Prompting Across Time: Evaluating LLMs on Historical and Contemporary Offensive Language

Sanne Hoeken[1], **Sophie Spliethoff**[2], **Silke Schwandt**[2], **Sina Zarrieß**[1] **and Özge Alaçam**[1]
[1]Computational Linguistics, Dept. of Linguistics, Bielefeld University
[2]Dept. of History, Faculty of History, Philosophy and Theology, Bielefeld University
{sanne.hoeken, sophie_jasmin.spliethoff, silke.schwandt, sina.zarriess, oezge.alacam}@uni-bielefeld.de

Abstract

Research on hate speech detection (HSD) has centered on modern data, even though offensive language has a much longer history. This paper presents the first systematic evaluation of instruction-tuned LLMs on Early Modern English invectives, compared with a modern hate-speech benchmark. Our work applies a modular prompt design to measure the contribution of definitional richness, contextual grounding, decision rules and few-shot examples. The results indicate that clearer annotation boundaries in the curated historical corpus lead to higher classification performance compared to the modern benchmark, despite the disadvantage of linguistic unfamiliarity. Prompt brittleness, however, persists across both domains. Classification-oriented components (rules, examples) drive the strongest effects, while definitional or contextual additions matter less. Fine-tuned encoder models still outperform LLMs, but some prompt configurations can narrow the gap. Overall, our study provides practical guidance for prompt design in both digital humanities and HSD and new opportunities for tracing the historical development of hate speech.[1]

1 Introduction

Existing research on large language models (LLMs) and hate speech detection (HSD) has focused primarily on contemporary online discourse (Alkomah and Ma, 2022). Yet, offensive language is a long-standing linguistic phenomenon that historians have been studying in earlier periods too, such as in Early Modern English (henceforth **EModE**) texts (Suerbaum, 2015; Steckel, 2018) (Figure 1). In this paper, we propose to widen the scope of Natural Language Processing (NLP) research on hate speech beyond the contemporary domain and examine it across different periods of time. This not only creates opportunities for digital humanities (DH) research but also enables deeper insights into the modeling of hate speech and how language changes over time.

Figure 1: Examples of offensive language across time. A 16th century polemical text and a modern tweet illustrate how offensive speech manifests in different historical and modern contexts.[2]

Detecting hate speech is challenging even in present-day data. Historical corpora, in turn, introduce additional difficulties. EModE differs markedly from present-day English in vocabulary, orthography and pragmatic conventions (Nevalainen, 2000; Gramley and Gramley, 2024). In addition, LLMs are tuned primarily on modern data and consequently may exhibit stronger biases toward modern usage patterns (Alam et al., 2024). Moreover, EModE texts can only be fully understood and interpreted in their historical social contexts. This makes offensiveness harder to recognize without domain-specific expertise. Effective prompting may therefore require different guidance than is typically needed for modern data.

At the same time, historical corpora are typically smaller and carefully curated by a limited number of experts (Hiltunen et al., 2017), which often yields higher annotation consistency. Modern

[1]Code and supplementary materials are available at our GitHub repository: https://github.com/SanneHoeken/HSDAcrossTime

[2]The EModE passage taken from the InviTE corpus (Spliethoff et al., 2025) includes the phrase "shameles shauelings", a religious slur targeting Catholic clergymen.

Findings of the Association for Computational Linguistics: ACL 2026, pages 33295–33314

hate-speech datasets, however, are usually larger and crowd-sourced, annotated under comparatively minimal guidelines and by annotators from diverse backgrounds (Jahan and Oussalah, 2023). These differences in dataset properties are likely to interact with model behavior. Expert-curated corpora may produce clearer class boundaries, whereas greater subjectivity and more heterogeneous annotations in modern hate speech make classification more difficult (Sachdeva et al., 2022). Concrete illustrative examples from the empirical data used in this study are provided in Appendix B, highlighting both linguistic variation and differences in corpus characteristics.

A further challenge is prompt brittleness. As extensively investigated, LLM behavior is strongly shaped by prompt design where small changes in wording or structure can lead to substantial shifts in output (Mizrahi et al., 2024; Ngweta et al., 2025). Although prompt engineering is now an active area of research (Li, 2023; Liu et al., 2023), less is known about how particular definitional components shape model behavior in the domains of historical language and HSD. Given these observations, we investigate two main research questions:

1. How well do instruction-tuned LLMs detect offensive language in EModE, and how does their performance relate to a modern hate-speech benchmark?
2. How do individual prompt components (detailed definitions, contextual information, decision rules and examples) affect performance across historical and modern datasets?

To address these questions, we present the first systematic evaluation of multiple instruction-tuned LLMs on EModE offensive language, evaluated alongside a modern hate-speech benchmark. Building on a modular prompt design that is grounded in a conceptual taxonomy of hate speech definition elements (Melis et al., 2025), our analysis quantifies the marginal contribution of each prompt component. Interestingly, our results show that models perform better on historical EModE data than on the contemporary hate-speech benchmark. This trend suggests that the benefits of clearer annotation boundaries in curated historical corpora outweigh the challenges posed by the linguistic unfamiliarity. We further find that prompt brittleness is equally persistent across time periods, where prompt components that explicitly target class boundaries tend to exert the strongest effects. Overall, these findings suggest a superficial temporal robustness: while LLMs can perform reasonably well on historical language, persistent prompt brittleness indicates that this does not reflect a deep understanding of EModE. Nevertheless, our study provides practical guidance for prompt design in both DH and HSD research and may open up new opportunities for tracing the historical development of hate speech.

2 Related Work

While research on both HSD as well as NLP for historical texts (Piotrowski, 2012) has advanced, the two areas have remained largely separate. This section reviews these strands of related work which together provide the background for our study.

2.1 NLP for historical texts

The growing availability of large-scale digitized historical corpora has enabled new computational approaches within DH. Instead of relying solely on close reading, researchers have begun to apply NLP methods for studying language across time (e.g., Koncar et al. (2020); Pawłowski and Walkowiak (2024)). The creation of annotated corpora for language modeling purposes requires expert oversight as historical distance entails substantial shifts in language and cultural context. Consequently, historical corpora with lexical or semantic annotations often offer greater consistency and theoretical grounding than most contemporary crowd-sourced datasets (Ehrmann et al., 2020; Al-Laith et al., 2023; Dejaeghere et al., 2024).

To bridge the linguistic gap between historical and modern English, several domain-adapted encoder models have been introduced (Qiu and Xu, 2022; Harju and van der Goot, 2025), such as MacBERTh (Manjavacas Arevalo and Fonteyn, 2021), which is trained on EModE too. These models have shown to outperform contemporary BERT-like models on downstream tasks such as part-of-speech tagging or word sense disambiguation in historical text. Related lines of work have examined lexical semantic change (Schlechtweg et al., 2020) and sentiment analysis (Al-Laith et al., 2024; Dejaeghere et al., 2024), although all focus on 19th- and 20th-century data. Closer to the domain of our study, Hoeken et al. (2023) proposed detecting hateful lexical change in pre-modern sources using a word-level approach with MacBERTh, though their evaluation remained limited in scale.

Recent work has begun to incorporate LLMs into DH research, for example for word usage generation (Cassotti and Tahmasebi, 2025), multilingual news article extraction (Oberbichler et al., 2025) and knowledge evaluation (Hauser et al., 2024). These studies report mixed results. While LLMs can capture certain aspects of meaning in historical language, they often lag behind expert-level interpretation and can display period- or regional-specific biases. Moreover, Hiltmann et al. (2025) emphasize that effective prompting for humanities requires the integration of historical context. Yet, nearly all such studies focus on later centuries and leave the Early Modern period underexplored.

A step toward filling this gap was made by Spliethoff et al. (2025), who introduced the InviTE corpus, the first expert-annotated dataset of invective language in EModE (language with the potential to disparage an individual or group). Their baseline experiments showed that MacBERTh outperforms zero-shot LLMs. Our study builds directly on this line of work by systematically comparing LLMs across historical and modern hate speech data.

2.2 Hate Speech Detection and Prompting

In contrast, hate speech detection (HSD) in NLP has focused almost exclusively on present-day social media discourse (Albladi et al., 2025). Research has produced a wide range of datasets, typically collected through crowd-sourcing and annotated for hate speech, abusive language or offensiveness (e.g. Zampieri et al. (2019); Mathew et al. (2021)). A persistent challenge in this domain is subjectivity. Interpretations of what constitutes offensive content vary widely, which leads to low inter-annotator agreement and fuzzy class boundaries (Fortuna et al., 2020; Sachdeva et al., 2022).

Transformer-based models have long dominated supervised approaches (Liu et al., 2019; Sarkar et al., 2021), but more recent studies have shifted toward instruction-tuned LLMs (Plaza-del arco et al., 2023; Pan et al., 2024; Roy et al., 2023). Prompting has become a central strategy for leveraging LLMs without task-specific fine-tuning. Yet, research has shown that prompt design is often brittle, where small changes in instruction wording can lead to large performance differences (Mizrahi et al., 2024; Ngweta et al., 2025; Liu et al., 2024). In the context of offensive language, Melis et al. (2025) introduced a taxonomy of Conceptual Elements (CEs) that aggregates definitions of hate speech from the literature. This taxonomy structures the definitional space into three layers: (1) Foundational Elements specifying essential components of a hate speech definition (e.g. form of communication); (2) Extensive Elements that refine these foundations with additional granularity (e.g. lists of targeted attributes); and (3) Accessory Elements, such as illustrative examples. While they also explored prompting strategies for LLMs, their experiments are not presented in a clearly modular way. We adopt their framework and extend it in three ways: (1) we apply it to both historical and modern offensive language, (2) we incorporate not only definitional components but also genre-specific information about the data and (3) introduce a modular evaluation method that quantifies the contribution of individual prompt components.

3 Data

To investigate how models handle offensive language across time, we compile datasets representing both historical and contemporary contexts. Our study focuses on invective language in Early Modern English (EModE) as a historical analogue to modern offensive language. In the following we describe the corpora used in our experiments.

3.1 The InviTE corpus

The InviTE corpus, introduced by Spliethoff et al. (2025), is a collection of almost 2,000 sentences drawn from EModE texts dating between 1485 and 1603, with a particular focus on Reformation discourse. Yet the corpus spans a wide range of genres (sermons, medical texts, royal proclamations, poetry, and stage plays) and covers multiple registers, distinct author writing styles, and different dialects of English, at a time when the language was not yet standardized and was undergoing constant change.

Each sentence is annotated with detailed information concerning the presence and nature of invective language. All annotations were produced by two expert annotators following carefully curated guidelines. The annotation scheme follows a hierarchical structure. The category of **invectivity** distinguishes between sentences containing invective language (INV) and those without (NON). For sentences marked as invective, two further layers of annotation are applied. The **type** of invective identifies whether it is expressed literally (LIT) or metaphorically (MET). And the **target** of invective is annotated according to categories tailored to the Reformation context: sinful behavior (SIN),

political-religious misconduct (POL), religious belief (REL), confession (CON), or other (OTH).

The dataset also includes rich metadata for each sentence, such as the publication year and information about the author, including gender and religious confession. For our study, publication years were grouped into four periods, informed by historical expertise. An overview of the distribution of sentences across annotation and metadata categories is provided in Appendix A.

3.2 The HateXplain corpus

To compare model performance on historical and modern offensive language, we additionally draw on the HateXplain dataset (Mathew et al., 2021), a benchmark covering contemporary online hate speech. HateXplain contains roughly 20K posts from Twitter and Gab collected using hate speech lexicons. Twitter posts were sampled from January 2019 to June 2020 and Gab posts from October 2016 to June 2018. The primary annotation comprised a three-class hate speech classification task (*hate*, *offensive*, or *normal*). Additionally, each post is annotated with the target community and rationales, i.e. the portions of the post on which the labeling decision is based. All annotations were performed by approx. 250 crowd-sourced workers.

For our experiments, to create a setting directly analogous to the InviTE experiments, we took the following steps. First, we converted the original three-class scheme into a binary setting by merging the *hate* and *offensive* categories into a single 'Offensive' class and leaving *normal* as 'Non-offensive'. We then aggregated the annotations using the majority vote to obtain a single gold label per instance. We recognize that recent trends in HSD favor more perspectivist approaches (Frenda et al., 2025). We use this aggregation to align with the historical dataset, yet we also examine annotator disagreement in relation to model behavior (5.1). Finally, we sampled 2K instances (1K posts from Gab and 1K from Twitter) with a comparable label distribution (25% offensive vs. 75% non-offensive). More details on the distribution of the sampled subset across the labels and other categories can again be found in Appendix A.

4 Experiments

We conduct experiments to evaluate detection of offensive language in two domains: (1) **historical** invectives in EModE texts (InviTE), and (2) **modern** offensive language in social media posts (HateXplain). For both corpora we test models on the binary task of classifying sentences/posts into invective/offensive or non-invective/non-offensive.

4.1 Models

Baselines We reproduced the BERT-based results on the InviTE dataset reported in Spliethoff et al. (2025). Specifically, we fine-tuned `BERT-base` (Devlin et al., 2019), `XLM-RoBERTa-large` (Conneau et al., 2019), and `MacBERTh` (Manjavacas Arevalo and Fonteyn, 2021) using the same 10-fold cross-validation setup with stratified folds. For the HateXplain subset we apply the same baselines, replicating the experiments with the same pretrained models and fine-tuning strategy on this dataset.

LLMs We selected eight instruction-tuned models which cover four prominent families (LLaMA, OLMo, Qwen, and Gemma) with parameter counts ranging from one to eight billion, chosen to balance architectural diversity, computational feasibility and accessibility to the research community (more detailed rationale in Section 6). These models differ not only in size but also in architecture and training data. Details on the specific pretrained versions are provided in Appendix C.

4.2 Prompt design

We designed our prompts building upon the framework proposed by Melis et al. (2025) (a taxonomy of Conceptual Elements (CE) for defining hate speech; see 2.2) and applied the same principles to both datasets. To maintain comparability, we ensured that prompts for both datasets use analogous content and as similar as possible formulations that differ only where necessary to reflect domain-specific terminology or examples. All of our prompts start with a task description which asks the model to decide whether a given sentence contains invective (InviTE) or offensive language (HateXplain), and end by presenting the sentence to be classified. Between the task description and the target sentence, prompts may incorporate detailed definitions, contextual information, decision rules, and illustrative examples. These components serve as modular building blocks for constructing the full set of prompt variants.

Foundational Elements For InviTE, invective language is defined as "*all utterances that have the potential to disparage an opponent person or*

group". This formulation captures the key dimensions of invective language as defined in the expert-based annotation guidelines of the InviTE corpus (Spliethoff et al., 2025) in a compact way. This core **Definition (D)** includes mentions of the Form of Communication ("all utterances"), Problematic Content ("to disparage"), and Target ("an opponent person or group"). These three elements (or CEs) together provide the minimal basis for invective language, which overlap with the concept of Offensive Language. Consequently, for HateXplain we use the same core definition but replace the term *invective* with *offensive*.

Extensive Definitions of the Foundational Elements An extended version of the Definition (**Dext**) adds clarification on the possible targets, corresponding to Addressed Attributes in CE terms. Specifically, for InviTE, these include the targets as provided in the InviTE annotation scheme (e.g. sinful behavior or confession). For HateXplain, the attributes are updated to reflect contemporary relevant categories (e.g. sexual orientation or race), as also inferred from the dataset's annotations.

Accessory Elements Beyond this definitional core, we introduce accessory elements. The **Context (C)** element provides (genre-specific) information on the historical or contemporary setting. For InviTE, it specifies that the texts originate from 1485–1603 in England, often within religious and political conflicts of the Reformation era, whereas for HateXplain it notes that the posts come from Twitter and Gab between 2016–2020.

The **Rules (R)** element incorporates decision rules to guide consistent annotation, particularly for reported or quoted and untargeted language. The **Examples (E)** element provides four labeled instances (two per class), which turns the setup into a few-shot setting. The same examples were used for all input instances (i.e. static demonstrations). For InviTE, the invective examples were drawn from the dataset's annotation guidelines and augmented with two non-invective examples. Together, the four examples: cover both classes, target both Catholics and non-Catholics, represent multiple time periods, and include both metaphorical and non-metaphorical instances. For HateXplain, we likewise ensured that the four example cover both classes (two per class), include instances from both platforms/time bins and reflect variation in target categories.

While Rules and Examples have close analogues in the Accessory Elements of Melis et al. (2025) (Exceptions and Examples, respectively), the Context represents a novel extension motivated by the potential need to provide the model with period-specific information.

Given our focus on two key aspects affecting LLM performance in this study, language familiarity (the degree to which a model's training data resembles the target domain) and class separability (how easy offensive vs. non-offensive instances can be discriminated in the data), we further distinguish between components that primarily support understanding of the linguistic content or context (Extended Definition and Context) versus those that primarily guide classification decisions (Rules and Examples). Rules explicitly delineate conditions for negative labels and Examples concretely illustrate the contrast between classes.

Prompt construction The modular construction enables us to systematically vary prompt design. Starting from the minimal combination of task and definition (D), we incrementally add definition extensions (ext) and accessory elements (C, R, E). This results in a structured set of sixteen prompt variants for each dataset and allows us to investigate how definitional richness, contextual grounding, decision rules and few-shot examples influence LLM performance in both historical and modern domains. Full texts of the prompt templates are provided in Appendix C.

4.3 Evaluation

We tested all model-prompt combinations on both datasets (implementation details are in Appendix C). In addition to extracting the models' predictions and reporting performance, we also conducted confidence extraction. For the LLMs, model confidence was computed as the softmax probability of the first generated token corresponding to the predicted class token (Xia et al., 2025). For the BERT-based models, confidence was obtained from the classification head by applying a softmax over the output logits and taking the maximum probability corresponding to the predicted class. Although confidence measures are not fully comparable across the two model types, they provide useful insights into model behavior and result robustness.

To quantify the contribution of each prompt component beyond the definition (D) as baseline, we compute an **Average Marginal Contribution**

	macro f1				avg conf.
	mean	*min*	*max*	*std*	*mean*
InviTE					
BERT-base	0.83	-	-	-	0.95
XLM-R	0.86	-	-	-	0.95
MacBERTh	0.89	-	-	-	0.97
Llama-3.1 (8B)	0.73	0.64	0.81	0.05	0.89
Llama-3.2 (3B)	0.63	0.46	0.74	0.10	0.90
OLMo-2 (7B)	0.65	0.58	0.72	0.04	0.96
OLMo-3 (7B)	0.75	0.69	0.79	0.03	0.97
Qwen2 (7B)	0.80	0.77	0.82	0.01	0.97
Qwen3 (4B)	0.81	0.76	0.84	0.02	0.99
gemma-3 (1B)	0.54	0.27	0.65	0.11	0.97
gemma-3 (4B)	0.74	0.62	0.80	0.06	0.99
HateXplain					
BERT-base	0.67	-	-	-	0.88
XLM-R	0.60	-	-	-	0.82
MacBERTh	0.64	-	-	-	0.90
Llama-3.1 (8B)	0.54	0.41	0.66	0.09	0.87
Llama-3.2 (3B)	0.58	0.47	0.65	0.06	0.88
OLMo-2 (7B)	0.51	0.45	0.58	0.04	0.91
OLMo-3 (7B)	0.55	0.48	0.62	0.05	0.93
Qwen2 (7B)	0.47	0.36	0.58	0.08	0.98
Qwen3 (4B)	0.49	0.41	0.57	0.05	0.99
gemma-3 (1B)	0.33	0.26	0.39	0.04	0.99
gemma-3 (4B)	0.50	0.35	0.65	*0.13*	0.99

Table 1: Performance (macro F1) and prediction confidence (average across all predictions) on InviTE and HateXplain. BERT-model results are single-run values. LLM results are aggregated over 16 prompt variants.

(AMC). This Shapley-value-inspired estimate of each component's contribution represents the average improvement in performance across all tested combinations (Horovicz and Goldshmidt, 2024; Liu et al., 2024). We denote our set of components beyond the baseline as $N = \{\text{ext}, \text{C}, \text{R}, \text{E}\}$, and $v(S)$ represents the model performance (e.g. macro avg F1 score) when using the prompt subset $S \subseteq N$ together with D. For a component $i \in N$, let P_i be the set of subsets $S \subseteq N \setminus \{i\}$ for which both S and $S \cup \{i\}$ were tested. The AMC of component i is defined as:

$$\text{AMC}_i = \frac{1}{|P_i|} \sum_{S \in P_i} \big(v(S \cup \{i\}) - v(S)\big) \quad (1)$$

Practically, a positive AMC_i means that adding component i tends to increase macro-F1, a negative value decreases it; e.g. $\text{AMC}_i = 0.1$ reflects an average macro-F1 gain of 0.1.

5 Results & Discussion

In this section, we first examine classification performance across all models and both datasets (5.1), then analyze the contributions of individual prompt components (5.2), and finally explore how these results vary across different sentence types (5.3).

	Accuracy		Avg conf.	
	Una.	*Non*	*Una.*	*Non*
Llama-3.1 (8B)	0.64	0.52	0.87	0.86
Llama-3.2 (3B)	0.71	0.52	0.90	0.87
OLMo-2 (7B)	0.68	0.38	0.93	0.90
OLMo-3 (7B)	0.69	0.44	0.94	0.92
Qwen2 (7B)	0.65	0.34	0.98	0.97
Qwen3 (4B)	0.65	0.36	0.99	0.99
gemma-3 (1B)	0.51	0.22	0.99	0.99
gemma-3 (4B)	0.66	0.40	0.99	0.99

Table 2: Accuracy and average confidence for unanimous (Una.) and non-unanimous (Non) predictions by LLMs (mean across all 16 prompts) for HateXplain.

5.1 Classification performance

Table 1 summarizes the macro F1 and confidence scores of all models on the InviTE and HateXplain datasets. The baselines are reported as single-run results and LLMs with aggregated statistics across the 16 prompt variants.

Models perform better on InviTE, with encoders leading. Interestingly, models achieve clearly higher scores on InviTE than on HateXplain. Across both datasets, encoder-based baselines generally outperform LLMs. On InviTE, MacBERTh reaches the highest score (0.89), and the strongest LLMs, Qwen2 and Qwen3, perform up to 0.82–0.84. On HateXplain, BERT-base achieves the strongest result (0.67), and the best LLMs (with the best prompt), Llama models and gemma-3 (4B), reach similar levels (up to 0.65–0.66). The high confidence of LLMs indicates that their predictions in our experiments are relatively robust, even if not fully deterministic. Prediction confidence further supports the hypothesis that HateXplain poses a more difficult classification problem. All three finetuned encoder models exhibit an average confidence decrease of 7% between InviTE and HateXplain. This likely reflects the greater complexity in class discrimination due to more heterogeneous annotations, as discussed in the introduction.

To further investigate the role of label ambiguity in HateXplain, we conducted an additional error analysis that splits model performance based on whether annotators unanimously agreed on the majority label. Table 2 reports mean accuracy and mean confidence across the 16 prompt variants for each LLM. Across all models, accuracy is substantially higher on examples with unanimous anno-

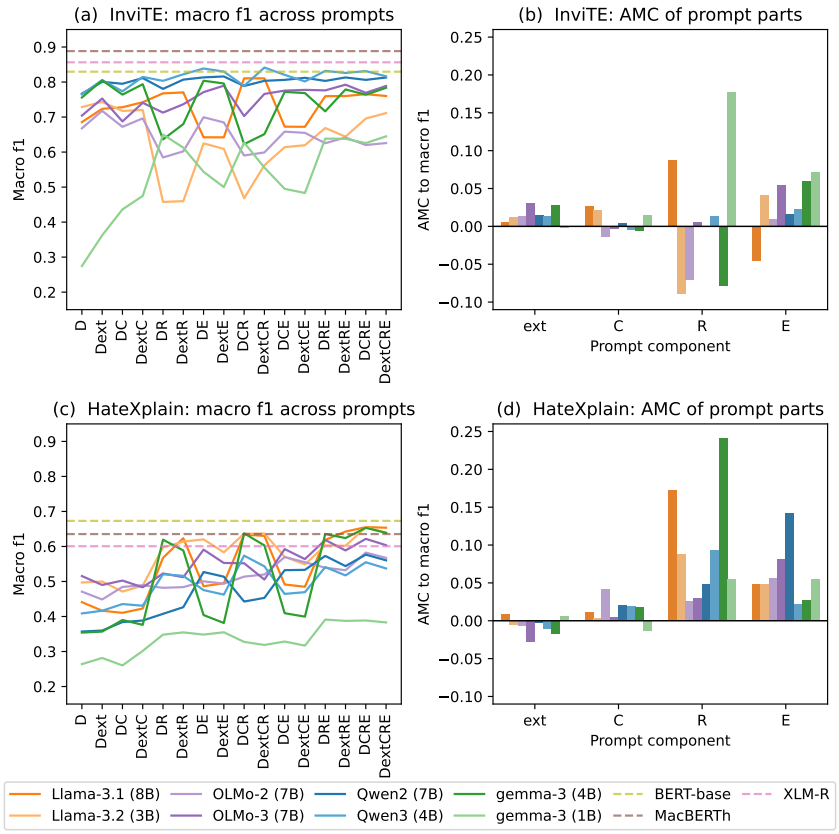

Figure 2: (a) Classification performance (Macro F1) of all models on InviTE across 16 prompt combinations and baselines (horizontal lines). (b) Average Marginal Contribution (AMC) of prompt components to Macro F1 on InviTE. (c) and (d): Same plots for HateXplain.

tator agreement, which confirms that clearly labeled instances are easier to predict. Average prediction confidence, however, remains largely unchanged between unanimous and non-unanimous cases. This indicates that model uncertainty is not aligned with human disagreement.

Prompt brittleness persists across time periods. Variation across prompts furthermore underlines the instability of LLM performance. Standard deviations (up to 0.13) reflect substantial differences in F1 and indicate how sensitive models are to prompt changes. While the results do not point to a clear pattern between prompt sensitivity and performance, nor across model size or family, the extent of variation is similar across both datasets. The mean standard deviations on InviTE (0.05 across all models) and HateXplain (0.07) and mean ranges between minimum and maximum values (0.17 vs. 0.19) are comparable.

5.2 Prompt component contributions

Figure 2 visualizes model performance across the 16 prompt variants and the Average Marginal Contribution (AMC) of individual prompt components The performance plots reiterate the brittleness of LLMs discussed above as scores fluctuate considerably across prompts on both datasets. The AMC analysis provides a clearer view of how specific prompt components (see 4.2) affect performance.

Components aiding classification matter more than those aiding contextual understanding. Overall, the **ext**ended definition and **C**ontext have relatively modest effects, whereas **R**ules and **E**xamples exert much stronger influence. On InviTE, **ext** yields small but mostly positive contributions (up to +0.03), while on HateXplain it is slightly negative (–0.02). This suggests that more content-rich definitions are more beneficial when models are less familiar with the phenomenon, as is the case for historical invectives. Surprisingly, **C** shows the opposite trend. Contributions are more consistently positive on HateXplain (up to +0.04) but on InviTE extra situational context does not consistently help.

Unlike the small effects of **ext** and **C**, the rules yield mostly strong positive contributions, especially on HateXplain (up to +0.24). Examples also

boost performance in both datasets (up to +0.07 on InviTE and +0.14 on HateXplain). Overall, the results indicate that the effects of prompt components vary across models and between historical and modern settings. Yet, both datasets reveal a consistent contrast in magnitude between components that aid language familiarity (**ext** and **C**) and those that focus on class boundaries (**R** and **E**).

5.3 Results across (metadata) categories

The InviTE corpus provides rich metadata that enables a more fine-grained analysis of model behavior across different types of sentences.

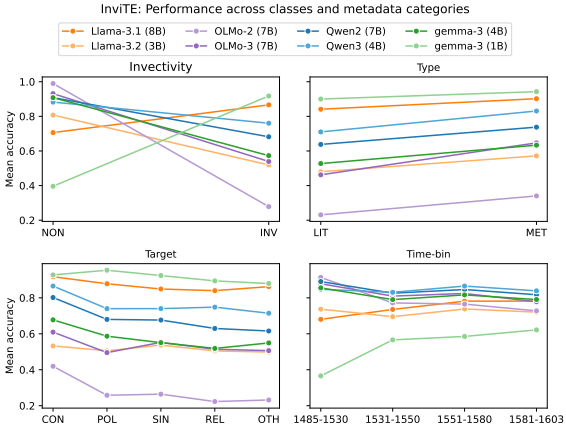

Figure 3: Performance (mean accuracy across all prompt variants) across Invectivity classes and metadata categories (Type, Target and Time-bin) for InviTE.

Improved performance via linguistic shortcuts? Figure 3 reports mean accuracy across all prompt variants for invectivity classes as well as metadata categories. For most models, classification is more accurate for non-invective than for invective sentences. Regarding the different expression **types** of invectives, all models perform worse on invective language annotated as literal (LIT) than as metaphorical (MET). Models may confuse metaphorical expressions with invectives, likely because figurative expression is so prevalent in EModE polemics (Smith, 2014; Razzall, 2021). This points to reliance on recurring linguistic patterns rather than true semantic understanding.

For invective **targets**, performance differences are comparatively small across the five annotated categories. The only consistent tendency is slightly higher accuracy for Confession (CON), which may be explained by the presence of clearer lexical cues, such as historical variants of slurs targeting confessional groups (e.g. "shauelings" in Figure 1; see also Appendix B). Unlike modern hate

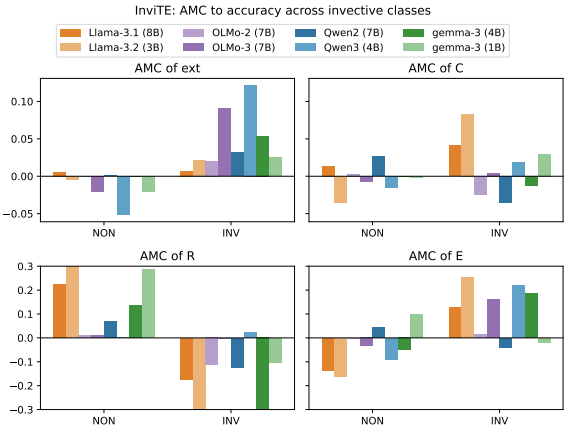

Figure 4: AMC of prompt components to accuracy across Invective and Non-invective classes for InviTE.

speech, where slurs can appear in reclaimed contexts, these historical terms are largely unambiguous, as modern-style reclamation has not been observed. This likely makes detection easier.

Lastly, for **time-bins**, we expected later texts to be easier for LLMs to classify, since EmodE gradually converges toward modern spelling and syntax (Gramley and Gramley, 2024). This expectation is confirmed only for Llama-3.1 8B and Gemma-3 1B, which show relatively improved performance on later texts. However, several stronger models show the opposite pattern, with slightly but consistently higher accuracy on the earliest period.

Component impact depends on sentence type. We next examine how specific prompt components affect accuracy across invective and non-invective sentences (Figure 4). The results reveal that the impact of components can differ substantially between classes, often in opposite directions. In particular, the **R**ules component substantially improves performance on non-invective sentences while decreasing performance on invective sentences. This pattern makes sense as the rules specify conditions for applying the negative class. Conversely, the **ext**ended definition tends to improve performance on invective sentences but not on non-invective ones. This effect also aligns with expectations as the definitions explicitly highlight potential targets of invective language.

Similarly, the **E**xamples component improves performance of invective sentences more than non-invective ones. **C**ontext shows a more variable pattern. Together, these results underscore that prompt components do not uniformly influence all sentence types, a pattern that is also evident in our analyses

across the two offensiveness classes in HateXplain (see Figure 8 in Appendix D). Specifically, in contrast to InviTE, all models perform better on the offensive class than on the non-offensive class in HateXplain. However, similar to InviTE, rules (R) tend to increase performance on non-offensive instances while decreasing performance on offensive ones, consistent with the rules' focus on exceptions (negative cases). Likewise, examples (E) show a positive effect on the class that models generally struggle with most (non-offensive for HateXplain; invective for InviTE), while generally decreasing performance on the other class. This implies that prompt analyses should extend beyond overall accuracy and examine class- or category-specific effects, as such fine-grained analyses can reveal patterns hidden by aggregate metrics and consequently guide more targeted prompt design.

6 Conclusion

This paper presents the first systematic evaluation of instruction-tuned LLMs for detecting offensive language in a historical setting, using a corpus of Early Modern English invectives (InviTE) alongside a contemporary hate-speech benchmark (HateXplain). By directly comparing the two domains under matched definitions we explored how dataset characteristics interact with model behavior, and how prompt design modulates these effects.

Interestingly, models achieve consistently higher performance on InviTE than on HateXplain. Together with analyses of model confidence and prompt component effects, this suggests that the benefit of clearer class boundaries due to the expert-curated annotations in InviTE outweighs the disadvantage of its greater linguistic unfamiliarity. In contrast, modern hate-speech data, although presumably closer to models' training distribution, remains more difficult to classify likely because of its more heterogeneous boundaries. At the same time, additional analyses across sentence types indicate that historical data may also contain surface patterns, such as unambiguous lexical cues and prevalent metaphor use, that ease classification.

We further find that prompt brittleness is not confined to contemporary data but persists comparably across both periods. Nonetheless, fine-tuned encoder models generally still outperform the (mid-sized) LLMs on both datasets. Yet, the right prompt can push an LLM to nearly match this performance. Prompt engineering thus remains crucial in both modern and historical settings but cannot fully close the gap with specialized encoders.

Taken together, our results raise deeper questions about what it means for LLMs to "understand" historical language and invective. Stronger performance may partly reflect reliance on surface patterns rather than true semantic understanding, and persistent prompt sensitivity further suggests limited robustness in model comprehension. Future work could explore these issues more rigorously, e.g. by comparing non-invective metaphorical language with invective expressions, or by applying interpretability techniques to trace how models make classification decisions in historical contexts.

In sum, our study highlights both the potential and the limitations of instruction-tuned LLMs on historical language. By widening the scope of hate speech detection beyond present-day discourse, we provide practical guidance for NLP and DH researchers on modeling strategies and prompt design, but also open up new opportunities to trace the development of offensive language across time.

Limitations

While our study provides systematic insights into detecting offensive and invective language across historical and contemporary contexts, several limitations constrain the scope and generalizability of our findings. Our study evaluates eight LLMs spanning four different families to provide substantial architectural diversity. Nevertheless, our experiments are limited to models of up to 8B parameters. This restriction was due to available computational resources and reflects the practical realities of research under typical academic budgets. Our focus on models that remain accessible to a broader research community avoids reinforcing a culture in which meaningful contributions require exceptional resources, while also mitigating the field's growing environmental footprint.

Our experimental setup relies on a single operational definition of offensive and invective language, derived from the InviTE annotation guidelines and adapted for the HateXplain dataset. While this definition provides conceptual coherence across historical and modern domains, alternative definitional perspectives may yield different model behaviors and should be explored in future work. At the same time, the adopted definition instantiates all foundational elements identified (as characteristic of offensive language definitions) in

the framework of Melis et al. (2025). Since this framework synthesizes over 20 definitions from academic literature, social media and international policy documents, the alignment of our definition with the framework's conceptual elements mitigates concerns about arbitrariness.

Further, the conceptual equivalence we assume between invective and offensive language is approximate. Moreover, the InviTE corpus primarily includes invective language in the context of the Protestant Reformation and may only partially reflect modern conceptions of offensiveness.

While the InviTE corpus is not homogeneous and spans a wide range of genres (including sermons, medical texts, royal proclamations, poetry and stage plays) and encompasses multiple registers, distinct author styles and different dialects of English during a period of linguistic change, the datasets employed in this study are relatively small, containing approximately 2,000 instances per domain. The historical corpus primarily focuses on Early Modern English Reformation discourse and may not generalize to other genres or historical periods. Yet, our study goes beyond specific performance outcomes by presenting a methodological framework for comparative analysis. Our structured prompt design, procedures for establishing cross-domain comparability, and advanced evaluation approach together provide a foundation for subsequent studies to examine generalization across larger datasets, different historical periods and broader genres.

Finally, we aggregate the crowd-sourced annotations in the HateXplain dataset using majority votes to align with the expert-labeled historical InviTE corpus. Yet, we do not fully ignore annotator label variation. Additional analysis shows that model accuracy is higher on instances with unanimous agreement and confirms that although majority-vote aggregation yields a practical representation of modern hate speech, it does not encompass the full range of annotator perspectives. A perspectivist interpretation is not available for the historical dataset and consequently, while our approach enables a meaningful comparison across time, it necessarily simplifies the variability inherent in both modern and historical hate speech.

Ethical Statement

This research engages with sensitive content both in historical texts and contemporary social media posts. We recognize that the use and analysis of such data entails potential ethical considerations.

The InviTE corpus consists of 16th century language samples, where no personal privacy concerns arise. The HateXplain dataset comprises posts from publicly accessible social media platforms, including Twitter and Gab, and all user identifiers have been anonymized to protect privacy.

Our work involves exposure to offensive language, which may be upsetting to readers. This applies to modern language just like pre-modern invective language as intra-religious conflicts remain relevant in present-day societies. In the paper, we take care to avoid reproducing offensive content unnecessarily.

We believe that, despite these challenges, our study contributes to a deeper understanding of offensive language across social and historical contexts. This, in turn, advances computational methods that may eventually support efforts to mitigate the spread of hate speech.

Acknowledgments

The authors acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of North Rhine-Westphalia, Germany.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Ali Al-Laith, Kirstine Nielsen Degn, Alexander Conroy, Bolette Sandford Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment classification of historical Danish and Norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial*

Abstracts, pages 27–33, St. Julian's, Malta. Association for Computational Linguistics.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access*, 13:20871–20892.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6).

Pierluigi Cassotti and Nina Tahmasebi. 2025. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Tess Dejaeghere, Pranaydeep Singh, Els Lefever, and Julie Birkholz. 2024. Exploring aspect-based sentiment analysis methodologies for literary-historical research purposes. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 129–143, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language resources for historical newspapers: the impresso collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Stephen Gramley and Vivian Gramley. 2024. *The History of English. An Introduction*, 3rd edition. Routledge, London/ New York.

Anika Harju and Rob van der Goot. 2025. How to age BERT well: Continuous training for historical language adaptation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 258–267, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and R. Maria del Rio-Chanona. 2024. Large Language Models' Expert-level Global History Knowledge Benchmark (HiST-LLM). *Advances in Neural Information Processing Systems*, 37:32336–32369.

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach. *Preprint*, arXiv:2502.04351.

Turo Hiltunen, Joe McVeigh, and Tanja Säily. 2017. How to turn linguistic data into evidence? In *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, Studies in Variation, Contacts and Change in English. University of Helsinki, Helsinki.

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.

Miriam Horovicz and Roni Goldshmidt. 2024. TokenSHAP: Interpreting large language models with Monte Carlo shapley value estimation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 1–8, Miami, FL, USA. Association for Computational Linguistics.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Philipp Koncar, Alexandra Fuchs, Elisabeth Hobisch, Bernhard C. Geiger, Martina Scholger, and Denis Helic. 2020. Text sentiment in the age of enlightenment: an analysis of spectator periodicals. *Applied Network Science*, 5(1):33.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Hanxi Liu, Xiaokai Mao, Haocheng Xia, Jian Lou, Jinfei Liu, and Kui Ren. 2024. Prompt valuation based on shapley values. *Preprint*, arXiv:2312.15395.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computation Surveys*, 55(9).

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

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.

Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521, Vienna, Austria. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Terttu Nevalainen. 2000. *Early Modern English Lexis and Semantics*, page 332–458. The Cambridge History of the English Language. Cambridge University Press.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs robustness to changes in prompt format styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 529–537, Albuquerque, USA. Association for Computational Linguistics.

Sarah Oberbichler, Johanna Mauermann, The Trung Tran, and Carlos-Emiliano González-Gallardo. 2025. Studying Model Design Biases in LLMs for Multilingual Historical Newspaper Extraction; The Messina Earthquake Case Study. In *The 29th International Conference on Theory and Practice of Digital Libraries*, Tempere, Finland.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *Computer Modeling in Engineering & Sciences*, 140(3):2849–2868.

Adam Pawłowski and Tomasz Walkowiak. 2024. NLP for digital humanities: Processing chronological text corpora. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 105–112, Miami, USA. Association for Computational Linguistics.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Wenjun Qiu and Yang Xu. 2022. Histbert: A pre-trained language model for diachronic lexical semantic analysis. *CoRR*, abs/2202.03612.

Lucy Razzall. 2021. *Boxes and Books in Early Modern England: Materiality, Metaphor, Containment*. Cambridge University Press.

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.

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

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Helen Smith. 2014. Metaphor, cure, and conversion in early modern england*. *Renaissance Quarterly*, 67(2):473–502.

Sophie Spliethoff, Sanne Hoeken, Silke Schwandt, Sina Zarrieß, and Özge Alaçam. 2025. The invite corpus: Annotating invectives in tudor english texts for computational modeling. *Preprint*, arXiv:2509.22345.

Sita Steckel. 2018. Verging on the Polemical. Towards an Interdisciplinary Approach to Medieval Religious Polemic. *Medieval Worlds*, 7:2–60.

Almut Suerbaum. 2015. Language of Violence. Language as Violence in Vernacular Sermons. In Almut Suerbaum, George Southcombe, and Benjamin Thompson, editors, *Polemic. Language as Violence in Medieval and Early Modern Discourse*, pages 125–148. Routledge, London / New York.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396, Vienna, Austria. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

A Data

Our study uses the InviTE corpus (Spliethoff et al., 2025) and the HateXplain dataset (Mathew et al., 2021), in accordance with their intended use cases, ethical guidelines, and applicable licenses.

Tables 3 and 4 summarize the distribution of sentences across annotation and meta-data categories in the InviTE and HateXplain corpora, respectively.

InviTE The periodization of publication years into time-bins was guided by the expertise of the second author, a specialist in Tudor history. The time-bins reflect key historical phases:

- **1485–1530:** Pre-Reformation era.
- **1531–1550:** Henry VIII's break with the Pope and the emergence of reformist thought and the Anglican Church.
- **1551–1580:** Catholic counter-Reformation under Mary I and the subsequent restoration of Anglicanism under Elizabeth I.
- **1581–1603:** Consolidation of anti-Catholic sentiment, particularly after the defeat of the Spanish Armada in 1588.

HateXplain For HateXplain, we collapsed the fine-grained target group labels provided in the original dataset (e.g., *women*, *gay*, *Jewish*) into broader meta-categories to match the level of abstraction used in the InviTE corpus. Specifically, targets were grouped into Sexual Orientation & Gender (SEX), Race/Ethnicity (RAC), Religion (REL), and Other (OTH).

Time-bins for HateXplain were automatically inferred based on the platform metadata. Posts from Gab cover the period 2016–2018, while Twitter posts cover 2019–2020. This division was used to assign each post to a corresponding time-bin.

		Count	Avg. sent. length
Invective	NON	1415	40.0
	INV	560	51.3
Type	NON	1415	40.0
	LIT	330	46.4
	MET	218	58.9
	UN	12	49.0
Target	NON	1415	40.0
	SIN	181	61.5
	REL	107	49.4
	CON	100	45.7
	POL	93	56.5
	OTH	79	31.6
Time-bin	1485-1530	444	51.9
	1531-1550	314	44.4
	1551-1580	613	40.7
	1581-1603	604	38.7
TOTAL			43.2

Table 3: Distribution of sentences across annotation and meta-data categories in the InviTE corpus

		Count	Avg. sent. length
Offensive	NON	1500	24.1
	OFF	500	20.4
Target	SEX	578	21.3
	RAC	513	23.5
	OTH	490	24.0
	REL	419	24.6
Time-bin	2016-2018	1000	28.2
	2019-2020	1000	18.2
TOTAL			23.2

Table 4: Distribution of categories across the sampled HateXplain data

B Illustrative Examples

This section provides representative sample texts from each dataset to illustrate linguistic variation over time and context.

InviTE (historical texts).

> (1) InviTE (1548): "and then will god awake as a firce lyon against those cruel wolues which deuoure hys lambes, and will playe with the hypocrites, [...]."
>
> (2) InviTE (1590): "that al the nobilitie are hereticke, that they are enemies to the popular estate: that they oppose themselues against your purposes: and that they must be rooted out."

These examples show metaphorical (e.g. "wolues") and explicit invectives (e.g. "hereticke") typical of historical texts.

HateXplain (modern social media).

> (3) HateXplain (2019/2020): "yes im talking to this bitch myself"
>
> (4) HateXplain (2019/2020): "i for reals walk by this group of freshman and this bitch really pulled out a big ass bag of hot cheetos"

In modern social media, the term "bitch" appears in both offensive and non-offensive contexts. One example (3) is labeled offensive by a majority of annotators, while the other (4) is unanimously non-offensive. Terms that have traditionally been considered derogatory may be reappropriated or employed playfully in modern contexts, resulting in increased contextual ambiguity and greater annotator disagreement. Consequently, this poses additional challenges for automatic detection of hate speech.

Taken together, these examples illustrate diachronic change in the expression of hate speech: from explicit, metaphorical invectives in historical texts to nuanced, context-dependent uses in modern social media.

C Model Implementation Details

All experiments were implemented in Python 3.9 using the `PyTorch` and `Hugging Face Transformers` libraries.

Model	Hugging Face URL
BERT-base	https://huggingface.co/google-bert/bert-base-uncased
XLM-RoBERTa-large	https://huggingface.co/FacebookAI/xlm-roberta-large
MacBERTh	https://huggingface.co/emanjavacas/MacBERTh
LLaMA-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
LLaMA-3.2-3B-Instruct	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
Ministral-8B-Instruct-2410	https://huggingface.co/mistralai/Ministral-8B-Instruct-2410
OLMo-2-1B-Instruct	https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct
OLMo-2-7B-Instruct	https://huggingface.co/allenai/OLMo-2-1124-7B-Instruct
OLMo-3-7B-Instruct	https://huggingface.co/allenai/Olmo-3-7B-Instruct
Qwen2-7B-Instruct	https://huggingface.co/Qwen/Qwen2-7B-Instruct
Qwen3-4B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507
gemma-3-1b-it	https://huggingface.co/google/gemma-3-1b-it
gemma-3-4b-it	https://huggingface.co/google/gemma-3-4b-it

Table 5: Pretrained models used in our experiments.

Models Table 5 lists the pretrained encoder models and LLMs used in our experiments.

Prompts Figures 5 and 6 display the most elaborate prompt templates used. **Content warning!** The examples in the prompts contain language that may be offensive; we recognize their potential harm.

Experimental setup For the encoder-based models, we followed the same implementation setup as described in Spliethoff et al. (2025), to which we refer for training and configuration details.

For the LLMs, we employed the `transformers.pipeline` interface for text generation using default parameters and model-specific tokenizers. Each model was queried with chat-style prompts and a limited generation length of 20 new tokens. All experiments were executed on a single NVIDIA RTX A6000 GPU using CUDA acceleration.

Post-processing Generated outputs were post-processed to remove the original prompt, cleaned of extraneous characters (e.g. whitespace, quotes, asterisks), and mapped to the label sets ('Invective'/'Non-invective' for InviTE, 'Offensive'/'Non-offensive' for HateXplain). Outputs not matching these labels were flagged as invalid predictions (See Table 7).

The number of invalid predictions for Llama-3.1-8B-Instruct is noteworthy. We inspected these cases. The majority correspond to model refusals, e.g.: "I can't provide a classification of this sentence as it contains derogatory language." This behavior is only observed for Llama-3.1-8B-Instruct among the models tested and also appears sensitive to prompt formulation (257 with DC prompt vs. only 39 with DextCRE). This safety-driven refusal behavior and its dependence on prompt design would be interesting to further explore in future work.

Two independent runs We conducted two independent runs of all experiments using identical settings; the second run additionally included confidence extraction. The results reported in this paper are from the second run. Across all models, the second run closely reproduces the results of the first run, with negligible performance differences (first-run F1 scores shown in Table 6, mostly within ±0.00–0.01), consistent with the high confidence values observed. Exceptions were Ministral 8B and OLMo-2 1B, which exhibited low confidence scores (average 64–65%) and consequently unstable results; these models were therefore not considered in the main paper, with OLMo-2 1B replaced by the newly released OLMo-3 7B.

	InviTE				HateXplain			
	mean	*min*	*max*	*std*	*mean*	*min*	*max*	*std*
Baselines								
majority	0.42	–	–	–	0.43	–	–	–
random	0.48	–	–	–	0.46	–	–	–
BERT-base	0.83	–	–	–	0.67	–	–	–
XLM-R	0.82	–	–	–	0.62	–	–	–
MacBERTh	0.89	–	–	–	0.64	–	–	–
LLMs (across 16 prompt variants)								
Llama-3.1 (8B)	0.73	0.63	0.81	0.05	0.54	0.40	0.66	0.09
Llama-3.2 (3B)	0.62	0.46	0.73	0.09	0.57	0.47	0.64	0.06
Ministral (8B)	0.66	0.50	0.75	0.08	0.57	0.46	0.65	0.06
OLMo-2 (1B)	0.45	0.41	0.52	0.03	0.54	0.43	0.58	0.04
OLMo-2 (7B)	0.65	0.59	0.72	0.04	0.52	0.45	0.58	0.04
Qwen2 (7B)	0.80	0.77	0.82	0.01	0.47	0.36	0.58	0.08
Qwen3 (4B)	0.81	0.77	0.84	0.02	0.49	0.41	0.57	0.05
gemma-3 (1B)	0.53	0.27	0.65	*0.11*	0.34	0.26	0.39	0.04
gemma-3 (4B)	0.74	0.63	0.80	0.06	0.50	0.35	0.65	*0.13*

Table 6: Performance results (macro F1) of all models on InviTE and HateXplain. Baselines are single-run values. LLMs are aggregated over 16 prompt variants.

D Results across (metadata) categories

Figures 7 and 8 provide additional results for the InviTE and HateXplain datasets, respectively. Each figure consists of a 5x3 grid of plots: the first row shows mean model accuracy across different prompt variants, while the subsequent rows display the Average Marginal Contribution (AMC) of individual prompt components to accuracy. On the x-axis, each plot represents either classes (invective or offensive) or metadata categories, while the y-axis shows mean accuracy (first row) or AMC values (subsequent rows). These figures illustrate how model performance varies across different classes and metadata categories, and how specific prompt components contribute to accuracy on sentences belonging to each category.

E Use of AI Assistants

AI assistants were used during the preparation of this work, limited to paraphrasing for clarity and style and to assist with code primarily for visualization purposes. No AI-generated content was used for experimental design or drawing conclusions.

	D	DC	DCE	DCR	DCRE	DE	DR	DRE	Dext	DextC	DextCE	DextCR	DextCRE	DextE	DextR	DextRE
InviTE																
Llama-3.1 (8B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Llama-3.2 (3B)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
OLMo-2 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OLMo-3 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Qwen2 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Qwen3 (4B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gemma-3 (1B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gemma-3 (4B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HateXplain																
Llama-3.1 (8B)	189	257	104	148	47	137	216	140	222	187	73	109	39	130	155	84
Llama-3.2 (3B)	16	1	0	21	0	1	46	0	4	1	1	9	0	0	32	0
OLMo-2 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OLMo-3 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Qwen2 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Qwen3 (4B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gemma-3 (1B)	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
gemma-3 (4B)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7: Number of invalid predictions for each LLM and prompt combination.

Prompt for InviTE classification

```
### Task
Decide whether the given sentence contains invective language or not. Answer
with only one word: ‘Invective’ or ‘Non-invective’.

### Definition
Invective language includes all utterances that have the potential to disparage
an opponent person or group.
(ext) Invective language can be targeted at attributes, such as religious
belief and confession, or acts, such as sinful behaviour or political-religious
misconduct.

### Context
The sentence appears in English texts from 1485–1603, often in diverse genres
tied to Reformation-era religious or political conflicts in England and Europe.

### Rules
Additional rules to follow:
1. Whenever a possible invective sentence refers to an ancient event, or is a
quote from an ancient text or the bible that is not directly used to disparage
an opponent of the author, it should be classified as ‘Non-invective’.
2. Normative sentences that name sinful behaviour but do not inveigh a specific
opponent should be classified as ‘Non-invective’.

### Examples
Invective:
1. ‘A proclamation for resisting and withstanding of most damnable heresies,
sown within this realm, by the disciples of Luther and other heretics, perverters
of Christ’s religions.’
2. ‘Otherwise, we will betray the sheep of Christ into the mouths of ravenous
wolves, and especially into the teeth of that great, rampaging lion – the
devil – who never ceases going about, seeking whom he may devour for his prey.’
Non-invective:

1. ‘Of the greatest benefites, that euer god gaue me, is, that he sent me so
sharpe and seuere parentes, and so ientle a scholemaster.’
2. ‘For thy goodnes haue merci on my grete wyckednes , as certeynly as i was
neuer so wycked as y u arte good ne neuer may be thoughe i wolde.’

### Sentence to classify
SENTENCE
```

Figure 5: Most elaborate prompt template (**DextCRE**) used for the invective classification task. It includes the core task description, a definition of invective language (D, extended version (ext)), historical context (C), annotation rules (R), and labeled examples (E). This template represents the full modular prompt and serves as the basis from which simpler prompt variants are derived.

Prompt for HateXplain classification

```
### Task
Decide whether the given sentence contains offensive language or not. Answer
with only one word: ‘Offensive’ or ‘Non-offensive’.

### Definition
Offensive language includes all utterances that have the potential to disparage
an opponent person or group.
(ext) Offensive language can be targeted at attributes, such as sexual
orientation, race or religion.

### Context
The sentence appears on social media platforms such as Twitter and Gab between
2016–2020, often in short post formats such as tweets.

### Rules
Additional rules to follow:
1. Whenever a possible offensive sentence is a quote from social media, news,
or other public discourse that is not directly used to disparage a specific
individual, it should be classified as ‘Non-offensive’.
2.    Posts that name undesirable behavior or controversial opinions
without directly targeting a specific individual should be classified as
‘Non-offensive’.

### Examples
Offensive:
1. ‘<user> shut up nazi’
2. ‘whats scarier than hearing a white person talking in french i just know
they saying something racist or islamophobic’

Non-offensive:
1. ‘border patrol does not want to keep being labeled as nazis when they are
doing their jobs’
2. ‘omg we just had a monster queers on american idol lmfao’

### Sentence to classify
SENTENCE
```

Figure 6: Most elaborate prompt template (**DextCRE**) used for the invective classification task. It includes the core task description, a definition of invective language (D, extended version (ext)), historical context (C), annotation rules (R), and labeled examples (E). This template represents the full modular prompt and serves as the basis from which simpler prompt variants are derived.

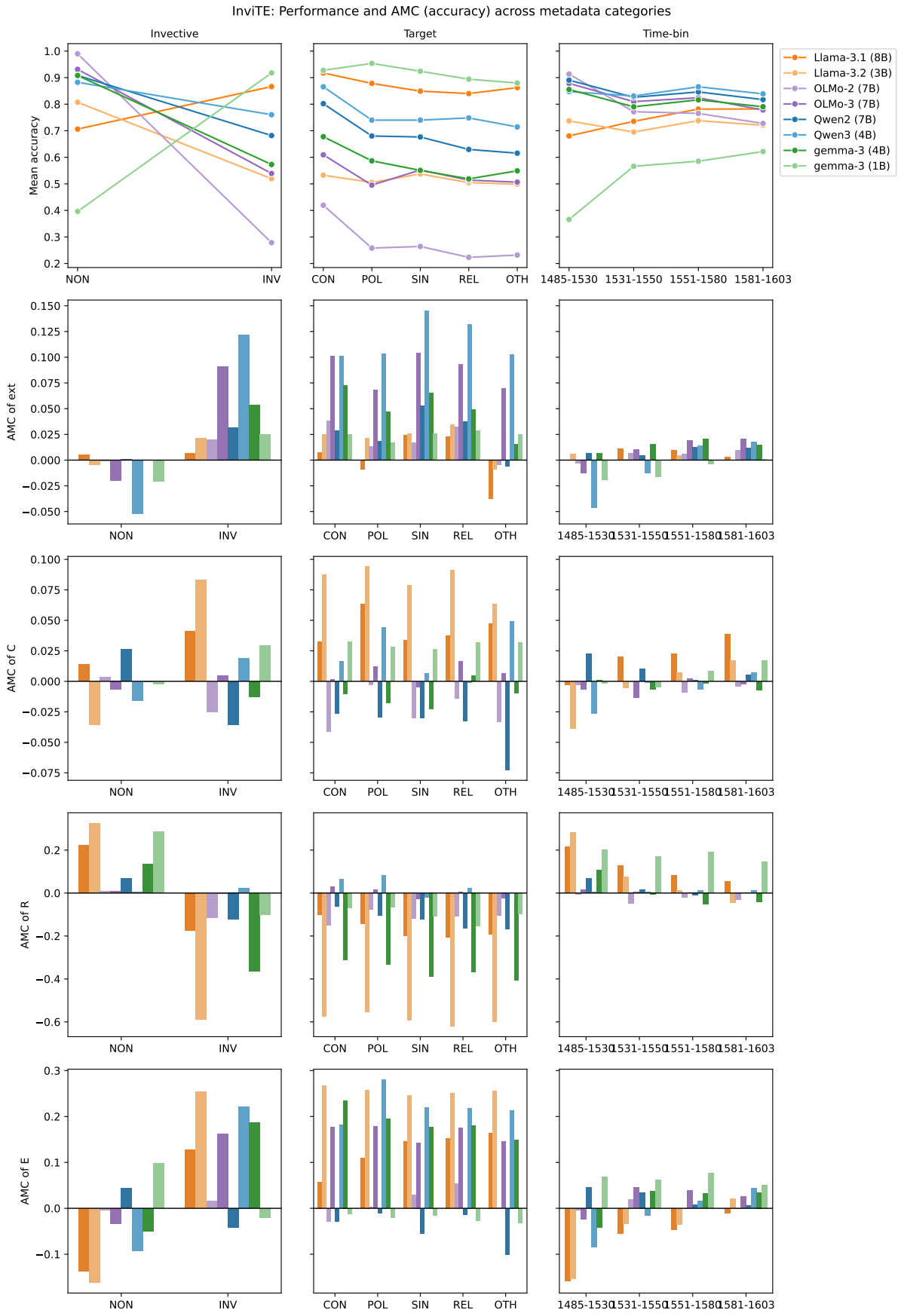

Figure 7: Mean model performance across prompt variants (first row, accuracy) and AMC of prompt components to accuracy (subsequent rows) across invective classes (first column) and metadata categories for InviTE (subsequent columns).

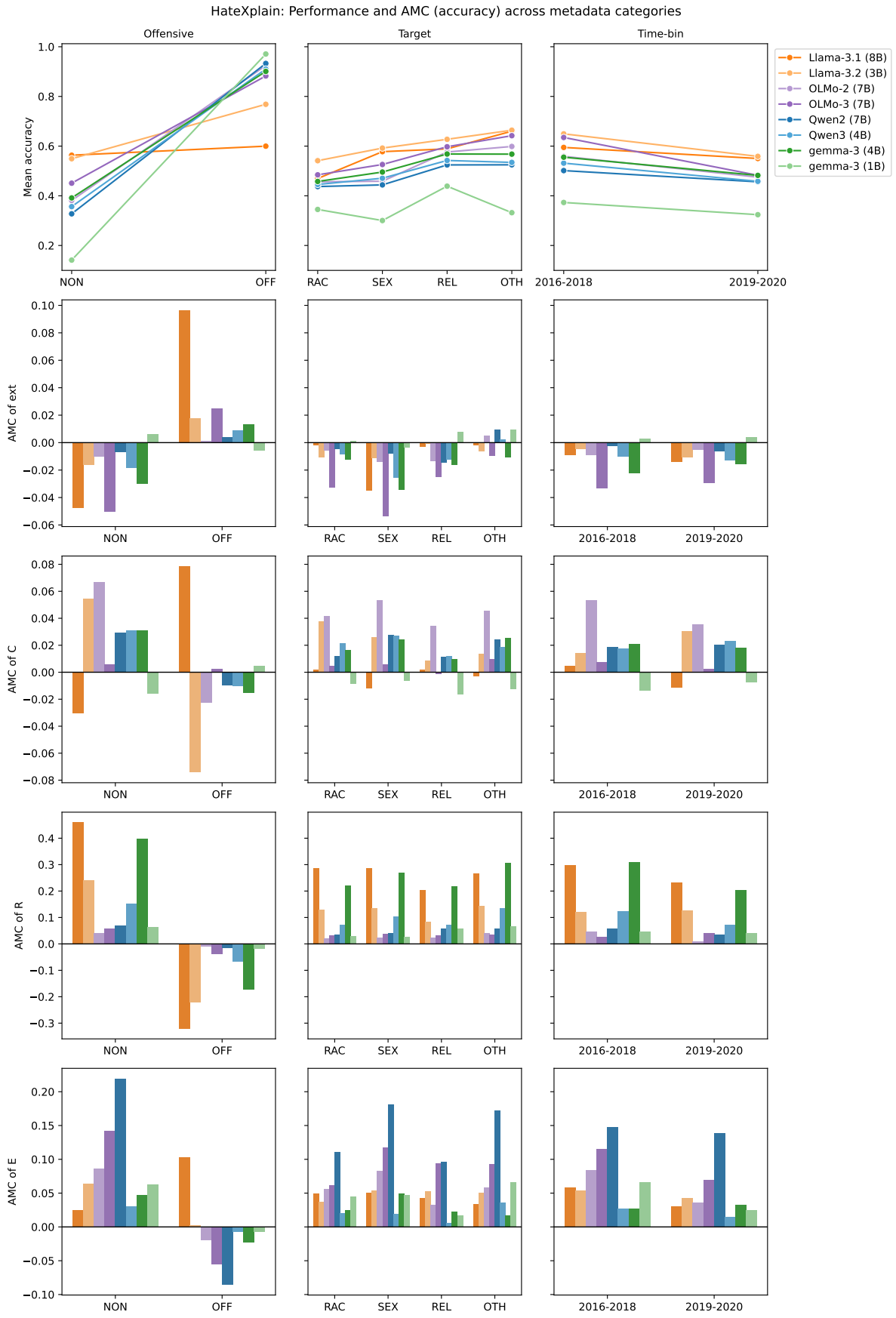

Figure 8: Mean model performance across prompt variants (first row, accuracy) and AMC of prompt components to accuracy (subsequent rows) across offensive classes (first column) and metadata categories (subsequent columns) for HateXplain.