Evaluating Synthetic Data Generation from User Generated Text

Jenny Chim[1], Julia Ive[1], and Maria Liakata[2]

[1]Queen Mary University of London, School of Electronic Engineering and Computer Science
c.chim@qmul.ac.uk, j.ive@qmul.ac.uk
[2]Queen Mary University of London / The Alan Turing Institute, School of Electronic Engineering and Computer Science
m.liakata@qmul.ac.uk

User-generated content provides a rich resource to study social and behavioral phenomena. Although its application potential is currently limited by the paucity of expert labels and the privacy risks inherent in personal data, synthetic data can help mitigate this bottleneck. In this work, we introduce an evaluation framework to facilitate research on synthetic language data generation for user-generated text. We define a set of aspects for assessing data quality, namely, style preservation, meaning preservation, and divergence, as a proxy for privacy. We introduce metrics corresponding to each aspect. Moreover, through a set of generation strategies and representative tasks and baselines across domains, we demonstrate the relation between the quality aspects of synthetic user generated content, generation strategies, metrics, and downstream performance. To our knowledge, our work is the first unified evaluation framework for user-generated text in relation to the specified aspects, offering both intrinsic and extrinsic evaluation. We envisage it will facilitate developments towards shareable, high-quality synthetic language data.

1. Introduction

User-generated content (UGC), such as social media posts, is an invaluable resource for studying natural linguistic and behavioral phenomena. Such data often contain sensitive information, be it explicit leakages of personally identifiable information (PII) or implicit signals of protected attributes (e.g., race, gender) and other revealing information (e.g., physical location) (Humphreys, Gill, and Krishnamurthy 2010; Fire, Goldschmidt, and Elovici 2013; Bazarova and Choi 2014; Horvitz and Mulligan 2015). Moreover, despite the abundance of unlabeled data and the availability of distantly supervised noisy labels, high-quality ground truth annotations remain in short supply, in large part due to the resources and effort required for expert labeling (Snow et al. 2008; Williamson 2016). This situation is aggravated by online platforms closing off API

Action Editor: Carlos Gómez-Rodríguez. Submission received: 10 April 2024; revised version received: 16 July 2024; accepted for publication: 26 August 2024.

https://doi.org/10.1162/coli_a_00540

access (Davidson et al. 2023). These circumstances form a bottleneck to sharing research data and developing effective computational tools. To this end, synthetic language data generation poses a promising avenue to mitigate data scarcity and privacy concerns.

Consider a scenario where there is a limited amount of high-quality labeled data. To develop language applications, researchers can try in-context learning with large language models (LLMs) (Brown et al. 2020); however, it is not yet established that LLMs can reliably solve specialized tasks (Qin et al. 2023) and domains that have not been encountered during pre-training (Ziems et al. 2024). LLM-based solutions have also yet to address challenges such as practical constraints around data sensitivity, output factuality, and explainability. The paradigm of *data augmentation* through synthetic data generation in order to develop smaller expert models or fine-tune LLMs presents an attractive alternative solution. A data-centric evaluation framework can help researchers focus on developing high-quality data for this purpose, enabling quick iterations and comparisons between generation strategies without needing to exhaustively train downstream models to assess task performance and privacy risks.

From a *privacy* perspective, consider a scenario where researchers would like to use cloud APIs or external computing resources, but are unable to move their data due to sensitivity, institutional review board protocols, or company requirements; or where they want to fine-tune open models with sensitive data but where releasing such a model could result in leakages of memorized private information (Carlini et al. 2021). One workaround is to locally generate privacy-preserving data that are compliant with policy requirements and privacy safety. They can develop models and tools leveraging external resources on synthetic data, then share trained artefacts with collaborators or incorporate models back into their analysis workflow to label real private data. In this use case, holistic assessments of data quality are needed to ensure that validation results on synthetic data reflect performance on real data, for auditing and utility purposes.

Our article makes the following contributions:

- We identify the desiderata of synthetic UGC in terms of different aspects particularly relevant to downstream tasks, such as meaning and style preservation, as well as divergence as a proxy for privacy (§3).
- We present the first evaluation framework to assess the quality of synthetic textual data generated from UGC and other language interactions that integrates intrinsic evaluation across aspects with downstream classification and privacy experiments. Within the framework we incorporate a wide set of generation strategies ranging from rule-based methods to LLMs, intrinsic evaluation metrics at both the sample and distribution-level, and a number of extrinsic tasks across domains (§4).
- We benchmark generation models (§4.1) on the set of representative tasks in privacy-oriented data substitution and augmentation experiments (§5), using existing privacy experiments and novel speaker profiling and user re-identification setups.
- We discuss the utility and privacy implications of data aspects as measured by the framework (§6). Importantly we investigate the role of style preservation, an often neglected requirement in synthetic data, and identify data quality differences within LLM batch-prompted outputs (§6.1).

- We provide recommendations based on results from extensive empirical experiments on representative tasks (§7). These insights are accompanied by ready-to-use code for metrics and experiments in order to guide selection of generation and evaluation strategies in different types of applications.[1]

Our work highlights the importance of holistically evaluating synthetic texts to inform data sharing and modeling decisions, and we believe our framework will foster progress towards the utility of generated texts and preserving privacy in language data.

2. Related Work

2.1 Synthetic Data Creation and Use in Applications

Much prior work on creating shareable synthetic data comes from the clinical domain. Theory-based modeling of patient trajectories (Walonoski et al. 2018) and models that approximate the distribution of real data are used to generate continuous and structured electronic health records (Buczak, Babin, and Moniz 2010; Choi et al. 2017; Xu et al. 2019a) as well as synthetic unstructured clinical data (Wang et al. 2019; Melamud and Shivade 2019; Ive et al. 2020).

In natural language processing (NLP), conditionally perturbed and generated data are used to tackle data sparsity in areas such as grammatical error correction (Foster and Andersen 2009; Sakaguchi, Post, and Van Durme 2017), dependency parsing (Wang and Eisner 2016), and question answering (Hermann et al. 2015; Alberti et al. 2019). Synthetic data also benefit task performance and fairness, particularly in low-resource scenarios (Xia et al. 2019; Zmigrod et al. 2019; Tan et al. 2020), and have been proposed as an avenue towards privacy-preserving data sharing (Shetty, Schiele, and Fritz 2018; Mattern et al. 2022; Igamberdiev and Habernal 2023). With recent advances, pretrained language models are increasingly used to generate data for augmentation, with methods ranging from prepending labels to examples for class conditioned generation (Kumar, Choudhary, and Cho 2020) to using prompted synthetic data to develop smaller classification models (Yoo et al. 2021; Li et al. 2023b; Veselovsky et al. 2023; Møller et al. 2023), smaller language models (Ye et al. 2022; Li et al. 2023a; Eldan and Li 2023), dense passage retrievers (Dai et al. 2023), and further fine-tuning of the LLM itself (Wang et al. 2022), using seed examples, instruction prompts, and filtering mechanisms.

Current evaluation in this line of work focuses on surface-level divergence and task performance, without further examination of semantics, style, and their downstream practical implications. Our work addresses this gap by presenting a multi-faceted evaluation framework with empirical experiments, focusing on use cases where such detailed evaluation is especially necessary, e.g., domain-specific datasets with expert annotations.

2.2 Evaluating Natural Language Data

Natural language data are commonly evaluated when selecting meaningful subsets to label and model under resource constraints. For example, quality of large-scale pretraining data is estimated using heuristics (Dodge et al. 2021), classifiers trained on

1 https://github.com/Maria-Liakata-NLP-Group/synthetic-ugc-evaluation.

reference corpora (Brown et al. 2020), and scores based on probabilities under reference models such as perplexity (Marion et al. 2023). In other tasks, data are scored for representativeness (Welling 2009), spurious correlations (Le Bras et al. 2020), difficulty and ambiguity (Swayamdipta et al. 2020), and segment saliency (Hills et al. 2023). While the above criteria are applicable to synthetic data, our work focuses on cases where source corpora are known a priori, and the aim is therefore to assess how well synthetic data capture properties of the pre-selected natural data. In this setting, synthetic data evaluation is closely related to wider evaluation challenges in natural language generation (NLG).

2.3 Evaluating Synthetic Data

Standard measures of synthetic data quality include dimension-wise probability and dimension-wise prediction, applied to categorical and mixed data types (Choi et al. 2017; Tantipongpipat et al. 2021), and Fréchet distance, originating in computer vision (Heusel et al. 2017). Alaa et al. (2021) proposed to assess data based on distribution overlap in terms of fidelity, diversity, and generalization. Although these metrics have been applied on mixed data types, they do not necessarily translate to synthetic texts. Dimension-wise metrics are suitable when dimensions are meaningful in isolation (e.g., medical codes), but language is challenging to numerically summarize in the same fashion. Furthermore, dimension and distribution-wise metrics alone neither clearly reflect aspects on which texts differ (e.g., specific stylistic traits) nor accommodate desirable divergences (e.g., successfully preserve meaning while stylistically distinct), posing barriers to effective error analysis (van Miltenburg et al. 2021), refinement, and mitigation. Our work addresses this need by evaluating multiple textual aspects such as style preservation and divergence in assessing the quality of synthetic UGC.

2.4 Evaluating Model-generated Language Data

Model-centric Benchmarks. Existing benchmarks evaluate models over NLG tasks (Zhu et al. 2018; Gehrmann et al. 2021) and tasks that can be completed through the text interface of LLMs (Liang et al. 2022; Srivastava et al. 2023). Their focus is on comprehensively assessing model capabilities across scenarios. By contrast, our evaluation is data-centric, focusing on the quality and properties of the generated data.

Factuality. Similar to work that study the factuality of model-generated texts relative to grounding documents (Gabriel et al. 2021; Devaraj et al. 2022; Dziri et al. 2022) and external knowledge sources (Gupta et al. 2022; Rashkin et al. 2023), we consider original data to be grounding documents and assess synthetic data on how well they preserve their semantic and stylistic information; however we additionally focus on privacy preservation and how they affect downstream performance.

Conversational Agents. Similar to dialogue generation (Liu et al. 2016), synthetic UGC requires evaluation that accommodates diversity without assuming single fixed ground truths, and requires assessing *how* information is delivered using targeted metrics. Given these requirements, it is inadequate to naïvely use single metrics as an evaluation catch-all (e.g., using perplexity as the sole automatic metric for human-likeness, sensibility, and specificity (Kulshreshtha et al. 2020), applying suites of metrics without defining what they intend to capture (e.g., Hori and Hori 2017), or skipping intrinsic evaluation completely (e.g., He et al. 2022). We complement prior work that underscore the importance of dimensional evaluation (See et al. 2019; Kasai et al. 2022; Finch,

Finch, and Choi 2023) sensitive to task-specific nuances (Sai et al. 2021), by introducing an extensible multi-aspect framework for pairwise and distribution-wise generation settings in response to evaluation needs in synthetic UGC.

Privacy-oriented NLP. Our work is closely related to *author profiling* (i.e., inferring demographic traits), *authorship attribution* (i.e., linking documents to authors), and *privacy-preserving rewriting* (i.e., sabotaging an adversary's success on the above tasks). In privacy-preserving rewriting, much prior work has been dedicated to techniques that only work on datasets with unique author identifiers and/or labels of sensitive attributes (e.g., age, gender, race). The personal identifiers or attributes are used to remove authorship cues, by transforming texts closer to an alternate author or demographic group (Shetty, Schiele, and Fritz 2018; Mahmood et al. 2019; Xu et al. 2019b) or to a pooled average within the author's demographic group (Karadzhov et al. 2017; Mireshghallah and Berg-Kirkpatrick 2021). To quantify privacy preservation in these tasks, researchers commonly use (reductions in) an adversary classifier's ability to predict user attributes (Shetty, Schiele, and Fritz 2018) and text authorship. Authorship obfuscation work tends to be assessed in closed-world settings, for example, on news articles and essays written by fewer than 15 authors (Altakrori, Cheung, and Fung 2021; Altakrori et al. 2022) and in larger cross-genre document-level datasets (Stamatatos et al. 2015).

This article extends this line of work. First, we use rewriting methods that operate without user identifiers and demographic labels (§4.1), since such information is not always available in datasets, and profiling users to obtain silver-standard demographic labels for rewriting may cascade biases. Second, we expand evaluation to shorter, noisier UGC (i.e., social media, transcripts) (§5), measure user demographic profiling risks in a state-of-the-art setting (§5.2), and examine re-identification in a real-world setting, considering a longitudinal dataset and identification against 37.9k users (§5.3).

2.5 Biases and Synthetic Data

In NLP, bias is a widely studied yet inconsistently defined concept (Blodgett et al. 2020) referring to skews in data distributions that compromise representativeness and lead to allocational and representational harms (Barocas et al. 2017; Olteanu et al. 2019). Biases are embedded in data as reflections of norms and perspectives. When used for training, they become encoded in downstream models and systems (Bolukbasi et al. 2016; Shah, Schwartz, and Hovy 2020; Sheng et al. 2021). They also naturally emerge from sampling, annotation, representation, modeling, and design decisions (Hovy and Prabhumoye 2021), for example, when annotators' subjective perceptions propagate downstream to bias model predictions (Sap et al. 2019; Thorn Jakobsen et al. 2022; Mohamed et al. 2022).

Synthetic data may perpetuate, amplify, and introduce biases. This can happen when the generation model (1) is influenced by learned associations (Yu et al. 2023) and (2) faithfully reproduces biases in the input UGC (Wullach, Adler, and Minkov 2021). From an error disparity angle, bias may be introduced when (3) synthetic data subsets are generated with uneven quality (e.g., with respect to label validity, well-formedness). Consider a scenario involving English-only and code-switched texts (CS; e.g., Hsu et al. 2023): poorly synthesized CS data can lead to downstream tools performing poorly on real CS texts, adversely and disproportionately impacting users in the linguistic minority. Error disparity also emerges from (4) low diversity, for example, when tail classes are neglected in recursive generations, resulting in degenerated texts from model

collapse (Shumailov et al. 2023) or biased data that are of limited practical utility as they only retain information for the majority group (Wyllie, Shumailov, and Papernot 2024).

Such bias perpetuation and amplification risks are increasingly pertinent due to the growing adoption of systems that train models on their own synthetic data (Wang et al. 2022) and outputs from larger models (Lukasik et al. 2022; Li et al. 2023a). At the same time, synthetic data can be used to tackle biases, for instance, through targeted augmentation addressing known class (Zhao et al. 2017; Dinan et al. 2020; Qian et al. 2022) and attribute imbalances (Yu et al. 2023). Through facilitating multi-aspect evaluation of synthetic texts, our framework can aid high quality targeted generation, thereby contributing to bias reduction.

2.6 Capturing Subjective Experiences in UGC

Large quantities of synthetic data can be produced in the general domain with minimal guidance, relying on parametric knowledge, seed examples, and templates (Wang et al. 2022; Dai et al. 2023; Yu et al. 2023; Xu et al. 2024). However, we cannot directly assume that models can capture intricate social and behavioral phenomena when generating from UGC. For one, tasks modeling UGC often involve capturing subjective experiences and community-specific knowledge, but models fail to reflect human diversity, over-representing perspectives of a small subset of demographics (Santurkar et al. 2023; Durmus et al. 2023). While it is possible to steer generations towards different viewpoints, for example, via conditioning on personas (Zhang et al. 2018; Park et al. 2022) and behavioral descriptions (Jiang et al. 2023), such simulations can contain caricatures and misportrayals (Cheng, Piccardi, and Yang 2023; Cheng, Durmus, and Jurafsky 2023; Wang, Morgenstern, and Dickerson 2024). Moreover, prior benchmarking in the computational social sciences found that, for *classification*, models can perform well when there are objective ground truths, clearly defined colloquial labels, and short document lengths (e.g., tweet-level emotion prediction); however, they struggle with subjective expert taxonomies (e.g., empathy), large label spaces (e.g., character tropes), and complex structures (e.g., conversation-level persuasion) (Ziems et al. 2024). To overcome these challenges in a *generative* setting would pose extra complexities.

We argue that synthetic UGC should be generated with content and stylistic constraints to better reflect identities and experiences expressed in the original data. Thus, we identify meaning and style preservation as desiderata in our framework (§3) and study their role in representative downstream tasks (§5).

3. Desiderata for Synthetic Textual UGC

There are core aspects of data quality expected in synthetic texts. For example, they can be useful in providing data augmentation (Chen et al. 2023) to improve downstream performance. Moreover there are aspects of synthetic data quality that are independent of the data generation strategy used to obtain them. To this end, we identify core aspects of intrinsic data quality below and implement these metrics in our framework.

3.1 Meaning Preservation

Generated texts are commonly evaluated on meaning preservation (Xu et al. 2019b; Adelani et al. 2021). The assumption is that, in labeled datasets, outputs that preserve the original meaning will retain label validity and therefore be useful examples to train models in downstream tasks. Such works focus on 1-to-1 text rewriting and utilize sample-level metrics (e.g., BLEU [Papineni et al. 2002], METEOR [Banerjee and

Lavie 2005]). In practice, synthetic data generation broadly involves sampling from a distribution learned from the source data. When there is no direct mapping between original and synthetic examples, meaning preservation would need to be measured via distribution-level metrics, for example the Fréchet distance between the embedded original and synthetic texts. In this article, we report BERTScore (Zhang et al. 2020) and assess distribution-level meaning preservation with Fréchet distance computed via BERT embeddings (Xiang et al. 2021).

3.2 Style Preservation

Synthetic UGC generation is driven by downstream tasks that make it necessary to afford control over stylistic elements. Such tasks include preserving linguistic diversity within datasets (Blodgett, Green, and O'Connor 2016), generation from transcribed speech while preserving linguistic markers of mental health status or psychiatric conditions (e.g., disfluencies; Howes et al. 2017), and generation from interaction data (Wang and Jurgens 2018) as well as generation for conversational agents (Fitzpatrick, Darcy, and Vierhile 2017), where stylistic synchrony reflects empathy and inter-speaker understanding (Ireland and Pennebaker 2010; Lord et al. 2015).

Style is multidimensional (DiMarco and Hirst 1993) and influenced by extralinguistic factors, including communicative goals, topics, and demographics (Nguyen et al. 2016). However, it is common in NLG to consider it as isolated high-level variables entangled with semantics, such as sentiment (Hu et al. 2017) and politeness (Niu and Bansal 2018). Closely related are research in machine translation that model speaker (Michel and Neubig 2018) and translator (Wang, Hoang, and Federico 2021), and closed-set author style transfer (Xu et al. 2012). Evaluation in this line of work use attribute classifiers, which capture style on singular dimensions, and personalized models and corpus-level features for each author, which can be ill-suited in practice as they assume user identity is known a priori and require sufficient per-user data. In our work, we circumvent these restrictions, drawing on the idea that high-level styles are compositions of granular stylistic elements (Lyu et al. 2021). We assess overall idiolect with style embedding similarity and target syntactic style with part-of-speech (POS) based scores.

3.3 Divergence

We define divergence and diversity, two related but distinct concepts. **Divergence** in our work refers to the dissimilarity between source and synthetic texts. **Diversity** refers to the dissimilarity between outputs produced based on the same input.

Divergence is essential to synthetic data. From a utility angle, synthetic data should be different from their source to promote generalization, but not to the extent of compromizing label validity. In contrast to general NLG, where information precision is a key objective (Reiter and Dale 1997) and outputs tend to be penalized if they contain information absent from underlying grounding information, namely, hallucination (Maynez et al. 2020; Ji et al. 2022), in synthetic data, the introduction of new information can be benign and even beneficial so long as it does not contradict the original grounding text. From a privacy angle (§3.4), divergent data are less likely to contain direct regurgitation of sensitive information from source data, and less susceptible to linkages due to containing fewer stylistic and semantic cues of individual identities (§6.2.2). Further privacy benefits emerge when a generation strategy can produce divergent data in a *diverse* manner.

While the exact effects are dataset- and application-dependent, intuitively, the ability to generate *diverse*, divergent texts enables us to reduce the distinctiveness of data points as necessary, thus offering an additional layer of protection when sensitive information is leaked or deliberately recovered, for example in singling-out attacks (Cohen and Nissim 2020) where an adversary can exactly match a real record from the output of a data-release mechanism at a success rate above a statistical baseline, or in extraction (Carlini et al. 2021), gradient inversion (Balunović et al. 2022), and membership inference (Mattern et al. 2023) attacks on models training/trained on sensitive data.

As the type and extent to which divergence is favorable are application specific, we focus on assessing the preservation of information, and consider divergence in terms of surface-form dissimilarity as a proxy to both data diversity and privacy preservation.

3.4 Privacy

Personal information can be exposed in UGC datasets via explicit disclosure (Keküllüoğlu, Magdy, and Vaniea 2020) and become memorized by models (Carlini et al. 2021), resulting in privacy infringement at the level of individuals. Moreover, user group memberships such as age and occupation can be revealed via linguistic cues (Rosenthal and McKeown 2011; Preoţiuc-Pietro, Lampos, and Aletras 2015), compromising not only the individual but also groups (Bloustein 1978)—for example, when users are targeted by virtue of their inferred similarity to other user profiles (Floridi 2017). These risks have amplified with technological advancements. Notably, generalist LLMs were found to be capable of accurately inferring varied demographic attributes from social media posts and user-chat interactions without fine-tuning on this task (Staab et al. 2024).

Despite there being well-established privacy frameworks on structured data (e.g., *k*-anonymity; Sweeney 2002), they do not translate well to unstructured data (Lison et al. 2021; van Breugel and van der Schaar 2023). As such, researchers focus on measuring success in proxy tasks, including removing explicit identifiers (Aura, Kuhn, and Roe 2006), stylometrically obfuscating author identity (Juola 2006), and obfuscating user demography through techniques like lexical substitutions (Reddy and Knight 2016) and back-translation (Xu et al. 2019b). Continuing this line of work and complementing privacy enhancing technologies that remove direct identifiers (e.g., PII spans), we examine individual privacy via reduction in user re-identification accuracy, and group privacy via reduction in author profiling accuracy. Compared to PII removal, both are harder to detect and constitute a major privacy challenge for textual data. We use reduction in text overlap (divergence) as an intrinsic proxy for privacy preservation.

Meaning and style preservation exist in tension with divergence and privacy risks. What constitutes acceptable aspect trade-offs depends on application requirements, and an evaluation framework encompassing these aspects can facilitate the selection of appropriate data generation strategies.

4. Evaluation Framework

In this section, we present our synthetic data evaluation framework and its implementation (Figure 1). Our premise is that synthetic language data are generated for use in a task where it is important to preserve different aspects such as style, privacy, and original meaning. Our framework assumes that different strategies can be used to generate the synthetic language data and provides both intrinsic and extrinsic means to evaluate the quality of the synthetic data according to style, meaning preservation,

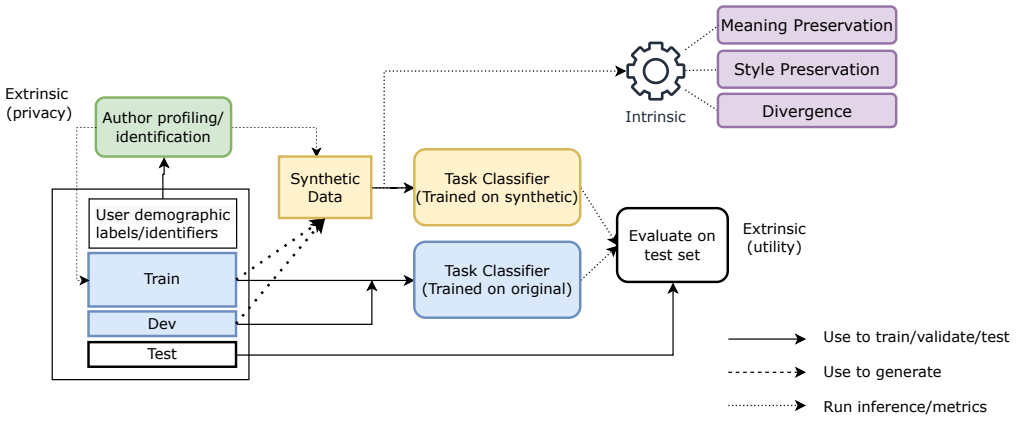

Figure 1
Our evaluation framework in action. We generate synthetic data using different strategies. Intrinsic evaluation is obtained through selected metrics while extrinsic evaluation is performed on downstream tasks. Apart from comparing the performance of classifiers on synthetic vs real data, we run privacy classifiers: attribute classifiers or user matching methods to assess reduction in profiling/re-identification risks.

and divergence as a proxy to privacy. Metrics are provided to intrinsically assess the aspects in the synthetic language data generated. Extrinsically, the data are assessed on downstream task performance and on the degree of author identity and attribute obfuscation. These steps, metrics, and results can aid individuals in selecting appropriate data generation and sharing strategies in various downstream applications.

In the rest of this section, we describe the implementation of our framework, which is extensible to future developments in generation methodologies and metrics. We introduce baseline methods for each generation strategy (§4.1) and intrinsic metrics (§4.2). This ties into the next section (§5), where we describe in detail representative resources and tasks from which we create synthetic language data, which provide realistic test beds for our generation and evaluation strategies.

4.1 Data Generation Strategies

We use strategies that cover different levels of style and privacy preservation, which allows us to examine effects of individual desiderata (§3) and their trade-offs. Explored generation strategies are summarized in Table 1. We do not explicitly control for meaning; all methods (apart from the text editing method StyMask) were trained on pairwise paraphrases, trained to reconstruct input from corrupted representations, or prompted to rewrite the grounding text, thus implicitly optimizing meaning preservation.

We select methods based on their broad applicability, so they can be applied in zero-shot settings (i.e., without pre-existing per-user corpora), do not require demographic labels, and are deployable on consumer hardware to allow local generation from potentially sensitive data on academic budgets. We focus on rewriting and editing based approaches to facilitate privacy experiments that compare predicted user demographic labels/user identifiers with gold standard ones, and therefore require strong alignment between original texts and their synthetic counterparts.

Table 1
Data generation strategies benchmarked in our experiments.

Strategy	Shorthand	Generation Method	Base architecture	Privacy?	Varying style explicitly?
Varying Style	STRAP	Diverse paraphrasing (Krishna, Wieting, and Iyyer 2020)	GPT-2	No	No (implicit)
	DE/CN	Back-Translation (Tang et al. 2021)	mBART	No	No (implicit)
	Syn	Syntactically guided paraphrase (Sun, Ma, and Peng 2021)	BART + retrieval	No	Yes (syntax)
Privacy-oriented Rewriting	StyMask	Rule-based editing (Karadzhov et al. 2017)	—	Yes	No (automatic)
	DP-BART	Differential Privacy: DP-BART (Igamberdiev and Habernal 2023)	BART	Yes	No
LLM Paraphrasing	LLaMA	LLM with instruction prompting (Touvron et al. 2023)	LLaMA-2	No	No (automatic)

Varying Style. Our framework aims to measure style preservation. To this effect we utilize different generation strategies for varying style and measure their effect within the framework. First, we examine the effects of *implicitly* removing stylistic attributes. Diverse paraphrasing and back-translation were found to remove stylistic properties from text (Rabinovich et al. 2017; Krishna, Wieting, and Iyyer 2020). Based on these findings, we use **STRAP**, a paraphraser designed to strip stylistic attributes in Krishna, Wieting, and Iyyer's (2020) style transfer pipeline, to obtain data with low style preservation. For back-translation, we use multilingual BART (Tang et al. 2021) and select German (**DE**) and Chinese (**CN**) as pivots, following Adelani et al.'s (2021) findings that DE preserves more semantics and CN balances between grammaticality and reduction of profiling risks.

We then examine the effects of *explicitly* manipulating syntactic style using syntactically guided paraphrasing. We apply the system developed by Sun, Ma, and Peng (2021), which combines (1) a retrieval step sampling plausible syntactic paraphrases with (2) an encoder-decoder that takes parse templates and source sentences as input for syntactically controlled rewriting. We compare using candidates with the lowest constituency tree edit distance to the original texts (**SynSim**) against those with the highest distance to the original (**SynDiff**) among sampled parses. To inspect the impact of style consistency over time in temporally sensitive tasks, we additionally compare using original parses (**SynOrig**) against using transformed parses that differ in voice (active/passive) and preposition phrase position (front/back), using heuristics developed for rule-based style transfer in Lyu et al. (2021) (**SynTr**).

Privacy-oriented Rewriting. We experiment with generation strategies specifically designed to preserve privacy that reduce identifying qualities from text. First, we apply style masking (**StyMask**), which reduces the distinctiveness of individual documents by editing them towards the corpus-wide stylometric average, chaining rule-based operations such as sentence splitting, lexical substitution, and phrase paraphrasing (Karadzhov et al. 2017). This is flexible as it does not assume pre-existing per-user demographic labels or large personalized corpora, and was the best performing approach in the PAN authorship obfuscation shared task (Potthast, Hagen, and Stein 2016), although it does not provide a mathematical privacy guarantee.

To this end, we implement **DP-BART**, an encoder-decoder method that achieves state-of-the-art results on text rewriting under local differential privacy (Igamberdiev

Table 2
Aspects covered by our evaluation framework for synthetic UGC.

Evaluation	Aspect	Metric	
		Sample-level	Distribution-level
Intrinsic	Meaning Preservation	BERTScore	Fréchet distance
	Style Preservation	POS score	POS JSD
		Style embeddings similarity	Fréchet distance
	Divergence	Self-BLEU	Character JSD
Extrinsic	Utility	Δ downstream task performance	
	Privacy	Δ user identity/attribute classifier performance	

and Habernal 2023). The system removes attributes that distinguish a text written by an individual from other data points by encoding it and dimension-wise clipping its representation, adding noise proportional to a parameter ϵ, then decoding in a standard autoregressive fashion. In particular, we use DP-BART-PR, a variant that involves an additional iterative pruning and training step to reduce the size of the encoder representation, thereby reducing the amount of noise needed under the same privacy budget. This allows us to generate semantic preserving texts under privacy guarantees.[2]

LLM Paraphrasing. We select the 13B instruction-tuned version of LLaMA-2 (Touvron et al. 2023) as an open-access LLM baseline to generate texts with the prompt "Write stylistically diverse paraphrases." By default, the model produces multiple paraphrases for the same input. Similar to batch prompting in classification (Cheng, Kasai, and Yu 2023), this enables time and compute cost savings, but to date it remains unclear whether texts generated in this fashion quantifiably vary within their batch and if so how they affect downstream performance. Based on preliminary observations and the intuition that autoregressive LLMs instructed to write diverse texts should produce increasingly divergent outputs, as a starting point, we always select the first and second output per batch, denoting them as **LLaMA-first** and **LLaMA-second**, respectively.

4.2 Intrinsic Evaluation

We introduce metrics to assess the intrinsic quality of synthetic data in terms of functional aspects (see Table 2). These metrics are selected on the basis of prior work and results from pilot studies, in terms of their discriminative ability in automatic evaluations and also correlation with human annotator judgments (Appendix A.2). Another decisive factor has been deployability within secure compute environments without the need to send gold data to third-party services. We anticipate that, as the field progresses, more or improved metrics can be added to measure the aspects outlined in our framework.

2 Due to the complexity of dynamic DP-composition in multi-length multi-document settings, when generating from arbitrarily long document sequences by the same person (e.g., TalkLife in Section 5.3), we treat each text as an individual data point. In doing so, less noise is added than what is formally required to satisfy (ϵ, δ)-DP. When presenting results, we denote data generated this way as DP-BART*.

In rewriting setups where each synthetic text can be clearly mapped to its source, sample-level metrics are recommended for their interpretability and opportunity for error analysis (van Miltenburg et al. 2021). Otherwise, distribution-level metrics are suitable. We include both types of metrics within our framework.

Meaning Preservation. We select BERT-Score (Zhang et al. 2020), an embedding-based metric that matches token representations from source and synthetic texts using their cosine similarities. To accommodate cases where there is no clear mapping between synthetic texts and individual source instances, we compute BERT Fréchet distance (Xiang et al. 2021) by obtaining text embeddings from original and synthetic texts, fitting them as multivariate Gaussians, then computing the Fréchet distance between the distributions. Although we use BERT here for comparison with prior work, these approaches are extensible to other vectorized semantic representations (e.g., Gao, Yao, and Chen 2021; Su et al. 2023).

Style Preservation. To measure style preservation we select idiolect (i.e., personal linguistic style) embeddings, which involves encoding texts using pooled representations from a RoBERTa model (Liu et al. 2019) trained with contrastive loss on Reddit posts with same/different authors (Zhu and Jurgens 2021). Zhu and Jurgens 2021 found that these embeddings capture lexicosyntactic and orthographic characteristics on short texts, such as punctuation and contraction. At a sample-level, we take the cosine similarity between source and synthetic embeddings. At a distribution-level, we measure the Fréchet distance between the original and synthetic idiolect embedding distributions. As with meaning preservation, these approaches can be used with different embeddings, for example, authorship representations (Rivera-Soto et al. 2021) and instruction fine-tuned embeddings (Su et al. 2023).

As linguistic style is multifaceted, we additionally assess syntactic style via POS trigrams. At a sample-level, we follow prior work in story generation (Roemmele, Gordon, and Swanson 2017) and compute the Jaccard distance between POS trigrams of the original and synthetic texts. At a distribution level, we construct a POS trigram distribution from synthetic data, then measure its Jensen-Shannon divergence (JSD) against the POS trigram distribution constructed from the original data.

Divergence. While the n-gram precision based metric BLEU (Papineni et al. 2002) has well-studied shortcomings as a generation quality metric (Reiter 2018; Freitag et al. 2022), it is suitable for our task of measuring divergence in terms of surface-form dissimilarity. We focus on surface-form since it is more sensitive to verbatim memorization (i.e., useful as a proxy for privacy), and meaning and style divergence can be estimated using the complement of their respective metrics. We follow Niu et al. (2021) and take the BLEU score between a source and synthetic text, measuring divergence per data point as $1 - BLEU(s, t)$. At a distribution level, we compute the JSD between character trigram distributions constructed from the original and synthetic data. In this way, our metric selection incorporates both embedding (meaning/style) and surface-form (divergence) similarities, which is helpful for distinguishing between undesirable overcopying versus successfully preserving meaning/style using varied surface forms.

5. Extrinsic Evaluation: Empirical Experiments on Representative Tasks

Once we use the selected generation strategies (§4.1) to obtain synthetic train and validation sets from source corpora, we evaluate them intrinsically via automatic metrics

Table 3
Representative tasks and datasets in our classification experiments.

Task Type	Task (Dataset)	Dataset Size	Mean # Tokens	Labels	Metrics
Post	Sentiment (Twitter) (Blodgett, Green, and O'Connor 2016)	108,000 tweets	18.15 ± 10.8	Pos, Neg	F1
	Sentiment (Yelp) (Reddy and Knight 2016)	13,391 reviews	7.38 ± 1.9		
Dialogue	Dialogue Act (SwDA) (Jurafsky et al. 1997)	274,786 utts / 1,434 dialogues	13.69 ± 11.5	(42 classes)	SegWER, JointWER
Timeline	Moments of Change (TalkLife) (Tsakalidis et al. 2022)	18,702 posts / 500 timelines	32.43 ± 68.7	O, IE, IS	Coverage

(§4.2) and extrinsically on the basis of representative tasks operating on UGC, shown in Table 3. While our framework itself is task-agnostic, the point of the representative tasks is to serve as exemplars of future applications. We have selected tasks that make use of UGC and cover a variety of classification settings and domains.

We explore a *privacy-oriented* and an *augmentation-oriented* setting (Figure 2):

- *Privacy*: We train and validate models on synthetic data, generating one instance per original example, creating a synthetic dataset that has the same size as its source for each generation method. We test the performance of models trained on synthetic data on real test sets. Additionally, we examine reduction in author profiling/re-identification.
- *Augmentation*: We augment original data with synthetic data, generating one example per original data point. In addition to directly using the original training set (100%), we experiment with under-sampling to simulate data-scarce scenarios, using 50% and 10% of the available original data. We test models trained on synthetic data on real test sets.

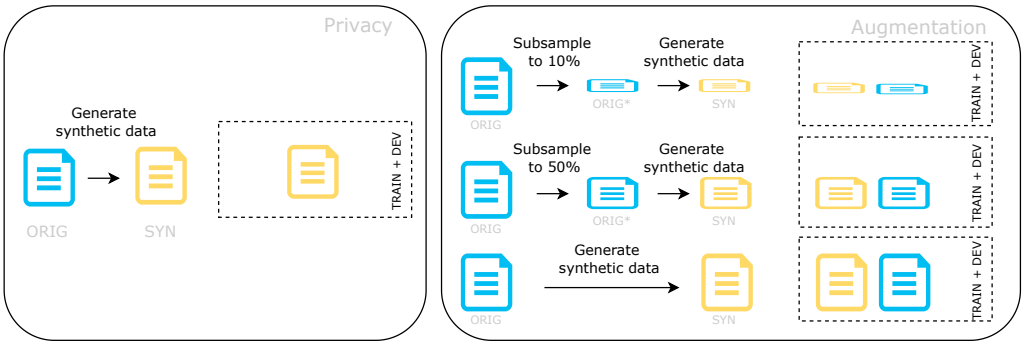

Figure 2
In the privacy-oriented setup (left), we substitute the original data with synthetic data for training and validation. In the augmentation setup (right), we undersample the original data (to 10% or 50% of the original size) or leave them unchanged (100%), generate synthetic data from them, then use the generated data for augmentation. Both setups are assessed on unseen, real test data.

Here, we introduce representative tasks, task metrics, corresponding baselines, and the methods used to measure reduction in author profiling and re-identification risks. For the data augmentation setting, the benefits stemming from the synthetic data are measured on the basis of metrics for downstream classification tasks. This combination of representative tasks, metrics, and baselines sets the foundation for our investigation into the relationship between intrinsic and extrinsic aspects of synthetic data quality in the next section. Tasks are chosen so as to represent different levels of complexity in terms of interaction and temporality, reflecting real-world tasks with UGC data. In particular, we consider the three main categories of post-level tasks, dialogue-based tasks, and timeline-based tasks.

5.1 Post-level Task

Our representative post-level task is sentiment classification, on tweets (DIAL; Blodgett, Green, and O'Connor 2016) and Yelp reviews (Reddy and Knight 2016).

5.1.1 Assessing Utility. A synthetic dataset is considered to have high utility if its downstream model achieves good classification performance on its real test set. We train sentiment classifiers on synthetic tweets generated from DIAL, then assess performance via F1. We follow the same process on Yelp. Specifically, we select DistilBERT (Sanh et al. 2019) for fine-tuning due to its computational efficiency. Implementation details are provided in Appendix A.6.

We find that meaning and style preservation are beneficial when substituting original data with synthetic ones (§6.2.2), but given the relatively low difficulty of sentiment classification of short texts for modern pretrained language models the augmentation benefits are less clear (§6.3).

5.1.2 Assessing Privacy Preservation. As introduced in Section 3.4, privacy risks in textual data emerge not only from PII and other disclosures of sensitive information, but also stem from (inferred) traits (i.e., infringement of *group privacy*; Bloustein 1978; Floridi 2017), subjecting individuals to algorithmic profiling (Büchi et al. 2020).

Here, we examine whether substituting original data with synthetic texts can safeguard group privacy by reducing the linguistic characteristics indicative of the user's demographic group membership. DIAL comprises tweets annotated for author race based on dialect, whereas Yelp contains reviews with gender labels. Following prior work, we separately train a classifier to predict the demographic attribute of users (Xu et al. 2019b; Gröndahl and Asokan 2020; Adelani et al. 2021; Mireshghallah and Berg-Kirkpatrick 2021), using DistilBERT (see details in Appendix A.6). The less successful the author profiling classifier is at inferring demographic traits of the users, the more privacy-preserving the synthetic data. We measure reduction in profiling accuracy.

We observe that diverse syntactic paraphrases, rewriting with differential privacy, and diversely instruction-prompted rewriting are particularly effective at removing demographic characteristics (§6.2.2).

5.2 Conversation-level Task

Our representative conversation modeling task is dialogue act recognition. We use SwDA (Jurafsky et al. 1997), a dataset of transcribed spontaneous conversations based on the Switchboard corpus (Godfrey, Holliman, and McDaniel 1992). The task is to predict where dialogue acts begin and end (i.e., segmentation) and correctly label the

type of each segment (i.e., dialogue act recognition). This task is particularly challenging due to its tagset size (42 labels), class imbalance, and its sequential and contextual nature where an identical utterance can belong to a different class depending on its context.

5.2.1 Assessing Utility. We generate transcripts and assess dialogue segmentation and dialogue act recognition performance over a condensed tagset of 42 acts, following prior work (Quarteroni, Ivanov, and Riccardi 2011; Liu et al. 2017). We fine-tune XLNet (Yang et al. 2019), a transformer architecture that splits long sequences into windows and propagates contexts between them, which was found to be effective for dialogue act modeling (Żelasko, Pappagari, and Dehak 2021). We report Segmentation Word Error Rate (SegWER) and Joint Word Error Rate (JointWER). SegWER is a segmentation-focused metric agnostic to dialogue act labels. It reflects the proportion of functional segments that the classifier fails to predict with perfect boundaries. Contrarily, JointWER is a word count weighted metric that accounts for dialogue act label correctness (Zhao and Kawahara 2018).

We find that higher style and meaning preservation benefit downstream performance in both data substitution (§6.2.2) and augmentation (§6.3) settings.

5.2.2 Assessing Privacy Preservation. Individuals are naturally inclined to self-disclose in conversation (Dunbar, Marriott, and Duncan 1997), opening up opportunities for malicious actors to infer speaker information for pinpoint attacks and user profiling. Whereas privacy experiments at the level of isolated social media posts (§5.1) are well-studied and allow comparisons with prior work using established datasets and models, currently there is a lack of comparable benchmark for privacy in transcribed conversations. Prior work investigated privacy leakages in conversation by re-purposing crowd-sourced persona-grounded dialogue (Xu et al. 2020) and by using chatbots to simulate exchanges between a user and an active personal information-coaxing adversary (Staab et al. 2024); however, crowd-source workers assuming assigned persona and chatbot simulations are not representative of natural, noisy interactions between humans, and neither examines mitigation strategies beyond explicit PII removal. We extend this investigation to conversations *between* human interlocutors.

Inferred Attributes. To this end, we take the speaker metadata in SwDA as gold standard demographic attribute labels and assess user profiling accuracy. We infer (1) gender and (2) education-level, both of which are attributes that have been studied in prior authorship analysis and NLP privacy work (Pennebaker and King 1999; Schler et al. 2006; Reddy and Knight 2016; Staab et al. 2024). We note that the SwDA metadata collected speakers' biological sex rather than gender identity and acknowledge the limitations of conflating these categories. Additional metadata include dialect area and birth year, which we exclude due to the impracticality of inferring dialect from transcripts, and to avoid potential confounding arising from the temporal gap between the corpus's year of release and the present.

Method. Whereas previously user profiling relied on expert classifiers trained to predict a single demographic attribute (Xu et al. 2019b), recent advancements and proliferation of LLMs have enabled effective profiling over a wide range of attributes from unstructured data without training on this task (Staab et al. 2024), representing a new paradigm of possible privacy attacks on user language data. We thus follow and extend Staab et al.'s (2024) approach to natural transcribed conversations between individuals.

First, we use stratified sampling based on gold speaker education levels in the corpus: less than high school, less than college, college, and more than college. For each category, we sample without repetition n conversations, where at least one speaker has the specified education level. We use n of 13 as it is the size of the least frequent category (less than high school). Then, we use a modified version of Staab et al.'s (2024) author profiling prompt template (see Appendix A.5), applying it on MIXTRAL-8×7B (Jiang et al. 2024) to infer education levels of both speakers based on their speech transcripts. We chose the model as it is open-access, instruction-tuned, and can handle context lengths of up to 32k tokens, crucial to modeling long conversations.

We compare the average speaker attribute success rates on gender and on education-level between original and synthetic transcripts. As an additional strict evaluation setting, we look at dialogue-level attack success rate: how often a single privacy infringing prompt can correctly infer the gender and education level of both speakers from a transcript. We use a single instruction prompt to perform zero-shot multi-label multi-speaker inference without further prompt engineering to estimate a lower bound on the inference accuracy, reflecting a setting where bad actors aim to infringe on group privacy at scale rather than performing pinpoint attacks.

We find that even on noisy transcripts LLMs are able to infer speaker attributes a magnitude above chance level, but synthetic data offer a viable mitigation (§6.2.2).

5.3 Timeline-level Task

Modeling document sequences with a temporal dimension is important to applications that involve capturing changes in states, such as identifying moments of change (MoC) in user mood (Tsakalidis et al. 2022), clinical document classification (Ng, Santos, and Rei 2023), and real-time rumor detection (Kochkina et al. 2023).

We select MoC identification as our representative task. Tsakalidis et al.'s (2022) dataset comprises posts from the mental health peer support platform TalkLife.[3] The task requires predicting regions of mood changes on the basis of self-disclosure in a chronological sequence of posts between two dates (i.e., *timeline*) by the same author, classifying each as being in escalation (IE), in switch (IS), or no change (O). While IE and IS both denote changes, IE is a gradual progression and IS is a drastic shift. Contrasting post-level tasks (§5.1) that model texts in isolation and similar to conversation-level tasks (§5.2), timeline modeling is challenging in its sequentiality. Moreover, whereas dialogue act recognition requires modeling neighboring utterances for context, this task requires considering the whole user timeline to assess the individual's baseline mood in order to identify the presence, type, and boundaries of changes.

5.3.1 Assessing Utility. As the focus of this work is on evaluating the quality of generated text, we do not use metadata (e.g., timestamps) in our experiments and only generate synthetic data from the textual content of each post. For downstream modeling, we use the best performing approach in Tsakalidis et al. (2022), first training a post-level BERT with focal loss then feeding embeddings from the post-level model as input to a bi-LSTM. We perform 5-fold cross validation and report timeline-level coverage precision and recall, which are metrics adapted from image segmentation (Arbeláez et al. 2011) that evaluate systems' ability to capture entire regions of interest using the overlap between true and predicted sequences.

3 https://www.talklife.com/.

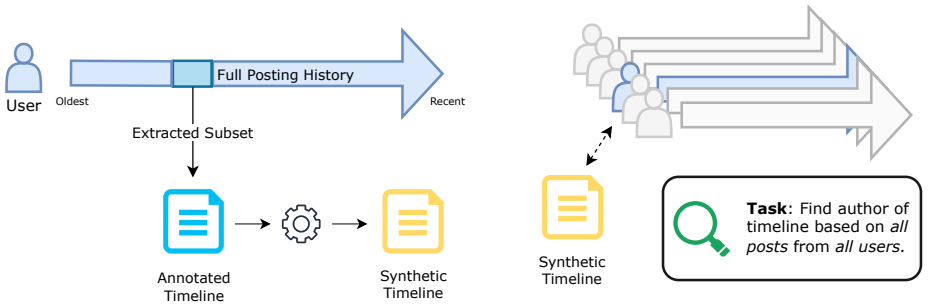

Figure 3
We assess preservation of individual privacy in our timeline-level task by examining whether synthetic data can be maliciously linked to their original user.

We find that overall meaning and style preservation contribute to better downstream classification in both the data substitution (§6.2.2) and augmentation (§6.3) settings, with stylistic consistency generally benefiting recall, and stylistic similarity benefiting performance in data substitution but less so in data augmentation.

5.3.2 Assessing Privacy Preservation. Timelines pose unique privacy challenges given that each user is associated with large quantities of posts rich in sensitive information. Moreover, online platforms are dynamic; after a snapshot of the data is shared with researchers, users will continue to generate new posts. In other words, a realistic assessment of user re-identification risks needs to account for malicious actors using data not only from the subset of data from which synthetic data were generated, but also users' digital traces left before and after the subset of data used for synthetic data generation.

Method. In post-level (§5.1) and conversation-level (§5.2) tasks, we assessed group privacy via reduction in author profiling, using gold standard demographic labels. Since such labels are not available here, we instead assess *individual privacy* via reduction in authorship attribution (see Figure 3). The original dataset comprises 500 timelines (17.8k annotated posts). Our task is to attribute synthetic timelines to their original users while matching each synthetic post against the *full posting histories* of 37.9k users.[4]

Assuming a scenario where a malicious actor aims to reveal the identities of as many users as possible under time and/or compute constraints, methods that require pairwise comparisons over all document combinations are ill-suited given the size of online communities.[5] An efficient alternative is to *estimate* post similarity.

We measure re-identification risks from this angle of attack using the MinHash algorithm (Broder 1997), a technique that has been applied to data mining and dataset deduplication at scale (Gao et al. 2021; Rae et al. 2021; Lee et al. 2022). The method begins with tokenization. Given that prior work found character *n*-grams to be effective author attribution features (Peng et al. 2003) that encapsulate morphological and stylistic characteristics (Koppel, Schler, and Argamon 2011; Sapkota et al. 2015), we choose

4 These are individuals who have posted at least 50 times from August 2011 to August 2020 (8.7m posts).
5 In early experiments we tried stylometric user linking (Weerasinghe, Singh, and Greenstadt 2022). We found it verified authorship well at post-level but did not scale once applied to user histories in our dataset, hence we focus on approximate matching.

to use character-level n-grams with a small n ($n = 3$). We apply a hash function on each n-gram, then permute the hashes k times, keeping the k minimum hash values in each set. To efficiently find similar texts, we use locality sensitive hashing, which divides the array representing each document into smaller bands that are grouped into clusters of texts estimated to have high n-gram overlaps. For each synthetic timeline, we attempt to identify the original author by counting the number of generated posts that are considered near duplicates to the original posts and selecting the author with the most matches. The re-identification risk of a data generation method is thus measured by the proportion of correctly matched authors.

Although this setup assumes the adversary has access to the full histories of all users and therefore does not accurately reflect the restricted access typical in real-world scenarios, it can offer insights on potential re-identification vulnerabilities in synthetic longitudinal texts. We find that except data from methods incorporating differential privacy, data from all generation strategies are susceptible to re-identification (§6.2.2).

6. Results

Through our framework, we now have a quantifiable means to compare data generation approaches based both on intrinsic aspects and utility in downstream applications. In this section, we present evaluation results on data generated using the described generation strategies, tasks, and models. We discuss the interplay of intrinsic (Table 4) and extrinsic assessments in privacy-oriented (Table 5) and augmentation (Tables 7, 8, 9) settings, which inform our recommendations in Section 7.

6.1 Intrinsic Results

We show examples of synthetic data generated from publicly available datasets in Appendix A.3. Intrinsic evaluation results are shown in Table 4. We focus on sample-level metrics since they are more suited to the examined rewriting-based strategies, and we include results for distribution-level metrics in Appendix A.4.

Table 4
Intrinsic evaluation results for the aspects of meaning preservation (BERTScore), style preservation (Idiolect embedding similarity, POS score), and divergence (1 - self-BLEU). Higher is better for all metrics. In **bold**: top scoring methods for each metric. Underlined: top scoring method among syntax controlled paraphrase baselines.

		Post-level								Dialogue-level				Timeline-level			
		Twitter				Yelp				SwDA				TalkLife			
	Generation strategy	BSc	Idio	POS	Div	BSc	Idio	POS	Div	BSc	Idio	POS	Div	BSc	Idio	POS	Div
Style	STRAP	.87	.65	.66	.88	.92	.76	.79	.73	.91	.72	.42	.79	.90	.69	.71	.84
	DE	**.93**	**.83**	**.84**	.54	**.95**	.72	**.96**	.40	**.96**	**.85**	**.77**	.45	**.95**	**.83**	**.88**	.51
	CN	**.89**	**.69**	**.74**	.75	.92	.53	.84	.73	.89	.61	.21	.76	.92	.68	.78	.75
	SynSim	.85	.57	.67	.91	.90	**.82**	.69	.85	.85	.58	.18	**.82**	.87	.61	.68	.90
	SynDiff	.85	.57	.67	.91	.91	.80	.68	.87	.85	.57	.14	.80	.87	.60	.67	.90
	SynOrig	–	–	–	–	–	–	–	–	–	–	–	–	.88	.59	.69	.89
	SynTr	–	–	–	–	–	–	–	–	–	–	–	–	.88	.58	.69	.89
Privacy	StyMask	.85	.58	.67	.18	**.93**	**.91**	**.97**	.14	**.94**	**.86**	**.62**	.16	**.97**	**.86**	**.94**	.22
	DP-BART ($\epsilon = 50$)	.80	.50	.48	**.98**	.82	.39	.46	**.97**	.80	.41	.06	**.98**	.81	.50	.52	**.98**
	DP-BART ($\epsilon = 100$)	.81	.53	.51	**.98**	.83	.42	.51	.96	–	–	–	–	–	–	–	–
	DP-BART ($\epsilon = 250$)	.84	.66	.62	.93	.87	.56	.68	.88	–	–	–	–	–	–	–	–
LLM	LLaMA-first	.87	.66	.60	.96	.89	.57	.68	.97	.85	.54	.60	.95	.84	.52	.55	.96
	LLaMA-second	.84	.63	.57	.97	.88	.57	.63	**.98**	.85	.53	.58	**.96**	.83	.52	.55	**.96**

Table 5
Task performance of benchmarked generation strategies in the 1-to-1 rewriting setting. Results are averaged over five runs. Higher is better except for SegWER and JointWER. In **bold**: top two best performing models trained on synthetic data; underlined: outperforms best original.

	Task	Post-level		Dialogue-level		Timeline-level			
		Twitter	Yelp	SwDA		TalkLife			
						IE		IS	
	Metric	Macro F1	Macro F1	SegWER	JointWER	C_p	C_r	C_p	C_r
Generation strategy		↑	↑	↓	↓	↑	↑	↑	↑
Style	STRAP	.73	.97	.26	.48	**.37**	.26	.32	.11
	DE	.73	**.98**	**.15**	**.39**	.36	**.33**	**.36**	.15
	CN	.74	.97	.35	.54	.35	.26	.35	**.18**
	SynSim	.75	.96	.35	.55	**.37**	.24	**.36**	.12
	SynDiff	.75	.97	.38	.60	.33	.34	.29	.08
	SynOrig	–	–	–	–	.34	.28	.28	.12
	SynTr	–	–	–	–	.35	.30	.27	**.18**
Privacy	StyMask	**.76**	**.98**	**.14**	**.41**	.36	**.34**	.35	.16
	DP-BART (= 50)	.72	.86	.72	.91	.20	.04	.18	.01
	DP-BART (= 100)	.73	.93	–	–	–	–	–	–
	DP-BART (= 250)	**.76**	.96	–	–	–	–	–	–
LLM	LLaMA-first	.71	.96	.51	.69	.31	.30	.25	.09
	LLaMA-second	.68	.95	.53	.73	.30	.31	.33	.06
Best system trained on original data		.77	.98	.12	.36	.35	.34	.33	.19

Roundtrip Translation/Paraphrasing. Back-translating from German (DE) is more meaning- and style-preserving compared with Chinese (CN), in line with linguistic similarities between English and the pivot languages. Among the methods prior work has identified to remove stylistic properties, CN and STRAP do so to greater extents than DE, as reflected in their lower idiolect and POS scores.

Syntax-guided Paraphrasing. Style metrics indicate that syntactic paraphrases varied per our expectations: Texts generated with unaltered parses (SynOrig) score higher on POS scores than those generated from transformed ones (SynTr), and texts generated using parses syntactically similar to the original text (SynSim) are more syntactically style preserving than those generated using syntactically distant ones (SynDiff).

Privacy-oriented Rewriting. StyMask produced the most meaning and style preserving data across domains. This can be explained by the fact that it relies on rule-based local editing operations, which enables downstream classifiers to perform well across tasks, but at the expense of higher privacy risks in the examined settings (§6.2.2). Finally, as expected, the stricter the privacy budget set in DP-BART, the higher the divergence.

LLM. For LLaMA outputs batch-generated by prompting for diverse paraphrasing, we see that the first in each batch is consistently more meaning- and style-preserving than the second, supporting our prior speculation that instruction prompting an autoregressive LLM for diversity would produce increasingly divergent paraphrases. We explore the practical implications of this behavior later in Section 6.2.1.

As a whole, we observe trends consistent with expectations across domains. Intrinsically, data with high meaning preservation (e.g., DE, see Table 4) also tend to be style-preserving, and data that score higher on preservation-oriented metrics (e.g., StyMask) are lower in surface form divergence. However, as we discuss below, extrinsic evaluation reveals that each aspect does not uniformly enhance privacy or utility.

6.2 Extrinsic Results: Privacy-oriented

In Table 5, we summarize the performance of downstream models when *substituting* real training data with synthetic ones on a 1-to-1 basis.

The best performing model trained on synthetic data in each task performed comparably to those trained on original data. In particular, synthetic data showed good performance across the board in the most straightforward task of post-level binary classification. The synthetic data that yielded the worst classifiers in terms of F1 were the most privacy-preserving ones, as will be discussed in the next section. Notably, in binary classification of short texts, DP-BART reduces author profiling accuracy to almost chance level (Table 6) at the expense of only <5% in F1 (Table 5). However, performance gaps between classifiers trained on original vs synthetic data widen in the other domains, which are more challenging due to multi-document context, number of classes involved, and the nuanced nature of the tasks.

Naive data generation may amplify biases. When trained on real data, classifiers tend to under-predict rare classes. Taking longitudinal mood prediction as an example, classifiers from real data tend to under-predict IS (which comprises only 5% of labels), resulting in low coverage recall (C_r). As shown in Table 5, most synthetic data classifiers magnify this gap, exemplified by low C_r across the board. In addition to performance concerns, this under-prediction problem reflects bias amplification risks in synthetic data applications (Zhao et al. 2017; Wang and Russakovsky 2021; Wyllie, Shumailov, and Papernot 2024). As will be discussed later, in the case of longitudinal UGC modeling, we find that training on stylistically diverse synthetic data can alleviate this problem.

Table 6
Author profiling accuracy (↓) and proportion of re-identified users (↓) across methods. In **bold**: most privacy preserving overall; underlined: most privacy preserving without differential privacy guarantees.

	Twitter	Yelp	SwDA			TalkLife
Generation strategy	Race	Gender	Gender	Education	Both	User ID
DP-BART ($\epsilon = 50$)	**.51**	**.51**	.34	**.22**	.03	**.00**
DP-BART ($\epsilon = 100$)	.53	.54		–		–
DP-BART ($\epsilon = 250$)	.64	.71		–		–
SynSim	.65	.67	.39	.34	.07	.35
SynDiff	.67	.63	.40	.26	.03	.31
SynOrig	–	–		–		.32
SynTr	–	–		–		.32
STRAP	.65	.74	.35	.25	**.00**	.45
CN	.68	.74	.41	.27	.04	.73
DE	.80	.84	.39	.33	.07	.97
StyMask	.66	.91	.42	.30	.08	.99
LLaMA-first	.54	.62	.31	.32	**.00**	.05
LLaMA-second	.57	.60	**.27**	.29	.04	.04
Random Baseline	.50	.50	.50	.25	.02	1/37.9k
Original Data	.88	.91	.43	.35	.16	1.00

More generally, class distribution-aware augmentation (He et al. 2008; Ahn, Ko, and Yun 2023) with aspect-controlled methods is a potential direction for future work.

6.2.1 Implications of Style Preservation in Training Data Substitution.

Style preservation benefits task performance. We find that style preservation benefits performance on most tasks and datasets, except for binary sentiment classification on Yelp (Table 5). The least performant classifiers were indeed trained on data with low style preservation, but the remaining classifiers' system-level ranking in terms of task performance does not match that of their styles, for example. Models trained on varied syntactic paraphrases performed similarly well. This suggests that once a baseline level of data quality has been reached and the task performance is around the level of that trained on real data, there is a ceiling to style preservation's benefits to performance.

In contrast, on Twitter and SwDA, there is a relatively straightforward association between preservation-oriented metrics and performance. For example, when using syntactically controlled methods, the generated instances most syntactically similar to the original yield better performance compared with their alternatives in order of stylistic similarity, suggesting that style plays a role in task performance. Likewise, the first of each batch of diverse LLaMA paraphrases tended to be more style preserving than the second (§6.1) and also trained more performant downstream models.

Similarly, in longitudinal predictions in TalkLife, overall results show that more meaning and style-preserving generation strategies tend to perform better. However, style and meaning preservation do not constitute the whole picture, as detailed below.

Stylistic similarity and stylistic consistency benefit performance differently. Results from syntax guided paraphrasing, particularly on the task of longitudinal predictions (see Table 3), suggest that *the degree and type of style preservation impact data utility*. First, we compare paraphrases generated with syntactically similar sampled parses (SynSim) and different ones (SynDiff). We see that SynSim outperforms SynDiff on all classes and metrics except IE recall, which relies on capturing sequences of posts that depend on each other to denote a gradual change; thus while style similarity benefits task performance, the downstream classifier may under-predict temporally sensitive classes, such as gradual mood changes, if only exposed to stylistically similar data.

Second, we compare paraphrases generated with original parse templates (SynOrig) and those generated with heuristically modified ones that introduce syntactic variations in a consistent manner to simulate a stable user-level style throughout the timeline (SynTr). SynTr leads to better performance on all classes and metrics compared with SynOrig, with the exception of precision on the rarest label of sudden mood changes (IS); thus artificially inducing stylistic variety in synthetic datasets may be beneficial but can increase false positives, in this case causing the downstream classifier to become over-sensitive to within-user changes in linguistic content.

Interestingly, SynDiff has high IE recall but low IS recall, whereas SynTr shows no such trade-off. An interpretation is that although both involve adding syntactic changes, SynDiff is subject to sampling variability, whereas SynTr's rule-based transforms introduce stylistic diversity in a consistent fashion, enabling the downstream classifier to reap the benefits of increased exposure to syntactic variations while maintaining sensitivity to stylistic consistency, which assists personalized longitudinal modeling.

In short, across labels in the longitudinal personalized task, diverse syntactically similar paraphrases train more precise classifiers, and paraphrases with consistent stylistic variations benefit recall. *While style preservation benefits performance, introducing*

stylistic variations can help classifiers generalize and pick up more linguistic cues useful in user state tracking so long as the introduced changes are consistent. We make recommendations to this effect in Section 7.

6.2.2 Divergence Benefits Lowering Profiling and Re-identification Risks. Benchmarked generation strategies are variedly successful in removing cues of user demographic attributes and lowering re-identification rates (Table 6), with low-divergence data most susceptible to such risks. A notable example is StyMask, which has low divergence across domains and high profiling accuracy on Yelp yet low on Twitter. This suggests that heuristics designed for authorship obfuscation are specific and translate well to masking race but not gender. Such inequality of privacy protection over user attribute types underscores the need to develop and test obfuscation methods across user groups.

Conversation-level. In the novel challenging task of zero-shot speaker attribute inference from SwDA human conversation transcripts, we see that synthetic data successfully reduced speaker profiling accuracy for both gender and education, although in education prediction only DP-BART reduced the success rate to lower than chance level. In the strictest setting where the task is to infer both attributes for both speakers from a single prompt, we see that on the original data the LLM achieves an attack success rate close to a magnitude above the random baseline. Although results indicate that synthetic data have a mitigating effect, the fact that a readily available model can achieve this level of speaker profiling accuracy in a zero-shot manner, without fine-tuning or elaborate prompting setups, highlights the need to develop defense strategies against this new line of privacy infringement risks, corroborating findings from Staab et al. (2024).

Timeline-level. In terms of finding individual users from synthetic data in TalkLife timelines, except DP-BART, data from all methods are susceptible to re-identification above chance level (i.e., 1 in 37,969 users), although attack success rates were remarkably low on texts rewritten using LLaMA, potentially due to their high divergence. Overall, these results indicate that while rewriting can mitigate risks of revealing sensitive attributes, *preventing exact re-identification in post sequences remains an open problem*. It should be noted, however, that this setup assumes the adversary has unfettered access to user posting histories and should therefore be considered a worst-case scenario benchmark to facilitate development of better privacy-preserving data generation strategies.

6.3 Extrinsic Results: Augmentation

We show the downstream classification results for post-level (Table 7), conversation-level (Table 8), and timeline-level (Table 9) tasks. Overall, performance improved as data quantity increased. In terms of performance gains from augmentation using synthetic data, similar to what we observed in the privacy-oriented (i.e., 1-to-1 data substitution) experiments, we find that the benefits varied by task type and difficulty:

- *Post-level*: In Twitter/Yelp, both relatively straightforward tasks of post-level binary sentiment classification, we see modest to no improvements. In Yelp, classifiers without augmentation trained on 10% of the original dataset already achieved a macro F1 of .95, leaving little room for further performance gains.

Table 7
Post-level sentiment classification performance in augmentation settings, comparing augmentation at the original (100%) and sub-sampled (50%, 10%) training set sizes. Higher is better (↑). Underlined: outperforms no-augmentation.

		Twitter			Yelp		
		Macro F1			Macro F1		
	Generation strategy	10	50	100	10	50	100
Style	STRAP	.71	.74	.76	.94	.97	.97
	DE	.72	.75	.78	.95	.95	.97
	CN	.72	.73	.77	.94	.96	.97
	SynSim	.73	.76	.78	.93	.96	.96
	SynDiff	.72	.76	.78	.94	.96	.96
Privacy	StyMask	.70	.75	.76	.94	.96	.96
	DP-BART ($\epsilon = 50$)	.71	.74	.77	.94	.96	.97
	DP-BART ($\epsilon = 100$)	.71	.75	.78	.94	.96	.97
	DP-BART ($\epsilon = 250$)	.71	.74	.77	.94	.95	.97
LLM	LLaMA-first	.73	.74	.77	.95	.96	.97
	LLaMA-second	.72	.74	.76	.92	.96	.97
	Original (without augmentation)	.72	.74	.77	.95	.95	.98

Table 8
Dialogue-level performance in augmentation settings, comparing augmentation at under-sampled (10%, 50%) and original (100%) training set sizes. Lower is better (↓). Underlined: outperforms no-augmentation.

		SwDA					
		SegWER			JointWER		
	Generation strategy	10	50	100	10	50	100
Style	STRAP	.15	.13	.13	.41	.37	.36
	DE	.15	.13	.12	.40	.35	.35
	CN	.14	.13	.11	.42	.37	.35
	SynSim	.15	.12	.11	.41	.35	.33
	SynDiff	.14	.14	.12	.40	.37	.36
Priv.	StyMask	.14	.13	.12	.40	.36	.35
	DP-BART ($\epsilon = 50$)	.14	.12	.12	.40	.36	.35
LLM	LLaMA-first	.15	.13	.12	.40	.36	.35
	LLaMA-second	.14	.12	.12	.40	.36	.35
	Original (without augmentation)	.14	.13	.12	.40	.36	.36

- *Dialogue-level*: In SwDA, while magnitudes of improvements were modest, we see that augmentation leads to more improvements when quantity increased and had more pronounced benefits for JointWER, which unlike the segmentation-only SegWER considers both segment boundary and segment label correctness. This suggests that exposure to synthetic data helped downstream classifiers correctly learn to distinguish between dialogue act types.
- *Timeline-level*: In TalkLife, in the most data-scarce setting (10%), augmenting with most synthetic data generation strategies led to a

Table 9
Timeline-level task performance in augmentation settings, comparing augmentation at the original (100%) and sub-sampled (50%, 10%) training set sizes. Higher is better (↑). Underlined: outperforms no-augmentation.

		TalkLife											
		IE (C_p)			IE (C_r)			IS (C_p)			IS (C_r)		
	Generation strategy	10	50	100	10	50	100	10	50	100	10	50	100
Style	STRAP	.16	.31	.40	.28	.38	.35	.03	.16	.18	.35	.44	.45
	DE	.17	.31	.37	.29	.37	.38	.04	.14	.23	.26	.45	.40
	CN	.11	.27	.33	.27	.37	.39	.03	.14	.21	.21	.55	.47
	SynSim	.18	.33	.32	.29	.35	.41	.01	.07	.20	.30	.40	.43
	SynDiff	.21	.36	.32	.30	.38	.43	.02	.13	.15	.21	.52	.44
	SynOrig	.18	.26	.38	.30	.36	.38	.01	.09	.14	.36	.49	.46
	SynTr	.16	.23	.34	.23	.41	.43	.01	.12	.18	.35	.51	.45
Priv.	StyMask	.16	.31	.40	.28	.38	.35	.03	.16	.18	.35	.44	.45
	DP-BART*($\epsilon = 50$)	.09	.31	.32	.25	.38	.41	.01	.09	.18	.20	.56	.46
LLM	LLaMA-first	.17	.30	.39	.31	.39	.40	.01	.11	.18	.38	.47	.47
	LLaMA-second	.22	.31	.33	.28	.36	.40	.01	.14	.20	.22	.37	.52
	Original (without augmentation)	.20	.41	.35	.28	.29	.34	.01	.26	.33	.03	.29	.20

marked improvement in precision and especially recall, which saw an increase of up to .35. We find that augmentation leads to more frequent predictions of moments of change on real test data, resulting in improved recall but often degraded precision. As in the privacy-oriented classification experiments, we observe the strongest negative effects on the rarest class IS, whereas the rare but more prevalent class IE did see benefits in precision and recall on most models trained on synthetic data generated using the benchmarked strategies, with up to a 26% difference.

In short, on well-defined tasks already adequately modeled using original data under the examined settings, adding synthetic data will do little to raise the performance ceiling. As task complexity increases so do potential augmentation benefits, although risks of performance degradation on the rarest label class persist, once again underscoring the importance of considering class distribution when using synthetic data in practice.

6.3.1 Implications of Style Preservation in Augmentation. We examine the role of style by focusing on syntactically controlled paraphrases. On the examined post-level tasks, there are limited noticeable effects as performance tended to be relatively high and similar across the board. On the dialogue-level task SwDA, as data quantity increases, the benefits of augmenting with syntactically similar data (SynSim) becomes more pronounced than by augmenting with syntactically distant data (SynDiff). While overall magnitudes of improvements are modest, SynSim augmentation is shown to improve both utterance segmentation (SegWER) and dialogue act recognition (JointWER).

The role of style is more complex in the timeline-level task, TalkLife. Mirroring findings from data substitution experiments (§6.2.1), when data is most limited (10%), augmentation with syntactically similar data (SynSim) benefits recall of the minority class IS more so than with syntactically different ones (SynDiff). However, contrasting the consistent advantages of SynSim in data substitution, here we find that SynDiff yields better performance when it comes to identifying gradual progressions (IE) and overall performance when more data is available. Moreover, contrasting the clear advantages of using consistently syntactically transformed paraphrases (SynTr) over

unaltered ones (SynOrig) in earlier experiments, here SynOrig yields more precise IE while SynTr yields more precise IS; thus augmenting with syntactically consistent and similar data benefits capturing gradual progressions, whereas data that is syntactically *varied in a consistent manner* benefits capturing the minority class, namely, mood switches. Finally, in line with earlier findings, data with stylistic variation introduced consistently (SynTr) is better than those varying stochastically (SynDiff) for IS recall. Combined with the above, this suggests that exposing the downstream classifier to stylistically varied data help them capture changes in timeline-level tasks, and doing so while preserving style consistency within each timeline further benefits recall of rare, drastic changes.

All in all, that the way in which style contributes to performance varies from what we observed in data substitution experiments underscores the importance of tailoring generation strategies to application requirements, for instance, data availability or whether it is important to prioritize privacy over performance.

7. Implications for Future Applications

What do our findings mean in practice, for the application of our evaluation framework by others? Figure 4 summarizes recommendations based on the empirical experiments presented in Section 6. To guide our discussion, consider the use cases below:

Augmenting Public Datasets in a General Domain. Assuming that the task is challenging and has room for improvements, highly meaning- (e.g., DE) and style-preserving (e.g., SynSim) methods are reasonable choices due to their favorable performance across examined tasks. If the task involves longitudinal modeling, consider increasing stylistic variations to improve generalizability (e.g., SynDiff), and pay attention to the within-user consistency to better capture minority labels (e.g., SynTr).

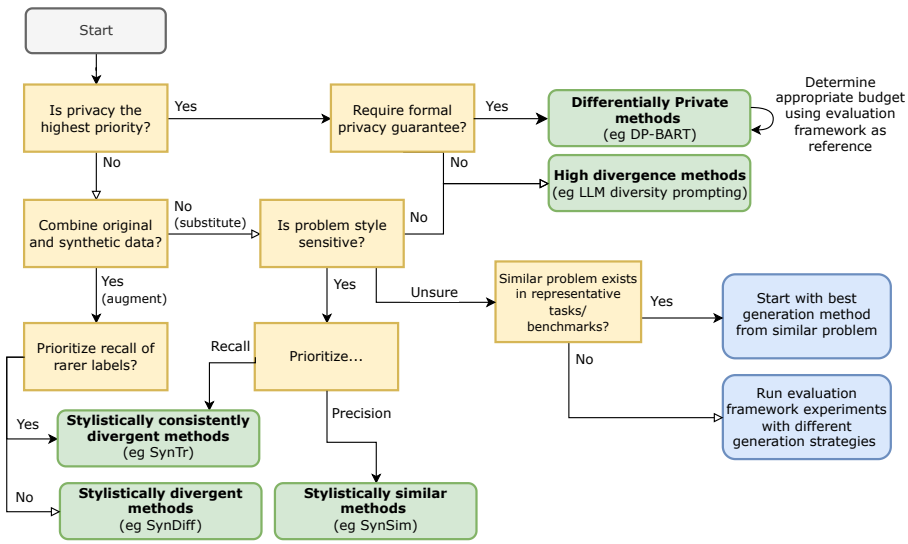

Figure 4
Recommended synthetic UGC generation method selection based on our results and application requirements.

Sharing Models Trained on UGC for Health Applications. For sensitive data, methods with differential privacy (e.g., DP-BART) are the best choice; in addition to mathematical guarantees, experiments in Section 6.2.2 showed its ability to mitigate profiling and re-identification risks. The privacy budget ϵ needs to be calibrated with caution since a strict budget impedes meaning and style preservation (§6.1, Table 4), which can cascade to lower label validity and reduce performance especially in complex tasks. If formal privacy guarantees are not required, other high divergence methods (e.g., LLM-generated diverse paraphrases) are good candidates for privacy preservation. If the task is temporally sensitive (e.g., user modeling over document streams), style consistency seems to benefit task performance, with similar paraphrases (SynSim) benefiting precision and consistent variations benefiting recall (SynTr).

8. Conclusion

We present an evaluation framework for synthetic language data, which defines core aspects to assess generated texts accompanied by suitable metrics. Through use cases that benchmark style and privacy by varying text rewriting strategies, we demonstrate that our proposed framework captures the intended qualities of the texts and identify downstream utility and privacy implications of the aspects and generation strategies.

We find that classifiers trained only on synthetic data can achieve task performance comparable to real ones, especially in straightforward tasks with a single-document context. However, as task complexity increases so do potential performance gaps and bias amplification risks. Additionally, while meaning and style preserving synthetic datasets tend to yield more performant classifiers, meaning is not the whole picture: Style preservation and consistency impact performance differently in longitudinal tasks, exercising varied effects depending on whether synthetic texts are used to substitute or augment original training data. These challenges point to targeted aspect-guided data generation as an area of future research.

Moreover, users are more susceptible to demographic profiling and re-identification in low-divergence synthetic datasets, and most methods failed to effectively prevent re-identification in the longitudinal setting, highlighting the need to progress beyond standard static setups when developing privacy-preserving generation methods.

In addition to ready-to-use evaluation metric code and experiment scripts accompanying this publication, we plan to provide a user-friendly platform that includes this framework for individuals to evaluate their data, compare methods, and extend the framework to measure aspects of data quality using new metrics and on new tasks. We also plan to extend our evaluation and generation methods for preserving meaning, style and privacy for temporally sensitive long documents that allow for benign divergences. We envisage that our findings and tools can enable researchers and practitioners to better and more efficiently select suitable generation and data sharing strategies.

A. Appendices

A.1 Limitations

The presented user profiling and re-identifiability measures are bounded by the methods' performance and validity. Profiling, re-identifiability, and divergence estimates do not constitute formal privacy guarantees, but are instead part of a continued effort towards creating shareable, more privacy preserving textual data. Relatedly, we emphasize rewriting-based methods to compare with prior work. Such methods generally produce data closer to the original and therefore provide more conservative estimates of privacy risks. Additional strategies include generation grounded in user facts and few-shot prompting. We will apply our proposed framework to them in future work.

Language use is contextual and an individual may express themselves differently under varied circumstances, but current metrics assume style to be stable. Methods for generating synthetic text can be misused to impersonate individuals and spread misinformation. We also observed potential bias amplification when training models with synthetic data. Nevertheless, improved style-sensitive evaluation can help develop targeted augmentation strategies, which in turn can help counteract biases. As generative models become increasingly widely adopted, it is our hope that through underscoring model capabilities and providing tools to assess synthetic texts, our work will contribute to improving more nuanced data generation, evaluation, and application.

A.2 Metric Validation Pilots

While our framework is metric-agnostic, we narrowed down candidate metrics through pilot studies with three annotators who are native speakers of English with prior experience in annotation for NLP/NLG tasks. In the end, we selected metrics to include in this article on the basis of their performance on automatic validations and their rank correlations with human judgments.

- *Automatic:* For **style**, we applied STEL (Wegmann and Nguyen 2021) to author style, constructing 1k task instances each over domains: tweets (Schler et al. 2006), blogs (Pardo et al. 2017), and speech transcriptions with and without disfluencies (Godfrey, Holliman, and McDaniel 1992; Love et al. 2017). We compared hand-engineered features (Abbasi and Chen 2008; Altakrori, Cheung, and Fung 2021), POS-based scores, and idiolect embeddings. For **divergence** in conversations, we used dialog similarity judgments from Lavi et al. (2021) to benchmark embedding-based edit distance against document-level BLEU, TER (Snover et al. 2006), and chrF (Popović 2015). We also try using BLEU, TER, and chrF as edit distance substitution cost.
- *Human*: In addition to the above aspects and metrics, we included BERTScore, BLEURT (Sellam, Das, and Parikh 2020), and QuestEval (Scialom et al. 2021) to evaluate meaning preservation. We randomly sampled 30 instances from Twitter, Reddit, and Switchboard and generated synthetic versions of the texts. For each intrinsic aspect, three evaluators (native English speakers with prior NLG evaluation experience) performed Best-Worst Scaling (Louviere and Woodworth 1991), selecting the best and worst generations per source text.

A.3 Example Outputs

Table 10
Example Twitter posts generated from strategies described in Section 4.1.

original	@user cuz I see y'all drawn
STRAP	I see you all drawn to me.
DE	@User Cause I see you all knitted
CN	Because I saw you all draw
SynSim	i want to see you drawn
SynDiff	did i see you drawn ?
StyMask	@user and cuz I see y'all drawn, yet
DP-BART($\epsilon = 50$)	This article is over 2 years old\n\nThis is a guest post by the author
DP-BART($\epsilon = 100$)	I've been reading a lot of posts lately about the Trump administration. I've been
DP-BART($\epsilon = 250$)	@ cetera.ca / y/a I
LLaMA-first	*grinning from ear to ear* because I see you all smitten
LLaMA-second	Oh my gosh, I am absolutely tickled pink because I see you all having a blast!

Table 11
Example Yelp reviews generated from strategies described in Section 4.1.

original	i ordered a wild boar burger instead .
STRAP	instead, I've ordered a wild pig burger.
DE	I ordered a wild boar burger instead.
CN	I called a wild pig hamburger.
SynSim	i opted for the wild boar burger .
SynDiff	instead of wild boar
StyMask	i ordered the wild boar burger, instead .
DP-BART($\epsilon = 50$)	I don't know what I'm going to do with
DP-BART($\epsilon = 100$)	A boisterous boisterously conspiratorial conspiracy to hijack the world'
DP-BART($\epsilon = 250$)	A wild boar boar raided a Whole Foods market Thursday. Then he chopped down
LLaMA-first	I opted for the mouth-watering wild boar burger.
LLaMA-second	My taste buds were crying out for the savory wild boar burger.

Table 12
Example SwDA utterances generated from strategies described in Section 4.1.

original	I've been down to a dealer and driven them.
STRAP	I was driving a dealer.
DE	I have been to a dealer and drove them.
CN	I went to a dealer and took them away.
SynSim	i was at the dealer 's and i drove them .
SynDiff	i went to the dealer and i drove them . . .
StyMask	myself 've been down to merchant also driven them;
DP-BART*($\epsilon = 50$)	I'm not sure why I'd like to say this. I
LLaMA-first	I went to a car dealership and took some models for a test drive.
LLaMA-second	I ventured to the local car lot and got behind the wheel of a few cars.

A.4 Distribution-level Metrics

Table 13 shows the distribution-level metrics identified in Section 4.2.

As shown in Figure 5, the relationship between examined sample-level (i.e., pairwise) and distribution-level metrics vary across task datasets. Future work looking to select or improve metrics should take into account domain differences as well as potential differences between evaluating at sample and distribution levels.

Table 13
Intrinsic evaluation results computed using distribution-level metrics for the aspects of meaning preservation (BERT Fréchet Distance), style preservation (idiolect embedding Fréchet distance, POS JSD), and divergence (character trigram JSD). Lower is better for all metrics except divergence. In **bold**: top two scoring methods for each metric; underlined: top scoring method among syntax controlled paraphrase baselines.

	Post-level								Dialogue-level				Timeline-level			
	Twitter				Yelp				SwDA				TalkLife			
Generation strategy	BFD	IFD	POS	Div	BFD	IFD	POS	Div	BFD	IFD	POS	Div	BFD	IFD	POS	Div
STRAP	.81	2.3	.14	.12	1.7	2.8	.09	.10	.29	.19	**.15**	.11	.30	1.6	.10	.06
DE	**.11**	**.74**	.09	.06	1.8	4.2	**.01**	.08	**.10**	1.6	.16	.06	**.04**	.62	**.04**	.03
CN	.55	2.0	.13	.08	2.0	8.1	.04	.13	.66	4.4	.23	.13	.14	2.5	.08	.04
SynSim	.31	6.1	.19	.18	.15	**1.7**	.13	.13	.31	5.2	.18	.22	.54	4.8	.09	.11
SynDiff	.35	6.9	.20	**.20**	**.11**	3.15	.09	.14	.31	7.2	**.14**	**.24**	.58	6.1	.13	**.13**
SynOrig	–	–	–	–	–	–		–	–	–		–	.94	4.9	.07	.10
SynTr	–	–	–	–	–	–		–	–	–		–	.92	4.9	.08	.10
StyMask	**.05**	**.59**	**.04**	.03	**.11**	**.98**	**.01**	.03	**.16**	1.0	.20	.08	**.08**	1.3	**.03**	.03
DP-BART ($\epsilon = 50$)	.41	5.7	.19	**.26**	.67	11.1	.28	**.28**	.46	8.8	.37	**.28**	.34	4.6	.19	**.18**
DP-BART ($\epsilon = 100$)	.29	4.5	.13	.17	.54	9.8	.22	**.20**	–	–		–	–	–		–
DP-BART ($\epsilon = 250$)	.12	2.2	**.07**	.12	.47	7.1	.07	.13	–	–		–	–	–		–
LLaMA-first	.30	2.6	**.07**	.11	.38	5.6	.11	.14	.60	**.43**	.22	.17	.82	**.46**	.50	.06
LLaMA-second	.42	2.3	.08	.12	.34	5.4	.13	.14	.59	**.43**	.23	.17	.78	**.48**	.50	.06

Figure 5
Absolute Spearman's rank correlation between sample-level and distribution-level metric values.

A.5 SwDA Privacy

Listing 1 shows the speaker attribute inference prompt adapted from Staab et al. (2024).

Listing 1
SwDA privacy experiments prompt as described in Section 5.2.

```
[INST] <<SYS>> You are an expert investigator with years of experience in online profiling
and text analysis. You work with an analytical mindset and try to answer questions as
precisely as possible.
<</SYS>>

Let us play a guessing game. The following are openly available everyday conversations by
two speakers. They have been carefully transcribed for open research purposes and
therefore contain natural disfluencies and fillers. Do not be distracted by these
characteristics of speech transcriptions. Your task is to guess speaker A and B's (1)
gender and (2) level of education.

Conversation:

{full_transcript}

First reason step-by-step for each value going over all information in the text as well as
language used. Then give your top guess for gender and top guess for education level based
on your reasoning. For this task, the accepted gender values are [''male'', ''female''].
The accepted education level values are: [''less than highschool'', ''less than college'',
''college'', ''more than college'''']. If you are unsure, make your best guess given the
provided information. Only give one answer value.
Note that someone who received a highschool diploma is considered less than college,
someone currently in college is considered less than college, and any postgraduate studies
are considered more than college.

Follow exactly the format specified below:

Speaker: A or B
Type: education or gender
Inference: Your detailed inference
Guess: Your top guess as a string.

Now, think carefully, choose verbatim from the valid options, and provide the response in
the specified format. [/INST]
```

A.6 Utility Experiments Classifiers

- *Twitter, Yelp:* We fine-tune DISTILBERT-BASE-CASED and train models (1) on real data annotated with demographic attributes (Reddy and Knight 2016; Blodgett, Green, and O'Connor 2016) to measure author profiling risks and (2) on synthetic data to assess utility. We train on cross-entropy loss for 3 epochs on a batch size of 32, using the AdamW optimizer and an initial learning rate of 1e-3 and ϵ of 1e-8.
- *SwDA, TalkLife:* We follow the best performing configurations reported in prior work (Żelasko, Pappagari, and Dehak 2021; Tsakalidis et al. 2022), shown in Table 14.

Table 14
Classifier implementation details.

	XLNet-base-cased	BERT-base-uncased	2 layer biLSTM, dense layer with softmax
LR	5e-5	2e-5	1e-3
Dropout	.10	.25	.25
Batch size	6	8	16
Optimizer	Adam	Adam	Adam
Loss	cross entropy	focal loss ($\gamma = 2$)	cross entropy
Epochs	10	3	100 (early stopping)

Acknowledgments

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant no. EP/V030302/1), Keystone grant funding to Maria Liakata and Julia Ive from Responsible AI (grant no. EP/Y009800/1), and the Alan Turing Institute (grant no. EP/N510129/1). Jenny Chim was supported by a Google DeepMind Studentship. We are grateful to our reviewers and thank them for their support.

References

Abbasi, Ahmed and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29. https://doi.org/10.1145/1344411.1344413

Adelani, David, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. 2021. Preventing author profiling through zero-shot multilingual back-translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8687–8695. https://doi.org/10.18653/v1/2021.emnlp-main.684

Ahn, Sumyeong, Jongwoo Ko, and Se-Young Yun. 2023. CUDA: Curriculum of data augmentation for long-tailed recognition. In *The Eleventh International Conference on Learning Representations*.

Alaa, Ahmed M., Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. 2021. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. *ArXiv*, abs/2102.08921.

Alberti, Chris, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173. https://doi.org/10.18653/v1/P19-1620

Altakrori, Malik, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256. https://doi.org/10.18653/v1/2021.findings-emnlp.359

Altakrori, Malik, Thomas Scialom, Benjamin C. M. Fung, and Jackie Chi Kit Cheung. 2022. A multifaceted framework to evaluate evasion, content preservation, and misattribution in authorship obfuscation techniques. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2391–2406. https://doi.org/10.18653/v1/2022.emnlp-main.153

Arbeláez, Pablo, Michael Maire, Charlotte Fowlkes, and Julien Malik. 2011. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:898–916. https://doi.org/10.1109/TPAMI.2010.161, PubMed: 20733228

Aura, Tuomas, Thomas A. Kuhn, and Michael Roe. 2006. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, WPES '06, pages 41–50. https://doi.org/10.1145/1179601.1179608

Balunović, Mislav, Dimitar I. Dimitrov, Nikola Jovanović, and Martin Vechev. 2022. LAMP: Extracting text from gradients with language model priors. In *Conference on Neural Information Processing Systems*, 18 pages.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT

evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*, page 1.

Bazarova, Natalya N. and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64:635–657. https://doi.org/10.1111/jcom.12106

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Blodgett, Su Lin, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130. https://doi.org/10.18653/v1/D16-1120

Bloustein, Edward J. 1978. *Individual and Group Privacy*. New Brunswick: Transaction Publishers.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364.

Broder, Andrei Z. 1997. On the resemblance and containment of documents. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*. pages 21–29.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Buczak, Anna L., S. Babin, and Linda J. Moniz. 2010. Data-driven approach for creating synthetic electronic medical records. *BMC Medical Informatics and Decision Making*, 10:59. https://doi.org/10.1186/1472-6947-10-59, PubMed: 20946670

Büchi, Moritz, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamò-Larrieux, Shruthi Velidi, and Salome Viljoen. 2020. The chilling effects of algorithmic profiling: Mapping the issues. *Computer Law & Security Review*, 36:105367. https://doi.org/10.1016/j.clsr.2019.105367

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 19 pages.

Chen, Jiaao, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211. https://doi.org/10.1162/tacl_a_00542

Cheng, Myra, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532. https://doi.org/10.18653/v1/2023.acl-long.84

Cheng, Myra, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875. https://doi.org/10.18653/v1/2023.emnlp-main.669

Cheng, Zhoujun, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model APIs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 792–810. https://doi.org/10.18653/v1/2023.emnlp-industry.74

Choi, E., Siddharth Biswal, Bradley A. Malin, Jon D. Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the*

2nd Machine Learning for Healthcare Conference, pages 286–305.

Cohen, Aloni and Kobbi Nissim. 2020. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352. https://doi.org/10.1073/pnas.1914598117, PubMed: 32234789

Dai, Zhuyun, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In the *Eleventh International Conference on Learning Representations*.

Davidson, Brittany I., Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen, and Alicia G. Cork. 2023. Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, 7(12):2054–2057. https://doi.org/10.1038/s41562-023-01750-2, PubMed: 37919445

Devaraj, Ashwin, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345. https://doi.org/10.18653/v1/2022.acl-long.506, PubMed: 36404800

DiMarco, Chrysanne and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–500.

Dinan, Emily, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188. https://doi.org/10.18653/v1/2020.emnlp-main.656

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98

Dunbar, Robin I. M., Anna Marriott, and Neil D. C. Duncan. 1997. Human conversational behavior. *Human Nature*, 8:231–246. https://doi.org/10.1007/BF02912493, PubMed: 26196965

Durmus, Esin, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083. https://doi.org/10.1162/tacl_a_00506

Eldan, Ronen and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*.

Finch, Sarah E., James D. Finch, and Jinho D. Choi. 2023. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071. https://doi.org/10.18653/v1/2023.acl-long.839

Fire, Michael, Roy Goldschmidt, and Yuval Elovici. 2013. Online social networks: Threats and solutions. *IEEE Communications Surveys & Tutorials*, 16:2019–2036. https://doi.org/10.1109/COMST.2014.2321628

Fitzpatrick, Kathleen Kara, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19. https://doi.org/10.2196/mental.7785, PubMed: 28588005

Floridi, Luciano. 2017. Group privacy: A defence and an interpretation. In Bart van der Sloot, Luciano Floridi, and Linnet Taylor, editors, *Group Privacy*. Springer Verlag. https://doi.org/10.1007/978-3-319-46608-8_5

Foster, Jennifer and Oistein Andersen. 2009. GenERRate: Generating errors for use in grammatical error detection. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90. https://doi.org/10.3115/1609843.1609855

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Gabriel, Saadia, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487. https://doi.org/10.18653/v1/2021.findings-acl.42

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB dataset of diverse text for language modeling. *CoRR*, abs/2101.00027. https://doi.org/10.18653/v1/2021.emnlp-main.552

Gao, Tianyu, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120. https://doi.org/10.18653/v1/2021.gem-1.10

Godfrey, John J., Edward Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520. vol.1. https://doi.org/10.1109/ICASSP.1992.225858

Gröndahl, Tommi and Nirmal Asokan. 2020. Effective writing style transfer via combinatorial paraphrasing. *Proceedings on Privacy Enhancing Technologies*, 2020:175–195. https://doi.org/10.2478/popets-2020-0068

Gupta, Prakhar, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801. https://doi.org/10.18653/v1/2022.acl-long.263

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

He, Xuanli, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842. https://doi.org/10.1162/tacl_a_00492

Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, volume 30, pages 6629–6640.

Hills, Anthony, Adam Tsakalidis, Federico Nanni, Ioannis Zachos, and Maria Liakata. 2023. Creation and evaluation of timelines for longitudinal user posts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3791–3804. https://doi.org/10.18653/v1/2023.eacl-main.274

Hori, Chiori and Takaaki Hori. 2017. End-to-end conversation modeling track in DSTC6. *ArXiv preprint arXiv:1706.07440*.

Horvitz, Eric and Deirdre Mulligan. 2015. Data, privacy, and the greater good. *Science*, 349(6245):253–255. https://doi.org/10.1126/science.aac4520, PubMed: 26185242

Hovy, Dirk and Shrimai Prabhumoye. 2021. Five sources of bias in natural language

processing. *Language and Linguistics Compass*, 15(8):e12432. https://doi.org/10.1111/lnc3.12432, PubMed: 35864931

Howes, C., Mary Lavelle, Patrick G. T. Healey, J. Hough, and Rosemarie McCabe. 2017. Disfluencies in dialogues with patients with schizophrenia. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Hsu, I Hung, Avik Ray, Shubham Garg, Nanyun Peng, and Jing Huang. 2023. Code-switched text synthesis in unseen language pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5137–5151. https://doi.org/10.18653/v1/2023.findings-acl.318

Hu, Zhiting, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Humphreys, Lee, Phillipa Gill, and Balachander Krishnamurthy. 2010. How much is too much? Privacy issues on Twitter. In *Conference of International Communication Association*.

Igamberdiev, Timour and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934. https://doi.org/10.18653/v1/2023.findings-acl.874

Ireland, Molly E. and James W. Pennebaker. 2010. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99(3):549. https://doi.org/10.1037/a0020386, PubMed: 20804263

Ive, Julia, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert J. Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ Digital Medicine*, 3:Article 69. https://doi.org/10.1038/s41746-020-0267-x, PubMed: 32435697

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 44 pages. https://doi.org/10.1145/3571730

Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *ArXiv preprint arXiv:2401.04088*.

Jiang, Guangyuan, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 22 pages.

Juola, Patrick. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334. https://doi.org/10.1561/1500000005

Jurafsky, Daniel, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. https://doi.org/10.1109/ASRU.1997.658992

Karadzhov, Georgi, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2017. The case for being average: A mediocrity approach to style masking and author obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 173–185. https://doi.org/10.1007/978-3-319-65813-1_18

Kasai, Jungo, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557. https://doi.org/10.18653/v1/2022.naacl-main.259

Keküllüoğlu, Dilara, Walid Magdy, and Kami Vaniea. 2020. Analysing privacy leakage of life events on Twitter. In *Proceedings of the 12th ACM Conference on Web Science*. https://doi.org/10.1145/3394231.3397919

Kochkina, Elena, Tamanna Hossain, Robert L. Logan, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023.

Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, 60(1):103116. https://doi.org/10.1016/j.ipm.2022.103116

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94. https://doi.org/10.1007/s10579-009-9111-2

Krishna, Kalpesh, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762. https://doi.org/10.18653/v1/2020.emnlp-main.55

Kulshreshtha, Apoorv, Daniel De Freitas Adiwardana, David Richard So, Gaurav Nemade, Jamie Hall, Noah Fiedel, Quoc V. Le, Romal Thoppilan, Thang Luong, Yifeng Lu, and Zi Yang. 2020. Towards a human-like open-domain chatbot. https://research.google/pubs/towards-a-human-like-open-domain-chatbot/

Kumar, Varun, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Lavi, Ofer, Ella Rabinovich, Segev Shlomov, David Boaz, Inbal Ronen, and Ateret Anaby Tavor. 2021. We've had this conversation before: A novel approach to measuring dialog similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1169–1177. https://doi.org/10.18653/v1/2021.emnlp-main.89

Le Bras, Ronan, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088.

Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445. https://doi.org/10.18653/v1/2022.acl-long.577

Li, Yuanzhi, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023a. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461. https://doi.org/10.18653/v1/2023.emnlp-main.647

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 525(1):140–146. https://doi.org/10.1111/nyas.15007, PubMed: 37230490

Lison, Pierre, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203. https://doi.org/10.18653/v1/2021.acl-long.323

Liu, Chia Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. https://doi.org/10.18653/v1/D16-1230

Liu, Yang, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178. https://doi.org/10.18653/v1/D17-1231

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Lord, S. P., Elisa Sheng, Zac E. Imel, John S. Baer, and David C. Atkins. 2015. More

than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior Therapy*, 463:296–303. https://doi.org/10.1016/j.beth.2014.11.002, PubMed: 25892166

Louviere, Jordan J. and George G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22:319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Lukasik, Michal, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. 2022. Teacher's pet: Understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*.

Lyu, Yiwei, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2116–2138. https://doi.org/10.18653/v1/2021.naacl-main.171

Mahmood, Asad, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using Mutant-X. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71. https://doi.org/10.2478/popets-2019-0058

Marion, Max, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564*.

Mattern, Justus, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873. https://doi.org/10.18653/v1/2022.emnlp-main.323

Mattern, Justus, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343. https://doi.org/10.18653/v1/2023.findings-acl.719

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Melamud, Oren and Chaitanya Shivade. 2019. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45. https://doi.org/10.18653/v1/W19-1905

Michel, Paul and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318. https://doi.org/10.18653/v1/P18-2050

Mireshghallah, Fatemehsadat and Taylor Berg-Kirkpatrick. 2021. Style pooling: Automatic text style obfuscation for improved classification fairness. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2009–2022. https://doi.org/10.18653/v1/2021.emnlp-main.152

Mohamed, Youssef, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785. https://doi.org/10.18653/v1/2022.emnlp-main.600

Møller, Anders Giovanni, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

Ng, Boon Liang Clarence, Diogo Santos, and Marek Rei. 2023. Modelling temporal

document sequences for clinical ICD coding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1640–1649. https://doi.org/10.18653/v1/2023.eacl-main.120

Nguyen, Dong, A Seza Doğruöz, Carolyn P. Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593. https://doi.org/10.1162/COLI_a_00258

Niu, Tong and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389. https://doi.org/10.1162/tacl_a_00027

Niu, Tong, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised paraphrasing with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5136–5150. https://doi.org/10.18653/v1/2021.emnlp-main.417

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13. https://doi.org/10.3389/fdata.2019.00013, PubMed: 33693336

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. https://doi.org/10.3115/1073083.1073135

Pardo, Francisco Manuel Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th Author Profiling task at PAN 2017: Gender and language variety identification in Twitter. In *CLEF*.

Park, Joon Sung, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. arXiv preprint arXiv:2208.04024*.

Peng, Fuchun, Dale Schuurmans, Vlado Keselj, and Shaojun Wang. 2003. Language independent authorship attribution with character level n-grams. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 267–274. https://doi.org/10.3115/1067807.1067843

Pennebaker, James W. and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–312. https://doi.org/10.1037/0022-3514.77.6.1296, PubMed: 10626371

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. https://doi.org/10.18653/v1/W15-3049

Potthast, Martin, Matthias Hagen, and Benno Stein. 2016. Author obfuscation: Attacking the state of the art in authorship verification. In *Conference and Labs of the Evaluation Forum*.

Preoţiuc-Pietro, Daniel, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764. https://doi.org/10.3115/v1/P15-1169

Qian, Rebecca, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521. https://doi.org/10.18653/v1/2022.emnlp-main.646

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384. https://doi.org/10.18653/v1/2023.emnlp-main.85

Quarteroni, Silvia, Alexei V. Ivanov, and Giuseppe Riccardi. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pages 5596–5599. https://doi.org/10.1109/ICASSP.2011.5947628

Rabinovich, Ella, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1074–1084. https://doi.org/10.18653/v1/E17-1101

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Rashkin, Hannah, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840. https://doi.org/10.1162/coli_a_00486

Reddy, Sravana and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26. https://doi.org/10.18653/v1/W16-5603

Reiter, Ehud. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401. https://doi.org/10.1162/coli_a_00322

Reiter, Ehud and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87. https://doi.org/10.1017/S1351324997001502

Rivera-Soto, Rafael A., Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919. https://doi.org/10.18653/v1/2021.emnlp-main.70

Roemmele, Melissa, Andrew S. Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*, pages 13–17.

Rosenthal, Sara and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772.

Sai, Ananya B., Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234. https://doi.org/10.18653/v1/2021.emnlp-main.575

Sakaguchi, Keisuke, Matt Post, and Benjamin Van Durme. 2017. Error-repair dependency parsing for ungrammatical texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195. https://doi.org/10.18653/v1/P17-2030

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678. https://doi.org/10.18653/v1/P19-1163

Sapkota, Upendra, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102. https://doi.org/10.3115/v1/N15-1010

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Scialom, Thomas, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604. https://doi.org/10.18653/v1/2021.emnlp-main.529

See, Abigail, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723. https://doi.org/10.18653/v1/N19-1170

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468

Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293. https://doi.org/10.18653/v1/2021.acl-long.330

Shetty, Rakshith, Bernt Schiele, and Mario Fritz. 2018. A^{4}NT: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1633–1650.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. https://doi.org/10.3115/1613715.1613751

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 95 pages.

Staab, Robin, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In the *Twelfth International Conference on Learning Representations*.

Stamatatos, Efstathios, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the PAN/CLEF 2015 evaluation lab. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283*, pages 518–538. https://doi.org/10.1007/978-3-319-24027-5_49

Su, Hongjin, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121. https://doi.org/10.18653/v1/2023.findings-acl.71

Sun, Jiao, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189. https://doi.org/10.18653/v1/2021.emnlp-main.420

Swayamdipta, Swabha, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293. https://doi.org/10.18653/v1/2020.emnlp-main.746

Sweeney, Latanya. 2002. *k*-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

https://doi.org/10.1142/S0218488502001648

Tan, Samson, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935. https://doi.org/10.18653/v1/2020.acl-main.263

Tang, Yuqing, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466. https://doi.org/10.18653/v1/2021.findings-acl.304

Tantipongpipat, Uthaipon Tao, Chris Waites, Digvijay Boob, Amaresh Ankit Siva, and Rachel Cummings. 2021. Differentially private synthetic mixed-type data generation for unsupervised learning. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–9. https://doi.org/10.1109/IISA52424.2021.9555521

Thorn Jakobsen, Terne Sasha, Maria Barrett, Anders Søgaard, and David Lassen. 2022. The sensitivity of annotator bias to task definitions in argument mining. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 44–61.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tsakalidis, Adam, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660. https://doi.org/10.18653/v1/2022.acl-long.318

van Breugel, Boris and Mihaela van der Schaar. 2023. Beyond privacy: Navigating the opportunities and challenges of synthetic data. *arXiv preprint arXiv:2304.03722*.

van Miltenburg, Emiel, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153. https://doi.org/10.18653/v1/2021.inlg-1.14

Veselovsky, Veniamin, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *ArXiv*, abs/2305.15041.

Walonoski, Jason, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238. https://doi.org/10.1093/jamia/ocx079, PubMed: 29025144

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

Wang, Angelina and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning*, pages 10882–10893.

Wang, Dingquan and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505. https://doi.org/10.1162/tacl_a_00113

Wang, Yue, Cuong Hoang, and Marcello Federico. 2021. Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1193–1199. https://doi.org/10.18653/v1/2021.naacl-main.94

Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Wang, Zixu, Julia Ive, Sumithra Velupillai, and Lucia Specia. 2019. Is artificial data useful for biomedical natural language processing algorithms? In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 240–249. https://doi.org/10.18653/v1/W19-5026

Wang, Zijian and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45. https://doi.org/10.18653/v1/D18-1004

Weerasinghe, Janith, Rhia Singh, and Rachel Greenstadt. 2022. Using authorship verification to mitigate abuse in online communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1075–1086. https://doi.org/10.1609/icwsm.v16i1.19359

Wegmann, Anna and Dong Nguyen. 2021. Does it capture STEL? A modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130. https://doi.org/10.18653/v1/2021.emnlp-main.569

Welling, Max. 2009. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128. https://doi.org/10.1145/1553374.1553517

Williamson, Vanessa. 2016. On the ethics of crowdsourced research. *PS: Political Science 38; Politics*, 49(1):77–81. https://doi.org/10.1017/S104909651500116X

Wullach, Tomer, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705. https://doi.org/10.18653/v1/2021.findings-emnlp.402

Wyllie, Sierra, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: Training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2113–2147. https://doi.org/10.1145/3630106.3659029

Xia, Mengzhou, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Xiang, Jiannan, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198. https://doi.org/10.18653/v1/P19-1579

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019a. Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32. https://doi.org/10.18653/v1/2021.findings-acl.193

Xu, Qiongkai, Lizhen Qu, Zeyu Gao, and Gholamreza Haffari. 2020. Personal information leakage detection in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6567–6580. https://doi.org/10.18653/v1/2020.emnlp-main.532

Xu, Qiongkai, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019b. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257. https://doi.org/10.18653/v1/W19-8633

Xu, Wei, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Xu, Zhangchen, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. *arXiv preprint arXiv:2406.08464*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

Ye, Jiacheng, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669. https://doi.org/10.18653/v1/2022.emnlp-main.801

Yoo, Kang Min, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2225–2239. https://doi.org/10.18653/v1/2021.findings-emnlp.192

Yu, Yue, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 45 pages.

Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. https://doi.org/10.18653/v1/P18-1205

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. https://doi.org/10.18653/v1/W18-5021

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989. https://doi.org/10.18653/v1/2021.emnlp-main.25

Zhao, Tianyu and Tatsuya Kawahara. 2018. A unified neural architecture for joint dialog act segmentation and recognition in spoken dialog system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–208. https://doi.org/10.1145/3209978.3210080

Zhu, Jian and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297. https://doi.org/10.18653/v1/2021.emnlp-main.25

Zhu, Yaoming, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1097–1100.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55. https://doi.org/10.1162/coli_a_00502

Zmigrod, Ran, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661. https://doi.org/10.18653/v1/P19-1161

Żelasko, Piotr, Raghavendra Pappagari, and Najim Dehak. 2021. What helps transformers recognize conversational structure? Importance of context, punctuation, and labels in dialog act recognition. *Transactions of the Association for Computational Linguistics*, 9:1163–1179. https://doi.org/10.1162/tacl_a_00420