

Bye Bye PERSPECTIVE API: Lessons for Building and Governing Measurement Infrastructure

David Hartmann^{1,2}, Manuel Tonneau^{1,3}, Angelie Kraft¹, LK Seiling¹, Dimitri Staufer²,
Pieter Delobelle^{4,5}, Jan Fillies⁶, Anna Ricarda Luther^{7,8},
Jan Batzner^{1,9,10,†}, Mareike Lisker^{11,12,†}

¹Weizenbaum Institute, ²TU Berlin, ³University of Oxford, ⁴KU Leuven, ⁵Pleias,
⁶Freie Universität Berlin, ⁷University of Bremen, ⁸ifib research, ⁹TU Munich,
¹⁰Munich Center for Machine Learning, ¹¹HTW Berlin, ¹²University of Hamburg

Correspondence: d.hartmann@tu-berlin.de

Abstract

PERSPECTIVE API closes at the end of 2026, removing the de facto standard for toxicity measurement and exposing researchers’ dependence on a tool they did not control. Drawing on this case, we argue that a research field must build and govern its own measurement infrastructure rather than borrow it. Surveying 241 papers that use or study PERSPECTIVE, we show what depending on it cost the research community: claims reaching past what the tool could support, results that shifted when its model was silently retrained, and disparities researchers could measure but not explain. These failures were amplified throughout the LLM lifecycle, where PERSPECTIVE supplied the labels, filtered the corpora, and graded the systems trained on each, rewarding errors rather than catching them. To keep the instrument open to study after shutdown, we release PERSPECTIVE scores for 5.9 million text snippets from 77 datasets. We further specify ten requirements for measurement infrastructure a field owns, and argue that what blocks such infrastructure is not technical capability but the value the field places on infrastructure work.

1 Introduction

Online abuse can have profound real-world consequences, including the suppression of marginalised voices, deterioration of mental health, and violence against marginalised groups (Marques, 2023; Luther et al., 2025). Research responded with automated classifiers that detect such content at scale (Vidgen and Derczynski, 2020). One became particularly influential: PERSPECTIVE API, developed by Google Jigsaw. Since 2017, it has offered a free, scalable API for classifying *toxicity*, an umbrella term for harmful content, and is now widely described as “a foundation for academic research on online abuse and incivility” (Nogara et al., 2025).

Despite its popularity, Jigsaw recently announced that PERSPECTIVE API will cease operation at the end of 2026, because, in its words, “As AI capabilities evolve, we are setting our sights on new challenges” (Jigsaw, 2026). For communities that built their work around it, the shutdown suddenly withdraws critical infrastructure, exposing how far the field had come to depend on a proprietary, externally controlled measurement tool.

Against that dependence, this position paper argues from the example of PERSPECTIVE API that a field must build and govern the measurement infrastructure it relies on. We analyse a decade of PERSPECTIVE API and what relying on it cost the research community: findings that outran what the tool was validated for, results that could not be compared over time, and biases that could not be explained. From this, we draw lessons for measurement infrastructure a field owns, from how its construct is defined to how the work is funded.

We make four contributions. *First*, we present a systematic bibliographic review of 241 papers that use or study PERSPECTIVE, tracing how it was used across hate speech detection, computational social science (CSS), and LLM development and evaluation, and what drove its adoption (§2). *Second*, judged against standard procedures of measurement validation, we show PERSPECTIVE was never fit for the central role it came to play (§3), and that these failures compound when the same tool filters pretraining data and grades the models built from it (§4). *Third*, from those failures, we derive five technical and five governance requirements (T1–T5, G1–G5), most of which apply to any measurement infrastructure, and assess existing alternatives against them (§5). *Fourth*, we release PERSPECTIVE scores for 77 datasets covering 5.9 million text snippets, keeping the instrument open to study after shutdown (§5). We close by asking why no alternative was built, and what would have to change for the field to build one (§6).

[†]These authors contributed equally.

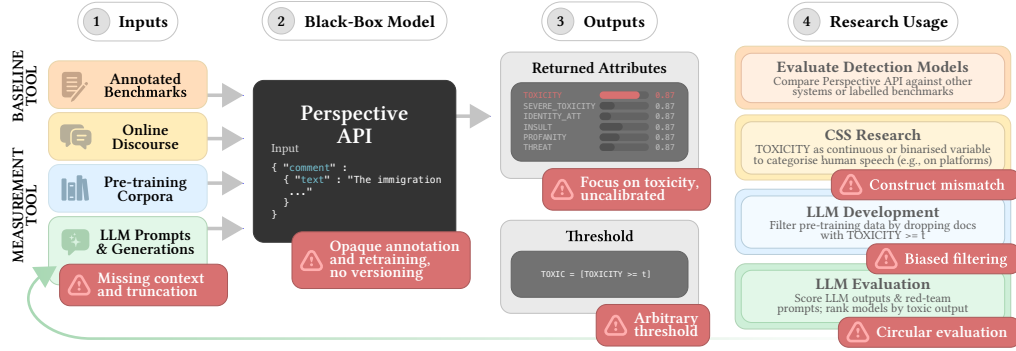


Figure 1: **How PERSPECTIVE API is used in research, as a *subject of evaluation* and as a *measurement tool*.** The figure shows choices in operationalisation and provision that are critiqued in this paper: ① **Inputs** are scored as isolated sentences, stripping sender, target, and conversational context. ② The **Black-Box Model** is opaque about model updates and annotations. ③ **Outputs** include six attributes, but downstream work retains only TOXICITY and binarises its uncalibrated score at an ad hoc threshold. ④ **Research Usage** covers evaluating detection models, CSS research, LLM development, and LLM evaluation, with issues such as construct mismatch, biased filtering, and circular evaluation. §3 and §4 develop each point and its consequences for research quality.

2 PERSPECTIVE API and Its Research Uptake

2.1 Use in Research

Within academic research, PERSPECTIVE was taken up across communities in two distinct ways (Figure 1): as a *subject of evaluation* and as a *measurement tool*. To measure this uptake, we conducted a systematic bibliographic search of research using PERSPECTIVE (see Appendix B for details). We found 241 papers, 18 of which are meta-critiques, and labelled the remaining 223 by usage type: baseline model, CSS measurement, pre-training filtering, or LLM evaluation.

PERSPECTIVE as a baseline model In *toxicity* and *hate speech detection* research, PERSPECTIVE was primarily treated as a system to be evaluated ($n = 83$). Pavlopoulos et al. (2019), for example, assess PERSPECTIVE as a baseline for the SemEval-2019 offensive language task. Evaluations of this kind measure agreement with a benchmark’s labels. Where the benchmark encodes a different construct, such as offensive language or hate speech, the result shows how closely PERSPECTIVE tracks that construct, not whether toxicity was measured well. Where the benchmark encodes toxicity, the labels usually descend from the same annotation scheme that produced PERSPECTIVE’s training data, so agreement demonstrates fidelity to that scheme rather than support for the construct the scheme encodes. Neither result establishes the validity of Jigsaw’s toxicity construct.

PERSPECTIVE as a measurement tool In contrast, CSS and LLM development and evaluation research primarily used PERSPECTIVE API as a tool to measure toxicity at scale. In CSS, researchers applied the API in large-scale studies of online discourse ($n = 50$, e.g., Theocharis et al., 2020; Kim et al., 2021; Chang et al., 2023; Frimer et al., 2023; Mosleh et al., 2024; Gervais et al., 2025). In LLM development, PERSPECTIVE was used to filter pre-training data by removing text flagged as toxic from the corpora used to train and align LLMs ($n = 29$, e.g., Welbl et al., 2021). In LLM evaluation (benchmarking, auditing, and red teaming), PERSPECTIVE was built into evaluation operationalisations to label model outputs and, in red teaming, to flag potentially offensive inputs ($n = 61$, e.g., Gehman et al., 2020; Liu et al., 2021; Lee et al., 2023; Liang et al., 2023). Across these three uses, PERSPECTIVE is not the object of study but the instrument that produces the finding. Whatever the tool captures or misses therefore passes into the published result, whether that is a claim about discourse or a safety score.

2.2 Drivers of Adoption

At its peak, PERSPECTIVE processed 500 million requests daily (Jigsaw, 2021), and was integrated into comment systems at major publishers, including the New York Times, Vox Media, and OpenWeb (Jigsaw, 2021; Rieder and Skop, 2021), while also becoming widely adopted in academic research (Horta Ribeiro, 2024). In what follows, we examine the factors behind this widespread adoption.

Timing PERSPECTIVE's launch in 2017 followed a surge in online abuse, including high-profile harassment campaigns (Massanari, 2017), the spread of conspiratorial narratives (Venturini, 2019), inter-ethnic violence and genocide (Mozur, 2018), and the mobilisation of political extremism online (Ganesh and Bright, 2020). Regulators responded: Germany's *Network Enforcement Act* (NetzDG), passed in 2017, required platforms to remove clearly illegal content. This moved the burden of moderation from users to platforms, which turned to centralised and increasingly automated systems (Gillespie, 2018; Gorwa, 2019; Gorwa et al., 2020). It also motivated academic research on automatic hate speech detection (Waseem and Hovy, 2016). PERSPECTIVE thus emerged at the intersection of regulatory pressure and a nascent research field, offering scale where human moderation could not (Davidson et al., 2017; Gorwa et al., 2020).

Cost and access PERSPECTIVE API was free throughout its decade of operation, whereas competing tools such as Amazon's RECOGNITION charge per image analysed (Amazon Web Services, 2026). Access was provided upon application to Jigsaw, though acceptance criteria remained opaque (Rieder and Skop, 2021). The API offered a 100ms response time and an expandable quota on request, making large-scale studies feasible at no cost. Its coverage of 18 languages, trained on 38 million records (Lees et al., 2022), further enabled research beyond English (Tonneau et al., 2024).

Ease of use, transparency, and performance PERSPECTIVE accepted single-sentence inputs without preprocessing and returned an interpretable scalar score, lowering the barrier to integration in research and production alike. Jigsaw provided model cards, threshold guidance, and partial architectural documentation (Jigsaw, 2022), a level of transparency uncommon among commercial APIs (Hartmann et al., 2025b). The team also engaged the research community through publications, benchmark datasets (Dixon et al., 2018; Lees et al., 2022), and responses to researcher feedback (Jigsaw, 2023). Though incomplete, as the next section details, this documentation positioned PERSPECTIVE API as a comparatively transparent and research-embedded proprietary tool. PERSPECTIVE also outperformed academic hate speech models in real-world evaluations, despite performing poorly in absolute terms (Tonneau et al., 2025).

Institutional credibility Adoption was also driven by institutional backing. PERSPECTIVE came from Jigsaw, a Google subsidiary, and grew out of a collaboration with Wikimedia (Wulczyn et al., 2017), two institutions whose standing lent credibility to the tool. Jigsaw also supplied the corpora and public competitions the field then built on (Jigsaw and Conversation AI, 2021). External validation followed: the New York Times, itself a recognised institution, adopted a moderation tool built with Jigsaw (Google, 2018). This institutional weight is what made PERSPECTIVE credible before it was validated, and it is likely why adoption continued despite the biases documented in §3.

Path dependence and network effects As Star (1999) argues, tools become infrastructure through their embeddedness in practice. Once taken up across a field, a measurement instrument shapes what is measurable, comparable, and credible, generating path dependence (Bowker and Star, 1999). PERSPECTIVE API exemplifies this dynamic: its integration into benchmarking suites such as *RealToxicityPrompts* (Gehman et al., 2020) and the *Holistic Evaluation of Language Models* (HELM) (Liang et al., 2023), as well as its use in CSS studies of political incivility and online harassment (Theocharis et al., 2020; Gervais et al., 2025), positioned it as a common reference point across research contexts. This standardisation enabled comparability across studies, which in turn encouraged continued use, creating a self-reinforcing dynamic consistent with network effects (Katz and Shapiro, 1994). In this sense, PERSPECTIVE API was not necessarily adopted because it was the most valid or best-performing tool, but because it had become an established point of reference, effectively positioning it over time as the de facto standard for toxicity detection.

3 Was PERSPECTIVE API Ever Fit for Purpose? A Measure Development and Infrastructural Critique

Becoming a standard through path dependence and network effects is not the same as being validated as one. In this section, we assess whether PERSPECTIVE API was fit for the central role it came to play in research by comparing its development to established procedures of measurement validation in the social sciences (Clark and Watson, 2019; Jacobs and Wallach, 2021; Wallach et al., 2025).

In this tradition, a *concept* (i.e., a broad and often contested notion) is first systematised into a *construct* (i.e., a more specific and explicitly defined understanding), specifying *what* is to be measured (Adcock and Collier, 2001). The construct is then operationalised through a *measurement instrument*, specifying *how* it is to be measured. Both steps, **systematisation** and **operationalisation**, are inherently situated and involve analytical choices (Wallach et al., 2025). They rely on assumptions that must be made explicit and supported theoretically and empirically.¹

We show that PERSPECTIVE provides limited grounding for key assumptions in its systematisation (§3.1) and operationalisation (§3.2), and that these limitations affect performance and research use (§3.3). We further identify structural issues in its application, particularly in cases where it is used to both define and measure toxicity (§4).

3.1 Systematisation

Contested behavioural assumption PERSPECTIVE’s construct of toxicity rests on a behavioural assumption that is empirically contested. Jigsaw defines toxicity as “a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion” (Dixon et al., 2018; Rieder and Skop, 2021), thereby tying toxicity to disengagement. However, empirical evidence suggests that this effect is not stable. For example, Beknazar-Yuzbashev et al. (2025) show that removing toxic content from social media feeds causally reduces time spent, ad impressions, and content consumption, indicating that its presence contributes to user engagement. Observationally, out-group animosity is the strongest predictor of shares and retweets on Facebook and Twitter (Rathje et al., 2021), and toxic conversations on Twitter tend to generate larger, wider, and deeper reply trees than non-toxic ones (Saveski et al., 2021). Taken together, this evidence challenges the disengagement premise: the same toxic content can repel some readers and engage others, depending on their relationship to it. A construct that ties toxicity to a single, universal behavioural effect therefore mistakes a reader-dependent phenomenon for a property of the text.

¹This is commonly referred to as *construct validation*, the process of establishing whether a measurement instrument captures what it is intended to measure or, in other words, of providing theoretical and evidential support for one’s operationalisation (Cronbach and Meehl, 1955).

Insufficiently specified construct Even beyond this contested behavioural assumption, the construct remains insufficiently specified. PERSPECTIVE API provides scores for multiple “toxic attributes”, including threat, insult, identity attack, severe toxicity, and profanity, suggesting a decomposition of toxicity into subcomponents. However, Jigsaw gives only one-sentence definitions and does not distinguish between the attributes, relate them to the overarching construct, or say whether the set is exhaustive (Rieder and Skop, 2021; Hartmann et al., 2025b). Both the scope of the construct and its relationship to these subcomponents therefore remain unclear, making it difficult to determine what is being measured or to assess construct validity (Wallach et al., 2025).

Construct-measure mismatch This lack of specification enables a mismatch between the intended construct and its use in downstream research, particularly in studies of hate speech. Toxicity and hate speech are related but distinct (Barth et al., 2023): toxicity covers rude or unreasonable language generally, while hate speech targets individuals or groups on protected characteristics (Anderson and Barnes, 2023; Tonneau et al., 2024). A measurement instrument designed for toxicity therefore cannot be assumed to validly capture hate speech, a conflation common across the field’s datasets and literature (Vidgen and Derczynski, 2020). In this regard, we observe three recurring usage patterns in the literature. The first uses PERSPECTIVE’s toxicity score to measure toxic language, aligning with the tool’s intended use, though constrained by its behavioural definition. The second uses attributes such as IDENTITY_ATTACK as proxies for hate speech, typically acknowledging the construct gap explicitly (e.g., Tonneau et al., 2025). The third frames research questions in terms of hate speech, political incivility, or LLM safety and operationalises them using the toxicity score without acknowledging the lack of theoretical alignment (Kim et al., 2021; Frimer et al., 2023). This last pattern is the most problematic: the instrument does not match the target construct and the mismatch goes unstated, a practice Gervais et al. (2025) document for political incivility research.

Lesson PERSPECTIVE tied toxicity to a contested behavioural effect and never fixed the scope of its attributes, leaving a construct too vague to validate and loose enough to absorb hate speech, incivility, and LLM safety alike. Any successor should state its construct explicitly and say for whom it holds (T2, G3).

3.2 Operationalisation and Provision

We now turn to how toxicity is operationalised and provisioned in PERSPECTIVE API. We focus on three aspects: the input representation, the model and its provision, and the resulting outputs.

Input: context-free PERSPECTIVE API takes a single sentence or short text as input and returns a toxicity score, thereby operationalising toxicity as a property identifiable in isolated text (Rieder and Skop, 2021; Hartmann et al., 2025b), as most classifiers of the period did. However, the meaning of a sentence depends on its conversational context, community norms, who is speaking, who is targeted, and the relationship between speakers and receivers (Mosca et al., 2021; Yoder et al., 2022; Anderson and Barnes, 2023). By design, PERSPECTIVE API strips this away, limiting the instrument to surface features and excluding the relational dimensions that define toxicity as a social phenomenon. The same text therefore receives the same score across contexts, even though its perceived toxicity depends on context-specific norms.

Model: misaligned and opaque The annotation scheme, while partially available (Conversation AI, 2017), diverges from Jigsaw’s public definition of toxicity (Dixon et al., 2018; Rieder and Skop, 2021). Whereas the public definition frames toxicity as a conjunctive criterion, namely a rude or disrespectful comment that is likely to make users leave a discussion, the annotation scheme applies these criteria disjunctively, allowing labels to be assigned based on rudeness or anticipated behavioural effects alone. As a result, the model is trained on labels that do not correspond to the definition users are given, creating a gap between construct and operationalisation. Opacity compounds the mismatch at the level of data, model, and input. The annotation process went undocumented, including annotator demographics, data sources, and labelling procedures (Rieder and Skop, 2021; Hartmann et al., 2025b). The model was silently retrained, an update warranted by evolving slurs and coded language (Siegel, 2020; Mendelsohn et al., 2023) but made without version pinning or changelogs, and announced only in a group Jigsaw said could not cover every update (Pozzobon et al., 2023). Finally, the system scored only the first 501 characters of each input, a limitation the published documentation omits (Talas and Hutchings, 2023).

Output: uncalibrated The outputs of PERSPECTIVE API are not calibrated to any interpretable real-world quantity. The API returns a score between 0 and 1, but this score does not correspond to prevalence, severity, or annotator agreement, as is true of scalar classifiers generally (Guo et al., 2017). According to Jigsaw’s technical documentation, scores are normalised against a balanced 50/50 toxic and non-toxic dataset rather than any real-world distribution (Jigsaw, 2017), meaning they reflect model confidence under an assumed even split (Nogara et al., 2025) rather than the prevalence of harm. Threshold guidance existed (Hua et al., 2020a,b), but those thresholds were validated for particular corpora and tasks, and in practice, the choice remained an undocumented decision made study by study. As a result, scores lack a stable interpretation and depend on arbitrary and often unreported thresholding choices.

Lesson PERSPECTIVE stripped context from its inputs, trained on labels that diverged from its published definition, retrained silently, and returned a score normalised to a distribution no real corpus resembles. Any successor should document and version every such design decision, and accept the context its construct depends on (T1, T3, T4, G2).

3.3 Consequences for Research Quality

As PERSPECTIVE API became embedded in research workflows, the limitations above produced systematic errors, inconsistent results, and biases across use cases.

Lacking validity and performance Construct validity issues translate into poor classification performance when PERSPECTIVE is applied beyond its intended scope. Gervais et al. (2025) show that PERSPECTIVE scores on congressional tweets achieve a maximum F1 of 0.33 against human-coded political incivility across all decision thresholds, and Mihaljević and Steffen (2022) find that it detects antisemitism only at a very basic level. Failures are most pronounced where interpretation depends on context: counterspeech is labelled hateful (Gligorić et al., 2024), reclaimed slurs are overmoderated (Dias Oliva et al., 2021; Röttger et al., 2021), and implicit hate speech, which relies on background knowledge, is systematically missed (Hartmann et al., 2025b). Finally, in real-world settings, PERSPECTIVE performs very poorly for hate speech detection but still outperforms academic models, pointing to a broader model miscalibration in the field (Tonneau et al., 2025).

Undermined reliability and reproducibility

Silent retraining made results from different times incomparable. Pozzobon et al. (2023) show that rescored *RealToxicityPrompts* with a later version of the API moved around 10K prompts from toxic to non-toxic and 232 in the opposite direction, a 49% net reduction in prompts labelled toxic. The same rescored produced a different ranking of the 37 models in the HELM toxicity benchmark. The shutdown makes this problem permanent: scores can no longer be recovered or verified, rendering PERSPECTIVE API-based benchmarks irreproducible.

Systematic bias and limited attribution Bias is evident in PERSPECTIVE’s outputs, but its sources cannot be attributed, given limited transparency about the data and annotation. In Jigsaw’s own early evaluations, benign sentences such as “*I am a gay man*” scored as toxic purely for naming an identity (Dixon et al., 2018). That Jigsaw published this failure and then ran a public competition on it (Jigsaw and Conversation AI, 2021) is rare among commercial providers. Later work broadened this: LGBTQIA+ content and reclaimed slurs are disproportionately flagged (Dias Oliva et al., 2021; Röttger et al., 2021), as is African-American English, a pattern documented for academic classifiers as well (Sap et al., 2019; Xu et al., 2021), while hate against marginalised groups is underdetected (Hartmann et al., 2025b). These patterns suggest reliance on surface features such as identity terms and dialect markers, combined with context-free inputs. Biases extend across languages and geographies: German text is rated more toxic than comparable content elsewhere (Nogara et al., 2025), and PERSPECTIVE scores an abusive Hindi word at 0.10 while the same insult in English scores 0.97 (Ghosh et al., 2021). Annotator demographics, beliefs, and guidelines shape what is labelled toxic in edge cases (Sap et al., 2022; Davani et al., 2023), and because PERSPECTIVE documents none of them (§3.2), the disparities above are measurable but not explicable: whether dialect is over-flagged because of who annotated, what they were told to do, or which text they saw cannot be established from outside.

Lesson PERSPECTIVE’s users published findings beyond its validated scope, could not compare results across time, and could not attribute the biases they observed. Any successor should validate performance within a stated scope, tie every result to a version, and document annotation well enough that disparities can be traced (T1, T2, G2).

4 Amplification and Circularity across the LLM Lifecycle

These issues extend beyond individual studies. Because PERSPECTIVE API is embedded at both ends of the LLM lifecycle, its errors propagate into the behaviour and standards of widely deployed LLMs.

Biased data filter A common step in LLM pre-training (§2) is to score the corpus with PERSPECTIVE and drop what it flags. Google and DeepMind report doing this (Welbl et al., 2021; Longpre et al., 2024), and Cohere a related pipeline (Ngo et al., 2021). The filtering happens before any model exists, so whatever the classifier misjudges is simply absent from what the model can learn. The downstream cost is not borne equally. As shown in §3.3, PERSPECTIVE systematically over-flags African-American English, LGBTQIA+ content, and reclaimed slurs. Filtering with it therefore removes these categories disproportionately from training corpora (Xu et al., 2021), leaving models worse at understanding these communities and at producing the speech and values their language carries. This compounds PERSPECTIVE’s silencing effect: communities whose speech the classifier already over-moderates may now also be under-served by the LLMs trained on its filtered output. The same tool then scores those models, so the check cannot see what the filter removed.

Circular evaluation Beyond filtering, PERSPECTIVE API also generated the training labels and served as the evaluation standard for classifiers built on them, producing a self-reinforcing circularity that runs through the field’s most influential resources, including *RealToxicityPrompts* (Gehman et al., 2020; Pozzobon et al., 2023). The loop has three recurring steps. *First*, researchers used PERSPECTIVE scores to generate training labels for toxicity and hate speech datasets, and for controllable-generation methods such as DExperts (Gehman et al., 2020; Liu et al., 2021). *Second*, models trained on these labels were evaluated against PERSPECTIVE outputs, directly or through benchmarks it had labelled (Gehman et al., 2020; Liang et al., 2023; Pozzobon et al., 2023). *Third*, PERSPECTIVE then graded the models built from its own labels (Liu et al., 2021; Bin Zia et al., 2022). At each step, ad hoc thresholds such as $t = 0.7$, often unreported (see stage ③ in Figure 1), turned uncalibrated scores into binary ground truth (Pozzobon et al., 2023; Hartmann et al., 2025b).

Across both filtering and evaluation, the field’s working definition of toxicity collapsed into a single system’s decision threshold. Because the same tool shapes both what models learn and what they are graded on, its systematic errors become invisible at both ends: models that inherit PERSPECTIVE’s biases score well on PERSPECTIVE-derived benchmarks because the errors are in the standard itself. The result is amplification rather than correction, since an error introduced when the corpus is filtered is confirmed when the model is graded. This is what [Blodgett et al. \(2020\)](#) describe as undeclared social assumptions laundered through technical framing until they appear neutral. Once PERSPECTIVE became the measure, it ceased to function as one ([Strathern, 1997](#)).

Lesson When one tool filters the training data and grades the models built from it, the field’s working definition of toxicity collapses into that tool’s decision threshold, and its errors become invisible at both ends. An instrument must be independent of what it evaluates (G1).

5 What Now? Towards Independent Research Infrastructure

PERSPECTIVE API emerged when platforms were investing in moderation under regulatory pressure (§2.2). That investment has since reversed. Since 2024, platforms have reduced moderation workforces ([McGuirk, 2024](#); [Gerken, 2025](#); [McMahon et al., 2025](#)), loosened content rules ([Biddle, 2025](#)), curtailed researcher data access ([Burnat and Davidson, 2026](#); [Hartmann et al., 2025a](#)), and dismantled trust and safety teams ([Moran et al., 2025](#)), withdrawing from earlier moderation commitments ([Magalhães et al., 2026](#)) even as toxic speech rises ([Hickey et al., 2025](#); [Törnberg and Chueri, 2025](#)) and remains implicated in offline violence ([CNN, 2025](#); [Euronews, 2025](#); [Venkataramakrishnan, 2025](#)). Regulation has expanded over the same period. Platforms in the EU now face the Digital Services Act ([European Parliament and Council of the European Union, 2022](#)) and the AI Act ([European Parliament and Council of the European Union, 2024](#)), and in the UK the Online Safety Act ([UK Parliament, 2023](#)). Each imposes duties that presuppose harmful content can be measured, since enforcement and compliance alike depend on it. Independent measurement capacity therefore matters more now than ever. We set out what the community should do before the shutdown, and the requirements that a successor must meet.

Immediate actions Before the shutdown on 31 December 2026, the community should archive PERSPECTIVE scores for all datasets and benchmarks, including full method documentation such as query date, threshold settings, and API version when available. These should be registered as versioned data artefacts. We contribute to this effort by releasing a snapshot of 77 publicly available toxicity datasets and major benchmarks, with 5.9 million scored text snippets ([Hartmann et al., 2026b](#)). The release keeps PERSPECTIVE open to study once the API is gone, and gives successor systems a fixed public baseline to compare against (Appendix C). These are short-term measures, not the successor itself. Until one exists, new projects should use open-weight models with clear versioning instead of closed endpoints, and any LLM used as a judge for toxicity detection should have its parameters and system prompts disclosed. The snapshot should serve only as a fixed reference for historical comparison. New evaluations should use human labels where possible, or calibrated open-weight classifiers reported alongside them.

Requirements and recommendations From the limitations in §3, we derive ten requirements for a successor infrastructure, grouped into **Technical** and **Governance** dimensions and mapped in Table 1 to the failures they address. Seven of the ten apply to any measurement infrastructure. Uncertainty (T4) and legitimacy (G3) also assume a construct that people disagree about, and only contextuality (T3) is specific to abusive language.

T1 • Reproducibility

Record the model, dataset, and metric behind every evaluation, and pin a citable version.

Versioning in a community-agreed schema ([Biderman et al., 2024](#); [Batzner et al., 2026](#)) is what makes a repeated query a repeated measurement, and what lets a result be rechecked years later.

T2 • Validity

State which construct the tool measures, under what conditions, and where that construct stops applying.

Without that boundary, a later user cannot tell when a research question has drifted beyond what the tool measures ([Jacobs and Wallach, 2021](#)).

T3 • Contextuality

Accept who is speaking, who is targeted, and prior context as scoring input.

Structured metadata makes counterspeech and reclaimed language separable from the attacks they resemble (Yu et al., 2022). Where it is missing, the tool should report its absence rather than score as though context were irrelevant.

T4 • Uncertainty

Make the score reflect how much people agree, not the model’s confidence.

Annotator disagreement is signal about a contested construct, not noise (Leonardelli et al., 2021; Uma et al., 2021). A score tied to that agreement makes a contested case visible, where one normalised to an artificial class balance does not.

T5 • Multilinguality

Validate and report performance separately for every supported language.

A single headline figure hides the disparities that matter most to the communities least well served, so per-language results need disclosure wherever performance falls below a defined threshold (Tonneau et al., 2025).

Existing classifiers do not meet these requirements DETOXIFY, the most used open alternative, trains on the three Jigsaw Kaggle challenges and so inherits Jigsaw’s construct (T2). It covers seven languages against PERSPECTIVE’s eighteen (T5), and its own documentation warns that profanity is scored toxic “regardless of the tone or the intent of the author”, with bias “towards already vulnerable minority groups” (Hanu and Unitary team, 2020). OpenAI’s moderation endpoint reproduces what created the original dependence: free at the point of use, a fixed thirteen-category taxonomy, and a single omni-moderation-latest alias behind which the model may change, meeting neither T1 nor T2 (OpenAI, 2026). Open-weight guard models such as LLAMA GUARD and SHIELDGEMMA improve on T3 by supplying the construct as a policy prompt at inference, but they move the choice of definition to whoever writes that prompt, and prompts are versioned no more carefully than models were (T1). No available classifier satisfies T1 and T2 together.

G1 • Independence

Keep the tool open-source, vendor-independent, and separate from what it evaluates.

The shutdown is the argument. Independence is not a preference for open tools but rather is what lets an instrument outlive the company that built it, and the community repair it when it breaks.

G2 • Transparency

Open model weights, training data, and annotation schemata to external audit.

Annotator demographics and beliefs shape what is labelled toxic in contested cases (Sap et al., 2022; Davani et al., 2023; Kraft et al., 2025): annotation is part of the instrument, and publishing it is what lets a disparity be traced rather than merely observed.

G3 • Legitimacy

Define toxicity together with affected communities, and preserve disagreement between them.

This should run through an open, documented process (Maronikolakis et al., 2022). Legitimacy does not require consensus. It requires that systematic disagreement, by annotator identity and target group, be documented and propagated rather than averaged away (Leonardelli et al., 2021; Sap et al., 2022; Davani et al., 2023).

G4 • Accountability

Audit continuously, and publish both what the audits find and what was changed in response.

Query-efficient methods now make external auditing tractable (Hartmann et al., 2026a), so what remains is a governance obstacle rather than a technical one. Accountability requires auditors independent of the maintainers, and a documented remediation for each finding.

G5 • Sustainability

Fund the infrastructure so that no single funder’s exit can end it.

PERSPECTIVE API was free at the point of use, which drove adoption and left no alternative when its funder exited. Free is not the same as sustainable: a successor needs funding spread across institutions, so that any one of them withdrawing slows maintenance rather than ending the service.

The governance templates exist Working instances of these requirements are in place elsewhere. EleutherAI’s *lm-evaluation-harness*, the widely used open-source evaluation framework that advanced the community standard, implements T1 directly: each evaluation task carries a version that contributors increment on any breaking change, recorded with date, pull request, and description (Gao et al., 2024). MLCommons is a template for G1, G4, and G5, being a 501(c)(6) non-profit whose AI Risk and Reliability working group spans industry, academia, and civil society, and which maintains AILUMINATE as a versioned public benchmark (Ghosh et al., 2025). Most directly, *Dynabench* began inside Meta and was transferred to MLCommons in 2022 rather than being shut down (Kiela et al., 2021), the migration PERSPECTIVE API never made. None of these are toxicity tools, but each shows that the governance a successor needs is already running somewhere.

Two constraints and a first step What none of these templates supplies is the measurement layer itself: versioned, per-language classifiers, and multiple construct-specific instruments under shared governance rather than one general toxicity tool. Two constraints bind its construction. The first is **privacy**. T3 asks a successor to score with sender, target, and conversation history, and G2 asks it to publish annotation schemata and annotator demographics, so both enlarge the personal data held about authors, targets, and annotators, on populations least able to absorb the risk. Our own release resolves a narrow version of this by publishing record identifiers and scores rather than the texts, and a successor faces the same choice at every layer. The second is **funding**. G5 asks for funding spread across institutions, but no durable model exists. Grant cycles are shorter than the maintenance horizon of infrastructure, and an instrument that must be revalidated as language changes cannot be funded as a one-off project. Both constraints are real, but the shutdown shows the cost of leaving them unsolved. As a first step, we call on ACL, ACM, and ICA to jointly establish a permanent working group covering versioned model oversight, annotation schema governance, and benchmark archiving, with a defined funding mechanism rather than merely institutional endorsement. Such a body could also take custody of what a provider retires, as MLCommons did for *Dynabench*, while PERSPECTIVE shows what is lost when none does.

A request to Jigsaw With that precedent in mind, we ask Google Jigsaw to release the per-language model versions, annotation guidelines, and remaining score distributions. Doing this at deprecation would cost very little, but if not, a decade of toxicity research will rely on a tool that can no longer be reviewed. Jigsaw could also support a successor with transitional funding or by handing the project to a group that can maintain it, rather than issuing a sunset notice alone.

6 Concluding Remarks

The closure of PERSPECTIVE API is an opportunity for researchers to be honest about how the field came to depend on a tool it did not control. The community adopted a corporate tool, incorporated it into benchmark suites, declared it a standard measure, and continued using it after its validity problems were known. Why no alternative was built is as much a question about our epistemic practices as our technical ones. This responsibility is not the community’s alone. Google Jigsaw built PERSPECTIVE API, encouraged its use in research, and is now withdrawing it. It should release the model weights, annotation schemata, and annotation documentation that made it function as infrastructure, where ownership permits. A decade of publicly consequential work should not disappear at shutdown. With this call for reflexivity and accountability, we join the active debate in critical research on dataset creation and use (Paullada et al., 2021; Sambasivan et al., 2021), while extending it to the measurement infrastructure that relies on and shapes those datasets. Past efforts at shared moderation infrastructure have not stalled for want of technical means but because such work counts as service labour. The community needs to recognise the scientific value of building measurement infrastructure, and to treat it as a form of care. Annotation schemas, inter-rater reliability scores, dataset construction decisions, and calibration evaluations deserve the same methodological rigour and visibility as model performance results. Funding bodies, programme committees, and tenure committees all participate in that judgement, and the community can change how such work is recognised and rewarded. Whether the field builds what replaces PERSPECTIVE API depends less on what is possible than on what it counts as research.

Limitations

No empirical evaluation This is a position paper laying out evaluation criteria, so we do not empirically evaluate successor infrastructure against the requirements we propose (T1–T5, G1–G5). Building and testing systems against these criteria is the natural follow-up to this scope. Relatedly, our claims about path dependence and the self-reinforcing loop (§4) rest on convergent qualitative evidence and the bibliographic search. Quantifying the drift and threshold sensitivity across the archived snapshot is a natural empirical extension. However, with the closure of the system, the focus should lie on learning from these issues.

Visibility bias Our critique of measurement modelling focuses on the well-known drawbacks of PERSPECTIVE API and may not fully reflect situations where the API worked well and met its goals. Our critique is also shaped by what is visible. PERSPECTIVE API published model cards, architecture papers, training corpora, and its own bias evaluations, so its failures are unusually well documented. No comparable evidence base exists for its closed competitors, and the density of criticism here should not be read as evidence that it was worse than tools nobody can inspect.

No costing We name funding as a binding constraint on the infrastructure we propose (§5) but do not price it. Compute, annotation, and long-term maintenance depend on scope, language coverage, and the institutional host, and we are not positioned to estimate them credibly from outside the bodies that would carry them. A costed proposal, developed with those bodies, is a necessary next step.

Partial stakeholder coverage The requirements we propose reflect the main priorities of the research community and may not include all stakeholder views, especially those of practitioners, affected communities, or people from non-Western and non-English-speaking backgrounds.

Ethical Considerations

This study does not involve human subjects. We did not collect data from participants or annotate any human-subject data. All labelling was done by our research team using published papers. The resource we share is a snapshot of Perspective API scores for 77 publicly available toxicity and hate-speech datasets. We only release record identifiers, scores,

and query metadata. The original texts are not included and remain available from each dataset’s original release under its own license. Some source datasets are already distributed as identifiers or have access restrictions set by their authors. By releasing only identifiers, we respect these conditions. However, this means that to reconstruct the scored texts, users must retrieve the source datasets, and some records from platform-based corpora may no longer be accessible. The scores are released under a CC BY 4.0 license, and each source dataset keeps its own license and citation. The source datasets were designed to include offensive and distressing language. By sharing identifiers instead of texts, we reduce but do not completely prevent the spread of this content, since identifiers may still link to material hosted elsewhere. This archive is meant for evaluating hate-speech detection systems and studying measurement infrastructure. It should not be used to train models to generate hateful content.

This paper critiques a measurement tool that over-moderates identity-related and reclaimed language and under-moderates implicit hate. These failures fall unevenly on marginalised communities. We document these failures to support the development of more equitable measurement infrastructure, not to prescribe in advance what it should look like. The requirements in §5 are offered in that spirit, since failures in hate-speech moderation can have serious consequences for the communities such moderation is meant to protect. As noted in our limitations, our account focuses on documented shortcomings and may under-represent contexts in which PERSPECTIVE API worked well. We also recognise that the requirements we propose reflect the priorities of our own research communities. As we argue under G3, legitimate successor infrastructure must be defined together with the communities most affected by toxicity measurement, and the requirements themselves should be validated against those communities’ priorities.

LLM-based Tools

We used writing assistance and LLM-based tools in a limited way during paper preparation and implementation. GitHub Copilot was used for code completion in the accompanying data release, Grammarly was used to suggest alternative phrasings, and Claude (Anthropic) was used for LaTeX formatting, copy-editing, and drafting suggestions on the abstract and Sections 2 to 5. All design de-

cisions and substantive writing were carried out by the authors, who reviewed and revised every AI-assisted edit.

Acknowledgments

This work was supported by the Weizenbaum Institute (grant number 16DII141), funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) and the State of Berlin. Here, D.H., A.K., L.S., and J.B. specifically acknowledge the support of the Weizenbaum Institute in context of the Short Project *Evaluating GenAI Evaluations*. This work also received funding from the European Union (EU) - NextGenerationEU and the Flemish Government under the “Onderzoeksprogramma Artificial Intelligence (AI) Vlaanderen” programme.

References

- Robert Adcock and David Collier. 2001. [Measurement validity: A shared standard for qualitative and quantitative research](#). *The American Political Science Review*, 95(3):529–546.
- Piush Aggarwal, Jawar Mehrabian, Weigang Huang, Özge Alacam, and Torsten Zesch. 2024. [Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 104–117, St. Julian’s, Malta. Association for Computational Linguistics.
- Amazon Web Services. 2026. [Amazon Rekognition pricing](#). AWS product documentation. Accessed: 2026-08-28.
- Luvell Anderson and Michael Barnes. 2023. [Hate Speech](#). In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, 2023 edition. Metaphysics Research Lab, Stanford University, Online.
- Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. [Resources for multilingual hate speech detection](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Somnath Banerjee, Sayan Layek, Pratyush Chatterjee, Animesh Mukherjee, and Rima Hazra. 2025. [Soteria: Language-Specific Functional Parameter Steering for Multilingual Safety Alignment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9347–9364, Suzhou, China. Association for Computational Linguistics.
- Pablo Barberá. 2015. [Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data](#). *Political Analysis*, 23(1):76–91.
- Niklas Barth, Elke Wagner, Philipp Raab, and Björn Wiegärtner. 2023. [Contextures of hate: Towards a systems theory of hate communication on social media platforms](#). *The Communication Review*, 26(3):209–252.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Jan Bartzner, Sree Harsha Nelaturu, Damian Stachura, Anastassia Kornilova, Jon Crall, Tommaso Cerruti, Yanan Long, Yifan Mai, Sanchit Ahuja, Asaf Yehudai, Marek Šuppa, John P. Lalor, Oluwagbemike Olowe, Jatin Ganhotra, Brian H. Hu, Eliya Habba, Andrew M. Bean, Chang Liu, Sander Land, and 29 others. 2026. [Every eval ever: A unifying schema and community repository for AI evaluation results](#). *Preprint*, arXiv:2606.14516.
- George Beknazar-Yuzbashev, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2025. Toxic content and user engagement on social media: Evidence from a field experiment. Technical report, CESifo Working Paper.
- Sam Biddle. 2025. [Leaked Meta rules: Users are free to post “Mexican immigrants are trash!” or “trans people are immoral”](#). *The Intercept*. Accessed: 2025-09-30.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2024. [Lessons from the trenches on reproducible evaluation of language models](#). *Preprint*, arXiv:2405.14782.
- Haris Bin Zia, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2022. [Toxicity in the decentralized web and the potential for model sharing](#). *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2):1–25.
- 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.

- Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. [Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets](#). 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 1004–1015, Online. Association for Computational Linguistics.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. [Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification](#). In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 491–500.
- Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA.
- Florian A. D. Burnat and Brittany I. Davidson. 2026. [The Accountability Paradox: How Platform API Restrictions Undermine AI Transparency Mandates](#), page 2416–2433. Association for Computing Machinery, New York, NY, USA.
- Ho-Chun Herbert Chang, James N. Druckman, Emilio Ferrara, and Robb Willer. 2023. [Liberals engage with more diverse policy topics and toxic content than do conservatives on social media](#). Technical Report WP-23-08, Institute for Policy Research, Northwestern University.
- Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, and 1 others. 2025. Integrative decoding: Improving factuality via implicit self-consistency. In *International Conference on Learning Representations*, volume 2025, pages 69573–69618.
- Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, and JinYeong Bak. 2025. [Unintended Harms of Value-Aligned LLMs: Psychological and Empirical Insights](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31742–31768, Vienna, Austria. Association for Computational Linguistics.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. [CONAN - COUNTER NARRATIVES THROUGH NICHE SOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.
- Lee Anna Clark and David Watson. 2019. [Constructing validity: New developments in creating objective measuring instruments](#). *Psychological Assessment*, 31(12):1412–1427.
- CNN. 2025. [Disorder breaks out in Northern Ireland for third straight night](#). CNN.
- Conversation AI. 2017. [Toxicity with subattributes: Crowdsourcing annotation scheme](#). https://github.com/conversationai/conversationai.github.io/blob/main/crowdsourcing_annotation_schemes/toxicity_with_subattributes.md. Accessed: 2026-04-17.
- Lee J Cronbach and Paul E Meehl. 1955. [Construct validity in psychological tests](#). *Psychological Bulletin*, 52(4):281–302.
- Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022. [HateCheckHIn: Evaluating Hindi Hate Speech Detection Models](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5378–5387, Marseille, France. European Language Resources Association.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and Play Language Models: A Simple Approach to Controlled Text Generation](#). In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.
- Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. [Hate speech classifiers learn normative social stereotypes](#). *Transactions of the Association for Computational Linguistics*, 11:300–319.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. [Automated hate speech detection and the problem of offensive language](#). In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515. AAAI Press.
- Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naqee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. [Multilingual and Explainable Text Detoxification with Parallel Corpora](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.
- Chiara Di Bonaventura, Barbara McGillivray, Yulan He, and Albert Meroño-Peñuela. 2025. [Hatevolution: What Static Benchmarks Don’t Tell Us](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17695–17707, Vienna, Austria. Association for Computational Linguistics.
- Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. [Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online](#). *Sexuality & Culture*, 25(2):700–732.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. [Measuring and Mitigating Unintended Bias in Text Classification](#). In

- Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73.
- Euronews. 2025. [How disinformation fuelled Spain's anti-migrant riots after attack on pensioner](#). Euronews.
- European Parliament and Council of the European Union. 2022. [Regulation \(EU\) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services \(Digital Services Act\)](#). *Official Journal of the European Union*, L 277.
- European Parliament and Council of the European Union. 2024. [Regulation \(EU\) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence \(Artificial Intelligence Act\)](#). *Official Journal of the European Union*, L series.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical Neural Story Generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. [A Hierarchically-Labeled Portuguese Hate Speech Dataset](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.
- Jeremy A. Frimer, Harinder Aujla, Matthew Feinberg, Linda J. Skitka, Karl Aquino, Johannes C. Eichstaedt, and Robb Willer. 2023. [Incivility is rising among American politicians on Twitter](#). *Social Psychological and Personality Science*, 14(2):259–269.
- Bharath Ganesh and Jonathan Bright. 2020. [Countering extremists on social media: Challenges for strategic communication and content moderation](#). *Policy & Internet*, 12(1):6–19.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. [Handling bias in toxic speech detection: A survey](#). *ACM Comput. Surv.*, 55(13s).
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.
- Tom Gerken. 2025. [TikTok puts hundreds of UK content moderator jobs at risk](#). *BBC*.
- Bryan T. Gervais, Connor Dye, and Amber Chin. 2025. [Incivility or invalidity? Evaluating Perspective API scores as a measure of political incivility](#). *American Politics Research*, 53(3):266–274.
- Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. [Detecting cross-geographic biases in toxicity modeling on social media](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.
- Shaona Ghosh, Heather Frase, Adina Williams, Sarah Luger, Paul Röttger, Fazl Barez, Sean McGregor, Kenneth Fricklas, Mala Kumar, Quentin Feuillade-Montixi, Kurt Bollacker, Felix Friedrich, Ryan Tsang, Bertie Vidgen, Alicia Parrish, Chris Knotz, Eleonora Presani, Jonathan Bennion, Marisa Ferrara Boston, and 83 others. 2025. [AILuminate: Introducing v1.0 of the AI risk and reliability benchmark from ML-Commons](#). *Preprint*, arXiv:2503.05731.
- Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. [Leveraging Multilingual Transformers for Hate Speech Detection](#). *Preprint*, arXiv:2101.03207.
- Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. [NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.
- Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. [Improving Adversarial Data Collection by Supporting Annotators: Lessons from GAHD, a German Hate Speech Dataset](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4405–4424, Mexico City, Mexico. Association for Computational Linguistics.
- Bastián González-Bustamante. 2024. [Benchmarking LLMs in Political Content Text-Annotation: Proof-of-Concept with Toxicity and Incivility Data](#). *Preprint*, arXiv:2409.09741.
- Google. 2018. [The New York Times: Using AI to host better conversations](#). Google Blog. Accessed: 2026-08-27.
- Robert Gorwa. 2019. [What is platform governance?](#) *Information, Communication & Society*, 22(6):854–871.

- Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. [Algorithmic content moderation: Technical and political challenges in the automation of platform governance](#). *Big Data & Society*, 7(1):2053951719897945.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pages 1321–1330.
- Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. [Word Embeddings Are Steers for Language Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430, Bangkok, Thailand. Association for Computational Linguistics.
- Laura Hanu and Unitary team. 2020. Detoxify. <https://github.com/unitaryai/detoxify>. Open-weight toxic comment classifiers trained on the Jigsaw Toxic Comment Classification, Unintended Bias in Toxicity Classification, and Multilingual challenge datasets.
- David Hartmann, José Renato Laranjeira de Pereira, Chiara Streitbürger, and Bettina Berendt. 2025a. [Addressing the regulatory gap: Moving towards an EU AI audit ecosystem beyond the AI Act by including civil society](#). *AI and Ethics*, 5:3617–3638.
- David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025b. [Lost in moderation: How commercial content moderation APIs over- and under-moderate group-targeted hate speech and linguistic variations](#). In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM.
- David Hartmann, Lena Pohlmann, Lelia Hanslik, Noah Gießing, Bettina Berendt, and Pieter Delobelle. 2026a. [Audit me if you can: Query-efficient active fairness auditing of black-box LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33673–33698, San Diego, California, United States. Association for Computational Linguistics.
- David Hartmann, Manuel Tonneau, Angelie Kraft, LK Seiling, Dimitri Staufer, Pieter Delobelle, Jan Filles, Anna Ricarda Luther, Jan Batzner, and Mareike Lisker. 2026b. [ByeByePerspectiveAPI-snapshot \(Revision 2f540d5\)](#). Dataset.
- Daniel Hickey, Daniel MT Fessler, Kristina Lerman, and Keith Burghardt. 2025. [X under Musk’s leadership: Substantial hate and no reduction in inauthentic activity](#). *PloS one*, 20(2):e0313293.
- Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. [The extended Moral Foundations Dictionary \(eMFD\): Development and applications of a crowd-sourced approach to extracting moral intuitions from text](#). *Behavior Research Methods*, 53(1):232–246.
- Manoel Horta Ribeiro. 2024. [Content Moderation in Online Platforms](#). Ph.D. thesis, École Polytechnique Fédérale de Lausanne.
- Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020a. [Characterizing twitter users who engage in adversarial interactions against political candidates](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Yiqing Hua, Thomas Ristenpart, and Mor Naaman. 2020b. [Towards measuring adversarial Twitter interactions against candidates in the US midterm elections](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):272–282.
- Abigail Z. Jacobs and Hanna Wallach. 2021. [Measurement and fairness](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 375–385, New York, NY, USA. Association for Computing Machinery.
- Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. [Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models](#). In *First Conference on Language Modeling*.
- Qi Jia, Xiang Yue, Tuney Zheng, Jie Huang, and Bill Yuchen Lin. 2025. [SimulBench: Evaluating Language Models with Creative Simulation Tasks](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8133–8146, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chonghe Jiang, Bao Nguyen, Anthony Man-Cho So, and Viet Anh Nguyen. 2025. [Probe-Free Low-Rank Activation Intervention](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2812–2824, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jigsaw. 2017. [Perspective API score normalization](#). Perspective API release notes. Accessed: 2026-08-24.
- Jigsaw. 2021. [Google’s Jigsaw announces toxicity-reducing API](#) Perspective is processing 500M requests daily.
- Jigsaw. 2022. [Perspective API model cards](#). Accessed 10 March 2026.
- Jigsaw. 2023. [Perspective API: Research into machine learning](#). Accessed 10 March 2026.
- Jigsaw. 2026. [Perspective API sunset announcement](#). Accessed 10 March 2026.
- Jigsaw and Conversation AI. 2021. [Jigsaw toxicity competitions on Kaggle](#). Kaggle. Accessed: 2026-08-27.

- Gauri Kambhatla, Matthew Lease, and Ashwin Rajadesingan. 2024. [Promoting Constructive Deliberation: Reframing for Receptiveness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5110–5132, Miami, Florida, USA. Association for Computational Linguistics.
- Michael L. Katz and Carl Shapiro. 1994. [Systems competition and network effects](#). *Journal of Economic Perspectives*, 8(2):93–115.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. [Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application](#). Preprint, arXiv:2009.10277.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ring-shia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. [The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity](#). *Journal of Communication*, 71(6):922–946.
- Hannah Kirk, Bertie Vidgen, Paul Röttger, Tristan Thrush, and Scott A. Hale. 2022. [Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-Based Hate](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.
- Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. [SemEval-2023 Task 10: Explainable Detection of Online Sexism](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.
- Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. 2023. [Pre-training Language Models with Human Preferences](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17506–17533. PMLR.
- Angelie Kraft, Judith Simon, and Sonja Schimmler. 2025. [Social bias in popular question-answering benchmarks](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1421–1438, Mumbai, India. AFNLP and ACL.
- Beomseok Lee, Hyunwoo Kim, Keon Kim, and Yong Suk Choi. 2024. [XDetox: Text Detoxification with Token-Level Toxicity Explanations](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15215–15226, Miami, Florida, USA. Association for Computational Linguistics.
- Deokjae Lee, JunYeong Lee, Jung-Woo Ha, Jin-Hwa Kim, Sang-Woo Lee, Hwaran Lee, and Hyun Oh Song. 2023. [Query-efficient black-box red teaming via bayesian optimization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. [A new generation of Perspective API: Efficient multi-lingual character-level transformers](#). In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.
- João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. [Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. [Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.
- Xun Liang, Hanyu Wang, Shichao Song, Mengting Hu, Xunzhi Wang, Zhiyu Li, Feiyu Xiong, and Bo Tang. 2024. [Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5797–5814, Bangkok, Thailand. Association for Computational Linguistics.

- Shaoyi Liaw, Fan Huang, Fabricio Benevenuto, Hae-woon Kwak, and Jisun An. 2023. [YouNICon: Youtube’s community of conspiracy videos](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1102–1111.
- Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. [Toxic Content Detection in online social networks: a new dataset from Brazilian Reddit Communities](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 472–482, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. [ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702. Association for Computational Linguistics.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. [DEXperts: Decoding-time controlled text generation with experts and anti-experts](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.
- Yang Liu. 2024. [Quantifying Stereotypes in Language](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1223–1240, St. Julian’s, Malta. Association for Computational Linguistics.
- Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. [A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276. Association for Computational Linguistics.
- Josephine Lukito, Bin Chen, Gina M. Masullo, and Natalie Jomini Stroud. 2024. [Comparing a BERT Classifier and a GPT classifier for Detecting Connective Language Across Multiple Social Media](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19140–19153, Miami, Florida, USA. Association for Computational Linguistics.
- Anna Ricarda Luther, Hendrik Heuer, Stephanie Geise, Sebastian Haunss, and Andreas Breiter. 2025. [Social media for activists: Reimagining safety, content presentation, and workflows](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI ’25*, New York, NY, USA. Association for Computing Machinery.
- João C. Magalhães, Clara Iglesias Keller, and Robert Gorwa. 2026. [The Great Sysop: Elon Musk, X, and the emergence of platform illiberalism](#). *New Media & Society*, page 14614448261424889.
- Antonis Maronikolakis, Axel Wisioerek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 2022. [Listening to affected communities to define extreme speech: Dataset and experiments](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.
- Teresa Marques. 2023. [The expression of hate in hate speech](#). *Journal of Applied Philosophy*, 40(5):769–787.
- Adrienne Massanari. 2017. [#Gamergate and The Fapening: How Reddit’s algorithm, governance, and culture support toxic technocultures](#). *New Media & Society*, 19(3):329–346.
- 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. AAAI Press.
- Rod McGuirk. 2024. [X Corp. has slashed 30% of trust and safety staff, an Australian online safety watchdog says](#). *AP News*.
- Liv McMahon, Zoe Kleinman, and Courtney Subramanian. 2025. [Facebook and Instagram get rid of fact checkers](#). *BBC*.
- Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. [From dogwhistles to bullhorns: Unveiling coded rhetoric with language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.
- Helena Mihaljević and Elisabeth Steffen. 2022. [How toxic is antisemitism? potentials and limitations of automated toxicity scoring for antisemitic online content](#). In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis*, pages 1–12, Potsdam, Germany.
- Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. [Mix and Match: Learning-free Controllable Text Generation using Energy Language Models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.
- Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. [ETHOS: a multi-label hate speech detection dataset](#). *Complex & Intelligent Systems*, 8(6):4663–4678.

- Robert Morabito, Jad Kabbara, and Ali Emami. 2023. [Debiasing should be Good and Bad: Measuring the Consistency of Debiasing Techniques in Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4581–4597, Toronto, Canada. Association for Computational Linguistics.
- Rachel Elizabeth Moran, Joseph Schafer, Mert Bayar, and Kate Starbird. 2025. [The end of trust and safety?: Examining the future of content moderation and upheavals in professional online safety efforts](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. [Understanding and interpreting the impact of user context in hate speech detection](#). In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.
- Mohsen Mosleh, Rocky Cole, and David G Rand. 2024. [Misinformation and harmful language are interconnected, rather than distinct, challenges](#). *PNAS Nexus*, 3(3):pgae111.
- Paul Mozur. 2018. [A genocide incited on Facebook, with posts from Myanmar’s military](#). *The New York Times*.
- Helen Ngo, Cooper Raterink, João G.M. Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. [Mitigating harm in language models with conditional-likelihood filtration](#). *Preprint*, arXiv:2108.07790.
- Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2025. [Toxic bias: Perspective API misreads German as more toxic](#). In *Proceedings of the 19th AAAI International Conference on Web and Social Media (ICWSM’25)*, volume 19, pages 1346–1357, Copenhagen, Denmark. AAAI Press.
- Felipe Oliveira, Victoria Reis, and Nelson Ebecken. 2023. [TuPy-E: detecting hate speech in Brazilian Portuguese social media with a novel dataset and comprehensive analysis of models](#). *Preprint*, arXiv:2312.17704.
- OpenAI. 2026. [Moderation](#). OpenAI API Documentation. Accessed: 2026-08-27.
- Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. [AustroTox: A Dataset for Target-Based Austrian German Offensive Language Detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11990–12001, Bangkok, Thailand. Association for Computational Linguistics.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. [Data and its \(dis\)contents: A survey of dataset development and use in machine learning research](#). *Patterns*, 2(11):100336.
- John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. [ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 571–576, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. [On the challenges of using black-box APIs for toxicity evaluation in research](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7595–7609, Singapore. Association for Computational Linguistics.
- Steve Rathje, Jay J. Van Bavel, and Sander Van Der Linden. 2021. [Out-group animosity drives engagement on social media](#). *Proceedings of the National Academy of Sciences*, 118(26):e2024292118.
- Traian Rebedea, Makesh Sreedhar, Shaona Ghosh, Jiaqi Zeng, and Christopher Parisien. 2024. [CantTalkAboutThis: Aligning Language Models to Stay on Topic in Dialogues](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12232–12252, Miami, Florida, USA. Association for Computational Linguistics.
- Bernhard Rieder and Yarden Skop. 2021. [The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API](#). *Big Data & Society*, 8(2).
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. [Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis](#). In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum, Germany.
- 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.
- Pratik S. 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.

- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- 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.
- Martin Saveski, Brandon Roy, and Deb Roy. 2021. [The structure of toxic conversations on Twitter](#). In *Proceedings of the web conference 2021*, pages 1086–1097.
- Alexandra A. Siegel. 2020. [Online hate speech](#). In Nathaniel Persily and Joshua A. Tucker, editors, *Social Media and Democracy: The State of the Field, Prospects for Reform*, pages 56–88. Cambridge University Press, Cambridge.
- Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025. [Multilingual Blending: Large Language Model Safety Alignment Evaluation with Language Mixture](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3433–3449, Albuquerque, New Mexico. Association for Computational Linguistics.
- Susan Leigh Star. 1999. [The ethnography of infrastructure](#). *American Behavioral Scientist*, 43(3):377–391.
- Marilyn Strathern. 1997. ‘Improving ratings’: Audit in the British university system. *European Review*, 5(3):305–321.
- Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. [On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.
- Anna Talas and Alice Hutchings. 2023. [Hacker’s paradise: Analysing music in a cybercrime forum](#). In *2023 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–14.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Stanford Alpaca: An Instruction-following LLaMA Model](#). Stanford Center for Research on Foundation Models.
- Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. [The dynamics of political incivility on Twitter](#). *Sage Open*, 10(2).
- Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. [From languages to geographies: Towards evaluating cultural bias in hate speech datasets](#). In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.
- Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2025. [HateDay: Insights from a global hate speech dataset representative of a day on Twitter](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2321, Vienna, Austria. Association for Computational Linguistics.
- Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. [Large-Scale Hate Speech Detection with Cross-Domain Transfer](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.
- Petter Törnberg and Juliana Chueri. 2025. [Elite political discourse has become more toxic in Western countries](#). *Preprint*, arXiv:2503.22411.
- UK Parliament. 2023. [Online Safety Act 2023](#). c. 50.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *J. Artif. Int. Res.*, 72:1385–1470.
- Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benvenuto. 2022. [HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.
- Siddharth Venkataramakrishnan. 2025. [From disinformation to violence: The UK far right and 2024 riots](#). *Counter Terrorist Trends and Analyses*, 17(2):8–15.
- Tommaso Venturini. 2019. [From fake to junk news, the data politics of online virality](#). In Didier Bigo, Engin Isin, and Evelyn Ruppert, editors, *Data Politics: Worlds, Subjects, Rights*, pages 123–144. Routledge, London.
- 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, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A. Hale, and Paul Röttger. 2023. [SimpleSafetyTests: a Test Suite for Identifying Critical Safety Risks in Large Language Models](#). *Preprint*, arXiv:2311.08370.
- Andreas Waldis, Vagrant Gautam, Anne Lauscher, Dietrich Klakow, and Iryna Gurevych. 2026. [Aligned Probing: Relating Toxic Behavior and Model Internals](#). In *Transactions of the Association for Computational Linguistics, Volume 14*, pages 271–291, Cambridge, MA. MIT Press.
- Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. [Position: Evaluating generative AI systems is a social science measurement challenge](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82232–82251. PMLR.
- Hongbo Wang, Mingda Li, Junyu Lu, Hebin Xia, Liang Yang, Bo Xu, Ruizhu Liu, and Hongfei Lin. 2024. [PcIGPT: A Large Language Model for Patronizing and Condescending Language Detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6913–6928, Miami, Florida, USA. Association for Computational Linguistics.
- Yifan Wang and Vera Demberg. 2024. [RSA-Control: A Pragmatics-Grounded Lightweight Controllable Text Generation Framework](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5561–5582, Miami, Florida, USA. Association for Computational Linguistics.
- Zeera Waseem and Dirk Hovy. 2016. [Hateful symbols or hateful people? predictive features for hate speech detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. [Challenges in detoxifying language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. [Ex machina: Personal attacks seen at scale](#). In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.
- Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. [Detoxifying language models risks marginalizing minority voices](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397. Association for Computational Linguistics.
- Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. [Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13967–13983, Bangkok, Thailand. Association for Computational Linguistics.
- Ziqing Yang, Yixin Wu, Rui Wen, Michael Backes, and Yang Zhang. 2026. [Peering Behind the Shield: Guardrail Identification in Large Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11663–11676, San Diego, California, United States. Association for Computational Linguistics.
- Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. [How hate speech varies by target identity: A computational analysis](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. [Hate speech and counter speech detection: Conversational context does matter](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.
- Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. [Conversations Gone Awry: Detecting Early Signs of Conversational Failure](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Requirements Overview Table

In Table 1, we present the complete list of ten requirements, each linked to the specific failure in §3 and §4 that inspired it. This mapping is meant to describe our process, not to serve as a final validation. It shows how we developed the requirements based on one tool’s documented failures. The research community will need to test and refine these requirements as they apply them to new systems.

B Bibliographic Review of Papers that used PERSPECTIVE

B.1 Method

We conducted a structured bibliographic search across ACL Anthology, Web of Science, and OpenAlex, identifying 768 records (383, 14, and 371 respectively). After removing 216 cross-database duplicates and 51 records excluded by rule-based screening (year < 2017, off-topic titles, or keyword-collision false positives), 501 records remained for title/abstract screening, of which full text was sought for 454 and retrieved for 397.

For each retrieved paper, we extracted every mention of “Perspective API”, “Jigsaw”, or a bare “perspective” occurring near a toxicity-related term, taking a ± 400 -character context window around each match. **All papers were labelled by hand by the first author**, using these extracted snippets to decide (a) whether the paper uses PERSPECTIVE in its own research and (b) its usage type – baseline model, CSS measurement, pre-training filtering, LLM evaluation, or meta-critique – consulting the full paper directly whenever the snippets left this ambiguous.

An LLM (gpt-5-mini), given the same snippets, was used only to cross-check the human labels, never to make the primary decision. Agreement between the LLM and the first author was 386/397 (97.2%) on inclusion and 80.3% on usage type. Every non-agreement was then reconsidered by hand against the full paper. Three usage-type labels were changed as a result; for all remaining cases the human label was retained.

Beyond the structured search, we separately traced the citations of every dataset we rescored (Appendix C) back to its own originating paper. This citation-chaining step is not part of the systematic search above, but surfaced 12 further papers not caught by the original search terms, itself a small illustration of how fragmented and hard to fully enumerate this literature already is.

In total we identify 241 papers using PERSPECTIVE in their own research (223 from the four categories discussed in the main text, plus 18 meta-critique papers, for which PERSPECTIVE is the object rather than the instrument of study). Table 2 gives the full breakdown. All paper categorisations can be found in the accompanying snapshot dataset (Hartmann et al., 2026b).

B.2 PRISMA Flow

Figure 2 reports identification through inclusion following PRISMA 2020: 768 records identified, 267 removed before screening (216 duplicates and 51 excluded by automation), 501 screened, 47 excluded at title/abstract, 454 sought for retrieval, 57 not retrieved, 397 assessed with full text, 168 excluded as not using PERSPECTIVE, and 229 included by human review. A second, non-systematic identification stream, citation-chaining from the archived dataset snapshot, added 12 further records outside this flow, as described above, giving **241 papers** that used PERSPECTIVE API. Most excluded papers used Jigsaw’s Kaggle datasets but did not use PERSPECTIVE for scoring.

C The Archived Snapshot

C.1 Method

Datasets were selected by first identifying every dataset cited in this paper’s corpus with a publicly available release and scoring it. We then traced outward in three further ways. (i) Known, curated Hugging Face releases of hate-speech and toxicity datasets, including several used to construct multilingual superset benchmarks. (ii) A scan of every GitHub, OSF, Zenodo, and figshare repository already retrieved as a code artifact of a paper in our corpus, using both filename-based and content-based text-column detection, the latter to catch datasets whose text field is not conventionally named. (iii) A targeted search, per paper without an identified code artifact, for a corresponding public release. Every dataset was manually checked against its cited source paper before inclusion. Ambiguous or unverifiable matches were excluded rather than guessed.

C.2 Overview

In total, the snapshot covers **77 datasets and major benchmarks, totalling 6,162,826 records, of which 5,897,618 (95.7%) were scored successfully**. Failures are concentrated in languages PER-

	Requirement	Recommendation	Addressed failure
T1	Reproducibility	All metadata components of evaluations should be documented, i.e., metadata on the model, benchmark dataset, and evaluation metric (Biderman et al., 2024), in a community-agreed schema, and each release should be pinned to a citable version (Batzner et al., 2026).	Irreproducible, incompatible results
T2	Validity	Evaluation tools should explicitly document which constructs, conditions, and input formats they measure, including limitations such as threshold selection and out-of-scope queries (Blodgett et al., 2021; Jacobs and Wallach, 2021; Wallach et al., 2025).	Construct underspecification
T3	Contextuality	Scoring should support structured metadata input, namely who is speaking, who is targeted, and prior context, to enable classification of counterspeech and reclaimed language (Yu et al., 2022).	Misclassification of identity-laden content
T4	Uncertainty	Evaluation tools should return a score that reflects how much annotators agree rather than model confidence, as that disagreement reflects contested constructs (Leonardelli et al., 2021; Uma et al., 2021).	Score unrelated to annotator agreement
T5	Multilinguality	Performance must be validated and documented per language, with explicit disclosure where it falls below a defined threshold (Nogara et al., 2025; Tonneau et al., 2025).	Uneven cross-lingual performance
G1	Independence	The infrastructure should be open-source and institutionally independent of any commercial entity, so no single organisation controls annotation schema, model weights, or the definition of toxicity, as PERSPECTIVE API demonstrated (Rieder and Skop, 2021; Pozzobon et al., 2023). No instrument should both label training data and grade the models built from it.	Single-vendor dependence, circular evaluation
G2	Transparency	Model weights, training data, and annotation schemata should be open and auditable, including annotator demographics, which influence group-level disparities (Sap et al., 2022; Davani et al., 2023; Kraft et al., 2025).	Opacity of training and annotation
G3	Legitimacy	Defining toxicity, selecting communities, and collecting training data must involve affected communities through an open, documented process (Maronikolakis et al., 2022). Legitimacy does not require consensus. It requires that systematic disagreement, by annotator identity and target group, be documented and propagated rather than averaged away (Leonardelli et al., 2021; Sap et al., 2022; Davani et al., 2023).	Contested disagreement averaged away
G4	Accountability	Auditing with published results and documented remediation across demographic groups, languages, and content types should be a continuous, open process run by auditors independent of the maintainers (Garg et al., 2023; Hartmann et al., 2026a).	Undetected systematic errors
G5	Sustainability	A long-term funding model must be put in place that does not reproduce the single-funder dependence that made PERSPECTIVE API’s shutdown so disruptive (Rieder and Skop, 2021).	Funding dependence

Table 1: Ten requirements and recommendations for successor toxicity measurement infrastructure, derived from the structural failures documented in §3.

SPECTIVE does not support at all (e.g. Turkish, Finnish), which we report explicitly per dataset rather than omitting.

Each dataset was scored via PERSPECTIVE’s `v1alpha1 comments:analyze` endpoint for all six production attributes (toxicity, severe toxicity, identity attack, insult, profanity, threat), with `doNotStore` set on every request. An explicit language hint was sent only for PERSPECTIVE’s documented production-supported languages. Otherwise the request relied on PERSPECTIVE’s own language auto-detection. Scoring took place between 2026-05-22 and 2026-08-28. PERSPECTIVE exposes no public model-version identifier,

so we record a per-record query timestamp as the closest available version pin. We release, per dataset: a stable dataset and record identifier (sufficient to merge back onto the original release), the six attribute scores, the declared and PERSPECTIVE-detected language, success/failure status and error reason where applicable, and the query timestamp, not the underlying text, which remains available from each dataset’s original release under its own license. Each dataset ships with a short README documenting its source, license pointer, and merge key. PERSPECTIVE scores are released under CC-BY 4.0. Full per-paper and per-dataset provenance, which paper each

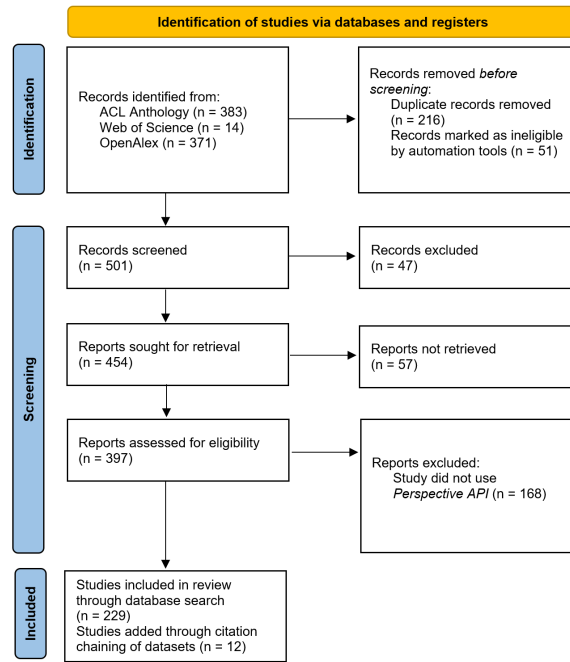


Figure 2: PRISMA flow diagram for the bibliographic search. Of 768 records identified across ACL Anthology, Web of Science, and OpenAlex, 229 were confirmed to use PERSPECTIVE in their own research. A further 12 were added through citation chaining from the datasets we rescored (Appendix C), giving 241 in total. Screening was done by the first author on verbatim full-text snippets of PERSPECTIVE API usage, with LLM-assisted screening used as a check on agreement (97.2% agreement on inclusion, 80.3% on usage type). After the check, all non-agreements were reconsidered and 3 papers were changed; for the rest, the human label was used.

Usage type	<i>n</i>
Baseline model	83
LLM evaluation	61
CSS measurement	50
Pre-training filtering	29
Meta-critique	18
Total	241

Table 2: Papers using PERSPECTIVE, by usage type (human-labelled).

dataset is linked to, its citation, and licensing, is provided as a table in the accompanying repository: <https://huggingface.co/datasets/dawiet/ByeByePerspectiveAPI-snapshot>. An overview of datasets can be seen in Table 3.

Dataset	Paper	Lang.	Category	<i>n</i> rows
CivilComments	Borkan et al. (2019) ; Dixon et al. (2018)	en	Baseline model	1,968,756
PolygloToxicityPrompts	Jain et al. (2024)	multi	LLM evaluation	847,420
HateDay	Tonneau et al. (2025)	multi	Baseline model	501,348
Hate Speech Superset (Arabic)	Tonneau et al. (2024)	ar	Baseline model	447,561
DATG	Liang et al. (2024)	en	LLM evaluation	372,275
RealToxicityPrompts	Gehman et al. (2020)	en	LLM evaluation	198,231
Toraman et al. Hate Speech	Toraman et al. (2022)	tr, en	Baseline model	197,849
Pile Toxicity (balanced)	Korbak et al. (2023)	en	Baseline model	196,759
XDetoX	Lee et al. (2024)	en	LLM evaluation	178,024
Wikipedia Detox	Wulczyn et al. (2017)	en	Baseline model	159,117
EFI	Jiang et al. (2025)	en	LLM evaluation	90,038
Wiki Madlibs	Dixon et al. (2018)	en	LLM evaluation	89,483
Unintended Harms	Choi et al. (2025)	en	LLM evaluation	59,814
Alpaca	Taori et al. (2023)	en	LLM evaluation	51,004
Hate Speech Superset (German)	Tonneau et al. (2024)	de	Baseline model	50,254
TuPy-E	Oliveira et al. (2023)	pt	Baseline model	43,210
Hate Speech Superset (Portuguese)	Tonneau et al. (2024)	pt	Baseline model	43,201
Hate Speech Superset (Turkish)	Tonneau et al. (2024)	tr	Baseline model	41,315
Hatevolution	Di Bonaventura et al. (2025)	en	Baseline model	41,133
Measuring Hate Speech	Kennedy et al. (2020) ; Sachdeva et al. (2022)	en	Baseline model	39,564
Lost in Moderation	Hartmann et al. (2025b)	en	Baseline model	35,491
WritingPrompts	Fan et al. (2018)	en	LLM evaluation	30,746
Hate Speech Superset (Spanish)	Tonneau et al. (2024)	es	Baseline model	29,646
Davidson et al. HSOL	Davidson et al. (2017)	en	Baseline model	24,782
PclGPT (Pcl-PT/SFT)	Wang et al. (2024)	en	Pre-training filter	24,052
Soteria	Banerjee et al. (2025)	ar	Pre-training filter	22,793
CantTalkAboutThis	Rebedea et al. (2024)	en	Baseline model	22,598
Aligned Probing	Waldis et al. (2026)	en	LLM evaluation	21,969
ToLD-Br	Leite et al. (2020)	pt	Baseline model	20,813
EDOS	Kirk et al. (2023)	en	Pre-training filter	20,000

Dataset	Paper	Lang.	Category	<i>n</i> rows
HatEval	Basile et al. (2019)	en, es	Baseline model	19,341
HateXplain	Mathew et al. (2021)	en	Baseline model	19,192
Hate Speech Superset (French)	Tonneau et al. (2024)	fr	Baseline model	17,879
LM-Steer	Han et al. (2024)	en	LLM evaluation	17,394
Capybara	—	en	Pre-training filter	15,935
Mix and Match	Mireshghallah et al. (2022)	en	LLM evaluation	15,144
Hate Speech Superset (Indonesian)	Tonneau et al. (2024)	id	Baseline model	14,184
Hate Speech Superset (India)	Tonneau et al. (2024)	multi	Baseline model	14,095
Lukito et al. Connective Data	Lukito et al. (2024)	en	CSS measurement	12,151
Ghosh Roy et al. Multilingual HS	Ghosh Roy et al. (2021)	en	Baseline model	12,072
CONAN	Chung et al. (2019)	en	Baseline model	11,641
GAHD	Goldzycher et al. (2024)	de	Pre-training filter	10,994
GAHD (mirror)	Goldzycher et al. (2024)	de	Pre-training filter	10,994
ToxBART	Yadav et al. (2024)	en	Baseline model	10,423
ToxicChat	Lin et al. (2023)	en	Pre-training filter	9,718
Arango Monnar et al. (Chilean)	Arango Monnar et al. (2022)	es	Baseline model	9,216
DiaSafety	Sun et al. (2022)	en	Baseline model	7,692
HateBR	Vargas et al. (2022)	pt	Baseline model	7,000
HateCheckHIn	Das et al. (2022)	hi	Baseline model	5,884
Fortuna et al. Portuguese HS	Fortuna et al. (2019)	pt	Baseline model	5,668
TextDetox (Japanese)	Dementieva et al. (2025)	ja	Baseline model	4,908
HatemojiBuild	Kirk et al. (2022)	en	Baseline model	4,723
HateCheck	Röttger et al. (2021)	en	Baseline model	3,728
Quantifying Stereotypes	Liu (2024)	en	Baseline model	2,960
YouNICon	Liaw et al. (2023)	en	CSS measurement	2,528
Reddit-BR Toxicity	Lima et al. (2024)	pt	Baseline model	2,500
Instructive Debiasing	Morabito et al. (2023)	en	LLM evaluation	2,398
Integrative Decoding	Cheng et al. (2025)	en	LLM evaluation	2,271
Suomi24 Toxicity	—	fi	Baseline model	2,236
Conversations Gone Awry	Zhang et al. (2018)	en	Pre-training filter	2,169
Perspective API Eval Set	Lees et al. (2022)	en	Baseline model	1,997
AustroTox	Pachinger et al. (2024)	de	Pre-training filter	1,983
AustroTox (mirror)	Pachinger et al. (2024)	de	Pre-training filter	1,983
Constructive Deliberation	Kambhatla et al. (2024)	en	Pre-training filter	1,956
RSA-Control	Wang and Demberg (2024)	en	LLM evaluation	1,507

Dataset	Paper	Lang.	Category	<i>n</i> rows
Chilean Protests (Spanish)	González-Bustamante (2024)	es	Baseline model	1,000
Chilean Constitution (Spanish)	González-Bustamante (2024)	es	Baseline model	1,000
ETHOS	Mollas et al. (2022)	en	Baseline model	997
Barberá Twitter Ideology	Barberá (2015)	en	Baseline model	945
SimulBench	Jia et al. (2025)	en	CSS measurement	753
eMFD	Hopp et al. (2021)	en	Pre-training filter	495
Ross et al. Refugee Tweets	Ross et al. (2016)	de	Baseline model	469
PPLM	Dathathri et al. (2020)	en	Pre-training filter	440
Text vs. Image Hate Detection	Aggarwal et al. (2024)	en	Baseline model	407
Multilingual Blending	Song et al. (2025)	multi	LLM evaluation	360
AP-Test	Yang et al. (2026)	en	Baseline model	150
SimpleSafetyTests	Vidgen et al. (2023)	en	LLM evaluation	100

Table 3: Archived snapshot: rescored datasets, their corresponding paper, dominant language(s), usage category, and row count. Rows labelled *Hate Speech Superset* are language- and geography-specific subsets of the superset released by [Tonneau et al. \(2024\)](#). Rows marked (*mirror*) are separate public releases of the same underlying data, scored and released independently; their records are included in the totals in §C.