LLMs as Annotators: Evaluating Model–Human Alignment in Detecting Contentious Language in Historical Corpora

Yahui Zhao, Clemencia Siro, Laura Hollink
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
{Yahui.Zhao, c.n.siro, l.hollink}@cwi.nl

Abstract

Historical texts often contain terminology that reflects outdated or harmful social values. Identifying such contentious terms is essential for the Galleries, Libraries, Archives, and Museums (GLAM) community, but manual annotation requires cultural expertise and is difficult to scale. This study evaluates whether large language models (LLMs) can support this process by aligning with human judgments of contentiousness in historical Dutch corpora. Using the Dutch *Contentious terms in Contexts Corpus* (*ConConCor*), we formalize the task as context-dependent binary classification and compare two LLMs across multiple prompt configurations and evaluation scenarios. The models achieve near-human-level agreement on explicit cases but diverge when contextual or historical reasoning is required. Analysis of disagreement patterns shows that LLMs capture overtly harmful expressions yet tend to over-predict contentiousness for identity-related and colonial terms and under-predict for semantically shifted or figurative uses. These findings suggest that LLMs can serve as auxiliary annotators for sensitive-language detection in historical materials, provided that human oversight and contextual interpretation remain central to annotation workflows.

Keywords: large language models, contentious language detection, automatic annotation, cultural heritage

Content Warning: This paper may contain examples with harmful or offensive language. Such language does not reflect the authors' views but is retained to represent the original sources and enable critical analysis faithfully.

1. Introduction

Historical texts offer invaluable insights into past societies, but also reflect the hierarchies and prejudices of their time. When institutions in the GLAM (galleries, libraries, archives, and museums) sector digitize these materials, they face the challenge of exposing terms that may today be considered outdated, biased, or offensive. Such language, often tied to colonial or discriminatory histories, can perpetuate harm or exclusion if presented without context (Modest, 2018). Responsible access, therefore, requires methods to identify such terms for curatorial review.

Existing workflows for identifying contentious terms rely on manual annotation by experts or trained crowdworkers (Kaldeli et al., 2021). This process is resource-intensive and difficult to scale to large corpora, as annotators must interpret not only the historical language but also the sociocultural context in which terms were used (Krusic, 2024). Judgments of offensiveness or contentiousness are therefore subjective and context-dependent. For example, the Dutch term *inheems* ("indigenous") is considered derogatory when describing a human character but neutral when referring to a plant. Beyond such interpretive variation, historical texts present additional challenges: prejudice and power relations are often conveyed indirectly through euphemisms or institutional language, and lexical meanings shift over time through semantic drift and polysemy (Yogarajan et al., 2023). These characteristics make it difficult for both humans and computational models to determine contentiousness from linguistic form alone, highlighting the need for context-sensitive methods (MacAvaney et al., 2019).

Recently, large language models (LLMs) offer a promising alternative, with strong contextual reasoning and natural-language explanation capabilities (Bamman et al., 2024; Ziems et al., 2024). Early work suggests that LLMs can approximate or even surpass human performance in ethical or value-sensitive classification tasks (Roy et al., 2023; Hamerlik et al., 2024), enabling scalable support for GLAM curators. Yet it remains unclear to what extent LLMs *align with human judgments* in identifying contentious terms, especially in historically and culturally sensitive contexts, raising concerns about reliability (Egami et al., 2023).

Therefore, in this work, we investigate: **Can LLMs assist in annotating contentious terms in a historical context and be consistent with human annotators in their judgments?**

To address this question, we frame the problem as a *context-dependent term classification* task, where the goal is to determine whether a target term in its surrounding historical context should be considered contentious by contemporary standards. We evaluate two LLMs, one closed and

the other open-source, on 2,090 annotated excerpts from the Dutch *Contentious terms in Contexts Corpus* (ConConCor) (Brate et al., 2021). The models are assessed under multiple prompt configurations and evaluation scenarios that compare their predictions against expert and crowd annotations. We measure alignment through inter-annotator agreement, accuracy with respect to human majority labels, and distributional similarity of label proportions, complemented by qualitative analyses of model–human disagreement and term-level sensitivity. This combined design allows us to examine not only how closely LLMs replicate human judgments, but also how their decision patterns differ across linguistic categories and historical usage contexts.

Our results show that LLMs achieve near-human-level reliability in explicit, unambiguous cases ($\alpha \approx 0.45$–0.53) but diverge from human annotators when interpretation depends on historical or contextual cues. Models capture overtly harmful expressions with high alignment with humans but tend to over-predict contentiousness for identity-related and colonial terms and under-predict for semantically shifted or figurative usages. These findings highlight both the potential and the current limitations of LLMs as auxiliary annotators for sensitive language detection in historical and cultural heritage materials.

Our contributions are threefold. First, we introduce a systematic evaluation of LLMs as potential annotators for identifying contentious terms in historical text, framing the task as context-dependent classification. Second, we conduct a comprehensive comparison across model types, prompt configurations, and annotation settings, combining quantitative agreement metrics with qualitative analysis of model–human disagreement. Finally, we provide empirical evidence on the strengths and limitations of current LLMs in handling socially and historically sensitive language, offering practical insights for their integration into annotation workflows within GLAM and other cultural heritage contexts.

2. Related Work

Within the ongoing societal debate around decolonization and representation of people and cultures, the terminology used to describe them has received growing attention (Ariese, 2020; Odumosu, 2020). Words Matter (Modest, 2018) advocates for addressing outdated or offensive terminology in GLAM collections, and provides English and Dutch glossaries of colonial and other terms. It contains information on the historical usage of each term, and proposes alternatives. Words Matter has become influential among both cultural heritage professionals and researchers. Building on it, Nesterov et al. (2023) created a machine-readable knowledge graph of contentious terminology, and Menis Mastromichalakis et al. (2025) extended this model to a multilingual resource within the De-BIAS project, incorporating, apart from Words Matter terms, many others in a wide range of languages.

Several computational methods have been proposed for the detection of harmful words in historical data. Brate et al. (2021) and Hoeken et al. (2023) applied supervised models to Dutch and English historical texts, while Nesterov et al. (2024) examined how existing knowledge graphs capture stereotyping or offensive language. The De-BIAS project further developed an API for large-scale multilingual detection of contentious terms (Menis Mastromichalakis et al., 2025).

Recent work has turned to LLMs as potential collaborators. Osti and Roke (2024) and Menis Mastromichalakis et al. (2025) propose using LLMs not merely as classification tools but as interpretive partners capable of providing natural-language explanations of why terms may be contentious, yet they either do not compare LLM judgments to human annotations or rely solely on accuracy-based evaluation. In parallel, Ziems et al. (2024) have shown that LLMs can reach human-level agreement in identifying hate speech and toxicity in contemporary texts, as measured by inter-annotator agreement and accuracy. Motivated by these successful use cases, we study how well open- and closed-source LLMs align with human annotators in recognizing contentious terminology in Dutch historical texts.

Beyond performance, several papers emphasize that disagreement among human annotators should be seen as signal rather than noise (Aroyo and Welty, 2015; Khurana et al., 2024). For offensiveness judgment, models should aim to reflect the distribution of human judgments rather than a single gold label, to not ignore minority voices (Sap et al., 2022). Building on this line of work, our study evaluates to what extent LLMs can reproduce the variability inherent in human judgment.

3. Methodology

This section describes the experimental design used to evaluate whether LLMs can align with human judgments when identifying contentious terms in historical Dutch texts. All experiments are conducted in a zero-shot setting to assess the models' intrinsic ability to reason about historical and contextual sensitivity without additional supervision. All code used in this study is available in

the accompanying GitHub repository.[1]

3.1. Task Definition and Data

We formulate the identification of contentious terms as a *context-dependent binary classification* task. Given a historical text excerpt x containing a highlighted target term t, the model predicts a binary label $y \in \{0, 1\}$, where $y{=}1$ denotes a contentious term and $y{=}0$ a non-contentious term:

$$f_\theta(x, t) \rightarrow \hat{y} \in \{0, 1\},$$

where $\hat{y}$ denotes the predicted label and θ the model parameters. The decision relies on contextual semantics rather than lexical identity. The key challenge lies in *contextual semantics*. Many terms acquire or lose sensitive connotations depending on the referent or author's intent, and polysemy or diachronic semantic change further complicates interpretation. Consequently, even expert annotators may disagree, making this a nuanced and inherently subjective classification problem.

Dataset. We use the Dutch Contentious terms in Contexts Corpus (ConConCor) (Brate et al., 2021), derived from the Europeana Dutch Newspaper Collection[2] spanning 1890–1941. The corpus contains 2,715 excerpts, each highlighting one potentially sensitive Dutch term and annotated by experts and crowdworkers as contentious or not. Most excerpts (2,395) received at least seven independent annotations. We use the human annotations as reference labels and adopt the original annotation guidelines as the baseline prompts for LLM evaluation, ensuring conceptual alignment between human and model judgments.

Following Brate et al. (2021), we perform a three-step filtering process to obtain reliable evaluation data. First, we remove low-quality annotators whose mean pairwise Krippendorff's α was below 0.2 or who failed more than half of the five control items used by Brate et al. (2021). Second, to mitigate noise from optical character recognition (OCR) errors, we exclude all excerpts marked as "Illegible OCR", retaining five control items. Third, we collapse the original four annotation categories into a binary scheme (*Contentious* vs. *Not contentious*) by discarding labels "I don't know" and "Illegible OCR".[3]

After filtering, the dataset comprises 2,090 excerpts annotated by 19 unique expert IDs and 329 crowdworker IDs, covering 88 distinct target terms. Among these, 1,254 excerpts exhibit complete human consensus (100% agreement), which we use as an unambiguous subset for comparison with model outputs. As shown in Figure 1, the remaining excerpts display varying dominant label proportions (50–99%), capturing subjective or ambiguous cases that are later analyzed in Section 4. We calculated the inter-human agreement of selected excerpts and annotations from ConConCor as our baseline: $\alpha{=}0.578$ among experts and $\alpha{=}0.510$ among crowdworkers.

All 88 target terms are Dutch unigrams sourced from Words Matter (Modest, 2018). They span several semantic domains, including racial and ethnic descriptors, colonial or geographical references, sexual-orientation, occupation, and body-related terms, and neutral alternatives. The interpretation of a term's offensiveness depends on its contextual usage rather than surface form alone, making the corpus suitable for evaluating models' sensitivity to *context-dependent semantics* and *sociohistorical variation in lexical meaning*.

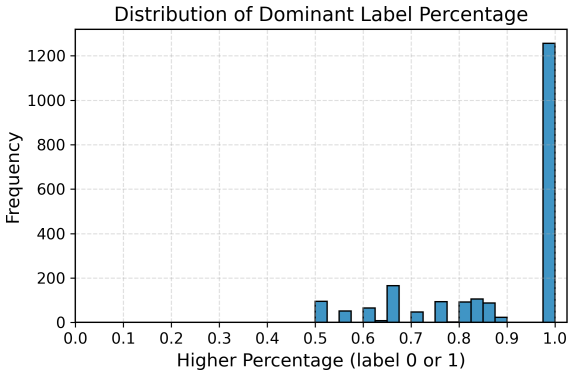

Figure 1: Distribution of dominant label percentage from each sample.

3.2. Experimental Setup

3.2.1. Models.

We evaluate two LLMs capable of understanding and generating Dutch text: GPT-4o (Achiam et al., 2023) and LLaMA-3-70B-Instruct (Grattafiori et al., 2024). These represent high-performance commercial and open-source models, respectively. Both are tested in a zero-shot setting without in-context examples to isolate general linguistic and contextual reasoning ability. We fix decoding parameters across all runs per model (temperature, top-p, and max tokens) for comparability.

[1] https://github.com/YahuiZo/LREC26-Evaluating-Model-Human-Alignment-in-Detecting-Contentious-Language-in-Historical-Corpora

[2] A subset of the Dutch National Library's historical newspaper archive. https://www.delpher.nl/

[3] Original Dutch categories: "Omstreden", "Niet Omstreden", "Weet ik niet", and "Onleesbare OCR".

3.2.2. Prompt Design.

We base all prompts on the original human annotation guidelines from ConConCor to maintain conceptual alignment with human raters. Each prompt consists of a task instruction and a Dutch text sample in which the target term is highlighted. We experiment with eight configurations varying along three controlled dimensions (Hamerlik et al., 2024; Zhao et al., 2021; Madsen et al., 2024; Huang et al., 2023): (i) **Instruction complexity:** original (ori) vs. simplified (simp) task definition; (ii) **Prompt language:** Dutch (nl) vs. English (en), while keeping the sample text in Dutch; (iii) **Output requirement:** label-only (onlyC) vs. label-plus-rationale (CandR). See Appendix A for all prompt template variations. Below is an example of a prompt with the English simplified task definition.

Example Prompt

```
Based on current standards, is
the bolded word **halfbloed**
in the following text delimited
by triple quotes, potentially
offensive, derogatory, or hurtful?
Answer only with Yes or No.
"""Historical Excerpt."""
```

These variations allow us to assess how prompt language, task complexity, and reasoning requirements affect model alignment with human judgments.

3.2.3. Evaluation Scenarios.

We design two complementary scenarios to examine model–human alignment and model consistency.

Scenario 1: Single-run evaluation. Each model produces one prediction per sample under all eight prompt settings, with `temperature`=0 for deterministic decoding. This setting simulates the use of an LLM as an additional annotator. We evaluate its agreement with individual human annotators and with the majority vote using Krippendorff's α and accuracy.

Scenario 2: Multi-run evaluation. Using the best-performing prompt from Scenario 1, each model generates multiple predictions per sample at `temperature`=1 to introduce stochasticity. This setup captures intra-model variability and allows comparison of model and human disagreement patterns, label distributions, and per-term sensitivity.

Implementation details. All experiments were conducted via API calls using identical decoding per model. Each model was queried independently per sample to prevent cross-sample memory effects.

4. Results

We present findings from the two evaluation scenarios introduced in Section 3.2. Scenario 1 evaluates LLMs as potential annotators by comparing their single-run predictions to human judgments, while Scenario 2 examines model variability and contextual sensitivity under multiple runs.

4.1. Agreement and Accuracy in Single-Run Evaluation

Both models achieve agreement levels comparable to human annotators, with GPT-4o slightly outperforming LLaMA-3-70B. Table 1 reports Krippendorff's α and accuracy averaged across all model–prompt configurations. GPT-4o and LLaMA-3-70B reach mean α values of 0.46 and 0.43, respectively, and accuracies around 0.78. The best-performing configuration, English instruction with rationale, achieves $\alpha{=}0.53$ for GPT-4o and $\alpha{=}0.50$ for LLaMA-3-70B, close to the inter-crowdworker baseline ($\alpha{=}0.51$). These results suggest that both models can approximate human-level consistency in identifying contentious terms, though they remain below expert agreement ($\alpha{=}0.58$).

4.2. Effect of Prompt Language, Complexity, and Rationale

Prompt formulation significantly affects model–human alignment. English instructions yield higher agreement than Dutch prompts, likely reflecting the models' training data bias toward English. Including rationales consistently improves agreement, especially for original (non-simplified) task descriptions. Simplified instructions reduce variance across prompt types, indicating that shorter and clearer formulations enhance robustness. As shown in Figure 2, most errors are false positives, revealing a shared tendency across both models to over-predict contentiousness rather than overlook offensive terms. Among all configurations, the English-with-rationale prompt provides the most balanced trade-off between recall and precision and is used for subsequent analyses.

4.3. Performance on Consensus and Non-Consensus Samples

Model alignment increases markedly on unambiguous, high-consensus samples. Table 2 compares model agreement with human annotations across all samples and the subset of 1,254 excerpts with perfect human consensus.

Prompt	GPT-4o		LLaMA 3 70B	
	IAA	Acc.	IAA	Acc.
ori-nl-onlyC	0.45 ±0.02	0.77 ±0.02	0.31 ±0.02	0.65 ±0.02
ori-nl-CandR	0.51 ±0.02	0.81 ±0.02	0.44 ±0.02	0.74 ±0.02
ori-en-onlyC	0.46 ±0.02	0.78 ±0.02	0.30 ±0.02	0.65 ±0.02
ori-en-CandR	**0.53 ±0.02**	**0.84 ±0.02**	0.50 ±0.02	0.79 ±0.02
simp-nl-onlyC	0.42 ±0.02	0.74 ±0.02	0.45 ±0.02	0.83 ±0.02
simp-nl-CandR	0.45 ±0.02	0.77 ±0.02	0.45 ±0.02	0.83 ±0.02
simp-en-onlyC	0.43 ±0.02	0.76 ±0.02	**0.50 ±0.02**	**0.85 ±0.02**
simp-en-CandR	0.42 ±0.02	0.75 ±0.02	**0.50 ±0.02**	**0.85 ±0.02**
Average	0.46	0.78	0.43	0.77

Table 1: IAA and accuracies across all prompts and models. For reference, the IAA scores among human annotators are 0.578 (experts) and 0.510 (crowdworkers). The bold scores are the highest in their columns.

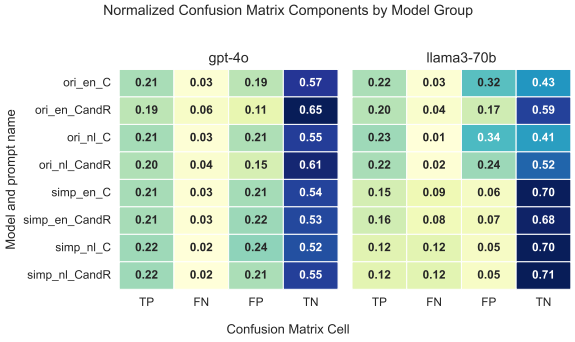

Figure 2: Relative proportions of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) across prompt types for each LLM. For each model, values are normalised so that the four confusion matrix components sum to 1.

On these clear cases, comparing with expert annotators, GPT-4o achieves α=0.97 and 0.99 accuracy, while LLaMA-3-70B reaches α=0.99 and 0.99 accuracy—approaching expert-level reliability. Performance declines on non-consensus samples, where subjectivity and contextual ambiguity dominate. This pattern indicates that while LLMs can replicate human judgments for overtly offensive or neutral terms, they struggle when meaning depends on subtle historical or social context.

4.4. Model Variability, Label Diversity, and Scatter-Level Alignment

LLMs exhibit limited internal variability and tend to produce overconfident, deterministic predictions. In the multi-run setting (`temperature`=1), we assess whether repeated predictions capture human-like disagreement. As shown in Figure 3, human annotations distribute across intermediate label proportions, whereas both models produce strongly polarized outputs (0 or 1). Despite stochastic decoding, neither model reproduces the range of human disagreement, suggesting low epistemic uncertainty and limited sensitivity to borderline cases.

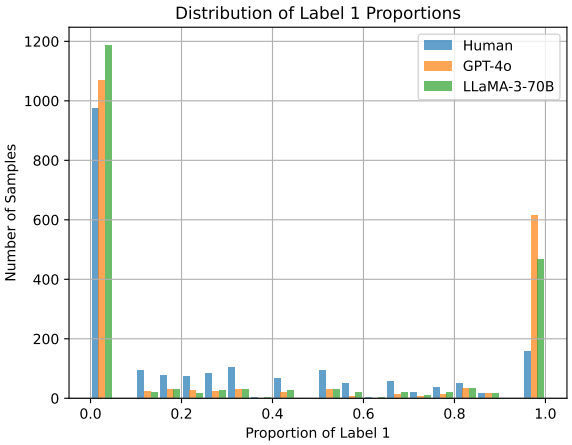

Figure 3: Distribution of contentious label proportions per sample for humans, GPT-4o, and LLaMA-3-70B (`ori-en-CandR`).

To examine this pattern more closely, Figures 4 and 5 show per-sample agreement between human and model predictions. Samples cluster along the extremes, indicating strong alignment on clearly contentious or non-contentious cases, but a higher density appears above the diagonal, showing that both models, particularly GPT-4o, more frequently label samples as contentious than humans do. This confirms an overall over-prediction bias rather than random disagreement.

4.5. Qualitative Analysis of Model–Human Disagreement

To complement the quantitative findings, we conducted a qualitative analysis focusing on (i) how

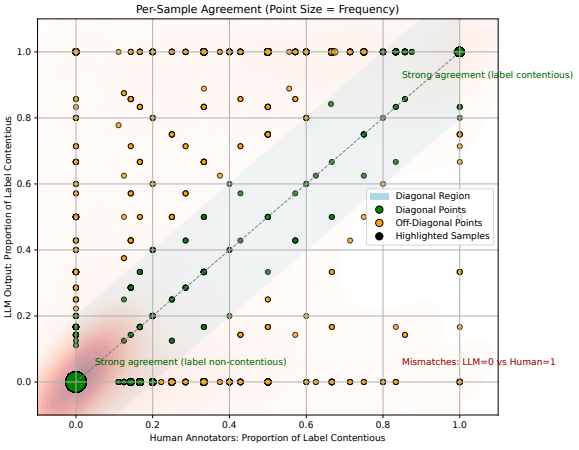

Figure 4: Per-sample agreement between human annotations and GPT-4o outputs (`ori-en-CandR`).

Sample	GPT-4o						LLaMA-3-70B					
	IAA			Acc.			IAA			Acc.		
	All	Exp	Crd	All	Exp	Crd	All	Exp	Crd	All	Exp	Crd
All	0.53 ± 0.02	0.65 ± 0.06	0.52 ± 0.02	0.84 ± 0.02	0.86 ± 0.06	0.83 ± 0.02	0.49 ± 0.02	0.65 ± 0.06	0.49 ± 0.02	0.79 ± 0.02	0.88 ± 0.06	0.79 ± 0.02
Consensus	0.76 ± 0.03	0.97 ± 0.02	0.74 ± 0.03	0.92 ± 0.01	0.99 ± 0.03	0.92 ± 0.02	0.62 ± 0.04	0.99 ± 0.02	0.60 ± 0.04	0.88 ± 0.02	0.99 ± 0.03	0.88 ± 0.02

Table 2: IAA of LLMs with all human annotators, experts, and crowd workers, and accuracy per model (All/Exp/Crd), on all samples and 1254 consensus samples.

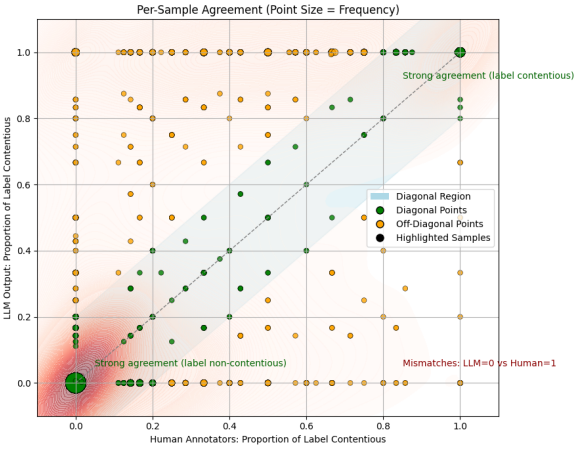

Figure 5: Per-sample agreement between human annotations and LLaMA-3-70B outputs (`ori-en-CandR`).

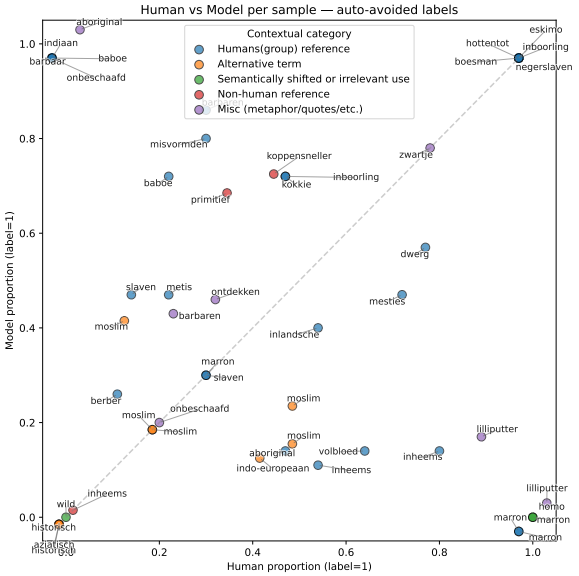

Figure 6: Selected samples in the strata, colored by contextual categories.

disagreement depends on contextual usage of terms and (ii) how it relates to broader lexical domains reflecting historical or social semantics.

4.5.1. Contextual Analysis

To further interpret model–human (dis)agreement patterns, we manually inspected fifty samples drawn from different regions of the scatterplot in Figure 4. Seven strata were defined according to the prediction difference, including exact agreement at both extremes (LLM = 0, human = 0 and LLM = 1, human = 1), two extremes of disagreement (LLM = 0, human = 1 and vice versa), regions of full and near-diagonal agreement ($|\text{LLM} - \text{human}| \leq 0.2$), and intermediate zones around the extremes. Figure 6 visualizes the selected samples in these strata, colored by their contextual category.

The inspected samples were grouped by the contextual use of the target term, using author-defined categories: human (group) reference, non-human reference, alternative term, semantically shifted or irrelevant use, and miscellaneous (e.g., metaphorical or unclear reference) cases. Human references represent the prototypical contentious situations where a term directly denotes a person or group. Terms such as *hottentot* (“Khoikhoi people”) and *boesman* (“Bushman”) occur in the upper-right region of the scatterplot, showing strong model–human agreement on offensiveness. Other terms, including *indiaan* (“Indian”) and *baboe* (“Indonesian female servant”), appear above the diagonal, indicating model oversensitivity, while *inheems* (“indigenous”) and *marron* (“Maroon”) occur below it, reflecting the models’ reduced sensitivity to historical context.

Non-human references, cases where the word refers to an object or abstract entity rather than people, tend to lie above the diagonal. For instance, *primitief* (“primitive”) describes the condition of a plantation, and *koppensneller* (“headhunter”) denotes a type of knife, yet both are often labeled contentious by the models. This suggests that models generalize harmfulness from lexical form rather than contextual meaning. Alternative terms, such as *moslim* (“Muslim”), mostly cluster near the lower-left quadrant, where both models and humans agree on non-contentiousness, though isolated uncertain cases indicate some sensitivity variation.

Semantically shifted or irrelevant uses reveal cases of lexical ambiguity or historical change, for example *marron* meaning “brown” and *homo* meaning “human”. The term *marron* originates

from an older sense of “brown”, which evolved in parallel with the later, contentious meaning “escaped slave”, while *homo* in some excerpts retains its historical meaning “human” rather than the modern, sensitive sense “homosexual”. Humans more often flagged such cases as harmful, implying greater sensitivity to historically charged associations or limited awareness of their older, non-contentious senses. Finally, miscellaneous cases mostly fall along the diagonal, showing strong agreement, with two notable outliers where humans perceived clear harm but the model did not. In these two cases, *lilliputter* (“Lilliputian”) was used metaphorically to describe a small institution or signify how small (all) humans are in comparison to the tower of Pisa.

Overall, the analysis shows that disagreement is systematic rather than random. LLMs align closely with human annotators when lexical offensiveness is explicit, but tend to overgeneralize harmfulness in non-human contexts and underestimate figurative or historically nuanced meanings.

4.5.2. Lexical Domain Analysis

Beyond contextual usage, we also examined the lexical thematic categories of the target terms to capture their historical and social-semantic origins. Each term was assigned to one or more overlapping domains defined by the authors: racial and colonial expressions, value-laden descriptors, dehumanizing or stereotypical terms, social identity labels, and neutral or reclaimed terms, as shown in Table 3.

The overall patterns mirror those observed in the contextual analysis. Racial and colonial terms show the strongest alignment between models and humans, both consistently identify overtly derogatory ethnonyms such as *hottentot* as harmful and recognize neutral forms such as *Aziatisch* as non-harmful. In contrast, value-laden descriptors often yield higher model sensitivity, suggesting overgeneralization of harmfulness of such terms.

In general, we see that model bias is structured by the historical origin and social sensitivity of word types rather than random variation. The interaction between contextual use and lexical domain reveals that disagreement depends jointly on how a term is used and what kind of social meaning it conveys. Models and humans generally align strongly on racial and colonial terms. However, when such terms have undergone semantic shifts, alignment is weaker. Similarly, alignment is weaker when terms are metaphorically used. Besides, value-laden and dehumanizing expressions are judged more harshly by models when stripped of explicit group reference, implying that lexical associations override contextual cues. Overall, model bias appears contextually modulated rather than purely lexical: models capture overt lexical offensiveness but lack the contextual flexibility that humans apply to historically and socially charged language.

Category	Example terms	Share (%)
Racial and colonial expressions	boesman (Bushman), inheems (indigenous), ontdekken (discover), slaven (slaves)	58%
Social identity labels	moslim (Muslim), homo (homosexual), baboe (Indonesian female servant)	15%
Value-laden descriptors	wild (wild), primitief (primitive), onbeschaafd (uncivilised), barbaren (barbarians)	12%
Dehumanizing or stereotypical descriptors	zwartje (diminutive form of zwart (black)), slaven (slaves), lilliputter (Lilliputian), misvormden (the deformed), homo (homosexual)	18%
Neutral or reclaimed terms	historisch (historical), aziatisch (Asian), inheems (indigenous), moslim (Muslim)	21%

Table 3: Lexical domain categories of the analyzed Dutch terms with representative examples (English translations in parentheses) and the proportion of terms assigned to each category. Totals may exceed 100% because some terms belong to multiple categories.

4.6. Term-Level Sensitivity and Bias Patterns

Building on the qualitative observations, we quantify model sensitivity across all 88 target terms. Figure 7 presents the proportion of samples labeled *contentious* for each term by human annotators, GPT-4o, and LLaMA-3-70B using the best-performing prompt configuration (*ori-en-CandR*). Terms are ordered by the proportion of human majority votes, providing a direct comparison of relative sensitivities.

Across the vocabulary, both models generally predict a higher proportion of contentious cases than humans, confirming the qualitative finding of model over-sensitivity. GPT-4o matches human proportions for 33 terms, assigns lower rates for 11, and higher rates for 44, while LLaMA-3-70B aligns on 23 terms, assigns lower rates for 13, and higher rates for 52. LLaMA-3-70B is therefore slightly more conservative in its predictions but also more likely to over-predict contentiousness overall. Despite these differences, both models exhibit high mutual consistency, sharing identical rates for 39

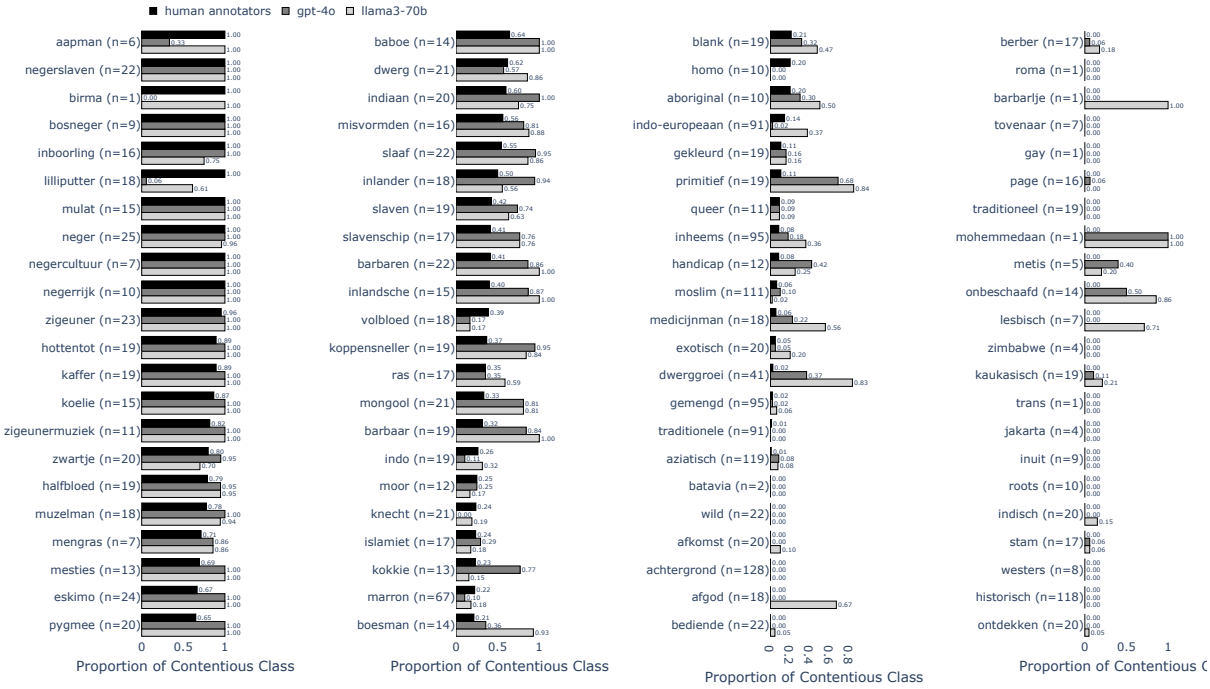

Figure 7: Proportion of samples labeled contentious for each target term by humans and models (`ori-en-CandR`).

terms.

Figure 7 highlights several consistent patterns. Highly offensive racial and colonial expressions—such as *kaffer* ("kaffir"), *hottentot*, and *negerrijk* ("empire of negroes")—are unanimously identified as contentious by both models and human annotators, reflecting strong alignment on explicit harms. Similarly, neutral alternatives like *achtergrond* ("background"), *traditioneel* ("traditional"), and *historisch* are consistently judged non-contentious, yielding few false positives. By contrast, religious and occupational designations (e.g., *koelie* ("coolie"), *muzelman* ("Muslim")) and outdated ethnonyms for Indigenous peoples (e.g., *eskimo* ("Eskimo"), *indiaan*, *aboriginal* ("Aboriginal")) are more frequently judged contentious by both models, indicating heightened model sensitivity to lexical cues of social identity. Conversely, context-dependent terms such as *lilliputter*, *volbloed* ("full-blooded"), *knecht* ("servant"), *marron*, and *homo* are underpredicted, suggesting that models struggle to account for pragmatic and historical nuance. LLaMA-3-70B tends to be slightly more sensitive across categories, while GPT-4o is comparatively conservative, resulting in smaller variance across terms.

Overall, the term-level analysis corroborates the contextual and lexical findings. LLMs reliably identify overtly harmful expressions but remain limited in capturing subtle or contextually modulated meanings. They tend to over-generalize from lexical form and under-estimate the interpretive flexibility that human annotators apply to historically or socially situated language.

In general, across both evaluation scenarios, large language models show strong alignment with human annotators for explicit and unambiguous cases but diverge when contextual interpretation is required. Their predictions are consistent across prompt variations and reach agreement levels comparable to human crowd annotators, yet they remain overconfident and less adaptable to context. Overall, the results indicate that model sensitivity is influenced by the presence of sensitive terms, particularly for racial and colonial terms, while human annotators display greater flexibility in interpreting meaning based on pragmatic and historical context.

5. Discussion

Our findings show that large language models can reach human-level reliability when judging clear-cut cases of contentious language but remain inconsistent in context-dependent settings that have a higher subjectivity level. Both GPT-4o and LLaMA-3-70B perform comparably to human annotators on unambiguous terms but fall short of agreement when interpretation depends on historical usage, speaker intention, or figurative meaning. This result aligns with prior work in contextual toxicity and bias detection, where models often rely on lexical patterns such as racial or gendered markers rather than reasoning over pragmatic or situational context (Kennedy et al., 2020; Abid et al.,

2021; Park et al., 2018).

The over-prediction of harmfulness for identity-related and colonial terms suggests that model sensitivity is dominated by token-level associations learned from contemporary data rather than from the historical framing of the corpus. Conversely, the consistent under-prediction for semantically shifted terms such as *marron* and *homo* indicates limited ability to recognize diachronic meaning change. These findings point to a systematic gap in how current models represent temporal semantics and interpret language with culturally specific connotations.

For GLAM institutions and other domains involving archival material, this means that LLMs can assist in identifying overtly harmful vocabulary but should not be relied upon for contextually or historically ambiguous cases. Integrating model predictions as part of a semi-automated workflow—where LLMs flag potential cases, and experts make final judgments, can balance scalability with interpretive accuracy. Future work should focus on improving calibration and contextual reasoning, for example, through multi-turn prompting, few-shot contrastive examples, or uncertainty-aware ensemble annotation frameworks.

6. Conclusion

This study investigated the potential of large language models to assist in annotating contentious terminology in historical Dutch texts, using 2,090 excerpts from the ConConCor corpus. GPT-4o and LLaMA-3-70B achieved agreement comparable to human annotators on clear cases but were less reliable for context-dependent judgments requiring cultural or temporal interpretation. Both models effectively recognized overtly offensive racial and colonial terms but tended to overlook figurative or historically shifted meanings, reflecting reliance on surface lexical cues rather than contextual reasoning. These results suggest that LLMs can support scalable first-pass annotation of sensitive language in GLAM collections, while human expertise remains essential for contextual validation. Future work should explore multilingual and diachronic modeling to better capture semantic change and contextual variation over time.

Acknowledgment

This research is funded by the Netherlands Organization for Scientific Research under the HAICu project (NWO NWA # 1518.22.105)

Ethics Statement

This work engages with sensitive historical language that may include offensive or discriminatory terms. All excerpts are drawn from publicly available archival sources and reproduced solely for research and analysis. The goal of this study is not to automate moral judgment but to investigate whether LLMs can assist human annotators in identifying potentially harmful language within historical materials. We emphasize that deploying such systems should include human oversight and contextual review, particularly in cultural heritage and educational applications.

Broader impacts of this research include promoting transparency and accountability in the digitization and annotation of archival materials. At the same time, we acknowledge that models trained on large-scale data may reproduce social and historical biases. By comparing model predictions with human annotations, our work aims to highlight rather than conceal such discrepancies and to provide evidence that can inform ethical curation practices. Future research should continue to engage historians, linguists, and affected communities to ensure that automated tools for heritage data respect cultural context, interpretive diversity, and contemporary ethical standards.

Limitations

A key limitation of this study concerns diachronic bias, referring to the mismatch between the historical language of our corpus and the largely contemporary data used to train modern LLMs. Although the models we employed include data up to 2024, their training emphasizes present-day language use and social norms, which may diverge from the meanings and connotations of terms in texts from 1890 to 1941. This temporal gap may lead models to project modern interpretations onto historically situated expressions, resulting in over- or underestimation of contentiousness depending on how a term's meaning has evolved.

Our study also focuses exclusively on Dutch historical materials, which limits both linguistic and cultural generalizability. To mitigate this, we relied on annotations from both expert and crowd annotators and used contextual excerpts rather than isolated words, allowing both human and model judgments to account for surrounding discourse. We evaluated two representative models, GPT-4o and LLaMA-3-70B, in a zero-shot setting to assess alignment without additional fine-tuning, although results may differ for other languages or domain-specific corpora and other LLMs.

Because the notion of contentiousness is inherently subjective and context-dependent, we com-

pared model outputs with both expert and crowd annotations and analyzed performance across consensus and ambiguous cases. Finally, our qualitative analysis covered a limited number of disagreement samples rather than exhaustive coverage. Future work should extend this framework to multilingual and multi-period corpora, incorporate temporally aware models that explicitly account for semantic change, and explore human-in-the-loop annotation pipelines to improve reliability and interpretability.

7. Bibliographical References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, An drey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao,

Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Csilla E. Ariese. 2020. Amplifying voices: Engaging and disengaging with colonial pasts in amsterdam. *Heritage & Society*, 13(1-2):117–142.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. *arXiv preprint arXiv:2410.12029*.

Ryan Brate, Andrei Nesterov, Valentin Vogelmann, Jacco Van Ossenbruggen, Laura Hollink, and Marieke Van Erp. 2021. Capturing contentiousness: constructing the contentious terms in context corpus. In *Proceedings of the 11th Knowledge Capture Conference*, K-CAP '21, page 17–24, New York, NY, USA. Association for Computing Machinery.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using large language model annotations for valid downstream statistical inference in social science: Design-based semi-supervised learning. *arXiv preprint arXiv:2306.04746*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Endre Hamerlik, Marek Šuppa, Miroslav Blšták, Jozef Kubík, Martin Takáč, Marián Šimko, and Andrej Findor. 2024. ChatGPT as your n-th annotator: Experiments in leveraging large language models for social science text annotation in Slovak language. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 81–89, Vienna, Austria. Association for Computational Linguistics.

Sanne Hoeken, Sophie Spliethoff, Silke Schwandt, Sina Zarrieß, and Özge Alacam. 2023. Towards detecting lexical change of hate speech in historical data. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 100–111, Singapore. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Eirini Kaldeli, Orfeas Menis-Mastromichalakis, Spyros Bekiaris, Maria Ralli, Vassilis Tzouvaras, and Giorgos Stamou. 2021. Crowdheritage: crowdsourcing for improving the quality of cultural heritage metadata. *Information*, 12(2).

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. Crowdcalibrator: Can annotator disagreement inform calibration in subjective tasks? *arXiv preprint arXiv:2408.14141*.

Lucija Krusic. 2024. Constructing a sentiment-annotated corpus of Austrian historical newspapers: Challenges, tools, and annotator experience. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 51–62, Miami, USA. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Jason Liartis, Kristina Rose, Antoine Isaac, and Giorgos Stamou. 2025. Don't erase, inform! detecting and contextualizing harmful language in cultural heritage collections. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21836–21850, Vienna, Austria. Association for Computational Linguistics.

Wayne Modest. 2018. *Words Matter: an unfinished guide to word choices in the cultural sector*. Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum.

Andrei Nesterov, Laura Hollink, Marieke van Erp, and Jacco van Ossenbruggen. 2023. A knowledge graph of contentious terminology for inclusive representation of cultural heritage. In *The Semantic Web*, pages 502–519, Cham. Springer Nature Switzerland.

Andrei Nesterov, Laura Hollink, and Jacco van Ossenbruggen. 2024. How contentious terms about people and cultures are used in linked open data. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4523–4533, New York, NY, USA. Association for Computing Machinery.

Temi Odumosu. 2020. The crying child: On colonial archives, digitization, and ethics of care in the cultural commons. *Current Anthropology*, 61(S22):S289–S302.

Giulia Osti and Elizabeth Russey Roke. 2024. Collaborating for change? assessing metadata inclusivity in digital collections with large language models (llms). In *2024 IEEE International Conference on Big Data (BigData)*, pages 2479–2488. IEEE.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. 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.

Vithya Yogarajan, Gillian Dobbie, Tim Pistotti, Joshua Bensemann, and Kobe Knowles. 2023. Challenges in annotating datasets to quantify bias in under-represented society. *ArXiv*, abs/2309.08624.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A. Prompt Design

This appendix lists the prompt templates used in our experiments.

Prompt type	Prompt template
ori-nl-CandR	`In deze annotatietaak vragen we u om te beoordelen of u de vetgedrukte term **term** omstreden vindt (contentious in het Engels) in de context van de tekst die eromheen staat, afgebakend door drievoudige aanhalingstekens. Voor het doel van deze taak beschouwen we een term als omstreden wanneer het volgens u mogelijk beledigend, denigrerend, kwetsend, of om een andere reden niet passend is naar huidige maatstaven. Om uw oordeel te vellen kunt u zich bijvoorbeeld afvragen of u de term in een soortgelijke zin zou gebruiken en/of u verrast zou zijn als de term op deze manier gebruikt zou worden door anderen. Als u een term in de gegeven zin een klein beetje omstreden vindt, antwoord dan 'omstreden naar huidige maatstaven'. Zo niet, antwoord dan 'niet omstreden'. En geef de reden.` `"""Historical Excerpt."""`
ori-nl-onlyC	`In deze annotatietaak vragen we u om te beoordelen of u de vetgedrukte term **term** omstreden vindt (contentious in het Engels) in de context van de tekst die eromheen staat, afgebakend door drievoudige aanhalingstekens. Voor het doel van deze taak beschouwen we een term als omstreden wanneer het volgens u mogelijk beledigend, denigrerend, kwetsend, of om een andere reden niet passend is naar huidige maatstaven. Om uw oordeel te vellen kunt u zich bijvoorbeeld afvragen of u de term in een soortgelijke zin zou gebruiken en/of u verrast zou zijn als de term op deze manier gebruikt zou worden door anderen. Als u een term in de gegeven zin een klein beetje omstreden vindt, antwoord dan 'omstreden naar huidige maatstaven'. Zo niet, antwoord dan 'niet omstreden'. Geef geen reden.` `"""Historical Excerpt."""`
simp-nl-CandR	`Naar huidige maatstaven, is het vetgedrukte woord **term**, in de onderstaande context afgebakend door drievoudige aanhalingstekens, mogelijk beledigend, denigrerend of kwetsend? Antwoord met Ja of Nee en geef een reden.` `"""Historical Excerpt."""`
simp-nl-onlyC	`Naar huidige maatstaven, is het vetgedrukte woord **term**, in de onderstaande context afgebakend door drievoudige aanhalingstekens, mogelijk beledigend, denigrerend of kwetsend? Antwoord alleen met Ja of Nee.` `"""Historical Excerpt."""`

Table 4: Dutch prompt templates used in the experiments.

Prompt type	Prompt template
ori-en-CandR	`In this annotation task, we ask you to judge whether you find the bolded term **term** contentious in the context of the surrounding text, delimited by triple quotes. For the purposes of this task, we consider a term to be contentious when you think it is potentially offensive, derogatory, hurtful, or otherwise inappropriate by current standards. In making your judgment, you might ask yourself, for example, whether you would use the term in a similar sentence and/or whether you would be surprised if the term were used in this way by others. If you find a term in the given sentence slightly contentious, answer 'contentious according to current standards'. If not, answer 'not contentious'. And give a reason. """Historical Excerpt."""`
ori-en-onlyC	`In this annotation task, we ask you to judge whether you find the bolded term **term** contentious in the context of the surrounding text, delimited by triple quotes. For the purposes of this task, we consider a term to be contentious when you think it is potentially offensive, derogatory, hurtful, or otherwise inappropriate by current standards. In making your judgment, you might ask yourself, for example, whether you would use the term in a similar sentence and/or whether you would be surprised if the term were used in this way by others. If you find a term in the given sentence slightly contentious, answer 'contentious according to current standards'. If not, answer 'not contentious'. Don't give a reason. """Historical Excerpt."""`
simp-en-CandR	`Based on current standards, is the bolded word **term** in the following text delimited by triple quotes potentially offensive, derogatory, or hurtful? Answer Yes or No and give a reason. """Historical Excerpt."""`
simp-en-onlyC	`Based on current standards, is the bolded word **term** in the following text delimited by triple quotes potentially offensive, derogatory, or hurtful? Answer only with Yes or No. """Historical Excerpt."""`

Table 5: English prompt templates used in the experiments.