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

Pietro Bernardelle[1] **Leon Fröhling**[2] **Stefano Civelli**[1] **Gianluca Demartini**[1]
[1]The University of Queensland, Australia
[2]GESIS - Leibniz Institute for the Social Sciences, Germany
{p.bernardelle,s.civelli,g.demartini}@uq.edu.au leon.froehling@gesis.org

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 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 LLMs 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 illustrate its usage with an example application and invite contributions to extend our initial release into a multi-construct benchmark suite for evaluating LLMs perspective alignment on natural language processing tasks.

Keywords: Dataset Standardization, Perspective Alignment in LLMs, Hate Speech Detection Resources

1. Introduction

The ever-increasing capabilities of large language models (LLMs) have enabled them to capture increasingly nuanced human perspectives (Bommasani et al., 2021; Brown et al., 2020). Researchers have begun exploring their potential for subjective tasks, with particular focus on "perspective alignment"—the ability of models to 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 decisions in socially sensitive domains where human perspectives naturally differ (Blodgett et al., 2020; Khamassi et al., 2024; Weidinger et al., 2021).

Recent work has explored how well LLMs can represent diverse human perspectives using two different approaches. The first approach examines whether models accurately predict how individuals (Argyle et al., 2023) or groups (Santurkar et al., 2023) would respond to surveys, a task Sorensen et al. (2024) describe as *distributional pluralism*. The second investigates whether aligned LLMs consistently reflect broader viewpoints across tasks (Agiza et al., 2024; Chen et al., 2024; Feng et al., 2023; Haller et al., 2024; He et al., 2024), aligning with what Sorensen et al. (2024) term *steerable pluralism*.

Survey prediction provides a natural evaluation setting: models' outputs—generated either through fine-tuning or persona-conditioning to represent specific perspectives—can be compared against authentic survey responses from individuals or subpopulations (Rupprecht et al., 2025). Because such datasets contain demographic information and corresponding answers, they create a clear benchmark: a well-aligned LLM should produce distributions that closely resemble real responses and can be evaluated using standard divergence or accu-

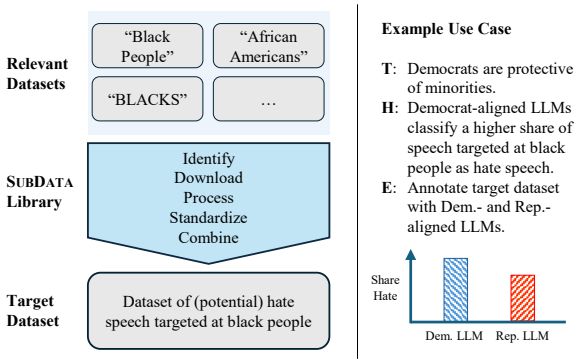

Figure 1: Overview of our proposed evaluation framework. SubData consolidates instances from diverse datasets into a unified resource. To assess LLM alignment with human perspectives from the combined dataset, we propose a workflow that tests theory-derived (**T**) hypotheses (**H**) through controlled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

Proceedings of NLPerspectives @ LREC 2026, pages 84–97
12 May 2026.

racy measures (Sorensen et al., 2024).

The broader challenge of task-independent alignment 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 toward different targets.

Despite these efforts, evaluating how perspective-aligned LLMs perform on subjective classification tasks remains challenging (Zheng et al., 2024), largely due to the lack of standardized resources that enable consistent comparison across viewpoints (Alipour et al., 2025). We address this gap by introducing a two-step framework that enables systematic evaluation of perspective-aligned language models.

(1) Dataset Standardization: SubData. We introduce SubData, an open-source Python library that collects and harmonizes heterogeneous datasets for subjective tasks.[1] Unlike general repositories, it unifies inconsistent annotation schemes and demographic categorizations, allowing researchers to build consistent collections for their needs. Our initial release focuses on hate speech detection, integrating ten datasets with a unified taxonomy of target groups (§3, §4). In doing so, we do not host or redistribute the datasets themselves, but we reviewed their licenses to ensure that our use aligns with creators' intentions of fostering hate speech research. Consistent with Vidgen and Derczynski (2020), we further emphasize the need to handle such material responsibly, with attention to privacy, personal data, and potential online harms.

(2) Theory-Driven Hypothesis Testing. Building on these standardized datasets, we propose a theory-driven approach to evaluate alignment (§5). 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. For instance, the workflow on the right side of Figure 1 illustrates testing whether Democrat-aligned LLMs classify more anti-Black content as hate speech than Republican-aligned ones, reflecting the popular hypothesis that Democrats prioritize minority protection theoretically derived by (Solomon et al., 2024). This framework enables quantitative measurement of alignment differences through controlled experimentation, and we further demonstrate its application in §6.

Our approach does not rely on subjective ground-truth labels; instead, it measures classification differences across models with distinct alignments, providing a direct lens on how perspective conditioning shapes downstream task behavior. While prior work has examined subjectivity in LLM annotation (Beck et al., 2024; Giorgi et al., 2024; Orlikowski et al., 2023), our framework extends this by systematically evaluating alignment effects in downstream applications.

2. Related Work

2.1. LLMs 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 LLMs alignment. Agiza et al. (2024), Chen et al. (2024) and Feng et al. (2023) 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 Fröhling et al. (2025) and Ge et al. (2024) demonstrated how synthetic personas can diversify model outputs and annotations. Building on this, Bernardelle et al. (2025a,b) mapped persona-prompted LLMs onto the PCT compass, providing a large-scale analysis of how these personas impact the distribution of language models across political ideological space. Similarly, Civelli et al. (2025a,b) revealed how politically-aligned persona-conditioned LLMs influence hateful content detection.

[1] All code is available open-source on GitHub and the library can be installed directly from PyPi.

Orlikowski et al. (2025) combined these approaches by fine-tuning models with socio-demographic 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 LLMs perspective alignment.

2.2. Evaluating LLMs Perspective Alignment

Evaluating alignment presents significant challenges, particularly for subjective tasks.

For survey response prediction, He et al. (2024) and Santurkar et al. (2023) 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, Giorgi et al. (2024) and Zheng 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., 2022). 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 library.

3.3. Dataset Characteristics

Table 1 provides an overview of the datasets included in SubData so far, 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

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.

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 §5 by creating consistency across disparate data sources.

4.1. Taxonomy Design Principles

The development of our taxonomy was guided by several key design principles tailored to the practical needs of researchers studying perspective alignment. We sought to balance specificity and generalizability, preserving critical distinctions between target groups while establishing categories broad enough to facilitate meaningful cross-dataset analysis. For instance, the target group "LGBTQ+" is commonly used in the literature to encompass a wide range of minority sexual and gender identities. While we recognize that this label can be overly broad, potentially obscuring the diverse experiences of the groups it covers, we decided against introducing every identity under this umbrella as a separate target group.

Importantly, our demographic categories were not arbitrarily chosen; they emerged from a bottom-up approach, derived directly from the categories present in the original datasets we sourced. This method ensures that our taxonomy reflects and unifies the actual structure of existing hate speech research, maintaining alignment with the data's inherent organization. Additionally, whenever possible, we preserved consistency with the original researchers' taxonomic decisions to honor 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.

For ambiguous cases, we consulted dataset documentation to determine the original authors' intent. For instance, determining whether the target "mexicans" should be mapped to the "latin" (race category) or "mexicans" (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. To validate the reliability of these decisions, we conducted an inter-annotator

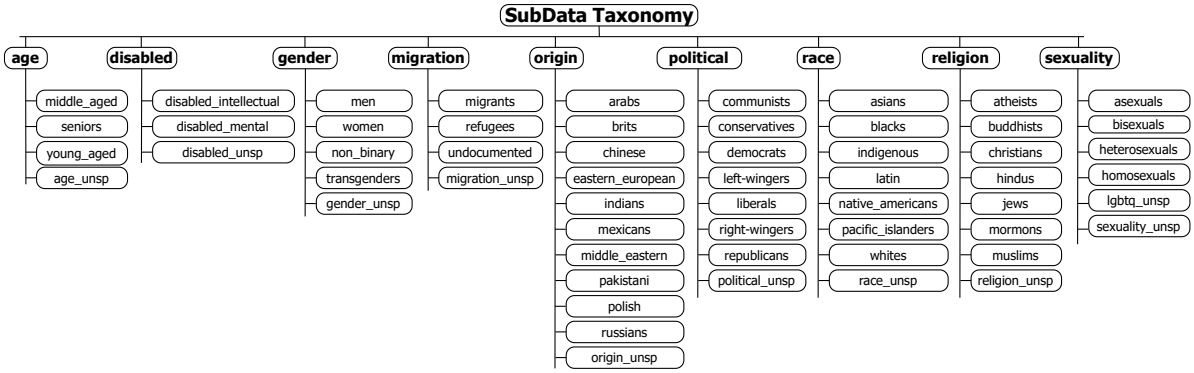

Figure 2: SubData taxonomy structure with target groups organized by category. *Note:* targets that should end in "_unspecified" have been abbreviated in the figure using "'_unsp."

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.

agreement check on the mapping process, obtaining a Cohen's κ of 0.986.

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.

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.

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 a 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. 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 LLMs 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 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.

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 (NLP), social science, and ethical AI development, potentially informing more nuanced approaches to model development and evaluation.

6. Example Use of SubData

To demonstrate the proposed framework, we present here a concrete use case. Feng et al. (2023) show that pretraining LLMs on partisan corpora shifts their political leaning, and that this shift propagates into downstream tasks such as hate speech detection, where left-leaning models tend to flag more content targeting minority groups than right-leaning ones (**T**). This connection is theoretically plausible because hate speech detection is not a politically neutral classification task: decisions about what should be flagged often reflect broader normative commitments around social harm, tolerance, and free expression. As a result, if political conditioning meaningfully changes a model's perspective, hate speech detection is one of the downstream settings where such differences should be especially visible.

Following their categorization (BLACKS, MUSLIMS, LGBTQ+, JEWS, LATIN, WOMEN, MEN, CHRISTIAN, WHITE), we hypothesize that a similar dynamic holds when partisan alignment is induced through persona-conditioning (**H**). Specifically, LLMs conditioned on left-leaning personas should produce higher detection rates for hate speech against minority groups, while LLMs conditioned on right-leaning personas should show the opposite tendency. The following subsections detail the experimental setup (**E**) and results.

6.1. Methodology

Data. We use the SubData library to collect and standardize the instances of interest from existing hate speech datasets. Specifically, we rely on its unified taxonomy of ten demographic groups: BLACKS, MUSLIMS, LGBTQ_UNSP, JEWS, ASIANS, LATIN, WOMEN, MEN, CHRISTIANS, WHITES.

This procedure enables us to construct a single merged dataset that ensures comparability across groups and facilitates controlled evaluation of perspective alignment. Because evaluating the full corpus across multiple models and persona conditions would be computationally prohibitive, we instead randomly sample 2,500 statements per target group after merging, yielding a balanced dataset of 25,000 instances for our experiments.

The sampling procedure chosen does not alter any original frequency distributions of different targets, given that our collection of publicly available datasets should not be treated as a ground-truth distribution of targets in the hate speech literature or in the real world. Furthermore, by choosing 2,500 instances per target group, we do not have to rely on oversampling or synthetic duplication to achieve this stratification. The smallest of the target groups studied in our use case features around 21,000 instances.

Language Models. We selected three open-source, instruction-tuned conversational LLMs for our analysis: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7B-Instruct (Team, 2025). These models were chosen for their open-source availability and moderate parameter size (7–8B), which

Target	Mistral-v0.3-7B				Llama-3.1-8B				Qwen-2.5-7B			
	Left	Right	OR	95% CI	Left	Right	OR	95% CI	Left	Right	OR	95% CI
blacks	**0.548**	**0.471**	1.359^{***}	[1.325,1.393]	0.625	0.614	1.050^{***}	[1.023,1.077]	0.319	**0.311**	1.037^{**}	[1.009,1.065]
muslims	0.471	0.391	1.389^{***}	[1.355,1.425]	0.599	0.576	1.100^{***}	[1.072,1.128]	0.228	0.204	1.152^{***}	[1.117,1.187]
lgbtq_unsp	0.313	0.247	1.389^{***}	[1.351,1.428]	0.391	0.372	1.086^{***}	[1.059,1.114]	0.168	0.157	1.084^{***}	[1.048,1.121]
jews	0.497	0.405	1.450^{***}	[1.415,1.487]	0.576	0.553	1.101^{***}	[1.074,1.129]	**0.320**	0.307	1.064^{***}	[1.036,1.093]
asians	0.343	0.260	1.481^{***}	[1.442,1.522]	0.469	0.446	1.096^{***}	[1.069,1.123]	0.197	0.176	1.146^{***}	[1.111,1.184]
latin	0.378	0.297	1.443^{***}	[1.405,1.482]	0.473	0.447	1.110^{***}	[1.082,1.138]	0.209	0.199	1.064^{***}	[1.032,1.097]
women	0.364	0.298	1.347^{***}	[1.312,1.383]	0.477	0.467	1.042^{**}	[1.016,1.068]	0.161	0.151	1.073^{***}	[1.037,1.110]
christians	0.202	0.167	1.263^{***}	[1.223,1.304]	0.298	0.290	1.040^{**}	[1.012,1.069]	0.064	0.061	1.057^{*}	[1.004,1.112]
men	0.299	0.250	1.279^{***}	[1.244,1.315]	0.442	0.429	1.056^{***}	[1.030,1.083]	0.111	0.104	1.081^{***}	[1.038,1.125]
whites	0.523	0.440	1.393^{***}	[1.359,1.429]	**0.656**	**0.649**	1.033^{*}	[1.006,1.060]	0.259	0.253	1.029^{*}	[1.000,1.059]
Overall	0.394	0.323	1.364^{***}	[1.352,1.375]	0.501	0.484	1.068^{***}	[1.060,1.077]	0.204	0.193	1.073^{***}	[1.063,1.084]

Table 3: Hate speech detection rates by target category and persona position across the three LLMs investigated. Each cell shows the average proportion of content flagged as hateful when targeting the specified group, using a persona-conditioned model with 20 left- and 20 right-oriented personas. Bold values indicate the highest detection rate for each model–condition pair across all targets. Odds Ratios (OR) quantify detection differences, with OR > 1 indicating higher rates for left personas. 95% confidence intervals are reported in a dedicated column. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

strikes a balance between reproducibility and diversity, allowing us to derive insights that generalize across architectures. We specifically use their conversational variants, fine-tuned for instruction following (Ouyang et al., 2022), as this aligns with our methodology: leveraging in-context prompts to condition models on different personas and evaluate their hate-speech detection behavior.

Experimental Setup. To simulate partisan perspectives, we adopt the political distributions introduced by Bernardelle et al. (2025a,b), which map persona-conditioned LLMs across the PCT ideological space. Following the approach of Civelli et al. (2025a,b), we select the 20 most left-leaning and 20 most right-leaning persona descriptions from each model distribution, yielding 40 personas in total per model. Each persona is then used as an in-context instruction to condition the LLMs for hate speech detection on the unified dataset (refer to Appendix A for more information on the prompt template).

To improve reliability and maintain consistent output formats, we adopted a structured output generation strategy throughout our experiments.[2] This approach constrains model generations to follow a predefined schema, thereby helping to prevent ill-formed or off-task completions. At inference time, schema adherence is enforced through dynamic vocabulary masking: at each decoding step, only tokens that keep the partial output consistent with the schema remain available for selection. This ensures that final outputs are both syntactically valid and semantically aligned with the intended task requirements. This design mitigated refusal behaviors and promoted consistent formatting across different models, addressing reproducibility issues observed in earlier work (Azzopardi and Moshfeghi, 2024; Röttger et al., 2024).

For every target group, we compare the classification behavior of left- and right-oriented personas, measuring the average proportion of instances labeled as hate speech. Given 25,000 statements and 40 personas, this amounts to a total of 1,000,000 inferences per model. This setup allows us to test whether persona-induced alignment reproduces the partisan effects previously observed in pretraining-based studies.

Computational resources. All experimental conditions were executed on a single H100 GPU. Each run required approximately one and a half hours to complete, resulting in a total runtime of roughly 5 hours across the full set of experiments.

6.2. Results

Table 3 reports the average proportion of hate speech detected across target groups for left- and right-oriented personas, along with odds ratios quantifying systematic differences.

Overall persona effects. Across all three LLMs, left-oriented personas consistently yield higher detection rates than right-oriented ones. This effect is evident not only for minority groups such as BLACKS, MUSLIMS, JEWS, and LGBTQ+, but also for majority groups including CHRISTIANS, MEN, and WHITES. The uniformity of this effect runs counter to our hypothesis that right-leaning personas would be more protective of majority groups. Instead, persona-conditioning on the left systematically raises sensitivity to hateful content, indicating a general tightening of classification thresholds rather than selective group protection. One possible explanation is that the relatively small parameter size of these models limits their ability to adopt nuanced persona

[2]We used the implementation from vLLM, though other toolkits offer equivalent functionality.

perspectives, producing broad shifts in classification behavior rather than the group-specific differences anticipated—a pattern also noted by Civelli et al. (2025a). Investigating the precise mechanism behind this asymmetry lies beyond the scope of the present study, but future work could leverage our framework with larger-scale models to test whether the hypothesized group-specific protection emerges under more expressive architectures.

Variation across models. Although the gap between left and right is consistent, its magnitude differs by model family. Mistral shows the strongest divergence (across all target groups overall OR = 1.364, $p < 0.001$). By contrast, Llama exhibits the smallest persona effect (OR = 1.068, $p < 0.001$), while Qwen falls in between (OR = 1.073, $p < 0.001$). When considering absolute protection levels, however, a different pattern emerges: averaging overall detection rates across left- and right-conditioned personas, Llama achieves the highest baseline detection (0.493), followed by Mistral (0.359), and Qwen the lowest (0.199). While odds ratios primarily capture the models' relative responsiveness to persona-conditioning, raw detection levels reflect their intrinsic tendency to classify statements as hateful.

Model-specific protection of Whites. Across models, detection rates are generally highest for statements targeting BLACKS, making this category the most consistently protected. The main exception arises with Llama, where WHITES receive the strongest protection (0.656 under left personas), surpassing BLACKS as the top category. This unusually high value—the largest single entry in the table—may reflect the model's U.S.-centric training distribution, where discourse around race often centers explicitly on contrasts involving WHITES. By contrast, Qwen and Mistral exhibit substantially lower absolute detection for WHITES, aligning more closely with the overall trend that prioritizes minority-group protection.

Remarks. Together, the results convey two concise points. First, persona-conditioning produces an asymmetric within-model effect: left-conditioned personas yield higher hate-speech detection rates relative to right-conditioned personas, consistent with a general tightening of classification thresholds rather than selective protection of particular groups. Second, the effect's magnitude and the models' baseline tendencies differ: Mistral is the most responsive to persona shifts, Llama shows a higher baseline detection with notable outliers, and Qwen is comparatively stable.

7. Conclusion

This paper introduces a two-step framework for the systematic evaluation of perspective alignment in LLMs. First, we present SUBDATA, an open-source library that standardizes heterogeneous datasets by unifying annotation schemes and demographic taxonomies, thereby enabling consistent evaluation across subjective NLP tasks. Second, we propose a theory-driven evaluation approach that leverages these standardized datasets to test hypotheses about how differently aligned models behave in downstream applications. We demonstrate the practical value of this framework through an experimental use case. This example illustrates how SUBDATA not only provides a resource for data integration but also facilitates rigorous, theory-grounded experimentation on LLMs perspective alignment.

Future Extensions. The most immediate extension of SUBDATA 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.

Moreover, 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. For misinformation datasets, the connection between theory and testable hypotheses will be grounded in the topical domain of the claims being evaluated. Specifically, different domains (e.g., politics, public health, or climate) are known to elicit systematically different judgments depending on individuals' prior beliefs and ideological commitments. This allows misinformation to be operationalized as a subjective construct, where disagreement is not merely noise but reflects underlying differences in perspective. Consequently, variation in model predictions across perspectives can be interpreted as meaningful evidence of alignment (or misalignment) with distinct human viewpoints.

Ultimately, we intend to develop an alternative approach for evaluating LLM 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, this narrow scope reflects the availability of suitable datasets. We chose to release the library early because alignment research is advancing rapidly but lacks standardized resources for downstream evaluation. Even in its current state, we believe SubData offers immediate value for studying LLM alignment with diverse perspectives.

Our unified taxonomy required pragmatic mapping choices that inevitably involve subjective judgment. Challenges 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), targets that are placed into different categories in different original datasets (e.g., "mexicans" either put into a race—latin—or an origin category) and intersectional groups (e.g., "blacks, women"). We applied our principles carefully to balance specificity and generalizability. While the mapping process is manual and limits scalability, we argue this effort is both necessary and valuable: meaningful taxonomies for subjective constructs require domain expertise and contextual sensitivity that automated methods often miss. It is also a one-time investment with lasting benefits, and our taxonomy already aligns with independent efforts, suggesting emerging consensus. Future versions may incorporate semi-automated clustering or embedding-based methods to propose candidate mappings, with human oversight ensuring contextual validity. At the same time, SubData supports customization—researchers can adapt, extend, or redefine taxonomies as needed—helping to mitigate the limitations of any single framework.

In our experimental setup, we are contrasting personas from the extremes of the ideological spectrum. While we made this decision deliberately in order to increase contrast and establish the functioning of our methodological contribution, future work should extend this analysis to intermediate and mixed identities, where distinctions are more subtle and potentially more informative.

Finally, the library inherits annotation errors and biases from its source datasets. SubData aggregates existing annotations without re-labeling or quality control, so we encourage users to verify annotation quality and consult original documentation where appropriate.

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.

Dataset Licensing and Access. SubData does not host, mirror, or redistribute any dataset included in its unified taxonomy. All data are obtained directly from their official public distribution endpoints (e.g., GitHub, Kaggle, project websites), and the library simply provides convenience functions that download or load these datasets using the user's own credentials when required. For datasets that mandate registration, authentication, or explicit acceptance of license terms, SubData does not bypass these access controls; users must obtain the data themselves under the original license conditions. Because the library only performs local standard ization of datasets that users already lawfully acquired, it does not create or distribute any derivative dataset, and all licensing obligations remain governed by the original providers. The licenses, access constraints, and redistribution permissions of the datasets available via SubData are detailed in Appendix (Table 4).

Acknowledgements

This work is partially supported by an Australian Research Council (ARC) Future Fellowship Project (Grant No. FT240100022) and by the Swiss National Science Foundation (SNSF) under contract number CRSII5_205975.

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 Tanu Mitra. 2025. Robustness and confounders in the demographic alignment of LLMs with human perceptions of offensiveness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22025–22047, Vienna, Austria. Association for Computational Linguistics.

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.

Leif Azzopardi and Yashar Moshfeghi. 2024. Prism: a methodology for auditing biases in large language models. *arXiv preprint arXiv:2410.18906*.

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, St. Julian's, Malta. Association for Computational Linguistics.

Pietro Bernardelle, Stefano Civelli, Leon Fröhling, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2025a. Political ideology shifts in large language models. *arXiv preprint arXiv:2508.16013*.

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2025b. Mapping and influencing the political ideology of large language models using synthetic personas. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 864–867.

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, Online. Association for Computational Linguistics.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How susceptible are large language models to ideological manipulation? pages 17140–17161.

Stefano Civelli, Pietro Bernardelle, and Gianluca Demartini. 2025a. The impact of persona-based political perspectives on hateful content detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1963–1968.

Stefano Civelli, Pietro Bernardelle, Nardiena A Pratama, and Gianluca Demartini. 2025b. Ideology-based llms for content moderation. *arXiv preprint arXiv:2510.25805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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, Online. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2025. Personas with attitudes: Controlling LLMs for diverse data annotation. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 468–481, Vienna, Austria. Association for Computational Linguistics.

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, Mexico City, Mexico. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17001–17019, Miami, Florida, USA. Association for Computational Linguistics.

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

Jigsaw. 2019. 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, Bangkok, Thailand. Association for Computational Linguistics.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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, Online. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jens Rupprecht, Leon Fröhling, Claudia Wagner, and Markus Strohmaier. 2025. German general personas: A survey-derived persona prompt collection for population-aligned llm studies. *arXiv preprint arXiv:2511.21722*.

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, Marseille, France. European Language Resources Association.

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. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

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.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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*.

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, Miami, Florida, USA. Association for Computational Linguistics.

A. Prompt Template

Prompt Template

You are an expert content analyst specializing in identifying hate speech according to specific guidelines. Your task is to analyze the provided text strictly from the perspective of **[PERSONA]**.

Based on your analysis, you must generate a single, valid JSON object that adheres to the following schema and logic. **Do not output any text other than the JSON object.**

Analysis Logic:

1. **Perspective:** Embody the given persona. Your entire analysis must reflect their likely values, biases, worldview, and sensitivity level to hateful content.
2. **Classification:** Determine if the persona would consider the text to be hate speech. This determines the value of `is_hate_speech`.

JSON Schema:

- `is_hate_speech` (*string*): Must be one of `"true"` or `"false"`.

Example 1 (Hateful):

```
{
    "is_hate_speech": "true"
}
```

Example 2 (Not Hateful):

```
{
    "is_hate_speech": "false"
}
```

Persona: **[PERSONA]**
Text: **[TEXT]**

Now, analyze the text from the persona's perspective and generate the JSON object.

Across all LLMs investigated, we employed a standardized prompt format to ensure comparability of results and to minimize confounding effects introduced by variations in wording or structure. We wrapped all inputs in the respective chat template of each model family. This preserved the conversational format they were optimized for during fine-tuning, reducing the risk of degraded performance or altered behavior due to mismatched prompting styles. For each combination of persona and statement (40 personas × 25,000 statements = 1,000,000 total pairs), we generated the corresponding prompt by substituting `[TEXT]` and `[PERSONA]` placeholders in the template with the actual text.

Dataset	License	Gated?	Redistribution Allowed?	Sourcing Method
Fanton et al. (2021)	Research-use only (not for redistribution)	×	×	Download via authors' GitHub (CONAN)
Hartvigsen et al. (2022)	Not explicitly stated	✓	×	Download via Hugging Face (requires form + token)
Jigsaw (2019)	CC0 1.0	×	✓	Download from Kaggle (Jigsaw Unintended Bias)
Jikeli et al. (2023a)	CC BY 4.0	×	✓	Download from Zenodo (open DOI)
Jikeli et al. (2023b)	CC BY 4.0	×	✓	Download from Zenodo (open DOI)
Mathew et al. (2021)	MIT	×	✓	Download via GitHub (HateXplain)
Röttger et al. (2021)	CC BY 4.0	×	✓	Download via GitHub (hatecheck-data)
Sachdeva et al. (2022)	CC BY 4.0	×	✓	Download via Hugging Face (open dataset)
Vidgen et al. (2021a)	CC BY 4.0	×	✓	Download via GitHub (Learning from the worst)
Vidgen et al. (2021b)	CC BY 4.0	×	✓	Download from Zenodo (open DOI)

Table 4: Licensing, access constraints, and redistribution permissions for datasets included in SUBDATA.