Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Context-Aware LLM Evaluators for Content Moderation Judgments in a Low-Resource Setting

Felix Krejca
University of Applied Sciences St. Pölten
Tobias Kietreiber
University of Applied Sciences St. Pölten
Michael Wiegand
University of Vienna
Sebastian Neumaier
sebastian.neumaier@ustp.at
University of Applied Sciences St. Pölten

Research Article

Keywords:

Posted Date: August 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10576121/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Context-Aware LLM Evaluators for Content Moderation Judgments in a Low-Resource Setting

Felix Krejca[1*], Tobias Kietreiber[1], Michael Wiegand[2], Sebastian Neumaier[1,3*]

[1*]University of Applied Sciences St. Pölten, Austria.
[2]University of Vienna, Austria.
[3]Vienna University of Economics and Business, Austria.

*Corresponding author(s). E-mail(s): felix.krejca@gmail.com; sebastian.neumaier@ustp.at;
Contributing authors: tobias.kietreiber@ustp.at; michael.wiegand@univie.ac.at;

Abstract

Large Language Models (LLMs) are increasingly used as automated evaluators in domains where expert annotation is scarce and expensive. Moderation of German-language newspaper forums is such a domain: reference annotations require professional moderators and reach only moderate inter-annotator agreement. We ask how reliably LLM evaluators reproduce these expert judgments, and how far that reliability depends on how the judgment is elicited rather than on model capability alone. On the One Million Post Corpus, we evaluate 5 context levels across 6 binary moderation categories, from the isolated comment to the full article and conversational thread, together with 2 retrieval-augmented variants supplying the most similar prior moderation decisions. The design is repeated over 3 generative models ranging from 1 to 31 billion parameters. Context improves agreement, but not in proportion to its volume, and no single level is best for every category. Retrieving 2 short prior decisions matches the benefit of the full context at a fraction of the prompt length. The evaluator matters more than the prompting strategy: only the largest model approaches the decision threshold the human moderators applied. Against supervised systems published on the same corpus and splits, our approach is stronger on the 3 rare categories that govern removal decisions and weaker on the 3 frequent constructive-content categories — without any annotated training data.

1 Introduction

Large Language Models (LLMs) are increasingly deployed not as generators of text but as **evaluators** of it: **automated judges that assign labels, assess quality, and render decisions that were previously the exclusive domain of trained human annotators**. This shift matters most in domains where a strong annotation tradition does not exist and labeled data are scarce. Content moderation of user-generated content (UGC) is a case in point. Moderation is not a surface-level detection problem but an evaluative one: a moderator judges whether a contribution is off-topic, discriminating, or constructive relative to the norms of a specific community and the discussion in which it is embedded. Producing such judgments at scale through human labor has become prohibitively time-intensive and economically inefficient, while shifting regulatory landscapes such as the EU Digital Services Act (DSA) mandate heightened algorithmic accountability and automated compliance [1, 2].

The case for LLM-based evaluation is strongest precisely where human annotation is costly and specialized expertise is required. **Moderation of a German-language newspaper forum** exemplifies this **low-resource setting**. The expert annotations of the One Million Post Corpus [3] required 4 professional moderators working across 3 structured annotation phases to label fewer than 12,000 posts, and inter-annotator agreement remained moderate (κ between 0.3 and 0.6), reflecting the genuine interpretive difficulty of the task. Annotation resources of this kind are rare, expensive, and difficult to extend to new outlets, new communities, or new moderation policies. If LLM evaluators can reliably approximate expert moderator judgments, they offer a route to scalable labeling and system evaluation in a domain where the human alternative does not scale at all.

Whether they can do so, however, depends on how the judgment is elicited. Prevailing automated approaches remain predominantly "context-blind," presenting the evaluator with an isolated comment and thereby withholding exactly the information a human moderator relies upon. Literature demonstrates that integrating surrounding textual context improves classification performance [4–6], but existing research typically isolates single context levels or focuses on English-language datasets [7, 8]. Similarly, while Chen et al. [9] demonstrate the potential of Retrieval-Augmented Generation (RAG) for grounding model judgments in relevant decisions, such methods remain unexplored for German-language newspaper comment moderation. The reliability of LLM evaluators in this setting is therefore an open empirical question, as is the extent to which it is governed by prompt design and model scale rather than by model capability alone.

To address these gaps, this study presents a **systematic evaluation of LLM-based moderation judgments against expert human annotations** across 6 binary moderation categories (*Off-Topic, Inappropriate, Discriminating, Possibly Feedback, Personal Stories, Arguments Used*). We **compare 5 context levels**, ranging from the isolated comment through article metadata and full article text to hierarchical conversational threads, and introduce the **first RAG-based approach for German-language forum moderation**, in which the evaluator is supplied with the most similar positive and negative human moderation decisions for the category under consideration. This design space is mapped across 3 generative models

ranging from 1 to 31 billion parameters (Llama-3.2-1B, Llama-3.1-8B, and Gemma-4-31B) [10, 11], and the resulting judgments are compared with the supervised systems previously reported on the same corpus and the same cross-validation folds [3, 12]. We investigate the following research questions:

- **RQ1 (Context):** To what extent does supplying LLM evaluators with varying levels of textual context, from article content to conversational threads, improve the agreement of their moderation judgments with expert human annotations?
- **RQ2 (Retrieval & Prior Decisions):** To what extent does grounding LLM judgments in retrieved human moderation decisions via RAG improve performance relative to unaugmented baselines?
- **RQ3 (Reliability & Scale):** How consistent are these effects across moderation categories, model sizes and model families, and what does this imply for the reliability of deploying LLM evaluators in low-resource moderation settings?

By addressing these questions, this work moves beyond context-blind evaluation and provides practical evidence on when LLM evaluators can substitute for, and when they must remain subordinate to, human moderation expertise in non-English digital public spheres.

The remainder of this paper is organized as follows. Section 2 places the study within work on LLM evaluators, defines our notion of automated content moderation, and surveys context-aware and German-language moderation research. Section 3 describes the One Million Post Corpus, its annotation procedure and the variables from which the context levels are constructed. Section 4 details the context extraction procedure, the 5 context levels and the 2 retrieval-augmented variants. Section 5 answers RQ1–RQ3 in turn, one factor at a time, and then compares the best prompting strategy with the supervised systems previously published on this corpus. Section 6 discusses limitations, and Section 7 concludes with design considerations for newsroom deployment and directions for future work.

2 Related Work

This section reviews the literature most relevant to the present study. To provide a clear structure, the discussion is divided into 4 parts. Section 2.1 reviews the literature on LLMs as evaluators and the reliability concerns identified in this body of work. Section 2.2 defines the notion of automated content moderation adopted in this research, which treats moderation as an evaluative judgment about a contribution's role in a community rather than as the detection of a fixed set of undesirable features. Section 2.3 surveys how contextual information has been used for moderation across languages and platforms. Section 2.4 then narrows to the German-language literature, with particular attention to the absence of both retrieval-augmented and evaluator-framed approaches in this domain.

2.1 Large Language Models as Evaluators

The use of LLMs as automated judges has developed rapidly across natural language processing and information retrieval. Zheng et al. [13] establish that strong LLMs

approximate human preference judgments on open-ended generation tasks closely enough to serve as a scalable proxy, and Gilardi et al. [14] report that instruction-tuned models can exceed crowdworker accuracy on several text-annotation tasks at a fraction of the cost. Ziems et al. [15] draw a complementary conclusion for computational social science: LLMs are useful as annotation aids and as zero-shot labelers for coarse constructs, but remain unreliable substitutes for trained coders on constructs that require sustained interpretive judgment.

Alongside these results, a body of work documents failure modes specific to the evaluator role. Wang et al. [16] show that LLM judgments are sensitive to the order in which candidates are presented, a position bias severe enough to reverse system rankings, and Panickssery et al. [17] show that evaluators recognize and favor their own generations.

Two properties of the moderation setting distinguish it from the tasks on which most of this literature is based. First, the judgment is not intrinsic to the item but relational: whether a comment is, for example, *Off-Topic* or constructive in the sense of *Arguments Used* — and the same holds for the remaining moderation categories — is defined with respect to an article and a surrounding discussion that the evaluator is usually not shown. Second, the reference standard is itself contested; the moderate inter-annotator agreement of the One Million Post Corpus places a ceiling on the agreement any evaluator, human or automated, can be expected to reach. Both properties motivate the 2 prompting strategies we study, supplying context and retrieving prior human decisions.

2.2 Defining "Automated Content Moderation"

The prompting strategies motivated above presuppose a clear notion of the judgment that is being automated. We therefore first fix what "automated content moderation" encompasses in this article. The term *automated* refers to the use of computational decision procedures to carry out moderation actions [18]. In practice, such programs work in 2 main ways: they match content against known items, for example to find exact duplicates, or they classify it, for example as spam or hate speech [18]. The term *content moderation*, however, covers more than the removal of unwanted content. This article follows James Grimmelmann, who defines content moderation as:

> "... the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [19].

Under this definition, moderation involves more than just sanctioning users. Because it includes the facilitation of cooperation, it also covers positive interventions that support a discussion, such as pinning an insightful comment. This broader view matters for the present study: 3 of our 6 moderation categories,*Possibly Feedback*, *Arguments Used* and *Personal Stories*, mark contributions as valuable rather than objectionable. Deciding whether they apply requires the evaluator to weigh a contribution against the norms of the community, not merely to detect a surface feature such as an offensive word.

2.3 The Usage of Contextual Information for Content Moderation

Judging a contribution against the norms of a community, as the definition above requires, presupposes access to the discussion in which it appears — the relational property identified in Section 2.1. Accordingly, incorporating contextual information enhances the performance of automated content classifiers [4, 8, 20–23]. Analyzing posts in isolation often fails to capture the nuances of online discourse, such as sarcasm, implicit bias, or ongoing conversational dynamics. Consequently, recent literature has shifted toward context-aware architectures. As summarized in Table 1, contemporary approaches use various forms of context, ranging from local conversational history to broader structural and user-centric networks, to improve the detection of hate speech, toxicity, and offensive language.

The use of the term "context" is correspondingly varied, reflecting the different communicative settings that arise across domains such as Twitter, Reddit, and news websites such as Der Standard, as well as across languages including English, German, Spanish, Portuguese, and Hindi. Table 1 maps this work along 2 axes that matter in this article: the kind of contextual data supplied, from immediate conversational threads and article metadata through to social graphs and author attributes, and the architecture that consumes it, from language-specific encoders to prompted LLMs.

Two strands of this literature bear directly on the design evaluated in this research.

Textual and conversational context. The most direct form of context is the text surrounding the target post. Yu et al. [4] use parent postings on Reddit to separate hate from counter-hate with RoBERTa, and Nayak and Joshi [7] and Schmid et al. [23] supply chronologically prior posts for hate and toxic speech detection on Twitter. Beyond conversational context, prior work has also explored the use of information originating from the source of a post. Pérez et al. [5] show that providing BETO, a BERT-based language model pretrained for Spanish, with the full news article in addition to the newspaper's original tweet and the target reply improves multi-label hateful content detection in Spanish. Similarly, Krejca et al. [6] incorporate article titles, article paths, and community guidelines from DerStandard to predict moderation outcomes. These approaches motivate the source-related context signals considered in Section 4.1.

Context supplied through the prompt. Where earlier work injected context architecturally, through fine-tuned language-specific encoders such as BERTimbau and ALBERTina for Portuguese [25], mBERT and indicBERT for Hindi and English [7], or BETO for Spanish [5], LLMs allow it to be supplied directly in the prompt. Han and Tang [27] use zero- and few-shot in-context learning with GPT-3 for binary hate speech detection, and Lee and Ram [24] have LLM agents assess hate speech collaboratively through readable reasoning steps. This is the strand our study extends, by treating what is placed in the prompt as the experimental variable rather than a fixed design choice.

A third strand models the social and behavioral footprint of users — vocabularies and network interactions [20], social-graph topology and posting histories [8, 21], and author identity attributes [22]. We do not pursue it: the One Million Post Corpus exposes user identifiers but no social graph, and confining the study to textual context

Table 1 Context-aware content moderation across languages and platforms. Scope: any language, any platform, any method, restricted to work that supplies a classifier with information beyond the target post itself.

Paper	Language	Dataset source	Context used	Task	Classifier
Davidson [22]	English	Artificial Data	Author Identity Variables, Social/Engagement Context, Topical Context	Comparison of 2 posts, which one is more likely to be in category	GPT 4o, Gemini 2.5 Flash, Qwen2-VL, Gemma3, InternVL3
Krejca et al. [6]	German	DerStandard	Article Title, Article Path, Online-Ratio, Community Guidelines	Online/Deleted	LSTM, CNN, GPT 3.5
Lee & Ram [24]	English	WikiHate, SupremacistHate, HateXplain (X), HateXplain (Gap)	In-Context Learning	Binary Hate Speech	MultiAgent Setup GPT4o-mini
Schmid et al. [23]	German	Twitter	Root Posting, Postings Before the Target	Hate Speech, Toxic Speech	GPT4o-mini, Grok3-mini, Claude 3 Haiku
Assis et al. [25]	Portuguese	Instagram, Twitter	Representative Keywords	Neutral, Offensive, Hate Speech	ChatGPT, MariTalk, BERTimbau, ALBERTina
Wang & Liao [26]	N/A	Reddit, Gab	Conversational Structure	Binary Hate Speech	SCNet
Ghosh et al. [8]	English	Reddit, GAB, DIALOCONAN, CAD, ICHCL, Latent Hatred	User's Previous Posts, Personal Social Context, Conversational Context	Binary Hate Speech	Context Synergized Hyperbolic Network
Nagar et al. [21]	English	Twitter	Post-User Postings, Profile Text, Social-Network Graph	Binary Hate	Variational Graph Autoencoder
Pérez et al. [5]	Spanish	Twitter	Textual: Comment, Original Tweet + Comment, Full News Article + Original Tweet + Comment	Binary Hateful; Multilabel Hateful	BETO
Han & Tang [27]	English	YouTube, Reddit	In-context learning	Binary Hate Speech	GPT-3
Yu et al. [4]	English	Reddit	Comment & Parent Comment	Hate, Counter-Hate, Neutral	RoBERTa
Mosca et al. [20]	English	Twitter	User Network; User Vocabulary	Multiclass (Racism/ Sexism/Neither) & (Hate/Offensive/ Neither)	Multi Layer Perceptron
Nayak & Joshi [7]	English, Hindi	Twitter	Conversational Context	Binary Hate; Binary Offensive	mBERT, indicBERT

keeps the comparison with the supervised systems of Section 5.4 like-for-like. Section 6 returns to this as a limitation.

Across all 3 strands, context is treated as a feature to be engineered into a classifier. None of this work asks how a model's agreement with expert judgment responds to what it is shown, which is the question Section 5 takes up.

2.4 Content Moderation in German Newspaper Datasets Using Transformers

The context-aware work surveyed in the previous subsection spans many languages and platforms. We now narrow the scope to the setting of this study, German-language newspaper forums. Table 2 lists the studies we are aware of that apply transformer-based models to content moderation tasks on such data.

Table 2 Transformer-based content moderation on German-language newspaper data. Scope: German only, newspaper fora only, irrespective of whether contextual information is used.

Paper	Dataset	Task	Model(s)
Krejca et al. [6]	OMP Corpus	Online/ Deleted	GPT-3.5-Turbo
Weber et al. [28]	HOCON34k	Binary HS	GPT-4o, Perspective API, OpenAI Moderation API
Keller et al. [29]	HOCON34k	Binary HS	BERT
Yadav & Milde [12]	NDR, OMP Corpus	Online/ Offline	BERT, forumBERT
Assenmacher et al. [30]	Rheinische Post, OMP Corpus	Online/ Offline	BERT
Goldzycher et al. [31]	GAHD	Binary HS	GPT-3.5, GPT-4, LeoLM, Perspective API, OpenAI Moderation API, GELECTRA-large
Justen et al. [32]	German Newspaper Corpus	Accepted/ Rejected	BERT
Bremm et al. [33]	GerMS-Detect	Sexism (Bin./Multi.)	GBERT, mE5
Krenn et al. [34]	GerMS-Detect	Sexism (Bin./Multi.)	GBERT
Kotarcic et al. [35]	Unpublished	Binary HS	GELECTRA, GBERT, mBERT, XLM-R
Gross et al. [36]	GerMS-Detect	Sexism (Bin./Multi.)	GBERT
Pachinger et al. [37]	Austrotox	Offensive, Target, Vulgar Spans	BERT, BERT-DBM, GBERT, GELECTRA, LeoLM, Mistral, GPT-3.5, GPT-4
Geiss & Zarcone [38]	GerMS-Detect	Sexism (Bin./Multi.)	GBERT, BERT
Akomeah et al. [39]	GerMS-Detect	Sexism (Bin./Multi.)	GBERT, BERT, GPT-3.5
Ruiter et al. [40]	M-Phasis	10 tasks	GBERT

The papers presented in Table 2 illustrate several critical trends and notable technical gaps within the field of automated content moderation. Most prominently, there is a clear dominance of BERT-based architectures, particularly models pre-trained on German-language corpora. Models such as GBERT and GELECTRA appear frequently throughout the literature, which shows the importance of language-specific pre-training for German moderation tasks.

In recent years, research has increasingly shifted toward the integration of generative LLMs. Studies by Krejca et al. [6], Weber et al. [28], Goldzycher et al. [31], and Pachinger et al. [37] demonstrate a growing reliance on proprietary models such as GPT-3.5 and GPT-4, as well as open-source alternatives including LeoLM. Notably, however, all of these studies use the generative model as a classifier and evaluate it as such; none frames the model as an evaluator whose agreement with expert judgment is the object of study, and none examines how that agreement responds to the way the judgment is elicited. It is precisely this evaluator framing that the present article adopts.

Despite this progress, the surveyed literature reveals a strong and persistent preference for binary moderation formulations. Of the 15 studies summarized in Table 2, 8 explicitly frame moderation as a binary classification task. Grouped by task type, these fall into 2 groups: binary hate speech detection, studied on HOCON34k by Weber et al. [28] and Keller et al. [29], on the GAHD dataset by Goldzycher et al. [31], and across multiple transformer architectures by Kotarcic et al. [35]; and binary moderation-outcome prediction, framed as Online/Deleted by Krejca et al. [6], as Online/Offline by Yadav and Milde [12] and Assenmacher et al. [30], and as Accepted/Rejected by Justen et al. [32].

In contrast, only a smaller subset of studies moves beyond this binary framing. 5 papers based on the GerMS-Detect Shared Task dataset, later published as GerMS-AT, introduce more fine-grained moderation formulations by combining binary classification with multiclass sexism detection [33, 34, 36, 38, 39]. However, even within these studies, binary classification remains a central evaluation setting, suggesting that richer label spaces have yet to become standard practice. Beyond the GerMS-AT corpus, only a limited number of studies adopt substantially more structured prediction settings. Pachinger et al. [37] move beyond binary labels by introducing span-level offensiveness detection and target identification within the AustroTox dataset. Similarly, Ruiter et al. [40] present the M-Phasis benchmark, which comprises 10 distinct moderation-related tasks and therefore extends beyond traditional binary or single-label classification paradigms. German-language content moderation research using transformer-based methods therefore remains overwhelmingly anchored in binary classification; only isolated efforts have explored more granular or structured alternatives.

Furthermore, a significant architectural gap is the complete absence of Retrieval-Augmented Generation approaches. Although RAG has become established across numerous NLP domains as a means of reducing hallucinations and incorporating external knowledge sources [41, 42], it has not yet been applied to German-language content moderation according to the surveyed literature. This represents a promising avenue, as RAG-based systems could allow moderation models to reference previous

moderation decisions and contextual examples without requiring costly retraining procedures [9]. This is a particularly attractive property in the evaluator setting, where the criterion is defined by an accumulating record of human decisions rather than by a fixed specification. The present work addresses this gap by incorporating a RAG-based approach for German-language content moderation, enabling the model to make use of previously collected moderation examples and decisions as external contextual knowledge. It is also evident that researchers have repeatedly relied on the One Million Post Corpus and GerMS-AT as benchmark datasets for German-language moderation tasks.

The role of context in German-language work is summarized by Krejca et al. [6], who highlight the varying degrees to which contextual information is incorporated across existing moderation datasets and models. Yadav and Milde [12] incorporate contextual information by jointly using article titles and user comments, and further improve their analysis through topic-based data partitioning. Similarly, Pachinger et al. [37] include article titles as a form of contextual signal in their experimental setup. Krejca et al. [6] report positive effects of incorporating contextual variables in classification tasks. In contrast, Keller et al. [29] explicitly highlight the limitations arising from insufficient contextual information during their annotation process, emphasizing the impact of context scarcity on dataset quality and downstream model performance. In comparison with the number of studies discussed above and in Table 2, this further illustrates the scarcity of context-aware approaches in the current literature on German-language content moderation.

In summary, while the field demonstrates considerable technical maturity with respect to model selection, it remains comparatively conservative in its architectural choices, favoring direct classification approaches over more complex, context-aware and retrieval-enhanced moderation systems, and it has not yet examined these systems as evaluators of human judgment.

3 Data

Assessing whether LLM judgments approximate expert moderation decisions requires a corpus in which such decisions have been recorded by trained moderators and in which the surrounding discussion is preserved. The One Million Post Corpus satisfies both conditions. This section describes the corpus and its annotation procedure (Section 3.1), and then details the post- and article-level variables from which the context levels used in our experiments are constructed (Section 3.2).

3.1 The One Million Post Corpus

To address the central research questions, this study uses the One Million Post Corpus [3], a large-scale dataset derived from the user discussion forums of the Austrian daily newspaper *Der Standard*. This corpus is one of the most extensive resources for German-language content moderation research, both in volume and in the granularity of its contextual data. It contains 1,000,000 posts carrying only the production moderation decision (*online* or *deleted*), alongside 11,773 posts manually annotated into 7 categories by expert forum moderators. The expert-annotated subset is used

throughout our analysis; the annotation categories are *Sentiment*, *Off-Topic*, *Inappropriate*, *Discriminating*, *Possibly Feedback*, *Personal Stories* and *Arguments Used*. Table 3 summarizes the corpus, and Table 4 defines the 6 moderation categories used in this article, as formulated by the moderators themselves, together with their expert label distribution.

Table 3 Key statistics of the One Million Post Corpus [3].

Total posts	1,011,773
Posts without expert annotation	1,000,000
Expert-annotated posts	11,773
Posts taken offline by moderators	62,320
Articles	12,087
Users	31,413
Min / median / max post length (words)	0 / 21 / 500

Unlike purely comment-focused datasets, this corpus enables the evaluation of context-aware architectures by providing rich contextual information, including user histories, article context, community votes, and the tree-like structure of user discussions.

Table 4 The 6 moderation categories used in this article, with the definitions formulated by the professional forum moderators of *Der Standard* and the resulting expert label distribution [3]. The categories fall into the 2 groups used throughout this article: those that govern removal decisions and those that identify constructive content. *Applies* is the number of posts judged to belong to the category, *Total* the number of posts labeled for it, and *Prev.* the prevalence of the positive (minority) class. Not every annotated post is labeled for every category, so the totals differ.

Category	**Definition**	**Applies**	**Total**	**Prev.**
Categories that govern removal decisions				
Off-Topic	Posts that digress too far from the topic of the corresponding article.	580	3,599	16 %
Inappropriate	Swearwords, suggestive or obscene language, insults, threats.	303	3,599	8 %
Discriminating	Racist, sexist, misogynistic, homophobic, antisemitic and other misanthropic content.	282	3,599	8 %
Categories that identify constructive content				
Possibly Feedback	Posts in which users ask questions or address feedback to the author or to the newspaper in general, and which may therefore require a reply.	1,301	6,038	22 %
Personal Stories	Users sharing personal experiences, anecdotes or stories relevant to the topic of the article.	1,625	9,336	17 %
Arguments Used	Users backing their statements with rational argumentation, reasoning or sources.	1,022	3,599	28 %

The annotation procedure for the manually annotated subset is described by Schabus et al. [3] and was structured in 3 phases to ensure consistency and to adequately capture rare types of posts. 4 professional moderators participated. A small dry run

was first performed to align the moderators' understanding of the category definitions. In the second phase, a random sample of 1,000 posts was annotated in parallel to assess inter-annotator agreement, measured by Cohen's κ and reported as between 0.3 and 0.6 depending on the category. The final phase increased the number of minority-class examples. Posts were selected using 3 targeted strategies:

1. news articles with a high share of filtered-out content, to capture likely *Off-Topic*, *Inappropriate* and *Discriminating* posts;
2. "share your thoughts" forum sections, to capture *Personal Stories*; and
3. posts either replied to by staff or flagged by a classifier, to enrich *Possibly Feedback* content.

This procedure yielded the 11,773 annotated posts.

Two features of this procedure matter for the present study. First, the effort involved, 4 trained professionals across 3 coordinated phases for fewer than 12,000 posts, is precisely the cost that makes the domain low-resource and motivates automated evaluation. The setting is, moreover, Austrian German newspaper discourse, a language variety for which annotated moderation resources are even scarcer than for standard German. Second, the reported agreement range establishes a ceiling: on categories where trained moderators agree only moderately, no evaluator should be expected to reproduce the reference labels exactly, and observed disagreement must be interpreted against that ceiling rather than against perfect agreement.

3.2 Dataset Characteristics

The dataset contains the post-related variables listed in Table 5, following the presentation of Krejca et al. [6]. As illustrated there, each post consists of 2 primary user-generated textual components: the headline and the comment body. For all experiments, these 2 fields were concatenated to form the complete comment representation used for subsequent processing.

In addition, each post carries a unique anonymized identifier (`Post ID`). Together with the `Parent Post` field, this identifier enables the reconstruction of conversational reply structures and discussion hierarchies. These hierarchical relations form the basis for the extraction of thread-level contextual information used in several experimental conditions.

Beyond post-level information, the corpus provides article-level metadata that can be linked through the `Article ID` variable. The available article attributes are summarized in Table 6. We use the article title and article body as contextual signals. While the article title is stored as plain text, the article body contains the full article content represented as an HTML string.

These article-level features constitute an additional source of contextual information and are incorporated into selected context levels; how they are assembled into prompts is described in Section 4.

We focus on the classification tasks defined over the 11,773 expert-labeled posts. As Table 4 shows, some categories suffer from considerably heavier class imbalance than others: positive-class prevalence ranges from 8 % (*Inappropriate*, *Discriminating*) to

Table 5 Post-level fields recorded in the One Million Post Corpus [6]. Rows are ordered as discussed in the text; the vote counts recorded by the corpus are omitted, as they are not used in this study.

Name	Description	Example
`Headline`	The headline of the post	NULL
`Body`	The main text of the post	Three packs of cards with 36 cards each per year
`Post ID`	Identifier of the post	81085
`Parent Post`	NULL for top-level comments; otherwise, the ID of the parent comment	80997
`Article ID`	The ID of the article under which the comment was posted	1212
`User ID`	The (anonymized) user identifier	7721
`Time Stamp`	The time when the post was created	2015-07-02 12:25:53.553
`Status`	Whether the post is "online" or "deleted"	online

Note: Examples originally written in German have been translated into English for readability.

Table 6 Article-level fields recorded in the One Million Post Corpus [6].

Name	Description	Example
`Article ID`	Identifier of the article	1212
`Path`	Breadcrumbs to the article	Newsroom/Panorama/Chronicle
`Date`	Date of publication	2015-07-02 05:30:00.00
`Title`	Title of the article	Ladies' convent in Innsbruck: Noblewomen who pray daily for the emperor
`Body`	Content of the article	*omitted for brevity*

Note: Examples originally written in German have been translated into English for readability.

28 % (*Arguments Used*), and in absolute terms the smallest categories provide roughly 300 positive examples in total.

All moderation categories (*Off-Topic*, *Inappropriate*, *Discriminating*, *Possibly Feedback*, *Personal Stories* and *Arguments Used*) of the expert-labeled subset were included, with the explicit exclusion of the *Sentiment* categories. *Sentiment* was excluded because it is an indicator of a shifting discussion rather than a judgment that directly prompts a reaction from the moderation team or the journalistic staff, as the other categories do.

4 Methods

Our experimental design varies 2 factors: the contextual information supplied to the evaluator, and whether prior human moderation decisions are retrieved and provided

alongside it. We therefore first describe how contextual features are extracted from the corpus and assembled into the 5 context levels (Section 4.1), and then the 2 retrieval-augmented variants (Section 4.2). Table 7 summarizes the notation used throughout; bold letters in its descriptions indicate the words from which each symbol is derived.

Table 7 Overview of the variables used throughout the Methods section. Bold letters in the descriptions indicate the terms from which each symbol is derived.

Symbol	Description
p	Target **p**ost to be classified
$T(p)$	Concatenated headline and body **t**ext of the target post
$A_{\text{title}}(p)$	**Title** of the **a**rticle under which p was posted
$A_{\text{body}}(p)$	**Body** text of the **a**rticle under which p was posted
$C(p)$	**C**onversational thread context of p, ordered chronologically, up to depth D
D	Maximum traversal **d**epth for conversational context extraction (default: 30)
I_s	**S**ystem-level **i**nstructions and output constraints
I_t	**T**ask-specific (category-level) moderation **i**nstructions [3]
$\mathcal{P}$	**P**rompt formatting/templating function
X_{CO}	**C**omment **O**nly level
X_{CAT}	**C**omment + **A**rticle **T**itle level
X_{FAC}	Comment + **F**ull **A**rticle **C**ontext level
X_{CC}	Comment + **C**onversational **C**ontext level
X_{FC}	**F**ull **C**ontext level
$E(p, k)$	The k reference **e**xamples retrieved for post p; throughout this study $k = 2$ (the most similar positive and the most similar negative example)
$X_{\text{CO-RAG}}$	RAG variant of the Comment Only level
$X_{\text{FC-RAG}}$	RAG variant of the Full Context level

4.1 Context Extraction and Prompt Construction

4.1.1 Extracting the Contextual Variables

To formalize the contextual data preparation, let p represent the target post. Based on the dataset features described above, we first define the directly accessible textual components:

- $T(p)$: The concatenated headline and body of the target post.
- $A_{\text{title}}(p)$: The headline/title of the parent article.
- $A_{\text{body}}(p)$: The main body text of the parent article.

While the article features can be merged directly via a relational join on the Article ID, constructing the conversational thread history requires iteratively traversing the database hierarchy. We define $C(p)$ as the chronologically ordered conversational context up to a maximum depth D. To construct this variable, we developed a conversational context extraction procedure, formally defined in Algorithm 1.

The upward traversal of the discussion tree relies on the explicit ceiling configuration (defaulting to $D = 30$) to safeguard against infinite loops emerging from potential database inconsistencies or cyclic references.

Algorithm 1 Conversational Context Extraction

Require: Target post p, Maximum depth D
Ensure: Chronological context string C

$C \leftarrow$ `""`
$x \leftarrow p.parent_post$
$depth \leftarrow 0$
while $x \neq \emptyset$ **and** $depth < D$ **do**
    $C \leftarrow$ `"["` $\cdot\ depth \cdot$ `"] "` $\cdot\ \text{text}(x) \cdot$ `"\n"` $\cdot\ C$
    $x \leftarrow x.parent_post$
    $depth \leftarrow depth + 1$
end while
return C

4.1.2 Assembling the Prompts

Our experiments use 3 open-weight generative LLMs: Llama-3.2-1B and Llama-3.1-8B [10], and Gemma-4-31B [11]. This selection follows 3 considerations. First, newspaper comment data are sensitive, and while proprietary APIs are one option, self-hosting open-weight models grants the highest degree of data sovereignty, keeping data within infrastructure the outlet controls; open weights also keep the study reproducible. Second, the 3 models range from 1 to 31 billion parameters, which is what allows RQ3 to isolate the effect of model scale, and they cover 2 model families, Llama and Gemma, so that the observed effects are not tied to a single family. Third, all 3 models support German and can be run on local hardware of the kind realistically available to a newsroom. All 3 models are prompted under an identical generation configuration so that observed differences reflect model capability rather than decoding settings. We use greedy decoding (`do_sample=false`) throughout: since each comment is assigned exactly one label, we want the single most probable output rather than a stochastically sampled one. This also matches evidence that greedy decoding outperforms sampling-based decoding on most evaluation tasks with a constrained output space and yields deterministic, reproducible outputs [43]. We run all models without quantization. Llama-3.2-1B and Llama-3.1-8B are loaded from their instruction-tuned Hugging Face checkpoints (`meta-llama/Llama-3.2-1B-Instruct` and `meta-llama/Llama-3.1-8B-Instruct`), and Gemma-4-31B from `google/gemma-4-31B-it`.

To evaluate these models, we adopted zero-shot prompting. This is the setting in which the low-resource argument for LLM evaluators actually applies: any few-shot or fine-tuned alternative presupposes the annotated data whose absence motivates automated evaluation in the first place. Zero-shot prompting is correspondingly the standard protocol in the LLM-as-a-judge literature [13, 14], and it keeps the prompting strategy, rather than the amount of supervision, as the only variable under study. To assess the contribution of contextual information, we evaluated 5 different context levels based on the extracted textual variables $T(p)$, $A_{\text{title}}(p)$, $A_{\text{body}}(p)$, and $C(p)$.

Prompt construction is governed by a formatting function $\mathcal{P}$ that combines task instructions and contextual information into a unified prompt. We distinguish between 2 instruction components:

- I_s: System-level instructions and output constraints.
- I_t: Category-specific moderation instructions derived from the annotation guidelines of Schabus et al. [3].

The formatting function $\mathcal{P}$ acts as a prompt templating engine, organizing contextual information into clearly separated markdown sections such as `[ARTICLE CONTENT]:`, `[DISCUSSION CONTEXT]:`, and `[COMMENT TO CLASSIFY]:`. This structure enables the model to distinguish the target comment from surrounding contextual information.

Using this framework, we define the following **5 context levels**[1]:

1. **Comment Only ($X_{\mathbf{CO}}$):**

$$X_{\mathrm{CO}} = \mathcal{P}\big(I_s, I_t, T(p)\big)$$

2. **Comment + Article Title ($X_{\mathbf{CAT}}$):**

$$X_{\mathrm{CAT}} = \mathcal{P}\big(I_s, I_t, T(p), A_{\mathrm{title}}(p)\big)$$

3. **Comment + Full Article Context ($X_{\mathbf{FAC}}$):**

$$X_{\mathrm{FAC}} = \mathcal{P}\big(I_s, I_t, T(p), A_{\mathrm{title}}(p), A_{\mathrm{body}}(p)\big)$$

4. **Comment + Conversational Context ($X_{\mathbf{CC}}$):**

$$X_{\mathrm{CC}} = \mathcal{P}\big(I_s, I_t, T(p), C(p)\big)$$

5. **Full Context ($X_{\mathbf{FC}}$):**

$$X_{\mathrm{FC}} = \mathcal{P}\big(I_s, I_t, T(p), A_{\mathrm{title}}(p), A_{\mathrm{body}}(p), C(p)\big)$$

Zero-shot inference was performed over the full expert-annotated subset. Because the retrieval-augmented variants require a retrieval pool disjoint from the evaluated instance, their predictions were produced under the original 10-fold split of Schabus et al. [3] and pooled across folds. Both settings therefore yield one prediction per annotated post, and the two are directly comparable. Model outputs were parsed into binary labels and compared against the expert annotations.

4.2 Retrieval-Augmented Generation (RAG)

To complement the zero-shot setup, we evaluated 2 Retrieval-Augmented Generation variants, built on the Comment Only (X_{CO}) and Full Context (X_{FC}) context levels.

[1] All prompts are available in our GitHub repository for reproducibility: https://github.com/FelixKrejca/HaSPI_content_moderation_w_context

These 2 are the endpoints of the context spectrum defined above: pairing retrieval with each of them isolates whether retrieved prior decisions substitute for context or complement it. Applying retrieval to the intermediate levels as well would have quadrupled the number of runs without addressing that question.

Semantic retrieval was performed using `embeddinggemma-300m` [44]. Evaluation followed the original 10-fold split structure provided by Schabus et al. [3]. For each moderation category, a model was evaluated across all folds, using 1 fold as the test set and the other 9 for retrieval, with final predictions aggregated over the complete cross-validation procedure. Retrieved examples were drawn exclusively from the 9 training folds of the corresponding cross-validation split. No examples from the held-out test fold were eligible for retrieval.

Following Chen et al. [9], each target post p was paired with 2 retrieved examples, denoted $E(p, 2)$ (see Table 7): the most similar positive example [Class X applies] and the most similar negative example [Class X does not apply] for the corresponding moderation category, using cosine similarity. These examples were inserted into the prompt under a dedicated `[REFERENCE EXAMPLES]:` section.

The resulting prompt formulations are:

$$X_{\text{CO-RAG}} = \mathcal{P}\big(I_s, I_t, E(p, 2), T(p)\big)$$

$$X_{\text{FC-RAG}} = \mathcal{P}\big(I_s, I_t, E(p, 2), T(p), A_{\text{title}}(p), A_{\text{body}}(p), C(p)\big)$$

Results were averaged across all folds and evaluated using the same procedure as the zero-shot experiments.

5 Results

Because the moderation categories differ substantially in prevalence, aggregate accuracy would obscure precisely the judgments of interest. We therefore report minority-class F_1 as the primary measure of agreement between model judgments and expert annotations, alongside minority-class precision and recall, which separate the 2 failure modes relevant to deployment: over-flagging acceptable contributions and overlooking those requiring intervention. Minority-class F_1 is also the measure reported by the prior work on this corpus against which we compare in Section 5.4, so the choice is dictated by comparability rather than preference.

The experimental design crosses 3 evaluators with 7 prompting strategies over 6 categories. The following subsections take one question at a time, varying a single factor while holding the others fixed, and address RQ1–RQ3 in turn before comparing the best prompting strategy with prior work. Figures 1 and 2 give the complete per-category picture, and the full matrix of precision, recall and F_1 for every cell is reported in Appendix A.

5.1 Which context level works best?

To isolate the effect of context we fix the evaluator to Gemma-4-31B, the strongest of the three, and set retrieval aside until Section 5.2. Table 8 reports the 5 context levels across the 6 categories.

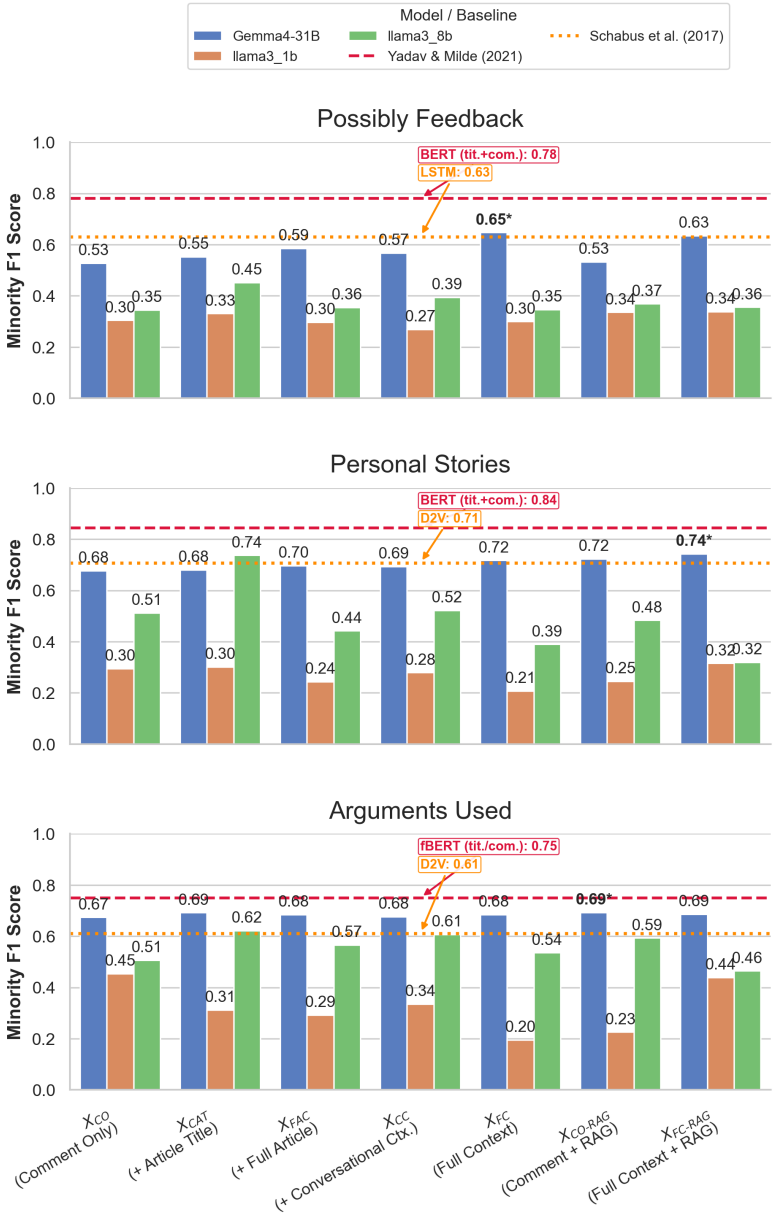


Fig. 1 Minority-class F_1 for the 3 categories associated with constructive content (*Possibly Feedback*, *Personal Stories*, *Arguments Used*). Bars show Gemma-4-31B, Llama-3.2-1B and Llama-3.1-8B across the 7 prompting strategies; horizontal lines mark the best result previously reported on the same subset and splits by Yadav and Milde [12] (dashed) and Schabus et al. [3] (dotted). Asterisks mark the best value per category.

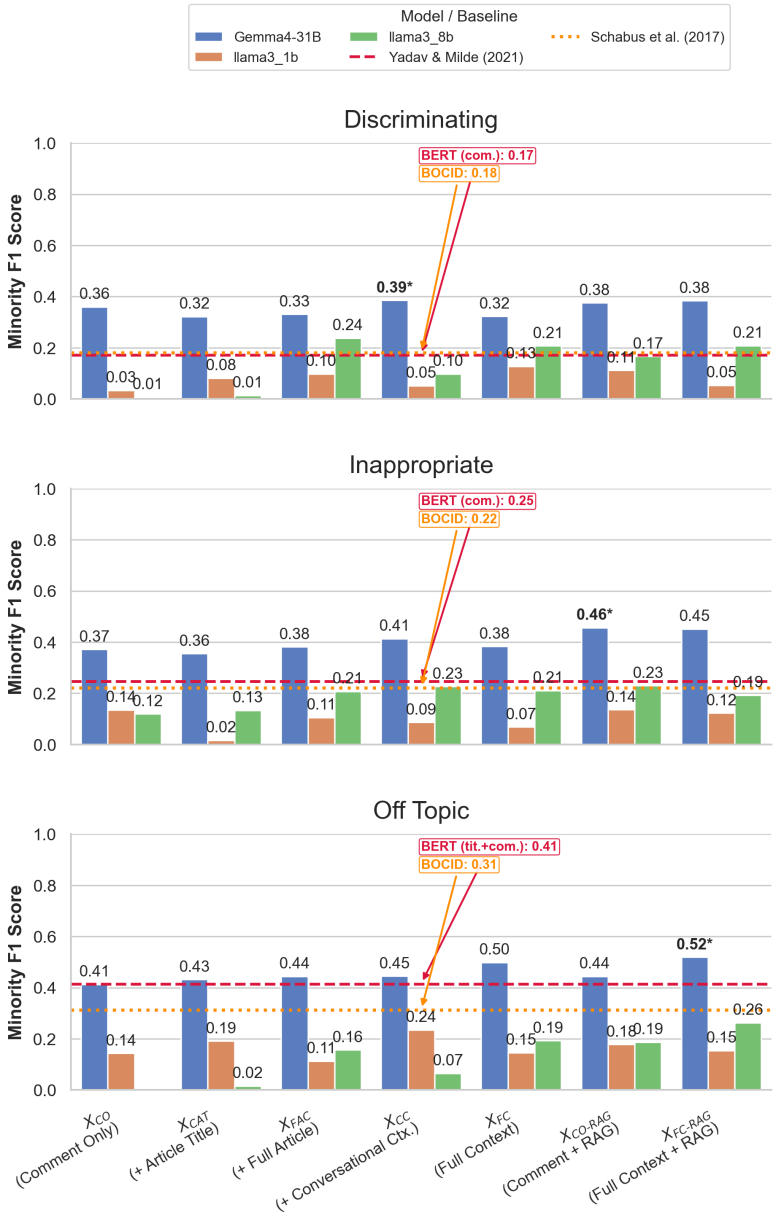


Fig. 2 Minority-class F_1 for the 3 categories that govern removal decisions (*Discriminating*, *Inappropriate*, *Off-Topic*). Bars show Gemma-4-31B, Llama-3.2-1B and Llama-3.1-8B across the 7 prompting strategies; horizontal lines mark the best result previously reported on the same subset and splits by Yadav and Milde [12] (dashed) and Schabus et al. [3] (dotted). Asterisks mark the best value per category.

Table 8 Comparison of the 5 context levels across the 6 moderation categories (RQ1): minority-class F_1 of **Gemma-4-31B**, with retrieval disabled. $\overline{F_1}$ is the mean across categories and Δ its change relative to the Comment Only baseline. Best value per column in bold. Category abbreviations: Arg. = *Arguments Used*, Disc. = *Discriminating*, Inapp. = *Inappropriate*, OffT. = *Off-Topic*, Pers. = *Personal Stories*, Feed. = *Possibly Feedback*.

Context level	Arg.	Disc.	Inapp.	OffT.	Pers.	Feed.	$\overline{F_1}$	Δ
Comment only (X_{CO})	0.674	0.360	0.372	0.412	0.677	0.528	0.504	—
+ article title (X_{CAT})	**0.692**	0.322	0.355	0.432	0.680	0.552	0.505	+0.002
+ full article (X_{FAC})	0.685	0.331	0.381	0.444	0.696	0.586	0.520	+0.017
+ conversation (X_{CC})	0.676	**0.386**	**0.414**	0.445	0.694	0.567	0.530	+0.026
+ article and conversation (X_{FC})	0.685	0.323	0.383	**0.498**	**0.717**	**0.647**	**0.542**	**+0.038**

Three observations follow. First, context helps, but modestly: the richest context level, X_{FC}, improves mean F_1 by 0.038 over the comment-only baseline, from 0.504 to 0.542. Second, the improvement is not monotonic in the amount of context supplied. Adding the article title alone is essentially inert (+0.002), and adding the full article (+0.017) is worth less than adding the conversation (+0.026), even though the article is by far the longer input. What the evaluator needs is the discussion the comment replies to, not the text the discussion is nominally about.

Third, and most important for deployment, no single context level is best everywhere. X_{FC} wins on *Off-Topic*, *Personal Stories* and *Possibly Feedback*; X_{CC} wins on *Discriminating* and *Inappropriate*; the article title alone wins on *Arguments Used*. The split is interpretable. *Off-Topic* is defined relative to the article, and supplying it raises agreement from 0.412 to 0.498, while *Possibly Feedback* gains most of any category from full context (0.528 $\rightarrow$ 0.647). *Discriminating* and *Inappropriate*, by contrast, are properties of the comment itself, and supplying the article actively hurts them: on *Discriminating*, every level that adds article text falls below the comment-only baseline of 0.360 — 0.322 for the title alone, 0.331 for the full article, 0.323 for both article and conversation — while the conversation alone raises it to 0.386. Only X_{CC} improves all 6 categories; every other level degrades at least one.

This ordering is a property of the largest evaluator. For Llama-3.1-8B the best level is the article title ($\overline{F_1}$ = 0.329), ahead of full context (0.314), and for Llama-3.2-1B every level falls below the comment-only baseline, from 0.228 to 0.174 under full context. These means are computed from the full result matrix in Appendix A; the corresponding per-category values are shown in Figures 1 and 2, and Section 5.3 returns to the role of the evaluator.

5.2 Does retrieval replace context, or add to it?

The 2 prompting strategies under study — supplying context and retrieving prior moderation decisions — provide the evaluator with different things. Context supplies the *situation* in which a comment was made; retrieval supplies the *standard* by which comparable comments were previously judged by human moderators. Whether these are substitutes or complements is an empirical question, and the design answers it directly: crossing context (none vs. full) with retrieval (off vs. on) yields the 2×2 in Table 9.

Table 9 Effect of combining context with retrieval of prior moderation decisions (RQ2): mean minority-class F_1 across the 6 categories for **Gemma-4-31B**. Marginal effects are given in the final row and column.

	Retrieval off	**Retrieval on**	**Effect of retrieval**
Comment only (X_{CO})	0.504	0.537	+0.034
Full context (X_{FC})	0.542	**0.570**	+0.027
Effect of context	+0.038	+0.032	

Table 10 Comparison of the 3 evaluator models under a fixed prompting strategy, retrieval combined with full context (RQ3). $\overline{F_1}$, $\overline{P}$ and $\overline{R}$ are means over the 6 categories; $\overline{F_1}^{\mathrm{CO}}$ is the comment-only baseline of the same model; $\overline{\Delta}$ is the mean change over all 36 model–strategy–category cells relative to that baseline, and *pos.* the number of those cells in which it is positive.

Evaluator	**Params**	$\overline{F_1}$	$\overline{P}$	$\overline{R}$	$\overline{F_1}^{\mathrm{CO}}$	$\overline{\Delta}$	**pos.**
Gemma-4-31B	31B	**0.570**	**0.520**	0.654	**0.504**	+0.030	32/36
Llama-3.1-8B	8B	0.301	0.182	**0.924**	0.248	**+0.073**	31/36
Llama-3.2-1B	1B	0.238	0.187	0.459	0.228	−0.024	16/36

Retrieval and context are near-equivalent substitutes and mildly complementary. Supplied on its own, retrieval recovers almost exactly what full context recovers: 0.537 against 0.542, a difference of 0.005. But it does so from a prompt containing 2 short comments rather than an entire news article plus the full ancestor chain, which makes it by far the cheaper of the 2 routes to the same agreement. Combining them is better than either alone, at 0.570, and both marginal effects remain positive in the presence of the other (+0.032 for context given retrieval, +0.027 for retrieval given context). The 2 effects overlap slightly: the joint gain of 0.066 is a little less than the 0.072 obtained by adding them separately, which is what one expects if the two partly convey overlapping information about how the category boundary is drawn.

Retrieval is also the more reliable of the two. Both retrieval variants improve all 6 categories over the comment-only baseline, whereas among the purely contextual levels only X_{CC} does. Isolating retrieval from context sharpens the picture further. Added to the comment-only prompt, retrieval helps *Inappropriate* most of any category (0.372 → 0.456); added on top of full context, it again helps *Inappropriate* most (+0.069), followed by *Discriminating* (+0.061), which every level that supplies article text degrades. These are the 2 categories in which the written annotation guideline leaves the decision boundary least determined, and where seeing how moderators actually drew it is more informative than seeing more of the discussion.

5.3 How much does the evaluator model matter?

Fixing the prompting strategy to the best one identified above, retrieval combined with full context, isolates the contribution of the evaluator itself. Table 10 reports the 3 models.

Table 10 shows a clear ordering: Gemma-4-31B delivers the best agreement on every summary measure, followed at a considerable distance by Llama-3.1-8B and then Llama-3.2-1B. The evaluator therefore matters more than the prompting strategy:

Table 11 Mean change in minority-class F_1 relative to each model's own comment-only level, by prompting strategy. Each entry is averaged over the 6 moderation categories. $\overline{\Delta}$ is the mean over all 36 model–strategy–category cells for that evaluator, and *pos.* the number of those cells in which the change is positive.

	Context				**Retrieval**			
Evaluator	X_{CAT}	X_{FAC}	X_{CC}	X_{FC}	$X_{\text{CO-RAG}}$	$X_{\text{FC-RAG}}$	$\overline{\Delta}$	**pos.**
Gemma-4-31B	+0.002	+0.017	+0.026	+0.038	+0.034	**+0.066**	+0.030	32/36
Llama-3.1-8B	+0.081	+0.079	+0.071	+0.066	**+0.090**	+0.052	+0.073	31/36
Llama-3.2-1B	−0.022	−0.037	−0.019	−0.054	−0.022	**+0.010**	−0.024	16/36
All evaluators	+0.020	+0.020	+0.026	+0.017	+0.034	+0.043	+0.027	79/108

the gap between the best and worst model under a fixed strategy is 0.332 in mean F_1, whereas the largest gap between strategies within Gemma-4-31B is 0.066. Prompt design is a refinement, not a substitute for capability.

The ranking is not proportional to scale, however. Without context, Llama-3.2-1B and Llama-3.1-8B are almost indistinguishable (0.228 and 0.248) despite an eightfold difference in parameters, while the step to 31B roughly doubles agreement. What scale buys is less raw capability than the ability to *use* additional information: the same context and retrieved examples that improve the 31B evaluator in 32 of 36 conditions fail to improve the 1B evaluator in 20 of 36, and degrade it by 0.024 on average. Table 11 breaks this observation down by prompting strategy: for each evaluator it reports the mean change in minority-class F_1 relative to that model's own comment-only baseline, averaged over the 6 categories.

Read across the rows, Table 11 also shows that retrieval combined with full context is the only single strategy that improves all 3 evaluators on average, and that some retrieval variant is the best available strategy for the 2 smaller models even though full context is not.

At first sight, Table 11 seems to suggest that the mid-sized model benefits most from prompting: Llama-3.1-8B records the largest mean gain of all 3 evaluators (+0.073). This impression is misleading. The gain is concentrated in 2 categories, *Off-Topic* and *Discriminating*, in which the model's comment-only baseline barely predicts the positive class at all (F_1 of 0.000 and 0.007). Starting from such a degenerate baseline, almost any change to the prompt produces an improvement.

Removing the 2 degenerate categories, *Off-Topic* and *Discriminating*, confirms this reading. Without them, the mean gain of Llama-3.1-8B drops from +0.073 to +0.036, essentially the level of Gemma-4-31B (+0.034). Moreover, what the model recovers to is not balanced judgment: under retrieval with full context it flags almost every comment, reaching a mean recall of 0.924 at a mean precision of only 0.182. Gemma-4-31B, at a precision of 0.520 and a recall of 0.654, is the only evaluator whose decision threshold comes close to that of the human moderators.

The 2 evaluators are therefore fit for different roles: only Gemma-4-31B is calibrated closely enough to the human threshold to generate silver labels unmonitored, whereas the high-recall Llama-3.1-8B setting is defensible as a triage filter ahead of human review, where a reviewer absorbs its false positives.

The best strategy is itself model-dependent: Gemma-4-31B peaks under retrieval with full context (0.570), Llama-3.1-8B under retrieval *without* context (0.338, ahead

of 0.301 with it), and Llama-3.2-1B barely above doing nothing (0.238 against 0.228). A prompt design validated at one scale does not transfer to another.

5.4 Comparison with Supervised Systems on the Same Corpus

Because this study trains no supervised model, the reference point for absolute agreement is provided by prior work on the same expert-annotated subset. Schabus et al. [3] and Yadav and Milde [12] both report minority-class F_1 per category on the same published 10-fold splits, which makes their results directly comparable to ours at the level of both the metric and the data. Table 12 reproduces their figures alongside ours; Figures 1 and 2 display the same baselines as horizontal lines next to our per-category results.

The comparison divides cleanly, and the division is not arbitrary. On the 3 categories that govern removal decisions, the LLM evaluator exceeds every previously reported system: *Off-Topic* (0.519 against a prior best of 0.415), *Inappropriate* (0.452 against 0.247) and *Discriminating* (0.384 against 0.173). These margins are not marginal; on *Discriminating* the figure is more than double the best previously published result. They are also obtained under a *single* prompting strategy, retrieval combined with full context, rather than by selecting the best prompting strategy per category, so they do not depend on tuning against the test labels.

On the 3 categories associated with constructive content the supervised systems remain ahead: *Personal Stories* (0.844 against our 0.743), *Possibly Feedback* (0.781 against 0.634) and *Arguments Used* (0.750 against 0.686). The gap is consistent across all 7 of our prompting strategies and is therefore a property of the task rather than of the prompt.

The 2 blocks differ in a way that explains the split. The categories we win have the lowest positive-class prevalence in the corpus, between 8 % and 16 %, so a supervised model sees between roughly 280 and 580 positive examples in total and far fewer per fold. The categories we lose have the highest prevalence, between 17 % and 28 %, and are also the most surface-detectable: personal narration and explicit argumentation leave lexical traces that fine-tuning captures well. Scarcity of positive supervision, not category difficulty as such, is what separates the 2 groups.

This is the low-resource argument in concrete form. It also understates the case, because the comparison is not like-for-like in our favor: the systems in columns 2–7 are trained on 9 folds of in-domain expert annotation, whereas our prompting strategies receive no annotated training data at all and, in the retrieval variants, only two retrieved examples per decision. Matching or exceeding a fine-tuned transformer on three of six categories without any parameter updates is what makes the approach viable for outlets, communities and moderation policies for which no annotated corpus exists — which is to say, for almost all of them.

The practical reading is a division of labor rather than a replacement. Where an annotated corpus already exists and the target category is well represented in it, a fine-tuned encoder remains the stronger and cheaper option. Where it does not, a retrieval-augmented LLM evaluator delivers usable agreement immediately, and its pool of retrievable decisions can be extended with new ones at negligible cost, whereas a fine-tuned model must be retrained.

Table 12 Minority-class F_1 on the expert-annotated subset of the One Million Post Corpus. Columns 2–4 are the baselines reported by Schabus et al. [3]; columns 5–7 the supervised transformer results of Yadav and Milde [12]. The final two columns are Gemma-4-31B in this study, first under a single fixed prompting strategy (RAG combined with full context) and then under the best of the 7 strategies evaluated. Best value per row in bold.

	Schabus et al. [3]			**Yadav & Milde [12]**			**This work**	
Category	BOW (com.)	D2V (com.)	LSTM (com.)	BERT (com.)	BERT (tit.+com.)	fBERT (tit./com.)	RAG + full ctx.	best config.
Off-Topic	0.255	0.310	0.268	0.339	0.415	0.368	**0.519**	**0.519**
Inappropriate	0.132	0.218	0.140	0.247	0.217	0.235	0.452	**0.456**
Discriminating	0.132	0.173	0.126	0.171	0.089	0.163	0.384	**0.386**
Possibly Feedback	0.616	0.601	0.630	0.719	**0.781**	0.771	0.634	0.647
Personal Stories	0.640	0.706	0.651	0.832	**0.844**	0.836	0.743	0.743
Arguments Used	0.563	0.610	0.604	0.725	0.748	**0.750**	0.686	0.693

6 Limitations

While this study provides a comprehensive evaluation of contextual and retrieval-augmented approaches for German-language content moderation, several limitations should be considered when interpreting the results.

The first limitation relates to the evaluation of Retrieval-Augmented Generation (RAG). The study relies on a single retrieval strategy based on Semantic Textual Similarity (STS) with one positive and one negative example. This restricts the ability to draw conclusions about alternative retrieval designs and their respective performance. We fixed this design deliberately: systematically varying retrieval strategy (e.g., number of examples, retrieval method, positive/negative balance) would multiply the experimental grid combinatorially across all 3 models, and our primary goal was to establish whether a minimal, realistically deployable retrieval configuration helps at all, rather than to optimize retrieval design itself.

Second, the generative models were primarily evaluated in zero-shot settings. The absence of systematically included fixed few-shot or one-shot prompting baselines limits the ability to disentangle performance gains from retrieval augmentation versus in-context learning effects.

Third, the study evaluates a relatively limited number of language models. To make stronger and more generalizable claims regarding the specific effect of model size on the efficacy of contextual moderation, future research would need to systematically compare a broader spectrum of model families across a wider range of parameter counts. For the current study, the ensemble of models was deliberately restricted to keep overall inference time within feasible bounds, given the extensive computational demands of the existing experimental matrix.

Fourth, this study only used the One Million Post Corpus. We specifically selected this dataset due to its rich contextual features, and its established presence in comparable reference literature, which facilitates robust benchmarking. Nevertheless, it would be highly valuable for future research to validate these findings across diverse datasets.

Finally, the reference annotations are themselves imperfect. Inter-annotator agreement among the professional moderators ranged from $\kappa = 0.3$ to 0.6, which places a

ceiling on attainable agreement: part of the residual disagreement we report reflects genuine ambiguity in the categories rather than evaluator error. We report agreement against the released majority labels and do not have per-annotator judgments that would place the evaluator on the same scale as the individual moderators.

7 Conclusion

This article examined whether LLMs can serve as evaluators of content moderation judgments in a low-resource, non-English setting. Using the expert-annotated part of the One Million Post Corpus, drawn from the discussion forums of the Austrian newspaper *Der Standard*, we compared the judgments of 3 generative models of different sizes with those of professional moderators across 6 moderation categories.

We varied the contextual information supplied to the model over 5 levels, from the isolated comment to the full article and conversational thread, and added 2 retrieval-augmented variants in which the model is shown the most similar prior human moderation decisions. The resulting judgments were compared with the expert annotations and with the supervised systems previously published on the same corpus. Four key findings emerged.

Context helps, but not in proportion to its volume. The conversation a comment replies to is more informative than the much longer article text, and no single context level is best for every category: categories defined relative to the article improve when the article is supplied, whereas categories that describe the comment itself are judged less accurately when it is added.

Retrieving prior moderation decisions is as effective as supplying full context, at a fraction of the prompt length. Two short retrieved examples — the most similar comment judged positively and the most similar comment judged negatively — match the benefit of the complete context, and combining the two strategies is better than either alone. Retrieval is also the more dependable of the two: it improves every moderation category, and it helps most in the categories where the written annotation guideline leaves the decision boundary most open.

The model matters more than the prompting strategy. The difference between the strongest and the weakest model is several times larger than the difference between prompting strategies within any single model. Larger models also profit more from additional information: the same context and retrieved examples that help the largest model leave the smallest one unchanged or worse, so a prompt design validated at one scale does not transfer to another. Only the largest model applies a threshold close to that of the human moderators, whereas the mid-sized model flags far too much content and is therefore suitable only as a pre-filter ahead of human review, not as an independent source of labels.

LLM evaluation pays where annotation is scarcest. Compared with the supervised systems previously published on this corpus, our approach is stronger on the 3 rare categories that govern removal decisions and weaker on the 3 frequent categories that describe constructive content — and it achieves this without any annotated training data.

For newsroom practice this implies 3 things: treat the prompting strategy as a per-category empirical question rather than a fixed pipeline; weigh gains against the inference cost of longer inputs, which on this corpus favors retrieving prior decisions over expanding context; and retain fine-tuned encoders where an annotated corpus already exists, since that is exactly where they remain ahead.

Several directions follow from the limitations above. Alternative retrieval strategies, larger sets of retrieved examples and controlled few-shot baselines are needed to separate the effect of retrieval from that of in-context examples as such; evaluator reliability should be measured more fully, by checking whether a model returns the same judgment on repeated runs and how far its judgments depend on the exact wording of the prompt; and the human-in-the-loop setting raises a question this study cannot answer, namely which evidence from an evaluator actually helps a moderator decide better.

The case for automating moderation judgment rests on knowing when such judgments can be trusted. In this setting that trust is conditional, and the conditions are clear: it depends on the category, on how the judgment is elicited, and above all on the capability of the model.

Declarations

Funding. This research was funded by the netidee Call 19 project 7207 "Hate Speech Prevention Through Imitation".

Competing interests. The authors declare no competing interests.

Data availability. The One Million Post Corpus analyzed in this study is publicly available at https://ofai.github.io/million-post-corpus under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license [3]. No new data were generated in the course of this study.

Materials and code availability. All code, prompt templates and evaluation scripts required to reproduce the experiments reported in this article are openly available at https://github.com/FelixKrejca/HaSPI_content_moderation_w_context.

Ethics, Consent to Participate, and Consent to Publish declarations. Not applicable. This study analyzes an existing, publicly released and anonymized corpus and involved no human participants, human subjects or animals.

References

[1] Drolsbach, C.P., Pröllochs, N.: Content moderation on social media in the eu: Insights from the dsa transparency database. In: Companion Proceedings of the ACM Web Conference 2024, pp. 939–942 (2024). https://doi.org/10.1145/3589335.3651482

[2] Council, E.P..: Regulation (EU) 2022/2065 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act). Official Journal of

the European Union, L 277 (2022). https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[3] Schabus, D., Skowron, M., Trapp, M.: One million posts: A data set of german online discussions. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1241–1244 (2017). https://doi.org/10.1145/3077136.3080711

[4] Yu, X., Blanco, E., Hong, L.: Hate speech and counter speech detection: Conversational context does matter. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5918–5930 (2022). https://doi.org/10.18653/v1/2022.naacl-main.433

[5] Pérez, J.M., Luque, F.M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P.S., Zajac, J., Miguel, P., Debandi, N., Gravano, A., *et al.*: Assessing the impact of contextual information in hate speech detection. IEEE Access **11**, 30575–30590 (2023) https://doi.org/10.1109/access.2023.3258973

[6] Krejca, F., Kietreiber, T., Buchelt, A., Neumaier, S.: Context-aware content moderation for german newspaper comments. arXiv preprint arXiv:2505.20963 (2025) https://doi.org/10.48550/arXiv.2505.20963

[7] Nayak, R., Joshi, R.: Contextual hate speech detection in code mixed text using transformer based approaches. arXiv preprint arXiv:2110.09338 (2021) https://doi.org/10.48550/arXiv.2110.09338

[8] Ghosh, S., Suri, M., Chiniya, P., Tyagi, U., Kumar, S., Manocha, D.: Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 6159–6173 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.377

[9] Chen, J., Shen, E., Bavalatti, T., Lin, X., Wang, Y., Hu, S., Subramanyam, H., Vepuri, K.S., Jiang, M., Qi, J., *et al.*: Class-rag: Real-time content moderation with retrieval augmented generation. arXiv preprint arXiv:2410.14881 (2024) https://doi.org/10.48550/arXiv.2410.14881

[10] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., *et al.*: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) https://doi.org/10.48550/arXiv.2407.21783

[11] Google DeepMind: gemma-4-31b-it. Hugging Face (2026). https://huggingface.co/google/gemma-4-31b-it

[12] Yadav, A., Milde, B.: forumbert: Topic adaptation and classification of contextualized forum comments in german. In: Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), pp. 193–202 (2021). https://aclanthology.org/2021.konvens-1.17/

[13] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023). https://doi.org/10.48550/arXiv.2306.05685

[14] Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30), 2305016120 (2023) https://doi.org/10.1073/pnas.2305016120

[15] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? Computational Linguistics **50**(1), 237–291 (2024) https://doi.org/10.1162/coli_a_00502

[16] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9440–9450 (2024). https://doi.org/10.18653/v1/2024.acl-long.511

[17] Panickssery, A., Bowman, S.R., Feng, S.: LLM evaluators recognize and favor their own generations. In: Advances in Neural Information Processing Systems (NeurIPS) (2024). https://doi.org/10.48550/arXiv.2404.13076

[18] Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society **7**(1), 2053951719897945 (2020) https://doi.org/10.1177/2053951719897945

[19] Grimmelmann, J.: The virtues of moderation. Yale JL & Tech. **17**, 42 (2015). https://yjolt.org/virtues-moderation

[20] Mosca, E., Wich, M., Groh, G.: Understanding and interpreting the impact of user context in hate speech detection. In: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (SocialNLP), pp. 91–102 (2021). https://doi.org/10.18653/v1/2021.socialnlp-1.8

[21] Nagar, S., Barbhuiya, F.A., Dey, K.: Towards more robust hate speech detection: using social context and user data. Social Network Analysis and Mining **13** (2023) https://doi.org/10.1007/s13278-023-01051-6

[22] Davidson, T.: Multimodal large language models can make context-sensitive hate

speech evaluations aligned with human judgement. Nature Human Behaviour **10**(3), 514–530 (2026) https://doi.org/10.1038/s41562-025-02360-w

[23] Schmid, L., Däniken, P., Giedemann, P., Tuggener, D., Bühler, J., Kamenowski, M., Girschick, K., Baier, D., Cieliebak, M.: HICC: A dataset for German hate speech in conversational context. In: Wartena, C., Heid, U. (eds.) Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers, pp. 168–176. HsH Applied Academics, Hannover, Germany (2025). https://aclanthology.org/2025.konvens-1.15/

[24] Lee, K., Ram, S.: Leveraging large language models for hate speech detection: Multi-agent, information-theoretic prompt learning for enhancing contextual understanding. Journal of Management Information Systems **42**(4), 1055–1086 (2025) https://doi.org/10.1080/07421222.2025.2561383

[25] Assis, G., Amorim, A., Carvalho, J., Oliveira, D., Vianna, D., Paes, A.: Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In: Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR), pp. 301–311 (2024). https://aclanthology.org/2024.propor-1.31/

[26] Wang, Y., Liao, Y.: Scnet: Hate speech detection leveraging syntactic information and conversational contexts. In: 2024 7th International Conference on Machine Learning and Natural Language Processing (MLNLP), pp. 1–6 (2024). https://doi.org/10.1109/mlnlp63328.2024.10800103 . IEEE

[27] Han, L., Tang, H.: Designing of prompts for hate speech recognition with in-context learning. In: 2022 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 319–320. IEEE, ??? (2022). https://doi.org/10.1109/csci58124.2022.00063

[28] Weber, M., Huber, M., Auch, M., Döschl, A., Keller, M.-E., Mandl, P.: Digital guardians: Can gpt-4, perspective api, and moderation api reliably detect hate speech in reader comments of german online newspapers? In: 2025 38th Conference of Open Innovations Association (FRUCT), pp. 318–326 (2025). https://doi.org/10.23919/fruct67853.2025.11239100 . IEEE

[29] Keller, M.-E., Auch, M., Döschl, A., Vlk, F., Quernheim, J., Hartmann, M., Mandl, P., Kaul, A., Franz, M.: Hocon34k: A corpus of hate speech in online comments from german newspapers. In: International Conference on Information Integration and Web Intelligence, pp. 212–226 (2024). https://doi.org/10.1007/978-3-031-78090-5_18 . Springer

[30] Assenmacher, D., Niemann, M., Müller, K., Seiler, M., Riehle, D.M., Trautmann, H.: Rp-mod, rp-crowd moderator-and crowd-annotated german news comment datasets. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[31] Goldzycher, J., Röttger, P., Schneider, G.: Improving adversarial data collection by supporting annotators: Lessons from gahd, a german hate speech dataset. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 4405–4424 (2024). https://doi.org/10.18653/v1/2024.naacl-long.248

[32] Justen, L., Müller, K., Niemann, M., Becker, J.: No time like the present: Effects of language change on automated comment moderation. arXiv preprint arXiv:2207.04003 (2022) https://doi.org/10.48550/arXiv.2207.04003

[33] Bremm, F., Blaneck, P.G., Bornheim, T., Grieger, N.: Detecting sexism in german online newspaper comments with open-source text embeddings (team gda, germeval2024 shared task 1: Germs-detect, subtasks 1 and 2, closed track). In: Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024), pp. 33–38 (2024). https://aclanthology.org/2024.germeval-2.5/

[34] Krenn, B., Petrak, J., Kubina, M., Burger, C.: Germs-at: A sexism/misogyny dataset of forum comments from an austrian online newspaper. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 7728–7739 (2024). https://doi.org/10.63317/5emqc8qourkr

[35] Kotarcic, A., Hangartner, D., Gilardi, F., Kurer, S., Donnay, K.: Human-in-the-loop hate speech classification in a multilingual context. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 7414–7442 (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.548

[36] Gross, S., Petrak, J., Venhoff, L., Krenn, B.: Germeval2024 shared task: Germs-detect–sexism detection in german online news fora. In: Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024), pp. 1–9 (2024). https://aclanthology.org/2024.germeval-2.1/

[37] Pachinger, P., Goldzycher, J., Planitzer, A., Kusa, W., Hanbury, A., Neidhardt, J.: Austrotox: A dataset for target-based austrian german offensive language detection. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 11990–12001 (2024). https://doi.org/10.18653/v1/2024.findings-acl.713

[38] Geiss, C., Zarcone, A.: Thaugs at germeval 2024 (shared task 1: Germs-detect): Predicting the severity of misogyny/sexism in forum comments with bert models (subtask 1, closed track and additional experiments). In: Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024), pp. 10–20 (2024). https://aclanthology.org/2024.germeval-2.2/

[39] Akomeah, K.O., Kruschwitz, U., Ludwig, B.: Team quabynar at the germeval 2024 shared task 1 germs-detect (subtasks 1 and 2) on sexism detection. In: Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024), pp. 26–32 (2024). https://aclanthology.org/2024.germeval-2.4/

[40] Ruiter, D., Reiners, L., d'Sa, A.G., Kleinbauer, T., Fohr, D., Illina, I., Klakow, D., Schemer, C., Monnier, A.: Placing m-phasis on the plurality of hate: a feature-based corpus of hate online. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 791–804 (2022). https://doi.org/10.63317/3ibvx8skdy85

[41] Arslan, M., Ghanem, H., Munawar, S., Cruz, C.: A survey on rag with llms. Procedia computer science **246**, 3781–3790 (2024) https://doi.org/10.1016/j.procs.2024.09.178

[42] Şakar, T., Emekci, H.: Maximizing rag efficiency: A comparative analysis of rag methods. Natural Language Processing **31**(1), 1–25 (2025) https://doi.org/10.1017/nlp.2024.53

[43] Song, Y., Wang, G., Li, S., Lin, B.Y.: The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 4195–4206 (2025)

[44] Vera, H.S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S.R., Smoot, S., Naim, I., Zou, J., Chen, F., *et al.*: Embeddinggemma: Powerful and lightweight text representations. arXiv preprint arXiv:2509.20354 (2025) https://doi.org/10.48550/arXiv.2509.20354

Appendix A Complete Result Matrix

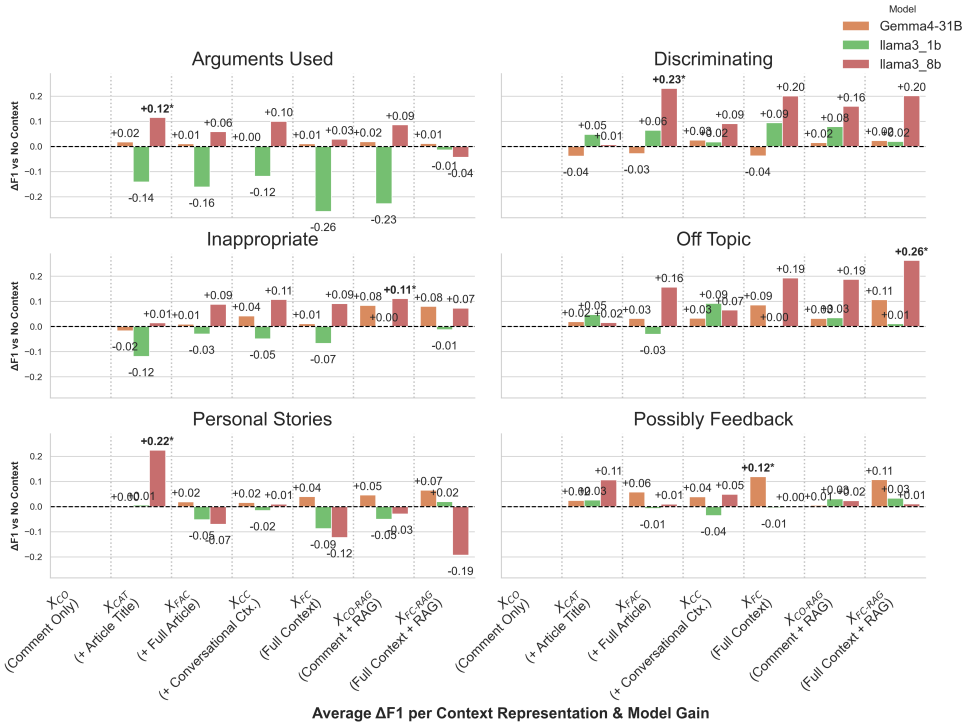
Table A1 reports the full set of results: minority-class precision, recall and F_1 for each of the 3 evaluators under each of the 7 prompting strategies, across all 6 moderation categories. Every figure quoted in Section 5 is derived from this table. Bold marks the best value in each column, so a bold entry identifies the highest precision, recall or F_1 obtained for that category by any model under any strategy; because the three metrics are marked independently, bold entries within a category need not belong to the same prompting strategy.

Table A1 Minority-class evaluation metrics for every evaluator, prompting strategy and moderation category. P = precision, R = recall. The best value in each column is set in bold.

Model	Context	Arguments Used			Discriminating			Inappropriate			Off-Topic			Personal Stories			Possibly Feedback		
		P	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
Gemma-4-31B	Full context (X_{FC})	0.632	0.747	0.685	0.347	0.301	0.323	0.377	0.389	0.383	**0.474**	0.526	0.498	0.565	0.980	0.717	**0.611**	0.688	**0.647**
	Full article context (X_{FAC})	0.640	0.736	0.685	0.313	0.351	0.331	0.350	0.419	0.381	0.345	0.624	0.444	0.540	0.981	0.696	0.489	0.730	0.586
	Conversational context (X_{CC})	0.582	0.806	0.676	0.348	0.433	**0.386**	0.338	0.535	0.414	0.379	0.540	0.445	0.536	0.985	0.694	0.459	0.740	0.567
	Article title (X_{CAT})	0.609	0.800	0.692	0.292	0.358	0.322	0.298	0.439	0.355	0.316	0.683	0.432	0.521	0.980	0.680	0.429	0.776	0.552
	Comment only (X_{CO})	0.588	0.791	0.674	0.312	0.425	0.360	0.292	0.512	0.372	0.317	0.591	0.412	0.516	0.984	0.677	0.403	0.766	0.528
	RAG + full context ($X_{\text{FC-RAG}}$)	0.648	0.742	0.686	**0.390**	0.396	0.384	0.415	0.515	0.452	**0.473**	0.579	**0.519**	0.621	0.971	**0.743**	0.572	0.719	0.634
	RAG + comment only ($X_{\text{CO-RAG}}$)	0.631	0.785	**0.693**	0.330	0.446	0.375	0.359	0.637	**0.456**	0.360	0.585	0.445	0.594	0.972	0.723	0.421	0.729	0.533
Llama-3.2-1B	Full context (X_{FC})	0.234	0.168	0.196	0.086	0.245	0.128	0.080	0.059	0.068	0.180	0.122	0.146	0.151	0.336	0.208	0.241	0.395	0.300
	Full article context (X_{FAC})	0.307	0.279	0.292	0.077	0.135	0.098	0.070	0.218	0.106	0.124	0.103	0.113	0.185	0.355	0.243	0.268	0.333	0.297
	Conversational context (X_{CC})	0.345	0.326	0.335	0.048	0.057	0.052	0.107	0.073	0.086	0.167	0.400	0.235	0.196	0.489	0.279	0.206	0.390	0.269
	Article title (X_{CAT})	0.334	0.294	0.313	0.145	0.057	0.082	0.061	0.010	0.017	0.142	0.291	0.191	0.240	0.407	0.302	0.231	0.587	0.331
	Comment only (X_{CO})	0.386	0.550	0.454	0.051	0.025	0.034	0.107	0.185	0.136	0.136	0.152	0.144	0.248	0.365	0.295	0.269	0.352	0.305
	RAG + full context ($X_{\text{FC-RAG}}$)	0.318	0.716	0.440	0.126	0.036	0.053	0.078	0.304	0.123	0.179	0.140	0.155	0.192	0.884	0.315	0.227	0.671	0.339
	RAG + comment only ($X_{\text{CO-RAG}}$)	0.241	0.213	0.226	0.080	0.188	0.112	0.099	0.222	0.136	0.176	0.183	0.178	0.180	0.385	0.245	0.250	0.510	0.336
Llama-3.1-8B	Full context (X_{FC})	0.372	0.957	0.536	0.137	0.433	0.207	0.122	0.752	0.211	0.153	0.262	0.193	0.242	0.991	0.390	0.217	0.856	0.347
	Full article context (X_{FAC})	0.412	0.902	0.566	0.174	0.372	0.238	0.122	0.700	0.208	0.151	0.164	0.157	0.285	0.983	0.442	0.226	0.829	0.355
	Conversational context (X_{CC})	0.470	0.857	0.607	0.156	0.071	0.098	0.181	0.304	0.227	0.137	0.043	0.066	0.377	0.849	0.522	0.270	0.732	0.394
	Article title (X_{CAT})	0.666	0.584	0.622	0.286	0.007	0.014	0.267	0.089	0.134	0.114	0.009	0.016	0.653	0.846	0.737	0.423	0.484	0.452
	Comment only (X_{CO})	0.695	0.399	0.507	0.077	0.004	0.007	0.277	0.076	0.119	0.000	0.000	0.000	**0.878**	0.362	0.512	0.533	0.255	0.345
	RAG + full context ($X_{\text{FC-RAG}}$)	0.303	**0.998**	0.465	0.119	**0.827**	0.208	0.108	**0.914**	0.193	0.157	**0.812**	0.263	0.190	**0.999**	0.320	0.216	**0.992**	0.355
	RAG + comment only ($X_{\text{CO-RAG}}$)	0.447	0.893	0.593	0.190	0.174	0.166	0.156	0.442	0.230	0.154	0.245	0.187	0.324	0.962	0.483	0.227	0.975	0.369

Appendix B Change in F_1 Relative to the Comment-Only Baseline

Figure B1 recasts the results of Table A1 as a change relative to each model's own comment-only configuration, broken down by category. It shows per category what Table 11 reports per strategy: the gains of Gemma-4-31B are small but positive in 32 of 36 cells, those of Llama-3.1-8B are larger but concentrated in the categories where its comment-only baseline is degenerate, and those of Llama-3.2-1B are positive in only 16 of 36 cells.



	X_{CAT}	X_{FAC}	X_{CC}	X_{FC}	$X_{CO\text{-}RAG}$	$X_{FC\text{-}RAG}$	Average Gain
Gemma4-31B	+0.002	+0.017	+0.026	+0.038	+0.034	+0.066	**+0.030**
llama3_1b	-0.022	-0.036	-0.018	-0.054	-0.022	+0.010	**-0.024**
llama3_8b	+0.081	+0.079	+0.070	+0.065	+0.090	+0.052	**+0.073**
Global Average	**+0.020**	**+0.020**	**+0.026**	**+0.017**	**+0.034**	**+0.042**	**+0.027**

Fig. B1 ΔF_1 relative to the comment-only configuration of the same model, per moderation category, with per-model and global averages. Positive values indicate that the contextual or retrieval-augmented variant agrees more closely with the expert annotations than the same model given the comment alone.