect of the Utility–Explanation Trade-Off (α). Across all datasets and DP setups, higher values of α (i.e., greater weight placed on utility) consistently yield higher composite scores. This is especially visible in the blue curves (α = 0.75), which dominate across most configurations. This trend reflects the stabilizing role of utility in the composite score, particularly under privacy-induced degradation.

Comparing DP Mechanisms. Among the DP strategies, DP-BART and DP-PROMPT (particularly at higher ε values such as 1500 and 165) achieve the highest average composite scores, especially for AG News and Trustpilot. In contrast, the TEM methods (T1 and T2) result in a substantial performance drop, most notably for *SST2*. As expected, DPB500 (strongest privacy budget) leads to the lowest performance across all datasets and α values, confirming the cost of tight privacy constraints.

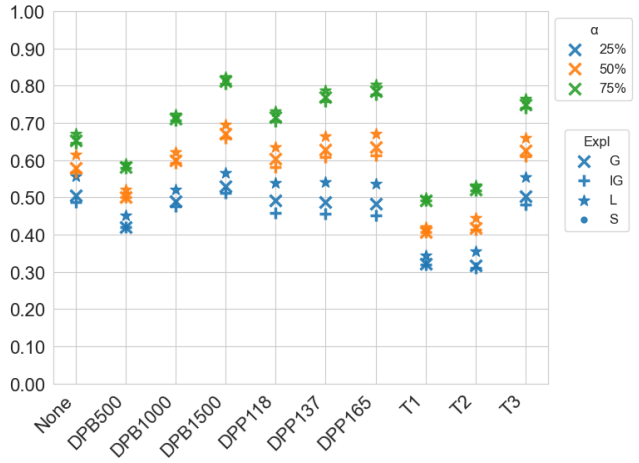
Dataset Sensitivity. SST2 consistently produces the lowest composite scores, reflecting its greater vulnerability



(a) SST2



(b) AG News



(c) Trustpilot

Figure 1: Composite Score by DP-ε, explainer, α, dataset over 5 models.

to privacy-preserving perturbations. This could be due to shorter input lengths or more subtle semantic cues required for sentiment analysis. In contrast, AG News emerges as the most robust dataset, maintaining high composite scores even under moderate DP conditions. *Trustpilot* performs moderately well and demonstrates steady recovery with increasing ε, particularly under DP-BART and DP-PROMPT.

Privacy–Performance Trade-Off. Overall, there is a clear upward trend in composite scores with increasing ε. This pattern is most pronounced in the DP-BART and DP-PROMPT curves, where performance improves steadily from $\varepsilon = 500$ to $\varepsilon = 1500/165$. This monotonic behavior confirms the expected trade-off: as privacy constraints are relaxed, utility and explanation quality are jointly improved.

4.3 Identifying Sweet Spots for Privacy-Utility-Explainability

We construct Table 7 using the results form Tables 1,2,3 and 4, where we identify and present the sweet spots based on the composite scores across privacy tiers and alpha values settings for two cases: explainer-dataset pairs (from Table 1,2,3) and the datasets with the highest average scores over the four explainers (From Table 3). Based on these result in Table 7, we observe the following:

Explainer–Dataset Consistency. Across all *privacy tiers*, LIME–AG News emerges as the top explainer–dataset pair in every case except the strictest privacy setting (ε small), where SHAP–AG News slightly outperforms it. This suggests that SHAP's token-level importance is more robust under heavy DP noise, but LIME generally provides the most faithful and useful explanations once privacy is relaxed.

Dataset-Level Robustness. When averaging across explainers, AG News consistently wins for no DP, medium privacy, and low privacy budgets. Under the tightest DP budget (small ε), however, Trustpilot takes the lead for all α values, indicating that explanations on Trustpilot degrade less, on average, under stringent privacy constraints.

Effect of Increasing ε. As ε increases (privacy is relaxed), composite scores rise monotonically for both the best explainer–dataset pair and the dataset-average. The largest recovery occurs between small and medium ε, especially for SHAP on AG News, highlighting that a moderate privacy budget recovers most of the lost explanation signal.

Utility-Explainability Trade-Off (α) Trends. Increasing α from 0.25 (explanation-focused) to 0.50 (balanced) to 0.75 (utility-focused) uniformly boosts all composite scores, since more weight is placed on classification F1. Notably, the ranking of sweet spots remains stable: LIME–AG News dominates once $\varepsilon \geq$ medium values, and *Trustpilot* remains the dataset of choice under the smallest ε.

Although high composite scores are expected with large ε values, it is noteworthy that even with medium and small ε (i.e., stricter privacy), some explainers still achieve strong performance. These results are promising, demonstrating that explainability can be achieved under tighter privacy constraints. For example, SHAP achieves this across all three α values representing different utility–explanation

trade-offs. This indicates that for certain combinations of dataset, explainer, and DP method, it is possible to attain high privacy while still maintaining strong utility and explanation quality, depending on the specific objective (i.e., lower or higher α).

4.4 Comparison Across Model Size

In Table 5, we investigate the effect of model size on the trade-off between privacy, utility, and explainability; in particular, whether this trade-off differs between smaller and larger models. To this end, we average results over the four explainers, but instead of aggregating across all five models (as in the previous tables), we group the results by model size. The *base* group includes BERT-BASE and ROBERTA-BASE, while the *large* group includes their corresponding larger variants. We exclude DeBERTa from this analysis, as we don't employ DeBERTa-large in our experiments. We consider this comparison between base and large variants of BERT and RoBERTa sufficient for a preliminary assessment of model size effects. Table 5 presents a comparison of (average across explainers) composite scores (mean_{std}) between base and large models, averaged across DP-ϵ values for each dataset and α. We learn the following:

Base Models Outperform Large Models. Across all datasets and α values, base models consistently outperform large models in terms of composite score. This is evident from the uniformly negative Δ (Large – Base) values, ranging from -0.119 to -0.286. The results suggest that increasing model size under DP leads to a decline in the quality of the resulting predictions and explanations.

Dataset Sensitivity. The *SST2* dataset shows the largest negative impact of model size. For instance, the gap reaches $\Delta = -0.286$ at $\alpha = 0.75$. In contrast, the performance drop for *AG News* and *Trustpilot* remains more moderate (around -0.12 to -0.16). This indicates that *SST2* may be more sensitive to privacy-induced noise, possibly due to shorter inputs or the nature of sentiment-based classification.

Effect of α (Utility–Explanation Trade-Off). The negative performance gap between large and base models generally widens as α increases (i.e., as more emphasis is placed on utility over explanation quality). This trend is particularly strong for *SST2*, where the difference grows from -0.188 ($\alpha = 0.25$) to -0.286 ($\alpha = 0.75$). This suggests that utility suffers more under large models when trained with DP.

Stability and Variability. The standard deviations are generally higher for large models, especially under $\alpha = 0.75$. This points to greater instability and inconsistency in the performance of large models under DP, in addition to their lower mean scores.

5 Discussion

We reflect on the main findings of our experiments, giving way to a collection of practical recommendations, as well as future points to consider.

Does dataset matter in the privacy-explainability trade-off? An initial review of the experimental results suggests that the *nature* of the dataset may contribute to different outcomes in terms of privacy versus explainability. One possible explanation for AG News outperforming SST2 and Trustpilot in terms of composite scores under multiple DP methods is that its topic-classification nature results in inputs that contain multiple contextually aligned keywords, which are often robust to DP-induced perturbations whereas SST2's short sentiment sentences and Trustpilot's informal user reviews may lose critical cues more easily. Additionally, PLMs achieve a higher baseline F1 on AG News, which could further buffer performance degradation under privacy constraints. Finally, the concentration of attribution weights on a few salient words in AG News may yield more faithful explanations compared to the more diffuse attributions required for sentiment or user-generated text. While these factors require further investigation, they offer a plausible account of AG News resilience in composite utility–explainability metrics with strict privacy budgets.

Despite the absolute differences exhibited by the AG News results in comparison to our other two datasets, one promising trend emerges. When looking at the composite score curves for any dataset in Figure 1 one can observe that a similar line is followed regardless of dataset, for any DP method or explainer. While these curves are not strictly uniform across datasets, the similar trends indicate a crucial point for future investigations, namely to determine the extent to which dataset matters in the combined study of privacy, utility, and explainability. Nevertheless, one must also keep in mind the differing performances of the various explainers we employ, which also seem to be impacted by the nature of dataset, and accordingly, the downstream task.

When privacy meets explainability. Harmonizing all of the experimental results, we converge on one important discussion point, which relates back to the initial research question we posed in this work: *how does DP impact explainability?* Diving deeper, we also reflect on the question of whether DP and explainability can co-exist, or if there possibly remains friction between these two important mandates.

A helpful starting point lies in the comparison of baseline results (i.e., those achieved without any DP) and those post-privatization. While in the majority of cases, the application of DP leads to decreases in explainability, this is not always the case. Indeed, in some scenarios for all three employed DP methods, some results outperform the non-private baselines, and this occurs at least once among all (DP method, ε) combinations for each explainer. Furthermore, this result is most pronounced in the setting where explainability is most preferred ($\alpha = 0.75$), suggesting that when utility loss is less important, the effects of DP gain a more positive light. In this, we uncover that in some scenarios, DP actually serves to *improve* explainability, a very promising prospect.

Beyond these results, we also wish to answer our research question from the angle of *which* DP methods have *which* impacts on downstream explainability. This must be approached carefully, as noted previously, since we cannot draw conclusions between DP mechanisms operating on dif-

ferent lexical levels and with different privacy guarantees. However, important trends do appear *within* mechanisms, such as the relative stability of the DP-PROMPT results as opposed to the much steeper fluctuations from TEM or DP-BART. We also conjecture that the choice of DP method also is tightly intertwined with the nature of the classification task, where generative methods (DP-PROMPT and DP-BART) generally lead to more favorable results in our two binary classification tasks, whereas TEM consistently outperforms in the four-class AG News task. These results point to the importance of DP method choice, which can be heavily reliant on the downstream task at hand. This is made especially evident in Table 4, where all DP methods achieve the best consolidated score in at least one configuration.

Recommendations for Practitioners. Based on our findings, we compile a collection of recommendations for practitioners at the intersection of data privacy and explainability.

Based on our "sweet spot" analysis, for *explanation quality* (low α), practitioners can use SHAP at the strictest privacy setting, otherwise default to LIME under moderate or low privacy. For *utility* (high α), LIME is the sweet spot across all but the strictest privacy settings.

Based on the analysis of Figure 1, *where we compare scores of each of the four explainers*, for applications requiring both strong privacy and explanation quality, LIME or SHAP combined with DPBart-1500 or DPPrompt-165 is advisable. In the case of SST2, aggressive privacy strategies such as TEM1 and TEM2 should be avoided, and moderate-DP setups using LIME are preferable. For AG News, LIME in combination with DPPrompt-165 or TEM3 performs robustly across all α values. Trustpilot demonstrates broad robustness and is a suitable candidate for real-world deployment in privacy-sensitive NLP scenarios.

When averaging *across the four explainers*, our results show that using DPBart-1500 or DPPrompt-165 is recommended to achieve the best trade-off between privacy, utility, and explanation quality. TEM1 and TEM2 should be avoided when explanation faithfulness is a priority, especially for sensitive datasets such as SST2. For utility-focused applications (high α), AG News paired with moderate-DP settings performs reliably well. In general, higher ε values are preferable when maintaining interpretability is critical.

Based on the *model size effect (base vs large model) analysis*, for privacy-preserving NLP tasks that *require both high utility and faithful explanations*, *base models* are consistently more reliable and effective than *large models*. This is especially true when utility is prioritized (high α) or for sensitive datasets such as *SST2*.

These recommendations based on our experiments can be generalized into the following set of guidelines that are important to consider in the research or practice of *explainable privacy* with natural language data:

1. **Decide on the importance of privacy vs. utility**: an important starting point is the setting of α, as this may be considerably different across various use cases.
2. **Consider the nature of downstream task**: our results indicate that dataset (and task) are important factors in the juxtaposition of privacy and explainability. This becomes especially important in the following point.
3. **Choose your privatization method wisely**: our initial findings suggest that for more complex tasks (e.g., multi-class classification), non-generative methods such as TEM may be more suitable. However, for more "colloquial" datasets, such as those stemming from user reviews, DP methods based on generative models may be more suitable to preserving both utility and explainability. While this guideline requires further validation, we emphasize the important interplay between choice of privatization method and nature of downstream task.
4. **Choose the smallest pretrained model acceptable for the given use case**: we learn that across our results, composite scores decrease as model size increases, showing how smaller models, if acceptable for a given use case, may be preferable in finding a balance between privacy, utility, and explainability.
5. **Measure on a variety of explainers**: we find that despite individual difference between explainers across all tested configurations, using averages and composite scores lead to clear emergent trends and interpretable differences between privacy levels (ε values) and datasets/tasks. We therefore recommend the usage of multiple explainers, tied together by a composite score, for a robust overview of performance differences between setups.

6 Conclusion

We conduct an investigation at the intersection of privacy and explainability, guided by the overarching methods of differentially private text rewriting and post-hoc explainability. In a series of experiments, we quantify the privacy-explainability trade-off, which lends interesting insights regarding the potential synergies between the two important topics. We are the first to conduct such an investigation in the context of natural language data, providing the foundations for further explorations into this interdisciplinary topic.

Future work: we envision a number of paths for future work based on our investigation, namely: (1) Further investigating "sweet spots", particularly via the integration of more robust proxies for privacy (i.e., beyond ε values). This could include for example the inclusion of membership inference testing, as performed by Shokri et al. (2021). (2) Extending our findings with experiments on additional datasets, Explainability and DP methods, and ε ranges. (3) Investigating the trade-offs with respect to non post-hoc explainability methods, as well as considering more recent frameworks for measuring faithfulness of explanations, such as that proposed by Zheng et al. (2025). (4) Including human evaluation (i.e, perceptions) into the calculation of composite score, namely to improve this score beyond automatic metrics.

Limitations: our study relies solely on quantitative evaluation without human assessment, focuses narrowly on post-hoc explainability and DP text rewriting, and omits nuances in DP mechanism design and guarantees. These constraints highlight the need for broader work to advance understanding of the explainability–privacy trade-off in NLP.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS23069 Software Campus 3.0 (TU München).

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.

Feyisetan, O.; Balle, B.; Drake, T.; and Diethe, T. 2020. Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, 178–186. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368223.

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.

Hu, L.; Habernal, I.; Shen, L.; and Wang, D. 2024. Differentially Private Natural Language Models: Recent Advances and Future Directions. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 478–499. St. Julian's, Malta: Association for Computational Linguistics.

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.

Vu, D. N. L.; et al. 2024. Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 507–527. Miami, Florida, USA: Association for Computational Linguistics.

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).