A Survey on Representing Linguistic Style: Challenges and Opportunities

Anna Wegmann
Utrecht University
a.m.wegmann@uu.nl

Cristina Aggazzotti
Johns Hopkins University

Rafael Rivera Soto
Johns Hopkins University

Dong Nguyen
Utrecht University

Abstract

Although representation learning has transformed semantic modeling in NLP, representations of linguistic style remain underexplored—partly due to conflicting definitions of style within and beyond NLP or unclear immediate advantages of separate style representations. In this survey, we provide an overview of style conceptualizations in NLP and (socio-)linguistics and suggest a definition of style for practitioners. Guided by this definition, we then review methods for creating and evaluating style representations. We conclude by discussing how style representations can make crucial contributions to the modern NLP pipeline (e.g., in dataset curation or evaluation) and to the application of NLP methods in other fields. We curate relevant datasets, models, and tools on our companion webpage at `https://anonymous.4open.science/w/StyleSurvey-60B0`. Throughout, we provide practical guidance, identify open research questions with concrete recommendations for addressing them and issue calls to action on critical gaps in current work.

1 Introduction

The Lego Grad Student[1] posted in July 2020:

> *Videoconferencing from his apartment with his advisor, the grad student feels like the victim of a home invasion.*

Now consider a rephrasing by GPT-5.5 using the Wikipedia-style prompt from Maini et al. (2024):

> *While videoconferencing with his advisor from his apartment, the graduate student feels as though his private home has been intruded upon.*

The style of the original post (e.g., more informal, compact) likely contributed to the 3k likes it received. *Style*, or how something is said, can affect a reader's perception as it can, e.g., influence engagement (Banerjee and Urminsky, 2025; Munaro et al., 2024) and change the persuasiveness of arguments (El Baff et al., 2020; Parhankangas and Renko, 2017). Similarly, style influences human quality ratings of generated content (Cai et al., 2024; Wu and Aji, 2025), including preference ratings on platforms like LMArena (Li et al., 2024). Thus, *if you care about LLMs, then style matters.*

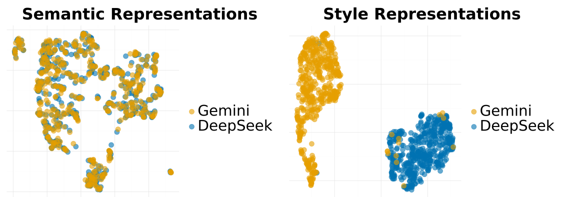

Figure 1: We compare reasoning traces from Muennighoff et al. (2025), generated with Gemini and DeepSeek for the same math problems. Style representations can distinguish between the two models—confirming results in Rivera Soto et al. (2023)—while semantic representations overlap. See §B.1 for details.

However, style is often disregarded in NLP. As a result, LLMs are brittle across style-like features: Rephrasing prompts in different styles can yield different reply qualities (Cao et al., 2026; Mizrahi et al., 2024; Wahle et al., 2024); LLM judges prefer long, formal, synthetic texts over relevance (Cao, 2025; Feuer et al., 2025; Wu and Aji, 2025); machine translations sound older and more male than the original (Hovy et al., 2020); models are biased against non-majority varieties (Fleisig et al., 2024; Hofmann et al., 2024; Liang et al., 2023) and perform worse on non-standard spellings (Ebrahimi et al., 2018; Li et al., 2019), simple and informal styles, and genres like poetry (Anschütz et al., 2025; Cao et al., 2026; Cao, 2025; Qi et al., 2021; Zhao et al., 2025). This brittleness might increase as we train on more synthetic data (Guo et al., 2024).

Style representations (i.e., vectors with entries that are optimized for style information) can help: They can improve model robustness by supporting the curation of stylistically diverse (post-)training

[1] An online creator on Twitter and Instagram that posts LEGO figures playing out scenes in a grad student's life.

datasets, support text generation in and evaluate adherence to a target style, help machine text detection, and enable new tasks (e.g., retrieval of documents in a target style). In the social sciences and humanities, they can support the analysis of literary texts and style dynamics in dialogue. A detailed discussion of these opportunities follows in §5.

Semantic representations are often also sensitive to style as word prediction tasks also need style information (Goldberg, 2019; Nguyen and Grieve, 2020; Miaschi et al., 2020; Tenney et al., 2018; Wegmann and Nguyen, 2021). However, semantic representations only show limited sensitivity to style (e.g., Mickus and Copot, 2024; Zhang et al., 2023b), in part because they are trained to prioritize semantic information, making it difficult to study style separately from content (cf. Fig. 1). Common benchmarks also do not include style-related tasks (Enevoldsen et al., 2025; Muennighoff et al., 2023)

In contrast to semantic representations, few community-vetted and broadly tested methods exist for representing the style of texts. **The main goals of this paper are** to promote the wider adoption of general linguistic style representations, survey existing work on them, and highlight key challenges and research directions. **We contribute**

- an overview of style definitions in linguistics and NLP, and our own unifying definition (§2)
- an overview of methods for representing (§3) and evaluating (§4) style representations
- a discussion of why style representations are useful for modern NLP and other fields (§5)
- open questions (❓), calls to action (📣), practical recommendations (🧭), and resources (🔧) including a companion webpage (🌐) collecting, among others, datasets and tools.

Style representation work has mostly focused on English, so naturally, so does most of the work we discuss. That said, approaches to create and evaluate representations largely transcend languages provided architectural components (e.g., tokenizers), evaluation datasets, and predefined features are adapted for optimal performance. We highlight applicability to and work in other languages, and, 📣 urge NLP researchers to study more languages.

Concurrent work This work surveys the theoretical background and existing work on style representations and highlights key challenges, such as the limitations of current benchmarks. Concurrent follow-up work (Anonymous, 2026) introduces a benchmark that addresses some of these limitations.

2 Style conceptualizations

> Linguists often define style as a distinctive pattern in language for some object of study (e.g., for an author or group), while NLP researchers often use "style" more loosely.

2.1 Style in linguistics

Researchers working with style often aim to describe a text's structural linguistic features (how something is said) rather than its semantic meaning[2] (what is said). This separation is contested within linguistics, where style and content are often seen as intertwined (cf. §C). Style can alternatively be understood as what makes a phrasing distinctive within a set of possibilities (Irvine, 2001), for instance, how speakers' linguistic choices reflect external factors like social background, identity, or register. Overall, we emphasize that *style is an elusive term that has been defined in many different, sometimes inconsistent, ways in linguistics (and other fields*, cf. §C). It overlaps with related terms like register, genre, and idiolect discussed below.

What are the objects of study? Style is usually studied through comparison; however, what is being compared varies. In (socio-)linguistics, style has been studied as different types of patterned variation: inter-individual variation—the idiosyncratic choices that potentially distinguish individuals, often referred to as their *idiolect* (Coulthard, 2004); intra-individual variation (Bell, 1984; Irvine, 2001; Labov, 2006; Meyerhoff, 2006; Wagner, 2025)—language patterns by the same speakers that differ across situations or contexts; and inter-group variation—language use by people in the same social group (Bell, 1984; Eckert, 2008; Irvine, 2001; Kristiansen, 2024). As an example of inter-group variation, *g*-dropping (*going* vs. *goin'*) may signal a person's association with a southern U.S. region (Campbell-Kibler, 2007). Genres (or domains) and registers[3] have also been objects of style research (Biber and Conrad, 2019;

[2] Or: referential meaning (Campbell-Kibler, 2011; Labov, 1972; Lavandera, 1978; Nguyen et al., 2016, 2021). Two variants have the same referential meaning if they are the same in a truth-conditional sense (i.e., true in exactly the same situations), while the social or stylistic significance might differ considerably (Labov, 1972; Weiner and Labov, 1983).

[3] *Genre* usually refers to conventionalized text types with specific structural and stylistic norms (e.g., journal articles with an abstract, introduction, etc.), while *register* describes language variation based on communicative purpose and context (e.g., academic prose that lacks rigid structure and is meant to communicate a thought/finding) (Biber and Conrad, 2019).

Grieve, 2023). Literature from a historical period, novels by a specific author, and blogs can display very different linguistic patterns, which might be called the style of a historical period, literary author, or blog (Biber and Conrad, 2019; Grieve et al., 2011; Hicke and Mimno, 2025; Irvine, 2001). More objects have been considered than we discuss, like the communication environment (e.g., speech in a courtroom in Ervin-Tripp, 2001) or the communicative manner (spontaneous vs. read speech in Williams and King, 2019). Scholars also study combinations of these objects (e.g., Tweets by one individual) or an object in specific contexts (e.g., a social group discussing a certain topic). For example, Holliday (2021) finds that biracial Black men displayed fewer African American[4] intonational features when discussing police narratives.

What is the function of style? Style can be studied through the function it serves. Some scholars argue that style is fundamentally embedded in social meaning, indexing social background and shaping social identity (Campbell-Kibler et al., 2006; Coupland, 2007; Eckert, 2008, 2012). For example, Labov (1972) found that differences in the pronunciation of /r/ correlated with social class, and Eckert (1989) found that self-identified "burnouts" at a Detroit school used more non-standard linguistic features (e.g., *gonna*) than college-bound "jocks" (e.g., *going to*). Labov originally viewed a speaker's vernacular as a reflection of their social identity, not an active choice (Labov, 1972), but more recent sociolinguistic approaches see style as more agentive, not only reflecting identity but performing and constructing it (Eckert, 2012). For example, the development of linguistic practices of trans activists can be tied to agency in creating their identity (Zimman, 2019) and speakers may choose styles for performative functions like getting attention (Ervin-Tripp, 2001). Style can serve communicative functions in an interaction (Coupland, 2007): Speakers may accommodate to or distance themselves from the style of interlocutors or audiences (Bell, 1984; Giles and Powesland, 1975; Giles et al., 1991; Khaleghzadegan et al., 2024), thereby shaping social relationships and interactions (Coupland, 2007). For example, police negotiators who match the linguistic style of hostage takers are more successful (Taylor and Thomas, 2008).

2.2 Style in NLP

Some work in NLP uses the term style in ways broadly consistent with linguistics, aiming to study formal/informal styles and literary authorial styles (e.g., Jhamtani et al., 2017; Rao and Tetreault, 2018; Wegmann and Nguyen, 2021); however, others increasingly use style as an umbrella term for general attributes of texts that vary across datasets (Jin et al., 2022) such as the sentiment of a text (Reif et al., 2022; Shen et al., 2017), but do not necessarily align with a typical linguistic definition of style.

Separating content and style As in linguistics, work in NLP finds that content and style are often correlated (Mikros and Argiri, 2007; Wang et al., 2023a), and learned representations frequently entangle the two (John et al., 2019; Man et al., 2026). Still, separating style and content tends to be a natural distinction for many NLP applications. For example, NLG systems intuitively have to determine what information to convey—the knowledge, or message—and what style to generate it in (Gatt and Krahmer, 2018). While neural NLG systems often handle content and style implicitly, generating texts end-to-end without explicit planning stages, the style-content distinction remains useful in practice, for example, when rephrasing and adapting texts, or evaluating the factual correctness of model outputs (cf. §5).

How to define style for practitioners We propose a definition of style[5] that integrates existing linguistic perspectives while clearly mapping onto NLP operationalizations (cf. §3 & §4):

> **Definition** A linguistic style consists of ***distinctive patterns in language use*** for an object of study (e.g., individuals, a group of authors, a given register) in its lexical, syntactic, morphological, orthographic, discourse, phonetic, etc. composition. The most distinctive patterns should not chiefly encode, but can correlate with, semantic meaning.

Consider an individual speaker as our object of study. When they shift from discussing American football to ballroom dance, the topic changes and

[4] While some features describe both, styles and dialects, dialects are typically distinguished from styles as language variation tied to speakers' social backgrounds and geographic regions (Biber and Conrad, 2019; Grieve et al., 2025). Nonetheless, some treat dialects as a "social style" (Coupland, 2007).

[5] Our definition does not exclude concepts like dialects, registers, or varieties for practical reasons: (i) the separation between such terms is inconsistent in linguistics, and (ii) computational style representations are often expected to be sensitive to dialect, register, and other variety information (cf. §4).

some linguistic features might change with it—e.g., they might talk more casually about football than ballroom dance. Yet some underlying linguistic features that remain consistent in both situations may be distinctive to the speaker, and carry social meaning (e.g., about the speaker's upbringing, cf. §2). If we instead take the informal register as our object of study, a different pattern counts as style: We might focus on what the casual football discussion has in common with other informal conversations. Following our definition, we survey methods for creating and evaluating style representation along three axes in the remainder of the paper: the object of study, the considered linguistic features and the relationship to semantic meaning. We consistently use the same colors for the same concepts.

3 Representing style

> Linguistic style is usually operationalized with patterns in linguistic features like function words or automatically-learned representations like neural text representations.

3.1 Predefined features

Style is often operationalized as systematic variation of linguistic features across linguistic levels, including morphology, orthography, syntax, and discourse (Biber and Conrad, 2019; Crystal and Davy, 1969; Grieve, 2007; Kniffka, 2007; Labov, 1972; Neal et al., 2017; Stamatatos, 2009). Predefined features are appealing because they are supported by linguistic theory, well-tested and interpretable. They can be used with statistical approaches like logistic regression or dimensionality reduction with factor analysis to determine how important each feature is. This transparency is especially important in high-stakes settings, like forensic linguistics, where explaining a model's decision is crucial (Argamon, 2018; Grant, 2022).

Stylometry A prominent feature-based approach is stylometry, which uses the frequencies of linguistic features to discriminate between author styles. There is no fixed set of features that work for every individual, despite much work attempting to find one (Juola, 2006; Nini, 2023); instead, the features often depend on the nature of the data (e.g., genre, register, amount of data, language) (Argamon, 2018). Still, function words and character n-grams, in particular, have proven effective at discriminating authors (Grieve, 2007; Houvardas and Stamatatos, 2006; Kestemont, 2014; Mosteller and Wallace, 1963; Nini et al., 2026; Peng et al., 2003) and speakers (Aggazzotti and Smith, 2025; Aggazzotti et al., 2024; Doddington, 2001; Sergidou et al., 2023; Tripto et al., 2023). N-grams (including characters, tokens/words, and part-of-speech tags) can work across many languages and can be particularly effective for morphologically-rich ones (HaCohen-Kerner et al., 2026; Puspitasari, 2026).

MDA Other feature operationalizations serve different purposes. For example, Multidimensional Analysis (MDA) (Biber, 1988) is used to determine how texts differ in their communicative function and originally relied on mostly grammatical category-related features (e.g. nouns, verbs); however, modern extensions (e.g., Clarke and Grieve, 2017; Grieve et al., 2011; Xu et al., 2026) additionally include more complex features, such as syntactic constructions and semantic classes.

🧭 **Choosing predefined features** There is no fixed, general set of most discriminative features, although n-grams are often used due to ease of implementation and coverage across linguistic levels. The right features depend on various factors, including the data (e.g., genre, length), task (e.g., authorship, machine text detection), and required computational efficiency. There are numerous feature extractions tools available, some of which have been developed specifically for such factors. We recommend reviewing previous work related to your data/task and testing different features on a held-out validation subset to determine which features discriminate best. See App. Tab. 2 for examples and 🔧 §D and 🌐 for feature extraction tools.

3.2 Automatically-learned features

By automatically-learned features or embeddings, we mean vector representations of text produced by (usually neural) models. Unlike predefined features, automatically-learned features do not rely on hand-crafted features but automatically discover style patterns. They tend to outperform predefined features on downstream tasks, but are usually less interpretable. Because style is difficult to operationalize, models are usually optimized on proxy downstream tasks, such as authorship verification or style transfer. 🔧 §D and 🌐 for model URLs.

Authorship verification The dominant approach to date trains models with a contrastive objective (Dong and Shen, 2018; Khosla et al., 2020) to learn representations where texts by the same author are close together in vector space and texts by different authors are far apart (Agarwal et al., 2025; Andrews and Bishop, 2019; Fincke and Boschee, 2024; Huertas-Tato et al., 2024; Khan et al., 2021; Kim et al., 2025; Man and Huu Nguyen, 2024; Rivera Soto et al., 2021; Sawatphol et al., 2022; Thakrar et al., 2025; Wang et al., 2023b; Wegmann et al., 2022). Such representations have been shown to encode style information (Wang et al., 2023b; Wegmann and Nguyen, 2021). Recently, Agarwal et al. (2025) extend the task to cross-genre authorship attribution (i.e., matching authors across domains like blogs and Reddit), fine-tuning a 12B decoder Mistral model instead of the 100M encoder RoBERTa model typical of prior work. Recent work extends style representations to multilingual settings: Qiu et al. (2025) cover 9 higher-resource languages (e.g., Arabic, German, Chinese), while Kim et al. (2025) train on 36 languages including low-resource languages like Kazakh and Georgian.

Content disentanglement Datasets may contain undesired correlations, for example, between style and content if an author only writes about one topic. To improve disentanglement, some work creates harder positive (same author) and negative (different author) pairs (Agarwal et al., 2025; Fincke and Boschee, 2024; Man and Huu Nguyen, 2024; Patel et al., 2025). Wegmann et al. (2022) use negatives roughly about the same topic, and Patel et al. (2025) supplement verification training data with synthetic paraphrases that vary along predefined features. Man et al. (2026) propose a novel disentanglement stage that follows contrastive pretraining where two distinct encoders for style and content are learned; their independence is enforced by training an LLM-based discriminator to predict whether latent pairs share the same author or topic, while also generating a natural language rationale for its decisions. In multilingual settings, negative pairs must be carefully constructed to avoid learning only trivial cross-lingual differences, for example, by ensuring negatives share the same language (Kim et al., 2025; Qiu et al., 2025).

Style transfer Another line of work learns representations as a byproduct of style-transfer, aiming to rewrite text to change a stylistic attribute without altering its content (Cheng et al., 2020b; John et al., 2019; Shen et al., 2017; Zhu et al., 2024). For example, a model trained to rewrite formal text as informal text—conditioned on the input and a learned embedding of the target style—results in embeddings with features indicative of informality. These methods usually rely on explicit style-content disentanglement and tend to learn representations that are more narrow in scope, often tied to single attributes (e.g., politeness) or differences between two corpora (Shen et al., 2017). John et al. (2019) train an auto-encoder to produce a style and a content vector, applying a style classification loss on the style and an adversarial style classification loss on the content vector, while Cheng et al. (2020b) minimizes the mutual information between the two.

LLM-guided embedding construction An increasing number of methods use LLMs to annotate or generate controlled training data (Kantharuban et al., 2026; Ma et al., 2025; Patel et al., 2023, 2025). LISA (Patel et al., 2023) stands out as it learns embeddings whose dimensions are *interpretable* features (e.g., active voice). It uses style annotations by GPT-3 to train EncT5 (Liu et al., 2022). Using LLMs for labelling offers a middle ground between predefined features (§3.1) and automatically-learned representations, but LLM brittleness on style may be a concern (§1).

Choosing a style representation We recommend models for different applications, cf. Tab. 5:

- For machine-text detection, LUAR has been used successfully (Rivera Soto et al., 2023).
- For authorship retrieval, LUAR is the standard for English and MSR extends to other languages (Kim et al., 2025).
- For style-transfer and disentangled style features, we recommend CISR (Wegmann et al., 2022) or StyleDistance, both trained for content-independence and evaluated on style benchmarks (Patel et al., 2025).

3.3 Challenges of style representations

We list open challenges and recommendations below; see §D.1 for more, including improved training and content-independence.

Finding available datasets While early representations were trained and evaluated on a single or few domains (Rivera Soto et al., 2021; Wegmann et al., 2022), recent work increasingly combines multiple domains (Huertas-Tato et al., 2024; Kim et al., 2025), though few include the same authors across domains (Ma et al., 2025). Common domains include e-mails, news, academic articles, blogs, social media, Q&A forums, reviews, Wikipedia, fanfictions, novels, and transcribed speech; several are combined in PAN shared tasks (Argamon and Juola, 2011; Bevendorff et al., 2025a) and training and evaluation setups across papers (Aggazzotti et al., 2024, 2025b; Huang et al., 2025; Huertas-Tato et al., 2024; Nini et al., 2026; Tyo et al., 2022). Non-English resources have only been used recently and mostly include Wikipedia, Reddit, or multilingual PAN tasks (Kantharuban et al., 2026; Kim et al., 2025; Israeli et al., 2025). Some common datasets (e.g., Reddit) are no longer accessible; we curate typical datasets with their availability and licensing (🌐). Without careful preparation, datasets might contain named entities, leakage between train and test sets (Brad et al., 2022; Sawatphol et al., 2024), or spurious topic correlations (Wegmann et al., 2022).

📣 **Define what you want to represent** Training representations on authorship verification implicitly defines style as the idiosyncrasies authors exhibit in certain corpora (Zhu and Jurgens, 2021). However, if the contrastive dataset never pairs authors from the same social group as negatives, representations may primarily encode group-level (e.g., discriminative of writers' native language in Fig. 2) rather than idiolectal features. Whether that is desirable depends on the intended use. We call on researchers to explicitly define their object of study (§ 2.2), use techniques like hard negatives to control for confounders (e.g., shared style in a social group), and evaluate whether embeddings primarily capture patterns for the intended object of study.

❓ **Extending to other languages** While n-grams can be applied across languages (§ 3.1), stylometric features must be adapted to each language. In NLP, predefined features have primarily been used for English, but some work develops them in languages like Chinese (Hou and Huang, 2019; Tang et al., 2019) and low-resource languages like Bengali (Hossain et al., 2020) and Urdu (Nazir et al., 2021). Further, while some embeddings have been learned for non-English languages (§ 3.2), they remain rare. Similar to semantic models, style embeddings for other languages face language-specific challenges like tokenization and fewer resources.

❓ **Building general-purpose embeddings** An open challenge is building general-purpose embeddings that are sensitive to multiple objects of study (e.g., individual, group, register, time) simultaneously. New training objectives could explicitly target multiple objects of study (e.g., recently, Kantharuban et al. (2026) target individual and dialect variation). Combining objectives might benefit from disentanglement approaches like minimizing the mutual information of two representations (Cheng et al., 2020a), adversarial training (John et al., 2019), or using VAEs (Bao et al., 2019; Chen et al., 2019). Such representations could eventually also capture multimodal and paralinguistic features like prosody and gesture.

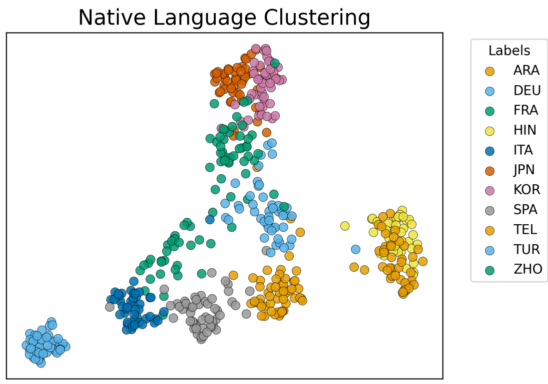

Figure 2: Stylistic representations trained on the "idiolectal" authorship verification task, cluster TOEFL (Test of English as a Foreign Language) essays by the native language of the writer. See §B.3 for more details.

❓ **Constructing interpretable embeddings** It is unclear how to learn representations with interpretable dimensions that are as performant as their uninterpretable counterparts. Promising directions include sparse autoencoders, which have recently been shown to automatically learn interpretable features (Huben et al., 2024), and further combining predefined features with neural training (Kantharuban et al., 2026; Patel et al., 2023).

4 Evaluating style representations

> To develop better style representations, we must be able to compare and evaluate them, but no standard currently exists.

On predefined features Representations (§ 3.2) can be evaluated on their sensitivity to predefined features (§ 3.1) using probing classifiers (Adi et al., 2017; Belinkov et al., 2017; Köhn, 2015; Shi et al., 2016). Alshomary et al. (2025b) probe style embeddings on morphology and syntax. However, probing has limitations, such as uncertainty over how to interpret classifier performance (Belinkov, 2022). Other approaches are sparse, but include studying performance loss on style tasks when removing style information (Anand et al., 2026; Zhu and Jurgens, 2021) or testing whether texts sharing

Model	Acc	TFLOPs	USD
StyleDistance	**.66**	≈ 13	0
GPT-5.2	.56	$\approx 89k$	1.61

Table 1: Style embeddings can outperform LLMs on authorship verification with fewer tera FLOPs. Acc = accuracy at threshold 0.8 for embeddings; USD = API cost in USD; see full setup in §B.4.

predefined features like contractions have higher cosine similarity (Patel et al., 2025; Wegmann and Nguyen, 2021)—this has recently been extended to multilingual settings (Qiu et al., 2025). Further, Terreau et al. (2021) test if representations can predict an author's distribution on 301 predefined features.

On objects of study Most works evaluate style representations on authorship attribution or verification tasks (Alkiek et al., 2025; Anand et al., 2026; Ding et al., 2019; Patel et al., 2025), and clustering of same-author texts (Hay et al., 2020; Huertas-Tato et al., 2024; Wegmann et al., 2022). They have also been evaluated on their ability to classify, distinguish or analyze other objects of study (§2) including: (i) literary or artificial authors (Hicke and Mimno, 2025; Man et al., 2026; Wang et al., 2023b), (ii) book genres (Hicke and Mimno, 2025; Maharjan et al., 2019), (iii) registers (Alkiek et al., 2025; Patel et al., 2025; Wegmann and Nguyen, 2021), (iv) dialects (Kantharuban et al., 2026), and (iv) author demographics like gender or age (Ding et al., 2019; Kang et al., 2019; Kang and Hovy, 2021). See §3.2 and 🌐 for available datasets.

Content-independence Content-independence is typically evaluated by studying performance changes on style tasks (e.g., authorship verification) when editing only the content (e.g., via masking "content words") or the style (e.g., via paraphrasing) of texts (Nini et al., 2026; Stamatatos, 2017; Wang et al., 2023b; Zhu and Jurgens, 2021). Other approaches test whether representations are more sensitive to style than to content changes (Patel et al., 2025; Qiu et al., 2025; Wegmann and Nguyen, 2021; Wegmann et al., 2022), whether they distinguish speakers discussing the same topic (Aggazzotti et al., 2024, 2025b), or whether they fail on semantic tasks like topic classification (Wang et al., 2023b). Generally, few representations score highly on content-independence (🔧 App. Tab. 5) and the field might benefit from more exhaustive content disentanglement (App. §D.1).

4.1 Challenges of style evaluation

We outline open challenges and recommendations in evaluating style representations. See App. §E.1 for more, like using measurement theory.

❓Developing standard benchmarks Researchers typically use varying combinations of tasks, data and domains, along with inconsistent splits, inaccessible datasets, and no or inconsistent style definition, making scores incomparable across publications. Few contributions aim to systematically evaluate representations on linguistic style, leaving the area behind semantic benchmarks like MTEB (Enevoldsen et al., 2025): Kang and Hovy (2021) collected the largest style classification dataset to date, but several of their classes (e.g., sentiment) would not be considered style by us. STEL (Wegmann and Nguyen, 2021) benchmarks sensitivity to single linguistic properties and registers via cosine similarities. Neither covers a wide range of styles or domains nor clearly defines an object of study (cf. §2.2). An open, accessible, high quality, and diverse style benchmark—probing sensitivity to predefined features, spanning multiple objects of study, and testing independence from content (see 🧭 below)—would be a significant contribution.

🧭 How to evaluate? We recommend combining evaluation approaches across our three axes (§2):

- Object of study: For authorship attribution, use PAN datasets for cross-publication comparability and your own dataset for domain-specific performance. For sensitivity to common style registers, use STEL or (multilingual) SynthSTEL (Patel et al., 2025; Qiu et al., 2025; Wegmann and Nguyen, 2021).
- Predefined features: Use probing approaches with tools like LFTK (Lee and Lee, 2023).[6]
- Content-Independence: Use the style-or-content variant of the STEL tasks (Patel et al., 2025; Qiu et al., 2025; Wegmann et al., 2022).

❓ Treat style as relative Style is inherently relative (Irvine, 2001). For example, it might be clearer or more relevant to judge if a text (e.g., *How are you?*) is more formal than another (e.g., *What's up?*) rather than if it is formal in isolation; see also App. Fig. 3. This may require new solutions in training and evaluation—for example, curating training data with hard positives or negatives positioned in

[6] LFTK is common, but no single tool is best (§3.1).

relation to each other or testing if representations correctly rank sentences along a style dimension.

5 Opportunities with style embeddings

> Style representations can make crucial contributions to the modern NLP pipeline and to applications of NLP methods.

We provide a selection of applications enabled by style representations with more in § F (e.g., authorship attribution, improving generalization). While generative models could also be used, style representations—similar to semantic embeddings (Enevoldsen et al., 2025; Warner et al., 2025)— can outperform much larger LLMs on discriminative tasks at a fraction of the computational cost. We demonstrate this for style embeddings in Tab. 1.

Diversify style in evaluation datasets Both style and content influence human preference judgments (Cai et al., 2024; Chen et al., 2024a; Li et al., 2024; Singhal et al., 2024). However, state-of-the-art performance is often established only on content tasks (mostly NLU and reasoning) using texts with limited stylistic variation (Guo et al., 2025; Truong et al., 2025), which may obfuscate a model's ability to generalize to other domains or generate diverse or preferred styles (Maini, 2023).With style representations, benchmarks could be composed based not only on what they test, but also on whether their datasets or tasks cover different or expected regions of the embedding space.

Style transfer & personalization Style representations can help generate text in a specific style (Zhang et al., 2023a) by providing a target style vector to condition (Ficler and Goldberg, 2017; Horvitz et al., 2024b,a) and evaluate generation models on (Chim et al., 2025; Horvitz et al., 2024b,a; Jangra et al., 2025; Liu et al., 2023). For personalization, style embeddings could be used to recognize the style of individuals and infer their preferences before adapting generated text to them (Liu et al., 2023, 2025; Zhang et al., 2025). Other applications include paraphrasing (Horvitz et al., 2024a), increasing accessibility (e.g., simplifying a text for a non-expert) (Anschütz et al., 2025; Cao et al., 2020; Surya et al., 2019), and style preservation in machine translation (Michel and Neubig, 2018; Rabinovich et al., 2017).

Curate diverse (multi-stage) training datasets Manipulating and diversifying the style of texts in in-context learning examples as well as pre- and post-training datasets—for example, by balancing across registers or rephrasing in different styles—has increased output diversity and performance across stylistic variation (Chen et al., 2024b; Lambert et al., 2025; Levy et al., 2023; Maini et al., 2024). Curriculum learning (or multi-stage training) has found renewed success recently (Allal et al., 2025; Ettinger et al., 2025; Walsh et al., 2025), but how stylistic diversity should vary across stages remains unclear. Style representations could be used to monitor the overall stylistic diversity of a dataset (Nguyen and Ploeger, 2025), rephrase training data in diverse styles (cf. style transfer), and select data points for training that increase or decrease stylistic diversity along a curriculum.

Machine text detection Recent work (Bevendorff et al., 2025b; Elkhatat et al., 2023; Gehrmann et al., 2019; Kumarage et al., 2023; Sun et al., 2025; Uchendu et al., 2020) shows that LLMs exhibit idiosyncrasies that distinguish their writing from human writing. Detectors that use style embeddings have been effective in in-domain and cross-domain settings (Kim et al., 2025; Rivera Soto et al., 2023).

Privacy On the flip side of attribution and detection is the task of obfuscating someone's identity (Potthast et al., 2018). Style representations can help determine if text that has been anonymized, such as via paraphrasing, sufficiently removes someone's style and protects their privacy (Aggazzotti et al., 2025a; Alperin et al., 2025; Bao and Carpuat, 2024; Shokri et al., 2025).

6 Conclusion

We hope to have demonstrated the potential of style representations for the NLP community. We believe that style representations, just as semantic embeddings, can become foundational across fields, with broad applications like retrieving documents with a (dis-)similar style to a query (Cao, 2025), tracking style shift in dialogue in sociolinguistics (Nguyen, 2025), or analyzing literary text in the digital humanities (Hicke and Mimno, 2025), with current embeddings already seeing significant adoption.[7] 📣 We call on researchers to use clearer definitions of style, create representations for languages other than English, more explicitly disentangle evaluation and training approaches, and develop evaluation methods into a standard.

[7] CISR counted >200k downloads in 01/2025 (bluesky).

Limitations

Style definitions are English-centric. We mainly discuss definitions of style considered by American scholars (cf. §2.1), and we give examples of predefined features mainly for English (cf. §3.1)—for instance, "*g*-dropping" is an English-specific marker. Different scripts and languages will usually need different predefined features and have a different history regarding style definitions and sociolinguistic research (see also Ball et al., 2023). While we include some considerations of predefined features for other languages, we leave considerations of other traditions of style definitions and detailed discussions of predefined features in other languages for future work.

We consider style mostly for text. Many of the examples and citations throughout this paper refer to text-based style since the limited style research in NLP has focused on written language, but linguistic style also manifests, and is perhaps better studied, in other modalities like speech (e.g., tone of voice), gestures, and vision (e.g., image generation). We leave considerations for representing style in other modalities for future work.

Why not use a different term instead of style?

> *...the extremely broad and ambiguous reference of the term [style] in everyday use has not made its status as a technical linguistic term very appealing.*
>
> — David Crystal

Scholars, such as Crystal (2011), have argued against using the term style at all due to its increasingly vague and colloquial use. Instead, researchers have opted to describe the specific phenomenon they are interested in (e.g., syntactic variation) and use less over-defined terms (e.g., language variation). While that can be helpful in some cases, we argue that using the term style is still worthwhile because (i) the term is used regularly in NLP (with 200 publications in the ACL Anthology mentioning "style" in the title or abstract in 2024) highlighting the interest in the term; (ii) style seems to provide a more concise and intuitive label than alternatives like "distinctive patterns in the used language varieties" or "systematic variation in textual features"; and (iii) the term style can draw from decades of theoretical foundation in stylometry and sociolinguistics.

Style is a concept used in many fields. Why focus on the ones discussed in the paper?

> *[...] the most immediate problem to be solved in the attack on sociolinguistic structure is the quantification of the dimension of style.*
>
> — William Labov in Labov (1972)

Next to NLP, we focus on definitions and concepts of style used in sociolinguistics, linguistics, stylometry, forensic linguistics, and corpus linguistics (§2; see an overview of the fields in §C). To the best of our knowledge, these are the most active areas already using, or intuitive areas that could profit from using, computational methods for analyzing style. Further, we believe that sociolinguistics is particularly relevant to consider, as its study of the interaction between language and society has unique potential to inform NLP methods (Nguyen, 2025), especially as NLP models are increasingly used within, and have growing impact on, society.

Ethical considerations

Style modeling is closely related to *author profiling* (cf. §4)—the task of recovering author characteristics based on the text they wrote (Nguyen et al., 2013; Rangel et al., 2013). Note that author profiling can be useful for improving performance on some NLP tasks (Hovy, 2015); however, identifying an author's gender, age, personality type, etc. has increasingly been criticized for bias and privacy concerns (Brennan et al., 2012; Elazar and Goldberg, 2018; Li et al., 2018; Lison et al., 2021).

Integrating more language diversity, and with it social factors, into NLP is a double-edged sword: There are clear advantages to integrating more diversity into NLP models and, specifically, representing minorities to increase the fairness and representativeness of NLP models (Bird and Yibarbuk, 2024; Grieve et al., 2025; Hovy and Yang, 2021; Markl et al., 2024); however, making NLP models more sensitive to social factors could also make them a threat to privacy across social groups. The performance of machine learning approaches on tasks like author profiling could increase, resulting in a large potential for misuse, such as the following examples: (1) Author profiles could be used to identify and profile individuals or political dissenters (Hovy and Spruit, 2016); (2) Author profiling could be used for predatory ad targeting, which might show gambling ads to vulnerable groups or not show job postings to certain social groups (Dudy et al., 2021); and (3) Author profiles could lead to data leakage, such as making health conditions recoverable for insurance companies that might increase their rates for certain individuals (Dudy et al., 2021).

This conflict between privacy and fairness has been described as one of the "dual-use problems" in NLP by Hovy and Spruit (2016). We aim to improve fairness without compromising individual privacy and safety but acknowledge that progress in one might sometimes come at the expense of the other. 📣 Therefore, we encourage researchers in the NLP community to engage with the dual-use problem more actively and work on techniques to make the design of language models more sensitive to human values, as suggested in Dudy et al. (2021), ideally without actively working on approaches to make sensitive data recoverable from texts. We further encourage researchers to actively anonymize datasets used for modeling and the evaluation of style representations.

We confirm that we have read and abide by the ACL Code of Ethics. Besides those mentioned, we do not foresee immediate risks of our work.

References

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

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations (ICLR)*.

Shantanu Agarwal, Joel Barry, Steven Fincke, and Scott Miller. 2025. Cross-genre authorship attribution via llm-based retrieve-and-rerank. *Preprint*, arXiv:2510.16819.

Cristina Aggazzotti, Nicholas Andrews, and Elizabeth Allyn Smith. 2024. Can authorship attribution models distinguish speakers in speech transcripts? *Transactions of the Association for Computational Linguistics*, 12:875–891.

Cristina Aggazzotti, Ashi Garg, Zexin Cai, and Nicholas Andrews. 2025a. Content anonymization for privacy in long-form audio. *arXiv preprint*. ArXiv:2510.12780 [cs].

Cristina Aggazzotti and Elizabeth Allyn Smith. 2025. A stylometric analysis of speaker attribution from speech transcripts. *Preprint*, arXiv:2512.13667.

Cristina Aggazzotti, Matthew Wiesner, Elizabeth Allyn Smith, and Nicholas Andrews. 2025b. The impact of automatic speech transcription on speaker attribution. *Transactions of the Association for Computational Linguistics*, in press.

Kenan Alkiek, Anna Wegmann, Jian Zhu, and David Jurgens. 2025. Neurobiber: Fast and interpretable stylistic feature extraction. *arXiv preprint*. ArXiv:2502.18590 [cs].

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son Nguyen, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, and 3 others. 2025. SmolLM2: When smol goes big — Data-centric training of a fully open small language model. In *Second Conference on Language Modeling (COLM)*.

Kenneth Alperin, Rohan Leekha, Adaku Uchendu, Trang Nguyen, Srilakshmi Medarametla, Carlos Levya Capote, Seth Aycock, and Charlie Dagli. 2025. Masks and mimicry: Strategic obfuscation and impersonation attacks on authorship verification. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 102–116, Albuquerque, USA. Association for Computational Linguistics.

Milad Alshomary, Narutatsu Ri, Marianna Apidianaki, Ajay Patel, Smaranda Muresan, and Kathleen McKeown. 2025a. Latent space interpretation for stylistic analysis and explainable authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1124–1135, Abu Dhabi, UAE. Association for Computational Linguistics.

Milad Alshomary, Nikhil Reddy Varimalla, Vishal Anand, Smaranda Muresan, and Kathleen McKeown. 2025b. Layered insights: Generalizable analysis of human authorial style by leveraging all transformer layers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10290–10303, Suzhou, China. Association for Computational Linguistics.

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin CM 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.

Vishal Anand, Milad Alshomary, and Kathleen McKeown. 2026. iBERT: Interpretable Embeddings via Sense Decomposition. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1413–1429, Rabat, Morocco. Association for Computational Linguistics.

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1684–1695, Hong Kong, China. Association for Computational Linguistics.

Anonymous. 2026. STEB: Style text embedding benchmark. *Under Review*.

Miriam Anschütz, Anastasiya Damaratskaya, Chaeeun Joy Lee, Arthur Schmalz, Edoardo Mosca, and Georg Groh. 2025. (Dis)improved?! How simplified language affects large language model performance across languages. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 847–861, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Tatsuya Aoyama and Nathan Schneider. 2022. Probeless probing of BERT's layer-wise linguistic knowledge with masked word prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 195–201, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ehsan Arabnezhad, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, and Julinda Stefa. 2020. A light in the dark web: Linking dark web aliases to real internet identities. In *Proceedings of the 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 311–321, Singapore, Singapore. Institute of Electrical and Electronics Engineers.

Shlomo Argamon. 2018. Computational forensic authorship analysis: Promises and pitfalls. *Language and Law/Linguagem e Direito*, 5(2):7–37.

Shlomo Argamon and Patrick Juola. 2011. Overview of the international authorship identification competition at PAN-2011. In *Notebook Papers of CLEF 2011 Labs and Workshops*, Amsterdam, Netherlands.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, and 5 others. 2022. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11699–11732.

Martin J. Ball, Rajend Mesthrie, and Chiara Meluzzi. 2023. *The Routledge Handbook of Sociolinguistics Around the World*, 2nd edition. Routledge, London, UK.

Akshina Banerjee and Oleg Urminsky. 2025. The language that drives engagement: A systematic large-scale analysis of headline experiments. *Marketing Science*, 44(3):566–592.

Calvin Bao and Marine Carpuat. 2024. Keep it Private: Unsupervised privatization of online text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8678–8693, Mexico City, Mexico. Association for Computational Linguistics.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019, Florence, Italy. Association for Computational Linguistics.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

Janek Bevendorff, Daryna Dementieva, Maik Fröbe, Bela Gipp, André Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, Artem Shelmanov, Efstathios Stamatatos, Benno Stein, Yuxia Wang, Matti Wiegmann, and Eva Zangerle. 2025a. Overview of PAN 2025: Generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection. In *Advances in Information Retrieval*, pages 434–441. Springer, Cham.

Janek Bevendorff, Matti Wiegmann, Emmelie Richter, Martin Potthast, and Benno Stein. 2025b. The two paradigms of LLM detection: Authorship attribution vs. authorship verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3762–3787, Vienna, Austria. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge University Press, Cambridge, UK.

Steven Bird, Ewan Klein, and Edward Loper. 2019. *Natural Language Processing with Python*. O'Reilly Media.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2014. ETS corpus of non-native written English LDC2014T06. Web Download. Philadelphia: Linguistic Data Consortium.

Florin Brad, Andrei Manolache, Elena Burceanu, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. Rethinking the authorship verification experimental setups. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5634–5643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. In *ACM Transactions on Information and System Security (TISSEC)*, volume 15, pages 1–22, New York, USA. Association for Computing Machinery.

John Burrows. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Na Cai, Shuhong Gao, and Jinzhe Yan. 2024. How the communication style of chatbots influences consumers' satisfaction, trust, and engagement in the context of service failure. *Humanities and Social Sciences Communications*, 11(1):687.

Kathryn Campbell-Kibler. 2007. Accent, (ING), and the social logic of listener perceptions. *American Speech*, 82(1):32–64.

Kathryn Campbell-Kibler. 2009. The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156.

Kathryn Campbell-Kibler. 2011. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, 22(3):423–441.

Kathryn Campbell-Kibler, Penelope Eckert, Norma Mendoza-Denton, and Emma Moore. 2006. The elements of style. In *Poster Session at New Ways of Analyzing Variation (NWAV)*, Columbus, USA.

Hongliu Cao. 2025. Writing style matters: An examination of bias and fairness in information retrieval systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 336–344, Hannover Germany. ACM.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2026. Out of Style: RAG's Fragility to Linguistic Variation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 280–318, Rabat, Morocco. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024b. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint*. ArXiv:2410.15226 [cs].

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. 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 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, and Dan Jurafsky. 2025. HumT DumT: Measuring and controlling human-like language in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25983–26008, Vienna, Austria. Association for Computational Linguistics.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020a. CLUB: A Contrastive Log-ratio Upper Bound of mutual information. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020b. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online. Association for Computational Linguistics.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating Synthetic Data Generation from User Generated Text. *Computational Linguistics*, 51(1):191–233.

Tanya Karoli Christensen and Torben Juel Jensen. 2022. *When Variants Lack Semantic Equivalence: Adverbial Subclause Word Order*, pages 171–206. Cambridge University Press, Cambridge, UK.

Eve V. Clark. 1992. Conventionality and contrast: Pragmatic principles with lexical consequences. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 171–188. Routledge, New York, USA.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada. Association for Computational Linguistics.

Jeff Collins, David Kaufer, Pantelis Vlachos, Brian Butler, and Suguru Ishizaki. 2004. Detecting collaborations in text comparing the authors' rhetorical language choices in the Federalist Papers. *Computers and the Humanities*, 38:15–36.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447.

Nikolas Coupland. 2007. *Style: Language Variation and Identity*. Cambridge University Press, Cambridge, UK.

David Crystal. 2008. *Txtng: The gr8 db8*. Oxford University Press, Oxford, UK.

David Crystal. 2011. *A Dictionary of Linguistics and Phonetics*, 6th edition. Blackwell Publishing, Malden, USA.

David Crystal and Derek Davy. 1969. *Investigating English Style*. Routledge, London, UK.

S. H. H. Ding, B. C. M. Fung, F. Iqbal, and W. K. Cheung. 2019. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

George R. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2521–2524.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Xingping Dong and Jianbing Shen. 2018. Triplet loss in siamese network for object tracking. In *Computer Vision – ECCV 2018*, pages 472–488, Cham. Springer International Publishing.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Penelope Eckert. 1989. *Jocks and Burnouts: Social Categories and Identity in the High School*. Teachers College Press, New York, USA.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41(1):87–100.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: A package for computational text analysis. *The R Journal*, 8(1):107–121.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Ahmed M. Elkhatat, Khaled Elsaid, and Saeed Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1):1–16.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Susan M. Ervin-Tripp. 2001. Variety, style-shifting, and ideology. In Penelope Eckert and John R. Rickford, editors, *Style and Sociolinguistic Variation*, pages 44–56. Cambridge University Press, Cambridge, UK.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, and 47 others. 2025. Olmo 3. Technical Report.

Qixiang Fang, Dong Nguyen, and Daniel L. Oberski. 2022. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, 11(1):39.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. 2025. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Steven Fincke and Elizabeth Boschee. 2024. Separating style from substance: Enhancing cross-genre authorship attribution through data selection and presentation. *Preprint*, arXiv:2408.05192.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68.

Howard Giles and Peter F. Powesland. 1975. *Speech Style and Social Evaluation*. Academic Press, London, UK.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint*. ArXiv:1901.05287 [cs].

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Tim Grant. 2022. *The Idea of Progress in Forensic Authorship Analysis*. Elements in Forensic Linguistics. Cambridge University Press.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Jack Grieve. 2023. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova. 2011. Variation among blogs: A multidimensional analysis. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web*, pages 303–322. Springer, Dordrecht, the Netherlands.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *arXiv preprint*. ArXiv:2412.10271 [cs].

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Yaakov HaCohen-Kerner, Eyal Seckbach, and Dan Bouhnik. 2026. Multi-task classification of Hebrew news articles: A comparative study of classical ML and BERT models in a morphologically rich, low-resource setting. *Applied Sciences*, 16(8).

Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

Rebecca M. M. Hicke and David Mimno. 2025. Looking for the inner music: Probing LLMs' understanding of literary style. *Computational Humanities Research*, 1:e3.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Nicole Holliday. 2021. Intonation and referee design phenomena in the narrative speech of Black/biracial men. *Journal of English Linguistics*, 49(3):283–304.

David I. Holmes. 1985. The analysis of literary style–A review. *Journal of the Royal Statistical Society: Series A (General)*, 148(4):328–341.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024a. ParaGuide: Guided diffusion paraphrasers for plug-and-play textual style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18216–18224, Vancouver, Canada. Association for the Advancement of Artificial Intelligence.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024b. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390, Miami, Florida, USA. Association for Computational Linguistics.

Anika Samiha Hossain, Nazia Akter, and MD. Saiful Islam. 2020. A stylometric approach for author attribution system using neural network and machine learning classifiers. In *Proceedings of the International Conference on Computing Advancements*, pages 1–7.

Renkui Hou and Chu-Ren Huang. 2019. Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering*, 26(1):49–71.

John Houvardas and Efstathios Stamatatos. 2006. N-gram feature selection for authorship identification. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA'06, page 77–86, Berlin, Germany. Springer.

Dirk Hovy. 2015. Demographic factors improve classification performance. 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 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can Large Language Models Identify Authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2024. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, 296:111867.

Judith T. Irvine. 2001. "Style" as distinctiveness: the culture and ideology of linguistic differentiation. In Penelope Eckert and John R. Rickford, editors, *Style and Sociolinguistic Variation*, pages 21–43. Cambridge University Press, Cambridge, UK.

Abraham Israeli, Shuai Liu, Jonathan May, and David Jurgens. 2025. The Million Authors corpus: A cross-lingual and cross-domain Wikipedia dataset for authorship verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25997–26017, Vienna, Austria. Association for Computational Linguistics.

Anubhav Jangra, Bahareh Sarrafzadeh, Silviu Cucerzan, Adrian de Wynter, and Sujay Kumar Jauhar. 2025. Evaluating Style-Personalized Text Generation: Challenges and Directions. *arXiv preprint*. ArXiv:2508.06374 [cs].

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models. *arXiv preprint*. ArXiv:1707.01161 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825 [cs].

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Patrick Juola, John Noecker Jr., Mike Ryan, and Sandy Speer. 2009. JGAAP 4.0–A revised authorship attribution tool. *Proceedings of Digital Humanities*.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Dongyeop Kang and Eduard Hovy. 2021. Style is NOT a single variable: Case studies for cross-stylistic language understanding. 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 2376–2387, Online. Association for Computational Linguistics.

Anjali Kantharuban, Aarohi Srivastava, Fahim Faisal, Orevaoghene Ahia, Antonios Anastasopoulos, David Chiang, Yulia Tsvetkov, and Graham Neubig. 2026. IDIOLEX: Unified and Continuous Representations for Idiolectal and Stylistic Variation. *arXiv preprint*. ArXiv:2604.04704 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint*. ArXiv:2001.08361 [cs].

Mike Kestemont. 2014. Function words in authorship attribution. From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Salar Khaleghzadegan, Michael Rosen, Anne Links, Alya Ahmad, Molly Kilcullen, Emily Boss, Mary Catherine Beach, and Somnath Saha. 2024. Validating computer-generated measures of linguistic style matching and accommodation in patient-clinician communication. *Patient Education and Counseling*, 119:108074.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5275–5287, Online. Association for Computational Linguistics.

Aleem Khan, Andrew Wang, Sophia Hager, and Nicholas Andrews. 2024. Learning to generate text in arbitrary writing styles. *arXiv preprint*. ArXiv:2312.17242 [cs].

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Junghwan Kim, Haotian Zhang, and David Jurgens. 2025. Leveraging multilingual training for authorship representation: Enhancing generalization across languages and domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34855–34880, Suzhou, China. Association for Computational Linguistics.

Hannes Kniffka. 2007. *Working in Language and Law: A German Perspective*. Palgrave Macmillan UK.

Arne Köhn. 2015. What's in an embedding? Analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Tore Kristiansen. 2024. Social variation in germanic. *Oxford Research Encyclopedia of Linguistics*.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of AI-generated text in Twitter timelines. *arXiv preprint*. ArXiv:2303.03697 [cs].

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

William Labov. 2006. *The Social Stratification of English in New York City*, 2nd edition. Cambridge University Press, Cambridge, UK.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling (COLM)*.

Beatriz R. Lavandera. 1978. Where does the sociolinguistic variable stop? *Language in Society*, 7(2):171–182.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. TextBugger: Generating adversarial text against real-world applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, San Diego, CA. Internet Society.

Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. 2024. Does style matter? Disentangling style and substance in chatbot arena.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Philip Lippmann and Jie Yang. 2025. Style over substance: Distilled language models reason via stylistic replication. In *Second Conference on Language Modeling (COLM)*.

Pierre Lison, 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, Online. Association for Computational Linguistics.

Frederick Liu, Terry Huang, Shihang Lyu, Siamak Shakeri, Hongkun Yu, and Jing Li. 2022. Enct5: A framework for fine-tuning t5 as non-autoregressive models. *arXiv preprint*. ArXiv:2110.08426 [cs].

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint*. ArXiv:2502.11528 [cs].

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023. RECAP: Retrieval-Enhanced Context-Aware Prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.

Stephan Ludwig, Ko de Ruyter, Max Friedman, Elisabeth Constantin Brüggen, Martin Wetzels, and Gerard Pfann. 2013. More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1):87–103.

Marcus Ma, Duong Minh Le, Junmo Kang, Yao Dou, John Cadigan, Dayne Freitag, Alan Ritter, and Wei Xu. 2025. CROSSNEWS: A Cross-Genre Authorship Verification and Attribution Benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24777–24785.

Suraj Maharjan, Deepthi Mave, Prasha Shrestha, Manuel Montes, Fabio A. González, and Thamar Solorio. 2019. Jointly learning author and annotated character n-gram embeddings: A case study in literary text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 684–692, Varna, Bulgaria. INCOMA Ltd.

Pratyush Maini. 2023. Phi-1.5 model: A case of comparing apples to oranges?

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.

Hieu Man and Thien Huu Nguyen. 2024. Counterfactual augmentation for robust authorship representation learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2347–2351, New York, NY, USA. Association for Computing Machinery.

Hieu Man, Van-Cuong Pham, Nghia Trung Ngo, Franck Dernoncourt, and Thien Huu Nguyen. 2026. Explainable Disentangled Representation Learning for Generalizable Authorship Attribution in the Era of Generative AI. *arXiv preprint*. ArXiv:2604.21300 [cs].

Nina Markl, Lauren Hall-Lew, and Catherine Lai. 2024. Language technologies as if people mattered: Centering communities in language technology development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10085–10099, Torino, Italia. ELRA and ICCL.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Miriam Meyerhoff. 2006. *Introducing Sociolinguistics*. Routledge, London, UK.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paul Michel 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, Melbourne, Australia. Association for Computational Linguistics.

Timothee Mickus and Maria Copot. 2024. Stranger than paradigms word embedding benchmarks don't align with morphology. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 173–189, Irvine, CA. Association for Computational Linguistics.

George K Mikros and Eleni K Argiri. 2007. Investigating topic influence in authorship attribution. In *SIGIR'07 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*.

Peter Millican. 2003. The Signature stylometric system. Web download. University of Oxford.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, Suzhou, China. Association for Computational Linguistics.

Ana Cristina Munaro, Renato Hübner Barcelos, Eliane Cristine Francisco Maffezzolli, João Pedro Santos Rodrigues, and Emerson Cabrera Paraiso. 2024. Does your style engage? Linguistic styles of influencers and digital consumer engagement on YouTube. *Computers in Human Behavior*, 156.

Zulqarnain Nazir, Khurram Shahzad, Muhammad Kamran Malik, Waheed Anwar, Imran Sarwar Bajwa, and Khawar Mehmood. 2021. Authorship attribution for a resource poor language—Urdu. *Transactions on Asian and Low-Resource Language Information Processing*, 21(3):1–23.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):86.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 439–448, Cambridge, USA. Association for the Advancement of Artificial Intelligence.

Dong Nguyen and Jack Grieve. 2020. Do word embeddings capture spelling variation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 870–881, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dong Nguyen and Esther Ploeger. 2025. We need to measure data diversity in NLP — better and broader. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8823–8832, Suzhou, China. Association for Computational Linguistics.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Andrea Nini. 2019. The multi-dimensional analysis tagger. In Tony Berber Sardinha and Marcia Veirano Pinto, editors, *Multi-Dimensional Analysis: Research Methods and Current Issues*. London; New York: Bloomsbury Academic.

Andrea Nini. 2023. *A Theory of Linguistic Individuality for Authorship Analysis*. Elements in Forensic Linguistics. Cambridge University Press.

Andrea Nini, Oren Halvani, Lukas Graner, Sophie Titze, Valerio Gherardi, and Shunichi Ishihara. 2026. Grammar as a behavioral biometric: using cognitively motivated grammar models for authorship verification. *Humanities and Social Sciences Communications*, 13(1):455.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. 2023. StyloMetrix: An open-source multilingual tool for representing stylometric vectors. *Preprint*, arXiv:2309.12810.

Annaleena Parhankangas and Maija Renko. 2017. Linguistic style and crowdfunding success among social and commercial entrepreneurs. *Journal of Business Venturing*, 32(2):215–236.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, page 267–274, USA. Association for Computational Linguistics.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin, Austin, USA.

Martin Potthast, Felix Schremmer, Matthias Hagen, and Benno Stein. 2018. Overview of the Author Obfuscation Task at PAN 2018: A New Approach to Measuring Safety. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, Avignon, France.

Drexel University PSAL. 2013. JSAN—The integrated JStylo and Anonymouth package. The Privacy, Security and Automation Lab (PSAL) at Drexel University.

Devi Ambarwati Puspitasari. 2026. Forming and authorship profile through n-gram tracing. *Kongres Internasional Masyarakat Linguistik Indonesia*, 2025.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! Adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Justin Qiu, Jiacheng Zhu, Ajay Patel, Marianna Apidianaki, and Chris Callison-Burch. 2025. mStyleDistance: Multilingual Style Embeddings and their Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16917–16931, Vienna, Austria. Association for Computational Linguistics.

Ella Rabinovich, 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, Valencia, Spain. Association for Computational Linguistics.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, Valencia, Spain. CEUR Workshop Proceedings.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John R. Rickford and McNair-Knox. 1994. Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In Douglas Biber and Edward Finegan, editors, *Sociolinguistic Perspectives on Register*, pages 235–276. Oxford University Press, New York, USA.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Y. Chen, Marcus Bishop, and Nicholas Andrews. 2023. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Rafael Rivera Soto, 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandra Camille Sandoval, Christabel Acquaye, Kwesi Adu Cobbina, Mohammad Nayeem Teli, and Hal Daumé Iii. 2025. My LLM might mimic AAE - But when should it? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5277–5302, Albuquerque, New Mexico. Association for Computational Linguistics.

Jitkapat Sawatphol, Nonthakit Chaiwong, Can Udomcharoenchaikit, and Sarana Nutanong. 2022. Topic-regularized authorship representation learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1076–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jitkapat Sawatphol, Can Udomcharoenchaikit, and Sarana Nutanong. 2024. Addressing topic leakage in cross-topic evaluation for authorship verification. *Transactions of the Association for Computational Linguistics*, 12:1363–1377. Place: Cambridge, MA.

Vageesh Kumar Saxena, Benjamin Ashpole, Gijs Van Dijck, and Gerasimos Spanakis. 2025. MATCHED: Multimodal Authorship-attribution To Combat Human trafficking in Escort-advertisement Data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4334–4373, Vienna, Austria. Association for Computational Linguistics.

Eleni-Konstantina Sergidou, Nelleke Scheijen, Jeannette Leegwater, Tina Cambier-Langeveld, and Wauter Bosma. 2023. Frequent-words analysis for forensic speaker comparison. *Speech Communication*, 150:1–8.

R.V. ShabbirHusain, Atul Arun Pathak, Shabana Chandrasekaran, and Balamurugan Annamalai. 2023. The power of words: Driving online consumer engagement in Fintech. *International Journal of Bank Marketing*, 42(2):331–355.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844, Red Hook, NY, USA. Curran Associates Inc.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Mohammad Shokri, Sarah Ita Levitan, and Rivka Levitan. 2025. Personalized author obfuscation with large language models. *arXiv preprint*. ArXiv:2505.12090 [cs].

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in RLHF.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2017. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Eivind Strøm. 2021. Multi-label style change detection by solving a binary classification problem. In *CLEF 2021: Conference and Labs of the Evaluation Forum*, pages 2146–2157.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in large language models. In *Forty-second International Conference on Machine Learning (ICML)*.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

Xuemei Tang, Shichen Liang, and Zhiying Liu. 2019. Authorship attribution of The Golden Lotus based on text classification methods. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, pages 69–72.

Paul J. Taylor and Sally Thomas. 2008. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, 1(3):263–281.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*.

Enzo Terreau, Antoine Gourru, and Julien Velcin. 2021. Writing style author embedding evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 84–93, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karishma Thakrar, Katrina Lawrence, and Kyle Howard. 2025. StAyaL | Multilingual style transfer. *arXiv preprint*. ArXiv:2501.11639 [cs].

Nafis Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti, and Dongwon Lee. 2023. HANSEN: Human and AI spoken text benchmark for authorship analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13706–13724, Singapore. Association for Computational Linguistics.

William M. K. Trochim, James P. Donnelly, and Kanika Arora. 2015. *Research Methods: The Essential Knowledge Base*. Cengage Learning, Boston.

Kimberly Truong, Riccardo Fogliato, Hoda Heidari, and Steven Wu. 2025. Persona-augmented benchmarking: Evaluating LLMs across diverse writing styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22687–22720, Suzhou, China. Association for Computational Linguistics.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv preprint*. ArXiv:2209.06869 [cs].

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Suzanne Evans Wagner. 2025. Style and social meaning across the lifespan. *Connecting the Individual and the Community in Sociolinguistic Panel Research*, page 96.

Jan Philip Wahle, Terry Ruas, Yang Xu, and Bela Gipp. 2024. Paraphrase types elicit prompt engineering capabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11004–11033, Miami, Florida, USA. Association for Computational Linguistics.

Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and

23 others. 2025. 2 OLMo 2 Furious. *arXiv preprint*. ArXiv:2501.00656 [cs].

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023a. Can Authorship Representation Learning Capture Stylistic Features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023b. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Janith Weerasinghe and Rachel Greenstadt. 2020. Feature vector difference based neural network and logistic regression models for authorship verification. In *Notebook for PAN at CLEF 2020*, volume 2695.

Anna Wegmann 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. Tokenization is sensitive to language variation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10958–10983, Vienna, Austria. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? Towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

E. Judith Weiner and William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics*, 19(1):29–58.

Jennifer Williams and Simon King. 2019. Disentangling style factors from speaker representations. In *20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language*, pages 3945–3949. ISCA.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE. Association for Computational Linguistics.

Yilin Xu, Eric Friginal, and Chunyu Hu. 2026. A Multi-Dimensional Analysis of English Tweets by Chinese and American automotive companies. *International Journal of Applied Linguistics*.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):64:1–64:37.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Yan Zhang, Zhaopeng Feng, Zhiyang Teng, Zuozhu Liu, and Haizhou Li. 2023b. How well do text embedding models understand syntax? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9717–9728, Singapore. Association for Computational Linguistics.

Jiaxu Zhao, Meng Fang, Kun Zhang, and Mykola Pechenizkiy. 2025. Unmasking style sensitivity: A causal analysis of bias evaluation instability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16314–16338, Vienna, Austria. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2022. Disentangled sequence to sequence learning for compositional generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4256–4268, Dublin, Ireland. Association for Computational Linguistics.

Jian Zhu 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kangchen Zhu, Zhiliang Tian, Jingyu Wei, Ruifeng Luo, Yiping Song, and Xiaoguang Mao. 2024. Style-Flow: Disentangle latent representations via normalizing flow for unsupervised text style transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15384–15397, Torino, Italia. ELRA and ICCL.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019(256):147–175.

Dimitrina Zlatkova, Daniel Kopev, Kristiyan Mitov, Atanas Atanasov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. An ensemble-rich multi-aspect approach for robust style change detection. In *Notebook for PAN at CLEF-2018*, page 3.

Chaoyuan Zuo, Yu Zhao, and Ritwik Banerjee. 2019. Style change detection with feed-forward neural networks. *Notebook for PAN at CLEF 2019*, 93.

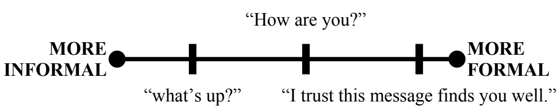

Figure 3: **Style is relative.** It might be more difficult or less interesting to make categorical judgments about a text's style in isolation than, for example, judging if a text is more formal than another on a formality continuum. As Irvine (2001) writes on page 22, "It is seldom useful to examine a single style in isolation" and "attention must be directed to relationships among styles—to their contrasts, boundaries and commonalities."

A Additional figures and tables

Tab. 2 shows an overview of various predefined feature style operationalizations (§3.1), and Fig. 3 portrays an example of why style may require new solutions (§4.1).

Type	Variable	Examples
PHONETIC	postvocalic /r/	more or less clear pronunciation of /r/ sound after vowel (Labov, 1972)
	intervocalic /t/	full/flapped /t/ voicing between two vowel sounds (*writer* → *rider*) (Bell, 1984)
	...	
MORPHO-LOGICAL	word endings	*g*-dropping (Campbell-Kibler, 2007), gerunds (Biber, 1988)
	nominalizations	ending in *-tion*, *-ment*
	verb morphology	*be* as a main or auxillary verb (Biber, 1988)
	...	
LEXICAL	word/token counts	number of words/tokens (Stamatatos, 2009)
	word/token ratios	ratio of types to tokens, ratio of short/long words to token count, etc. (Altakrori et al., 2021)
	word/token n-grams	for n of various lengths (Abbasi and Chen, 2008; Stamatatos, 2009)
	word length	average word length (Biber, 1988), also cf. Grieve (2007)
	sentence length	distribution of average sentence length, cf. Grieve (2007)
	vocabulary richness	hapax (dis)legomena, Yule's I/K, number of unique tokens (Abbasi and Chen, 2008; Stamatatos, 2009)
	function words	grammar-functioning words, e.g., *the*, *be*, *to* (Abbasi and Chen, 2008; Mosteller and Wallace, 1963; Stamatatos, 2009)
	pronoun use	word frequency distributions of 1st, 2nd,... person pronouns (Biber, 1988; Pennebaker et al., 2015)
	hedge words	*at about*, *something like* as hedges in Biber MDA features; *maybe*, *perhaps* in tentative dimension in LIWC
	quantifiers	*each*, *all* as quantifier words or *everybody*, *anybody* as quantifier pronouns (Biber, 1988)
	...	
SYNTACTIC	POS counts	noun, verb, adjective,... (Abbasi and Chen, 2008; Biber, 1988)
	POS n-grams	for various n (Abbasi and Chen, 2008; Weerasinghe and Greenstadt, 2020)
	passive voice	agentless passives (Biber, 1988)
	subordination features	*that* relative clause vs. *wh-* relative clause (e.g., *the dog that* vs. *the dog who*) (Biber, 1988)
	negation	*need no water* as negative concord (Eckert, 2008); *not* in analytic negation (Biber, 1988), negation words in LIWC
	invariant *be*	*He be working* (Rickford and McNair-Knox, 1994)
	zero copula	*She nice* (Rickford and McNair-Knox, 1994)
	...	
DISCOURSE	contraction use	*can't* vs. *cannot* (contractions list[1] Biber (1988))
	discourse particle	*well*, *now* (Biber, 1988)
	readability	Flesch Reading Ease, Flesch Kincaid Grade Level, etc. (Python's `textstat`[2])
	compression	train a compression model on one text and use it to estimate how similar in style another text is, cf. Stamatatos (2009)
	...	
ORTHO-GRAPHIC	character types	hashtags, emojis, exclamation marks (Clarke and Grieve, 2017); uppercase characters, digits (Stamatatos, 2009)
	character n-grams	for various n (Abbasi and Chen, 2008; Stamatatos, 2009)
	lengthening	*cooool* (Nguyen and Grieve, 2020)
	number substitutions	*2day* (Crystal, 2008)
	misspellings	common misspellings list[3]
	acronyms/abbreviations	common shortened forms list[4]
	...	

1 `https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions`
2 `https://pypi.org/project/textstat/`
3 `https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines`
4 `https://en.wikipedia.org/wiki/SMS_language`

Table 2: **Overview of predefined feature style operationalizations used in different fields.** Specific linguistic features that have been used to operationalize style and examples of each are categorized by linguistic level: phonetic (i.e., pronunciation and sound patterns), morphological (i.e., word structure and inflection), lexical (i.e., word choice), syntactic (i.e., sentence structure), discourse (i.e., larger structure), and orthographic (i.e., spelling and punctuation). Note that the categorizations might overlap, e.g., *g*-dropping might also be considered an orthographic or phonological variable, and character n-grams might encode different morphemes. These features have been investigated separately (Campbell-Kibler, 2009) and collectively (e.g., Abbasi and Chen, 2008; Biber, 1988; Neal et al., 2017; Stamatatos, 2009). This table was inspired by and partially filled with elements from other tables of stylometric features in these and other sources. For further references and examples consider also Grieve (2007) and Biber (1988).

B Motivating examples

B.1 Reasoning traces in the s1 dataset

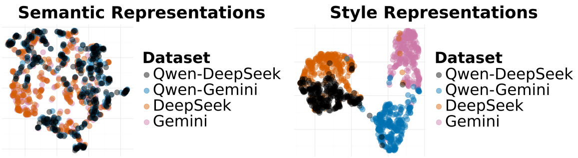

Figure 4: **Style representations of distilled Qwen models are close to teacher models** We compare reasoning traces on s1 for DeepSeek and Gemini models (Muennighoff et al., 2025) and reasoning traces on Math500 (Hendrycks et al., 2021) generated by models distilled on the s1 DeepSeek and Gemini reasoning traces respectively. The style representations (right) group the style of the student model closer to the style of the teacher model, while the semantic representations (left) overlap.

We created Fig. 1 using the first 500 elements of the s1 datasets provided by Muennighoff et al. (2025) with reasoning traces generated by Gemini Flash Thinking Experimental and DeepSeek R1.[8] We used a semantic representation model[9] and a style representation model[10] and UMAP (McInnes et al., 2018) with default settings.

Pioneering work found that the style of reasoning traces might be important to consider for the performance of reasoning models (Lippmann and Yang, 2025). Note, however, that their definition of style does not fully align with the definition used in this paper (e.g., including "non-linear thinking" as a style). In an ablation, we compare the semantic and style representations of the DeepSeek and Gemini teacher models and the distilled Qwen models on DeepSeek and Gemini. While the original Muennighoff et al. (2025) paper trains Qwen models only on Gemini reasoning traces, the authors later experimented with DeepSeek reasoning traces and found them to lead to better performance.[11] We take the first 270 s1 reasoning traces as provided by Muennighoff et al. (2025) and use the fine-tuned

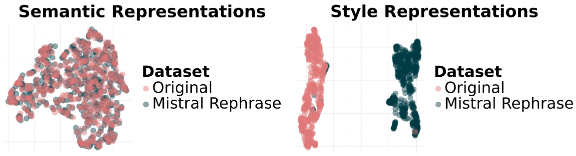

Figure 5: **Comparing semantic and style representations of LLM-rephrases** We compare MRPC sentences (Dolan and Brockett, 2005) and their LLM-generated "Wikipedia-style" rephrases using prompts from Maini et al. (2024). Style embedding models (right) can easily distinguish between the original and the LLM-rephrased sentences, while semantic embeddings (left) overlap. Studying stylistic diversity of LLM-rephrases is relevant as stylistic rephrasing is increasingly used in dataset curation for pre- and post-training.

Qwen models on Gemini[12] and DeepSeek[13] reasoning traces to generate reasoning traces[14] for the first 270 Math500[15] problems (Lightman et al., 2023). We use a different dataset from s1 to query student models to avoid artifacts of memorization. We choose Math500 as the distilled s1 Qwen models were also evaluated on it. See the results in Fig. 4 using UMAP visualization as before. We show that the style of the model distilled on Gemini reasoning traces is also closer in style to the Gemini reasoning traces than to the DeepSeek reasoning traces. Thus, the student model is effectively adopting the style of the teacher model (same for the DeepSeek model).

B.2 Rephrases of the MRPC dataset

Using synthetic data in pre- and post-training is increasingly common. We take the prompt from Maini et al. (2024) and use the Mistral-7B-Instruct-v0.1 model[16] (Jiang et al., 2023) to create Wikipedia-style rephrases of the first 500 elements of the MRPC dataset (Dolan and Brockett, 2005). We use the same models as in §B.1 for the semantic and style representations as well as hyperparameters for the UMAP visualization. Style representations clearly distinguish the LLM rephrases from the original sentences, while semantic representa-

[8]"gemini_thinking_trajectory" and 'deepseek_thinking_trajectory" column in `https://huggingface.co/datasets/simplescaling/s1K-1.1`

[9]Hugging Face's sentence-transformers/all-mpnet-base-v2

[10]Hugging Face's AnnaWegmann/Style-Embedding

[11]`https://x.com/Muennighoff/status/1886405528777073134`

[12]`https://huggingface.co/simplescaling/s1-32B`

[13]`https://huggingface.co/simplescaling/s1.1-32B`

[14]By preceding the response with "\n<|im_start|>think\n"

[15]`https://huggingface.co/datasets/HuggingFaceH4/MATH-500`

[16]`https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1`

Model	AUC	Acc@0.8	TFLOPs	USD
StyleDistance	**.69**	**.66**	≈ 13	0
GPT-5.2	.56	.56	$\approx 89k$	1.61

Table 3: Style embeddings can outperform LLMs on authorship verification with far fewer FLOPs (see §B.4). Acc@0.8 = Accuracy at threshold 0.8

tions do not (Fig. 5).

B.3 Clustering writers of English by native language

We created Fig. 2 using the ETS Corpus of Non-Native Written English (LDC2014T06) (Blanchard et al., 2014). The corpus is comprised of English essays written by speakers of 11 non-English native languages as part of an international test of academic English proficiency, TOEFL (Test of English as a Foreign Language). We used LUAR[17] (Rivera Soto et al., 2021), a style representation model trained on the authorship verification task. Each point in the figure is an embedding of 5 TOEFL essays written by authors of the same native language picked at random. We reduce the dimensionality to two components using UMAP (McInnes et al., 2018) with default settings. Although the style representation was initially trained on the "idiolectal" authorship verification task (distinguishing authors based on their distinctive language use), Fig. 2 reveals that it also captures features indicative of the writer's native language.

B.4 Comparing to prompting

We compare StyleDistance embeddings[18] (Patel et al., 2025) and GPT-5.2 with linguistically informed (LIP) prompts taken from Huang et al. (2024). We compare both on the authorship verification task using 500 pairs of Reddit comments. Specifically, we take the first 250 triples of the test split used in Wegmann et al. (2022), split into two authorship verification pairs that share one text (e.g., (A, B, C) $\rightarrow$ (A, B), (A, C)). The original dataset was constructed to make content cues less helpful to solve the task; as a result, we assume that models need to have learned a certain level of content-independence to perform well.

Results Results are shown in Tab. 1 and again with AUC in App. Tab. 3. We measure the area under the receiver operating curve (AUC) and the accuracy at a threshold of 0.8 (Acc@0.8), finding that StyleDistance outperforms GPT-5.2 while being more efficient than GPT-5.2 as well. We report order-of-magnitude TFLOPs using 2*N(active parameters)*T(tokens) based on Kaplan et al., 2020. This result shows that style embeddings can be orders of magnitude more efficient as well as more performant than SOTA LLMs at style tasks like authorship verification.

[17] https://huggingface.co/rrivera1849/LUAR-MUD

[18] https://huggingface.co/StyleDistance/styledistance

C Additions to style conceptualizations

Fully separating style and semantic meaning might be impossible.

> *Sociolinguists generally think of styles as different ways of saying the same thing. In every field that studies style seriously, however, this is not so.*
>
> — Penelope Eckert

A precise separation of semantic meaning and style poses practical challenges. It has been argued that, for example, only Labov (1972)'s original object of study—phonological variables—can leave semantic meaning untouched, whereas all other variables, including lexical and syntactic variables, will necessarily change the semantic meaning (Campbell-Kibler, 2011; Lavandera, 1978; Sun et al., 2023). Additionally, Eckert (2008, 2012) argues that using a certain style systematically connects an utterance to the social world, and that style thus influences social meaning. Others argue that any two forms must necessarily contrast in meaning (Clark, 1992). Some work in sociolinguistics sidesteps the problem of meaning equivalence by studying when speakers use one form rather than another in a given context without claiming the forms are equivalent in general (Campbell-Kibler, 2011; Christensen and Jensen, 2022). Nonetheless, we believe that attempting to separate style and semantic meaning has practical uses (see §2.2 or Weiner and Labov, 1983).

Style across research fields Several fields study linguistic style in some capacity. As discussed in the paper, *sociolinguistics* examines the relationship between language and society with a focus on language change and variation (Eckert, 2008; Labov, 1972). *Corpus linguistics* is the descriptive study of how language is actually used by analyzing text corpora (e.g. Biber, 1988; Biber and Conrad, 2019). Typical applications might include comparing language between different genres like scientific papers and news articles. *Forensic linguistics* involves the study of style in the context of law and crime investigation and is typically interested in recognizing a style or *idiolect* that helps distinguish an investigated individual (Coulthard, 2004). Practical insights from forensic linguistics also reciprocally influence *stylistics* and *stylometry*, which more generally study linguistic style in language. Stylometry applications include investigating the style of literary authors (Holmes, 1985) or attributing disputed literary works (Burrows, 2002; Mosteller and Wallace, 1963; Stamatatos, 2009). Style in *NLP* has been investigated to characterize authors (e.g., age or gender in Koppel et al., 2002; Nguyen et al., 2013), detect stylistic inconsistencies (Collins et al., 2004; Stamatatos, 2009), and adapt styles in machine translation (Niu et al., 2017, 2018; Rabinovich et al., 2017). Linguistic style also plays a significant role in related fields like *psycholinguistics*, or even in *communication* and *marketing*, such as by influencing consumer engagement (Munaro et al., 2024; ShabbirHusain et al., 2023) and purchases (Ludwig et al., 2013).

Note that these fields are not strictly separable. Methods from corpus linguistics can inform sociolinguistics, forensic linguistics can use methods from stylometry, and so on. Further, there are several fields that can be connected to linguistic style that we do not specifically discuss here, such as *discourse analysis*, *digital humanities*, *linguistic anthropology*, and *sociology*.

D Additions to representing style in NLP

Available predefined feature extraction tools There are a multitude of tools available that automatically extract predefined features from text. The choice of tool and feature set, though, depends on various factors, such as preferred programming language, the nature of the data, and the goal of the task. Therefore, best practice is to systematically compare multiple feature sets, sometimes across tools, for each specific use case. Python tools include but are not limited to spaCy (Honnibal et al., 2020), Stanza (Qi et al., 2020), and NLTK (Bird et al., 2019) for general text processing, PAN submissions for authorship attribution (Weerasinghe and Greenstadt, 2020) and style change detection tasks (Strøm (2021), Zlatkova et al. (2018), LFTK (Lee and Lee, 2023) and StyloMetrix (Okulska et al., 2023) for extracting numerous stylometric features (but not n-grams), NeuroBiber and BiberPlus (Alkiek et al., 2025) for extracting Biber-style features, and StyloSpeaker (Aggazzotti and Smith, 2025) for extracting speech transcript features. Non-Python stylometric authorship tools include Stylo in R (Eder et al., 2016) and JStylo in Java (PSAL, 2013). Software that does not require programming includes LIWC (Pennebaker et al., 2015), which groups words into linguistically and psychologically meaningful categories; JGAAP (Juola et al., 2009) and Signature (Millican, 2003), which extract stylomet-

Tool	Original Purpose	Language / Platform	Type	Link
spaCy (Honnibal et al., 2020)	General text processing	Python	library	github.com/explosion/spaCy
Stanza (Qi et al., 2020)	General text processing	Python	library	https://github.com/stanfordnlp/stanza
NLTK (Bird et al., 2019)	General text processing	Python	library	github.com/nltk/nltk
PAN 2020 AV (Weerasinghe and Greenstadt, 2020)	AV	Python	Task subm.	github.com/janithnw/pan2020_authorship_verification
PAN 2021 SCD (Strøm, 2021)	SCD	Python	Task subm.	github.com/eivistr/pan21-style-change-detection-stacking-ensemble
PAN 2019 SCD (Zuo et al., 2019)	SCD	Python	Task subm.	github.com/chzuo/PAN_2019
PAN 2018 SCD (Zlatkova et al., 2018)	SCD	Python	Task subm.	github.com/machinelearning-su/style-change-detection
LFTK (Lee and Lee, 2023)	Stylometric feature extraction (no n-grams)	Python	library	github.com/brucewlee/lftk
StyloMetrix (Okulska et al., 2023)	Stylometric feature extraction (no n-grams)	Python	library	github.com/NASK-NLP/StyloMetrix
BiberPlus (Alkiek et al., 2025)	Biber-style feature extraction	Python	library	github.com/davidjurgens/biberplus
NeuroBiber (Alkiek et al., 2025)	Biber-style feature extraction	HF	Model	huggingface.co/Blablablab/neurobiber
MAT (Nini, 2019)	Biber-style feature extraction	Python	library	github.com/andreanini/multidimensionalanalysistagger
StyloSpeaker (Aggazzotti and Smith, 2025)	Speech transcript feature extraction	Python	library	github.com/caggazzotti/styloSpeaker
Stylo (R) (Eder et al., 2016)	Stylometric authorship analysis	R	library	github.com/computationalstylistics/stylo
JStylo (Java) (PSAL, 2013)	Stylometric authorship analysis	Java	App	github.com/psal/jstylo
LIWC (Pennebaker et al., 2015)	Ling./psych. categories	SW (GUI)	App	www.liwc.app/
JGAAP (Juola et al., 2009)	Stylometric + n-gram features	SW (GUI)	App	evllabs.github.io/JGAAP/
Signature (Millican, 2003)	Stylometric + n-gram features	SW (GUI)	App	www.philocomp.net/texts/signature.htm
Coh-Metrix (Graesser et al., 2004)	Text cohesion and discourse features	SW (GUI)	App	soletlab.asu.edu/coh-metrix/

Table 4: **Comparison of common tools for extracting predefined features** The table summarizes their original purpose, programming language or platform, type of resource, and URL. Abbreviations: **AV** = authorship verification, **Task subm.** = shared-task submission, **SCD** = style change detection, **HF** = Hugging Face, **App** = standalone application, **SW** = non-programming software. These tools particularly work for English, but see our Github for tools/papers for other languages: https://anonymous.4open.science/w/StyleSurvey-60B0/.

ric and n-gram features; and Coh-Metrix, which can measure more complex features like text cohesion (Graesser et al., 2004). We summarize these tools in Tab. 4.

Available automatically-learned models To the best of our knowledge, the available learned style representation models on HuggingFace are CISR[19] (Wegmann et al., 2022), StyleDistance[20] (Patel et al., 2025), mStyleDistance[21] (Qiu et al., 2025), LUAR[22] (Rivera Soto et al., 2021), Multilingual Style Representation or MSR[23] (Kim et al., 2025) and STAR[24] (Huertas-Tato et al., 2024). Another model available via a private sharing site is LISA[25] (Patel et al., 2023). Following the discussion in § 3.2, some style representations may capture more semantic features than others, and thus may prove to be more useful for different downstream tasks. We summarize these models in Tab. 5.

D.1 Additions to the challenges of style representations

❓ **Improving training** There are several other areas in training that remain underexplored, including fine-tuning newer encoder models like ModernBERT (Alshomary et al., 2025b), designing tokenizers specifically for style representations (Wegmann et al., 2025), and pooling not only the last, but several or all, encoder layers (Alshomary et al., 2025b).

❓ **Automatic feature selection** There is no standard set of predefined features that works across tasks and domains (§ 3.1). Future work could attempt to either develop strategies to select predefined features that work best for different kinds of tasks and data or to develop an ensemble method that combines features dynamically.

❓ **Including language modeling objectives** Previous work found that fine-tuning pretrained transformer models on style tasks can curiously lead to reduced performance on some style tasks compared to the pretrained base model (Patel et al., 2025; Wegmann and Nguyen, 2021). This might be connected to a difference in the object of study for the training and evaluation tasks. For example, using individuals as the object of study (e.g., using authorship verification as the training task) can lead to unlearning stylistic attributes that can vary for the same individual (e.g., the formality of their writing across online forums, job applications, and other contexts). When aiming to learn general-purpose style representations, it might be necessary to include further stylistic or continued language modeling objectives like masked language modeling.

❓ **Improve content-independence** This was already mentioned in the main paper, but we highlight this point for more clarity again. "Generally, few style representations reach high scores on content-independence (🔧 App. Tab. 5) and might benefit from more exhaustive content disentanglement.", see Tab. 4. Consider current content-disentanglement strategies in § 3.2.

E Additions to evaluating style representations

E.1 Additions to challenges in evaluating style representations

📣 **Evaluate interpret- and explainability** The evaluation of learned style representations on predefined features is not yet systematic, but is promising to pursue, as it can build on rich literature in linguistics and stylometry (§ 2.1, § 3.1) and can help make learned representations more interpretable. Further, there is only limited work on explaining learned style spaces. Alshomary et al., 2025a pioneer this direction by generating explanations on why embeddings cluster certain authors. It could also be useful to further evaluate style representations using mechanistic interpretability approaches to make sense of neuron activations and hidden representations, e.g., Vulić et al. (2020); Aoyama and Schneider (2022).

❓ **Leverage measurement theory** In the social sciences, measures are commonly assessed for *reliability* (do measures return the same result with repeated measurement?) and *validity* (do measures capture the concept of interest?). Measurement theory could provide the evaluation of style embeddings and the construction of style benchmarks with a theoretical framework, highlight gaps, and

[19] https://huggingface.co/AnnaWegmann/Style-Embedding
[20] https://huggingface.co/StyleDistance/styledistance
[21] https://huggingface.co/StyleDistance/mstyledistance
[22] https://huggingface.co/rrivera1849/LUAR-MUD
[23] https://huggingface.co/Blablablab/multilingual-style-representation-Llama-3.2
[24] https://huggingface.co/AIDA-UPM/star
[25] https://ajayp.app/posts/2023/11/learning-interpretable-embeddings-via-llms/

Model	Training Task	Languages	Content / Style Disentanglement	Interpretable?	Tasks*
LUAR (Rivera Soto et al., 2021)	AV	English	Weak	No	AR, MTD
CISR (Wegmann et al., 2022)	AV	English	Medium	No	AV, MTD
StyleDistance (Patel et al., 2025)	AV	English	Strong	No	AV, ST
mStyleDistance (Qiu et al., 2025)	AV	Multiple	Strong	No	AV, ST
LISA (Patel et al., 2023)	AV	English	Strong	Yes	Unknown
MSR (Kim et al., 2025)	AV	Multiple	Medium	No	AR, MTD
STAR (Huertas-Tato et al., 2024)	AV	English	Weak	No	AR, CL

Table 5: **Comparison of open-source learned style representation models** The categorization is based on key dimensions including the languages supported, the measured strength of content/style disentanglement, interpretability, and the specific downstream tasks the models are have been found useful for. *Note that the models may be useful for more tasks than stated here, the analysis is based on the authors' experience with them. Acronym Definitions: **AR** = Authorship Retrieval, **AV** = Authorship Verification, **MTD** = Machine-Text Detection, **ST** = Style Transfer, **CL** = Clustering

provide inspiration for future methods. 🔧 See Trochim et al. (2015) for more on measurement theory. See Fang et al. (2022) for examples of how to apply measurement theory to embeddings and Bean et al. (2025) for recommendations on how to construct valid benchmarks. We give some concrete examples of how measurement theory might be applied for style embeddings and benchmarks below.

For style embeddings, *convergent validity* (i.e., does the measure show similar measurement for similar concepts?) might be assessed by testing that texts that have a similar style have similar representations. This could be done by perturbing texts in stylistically inconsequential ways (e.g., by swapping out named entities like "Maria has style." to "Emma has style.") and comparing their embeddings. *Discriminant validity* (i.e., Is the measure not sensitive to concepts it should not be related to?) might be assessed by confirming that texts that change in other aspects than style (e.g., content) are still embedded similarly. This has been assessed before by evaluating content-independence (§ 4). *Predictive validity* (i.e., Can the measure be used to predict something that it should be predictive of?) might be assessed by evaluating performance on downstream tasks that make use of style representations, such as style classification or style transfer tasks (§ 4). 🔧 See also Fang et al. (2022) for further inspiration.

For style benchmarks, *reliability* (i.e., Is the measure giving the same results with repeated measurement?) might be improved by making sure that the same seeds are used when applying the benchmarks—for example, when using style classification tasks and a classifier is trained on top of embeddings. 🔧 See also Bean et al. (2025) for further inspiration related to benchmark *validity*—for example, they suggest to employ sampling strategies like stratified sampling that are representative of the task space.

F Additions to opportunities with style embeddings

Authorship attribution Style representations can enable authorship verification and attribution tasks, including historical authorship attribution of disputed texts (Mosteller and Wallace, 1963), identifying harmful actors (Arabnezhad et al., 2020; Saxena et al., 2025), detecting plagiarism in educational contexts (Elkhatat et al., 2023), and attributing speakers from speech transcripts (Aggazzotti et al., 2024, 2025b; Tripto et al., 2023).

Disentangle internal representations It may be useful to disentangle LLM-internal representations of style to allow models to turn style information on or off as needed. Disentanglement approaches have helped cross-domain generalization (Yang et al., 2023; Zheng and Lapata, 2022) and might also help cross-style generalization. This can be especially relevant for stylistic tasks (e.g., authorship attribution) that should rely on style information, and for semantic tasks (e.g., reasoning) that should not rely on style information (Wegmann et al., 2025). Disentanglement might work especially well with mixture-of-experts approaches (Artetxe et al., 2022), with style-specific architectures (e.g., tokenizers) for relevant experts.

Considering style in annotations Human-written texts and labels can include spurious correlations as a result of annotation instructions

(Gururangan et al., 2018). Style representations could be used to monitor the output of annotation efforts, and ultimately, to distinguish instructions that evoke more stylistically diverse annotations.

Bias identification and reduction As mentioned (§ 1), language models are often biased against certain styles, including those associated with marginalized groups. Approaches detailed in § 5, like curating training and evaluation datasets with more diverse styles, can improve performance across styles and thus reduce model bias. Further, it might be possible to use style representations to identify biased behavior of a trained model: For example, representations might be used to cluster question texts of similar styles and compare model performances across them.

Develop style measures With style measures we mean the broader class of methods and metrics that include style representations. One might, for example, develop a metric that measures the formality of a text, returning values between 0 and 1. Style representations are similarly quantitative measures of stylistic properties, but they typically encode (latent) stylistic dimensions in a vector space. In this study, we focus on style representations, but they can be applied to develop style metrics.

F.1 Open questions in the application of style representations

We add open challenges in the application of style representations to different problems.

Circular evaluation in style transfer When performing generative tasks conditioned on style representations, such as authorship style transfer, difficulties can arise when comparing models. Various works (Horvitz et al., 2024a,b; Khan et al., 2024) train authorship style transfer models with the aid of style embedding models (§ 3.2) but also evaluate the adherence to the target style using style embedding models. When comparing two systems like ParaGuide (Horvitz et al., 2024a) and StyleMC (Khan et al., 2024), the former trained with CISR embeddings (Wegmann et al., 2022) and the latter with LUAR embeddings (Rivera Soto et al., 2021), it remains unclear which embedding space to use for evaluation without giving either model undue advantage. We encourage the community to investigate additional possibilities for evaluation (e.g., based on predefined features, cf. § 4) or establish a standard representation for training as well as evaluation.

Should we even care about styles for user-facing LLMs? Some recent work shows that more human-like outputs by LLMs might be dispreferred by humans and might lead to increased anthropomorphism (Cheng et al., 2025; Sandoval et al., 2025). This hints at a complex set of desiderata NLP researchers should consider when building LLMs and when using representations to steer LLMs toward generating texts in different styles. However, what style of output is preferred remains highly contextual (i.e., dependent on the setting) (Sandoval et al., 2025), and we believe that training on stylistically diverse corpora remains essential for LLMs to understand and engage with diverse human styles.

G Intended use and licenses for used artifacts

We only use models and datasets for motivating examples in our survey. We discuss their licenses and intended use below.

G.1 Datasets

s1k We use the s1k dataset provided by Muennighoff et al. (2025) and accessed at https://huggingface.co/datasets/simplescaling/s1K-1.1. The dataset was shared with an MIT license, which we adhere to.

MRPC We use the MRPC dataset provided by Dolan and Brockett (2005). The dataset is available on the Microsoft website at https://www.microsoft.com/en-us/download/details.aspx?id=52398. No license information is easily available. However, it is a widely used and shared dataset, and the paper mentions it is for the express purpose of stimulating research.

Math500 We use the Math500 dataset provided by Lightman et al. (2023). It was shared with an MIT license by OpenAI. See https://github.com/openai/prm800k/.

Style Embedding Data We use the Style Embedding data provided by Wegmann et al. (2022) at https://huggingface.co/datasets/AnnaWegmann/StyleEmbeddingData. No license information is easily available but it has been publicly shared.

ETS Corpus of Non-Native Written English We use the ETS Corpus of Non-Native Written English (also known as TOEFL11 or LDC2014T06) provided by Blanchard et al. (2014). It is accessed via the Linguistic Data Consortium (LDC) at https://catalog.ldc.upenn.edu/LDC2014T06. The dataset is distributed under a specific LDC user license agreement restricted to non-commercial research use, which we adhere to.

G.2 Models

CISR We use Wegmann et al. (2022)'s CISR model at https://huggingface.co/AnnaWegmann/Style-Embedding. No clear license information is given, but the model was published in a research paper encouraging further use.

LUAR We use Rivera Soto et al. (2021)'s LUAR model at https://huggingface.co/rrivera1849/LUAR-MUD, shared with an Apache 2.0 license, which we adhere to.

SBERT We use an SBERT (Reimers and Gurevych, 2019) semantic representation model, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, shared with an Apache 2.0 license, which we adhere to.

Mistral We use Jiang et al. (2023)'s Mistral-7B-Instruct-v0.1 model, https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1. The model was shared with an Apache 2.0 license, which we adhere to.

s1 models We use Muennighoff et al. (2025)'s fine-tuned Qwen models on Gemini (https://huggingface.co/simplescaling/s1-32B) and DeepSeek (https://huggingface.co/simplescaling/s1.1-32B). Both models are shared with an Apache 2.0 license, which we adhere to.

StyleDistance We use Patel et al. (2025)'s StyleDistance model at https://huggingface.co/StyleDistance/styledistance. It was shared with a MIT license, which we adhere to.

H Identifying or offensive content in datasets

We use small existing datasets only for motivating examples (see § B). We do not release datasets. The small used Reddit slice might contain some offensive content. We do not expect the other used datasets (§ G.1) to include offensive content as they are reasoning datasets, crowd-worker created paraphrases, and TOEFL essays. However, the TOEFL essays might include some personally identifying content. We did not take steps to remove identifiable cues or offensive content. We hope that the effect is negligible as the datasets were already publicly accessible and we only use them as motivating examples.

I Use of AI Assistants

We used ChatGPT, GitHub Copilot, and Claude Code for coding, to look up commands, and to generate individual functions for plotting. Generated functions were tested w.r.t. expected behavior. We used AI assistants (mostly Claude and ChatGPT) for concise rephrasing and grammatical error correction in writing.

J Model Size and Budget

We only ran motivating experiments (see models and experimental setup in §B). Note that the experiments are not part of our main contributions. We used a total of ≈1GPU hour with one A100.