SubData: Bridging Heterogeneous Datasets to Enable Theory-Driven Evaluation of Political and Demographic Perspectives in LLMs

Anonymous ACL submission

Abstract

As increasingly capable large language models (LLMs) emerge, researchers have begun exploring their potential for subjective tasks. While recent work demonstrates that LLMs can be aligned with diverse human perspectives, evaluating this alignment on actual downstream tasks (e.g., hate speech detection) remains challenging due to the use of inconsistent datasets across studies. To address this issue, in this resource paper we propose a two-step framework: we **(1)** introduce SubData, an open-source Python library designed for standardizing heterogeneous datasets to evaluate LLM perspective alignment; and **(2)** present a theory-driven approach leveraging this library to test how differently-aligned LLMs (e.g., aligned with different political viewpoints) classify content targeting specific demographics. SubData's flexible mapping and taxonomy enable customization for diverse research needs, distinguishing it from existing resources. We invite contributions to add datasets to our initially proposed resource and thereby help expand SubData into a multi-construct benchmark suite for evaluating LLM perspective alignment on NLP tasks.

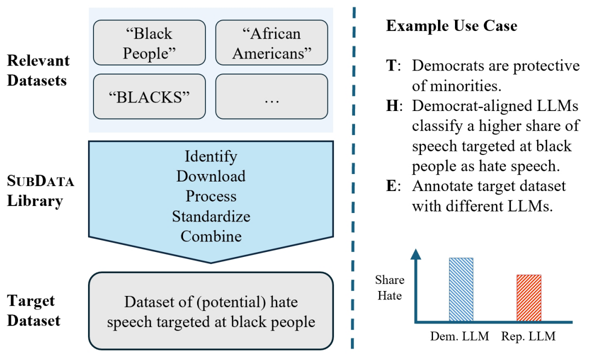

Figure 1: Overview of our proposed evaluation framework. The SubData library consolidates instances from diverse datasets into a unified resource. To assess LLM alignment with human perspectives from the combined dataset, we propose a workflow that tests theory-derived (T) hypotheses (H) through controlled experiments (E), measuring how accurately LLMs reflect viewpoints of different demographic and ideological groups.

1 Introduction

The ever-increasing capabilities of today's large language models (LLMs) have enabled these systems to represent increasingly nuanced human perspectives (Brown et al., 2020; Bommasani et al., 2021). Researchers have begun exploring these models' potential for subjective tasks, with particular focus on "perspective alignment"—the ability of these models to accurately reflect diverse human viewpoints across different contexts (Durmus et al., 2023; Kirk et al., 2024). Ensuring robust evaluation of this alignment is crucial as LLMs increasingly mediate information access and influence decision-making in socially sensitive domains where human perspectives naturally differ (Blodgett et al., 2020; Weidinger et al., 2021; Khamassi et al., 2024).

Recent research has explored how well LLMs can represent diverse human perspectives using two different approaches. The first approach evaluates whether these models accurately predict how specific individuals (Argyle et al., 2023) or groups (Santurkar et al., 2023) would respond to surveys, similar to what Sorensen et al. (2024) introduce as *distributional pluralism* in their position paper on pluralistic LLM alignment. The second approach evaluates whether aligned LLMs consistently reflect broad viewpoints across a range of tasks (Feng et al., 2023; Agiza et al., 2024; Chen et al., 2024; Haller et al., 2024; He et al., 2024), similar to what Sorensen et al. (2024) call *steerable pluralism*.

For survey response prediction, researchers can directly evaluate alignment by comparing the actual survey responses provided by individuals or subpopulations (the "ground truth") with the predictions generated by LLMs attempting to represent these perspectives (either via fine-tuning or persona-based prompting). Existing survey datasets are particularly valuable for this task because they contain both demographic information about respondents and their authentic responses. This creates a clear evaluation framework: a well-aligned LLM should produce outputs that closely match what the real individuals or groups actually said in their survey responses. As suggested by Sorensen et al. (2024), these types of survey prediction can be "compared to the population distribution using any distributional divergence metrics [..] or hard measures [..]" and are thus relatively easy to evaluate.

The broader challenge of task-independent alignment—ensuring LLMs accurately represent diverse perspectives across different contexts—has inspired various evaluation methodologies. Political alignment studies by Agiza et al. (2024) and Chen et al. (2024) use the Political Compass Test (PCT)—a widely used questionnaire for mapping political beliefs along economic and social axes—to verify whether models aligned to specific ideologies position themselves appropriately on the PCT map. He et al. (2024) compare model answers to multiple-choice questions against positions expressed by relevant subgroups. Sorensen et al. (2024) propose direct human annotations or reward models to measure whether generated responses correctly reflect specific attributes. More closely related to our conceptualization of alignment evaluation, Haller et al. (2024) assess sentiment in open-ended generations when prompted about different demographics, while Feng et al. (2023) examine how political alignment affects hate speech detection performance toward different targets.

While these evaluation methods help verify alignment at a general level, evaluating how perspective-aligned LLMs perform on subjective classification tasks remains challenging (Zheng et al., 2024), primarily due to the lack of standardized resources that enable consistent comparison across different human viewpoints (Alipour et al., 2024). We address this gap by introducing a two-step framework that enables systematic evaluation of perspective-aligned language models.

(1) Dataset Standardization: SubData We contribute SubData, an open-source Python library that collects, combines, and standardizes heterogeneous datasets for subjective tasks. Unlike general repositories that provide access to raw data, SubData automates the unification of inconsistent annotation schemes and demographic categorizations, enabling researchers to create consistent collections tailored to specific research needs. Our initial implementation focuses on hate speech detection, integrating ten diverse datasets with a unified taxonomy of target groups (§3, §4, §5). While we developed SubData primarily for evaluating LLM perspective alignment (as detailed in subsequent sections), its harmonization of hate speech taxonomies connects to broader research efforts. Fillies and Paschke (2025) showed that unifying datasets and taxonomies directly enhances classification performance when training task-specific models. Moreover, SubData enables empirical investigations like those by Yu et al. (2024) on dataset creation dynamics, revealing discrepancies between operationalized targets and those actually represented in resulting resources.

Importantly, SubData does neither produce any novel annotations nor does it check the quality of the existing annotations. The library serves the purpose of combining existing annotated datasets into novel resources based on the relevant unit of interest. We facilitate the access to existing datasets and thereby maintain the original purpose of fostering hate speech research expressed by the dataset creators when making their data available to the community. Even though we do not host or redistribute any datasets ourselves, we still consulted the licenses under which the datasets are released to make sure we are not acting against dataset creators' intentions. We echo Vidgen and Derczynski (2020) with their recommendation to consider the social implications of personally-identifying information and offensive content for issues such as privacy and online harm when using this type of data.

(2) Theory-Driven Hypothesis Testing Building on these standardized datasets, we propose a theory-driven approach to evaluate alignment (§6). As illustrated in Figure 1, our framework follows a systematic process: researchers first formulate hypotheses (H) based on established social or political theory (T), then design experiments (E) to test whether differently-aligned models behave as expected. The right side of Figure 1 demonstrates the proposed workflow with a possible use case—testing the hypothesis that

Democrat-aligned LLMs will classify more content targeting Black people as hate speech compared to Republican-aligned LLMs, based on research suggesting Democrats prioritize protecting minorities (Solomon et al., 2024). The visualization shows how our framework would enable the quantitative measurement of these alignment differences through controlled experimentation, with the bar chart illustrating potential findings.

Our theory-grounded approach does not require ground truth labels, thus circumventing the inherent subjectivity of human annotations for subjective constructs. Instead, it directly measures classification differences between models aligned with different perspectives, providing a clear assessment of alignment effects. Although existing work has examined subjectivity in LLM annotations (Orlikowski et al., 2023; Beck et al., 2024; Giorgi et al., 2024), our framework specifically addresses the evaluation of perspective alignment on downstream tasks.

2 Related Work

Approaches to LLM Perspective Alignment

Research on aligning LLMs with diverse human perspectives has followed two main approaches: fine-tuning models on perspective-specific data and using persona-based prompting.

Several studies have explored fine-tuning approaches for task-agnostic LLM alignment. Feng et al. (2023), Agiza et al. (2024) and Chen et al. (2024) investigated how political alignment and data selection affect model biases and downstream tasks like hate speech detection. Similarly, Haller et al. (2024) developed OpinionGPT by fine-tuning models on ideologically diverse data to represent explicit biases.

As an alternative to these resource-intensive post-training methods, persona-based prompting has emerged as a more efficient technique for task-specific perspective alignment. Argyle et al. (2023) showed that LLMs can accurately simulate survey responses across demographic groups, while Ge et al. (2024) and Fröhling et al. (2024) demonstrated how synthetic personas can diversify model outputs and annotations. Building on this, Bernardelle et al. (2024) mapped persona-prompted LLMs onto the political compass, providing a large-scale analysis of how these personas impact the distribution of language models across political ideological space. Similarly, Civelli et al. (2025) revealed how politically-aligned persona-prompted LLMs influence hateful content detection.

Orlikowski et al. (2025) combined these approaches by fine-tuning models with sociodemographic attributes to represent individual annotators, finding that persona-based prompting barely improves the models' ability to predict individuals' annotations and that improvements from fine-tuning mainly come from demographic profiles serving as identifiers for individual annotators. Liu et al. (2024) identified further limitations in this technique, showing that models struggle with "incongruous personas" and default to stereotypical stances when predicting responses for personas with contradicting traits. The conflicting evidence seen in the literature regarding the models' ability to consistently represent different subjective perspectives serves as further motivation to develop comprehensive resources for the evaluation of this type of LLM perspective alignment.

Evaluating LLM Perspective Alignment

Evaluating alignment presents significant challenges, particularly for subjective tasks.

For survey response prediction, Santurkar et al. (2023) and He et al. (2024) compared model predictions against actual responses from specific demographic groups. Castricato et al. (2025) built on the PRISM dataset (Kirk et al., 2024) to create a test bed for evaluating pluralistic alignment using preference pairs from personas sampled from census data.

For downstream tasks, Zheng et al. (2024) and Giorgi et al. (2024) assessed how personas affect model performance and biases in content classification. Despite these advances, evaluating perspective-aligned LLMs on subjective classification tasks remains challenging due to the lack of standardized resources that enable consistent comparison—a gap our proposed framework addresses.

3 SubData Construction

3.1 Dataset Selection Criteria

Our approach to evaluating perspective alignment in LLMs necessitates datasets with specific characteristics suited for this analysis. We require datasets that address subjective constructs such as hate speech, toxicity, or abusive language—domains where human interpretations naturally diverge across demographic and ideological lines Sap et al.

Dataset \ Category	age	disabled	gender	migration	origin	political	race	religion	sexuality	Dataset size
Fanton et al. (2021)	0 (0)	175 (1)	560 (1)	637 (1)	0 (0)	0 (0)	301 (1)	1,402 (2)	465 (1)	3,540
Hartvigsen et al. (2022)	0 (0)	19,631 (1)	19,563 (1)	0 (0)	62,458 (3)	0 (0)	80,979 (4)	41,014 (2)	21,344 (1)	244,989
Jigsaw (2019)	0 (0)	18,602 (3)	178,266 (4)	0 (0)	0 (0)	0 (0)	94,334 (5)	132,734 (7)	29,115 (4)	453,051
Jikeli et al. (2023a)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6,439 (1)	0 (0)	6,439
Jikeli et al. (2023b)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3,012 (3)	2,315 (2)	0 (0)	5,327
Mathew et al. (2021)	0 (0)	153 (1)	5,584 (2)	1,701 (1)	1,855 (2)	0 (0)	7,684 (5)	6,106 (6)	2,750 (4)	25,833
Röttger et al. (2021)	0 (0)	510 (1)	1,020 (2)	485 (1)	0 (0)	0 (0)	504 (1)	510 (1)	577 (1)	3,606
Sachdeva et al. (2022)	2,355 (4)	1,801 (3)	22,535 (5)	5,473 (2)	11,637 (2)	0 (0)	21,024 (7)	12,461 (8)	14,934 (4)	92,220
Vidgen et al. (2021a)	41 (2)	414 (3)	689 (3)	45 (2)	164 (5)	688 (7)	397 (4)	273 (4)	472 (3)	3,183
Vidgen et al. (2021b)	23 (1)	521 (1)	3,630 (4)	1,507 (2)	862 (6)	0 (0)	3,881 (5)	2,384 (2)	1,437 (3)	14,245
All Datasets	2,419 (4)	41,807 (3)	231,847 (5)	9,848 (4)	76,976 (11)	688 (8)	212,116 (8)	205,638 (8)	71,094 (6)	852,433

Table 1: Overview of hate speech datasets in SUBDATA, showing the number of instances and unique target groups (in parentheses) per target category. *Note*: The "All Dataset" row reports the total unique target groups per category across all datasets. When the total equals the maximum from a single dataset (e.g., disabled: 3, matching Jigsaw (2019)'s 3), that dataset fully accounts for the category's unique target groups. When the total exceeds the maximum (e.g., origin: 11, exceeding Hartvigsen et al. (2022)'s 3), multiple datasets contribute distinct target groups, increasing the total.

(2021). This subjectivity is essential as it creates the interpretive space where different perspectives become measurable. Additionally, these datasets must provide explicit annotations identifying which specific demographic groups are targeted by the content (for example, specifying when content targets Jews, women, or immigrants), rather than merely indicating that some unspecified group was targeted. This granular targeting information is crucial because it enables us to test theory-driven hypotheses about how LLMs aligned with different perspectives might classify content targeting specific demographics differently.

3.2 Data Collection Methodology

Because of the lack of a single repository that stores and documents the properties of datasets, identifying the set of relevant datasets is an inherently difficult challenge. We therefore employed a multi-phase approach to identify suitable datasets.

First, we leveraged our existing knowledge of hate speech detection literature to identify candidate datasets, drawing on our team's established expertise in this domain. Second, we examined existing repositories including hatespeechdata.com (Vidgen and Derczynski, 2020) and toxic-comment-collection (Risch et al., 2021), which provided structured access to multiple potentially relevant datasets. Third, we conducted systematic searches with keyword combinations of "target[ed]" and "hate speech" on scholarly databases to identify related literature that might present or reference additional resources. Finally, we individually assessed each dataset through manual verification to confirm it contained explicit target group annotations that satisfied our criteria.

This process yielded ten datasets that meet our requirements. While we have striven to make our initial dataset collection comprehensive, we acknowledge that this collection is not exhaustive and that some relevant sources may have been overlooked. Rather than seeing this as a limitation, we consider it an opportunity to build a collaborative research community focused on annotation subjectivity. We actively encourage researchers to contact us with suggestions for additional datasets that satisfy our outlined criteria to be included in the SUBDATA library.

3.3 Dataset Characteristics

Table 1 provides an overview of the datasets included so far in SUBDATA, categorizing targets across nine demographic dimensions (age, disability, gender, migration, origin, political, race, religion, and sexuality). All target categories are organized according to the unified taxonomy we detail in §4, which standardizes the heterogeneous labels from original sources. This standardized categorization enables researchers to quickly identify suitable datasets for specific research questions regarding perspective alignment, highlighting both the strengths and limitations of current hate speech detection resources.

We would like to point out that the number of entries in some datasets of Table 1 may differ from those reported in the original publications because of our focus on targeted hate speech. When entries in source datasets had multiple targets in a single annotation (e.g., "[bla, jew]"), we created separate instances for each target, thereby increasing the number of entries. Conversely, we excluded entries without specific target groups (e.g., labeled as "other"), resulting in datasets that sometimes contain fewer instances than the originals. We also deduplicate instances, removing repeated entry-target pairs even when these duplications might be intentional in the original dataset—such as in Fanton et al. (2021) where identical hate speech instances appear multiple times with different counterspeech responses. Since our research focuses specifically on targeted hate speech, we treat these as functional duplicates.

4 SubData Unified Taxonomy

Following our dataset selection and collection methodology, SubData implements a standardized taxonomy that addresses the inconsistencies in how target groups are labeled across hate speech datasets. This allows to leverage the systematic evaluation framework described in §6 by creating consistency across disparate data sources.

4.1 Taxonomy Design Principles

The development of our taxonomy was guided by several key design principles that reflect the practical needs of researchers studying perspective alignment. We aimed to balance specificity and generalizability by preserving important distinctions between target groups while creating categories broad enough to enable meaningful cross-dataset analysis. As an example of such a consideration serves the target group "LGBTQ+", oftentimes used in the literature to represent any minority sexual or gender identity. While we consider this too broad and diverse a label as to meaningfully represent the very different types of target groups it covers, we also decided against introducing every identity group that identifies with this umbrella term as an independent target group. In the end, we chose to be practical by using the LGBTQ+-related target groups frequently used in the literature. When possible, we maintained consistency with the original researchers' taxonomic decisions to preserve their methodological choices and conceptual frameworks.

4.2 Target Group Mapping

The mapping process converts heterogeneous target labels from original datasets into our standardized taxonomy. This involves both direct equivalences (e.g., "Jewish people" → "jews") and more complex decisions requiring contextual judgment. Table 2 provides a sample of our mapping strategy across multiple datasets, illustrating how diverse original terminology is standardized in SubData.

Dataset	Original Keyword	Target
Fanton et al. (2021)	"JEWS"	jews
Hartvigsen et al. (2022)	"jewish"	jews
Jikeli et al. (2023a)	"Kikes"	jews
Vidgen et al. (2021a)	"jewish people"	jews
Vidgen et al. (2021b)	"bla, jew"	jews blacks
Vidgen et al. (2021b)	"bla, african"	blacks
Jigsaw (2019)	"black"	blacks
Jikeli et al. (2023b)	"Blacks"	blacks
Röttger et al. (2021)	"black people"	blacks

Table 2: Standardization of target terminology across datasets using SubData's mapping system. The table provides examples of how diverse original keywords from multiple hate speech datasets are normalized into consistent target categories.

For ambiguous cases, we consulted dataset documentation to determine the original authors' intent. For instance, determining whether the target "africans" should be mapped to "blacks" (race category) or "africans" (origin category) required careful contextual judgment. When documentation clarified the original creators' intended meaning, we followed their categorization. When such guidance was unavailable, we applied consistent principles across similar cases.

As part of our approach, for each category we designated target groups with the suffix "_unspecified" (e.g., "disabled_unspecified," "race_unspecified") to handle cases where the original dataset used generic terminology without specifying subtypes.

Figure 2 illustrates the complete taxonomy structure with all target groups organized by category.

4.3 Taxonomy Limitations and Customization

Despite our efforts to create a comprehensive framework, we acknowledge several limitations in our taxonomy that primarily stem from the inherent challenges associated with the matching we are performing (Shvaiko and Euzenat, 2011). These include the LGBTQ+ target group heterogeneity that mixes gender identities and sexual orientations, blurred distinctions between racial identity and geographic origin, and simplified representations of demographic intersectionality mapped to single-attribute target groups (e.g., "blacks,women"). Independent from our work, Fillies and Paschke (2025) point to the same challenges when developing their targeted hate speech taxonomy, relying on similar strategies to solve them.

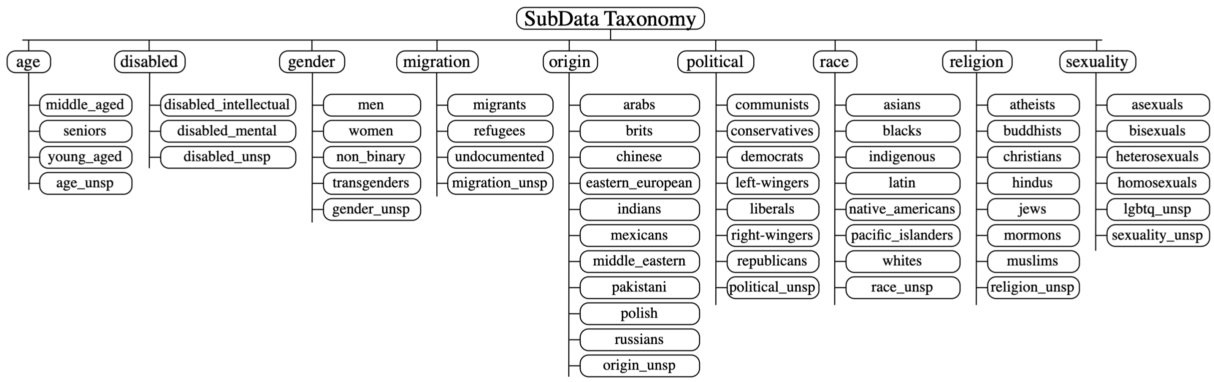

Figure 2: SubData taxonomy structure with target groups organized by category. *Note:* targets that should end in "_unspecified" have been abbreviated in the figure using "'_unsp."

We are confident that our taxonomy represents a useful basis for different research purposes and take the large overlap with the unified taxonomy proposed by Fillies and Paschke (2025) as evidence for convergence on a generally accepted targeted hate speech taxonomy. However, recognizing that no single taxonomy can satisfy all research needs, SubData provides several customization functions that give researchers flexibility in adapting the framework to their specific requirements (more about it in §5). While this customizability is valuable, it creates challenges for maintaining comparability across studies when researchers modify the taxonomy. To address this issue and increase transparency, we implemented functionality to export a LaTeX version of the taxonomy (and all other modifiable resources) that researchers can include directly in their manuscripts, clearly documenting any modifications they have made.

5 SubData Library

While Figure 1 gives an abstracted overview of the SubData library's basic workflow, the user-facing functionalities are documented in the following subsections. Building upon the dataset selection strategy outlined in §3 and the taxonomy and mapping strategies described in §4, the library offers a flexible framework that enables researchers to: **(1)** access instances targeting specific demographic groups across multiple datasets; **(2)** customize the taxonomy and mapping according to specific research needs; and **(3)** generate consistent datasets for evaluating LLM perspective alignment.

5.1 Core Functionalities

The library's functionality can be organized into three main categories:

Dataset Creation and Access

- `create_target_dataset()`: Generates a dataset containing instances targeting the specified valid target group (e.g., "jews", "blacks") from all available datasets. Returns a dataframe with instance ID, text, target name, and source dataset.

- `create_category_dataset()`: Assembles instances targeting all groups within a specified category (e.g., "religion," "race"). Downloads and processes all datasets containing any target groups within the specified category.

- `get_target_info()`: Displays available instances for specific target groups, showing distribution across datasets and availability status. Displays the total number of instances available, lists source datasets with counts, and provides access requirement information for restricted datasets.

- `get_category_info()`: Provides an overview of available instances for all groups within a category. Displays total instance counts across all target groups in the category, breaks down counts per target, and shows dataset availability information.

Taxonomy Customization

- `show_taxonomy()`: Displays and exports the specified taxonomy. Either returns the full taxonomy or only the specified categories (either "all" or a list of category names). When LaTeX export is enabled, the function generates formatted tables in a txt file, making it convenient to include taxonomy details in academic papers.

- `update_taxonomy()`: Reorganizes target groups across categories or creates new categories. This function accepts a dictionary of taxonomy changes (specifying which targets to move from which categories to which new categories) and stores the modified taxonomy under the provided name. If a target is moved to a non-existent category, a new category will be created automatically, allowing for flexible taxonomy extension.

- `add_target()`: Creates entirely new target groups when needed. This function requires three parameters: the name of the new target, the existing category to place it in, and a list of original dataset keywords that should map to this new target. Stores the modified taxonomy and mapping under the provided names.

- `update_overview()`: Updates the internal dataset overview that informs the dataset creation and information functions. This function should be called after any taxonomy or mapping modifications to ensure that future calls of the dataset-generating functions access the correct resources.

Mapping Modification

- `show_mapping()`: Displays and exports the specified mapping between original dataset keywords and standardized target groups. Either returns the individual mappings for all datasets or only for those specified (either "all" or a list of dataset names). The LaTeX output consists of separate tables for each dataset, clearly documenting the keyword-to-target transformations used in the research pipeline.

- `update_mapping_specific()`: Modifies mappings for individual datasets, allowing dataset-specific customization of how original dataset labels map to standardized target groups. This function accepts a nested dictionary specifying which keywords in which datasets should be mapped to which target groups. Stores the modified mapping under the provided name.

- `update_mapping_all()`: Applies mapping changes consistently across all datasets, ensuring uniform treatment of keywords. This function takes a dictionary mapping original keywords to new target groups, affecting all datasets where those keywords appear. It stores the modified mapping under the provided name.

Dataset Overview

- `update_overview()`: Updates the internal dataset overview that informs the dataset creation and information functions. This function should be called after any taxonomy or mapping modifications and accepts parameters for naming the modified configurations (overview_name, mapping_name, taxonomy_name), as well as an optional authentication token (hf_token) for accessing restricted datasets.

- `show_overview()`: Displays and exports the specified overview based on the specified taxonomy. This function accepts an overview name and taxonomy name as parameters, with boolean options to control JSON export (export_json) and LaTeX table export (export_latex). When LaTeX export is enabled, the function generates formatted tables in a txt file. The function returns the overview as a dictionary.

5.2 Implementation and Availability

All code is available open-source on GitHub[1] and the library can be installed directly from PyPi[2]. The library handles dataset availability transparently—if a dataset is not openly available, the functions inform users how to access it, either by providing authentication credentials or manually downloading and storing datasets in a specified location.

6 Theory-Driven Hypothesis Testing

The SubData library not only provides standardized datasets but also serves as a foundation for a theory-driven approach to evaluating LLM perspective alignment. This approach follows the process illustrated in Figure 1:

1. **Theory (T)**: Researchers begin by identifying established social or political theories that predict differences in how various demographic

[1] https://github.com/Subdata-Library/Subdata/
[2] https://pypi.org/project/subdata/

or ideological groups differ in their perception of subjective constructs.

2. **Hypothesis (H)**: Based on these theories, researchers formulate testable hypotheses about how LLMs aligned with different perspectives might classify content.

3. **Experiment (E)**: Using SUBDATA's standardized datasets, researchers design controlled experiments to test these hypotheses by measuring classification differences between differently-aligned models.

6.1 Advantages of the Framework

The theory-driven framework we propose offers substantial benefits for researchers studying LLM perspective alignment. By focusing on comparative model behavior rather than adherence to supposedly objective standards, our approach **(1) elegantly circumvents the persistent challenge of subjectivity in human annotations**. When dealing with inherently subjective constructs like hate speech, the framework does not require consensus on "ground truth" labels—which are often contested and vary across demographic and ideological lines—but instead directly measures differences between models aligned with distinct perspectives. This shift in evaluation methodology acknowledges the fundamental subjectivity of these tasks while still enabling rigorous analysis by grounding the tested hypotheses directly in theory.

Furthermore, our approach **(2) enables precise quantitative measurement of alignment effects on classification behavior**. Researchers can measure exactly how much perspective alignment influences model outputs when classifying content targeting specific demographics, providing concrete metrics rather than relying on qualitative assessments. This quantitative foundation makes evaluations more rigorous and facilitates meaningful comparisons across different studies, contributing to more cumulative research in this emerging field.

The framework's versatility extends beyond its primary application in political alignment evaluation. It **(3) naturally supports diverse research directions**. This flexibility makes our approach valuable for researchers working at the intersection of natural language processing, social science, and ethical AI development, potentially informing more nuanced approaches to model development and evaluation.

6.2 Key Use Cases

We naturally see the main use case for the presented SUBDATA library and its associated theory-driven evaluation framework in what we built it for, the evaluation of the alignment of LLMs with different human perspectives on downstream tasks. However, another strain of research that would likely benefit from the provided standardized access to targeted hate speech datasets is research on hate speech itself. While there has been very notable work to create a repository of hate speech datasets by Vidgen and Derczynski (2020) and to even facilitate and standardize access to them by Risch et al. (2021), no such resource is available for targeted hate speech specifically. The need for a stronger focus on the targets of hate speech has recently been presented by Recently, Yu et al. (2024) have argued for a stronger focus on the targets of hate speech. By making the target group the unit of interest based on which the data is ultimately downloaded and assembled, we think that our SUBDATA library is a natural fit for this emphasis on the target groups in hate speech research. Aggregating different source datasets into new datasets based on their target groups further increases the reusability of the existing datasets for novel applications.

7 Future Extensions

The most immediate extension of our SUBDATA library is the inclusion of additional datasets, both those that we may have overlooked in our initial collection as well as those that are yet to be released. In parallel, we aim to cultivate a community of researchers interested in aligning LLMs with diverse human viewpoints, which would naturally accelerate the inclusion of additional datasets.

Beyond including more dataset, we plan to broaden the scope of SUBDATA by introducing additional subjective constructs. Our next priority is misinformation, for which we have already compiled an initial collection of datasets that will soon be accessible through the library.

Ultimately, we intend to develop an alternative evaluation approach for LLMs alignment with different human viewpoints, focusing on annotator characteristics rather than instance features. Through these initiatives, we aspire to evolve SUBDATA into a comprehensive multi-construct benchmark suite for evaluating how well LLMs align with humans across various downstream tasks.

Limitations

While the initial implementation of SUBDATA focuses on hate speech detection with a unified taxonomy of target groups, we acknowledge certain limitations in our approach. A first limitation is the currently narrow focus on targeted hate speech, as these are the only datasets already available through the library. However, we decided to already publish SUBDATA because the alignment of LLMs is a very recent and relevant topic in NLP with novel methods being presented frequently, but lacking resources for the systematic evaluation of the quality of the alignment, particularly for downstream tasks. We are confident that the library in its current state will thus already proof to be helpful for researchers studying the alignment of LLMs with diverse perspectives.

We made pragmatic choices in mapping target groups across datasets, which necessarily involve subjective judgments about categorization. As already laid out in §4, the unification of different taxonomies is a challenging endeavor. The inherent limitations we acknowledge include the existence of target groups in the literature that conflate targets from different categories (e.g., "LGBTQ+" for minority gender identities and sexual orientations) or that are put into different categories in different original datasets (e.g., "africans" either put into a race or an origin category). Lastly, there are datasets that combine multiple target groups to represent intersectional target groups (e.g., "blacks, women"). While we inherit these challenges from the different source datasets, we tried to apply our taxonomy principles carefully and consistently to create a comprehensive taxonomy that balances specificity and generalizability. Additionally, our framework provides flexibility for researchers to customize these mappings according to their specific research needs—for instance, adjusting categories to focus on particular demographics or redefining target groups entirely. This customization capability mitigates the limitation of any single taxonomic approach.

In addition to inheriting the challenges associated with mapping target groups into a unified taxonomy, the aggregated datasets inherit any annotation errors and biases from the individual source datasets. SUBDATA aggregates the annotated instances from the featured source datasets into novel datasets built around a specified target group. This process does not produce any new annotations nor does it check the quality of the source dataset annotations. We therefore encourage users to perform their one quality checks of the source dataset annotations, as well as to consult the original datasets' documentation if in doubt.

Ethical Considerations

While SUBDATA provides valuable datasets for evaluating LLMs perspective alignment, we acknowledge potential ethical concerns. The library's aggregation of hate speech datasets creates a concentrated collection of offensive content that could be misused to train hateful models or generate toxic content. Additionally, our framework's ability to test how differently-aligned LLMs classify content targeting specific demographics could be misused to intentionally create biased systems. We emphasize that SUBDATA's purpose is to improve evaluation transparency and understanding of perspective alignment, not to enable harmful applications. We recognize that the target groups represented in these datasets face real discrimination and harassment. Research using SUBDATA should be conducted with sensitivity to the lived experiences of these communities, and findings should be communicated in ways that avoid reinforcing harmful stereotypes or creating additional psychological harm.

References

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 2–12.

Shayan Alipour, Indira Sen, Mattia Samory, and Tanushree Mitra. 2024. Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness. *arXiv preprint arXiv:2411.08977*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2024. Mapping and influencing the political ideology of large language models using synthetic personas. *arXiv preprint arXiv:2412.14843*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How susceptible are large language models to ideological manipulation? *arXiv preprint arXiv:2402.11725*.

Stefano Civelli, Pietro Bernardelle, and Gianluca Demartini. 2025. The impact of persona-based political perspectives on hateful content detection. *arXiv preprint arXiv:2502.00385*.

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

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762.

Jan Fillies and Adrian Paschke. 2025. Improving hate speech classification with cross-taxonomy dataset integration. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 148–159.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2024. Personas with attitudes: Controlling llms for diverse data annotation. *arXiv preprint arXiv:2410.11745*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *arXiv preprint arXiv:2410.07991*.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2024. Opiniongpt: Modelling explicit biases in instruction-tuned llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 78–86.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Zihao He, Minh Duc Chu, Rebecca Dorn, Siyi Guo, and Kristina Lerman. 2024. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. *arXiv preprint arXiv:2406.12074*.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Accessed: 2024-11-19.

Gunther Jikeli, Sameer Karali, Daniel Miehling, and Katharina Soemer. 2023a. Antisemitism on Twitter: A Dataset for Machine Learning and Text Analytics.

Gunther Jikeli, Sameer Karali, and Katharina Soemer. 2023b. Hate Speech and Bias against Asians, Blacks, Jews, Latines, and Muslims: A Dataset for Machine Learning and Text Analytics.

Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports*, 14(1):19399.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9832–9850.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *arXiv preprint arXiv:2502.20897*.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029.

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021. Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

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

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Pavel Shvaiko and Jérôme Euzenat. 2011. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176.

Brittany C Solomon, Matthew EK Hall, Abigail Hemmen, and James N Druckman. 2024. Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it. *Proceedings of the National Academy of Sciences*, 121(39):e2402428121.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, page 08944393241258771.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.