



Eylon Caplan, Tania Chakraborty, & Dan Goldwasser

Department of Computer Science

Purdue University

West Lafayette, IN, USA

{ecaplan, tchakrab, dgoldwas}@purdue.edu

Abstract

Variation in language use, shaped by speakers’ sociocultural background and specific context of use, offers a rich lens into cultural perspectives, values, and opinions. For example, Chinese students discuss *healthy eating* with words like *timing*, *regularity*, and *digestion*, whereas Americans use vocabulary like *balancing food groups* and *avoiding fat and sugar*, reflecting distinct cultural models of nutrition (Banna et al., 2016). The computational study of these Sociocultural Linguistic Phenomena (SLP) has traditionally been done in NLP via tailored analyses of specific groups or topics, requiring specialized data collection and experimental operationalization—a process not well-suited to quick hypothesis exploration and prototyping. To address this, we propose constructing a “sandbox” designed for systematic and flexible sociolinguistic research. Using our method, we construct a demographically/topically split Reddit dataset, **SPLITS!**, validated by self-identification and by replicating several known SLPs from existing literature. We showcase the sandbox’s utility with a scalable, two-stage process that filters large collections of *potential* SLPs (PSLPs) to surface the most promising candidates for deeper, qualitative investigation.¹

1 Introduction

How we speak is fundamental to who we are. Language is a powerful window into culture, allowing us to investigate the values and shared perspectives that shape a community’s identity (Bucholtz and Hall, 2005). Such indicative language manifests across many settings, in dialects with distinct grammar and vocabulary, like African American Vernacular English (AAVE) (Tia and Aryani, 2020a; Smitherman, 2007), but may also appear in more subtle, context-specific differences. For example,

¹We release our [code](#), our [data](#), and a [sandbox demo](#).

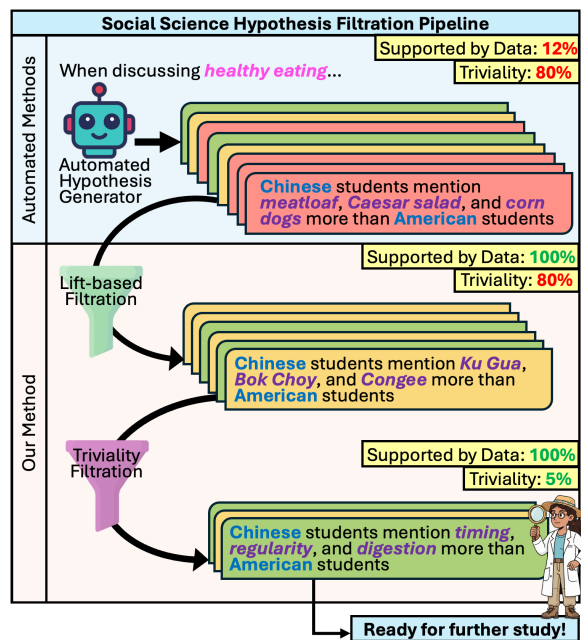


Figure 1: Automated methods can generate a large pool of hypotheses. Our goal is to surface the ones that are likely to be true (*supported by data*) and interesting (*non-trivial*). E.g., Chinese students mentioning Chinese food items is not surprising, but emphasis on *timing* and *regularity* in the context of “healthy eating” offers cultural insights.

Figure 1 shows an example of a *sociocultural linguistic phenomenon* (SLP) wherein Chinese and American students discuss “healthy eating” with different vocabulary, reflecting how Americans are influenced by media/school nutrition messages, whereas Chinese students are influenced by Traditional Chinese Medicine (Banna et al., 2016).²

Computational social science provides powerful tools to study specific SLPs within social media, yielding valuable insights into how different groups view and discuss contexts like politeness (Li et al., 2020; Havaladar et al., 2025), values/moral foundations (Roy and Goldwasser, 2023; Borenstein et al.,

²This work includes sensitive issues; discussion of demographics, representation, and misuse is in the Ethics Statement.

2024; Caplan and Goldwasser, 2025), individualism (Havaldar et al., 2024), sustainability (Reuver et al., 2024), and health policy (Alliheibi et al., 2021; Pacheco et al., 2023). These are often tailored, deep dives of a single group/context, requiring significant effort in specialized data collection and experimental operationalization. Such methods are rigorous and powerful, but costly, making them difficult to extend to quick hypothesis exploration and idea prototyping.

The exhaustive and low-cost nature of a quick hypothesis exploration system would enable initial investigations of previously under-studied groups and contexts without requiring significant upfront investment. These investigations could identify promising phenomena worthy of deeper, bespoke study. To enable this rapid idea exploration, we envision a “sandbox” for systematically exploring the hypothesis space. Specifically, we seek to span the space of *groups* that use language, and the space of *contexts* in which language is used.

To this end, we introduce a method for constructing such a “sandbox” using Reddit: with which we create the **SPLITS!** dataset—split by both **user demographics** and **use contexts** operationalized as discussion topics (e.g., photography, travel, humor). We demonstrate its flexibility across various analytical approaches; for instance, we demonstrate one straightforward application by using **SPLITS!** to confirm well-documented SLPs, such as *code-switching* patterns of Black AAVE speakers.

Recent work has focused on automatedly *generating* scientific hypotheses, including social science hypotheses (Yang et al., 2024; Manning et al., 2024; Peterson et al., 2021), yet a key bottleneck remains: how can a social scientist efficiently sift through thousands of computer-generated ideas to find the ones worth pursuing (Figure 1)? As a second application of **SPLITS!**, we address this by filtering for “promising” phenomena (worthy of further study) among a large pool of *potential* SLPs (PSLPs).

To do this, we propose a scalable, two-step process (Figure 1) that takes as input a large pool of candidate PSLPs, and surfaces ones which are (1) supported by the data in the sandbox and (2) more likely to be unexpected, avoiding trivial PSLPs like “Chinese students mention Chinese foods more than Americans”. Our contributions are as follows:

- Propose a method for constructing a flexible, extensible “sandbox” for sociocultural linguistic investigation, and use it to create **SPLITS!**, a 9.7 million-post dataset.

- Demonstrate the dataset’s flexibility by reproducing known, literature-backed phenomena (AAVE code-switching and others).
- Propose a two-step process using *lift* and *triviality* to surface promising hypotheses from a large pool—intended for further expert investigation.

2 Background and Related Work

2.1 SLPs and Variation

We define a *sociocultural linguistic phenomenon* (SLP) as any distinctive speech characteristic exhibited by a particular group. This concept is rooted in the field of sociolinguistics, which seeks to find “correlations between social structure and linguistic structure” (Gumperz, 1971). Specifically, our work connects to **variationist sociolinguistics**, which focuses on the social evaluation and use of linguistic variants through hypothesis-formation and statistical testing (Chambers, 2002).

Traditionally, variationist studies focus on how different linguistic forms are used to express the **same meaning** (e.g., phonetic variants of [r]) across different social groups (Wardhaugh and Fuller, 2021). However, more recent “third-wave” approaches to variation have expanded this view to include **stancetaking**, where interlocutors use language to position themselves relative to the topic, each other, and broader social identities (Eckert, 2012; Jaffe, 2009). In this broader view, the fixed variable is not the meaning but the **context of use**, and the observed variation can be syntactic, lexical, or even semantic. Our definition of an SLP aligns with this broader conception: we hold the topic of discussion constant and analyze how different demographic groups vary their language to express different perspectives, priorities, and stances.

2.2 Hypothesis Testing in Sociolinguistics

Sociolinguistics is an empirical science, where the study of SLPs traditionally begins with **expert-led hypothesis formation**. Hypotheses are formed based on deep ethnographic knowledge gathered through methods like fieldwork and interviews (Wardhaugh and Fuller, 2021; Hernández-Campoy, 2014). More recently, computational social science (CSS) has leveraged large-scale social media data to complement this process, enabling the generation and evaluation of social science hypotheses at a new scale (Yang et al., 2024; Manning et al., 2024; Peterson et al., 2021). While the source of data (e.g., interview transcripts vs. online posts) and

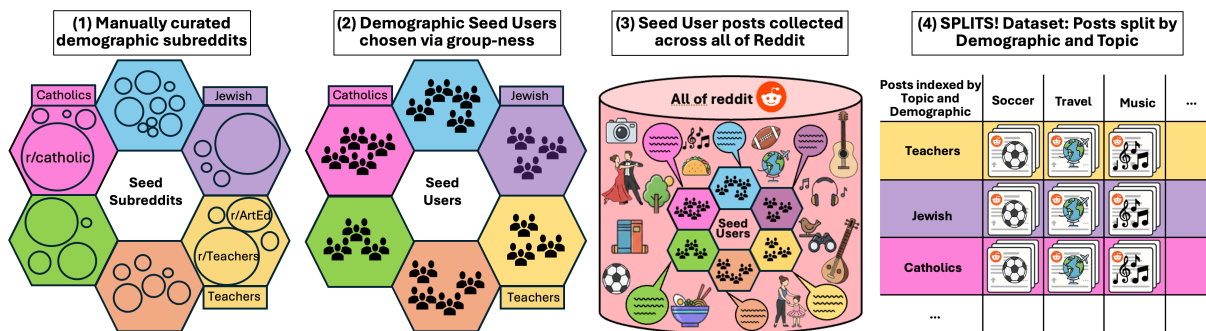


Figure 2: **SPLITS! dataset creation:** (1) Set of seed subreddits are chosen for each demographic using manual inspection aided by statistical similarity metrics. (2) From all users in the seed subreddits, those with high "group-ness" are selected as seed users. (3) We collect all posts made by the seed users across all of reddit. (4) Posts are then labeled by content topic (independent of the source subreddit). The final dataset consists of these posts aggregated by demographic as well as topic.

hypotheses (expert-derived vs. machine-generated) may differ, the core analytical step of using **correlational studies** and **statistical evaluation** to validate findings remains consistent.

2.3 Lexica

In this work, we represent SLPs as lexica. A *lexicon* is a curated collection of words or phrases widely used for interpretable models in the social sciences (Hayati et al., 2021; Pryzant et al., 2018; Boyd et al., 2022). Lexica are a scalable and explainable method for analyzing large datasets (Havaladar et al., 2024) and have been used to study sentiment, emotions, and mental health (Geng et al., 2022). This approach is distinct from feature importance analysis (e.g., SHAP) (Ribeiro et al., 2016; Kim et al., 2020; Lundberg and Lee, 2017), which aims to interpret a model’s predictions. Our focus, in contrast, is on using lexica to directly test hypotheses about the world as reflected in the data (Geng et al., 2022; Havaladar et al., 2024).

2.4 Perspectivism and Social Media Corpora

A significant body of sociolinguistically informed NLP research adopts a "Perspectivist" view: that language varies with a speaker’s background and context (Nguyen et al., 2021; Blodgett et al., 2020; Liu et al., 2025; Joshi et al., 2024; Frenda et al., 2024). This view has spurred the creation of perspective-aware datasets (Fleisig et al., 2024; Aroyo and Welty, 2015) and models for tasks like stance detection (Santurkar et al., 2023; Ceron et al., 2024; Pujari et al., 2024). Many of these resources are social media corpora curated to study identity, often for tasks like demographic inference or bias detection (Sachdeva et al., 2022; Wood-Doughty et al., 2021; Sap et al., 2020; Nadeem

et al., 2020; Preoțiu-Pietro and Ungar, 2018; Tiginova et al., 2020). Our work builds on this tradition but differs in two key ways: our focus is on creating data splits by both demographic group and topic, and we do not attempt to attribute demographic authorship based on linguistic content.

3 SPLITS! Dataset Creation Method

In this section, we present our method for constructing a Reddit “sandbox” for systematic sociolinguistic research. For the systematic approach, we ground our dataset’s design in the two axes of variation described by Maclagan (2005):

- **User-level factors:** a speaker’s long-term attributes (ethnicity, occupation, religion). We instantiate this axis using user **Demographic**.
- **Use-level factors:** the context of an utterance (setting, purpose). We instantiate this axis by splitting our data by discussion **Topic**.

As such, we define a **demographic** as a collection of people with a shared, long-term user-level quality. We construct our dataset, **SPLITS!**, from a corpus of Reddit data spanning 2012–2018 (Chang et al., 2020), collecting the top 50k subreddits by size. The final dataset contains 9.7 million posts, split across the 6 demographics and 89 topics.

3.1 Curating Demographic subreddits

In order to build the dataset, it was crucial to obtain posts that were truly written by some demographic group. To do this, we first identified subreddits that would be *almost exclusively* used by people of some demographic, a set of subreddits which we refer to as the *seed set*, shown in the first panel of Figure 2. For example, for the Catholic demographic, we included the subreddit r/CatholicDating, as it is likely to contain a very

high proportion of Catholics. We define a *user* of a subreddit as a Reddit account which at some point commented or posted in that subreddit. We began with a single high-quality seed subreddit and iteratively expanded this set by computing the user-overlap similarities (Jaccard, Cosine) to find related subreddits, which were then manually reviewed for inclusion. This process continued until no new relevant subreddits were found (visualized in Figure 7, App. A.1).

We manually crafted 6 demographic seed sets: 2 ethnicities (African American/Black, Jewish descent), 2 occupations (teacher, construction worker), and 2 religious groups (Catholic, Hindu/Sikh/Jain). These demographics were chosen based on having relatively clean seed sets and being within an order of magnitude in total number of users. Broader discussion of demographic choices, including dimensions like gender and age are in App. A.1, with exact user and post counts.

3.2 Demographic Seed Users via Group-ness

The next step, as shown in part 2 of Figure 2 was to choose the *seed users* for each demographic. For each demographic, taking all posts in the union of the seed subreddits gave a seed demographic corpus SD , e.g. SD_{catholic} . We began by taking all unique users of the posts in $SD_{\text{demographic}}$, and got a demographic user set U , e.g. U_{catholic} for SD_{catholic} . To make sure that the users of these posts truly belonged to the demographic, we devised a metric to measure a user’s likely ‘group-ness’. The final seed user set for each demographic only consists of users that pass the group-ness threshold. The intuition was that a user is more likely to be in the target demographic if (1) they have many posts in the demographic seed set and (2) their posts are spread out across several subreddits (which we viewed as used *almost exclusively* by the target demographic). We do this with the metric in Eq. 1, in which we reward the total amount of activity *and* diversity across subreddits in the seed set.

$$\text{group-ness}(u) = \sum_{s \in S_D} \log(1 + c_{u,s}) \quad (1)$$

Here u is a user, S_D is the set of seed subreddits for the demographic D , and $c_{u,s}$ is the count of posts by user u in subreddit s . We hypothesize that users with higher group-ness metric scores are likelier to be members of the target demographic. This is further validated in the next subsections.

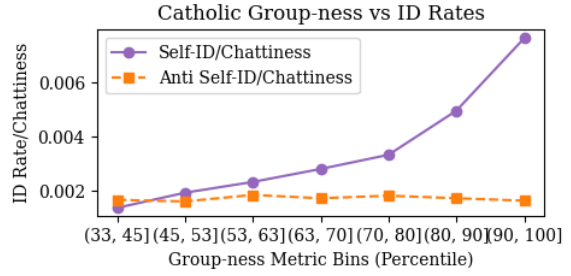


Figure 3: Does increasing group-ness isolate target demographic users? We plot self-identification rates vs. group-ness of the Catholic demographic, showing that it does.

3.3 Collecting Seed User Posts Across Reddit

Once we had the set of seed users for each demographic, the next step was to gather all the posts by the users on a variety of topics (Fig 2, part 3). We tracked the seed users across Reddit and collected all of their posts among **all subreddits** (not just the seed set). This yielded a set of posts C for each demographic, where $SD_i \subset C_i$ for each group i .

Validating group-ness. We wanted to ensure our group-ness metric guaranteed that the authors of the posts for each C_i truly belonged to the demographic i . To test this, we created a set of self-identification phrases (e.g. for Catholic, “I am a Catholic”, “I’m Catholic”, etc.), and anti-self-identification phrases (e.g. “I’m not a Catholic”, etc. and also “I’m a Baptist”, “I’m Jewish”, etc.). Some phrases can be found in Appendix A.1. Next, we searched for these phrases among all posts in a demographic’s C , and to avoid false positives (e.g. quoted phrases, sarcasm, irony), we used an LLM to verify the context of these phrases (see prompt in App. E, human validation in App. A.4). Finally, we checked to see whether the group-ness metric correlated with higher self-identification rates, and whether it negatively correlated with anti-self-identification rates. To account for users with extremely high post count (chattiness) self-identifying very often, we normalized the self-identification rate by the average chattiness.

Figure 3 shows the result for the Catholic demographic. Similar successful validation trends were observed for all six retained demographics (see Appendix A.1 for full plots). We note that we attempted a 7th demographic (Korean) which failed this validation due to insufficiently exclusive seed communities, highlighting a limitation of our methodology (see App. A.2). For the successful groups, past some threshold (usually about

the 75th percentile in the group-ness metric), as a user’s group-ness metric increases, their likelihood of self-identifying increases, and their likelihood of anti-self-identifying (i.e. saying they *aren’t* in the target demographic) goes down or stays the same. These thresholds of group-ness are chosen heuristically, as they reflect an inherent, necessary trade-off between demographic purity (precision) and dataset volume (recall). We select these manually based on empirical separation, and the exact threshold values are reported in our released code.

We make the following assumptions: (1) Redditors are generally honest when self-identifying, (2) a user’s probability of explicitly self-identifying or anti-self-identifying in a given post is independent of their total posting frequency, and (3) honestly stating a self-identification or anti-self-identification phrase typically indicates genuine membership or non-membership. With these assumptions, the increasing separation we see between self-identification and anti-self-identification must mean that **users with a higher group-ness metric level are likelier to be in the target demographic**. That is, past some threshold, the datasets are ‘cleanly’ posted by the target demographic.

This approach inherently focuses our analysis on users **highly engaged in identity-centric subreddits**; while this creates a known selection bias, it provides a robust, high-confidence proxy for studying the language of these active online communities, a trade-off further discussed in Limitations.

Intersectionality. As with all identities, the demographics we have selected may overlap in any individual. In practice, our dataset consists of demographics selected primarily for volume and exclusivity of subreddits. While some intersections are nearly nonexistent due to mutually excluding definitions (e.g. Catholic and Hindu/Jain/Sikh), others could have nontrivial intersections.

To assess the empirical intersectionality of the selected demographic groups, we computed pairwise Jaccard similarity between sets of users in each group (full heatmap in App. A.1). Given the minimal empirical overlap observed (the greatest Jaccard was 0.0081), we chose to analyze the demographic groups separately in this work. We stress that this is a consequence of our sampling methodology and not a claim about the non-existence of intersectional identities in the real world. Acknowledging this limitation, we leave the computational study of intersectional SLPs as critical future work.

Topic	Post Count		AAVE Use (% posts)	
	Black	Non-Black	Black	Non-Black
Hip-Hop	56k	18k	3.16* [†]	2.00*
Professional	101k	673k	0.33 [†]	0.23

Topic	Post Count		Yiddish Use (% posts)	
	Jewish	Non-Jewish	Jewish	Non-Jewish
Judaism	97k	64k	0.19* [†]	0.07*
Professional	135k	639k	0.01 [†]	0.01

Topic	Post Count		‘Dance’ Use (% posts)	
	Hindu/Sikh/Jain	Non-Hindu/Sikh/Jain	Hindu/Sikh/Jain	Non-Hindu/Sikh/Jain
Personal Cultural Identity	88k	381k	0.44*	0.36

Table 1: Can **SPLITS!** replicate five known SLPs using lexical proportions? We show lexicon usage by demographic; *denotes p-value < 0.001, between the two demographics (within topic). [†]denotes significance for the primary demographic’s code-switching across topics. We show that all five SLPs are replicated.

3.4 Topics

The final step was to label the posts for their topics. To create topic splits, we began with 11 categories (e.g., ‘Sports’, ‘Entertainment’), using an LLM to generate about 20 specific topics (e.g., ‘basketball’, ‘sci-fi’) and corresponding query keywords for each (App. A.1). For each demographic’s post collection (C_i), we filtered for users with high ‘group-ness’ scores, setting the threshold for each group based on the observed separation between self- and anti-identification rates (Fig. 3). We then used the keywords to retrieve documents from these users’ posts with the ColBERT retrieval model (Khattab and Zaharia, 2020). Finally, we used an LLM-based system to assess the topic relevance of each post (prompt in App. E, human validation in App. A.5). We removed topics with too few posts after this step. The final dataset contains 9.7 million posts across 89 topics. We denote the set of posts for demographic d on topic t as $C_{d,t}$. A visual of the final **SPLITS!** dataset is shown in part 4 of Fig 2.

3.5 Replicating Case Studies with **SPLITS!**

To validate the richness and demonstrate one method of using **SPLITS!** for answering sociocultural linguistic questions, we analyzed 5 literature-backed SLPs and show they are captured. For brevity, detailed explanations of each phenomenon, including literature and lexica are in App. B.

Table 1 shows the results of using **SPLITS!** to count post proportions that use lexica from existing literature. Analysis with our dataset replicates:

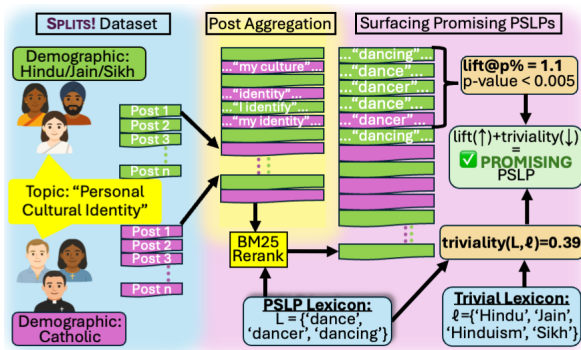


Figure 4: Main steps to determine if a PSLP is *promising* using **SPLITS!**: (1) Aggregate posts from 2 demographics about the same topic (2) Use the PSLP’s lexicon to rerank the aggregated posts (3) Compute lift and trivality for the PSLP (4) Promising PSLPs achieve high lift & low trivality.

- AAVE³: Black users use AAVE features more frequently than non-Black users
- AAVE code-switching: Black users themselves use AAVE more when they discuss ‘Hip-Hop’ than ‘Professional’ topics
- Yiddish: Jewish users use Yiddish terms more frequently than non-Jewish users
- Yiddish code-switching: Jewish users themselves use Yiddish terms more when they discuss ‘Judaism’ than ‘Professional’ topics
- ‘Dance’ as Identity in South Asians: when discussing ‘Personal Cultural Identity’, Hindus/Jains/Sikhs mention ‘dance’ and ‘dancing’ more than non-group members

4 Two-Stage PSLP Filtration Process

This section details another method of using **SPLITS!**: quickly filtering for PSLPs’ (1) alignment with data and (2) non-triviality—worthiness for future study. The method relies on the *specific class* of PSLPs testable by this lexical form: “demographic A uses set L of words and phrases more than demographic B when discussing topic t ”. We represent this as $\text{PSLP}_{L,A,B,t}$.

Figure 4 shows the high level: we propose pairing the corpora of 2 demographics discussing the *same* topic, indexing them together, and measuring how much better than random a lexicon is able to rerank the set. This method is motivated by the *speed* and *reusability* conferred by indexing for rapid testing of PSLPs. We stress that **not all SLPs can be put in this form**; our pipeline is intentionally scoped to test lexical-level hypotheses for tractability and interpretability.

³African American Vernacular English

4.1 Quantifying Validity with Lift

To ensure enough posts for each topic, topic/demographic post sets $C_{d,t}$ with fewer than 2,000 posts were dropped. Then, for every topic t , and two demographics A, B , the two post sets $C_{A,t}$ and $C_{B,t}$ were combined and indexed together using the pyserini BM25 implementation (Lin et al., 2021). A lexical method (BM25) was chosen as a non-‘black box’ model, conferring interpretability compared to neural models. Hence there were 781 unique indices $I_{A,B,t}$ for every combination of demographics A, B and topic t . App. A.1 visualizes the pairings, the number of posts by demographic, and by topic category.

Given a PSLP’s lexicon L , we can then test its validity over our dataset by doing the following: Using the relevant index $I_{A,B,t}$ to the PSLP, we rerank the whole set using L as a query and the phrase-aware BM25 algorithm (Figure 4). To measure L ’s success in pulling demographic A upward, we use the metric of *lift* from Data Mining (Tufféry, 2011). We define lift@ $p\%$ of demographic A in Eq. 2.

$$\text{lift}@p\% = \frac{\#A \text{ posts}@p\% / \# \text{ posts}@p\%}{\#A \text{ posts overall} / \# \text{ posts overall}} \quad (2)$$

Here $p\%$ is the top p percent of posts in the ranking. Lift@ $p\% > 1$ indicates that L has pulled A over B more than random. To guarantee significance, we also perform a one-tailed hypergeometric test, which is the exact distribution of randomly selected top $p\%$ posts. The magnitude of the lift indicates the strength of L pulling up A over B , and the p-value indicates the significance of the PSLP overall. Throughout this paper, we report lift@0.5%, though we compute it for 1%, 2%, 5%, and 10% as well. Generally, smaller values of $p\%$ capture more rare/subtle phenomena, while higher values only capture very prominent phenomena.

4.2 Quantifying Triviality

High significance lift is not enough to indicate a promising PSLP for future study. Despite each index being limited to a specific topic, any demographic is inherently more likely to use words that are ‘trivial’ to that demographic. By this, we mean words that are definitional in nature (e.g. “Jewish”, “Jew”, “Judaism” for the Jewish demographic), or words that are exclusive/nearly exclusive to the demographic (e.g. “Diwali”, “kirpan”, “ahimsa” for the Hindu/Jain/Sikh demographic).

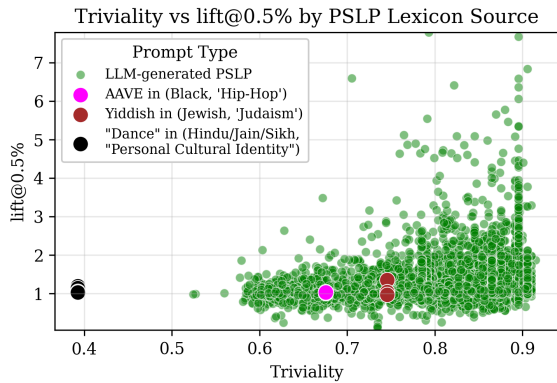


Figure 5: Are higher-lift PSLPs more trivial? We plot LLM PSLPs (subsamped) along with case studies. The upward trend indicates many obvious hypotheses achieve high lift.

To filter out trivially lifting PSLPs, we propose a ‘triviality’ metric. At a high level, this measures the similarity between the PSLP’s lexicon L and the target demographic A (Figure 4). To operationalize this, we manually authored a small lexicon ℓ for each demographic, containing 5-10 terms (e.g. ℓ_{jewish} as “Jewish”, “Jew”, “Judaism”, “Jewish holidays”, ...; see App. C for full lists).

We use the subspace recall-like ‘ $R_{subspace}$ ’ score introduced in Ishibashi et al. (2024) to measure the similarity between two sets of words (utilizing bert-base-uncased embeddings), scored within $[0, 1]$. We precisely define triviality as

$$triv(PSLP_{L,A,B,t}) := R_{subspace}(L, \ell_A). \quad (3)$$

As such, the more words in the lexicon L that are semantically similar to the target demographic A as a whole, the more trivial it becomes. We note that the ‘triviality’ metric can be easily tuned to fit specific use cases. We acknowledge that with this definition, *non*-triviality does not perfectly map onto “interesting-ness” or “unexpectedness”, but in sections § 5.2 and § 5.3, we show that it correlates with human judgments of “unexpectedness”, and the “interesting-ness” of research findings.

5 Validation of Lift and Triviality

We now motivate and validate our metrics.

5.1 Why Statistical Significance is Insufficient

To demonstrate our filtration process’s ability to filter a large and noisy set of candidates, we used an LLM to generate over 23,000 PSLPs across all demographic-pair-topic combinations. We stress

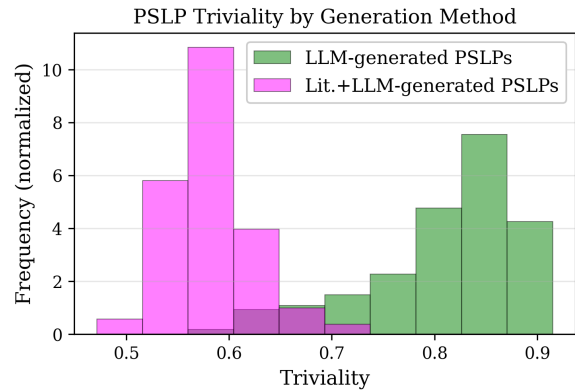


Figure 6: Are PSLPs inspired by social science findings less trivial? We plot the triviality of these PSLPs (pink) against normal LLM PSLPs (green), showing they are far less trivial.

that **our goal was not to design a novel hypothesis generator**, but to produce a large, diverse set of candidate PSLPs for our proposed filtration process at scale. Further candidate details in Appendix C.

Figure 5 plots the resulting Lift versus Triviality score for each of these 23,000+ PSLPs, with our case studies from § 3.5 included for reference (case study demographic vs. demographic lifts are in App. B). The plot reveals the central motivation for our proposed filtration process: there is a significant positive correlation (0.32 Spearman for lift@0.5%) between a PSLP’s triviality and its statistical lift. This finding confirms that **relying on statistical significance alone is insufficient**, since **many commonsense or obvious hypotheses also achieve high lift**. This necessitates our process’s second stage to isolate the more promising, non-trivial candidates for expert review.

5.2 Human Validation of the Triviality Metric

To validate our automated Triviality metric, we measured its alignment with human judgments of “unexpectedness.” We conducted a study where 9 demographically diverse annotators, including members of 4 of the 6 groups in our dataset, rated 500 PSLPs. For each PSLP (a lexicon, a demographic pair, and a topic), they rated its unexpectedness on a 1-5 scale (5 being highly unexpected). The annotation task proved reliable, achieving an Intraclass Correlation Coefficient (ICC(2,k)) of 0.74, indicating good agreement. As hypothesized, we found a significant negative correlation (Spearman’s $\rho = -0.38$) between our Triviality score and the average human score. This confirmed that our metric can be a useful (though imperfect) heuristic filter for prioritizing unexpected candidates.

Target	Contrast	Topic	Lexicon	Triviality (percentile) ↓	Human Score ↑ [1, 5]
Hindu/Jain/Sikh	Teacher	Books/Literature	powerful woman, making her own choices, emotional complexity, ...	8.80	*
Jewish	Catholic	Healthcare	early detection, preventative care, screening, proactive, ...	12.10	4.67
Catholic	Black	Healthcare	papal authority, church teachings, bishop, doctrine, catechism, ...	99.38	1.00

Table 2: What do filtered PSLPs look like? Row 1: literature-inspired PSLP (Jain, 2016) (interesting), Row 2: LLM-generated PSLP (non-trivial, unexpected), Row 3: LLM-generated PSLP (trivial, expected).

5.3 Validating Triviality Against Published Sociocultural Research

To further validate the Triviality metric, we test its ability to distinguish standard LLM-generated hypotheses from those grounded in existing academic literature. Our rationale is that *findings published in peer-reviewed social science papers represent a reliable proxy for what domain experts consider ‘interesting’*. If our Triviality metric is effective, it should assign significantly lower scores to hypotheses derived from expert-vetted sources.

Literature-Inspired PSLPs. To construct this test set, we manually curated 11 social science papers by searching Google Scholar for literature on specific demographic-topic intersections present in our dataset (e.g., ‘Teacher Anime/Manga’, ‘Catholic Video Games’) (see App. F). Selected papers had findings that were directly relevant to the target demographic and topic. For each paper, we prompted an LLM to first summarize its core findings and then, based on that summary, generate 10-15 lexica that operationalize the paper’s conclusions (prompt in App. E). The result was 132 lexica. A human validation study confirmed that these generated lexica faithfully reflect the findings of their source papers (see App. F.1). To create testable PSLPs, we paired these lexica with all available contrast demographics in our dataset, yielding 591 total *Literature-Inspired PSLPs*.

Figure 6 compares the Triviality score distributions for standard LLM-generated PSLPs vs. our Literature-Inspired PSLPs. The distinction is stark: **Literature-Inspired PSLPs exhibit a markedly lower distribution of Triviality scores** (mean of 0.585 vs. 0.810). This result demonstrates that our Triviality metric generally *aligns with the concept of academic sociocultural ‘interestingness,’* serving as a valuable heuristic to prioritize hypotheses worthy of deeper investigation.

6 Surfacing Promising PSLPs

We now apply our pipeline on a large pool of LLM-generated hypotheses. We quantify the efficiency

of our two-step filtration and then present a qualitative analysis of some surfaced PSLPs.

6.1 Quantitative Analysis of Filtration

Our pipeline is designed to help researchers efficiently navigate vast hypothesis spaces. To demonstrate this, we applied our two-step filtration process to the over 23,000 PSLPs generated by an LLM. First, we filtered for statistical significance ($p < 0.025$), which by itself **reduced the candidate pool by over 10x**, identifying $\sim 2,300$ PSLPs that are supported by the data. The subsequent challenge is to find the truly “promising” candidates within this set—those that are not only supported by data but also deemed unexpected by human experts (avg. score $> 3/5$).

This is where our Triviality metric provides a crucial second filter. By using it to prioritize the inspection of the $\sim 2,300$ supported candidates, we reduce the manual “effort” (number of PSLPs inspected per promising find) by **an additional 1.5-1.8x** (Table 3). In this particular LLM-based generation scenario, the full two-step process provides a **combined 15-18x reduction in effort**, showcasing its value in helping researchers quickly isolate a high-potential set of hypotheses from a large, noisy initial pool for deeper qualitative study.

6.2 Qualitative Analysis of Promising PSLPs

Table 2 showcases significantly lifting PSLPs surfaced by our pipeline, illustrating the distinction between those deemed promising and those filtered out as trivial. A particularly compelling, non-trivial example suggests that Jewish users, when discussing healthcare, employ a lexicon of ‘preven-

Percentile threshold	Precision	Recall	F1	Effort (# inspected/promising)	Speed-up	p-value
1 (Baseline)	0.270	1.000	0.425	3.70	1.00×	–
0.1	0.480	0.178	0.259	2.08	1.78×	0.001
0.2	0.455	0.341	0.390	2.20	1.68×	$< 10^{-3}$
0.3	0.447	0.496	0.470	2.24	1.65×	$< 10^{-3}$
0.4	0.425	0.630	0.507	2.35	1.57×	$< 10^{-3}$
0.5	0.398	0.741	0.518	2.51	1.47×	$< 10^{-3}$

Table 3: Triviality used for ‘unexpectedness’ classification of LLM PSLPs. The Percentile Threshold is the Triviality cut-off for classifying “unexpected”.

tative care’, ‘early detection’, and being ‘proactive’ significantly more than Catholic users.

This linguistic pattern may point to deeper, culturally-ingrained perspectives and can provide a concrete, data-driven **starting point** for social scientists to further explore this space. For example, theological scholarship contrasts Judaism’s focus on the present world with Christianity’s historical emphasis on the afterlife (McDermott, 2015; Schwartz, 2011; Eckardt, 1972). We stress again that PSLPs are only *potential* SLPs—a **first step** for further study, not foregone conclusions.

7 Conclusion

We presented a method for building a flexible sociolinguistic “sandbox” for the rapid exploration of cultural language use, creating the **SPLITS!** dataset as an instance. We demonstrated its utility through a two-stage filtration process that uses *lift* and *triviality*. This approach effectively narrows a vast space of computer-generated hypotheses to a manageable set of promising candidates, bridging the gap between large-scale computational analysis and nuanced, expert-led sociolinguistic inquiry.

Future work could attempt to capture richer linguistic dimensions beyond lexica, such as syntactic structures, semantic framing, or pragmatic features. However, transitioning to more semantically complex or neural representations must be approached cautiously; allowing for non-lexical features introduces the risk of uninterpretable model bias.

Ethics Statement

Research into language and its relation to identity carries significant ethical responsibilities. We have sought to address these responsibilities throughout our research design, dataset curation, and framework development.

Risk of Stereotyping and Social Essentialism

This paper investigates linguistic variation across demographic groups. Any such work risks engaging in or enabling harmful stereotyping. We are particularly concerned with avoiding Social Essentialism—the fallacious belief that social groups possess distinct, inherent essences (Gelman, 2004; Rhodes et al., 2012). In AI, this can manifest as systems that unfairly generalize or reinforce societal biases (Allaway et al., 2023b).

Our methodology is explicitly designed to counter this risk. Following recent work in cog-

nitive science and NLP that advocates for more nuanced models of identity (Benitez et al., 2022; Allaway et al., 2023a), the **SPLITS!** dataset is structured not only by demographic but also by 89 distinct topics of discussion. This design is rooted in our guiding principle: **language is expressed and interpreted in nuanced ways that extend well beyond a person’s demographic membership**. By enabling analysis that is always conditioned on a specific context, our framework encourages a perspectivist view of language and has the potential to reveal nuanced differences that actively counter broad, essentialist stereotypes.

Data Source and Privacy

This work is built on publicly available Reddit posts sourced from the Convokit project (Chang et al., 2020). While the data is public, we acknowledge that users may not anticipate their posts being used in academic research. To mitigate privacy risks, the public release of **SPLITS!** does not contain Reddit usernames. To enable user-level analysis, we provide pseudonymized user IDs.

Intended Use and Prohibited Misuse

We provide this dataset and framework with clear guidelines on its intended and prohibited uses.

Intended Use: **SPLITS!** is designed for non-commercial, academic research into observational sociocultural linguistics. Its purpose is to help researchers explore how language use varies across different contexts and to generate hypotheses for further qualitative or quantitative investigation.

Prohibited Misuse: This dataset is not suitable for any application that assigns scores, makes judgments about, or could otherwise stereotype individuals based on their perceived group membership.

Limitations

We list the limitations of our work here.

Representational Bias. The data is sourced from Reddit and is not representative of the global population or even the full population of the demographic groups studied. Therefore it inherently contains selection bias.

The dataset reflects the language of English-speaking, active Reddit users within these communities. It cannot capture the full richness or complexity of their lived experiences and should not be interpreted as doing so.

Temporal Bias. The data spans from 2012 to 2018. Language, cultural norms, and the topics of online discourse evolve rapidly. This dataset provides a valuable historical snapshot, but the patterns observed may not reflect the contemporary linguistic practices or views of these communities.

Platform-Specific Discourse. Reddit has a unique culture with its own jargon, memes, and conversational norms. The linguistic phenomena identified may be intertwined with the specific communicative conventions of the Reddit platform and may not generalize to other social media contexts or offline conversations.

Demographic Labeling and Intersectionality. Our demographic labels are high-confidence proxies derived from user activity in “seed” subreddits, not from direct self-declaration. This method may select for users who are particularly active and vocal in identity-centric communities. A key risk of this approach is that **it may amplify perceived linguistic differences between groups**, as our sample may not be representative of the broader demographic, but rather a subset with stronger in-group identification. Furthermore, our broad demographic categories (e.g., “Black,” “Catholic”) do not capture the immense intra-group diversity. Our current framework treats these groups as discrete and is not designed to analyze intersectional identities, where multiple identity facets jointly shape language use. The computational study of such intersectional phenomena is a critical direction for future work.

Seed and ℓ Selection. Our methodology relies on two main manual decisions: demographic seed subreddits, and ‘trivial’ demographic lexica. For the demographic seed sets, we relied on subreddit names and descriptions. We believe this process is transparent and replicable by any informed researcher without requiring privileged “insider” knowledge of a group. Similarly, the Triviality lexica (ℓ_A) were constructed using terms that are fundamentally definitional to each demographic, ensuring broad agreement and minimizing subjectivity.

LLM-based Topic Creation. Our topic-splitting process uses LLMs for keyword generation and relevance filtering. While powerful, these models are not infallible and may introduce noise or systematic biases into the topic definitions, potentially affecting the validity of cross-topic comparisons.

PSLP Representation. Our framework operationalizes PSLPs as lexica of words and phrases. This was a deliberate design choice to ensure scalability and interpretability. However, this simplification cannot capture more complex phenomena rooted in deeper semantic or pragmatic nuance, such as sarcasm, narrative structure, or connotative meaning. Our framework should therefore be seen as a tool for identifying lexical-level differences, which can serve as a starting point for more nuanced qualitative or computational investigation.

LLM-Generated Hypothesis Bias. While our pipeline utilizes LLMs to rapidly generate a large pool of candidate PSLPs, researchers must remain highly critical of these models when generating hypotheses. LLMs inherently encode biases from their pre-training data and may hallucinate, reproduce harmful stereotypes, or over-represent majority viewpoints. Consequently, the hypotheses surfaced by our pipeline should never be taken as factual sociolinguistic claims, but rather treated strictly as exploratory starting points for rigorous human-led validation.

Embedding Bias in Triviality. Our automated Triviality metric relies on BERT embeddings to measure the semantic similarity between a PSLP lexicon and a target demographic lexicon. Because pre-trained models contain inherent representational biases, these embeddings may possess certain blind spots. Consequently, the Triviality metric might misjudge or fail to capture the semantic nuance of culturally specific, historical, or marginalized concepts, potentially filtering out valid phenomena or missing hidden trivialities.

Acknowledgments

We thank the anonymous reviewers for helping to strengthen the paper, and our annotators for generously volunteering their time. This work was supported by an NSF CAREER award IIS2048001.

References

- Co Airhihenbuwa, S Kumanyika, Td Agurs, A Lowe, D Saunders, and Cb Morssink. 1996. [Cultural aspects of African American eating patterns](#). *PubMed*.
- Emily Allaway, Nina Taneja, Sarah-Jane Leslie, and Maarten Sap. 2023a. [Towards countering essentialism through social bias reasoning](#). *Preprint*, arXiv:2303.16173.

- Emily Allaway, Nina Taneja, Sarah-Jane Leslie, and Maarten Sap. 2023b. [Towards countering essentialism through social bias reasoning](#). *arXiv preprint arXiv:2303.16173*. Preprint.
- Fahad M. Alliheibi, Abdulfattah Omar, and Nasser Al-Horais. 2021. [Opinion Mining of Saudi Responses to COVID-19 Vaccines on Twitter](#). *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(6). Number: 6 Publisher: The Science and Information (SAI) Organization Limited.
- Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *The AI Magazine*, 36(1):15–24.
- Puput Puji Astuti. 2018. [THE USE OF AFRICAN-AMERICAN VERNACULAR ENGLISH \(AAVE\) IN LOGIC’S EVERYBODY](#).
- Jinan C. Banna, Betsy Gilliland, Margaret Keefe, and Dongping Zheng. 2016. [Cross-cultural comparison of perspectives on healthy eating among Chinese and American undergraduate students](#). *BMC Public Health*, 16(1):1015.
- Jaclyn Benitez, R. A. Leshin, and Marjorie Rhodes. 2022. [The influence of linguistic form and causal explanations on the development of social essentialism](#). *Cognition*, 229:105246.
- Sarah Bunin Benor and Steven M Cohen. [Talking Jewish: The “Ethnic English” of American Jews](#).
- S.B. Benor. 2012. *Becoming Frum: How Newcomers Learn the Language and Culture of Orthodox Judaism*. Jewish Cultures of the World. Rutgers University Press.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in nlp](#). *Preprint*, arXiv:2005.14050.
- Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2024. [Investigating Human Values in Online Communities](#). *arXiv preprint*. ArXiv:2402.14177 [cs].
- Rod L. Boyd, Aneesh Ashokkumar, Sara Seraj, and James W. Pennebaker. 2022. *The Development and Psychometric Properties of LIWC-22*. University of Texas at Austin, Austin, TX.
- Pavel Brunssen. 2023. *The Making of “Jew Clubs”: Performing Jewishness and Antisemitism in European Soccer and Fan Cultures*. Thesis. Accepted: 2023-05-25T14:56:53Z.
- Mary Bucholtz and Kira Hall. 2005. [Identity and interaction: a sociocultural linguistic approach](#). *Discourse Studies*, 7(4-5):585–614. Publisher: SAGE Publications.
- Heidi A. Campbell, Rachel Wagner, Shanny Luft, Rabia Gregory, Gregory Price Grieve, and Xenia Zeiler. 2016. [Gaming Religionworlds: Why Religious Studies Should Pay Attention to Religion in Gaming](#). *Journal of the American Academy of Religion*, 84(3):641–664. Publisher: [Oxford University Press, American Academy of Religion].
- Eylon Caplan and Dan Goldwasser. 2025. [Concept-Carve: Dynamic realization of evidence](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20792–20809, Vienna, Austria. Association for Computational Linguistics.
- Jennifer Caplan. 2021. [Public Heroes, Secret Jews: Jewish Identity and Comic Books](#). *Journal of Jewish Identities*, 14(1):53–70. Publisher: Johns Hopkins University Press.
- Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. [Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms](#). *Preprint*, arXiv:2402.17649.
- J. K. Chambers. 2002. [Studying language variation: An informal epistemology](#). In J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, editors, *The Handbook of Language Variation*. Blackwell, Oxford.
- Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. [Convokit: A toolkit for the analysis of conversations](#). In *Proceedings of SIG-DIAL*.
- Paula Chesley. 2011. [You Know What It Is: Learning Words through Listening to Hip-Hop](#). *PLoS ONE*, 6(12):e28248.
- Kelly Cheung. 2015. *Learning past the pictures in the panels: teacher attitudes to manga and anime texts*. thesis, Macquarie University.
- Cuevas Ag, O’Brien K, and Saha S. 2016. [African American experiences in healthcare: “I always feel like I’m getting skipped over”](#). *PubMed*.
- Monika Czerwonka and Maria Pietrzak. 2024. [Application of Catholic Social Teaching in Finance and Management](#). *The Person and the Challenges. The Journal of Theology, Education, Canon Law and Social Studies Inspired by Pope John Paul II*, 14:295–313.
- A. Roy Eckardt. 1972. [Death in the Judaic and Christian Traditions](#). *Social Research*, 39(3):489–514. Publisher: The New School.
- Penelope Eckert. 2012. [Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation](#). *Annual Review of Anthropology*, 41:87–100.

- Kevin Fellezs. 2012. [Black Metal Soul Music: Stone Vengeance and the Aesthetics of Race in Heavy Metal](#). pages 23–36.
- Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. [The perspectivist paradigm shift: Assumptions and challenges of capturing human labels](#). 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 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. [Perspectivist approaches to natural language processing: a survey: Perspectivist approaches to natural language processing...](#) *Lang. Resour. Eval.*, 59(2):1719–1746.
- Susan A. Gelman. 2004. [Psychological essentialism in children](#). *Trends in Cognitive Sciences*, 8(9):404–409.
- Yilin Geng, Zetian Wu, Roshan Santhosh, Tejas Srivastava, Lyle Ungar, and João Sedoc. 2022. [Inducing Generalizable and Interpretable Lexica](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4430–4448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- John J. Gumperz. 1971. *Language in Social Groups*. Stanford University Press, Stanford, CA.
- Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Young-Min Cho, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. [Building Knowledge-Guided Lexica to Model Cultural Variation](#). *arXiv preprint*. ArXiv:2406.11622 [cs] version: 2.
- Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2025. [Comparing Styles across Languages: A Cross-Cultural Exploration of Politeness](#). *arXiv preprint*. ArXiv:2310.07135 [cs].
- Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. [Does BERT Learn as Humans Perceive? Understanding Linguistic Styles through Lexica](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Juan Manuel Hernández-Campoy. 2014. [Research methods in Sociolinguistics](#). Publisher: John Benjamins.
- Xiaoxiao Hu, Seth Kaplan, and Reeshad S. Dalal. 2010. [An examination of blue- versus white-collar workers’ conceptualizations of job satisfaction facets](#). *Journal of Vocational Behavior*, 76(2):317–325.
- Yoichi Ishibashi, Sho Yokoi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. [Subspace Representations for Soft Set Operations and Sentence Similarities](#). 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 3512–3524, Mexico City, Mexico. Association for Computational Linguistics.
- Alexandra Jaffe, editor. 2009. *Sociolinguistic Perspectives on Stance*. Oxford University Press, Oxford.
- Mahima Jain. 2016. [Reinterpretation of Hindu Myths in Contemporary Indian English Literature](#). *International Journal of Engineering, Science and Humanities*, 6(2):08–14.
- Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. [Natural language processing for dialects of a language: A survey](#). *Preprint*, arXiv:2401.05632.
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. [Challenges of studying and processing dialects in social media](#). In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.
- Omar Khattab and Matei Zaharia. 2020. [CoBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT](#). *arXiv preprint*. ArXiv:2004.12832 [cs].
- Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. [Interpretation of NLP models through input marginalization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.
- Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020. [Studying Politeness across Cultures Using English Twitter and Mandarin Weibo](#). *arXiv preprint*. ArXiv:2008.02449 [cs].
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. [Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations](#). In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Scott Lundberg and Su-In Lee. 2017. [A Unified Approach to Interpreting Model Predictions](#). *arXiv preprint*. ArXiv:1705.07874 [cs].

- Margaret MacLagan. 2005. [Regional and Social Variation](#). In *Clinical Sociolinguistics*, pages 15–25. John Wiley & Sons, Ltd. Section: 2.
- Benjamin S. Manning, Kehang Zhu, and John J. Horton. 2024. [Automated Social Science: Language Models as Scientist and Subjects](#). *arXiv preprint*. ArXiv:2404.11794 [econ].
- Gerald McDermott. 2015. [A Thumbnail Sketch of Judaism for Christians](#). *C.S. Lewis Institute*.
- John McWhorter. 2009. [Word On The Street: Debunking The Myth Of A Pure Standard English](#). Basic Books.
- John McWhorter. 2013. [Talking like that](#). *Jewish Review of Books*.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. [Stereoset: Measuring stereotypical bias in pretrained language models](#). *Preprint*, arXiv:2004.09456.
- 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.
- Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. [Interactive concept learning for uncovering latent themes in large text collections](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080, Toronto, Canada. Association for Computational Linguistics.
- Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. 2021. [Using large-scale experiments and machine learning to discover theories of human decision-making](#). *Science*, 372(6547):1209–1214. Publisher: American Association for the Advancement of Science.
- Daniel Preotiuc-Pietro and Lyle Ungar. 2018. [User-level race and ethnicity predictors from Twitter text](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. [Deconfounded Lexicon Induction for Interpretable Social Science](#). 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 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.
- Rajkumar Pujari, Chengfei Wu, and Dan Goldwasser. 2024. ["we demand justice!": Towards social context grounding of political texts](#). *Preprint*, arXiv:2311.09106.
- Myrthe Reuver, Alessandra Polimeno, Antske Fokkens, and Ana Isabel Lopes. 2024. [Topic-specific social science theory in stance detection: a proposal and interdisciplinary pilot study on sustainability initiatives](#). In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 101–111, Vienna, Austria. Association for Computational Linguistics.
- Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. 2012. [Cultural transmission of social essentialism](#). *Proceedings of the National Academy of Sciences*, 109(34):13526–13531.
- Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["Why Should I Trust You?": Explaining the Predictions of Any Classifier](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, San Diego, California. Association for Computational Linguistics.
- Shamik Roy and Dan Goldwasser. 2023. ["A Tale of Two Movements": Identifying and Comparing Perspectives in #BlackLivesMatter and #BlueLivesMatter Movements-related Tweets using Weakly Supervised Graph-based Structured Prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10437–10467, Singapore. Association for Computational Linguistics.
- Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, and Chris Kennedy. 2022. [Targeted identity group prediction in hate speech corpora](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 231–244, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) *Preprint*, arXiv:2303.17548.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). *Preprint*, arXiv:1911.03891.
- Amy E. Schwartz. 2011. [Is There Life After Death? Jewish Thinking on the Afterlife](#). *Moment Magazine*.
- Philippa Shoemark, James Kirby, and Sharon Goldwater. 2018. [Inducing a lexicon of sociolinguistic variables from code-mixed text](#). In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 1–6, Brussels, Belgium. Association for Computational Linguistics.
- Geneva Smitherman. 2007. [African American English](#). GRIN Verlag.
- Leonard Stein. 2019. [Jewish Flow: Performing Identity in Hip-Hop Music](#). *Studies in American Jewish Literature (1981-)*, 38(2):119–139. Publisher: Penn State University Press.

Ian Stewart. 2014. [Now We Stronger than Ever: African-American English Syntax in Twitter](#). In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg, Sweden. Association for Computational Linguistics.

Ichwan Suyudi, Agung Prasetyo Wibowo, and Luthfi Chanafiah Pasha. 2023. [Grammatical Analysis of African American Vernacular English in The Eminem Show Album: A Linguistics Perspective](#). *Langkawi: Journal of The Association for Arabic and English*, page 56. Publisher: Institut Agama Islam Negeri Kendari.

Intan Tia and Intan Aryani. 2020a. [African American Vernacular English \(AAVE\) Used by Rich Brian: A Sociolinguistic Investigation](#). *Language Circle Journal of Language and Literature*, 15:67–72.

Intan Tia and Intan Aryani. 2020b. [A Sociolinguistic Investigation](#). *Language Circle Journal of Language and Literature*, 15:67–72.

Anna TGUNOVA, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. [RedDust: a Large Reusable Dataset of Reddit User Traits](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6118–6126, Marseille, France. European Language Resources Association.

S. Tufféry. 2011. *Data Mining and Statistics for Decision Making*. Wiley Series in Computational Statistics. Wiley.

R. Wardhaugh and J.M. Fuller. 2021. *An Introduction to Sociolinguistics*. Blackwell Textbooks in Linguistics. Wiley.

Zach Wood-Doughty, Paiheng Xu, Xiao Liu, and Mark Dredze. 2021. [Using noisy self-reports to predict Twitter user demographics](#). In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 123–137, Online. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. [Large Language Models for Automated Open-domain Scientific Hypotheses Discovery](#). *arXiv preprint. ArXiv:2309.02726* [cs].

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. [VALUE: Understanding Dialect Disparity in NLU](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

A Dataset Details

A.1 Demographic Processing

Here we provide more details about the **SPLITS!** dataset construction. Using both similarities (Jac-

Demographic	Seed subreddits
African-American	AfricanAmerican, BlackAtheism, BlackHair, BlackLadiesFitness, BlackWomens, Blackfellas, Blackpeople , Blerds, Dreadlocks, EbonyImagination, JustProBlackThings, Natural_Hair, blackculture, blackgirlgamers, blackgirls, blackinamerica, blackladies, blackpower, blackyoutubers
Catholics	AnglicanOrdinate, Catholic , CatholicDating, CatholicGamers, CatholicMemes, CatholicParenting, CatholicPhilosophy, CatholicPolitics, CatholicVideos, CatholicWomen, Catholic_News, Catholicism, DeusVult, EasternCatholic, FAMnNFP, MarriedCatholics, OrthodoxChristianity, RCIA, RealCatholicMen, Roman_Catholics, TraditionalCatholics, TrueCatholicPolitics, VideoSancto, catechism, divineoffice, eRetreats, homilies, knightsofcolumbus, modelCatholicChurch
Teachers	teachingresources, ELATeachers, Teachers , historyteachers, TeacherTales, SubstituteTeach, teaching, Teacher, ScienceTeachers, ArtEd, ECEProfessionals, SpanishTeachers
Construction Workers	Builders, Concrete, Contractor, Insulation, drywall, Welding, Ironworker, Construction , BlueCollarWomen, Roofing, WeldPorn, weldingjobs, ConstructionPorn, BadWelding
Hindus	AdvaitaVedanta, hindu ,
Jains	krishna, IndiaRWResources,
Sikhs	hinduism, KrishnaConsciousness, hindurashtravad, Sikh, Jainism, truehinduism, diwali, bhajan

Table 4: Demographics and corresponding seed subreddits (first seed in bold)

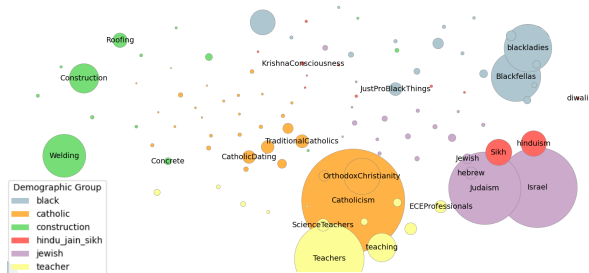
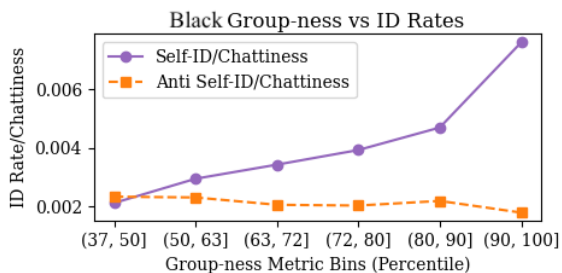


Figure 7: Visualization of the seed subreddit discovery process. Each bubble is a seed subreddit, sized by post volume and positioned by user overlap with other seeds, and clustering validates our iterative expansion method. Crucially, this plot shows raw user overlap between subreddits, not final demographic user groups, which are filtered to be nearly disjoint.



card/cosine) accounted for the skewed cases when one subreddit was very large while the other was very small. This is because the Jaccard index normalizes the intersection by the number of unique users in both subreddits, while the Cosine similarity normalizes by the geometric mean of the number of users. Table 4 details the seed subreddits that were annotated for each demographic, and these subreddits are pictured in Figure 7. Table 5 details the self-identification and anti-self-identification phrases. We note that each C_i was filtered to remove posts by known bots and then the top 1% of users by number of posts were also filtered out to remove unknown bots and spam users. We also show the self-ID vs. anti-self-ID plots for the remaining demographics in Figure 8, and the exact thresholds used were: Jewish \rightarrow 90, Black \rightarrow 75, Catholic \rightarrow 75, Construction \rightarrow 90, Teacher \rightarrow 75, Hindu/Jain/Sikh \rightarrow 80, chosen based on separation in the plots. The full heatmap of demographic-demographic Jaccard similarities is in Figure 10.

Two very common demographic dimensions, gender and age, were considered, but ultimately not used in **SPLITS!** Because we aimed to limit intersectional identities, we opted not to use non-minority groups like ‘men’ and ‘women’ because they would result with huge amounts of overlap with other demographics. In addition,

Demographic	Self-ID Phrases	Anti-Self-ID Phrases
African-American	I am black I'm black I am African American ... As a black man As a black woman	I am not black I'm not black I'm white ... I am Caucasian I am Asian
Catholics	I am Catholic I'm Catholic ... As a Roman Catholic I belong to the Catholic Church I am part of the Catholic Church	I am not Catholic I'm not Catholic ... I am a Baptist I'm Methodist ... I am agnostic As an ex-Christian As an ex-Catholic
Teachers	I am a teacher I'm a teacher I teach I am an educator ... I'm in education I am a school teacher I love teaching	I am not a teacher ... I am a student ... I'm an accountant I work in accounting I am a scientist ... I work in pharmacy
Construction Workers	I am in construction I'm in construction ... I work in the construction industry I'm in the construction industry ... I work with drywall I work in insulation ... I am an ironworker I'm an ironworker	I am not in construction I don't work construction ... I have a desk job I work in tech I work in IT I am a programmer ... I am an artist I'm an artist I work in retail I'm in retail
Hindus, Jains, and Sikhs	I am Hindu I'm Hindu As a Hindu I identify as Hindu ... I practice Sanatana Dharm I follow Sanatana Dharm I am a follower of Sanatana Dharm ... I'm a follower of Sanatana Dharm As a follower of Sanatana Dharm I believe in Sanatana Dharm I am Jain I'm Jain I am a Jain I'm a Jain ... I am Sikh I'm Sikh ... I practice Sikhism As a follower of Sikhism	I am not Hindu I'm not Hindu I am not a Hindu I'm not a Hindu ... I am Christian I'm Christian I am a Christian ... I am agnostic I have no religion As an ex-Hindu As an ex-Jain As an ex-Sikh

Table 5: Self-identification and anti self-identification phrases for various demographics

‘women’ subreddits were often labeled specifically for women (e.g. ‘r/TwoXChromosomes’, ‘r/womenintech’), whereas they were very rarely for men, meaning selecting users in this way would almost certainly skew the distributions away from the real world.

For age, the single main issue was the demographic of people who use Reddit in general. Specifically, very few older-age subreddits exist, and the ones that exist have very few posts. Because of this, ‘age’ as a demographic dimension was not used.

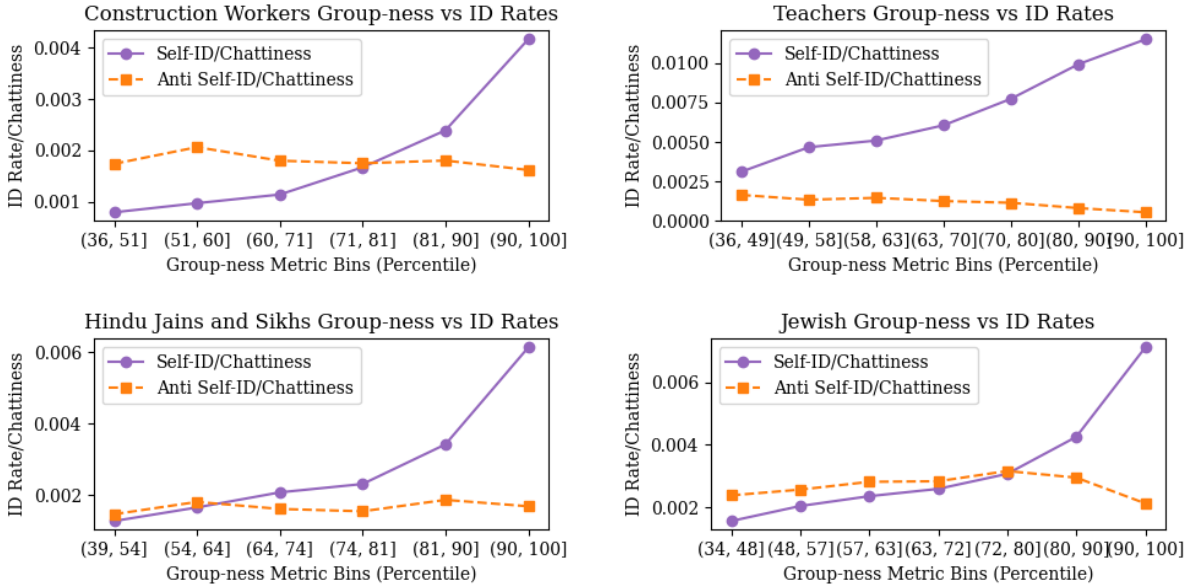


Figure 8: Normalized self-identification rate vs. group-ness of the remaining demographics.

A.2 When Demographic Annotation Fails: Koreans

Of the 7 demographic groups that we attempted to annotate, all were successful in having group-ness separate between self-ID and anti self-ID, *except* for the group *Koreans*. We annotated this set just as the others, starting with seed subreddit *r/korea*, and ending with seed set {KLeague, KoreanFood, South_Korea, busan, gyopo, hanguk, korea, seoul, southkorea}. However, after collecting all posts across all Reddit, and computing the self-ID and anti self-ID rates, we saw a negative result (Figure 9). Here, we see that the anti self-ID rate does *not* decrease as we turn up the group-ness, indicating that the group-ness metric here did not properly capture the target demographic. Inspecting the seed set, we hypothesize that this is because many non-Korean users were active in these groups such as Korean language learners, fans of Korean teams, or people interested in Korean culture like KoreanFood.

This demonstrates that our methodology using group-ness works only when the seed sets are ‘clean’ enough so as to properly define a powerful group-ness metric for that demographic. Meaning, demographics unsuitable for our methodology are ones which do not have enough independent, clean sub-communities, making them difficult to isolate.

A.3 Topic Processing

Table 6 lists some categories and topics, while Table 8 lists some select examples of topics within

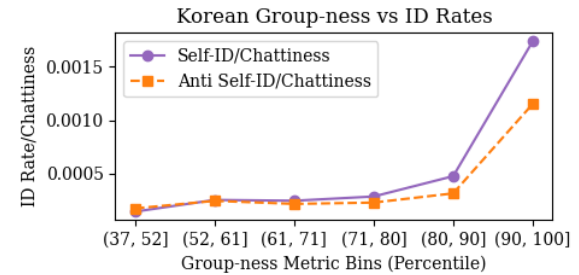


Figure 9: Failure case: normalized self-identification rate and anti self-identification rate do *not* separate as group-ness increases.

their categories, and shows the respective keywords used for the ColBERT topic relevance search.

To cheaply ensure any given split’s posts were truly about the topic in question, we applied the following process for each topic: first, we retrieved the top 100 thousand posts within each demographic’s C_i using ColBERT. We then combined all demographics’ posts on the topic and ordered them by the ColBERT relevance score (600 thousand posts). Then we applied the LLM sliding window-binary search algorithm. The idea is as follows: in general, the ColBERT relevance score tells us relative relevance of documents to the topic. For example, in general it can tell us that post x is more relevant to topic t than post y . However, it does not at all tell us at what point posts are no longer relevant at all. Consider a contrived topic query like “discussion of Christmas trees on the moon fighting over a purple golf club”. While we can assume there are either 0, or very few posts actually about this

Category	Topics
Sports	basketball, soccer, football...
Entertainment	superheroes, sci-fi, fantasy...
Tech/Gaming	pc builds, coding, AI...
Careers	jobs, resumes, freelance...
Hobbies	gardening, cooking, crafts...
Finance	budgets, stocks, retiring...
Education	college, study tips, exams...
News	global, politics, environment...
Travel	budget, luxury, backpacking...
Humor	memes, satire, animals...
Political/Social	abortion, Russia, taxes...

Table 6: Categories and (some) topics.

topic, the ColBERT retrieval will still give us 100 thousand posts for each demographic. So the point is to find the cutoff at which posts are (on average) no longer about the topic in question.

The binary search algorithm works as follows over a set of documents D and topic t , given a window size w , a cutoff proportion p , and an error threshold e : the top w posts are fed into an LLM one at a time, prompting it to say whether the post is within topic t (yes/no). Among all posts in the window, the proportion of ‘yes’s is computed. If the proportion is lower than p , then the window moves halfway upward between the current upper and lower bounds. Likewise, if the proportion is higher than p , then the window is moved halfway downward between the current lower and upper bound. This process continues until either a window’s proportion is within e of the desired cutoff p , or the bounds are exhausted. Once the process terminates, all posts ranked higher than the current window are considered above the cutoff, and included as ‘topical’.

For our processing, we used the gpt-4.1-nano-2025-04-14 model, $w = 100$, $p = 0.8$, and $e = 0.03$, meaning the bottom 100 posts ranked by relevance of any split must have *at least 77%* topical posts as judged by the LLM. The underlying assumption for this algorithm is that *on average*, higher ranked documents are more likely to be on topic than lower ranked documents. If it were exactly true, then we would not need a window. Due to the noisiness of the real-world data, we use a method more robust to this noise, while still ensuring that our LLM calls do not scale linearly with the number of posts (10’s of millions).

Figure 11 depicts the sizes of each $C_{d,t}$ split, both by demographic, topic category, and overall. The “Political” category is a catch all for many political and social topics of discussion, and includes more topics than the other categories (which is why

Demographic	# Posts (M)	# Users (K)
Teacher	21.6	19.1
Catholic	21.5	13.5
Black	16.6	9.0
Construction Worker	10.7	4.0
Jewish	15.2	5.3
Hindu/Jain/Sikh	3.8	2.6

Table 7: Pre-topic filtration data stats.

it also has more posts). After removing splits with fewer than 2,000 posts, we also removed ‘orphan’ splits that had only one demographic per topic. Figure 12 shows the number of each pairing once the data is combined and indexed together. All pairs are very close in number. Figure 16 shows the exact distributions of triviality among all 5 kinds of prompts. This clearly shows how the more creative prompt leads to more non-trivial lexica, and that including information about the topic in the prompt helps reduce triviality.

Category	Specific Topic	Keywords
Sports & Fitness	Basketball	basketball, hoop, net, dunk, dribble, NBA...
Sports & Fitness	Soccer	soccer, football, goal...
Entertainment & Media	Superheroes/Comic Book Media	superheroes, comic books, Marvel...
Entertainment & Media	Fantasy TV/Movies	fantasy, magic, sword and sorcery, medieval...
Hobbies & Special Interests	Gardening	gardening, garden tips, plant care...
Hobbies & Special Interests	Cooking/Baking	cooking, baking, recipes, food blog...
Education & Academia	College Applications & Admissions	college applications, university admissions, application process, admission requirements, college essay...
Education & Academia	Study Techniques & Productivity	study techniques, productivity tips, time management, study schedule, note taking, active recall...
Education & Academia	Exam Preparation & Test-Taking Strategies	exam preparation, test strategies, study tips, exam study guide, test taking techniques...

Table 8: Select example neutral categories, sub-topics and keywords

A.4 Human Validation of Self-Identification Classification with LLM

To compute self-identification and anti self-identification, we used first a set of regex expressions to retrieve posts, followed by an LLM filtration step to remove invalid matches (e.g. quotations, part of a bigger phrase, sarcasm, etc.). To assess the quality of this LLM filtration, we had an annotator manually examine 112 posts: 56 self-identification posts, and 56 anti self-identification posts, with each set having been labeled as “valid

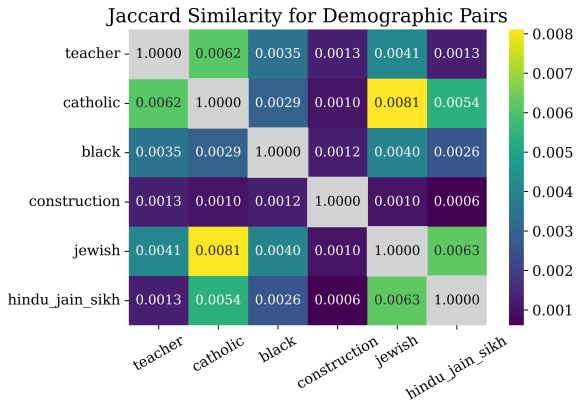


Figure 10: User intersectionality in the **SPLITS!** dataset.

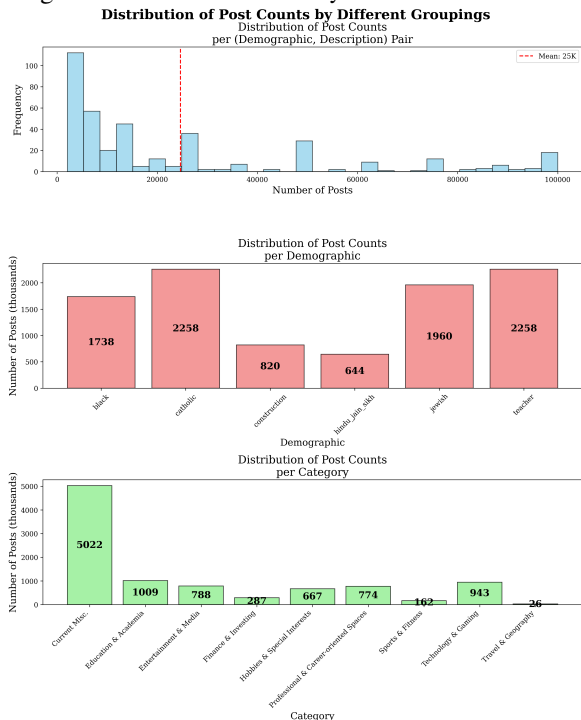


Figure 11: Distributions over different groupings of **SPLITS!**.

matches” by the LLM and the other half “invalid”. The annotator also assessed the validity, and we report the results of using the *regex only* vs. *regex + LLM* in Table 13.

Method	Precision	Recall	F1-Score
Regex	0.536	1.000	0.698
Regex + LLM	0.964	0.898	0.930

Table 9: Overall performance comparison of human annotation for self-ID/anti self-ID match validation. Adding the LLM greatly improves the precision, and therefore the F1 for less noisy results.

A.5 Human Validation of Topic Classification with LLM

To ensure the quality of the topic annotation process, we randomly sampled 100 posts from diverse

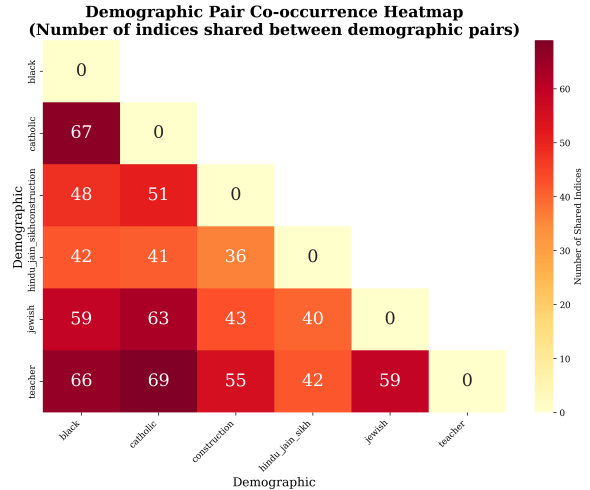


Figure 12: Heatmap of combined indices by demographic.

topics and demographics such that 50 were excluded as ‘not topical’ and 50 were included as ‘topical’. We had a human annotator classify each post, and we compute the agreement with the LLM as a classifier. The result is [Precision: 0.96, Recall: 0.96, F1: 0.96]. Therefore we can say that the vast majority of topical annotations by the LLM are likely correct.

B Case Studies of Known SLPs

AAVE African American Vernacular English (AAVE) is a well-known and studied SLP (Stewart, 2014; Shoemark et al., 2018; Jørgensen et al., 2015; Ziems et al., 2022), which often involves *code-switching*—using the vernacular more in some settings than others (Mcwhorter, 2009). For instance, AAVE is prevalent within the domain of Hip-Hop and Rap music (Tia and Aryani, 2020b; Suyudi et al., 2023; Astuti, 2018; Chesley, 2011), while being used much less in professional settings. We test if our dataset captures these two patterns using a lexicon with features from Ziems et al. (2022) and Smitherman (2007) (e.g., immediate future ‘finna’, ‘ass’ camouflage, copula deletion). We find that Black users use AAVE significantly more than non-Black users across both ‘Hip-Hop’ and ‘Professional’ topics. Furthermore, Black users themselves use AAVE features significantly more when discussing Hip-Hop than in professional contexts. Together, these results ($p < 10^{-5}$) confirm that **SPLITS!** is rich enough to capture both between-group linguistic phenomena and within-group code-switching.

Jewish English Benor and Cohen studied the vocabulary of American Jews, noting a difference

in the usage of certain Yiddish and Hebrew words. Further, Benor (2012); McWhorter (2013) study how Jews tend to use such in-group language less often when discussing secular topics as compared to more religious ones. Just as with Black AAVE use, we use the lexicon from Benor and Cohen to distinguish Jews from non-Jews, both on the topic of ‘Judaism’, and on ‘Professional’ topics. This again served the purpose of (1) verifying that the richness of the dataset captures the SLP of Jewish use of Yiddish words and (2) that this usage is code-switched depending on context.

We used the features from Benor and Cohen, including only lexicon entries that were used more by Jews than non-Jews. This included Yiddish/borrowed Hebrew terms. Fixing topic t first as ‘Judaism’, and then as ‘Professional’ topics, for each demographic d we compute the proportion of posts in $C_{d,t}$ with at least one Yiddish lexicon entry. As seen in Table 10, we see that Jewish people in the dataset use Yiddish significantly more than non-Jewish people, both when discussing Judaism and also Professional topics, aligning with the SLP of Jewish people’s Yiddish use. This answers question (1): the dataset *is* sufficiently rich that Yiddish lexical features significantly distinguish between Jewish and non-Jewish users.

To answer question (2), we compared the proportions of only Jewish people’s Yiddish use on the topic of Judaism vs. Professional topics. In that comparison, we also see a significantly higher use of Yiddish in Judaism than in Professional topics. This means that not only do Jewish people use Yiddish features more than non-Jewish users, but they use these features far more in certain contexts. These two results together show that the dataset captures the known SLP of (1) Jewish Yiddish use and (2) Jewish code-switching.

Group	Topic	Total posts	Prop. Yid/Heb
J	Jud.	96,880	1.90×10^{-3}
¬J	Jud.	64,400	7.30×10^{-4}
J	Prof.	135,034	1.41×10^{-4}
¬J	Prof.	639,210	5.95×10^{-5}

Hypothesis	p-value
$p(J, Jud.) > p(\neg J, Jud.)$	$< 10^{-5}$
$p(J, Jud.) > p(J, Prof.)$	$< 10^{-5}$
$p(J, Prof.) > p(\neg J, Prof.)$	0.00079
$p(\neg J, Jud.) > p(\neg J, Prof.)$	$< 10^{-5}$

Table 10: Yiddish/Hebrew usage stats and hypothesis tests

Jewish: (avg. 1.124@0.5%), but are about average in triviality (0.746 Triviality)

Group	Topic	Total posts	Prop. dance
HJS	PCI	87,514	0.004365
¬HJS	PCI	381,181	0.003618

Hypothesis	p-value
$p(HJS, PCI) > p(\neg HJS, PCI)$	0.00054

Table 11: “dance” usage stats and hypothesis test

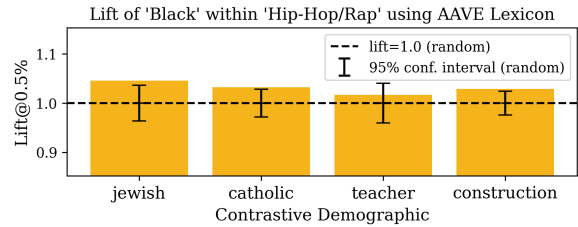


Figure 13: Lift at 0.5 of the Black demographic when talking about Hip-Hop/Rap using AAVE lexicon, as contrasted with 4 other demographics.

C PSLP Automatic Evaluation

Table 12 shows the lexica ℓ that were used to compute the Triviality for each demographic.

We used an LLM to generate over 23,000 candidate PSLPs. We used five prompt variations that gave the model increasing levels of context: from just the target demographic (A), to the demographic pair (A, B), to the full context including the topic (t). A final “creative” prompt also used the full context but was instructed to generate more novel lexica (all prompts are in App. E).

We see the difference in triviality induced by the prompt: excluding the topic (as in the Demo and Demo/Demo prompts) gives more trivial PSLPs, while the ‘Creative’ prompt gives the most non-trivial PSLPs. The distribution of each prompt’s triviality (without lift) can be found in Figure 16, and lift can be found in Table 9.

Demographic	Lexicon (ℓ)
Teacher	teacher, educator, education, teaching, teach, schoolteacher
Catholic	Catholic, Catholicism, Catholic Church, Mass, Eucharist, Catechism, Catholic priest
Black	Black, African American, Black history, Afro-American, Black people, Black person
Construction Worker	construction worker, construction, builder, construction site, contractor, building, laborer
Jewish	Jewish, Jew, Judaism, Jewish holidays, Torah, synagogue, Kosher, Shabbat, Rabbi
Hindu/Jain/Sikh	Hindu, Hinduism, Jain, Jainism, Sikh, Sikhism, puja, Shiva, Vishnu, ahimsa, karma, Gurdwara

Table 12: Demographic lexica for triviality calculation.

D Human Validation of the Triviality Metric

This section provides a detailed account of the human validation study conducted to assess the effectiveness of our Triviality metric as a proxy for human judgments of “unexpectedness.” The study

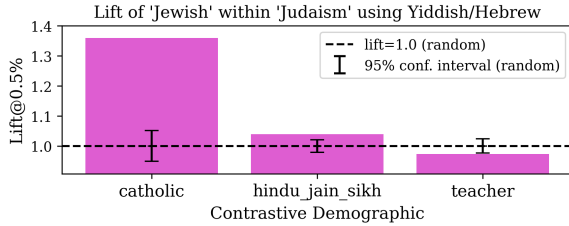


Figure 14: Lift at 0.5 of the Jewish demographic when talking about Judaism using Yiddish/Hebrew, as contrasted with 3 other demographics.

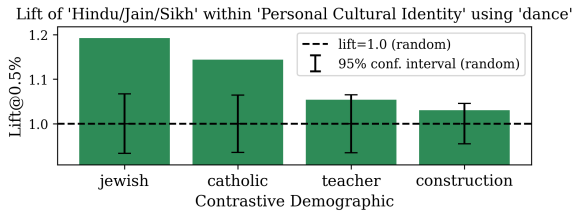


Figure 15: Lift at 0.5 of the Hindu Jain Sikh demographic when talking about Personal Cultural Identity using "dance", as contrasted with 4 other demographics.

Prompt Type	mean	min	q25	q75	max
D1 + D2	1.28	0.43	1.05	1.37	6.84
D1 + D2 + Topic	1.32	0.25	1.03	1.40	7.78
D1 + Topic	1.20	0.11	1.00	1.29	5.09
D1	1.44	0.45	1.11	1.56	7.67
D1 + D2 + Topic (Theory)	1.07	0.05	0.96	1.13	6.71

Table 13: Distribution of lift@0.5 by prompt type.

involved 9 annotators who collectively spent approximately 7 hours on the task. The annotators were from diverse demographic backgrounds, including members of 3 of the 6 groups studied in our dataset, and all held at least a Bachelor’s degree with fluency in English. After filtering for significant lift PSLPs ($p < 0.025$), we randomly sampled 500 PSLPs for this study, with each PSLP being rated by 3 of the 9 annotators to ensure reliable judgments.

The annotation task was designed to elicit an intuitive judgment of surprise. For each instance, annotators were presented with a Target Demographic (e.g., “Jewish”), a Contrast Demographic (e.g., “Catholic”), a Topic (e.g., “Elections”), and a Lexicon (e.g., “ballot access”, “voter registration”, “gerrymandering”). They were then instructed to rate the lexicon based on the following instructions:

1. *First think about what you know about the two demographics A and B, especially when they talk about the given topic. What kinds of words/phrases might Demographic A use that Demographic B would not? Specifically, we care about such words/phrases that are not obvious, or unexpected.*

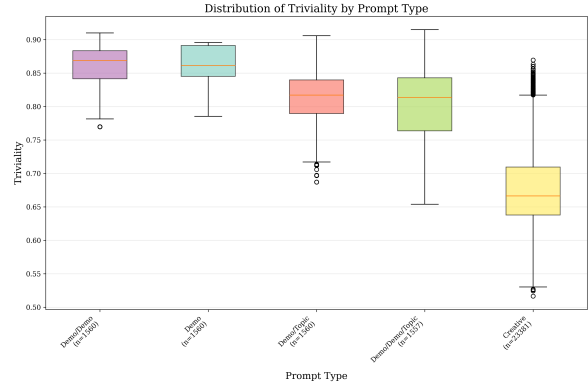


Figure 16: Distributions of Triviality by prompt type.

Example: When talking about "recipes", Indian people when contrasted with American people may use words (ingredients) that are common in Indian recipes (cardamom or saffron, turmeric, etc.)

Note: When Indian and Bangladeshi people are contrasted, this might not work because such words are shared across the two groups.

2. *Once you are ready, think about how these keywords compare to what you came up with. Were you surprised that the keywords worked in distinguishing the two groups?*

To measure the consistency and quality of the annotations, we calculated the Intraclass Correlation Coefficient. Using the ‘ICC(2,k)’ two-way random effects model, which assesses the reliability of the average ratings, we achieved a score of **0.738**. This indicates “good” reliability and confirms that annotators shared a consistent understanding of the task.

Our analysis yielded two key results. First, we computed the Spearman rank correlation (ρ) between our automated Triviality score and the average human unexpectedness score for all 500 PSLPs, finding a significant negative correlation of $\rho = -0.38$ ($p < 0.001$). This supports our hypothesis that as a PSLP becomes more trivial according to our metric, it is perceived as less unexpected by humans. Second, we evaluated the metric’s utility as a filter. We defined a PSLP as “promising” if all 3 annotators assigned it an unexpectedness score of 3 or higher, which applied to 135 of the 500 instances (27%). The **1.5-1.8x improvement factor** reported in the main paper represents the reduction in manual labor; it is the ratio of PSLPs one must inspect to find a “Promising” one without the filter versus with the filter. For example, if a researcher must review 100 statistically significant PSLPs to find 10 “Promising” ones (a 10:1 effort), and our Triviality filter narrows the pool to 60 candidates

that still contain 9 of the “Promising” ones (a 6.7:1 effort), the reduction in effort is $10/6.7 \approx 1.5x$. The 1.5-1.8x range reflects this practical speedup at reasonable operating thresholds for the filter.

E Prompts

All LLM calls were done using gpt-4.1-nano-2025-04-14, except for PSLP generation, where we used gpt-4.1-2025-04-14.

```

### Task Overview:
As a social media analysis assistant, your task is to analyze a social media post and determine if the user has self-identified themselves. You will be given a post, and a target demographic (e.g. "Black", "Teacher", etc). Your task is to read the post and determine with high confidence whether the user has self-identified themselves as the demographic (e.g. "I am a black man"). In addition, you must determine whether the user has self-identified as a demographic that is mutually exclusive to the target demographic (e.g., for "black", this could be saying "I am a white woman" or "I am not black"; for "teacher", this could be saying "I work in construction" or "I am not a teacher").

### Response Format:
Your response must adhere to the format below:
User self-identifies as demographic: yes OR no
User self-identifies as mutually exclusive demographic: yes OR no

### Demographic
{demographic}

### Social Media Post
{post}

### Response

```

Figure 17: Prompt for self-id or anti-self-id.

F Literature-Inspired PSLPs

Table 14 contains the 11 demographic/topic pairs and academic papers which were used to generate the Literature-Inspired PSLPs.

F.1 Human Validation of Literature-Inspired PSLPs’ Alignment with Papers

When creating the Literature-inspired PSLPs, the LLM was conditioned on published social science papers. Despite this conditioning to produce the lexica, we wanted to test whether the generated lexica truly aligned with the results of the paper. To do this, we had a human annotator inspect 44 of the Literature-inspired lexica. For each one, we showed the topic, demographic, and the lexicon, mixed with 3 other lexica for the that topic/demographic that were *not* conditioned on the paper. The annotator was allowed to view the relevant paper, and was asked to pick which lexicon of the 4 truly came from the results of the paper. Of the 44, the annotator picked correctly 42 times (95.45%), showing that in some way, the Literature-inspired PSLPs were truly inspired by the findings of their respective papers.

```

### Task Overview:
You are a socio-linguistic scientist. You will answer questions posed by the user, taking into consideration every detail of their request. Format the output in the same way as the example provided.

### Task Details:
You are given two demographics; a target demographic and a contrast demographic. Your task is to generate words and phrases that would be used in posts by the target demographic but are unlikely to be used by the contrast demographic. The posts will be mostly in the English script.

### Strict Rules:
1. ONLY output words (and phrases) that can be directly searched. No theories like "they might use..."
2. Output as many words and phrases as you think are appropriate:
– If there are many differentiating things, output many words. Otherwise output fewer but high quality words.
3. First think about the demographics given out loud. Mention out loud what you know about them and then generate the words.
4. Don't output reasoning with the words, just the words and phrases.

### Example Input:
Target demographic: Chinese
Contrast demographic: Russian

### Example Output:
The target demographic is Chinese, and the contrast demographic is Russian. The goal is to find English-language words and phrases that appear in posts by people from the Chinese demographic but are unlikely to appear in posts by Russians. The words must be searchable, i.e., things they would actually write or say online, in English. People from mainland China or ethnic Chinese individuals posting in English may exhibit certain linguistic, cultural, or stylistic markers that set them apart from Russian posters.

Key distinguishing features to think about:
* Cultural references: Chinese users may refer to cultural institutions, apps, celebrities, or ideologies unique to China.
* Language influence: Certain phrasings that come from direct translation of Chinese idioms or grammar.
* Internet platforms: Use of uniquely Chinese platforms like WeChat, Weibo, or Bilibili.
* Education/study culture: Common English phrases around Gaokao, overseas study, etc.
* Political language: Specific phrasing from Chinese state media, such as "core socialist values" or "Western hostile forces", not likely to be used by Russians.

Words and Phrases:
1. WeChat
2. Douyin
3. Gaokao
4. 996 work culture
5. Spring Festival Gala
6. Bilibili
7. Little Red Book
8. Weibo
9. Tsinghua University
10. C-pop
...

### Task Input:
Target demographic: {target}
Contrast demographic: {contrast}

### Task Output:

```

Figure 18: Prompt when 2 demographics are given.

Demographic	Topic	Lexica (combined, sampled)	Paper
Black	Cooking/ Baking	pork chops, family dinner table, eat at home, soul food, adapting recipes, switching ingredients, breaking bread together, Sunday dinner	Cultural aspects of African American eating patterns (Airhihenbuwa et al., 1996)
	Healthcare	didn't listen, talked over me, second opinion, questioning everything, really listens, understands me	African American experiences in healthcare: "I always feel like I'm getting skipped over" (Cuevas Ag et al., 2016)
	Metal/ Rock Music	something else, don't fit in, always a statement, the source, the right look, don't fit the image, technical skill, prove myself, safe to like, empty praise	Black Metal Soul Music: Stone Vengeance and the Aesthetics of Race in Heavy Metal (Fellezs, 2012)
Jewish	Superheroes/ Comic Book Media	secret identity, living two lives, finally revealed, true self, feeling like a monster, outsider status, tragic backstory, shaped by his past	Public Heroes, Secret Jews: Jewish Identity and Comic Books (Caplan, 2021)
	Soccer	badge of honor, form of solidarity, feel comfortable, remembering our history, in contrast to them, what separates us,	The Making of "Jew Clubs": Performing Jewishness and Antisemitism in European Soccer and Fan Cultures (Brunssen, 2023)
	Hip-Hop/ Rap Music	prove myself, not like that, same struggle, our histories, out of place, I know it's weird, new path, found my way	Jewish Flow: Performing Identity in Hip-Hop Music (Stein, 2019)
Catholic	Video Games	moral compass, personal conviction, fighting darkness, a greater good	Gaming Religionworlds: Why Religious Studies Should Pay Attention to Religion in Gaming (Campbell et al., 2016)
	Banking & Financial Institutions	stewardship, social responsibility, moral screening, serving the team, people first	Application of Catholic Social Teaching in Finance and Management (Czerwonka and Pietrzak, 2024)
Hindu/Jain/Sikh	Book/ Literature Discussions	political dynasty, her side of the story, a woman's perspective, a fresh take, a reluctant hero, our cultural heritage, passing down the stories	Reinterpretation of Hindu Myths in Contemporary Indian English Literature (Jain, 2016)
Teacher	Anime/ Manga	gateway text, a good start, visual literacy, film techniques, student-led discussion, the real experts, the manga kids, their little group	Learning past the pictures in the panels: teacher attitudes to manga and anime texts (Cheung, 2015)
Construction	Public Sector/ Government Jobs	a good job, honest living, got my back, good crew, fair wage, pays the bills	An examination of blue- versus white-collar workers' conceptualizations of job satisfaction facets (Hu et al., 2010)

Table 14: Literature-inspired PSLPs across demographics and topics.

```

### Task Overview:
You are a socio-linguistic scientist. You will answer questions posed by the user, taking into consideration every detail of their request. Format the output in the same way as the example provided.

### Task Details:
You are given a target demographic. Your task is to generate words and phrases that would be used in posts by the target demographic. The words and phrases should be relatively unique to this demographic so that we can identify documents authored by this demographic. The posts will be mostly in the English script.

### Strict Rules:
1. ONLY output words (and phrases) that can be directly searched. No theories like "they might use..."
2. Output as many words and phrases as you think are appropriate:
- If there are many differentiating things, output many words. Otherwise output less but high quality words.
3. First think about the demographics given out loud. Mention out loud what you know about them and then generate the words.
4. Don't output reasoning with the words, just the words and phrases.

### Example Input:
Target demographic: Sri Lankan

### Example Output:
Sri Lankans are a South Asian demographic with diverse linguistic and cultural backgrounds, primarily Sinhala and Tamil speakers, but English is widely used in online posts, especially among urban youth and diaspora communities. Common topics include cricket, politics, local food, and religious/cultural festivals like Vesak, Sinhala and Tamil New Year, and Poson. There's also heavy usage of local slang, transliterated Sinhala/Tamil phrases, and references to places, politicians, and celebrities. English posts by Sri Lankans often include code-switching (Sinhala/Tamil-English hybrid sentences), slang like "machan" (buddy), and references to foods like kottu, hoppers, and short eats. The political discourse often includes names of local parties (SLPP, UNP, JVP), figures (Gotabaya, Mahinda, Ranil), and issues like "aragalaya" (protest movement). Diaspora posts may talk about missing Sri Lanka, family back home, and visiting "lanka" during holidays.

Words and Phrases:
1. machan
2. kottu
3. hoppers
4. short eats
5. poya day
6. Vesak
7. Avurudu
8. Poson
9. parippu
10. rice and curry
11. dhal curry
12. sambol
13. thambili
14. aragalaya
15. Gota
16. Mahinda
17. Ranil
...

### Task Input:
Target demographic: {target}

### Task Output:

```

Figure 19: Prompt when only target demographics is given.

```

### Task Overview:
You are a socio-linguistic scientist. You will answer questions posed by the user, taking into consideration every detail of their request. Format the output in the same way as the example provided.

### Task Details:
You are given a topic and a target demographic. Your task is to generate words and phrases that are likely to be used in posts by the target demographic when talking about the given topic. The words and phrases should be such that they help identify the demographic as the author of documents about the topic. The posts will be mostly in the English script.

### Strict Rules:
1. ONLY output words (and phrases) that can be directly searched. No theories like "they might use..."
2. Output as many words and phrases as you think are appropriate:
- If there are many differentiating things, output many words. Otherwise output less but high quality words.
3. First think about the demographics given out loud. Mention out loud what you know about them and then generate the words.
4. Don't output reasoning with the words, just the words and phrases.

### Example Input:
Target demographic: Sri Lankan
Topic: cricket

### Example Output:
Sri Lankans are deeply passionate about cricket—it's the most popular sport in the country and a major source of national pride. Sri Lankan cricket fans often reference legendary players, local teams, and emotional highs and lows of key matches, especially against rivals like India, Pakistan, and Australia. They use Sinhala or Tamil terms occasionally, but in English posts, they tend to blend cricket lingo with cultural pride. They also often refer to specific tournaments (e.g., Asia Cup, World Cup), iconic moments, and local slang or nicknames for players. There is a strong tendency to mention players from different eras (e.g., Sanath, Sanga, Mahela, Malinga, etc.) and to express strong emotions about wins, losses, and team selections.

#
Words and Phrases:
1. Lion pride
2. Lankan Lions
3. Mahela magic
4. Sanga class
5. Sanath sixes
6. Malinga yorkers
7. Dilshan scoop
8. Kusal form
9. Lankan fighting spirit
10. Island nation roar
11. SL cricket
12. Come on Sri Lanka
13. Lankan powerplay
14. Proud to be Sri Lankan
...

### Task Input:
Target demographic: {target}
Topic: {topic}

### Task Output:

```

Figure 20: Prompt when target demographic and topic is given.

Task Overview:
 You are a socio-linguistic scientist. You will answer questions posed by the user, taking into consideration every detail of their request. Format the output in the same way as the example provided.

Task Details:
 You are given a **topic** and two demographics; a **target demographic** and a **contrast demographic**. Your task is to generate words and phrases that are likely to be used in posts by the target demographic but unlikely to be used by the contrast demographic, for the given topic. The posts will be mostly in the English script.

Strict Rules:
 1. **ONLY** output words (and phrases) that can be directly searched. No theories like "they might use..."
 2. Output as many words and phrases as you think are appropriate:
 - If there are many differentiating things, output many words. Otherwise output less but **high quality** words.
 3. First think about the demographics given **out loud**. Mention out loud what you know about them and then generate the words.
 4. Don't output reasoning with the words, just the words and phrases.

Example Input:
 Target demographic: Chinese
 Contrast demographic: Russian
 Topic: cooking

Example Output:
Reasoning:
 Chinese cooking culture emphasizes diverse regional cuisines like Sichuan, Cantonese, Hunan, and Shanghainese. It includes techniques such as stir-frying, steaming, braising, and focuses on ingredients like soy sauce, ginger, garlic, and fermented products (e.g., doubanjiang, fermented tofu). Terms are often borrowed from Chinese languages even when writing in English, particularly for dish names, ingredients, and cooking methods. Russian cooking, on the other hand, leans toward root vegetables, dairy, baking, and stewing. There's less focus on wok cooking or umami-rich fermented ingredients. While both cultures value home cooking, the vocabulary around the ingredients and cooking techniques differs significantly. Therefore, we can identify words and phrases commonly used in English-language posts by Chinese individuals that reflect unique elements of Chinese culinary tradition, which are much less likely to appear in Russian cooking posts.
 Words and Phrases:
 1. wok hei
 2. doubanjiang
 3. red braised pork
 4. mala hotpot
 5. liangpi
 ...

Task Input:
 Target demographic: {target}
 Contrast demographic: {contrast}
 Topic: {topic}

Task Output

Figure 21: Prompt when both demographics and topic is given.

Task Details:
 You are given a **topic** and two demographics; a **target demographic** and **contrast demographic**. Your task is to come up with **15 cultural, sociological, or linguistic theories** about how the **target group** talks about the topic, especially as opposed to the contrast demographic. Then for each theory, come up with keywords and phrases to help retrieve posts from the target demographic. Output at least 10 words for each theory. Follow the response format specified below.

Instruction for Theories:
 - The theories should be **nuanced and sophisticated**. Don't just combine demographic + topic.
 — Some bad examples for the topic "Bread":
 · Hindus eat naan, roti, paratha, etc.
 · Catholics talk about eucharist, host, communion, etc.
 - Start by thinking about the factors that are relevant to the topic (which is not demographic dependent).
 — For example, for the topic "Health", the factors that would be relevant are genetics, age, diet, physical activity, substance use, healthcare access, financial status, etc.
 - Then think about how these factors manifest in the target demographic, especially as opposed to the contrast demographic.
 — For example, for the factor "genetics", target demographic "Hispanics", and contrast demographic "Black":
 Good example: gallbladder disease (considered high predisposition in Hispanics but low in Black people).
 Bad example: Type 2 Diabetes (both groups have high predisposition to this).
 - The theories will be validated by real sociolinguists and social scientists, so try to come up with new theories that most people would not think of. Don't be conservative, instead get creative!

Instruction for Keywords and Phrases:
 - Each word/phrase will be used for a direct look-up so focus on high quality words.
 - Make sure each word/phrase is likely to be used a lot.
 — Bad example: "spirit flowing through the bones"
 — Good example: "warm tonic"
 - Don't use words or phrases that could only ever be used by one demographic (e.g. "ahimsa", "moksha", "kosher", "mitzvah", "catechesis", "sacrament", etc.). This includes demographic-specific foreign language words, holidays, traditions, etc. Do not use these words.

Response Format (angle brackets are placeholders for your output):
 Reasoning: <The factors I think are relevant are... This is what I know about this demographic related to these factors... contrasted with the contrast demographic I think... therefore...>

Final Outputs:
 1. Theory 1: <your first theory>
 Keywords and Phrases: <word>, <phrase>, ...
 .
 .
 15. Theory 15: <your last theory>
 Keywords and Phrases: <word>, <phrase>, ...

Input:
 Target demographic: {target}
 Contrast demographic: {contrast}
 Topic: {topic}

Figure 22: Prompt when both demographics and topic is given to generate creative lexicon.

```

### Task Details:
You are given a research paper that studied a target demographic in a
certain context. Your task is to extract all cultural, sociological, or
linguistic findings that were demonstrated in the paper. Then, as a direct
consequence of each finding being true, come up with 3 sets of keywords and
phrases that would be used by target demographic within that context. Output at
least 5 words/phrases in each set. Follow the response format specified below.

### Instruction for Findings:
- The findings should be direct results demonstrated in the paper.

### Instruction for Keywords and Phrases:
- Given that you know the findings are proven, what set of keywords/phrases
would be used by the target demographic more than other demographics within
the context?
- Each word/phrase will be used for a direct look-up so focus on high quality
words.
- Make sure each word/phrase is likely to be used a lot.
- Bad examples: *spirit flowing through the bones*, *it felt like such a long
day*, *wasn't even making eye contact with me*
- Good examples: *warm tonic*, *haven*, *supported*, *worth exploring*
- Don't use words or phrases that could only ever be used by one
demographic (e.g. *ahimsa*, *moksha* (Hindus), *kosher*, *mitzvah*
(Jewish people), *catechesis*, *sacrament*, (Catholics) etc.). This includes
demographic-specific foreign language words, holidays, traditions, etc. Do not
use these words.

### Response Format (angle brackets are placeholders for your output):
1. Finding: <extracted finding>
- Keywords and Phrases: <word>, <phrase>, ... (at least 5)
- Keywords and Phrases: <word>, <phrase>, ... (at least 5)
- Keywords and Phrases: <word>, <phrase>, ... (at least 5)
2. Finding: <extracted finding>
- Keywords and Phrases: <word>, <phrase>, ... (at least 5)
- Keywords and Phrases: <word>, <phrase>, ... (at least 5)
- Keywords and Phrases: <word>, <phrase>, ... (at least 5)
...
### Input: Target demographic: {target}
Context: {topic}
Paper: (attached)

```

Figure 23: Prompt to generate a lexicon aligning with the results of a published research paper.