Linguistic Identity Leakage: When Language Reveals Identity in Anonymized Text

Wajdi Zaghouani
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Privacy-preserving natural language processing (NLP) typically focuses on removing explicit identifiers such as names, addresses, and phone numbers. We argue that this approach overlooks a key risk: natural language itself encodes signals about a speaker's geographic origin, social background, and community membership that persist after anonymization. We introduce **Linguistic Identity Leakage** (LIL), defined as the inference of personal or demographic attributes from linguistic features in text where explicit identifiers have been removed. We further introduce **Linguistic Personally Identifiable Information** (L-PII) to denote the linguistic features that enable such inference. Drawing on sociolinguistics, stylometry, and NLP privacy research, we propose a taxonomy of linguistic identity signals across five categories and examine implications for dataset release, language model training, and privacy auditing. Using examples from Arabic dialectal variation and other multilingual contexts, we present the **Identity Inference Risk** (IIR) framework for assessing residual privacy risk in NLP systems and discuss how contemporary LLMs amplify these risks. Our goal is to encourage broader recognition of the gap between conventional anonymization practices and the linguistic reality of natural language data.

1 Introduction

A text can reveal who wrote it even after every name, address, and phone number has been removed. Dialectal vocabulary places an author in a region; code-switching patterns encode education and community; stylometric regularities fingerprint the individual. These signals constitute what we call Linguistic Identity Leakage (LIL), and they survive many of the anonymization pipelines currently in wide use across NLP.

Privacy protection in NLP often relies on a long-standing operational assumption: once explicit identifiers have been removed from a text, the resulting document is sufficiently de-identified to protect the author's privacy. This assumption motivates the standard pipeline of named entity recognition followed by redaction or pseudonymization, and it underpins widely used anonymization frameworks in both clinical NLP and general-domain dataset release (Lison et al., 2021). We do not claim that all of privacy-preserving NLP rests on this assumption: the Text Anonymization Benchmark of Pilán et al. (2022) extends evaluation to indirect and quasi-identifiers, and a growing body of research applies differential privacy and language-model-based rewriting to text (Feyisetan et al., 2020; Igamberdiev and Habernal, 2023; Mattern et al., 2022; Klymenko et al., 2022). Even so, span-level redaction remains the dominant deployed practice, and the main threat models in NLP privacy research focus on either explicit identifier disclosure, training data memorization, or membership inference.

The assumption that identity is localized in a finite set of named spans is increasingly difficult to sustain given insights from two adjacent fields. Stylometric research has shown for decades that writing style alone can identify authors with high accuracy from distributional properties of word choice, punctuation, and syntactic structure (Stamatatos, 2009). Sociolinguistic work establishes that dialect, register, and code-switching are not random noise but systematic reflections of the speaker's geographic origin, social class, educational background, and community affiliation (Labov, 1972).

These two observations jointly imply that natural language text, even after explicit identifiers are stripped away, can serve as a rich source of identity inference. More recently, empirical work has shown that LLMs can infer personal attributes such as location, income, and gender from ostensibly anonymized social media posts with surprisingly high accuracy, without any explicit disclosure of those attributes by the author (Staab et al., 2024).

Proceedings of the Seventh Workshop on Privacy in Natural Language Processing, pages 107–117

Yet NLP privacy research has not articulated a unified framework that connects these threads. Research on anonymization beyond NER focuses on quasi-identifier span detection and disclosure risk over a closed set of attributes (Lison et al., 2021; Pilán et al., 2022); research on differential privacy focuses on calibrating noise to individual data contributions (Dwork, 2006; Feyisetan et al., 2020); and research on training data extraction focuses on memorization of verbatim text strings (Carlini et al., 2021). None of these lines of work treats the structural linguistic surface of the text itself, by which we mean the dialectal, sociolectal, code-switching, and stylometric patterns distributed across a document, as a privacy surface in its own right. This is what we mean by the *linguistic fabric* of a text.

This paper addresses that gap. We make four contributions. First, we introduce **Linguistic Identity Leakage** (LIL) as a distinct privacy risk category, separate from explicit identifier disclosure, training data memorization, and membership inference. Second, we introduce **Linguistic Personally Identifiable Information** (L-PII) to denote the linguistic features that contribute to this risk. Third, we develop a five-category taxonomy of such features with examples from Arabic and other languages, presented as an initial synthesis rather than a closed classification. Fourth, we propose the **Identity Inference Risk** (IIR) framework that operationalizes LIL as a measurable property of NLP systems and datasets. We also discuss LLM amplification of LIL risk and propose mitigation directions. Throughout, we use Arabic dialectal data as a running example, given that Arabic presents one of the most vivid cases of geographically and socially stratified linguistic variation globally (Holes, 2004; Bouamor et al., 2018), and we comment on cross-linguistic generalization in Section 8.

2 Background and Related Work

2.1 Privacy-Preserving NLP Beyond Span Redaction

NLP privacy research is broader than NER-based redaction. Lison et al. (2021) survey anonymization models for text data, arguing for a move beyond sequence labeling toward methods that explicitly model disclosure risk. Their framework distinguishes direct identifiers (names, identification numbers) from quasi-identifiers (age, occupation, location). The Text Anonymization Benchmark of Pilán et al. (2022) extends this with annotated spans for both direct and indirect identifiers and evaluation metrics that target disclosure risk rather than sequence labeling accuracy.

Differential privacy (DP) is another major formal privacy framework for NLP (Dwork, 2006; Feyisetan et al., 2020). Feyisetan et al. (2020) proposed metric local DP for text via calibrated noise on word embeddings; Igamberdiev and Habernal (2023) extended this to sentence-level rewriting with DP-BART; Mattern et al. (2022) show that word-level DP mechanisms face mathematical constraints that limit their practical privacy guarantees; and Klymenko et al. (2022) review DP for NLP and the challenge of preserving coherence and utility alongside privacy. These methods protect content at the level of individual tokens or short spans, with paraphrase and rewriting taking steps toward stylistic protection. What remains missing is a unified treatment of the dialectal, sociolectal, code-switching, and stylometric signals distributed throughout the text at a structural level. Our framework is intended to complement this line of work, not replace it.

Training data extraction is another threat model. Carlini et al. (2021) showed that LLMs can memorize and reproduce verbatim personal information such as names, phone numbers, and email addresses. This is conceptually distinct from LIL: extraction targets verbatim reproduction from memorized training data, whereas LIL targets identity inference from the linguistic character of held-out anonymized text. The two can interact, since a model that has seen many examples of a dialect or sociolect is plausibly better at recognizing it later, but neither reduces to the other.

2.2 LLMs as Inference Engines for Personal Attributes

A critical recent development that motivates our framework is the finding that LLMs can infer sensitive personal attributes from text at inference time. Staab et al. (2024) construct a dataset of real Reddit profiles and show that current LLMs infer attributes such as geographic location, income, and sex with up to 85% top-1 accuracy at a fraction of the cost and time required by human annotators. They also show that common mitigations such as text anonymization and model alignment are currently ineffective against this form of inference, and provide qualitative examples in which geographic origin was inferred from region-specific expressions the user never intended as location dis-

closures. Subsequent work in the same line has extended these findings to images and to synthetic profiles, highlighting that attribute inference scales with model capability (Staab et al., 2025).

The implication for our framework is that LIL is not merely theoretical; contemporary LLMs already constitute a practical adversary capable of exploiting it at scale, without needing access to the author's training data. We discuss the precise relationship between LIL and the broader notion of LLM attribute inference in Section 3.

2.3 Author Profiling and Attribute Inference from Linguistic Features

A long line of NLP research has studied demographic inference from writing style independently of explicit identifier removal. Schler et al. (2006) analyzed a large blog corpus and found that lexical and stylistic features could infer author age and gender with meaningful accuracy. Subsequent studies confirmed similar patterns on Twitter, online forums, and other platforms (Neal et al., 2017). Stylometric research more broadly establishes that writing style constitutes an authorial fingerprint (Stamatatos, 2009), with character n-grams, function word distributions, and syntactic complexity measures identifying authors with high accuracy across genres and domains. Zhai et al. (2022) demonstrate that adversarially trained attributors can degrade the effectiveness of existing obfuscation approaches from 20–30% success to 5–10%, underscoring that naive stylometric flattening is not a reliable mitigation against a determined adversary. A recent survey of privacy risks across social media NLP tasks argues that dialect and native-language identification exhibit high re-identification potential through linguistic fingerprinting and that, in the cases analyzed, anonymized text often remained traceable through stylistic and syntactic cues (Goswami et al., 2026).

2.4 Sociolinguistic Variation

Sociolinguistics provides the theoretical grounding for why language encodes social identity. Labov (1972) established that phonological, lexical, and syntactic choices correlate systematically with speakers' social characteristics including class, age, gender, and geographic region. Code-switching, alternating between two or more languages, encodes information about bilinguals' community memberships and educational histories (Myers-Scotton, 1993). Resources such as the MADAR Arabic Dialect Corpus (Bouamor et al., 2018), which documents systematic lexical differences across 25 Arabic cities, make concrete how fine-grained the geographic information encoded in dialectal variation can be: written text can in principle disclose not just a speaker's country but their city of origin.

2.5 The Research Gap

Despite these converging lines of evidence, NLP privacy research lacks a framework that synthesizes them into a unified privacy risk model. Existing work treats stylometry, sociolinguistics, anonymization, and LLM inference largely in isolation. Our paper provides that synthesis, connecting anonymization research, differential privacy, stylometry, sociolinguistics, and LLM inference into a coherent account of the linguistic surface as a privacy risk that span-level methods do not target.

3 Linguistic Identity Leakage

3.1 Definition

We define **Linguistic Identity Leakage** (LIL) as follows:

> *Linguistic Identity Leakage occurs when linguistic features present in a text allow an adversary to infer personal or demographic attributes about the author even after all explicit personally identifiable information has been removed from that text.*

This definition has three key components. The phrase *linguistic features* refers to properties of the text that arise from the author's natural language choices rather than from explicit declarations of identity. The phrase *personal or demographic attributes* includes geographic origin, social class, community membership, ethnicity, educational background, and professional identity, among others. The phrase *even after all explicit PII has been removed* anchors LIL as a residual privacy risk, distinct from failures of explicit identifier removal.

3.2 Distinction from Related Privacy Risks

Table 1 situates LIL in relation to existing privacy risk categories in the NLP literature.

The relationship between LIL and the LLM attribute inference threat of Staab et al. (2024) deserves attention. We see it as one of subset and complementary emphasis. Staab et al. (2024) show that

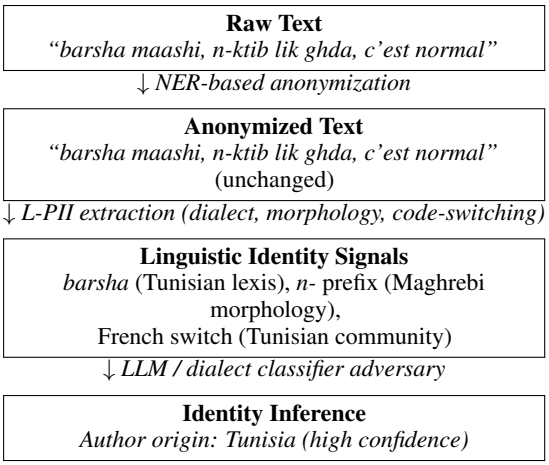

Figure 1: The LIL pipeline. Standard NER-based anonymization leaves dialectal, morphological, and code-switching L-PII entirely intact, enabling downstream identity inference. The example text means "a lot is going on, I'll write to you tomorrow, it's fine" and would not be altered by any current span-based anonymization system.

Privacy Risk	Mechanism
Explicit identifier leakage	Names, phone numbers, addresses in text
Training data memorization	Model reproduces verbatim training sequences
Membership inference	Adversary determines whether a record was in training data
Demographic inference from content	Explicit statements reveal sensitive attributes
LLM attribute inference	LLM infers attributes from contextual cues at inference time
LIL (this paper)	**Identity inference from structural linguistic features after explicit PII removal**

Table 1: LIL in relation to existing NLP privacy risk categories. LIL is distinct because it operates on the implicit linguistic fabric of text rather than on explicit content, model memorization, or inference-time semantic context alone.

LLMs infer attributes from text using a broad mixture of semantic content, stylistic cues, and world knowledge, and aggregate across these signals. LIL is the portion of that risk traceable to structural linguistic features (dialect, register, code-switching, stylometric patterns) rather than to semantic content or world-knowledge reasoning, and it is the portion span-based anonymization is least equipped to address. The framework value of LIL is therefore not the existence of the threat, which Staab et al. (2024) establish empirically, but the analytical separation: the appropriate mitigations (dialect neutralization, stylometric flattening, structurally aware DP rewriting) differ from those for content-driven disclosure (semantic redaction, topic suppression). This separation is what enables the topic-controlled evaluation in the IIR framework (Section 7).

3.3 Linguistic Personally Identifiable Information

Building on LIL, we introduce **Linguistic Personally Identifiable Information** (L-PII): the set of linguistic features in a text that enable identity or demographic inference. The concept is modeled on the notion of quasi-identifiers in the database privacy literature (Sweeney, 2002), where attributes that are individually innocuous can collectively enable re-identification. In the linguistic domain, a single dialectal word may not suffice to localize a speaker, but a consistent pattern of such choices across a document produces a profile that functions as a combination of quasi-identifiers.

L-PII differs from conventional quasi-identifiers in two respects. First, L-PII arises from implicit communicative behavior that is not consciously controlled by the author, making it harder to remove through self-censorship or deliberate style modification. Second, the combination of L-PII features across a document creates a high-dimensional profile whose de-identification cannot be achieved by suppressing a fixed list of span types, since the identity signal is distributed throughout the text rather than localized in specific named entities or quasi-identifier fields.

4 A Taxonomy of Linguistic Identity Signals

We propose a five-category taxonomy of linguistic features that constitute L-PII. The taxonomy is constructed conceptually, drawing categories from the four anchor literatures of stylometry (Stamatatos, 2009; Neal et al., 2017), sociolinguistic variation (Labov, 1972; Holes, 2004), code-switching theory (Myers-Scotton, 1993), and computational dialect identification (Bouamor et al., 2018; Abdul-Mageed et al., 2024). The five categories are intended to cover the main types of linguistic signal that current dialect, sociolect, and stylometric NLP systems already operate on. We do not claim exhaustiveness: a more systematic taxonomy construction in the sense of Nickerson et al. (2013), with explicit ending conditions and iterative empirical-to-conceptual refinement, is a natural follow-up that we leave to future work. For

Feature	Gulf	Levant	Tunisia
Lexical (meaning: very / now / what)			
very	marra	kteer	barsha
now	haaliin	halla'	daaba
what	shnoo	shuu	shniyya
Verbal morphology (imperfect prefix)			
I write	aktib	b-ktub	n-ktib
Code-switching ("it's fine")			
	it's fine	okay ya3ni	c'est normal

Table 2: Lexical, morphological, and code-switching L-PII signals across three Arabic dialect groups (transliterated). Each signal type survives NER-based anonymization and enables geographic inference (Abdul-Mageed et al., 2024).

each category we describe the type of identity inference it enables and provide illustrative examples.

4.1 Dialect Markers

Regional dialects carry systematic lexical, morphological, and code-switching differences that vary geographically. Table 2 illustrates several signal types for Arabic across three dialect groups. The MADAR Arabic Dialect Corpus (Bouamor et al., 2018) documents city-level variation across 25 Arabic cities, a geographic resolution far exceeding anything targeted by conventional anonymization.

NLP dialect identification systems are, from a privacy standpoint, identity inference engines operating on features that survive standard anonymization (Abdul-Mageed et al., 2024). The verbal morphology row in Table 2 illustrates a signal that is particularly hard to redact: the imperfect prefix *n-* is characteristic of Maghrebi Arabic, while *b-* marks Levantine imperfect aspect, and these prefixes appear on every verb in the text. Analogous structural patterns exist in many other languages and language families (Labov, 1972).

4.2 Sociolect Markers

Sociolects are language varieties associated with particular social groups defined by class, profession, age, or education. Sociolect markers include register choices, specialized vocabulary, slang, and formality patterns. In the Arabic context, the choice between Modern Standard Arabic and colloquial forms in formal writing is itself a sociolect marker, encoding educational background and familiarity with written registers.

Research on author profiling demonstrates that these markers are empirically potent. Schler et al. (2006) analyzed a large blog corpus and found significant differences in writing style and content between age groups and between male and female bloggers, differences strong enough to enable automated inference of both attributes. This result, replicated across different platforms and languages (Neal et al., 2017), confirms that sociolect markers are not merely theoretically useful for identity inference but already exploited by existing NLP systems. In English, patterns such as hedging density, nominalization frequency, and passive voice usage distinguish professional from lay writing and have been studied in the context of authorship attribution (Stamatatos, 2009).

4.3 Code-Switching Patterns

Code-switching between languages within a document constitutes a particularly informative L-PII feature. The choice of which language to switch into, the domains that trigger switching, and the syntactic positions at which switching occurs all encode information about the bilingual speaker's background (Myers-Scotton, 1993).

In Arabic social media, Arabic-English code-switching is pervasive and encodes signals about urban identity and higher education exposure. A sentence that embeds the English word *sensitive* inside an Arabic matrix clause ("al-mawdoo' sensitive shway", meaning 'the topic is a bit sensitive') reflects a sociolinguistic pattern associated with educated urban Arab bilinguals in the Gulf and Levant regions, whereas Tunisian-French code-switching is diagnostic of a different community entirely. These patterns persist in anonymized text even when all named entities are removed, and an LLM prompted to infer the author's regional background would have access to precisely this signal (Staab et al., 2024).

4.4 Cultural and Community-Specific References

Natural language texts contain implicit references to cultural institutions, practices, and norms that index community membership. In Arabic, kinship terms used as forms of address (such as *Abū Khālid*, 'father of Khalid') function as sociolinguistic conventions whose use and pragmatic context vary by region, serving as identity markers even without an accompanying proper name. When combined with dialect features and code-switching patterns, these cultural references compound the L-PII profile already present in the text. In small communities,

this combination may uniquely identify a speaker even after all named entities have been removed.

4.5 Stylistic and Discourse-Level Patterns

Consistent patterns in discourse organization, punctuation, sentence length, and discourse marker usage constitute the stylometric layer of L-PII: function word frequencies, character n-gram distributions, and syntactic complexity measures. These signals tend to be robust even against deliberate obfuscation attempts (Stamatatos, 2009; Neal et al., 2017), and adversarially aware attributors can defeat naive obfuscators (Zhai et al., 2022). Crucially, they operate at a level of abstraction entirely invisible to span-level anonymization: a named entity recognizer correctly redacts a name while leaving untouched the author's characteristic sentence structure, connective preferences, and function word distributions, all of which an LLM or stylometric system can exploit (Staab et al., 2024).

5 Illustrative Scenarios

We present three illustrative scenarios grounding LIL in real-world privacy contexts. We frame these as scenarios rather than empirical case studies, since we do not provide measurements; their function is to make the privacy risks concrete and to anchor the IIR framework in Section 7.

Scenario 1: Arabic Health Forum Data. Consider a corpus of anonymized patient forum posts in Arabic, from which names, hospital names, and city references have been removed before public release. Dialectal lexis (*barsha* for Tunisia, *marra* for Gulf) and the imperfect verbal prefix *n-* (Maghrebi) immediately signal regional origin, which can correlate with ethnicity or migrant status in multi-community healthcare contexts. An IIR audit using a dialect classifier from the NADI series (Abdul-Mageed et al., 2024) would detect this residual signal even after NER-based de-identification, demonstrating that span-level anonymization alone is insufficient for such corpora.

Scenario 2: Anonymous Whistleblower or Activist Text. An anonymous online comment uses Gulf Arabic lexis, Arabic-English code-switching in the pattern *"this is sensitive shway"*, and kinship address forms typical of Gulf communities. Even with all explicit identifiers removed, these markers jointly narrow the author's likely origin to a sub-national region. Stylometric features (sentence length distributions, function word profiles (Stamatatos, 2009)) may further link the post to a known activist's writing style. The political risks of such re-identification are severe, yet anonymization pipelines that stop at NER offer no protection against it.

Scenario 3: Anonymous Employee Feedback. An employee submits anonymous workplace feedback. High nominalization density, hedging patterns, and formal register choice signal higher educational background and professional seniority (Schler et al., 2006). In a small team, the combination of sociolect markers with code-switching patterns, such as embedding English technical terms into an Arabic sentence, can uniquely identify the submitter. No span-based anonymization tool targets these sociolect-level quasi-identifiers, leaving workplace privacy guarantees hollow.

6 LLMs as Amplifiers of Linguistic Identity Leakage

The risks we describe are not hypothetical. Staab et al. (2024) demonstrate that LLMs already constitute practical adversaries capable of exploiting L-PII at scale. They found LLMs infer personal attributes from text with substantially higher accuracy and lower cost than human annotators, and that this capability is not defeated by standard text anonymization. In one illustrative case, an LLM correctly inferred a user's city of origin from a comment mentioning a traffic maneuver specific to Melbourne, which is structurally analogous to dialect-based LIL. Their main study covers English Reddit content, so we treat the findings as strong evidence for English social media and a working hypothesis for other languages and domains.

This has three consequences. First, the threat model is realizable today: an adversary with API access to a state-of-the-art LLM can extract demographic profiles from anonymized corpora at scale. Second, adversary capabilities in the IIR framework should be calibrated to LLM-level inference power, not simpler classifiers. Third, because LLM capabilities continue improving, LIL risk is unlikely to diminish without active mitigation.

There is also a subtler training-time consequence. Models trained on dialectally and demographically diverse corpora internalize associations between linguistic features and demographic attributes as a by-product of training. This is logically distinct from the verbatim memorization studied by Carlini

et al. (2021): LIL does not require the target text to have been seen during training, only that sufficient similar patterns have been observed. Memorization and demographic profiling capability are therefore separable risks of LLM training, even though they share a common cause in large-scale exposure to user-generated text.

7 Implications for Privacy-Preserving NLP

7.1 Dataset Release and Anonymization Pipelines

Span-based anonymization pipelines focus primarily on named entity recognition and redaction or pseudonymization of direct and quasi-identifier spans (Lison et al., 2021). Even quasi-identifier-aware benchmarks such as TAB (Pilán et al., 2022) treat identity as localized in span-level evidence over a closed attribute set. Our framework shows that this assumption does not always hold: identity signals are distributed throughout the linguistic fabric of a text, and LLMs can already exploit these signals at scale (Staab et al., 2024).

A practical implication is that dataset release protocols should include a *linguistic privacy audit* alongside traditional anonymization. Such an audit would characterize the linguistic identity signals present in the dataset (dialect variety, code-switching patterns, sociolect markers) and assess the degree to which these signals enable inference of attributes that were intended to be protected. For Arabic datasets, resources like MADAR (Bouamor et al., 2018) provide a reference for characterizing the geographic specificity of dialect features. This is especially important for datasets drawn from marginalized communities, where the combination of community-specific linguistic features and small population sizes increases the risk of individual identification (Gebru et al., 2021).

7.2 Language Model Training

LLMs trained on large multilingual corpora internalize statistical associations between linguistic features and demographic categories. This creates a pathway for LIL at the model level that is distinct from the memorization-based threat studied by Carlini et al. (2021): the concern is not that the model reproduces specific training examples but that it acquires demographic profiling capabilities from aggregate patterns, and these capabilities can later be invoked at inference time on previously unseen text. Privacy audits of LLMs should therefore include assessments of the models' ability to infer protected attributes from linguistic features in held-out text, complementing the membership inference and extraction attack evaluations that are more commonly reported.

7.3 The Identity Inference Risk Framework

We propose the **Identity Inference Risk** (IIR) framework as an operational tool for assessing LIL in NLP datasets and systems. The framework has three components.

Feature characterization. For a given dataset or model output, identify the distribution of L-PII features across the five categories in Section 4: dialect markers, sociolect markers, code-switching patterns, cultural references, and stylistic/discourse patterns.

Inference risk estimation. For each L-PII category, estimate the adversary's advantage using concrete, measurable proxies: the area under the ROC curve (AUC) of a classifier trained to predict a protected demographic attribute (e.g., regional dialect, age group, gender) from text in that category; the mutual information between the feature distribution and the attribute label; and the calibrated confidence of an LLM adversary at inference time, following the methodology of Staab et al. (2024). Topic-normalized splits or neutral paraphrases should be used as controls to isolate structural linguistic signals from topical or semantic content cues, so that IIR scores reflect L-PII specifically rather than the semantic content of the text. IIR assessment should include LLM-based adversary simulation as a core component, since LLMs currently represent the most capable practical adversary (Staab et al., 2024).

Aggregated risk score. Combine the per-category AUC and mutual information estimates into an aggregated IIR score using a risk-sensitive composite (e.g., weighted maximum or calibrated sum), supporting comparisons across datasets and evaluation of mitigation effectiveness. Practitioners should report per-attribute and per-category subscores alongside the aggregate to preserve interpretability. A standardized IIR audit checklist (specifying the protected attributes assessed, the attacker class used, the calibration routine, the topic-control method, and utility metrics on downstream tasks) would constitute a concrete, reproducible deliverable for dataset and model releases. The IIR framework is deliberately framed as a risk assess-

ment tool rather than a formal privacy guarantee, since the degree of LIL depends on adversary capabilities that vary across languages and over time.

To make the framework concrete, we sketch how it would apply to an existing publicly available dataset. Goswami et al. (2026) report that, in the cases they examined, more than half of anonymized dialect datasets remained traceable through stylistic and syntactic cues, with Arabic, English, and code-mixed corpora particularly vulnerable. We treat these figures as preliminary evidence motivating IIR audits rather than a settled empirical claim. An IIR audit would proceed as follows. In *feature characterization*, a dialect identification model from the NADI series (Abdul-Mageed et al., 2024) classifies each document by region, yielding a per-document dialect signal profile. In *inference risk estimation*, a logistic regression and an LLM adversary (following Staab et al. 2024) each attempt to predict geographic region from documents whose named entities have been removed; AUC on held-out topic-matched pairs quantifies the residual structural risk. In *aggregated risk score*, subscores across dialect, sociolect, and code-switching categories combine into a composite IIR value. An IIR score substantially above 0.5 after topic control would constitute evidence of actionable L-PII, triggering a recommendation for dialect neutralization or tiered-access restriction before public release.

7.4 Mitigation Directions

We sketch three approaches and situate them relative to existing techniques. *Dialect neutralization* rewrites dialectal text into a standard form before release (e.g., Arabic dialectal content into MSA), at the cost of removing dialectal authenticity and potentially harming dialect-sensitive downstream tasks. *Stylometric flattening* post-processes text to reduce authorial fingerprint identifiability, building on authorship obfuscation (Stamatatos, 2009); however, Zhai et al. (2022) show that adversarially trained attributors largely defeat naive obfuscators, motivating defenses robust to adversarial re-identification. *Adversarial LIL auditing* uses an LLM adversary (Staab et al., 2024) in the anonymization pipeline, iteratively revising text until the adversary fails to infer demographic attributes with confidence; recent work explores related ideas in the broader anonymization setting (Staab et al., 2025). Existing DP rewriting methods (Feyisetan et al., 2020; Igamberdiev and Habernal, 2023; Mattern et al., 2022) provide baselines against which the privacy gain of each approach can be quantified; how much L-PII survives after DP rewriting is a key open question, since these methods protect token-level content rather than structural linguistic features. All three directions face the same privacy-utility tension: L-PII is woven into the communicative fabric of the text itself.

8 Discussion

The asymmetry between suppression and utility, and differential exposure. Suppressing L-PII requires altering the linguistic fabric of a text: a stylometrically flattened or dialect-neutralized text loses the author's voice and dialectal authenticity. This tension is more fundamental than for explicit identifier removal, and falls unevenly across communities. Speakers of minority or under-resourced varieties face greater LIL risk, a distributional justice concern that runs counter to the typical goals of NLP privacy research.

Generalization beyond Arabic. We use Arabic as a running example because it offers an unusually clear case of stratified variation (Holes, 2004; Bouamor et al., 2018), but the framework is not Arabic-specific. Geographic dialect variation, register stratification, and code-switching are documented across most language families, and stylometric authorial fingerprints (Stamatatos, 2009; Neal et al., 2017) are language-agnostic by design. The English Reddit results of Staab et al. (2024) and the cross-lingual coverage discussed by Goswami et al. (2026) indicate that L-PII operates in languages other than Arabic. What changes across languages is the resource landscape: languages with fewer dialect classifiers and stylometric tools are harder to audit, not necessarily safer.

Operationalizing structural vs. semantic signals. A key challenge for IIR is isolating structural linguistic signals from semantic content. We propose two controls: topic-normalized evaluation (comparing inference accuracy within topic-matched document pairs) and neutral paraphrase baselines (replacing dialect-specific lexis with standard equivalents and measuring the AUC drop). Residual AUC after both controls is the operational evidence of structural L-PII.

Does generative AI reduce LIL? A natural question is whether widespread use of generative AI

to produce or polish text reduces LIL by homogenizing individual style. This is plausible at the stylometric layer, where heavy LLM rewriting can flatten authorial idiosyncrasies. It is less plausible at the dialectal and code-switching layer, since most general-purpose chat assistants default to standard written varieties and cannot reliably reproduce dialectal lexis or morphology. We would therefore expect heavy LLM mediation to shift the L-PII profile toward stylistic flattening while leaving dialect and code-switching signals largely intact in user-authored portions, but verifying this is an open empirical question.

Tools, consent, and governance. The feasibility of LIL grows as dialect identification (Abdul-Mageed et al., 2024; Bouamor et al., 2018), stylometry (Stamatatos, 2009), and LLM inference (Staab et al., 2024) improve. Consent assurances that "personal information will be removed" are incomplete when linguistic variety constitutes a quasi-identifier (Sweeney, 2002). Governance should include tiered-access controls, releasing anonymized versions publicly while restricting dialectally rich subsets to vetted researchers under data use agreements, and Datasheets for Datasets (Gebru et al., 2021) should include LIL risk disclosures.

9 Future Research Directions

Our paper is conceptual and does not provide empirical measurements of LIL magnitude, but several empirical directions follow directly. First, quantitative studies should measure residual identity inference from anonymized NLP datasets across languages and domains using LLM-based adversary simulation (Staab et al., 2024) on topic-matched splits. Second, L-PII suppression methods need evaluation protocols measuring both privacy gain and utility on downstream tasks, using existing DP rewriting approaches (Igamberdiev and Habernal, 2023; Mattern et al., 2022) as baselines. Third, evaluation benchmarks should assess whether demographic profiling capability is a general property of pretrained multilingual models. Fourth, the homogenization hypothesis (Section 8) should be tested directly. Finally, a more systematic taxonomy construction in the sense of Nickerson et al. (2013), with empirical-to-conceptual iteration over an annotated corpus, would put the five-category structure on a stronger methodological footing.

10 Conclusion

We have introduced Linguistic Identity Leakage (LIL) as a privacy risk category systematically underaddressed in NLP, proposed Linguistic Personally Identifiable Information (L-PII) with a five-category taxonomy, grounded the framework in three illustrative scenarios, and proposed the Identity Inference Risk (IIR) framework as an operational auditing tool. Removing names and addresses from a text is not sufficient to protect privacy when dialect, morphology, and stylometric patterns still encode the author's origin and social background. This is not theoretical: Staab et al. (2024) demonstrate that LLMs already infer these attributes at scale on English social media with accuracy that standard mitigation strategies cannot defeat.

Key Takeaways. The following points summarize our contribution. Points (1) and (2) reinforce existing findings from sociolinguistics and stylometry, but make explicit their consequences for NLP privacy practice; points (3) and (4) are framework-level recommendations specific to LIL. **(1)** Text anonymization must go beyond NER-based redaction; dialect, morphology, and stylometric patterns are invisible to span-level de-identification. **(2)** Dialect, sociolect, and stylometric signals must be treated as quasi-identifiers in dataset documentation. **(3)** Dataset releases should include an IIR linguistic privacy audit reporting per-category inference risk scores. **(4)** LLM privacy evaluations should test demographic inference from linguistic structure, not only memorization and membership inference.

Limitations

This paper is conceptual and does not present experimental results. The taxonomy of L-PII features is based on existing research in sociolinguistics, stylometry, code-switching, and computational dialect identification, and is not exhaustive; additional categories may exist beyond those addressed here, and the categories were derived through conceptual synthesis rather than the more systematic taxonomy construction of Nickerson et al. (2013). The Arabic examples are drawn from the existing literature (Abdul-Mageed et al., 2024; Bouamor et al., 2018) and are intended as illustrations rather than empirical claims about identifiability rates.

The IIR framework is a design sketch rather

than an implemented system; practical deployment would require empirical calibration of inference risk estimates for each L-PII category. We do not address the formal privacy properties of any proposed mitigation strategy. While Staab et al. (2024) provide strong evidence that LLMs already constitute a practical LIL adversary for English-language social media text, the degree of this risk for other languages and domains remains an open empirical question. The figures we cite from Goswami et al. (2026) are taken from a single recent survey and should be treated as preliminary evidence rather than a settled empirical fact. Finally, our discussion of how generative AI may homogenize style and reduce LIL is a hypothesis, not an empirical finding.

Ethical Considerations

The concepts introduced in this paper describe a privacy risk and are intended to support more robust privacy protection in NLP research. We are aware that a taxonomy of linguistic identity signals could be misused to de-anonymize individuals; we have deliberately refrained from providing system-level specifications that would lower that barrier. The mitigation directions we sketch are oriented toward protection, not exploitation.

The LIL framework has particular implications for research on marginalized linguistic communities. Dialect and code-switching data collected from such communities often carries higher LIL risk precisely because the communities are smaller and their linguistic features are more distinctive. NLP datasets collected from under-resourced language communities are already published and in active use, and the privacy risks we describe apply to them today. Researchers working with such data should consider LIL risk as part of their ethical review process and design governance policies accordingly.

We encourage Datasheets for Datasets (Gebru et al., 2021), Institutional Review Board protocols, and informed consent processes for linguistic data collection to be updated to reflect the residual privacy risks that persist after explicit identifier removal, so that data contributors can make genuinely informed decisions about their participation.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand, 2024. Association for Computational Linguistics. https://aclanthology.org/2024.arabicnlp-1.79

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3387–3396, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). https://aclanthology.org/L18-1535

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, Berlin, Heidelberg, 2006.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM 2020)*, pages 178–186. ACM, 2020. https://dl.acm.org/doi/10.1145/3336191.3371856

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. https://dl.acm.org/doi/10.1145/3458723

Dhiman Goswami, Jai Kruthunz Naveen Kumar, and Sanchari Das. NLP privacy risk identification in social media (NLP-PRISM): A survey. *arXiv preprint arXiv:2602.15866*, 2026. https://arxiv.org/abs/2602.15866

Clive Holes. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press, Washington, D.C., revised edition, 2004.

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

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. Differential privacy in natural language processing: The story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States, 2022. Association for Computational Linguistics. https://aclanthology.org/2022.privatenlp-1.1

William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, PA, 1972.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 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, Online, 2021. Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.323

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States, 2022. Association for Computational Linguistics. https://aclanthology.org/2022.findings-naacl.65

Carol Myers-Scotton. *Social Motivations for Codeswitching: Evidence from Africa*. Clarendon Press, Oxford, 1993.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):86:1–86:36, 2017. https://dl.acm.org/doi/10.1145/3132039

Robert C. Nickerson, Upkar Varshney, and Jan Muntermann. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359, 2013. https://doi.org/10.1057/ejis.2012.26

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4): 1053–1101, 2022. https://aclanthology.org/2022.cl-4.19

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03*, pages 199–205, Stanford, CA, 2006. AAAI.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. https://openreview.net/forum?id=kmn0BhQk7p

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. https://onlinelibrary.wiley.com/doi/10.1002/asi.21001

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. https://dl.acm.org/doi/10.1142/S0218488502001648

Wanyue Zhai, Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. A girl has a name, and it's ... adversarial authorship attribution for deobfuscation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7372–7384, Dublin, Ireland, 2022. Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.509