Severity-Dependent Bias in LLM Evaluators: A Span-Level Audit of Polarizing Language Detection in Everyday News

Prerana Khatiwada*
preranak@udel.edu
University of Delaware
Newark, DE, USA

Kathleen Higgins*
kathigg@udel.edu
University of Delaware
Newark, DE, USA

Ashrey Mahesh
mahesha@udel.edu
University of Delaware
Newark, DE, USA

Varun Pappu
varunp@udel.edu
University of Delaware
Newark, DE, USA

Benjamin E. Bagozzi
bagozzib@udel.edu
University of Delaware
Newark, DE, USA

Matthew Louis Mauriello
mlm@udel.edu
University of Delaware
Newark, DE, USA

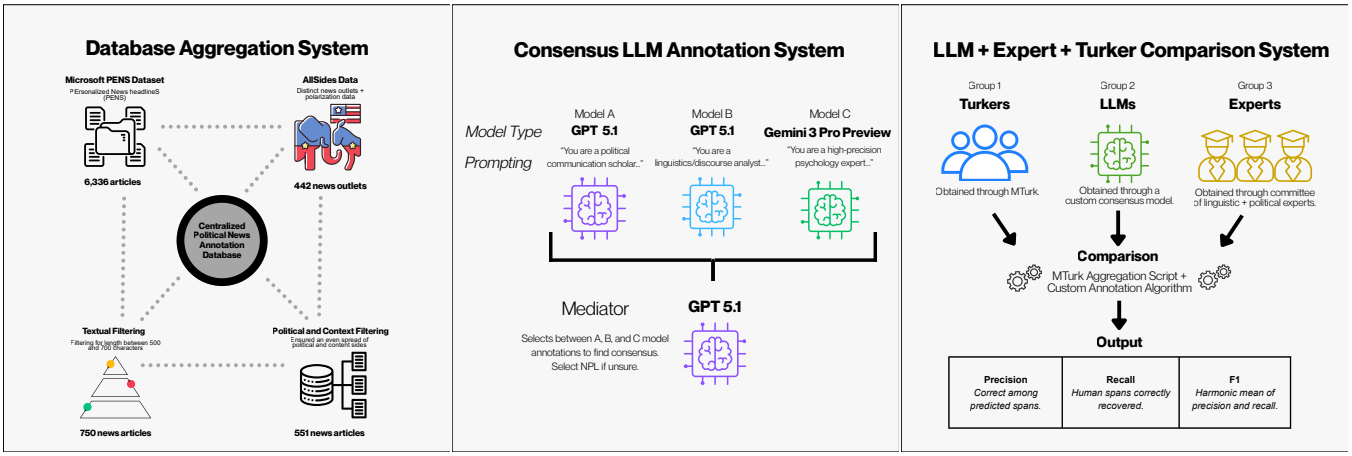

Figure 1: Dataset Construction and Consensus LLM Annotation Framework. *Left*: We construct a centralized news annotation database by aggregating news articles from the Microsoft PENS dataset and AllSides *Center*: Three differently prompted LLMs generate span-level annotations, and a mediator adjudicates consensus across model outputs. *Right*: LLM consensus annotations are compared against crowd (MTurk) and expert committee annotations using a custom aggregation algorithm.

Abstract

As large language models (LLMs) are increasingly deployed as evaluators of online content, questions arise about their reliability as auditing agents, particularly in high-stakes domains such as misinformation detection. We develop a human-informed taxonomy to classify polarizing language and examine whether safety-aligned (i.e., RLHF- and instruction-tuned) commercial LLMs exhibit severity-dependent detection patterns. We operationalize severity using two label categories, persuasive propaganda (lower severity) and inflammatory language (higher severity), and measure detection patterns as variation in model labeling across these categories in everyday news media. Using a multi-annotator LLM committee with adjudication, we compare LLM performance against both expert in-house annotations and real-world crowd annotations in a free-form minimum-one-span setting, where at least one annotation must be made per paragraph.

In pilot evaluations of our annotation tool and span-level pipeline, LLMs achieve an article-level F1 of 0.707 and category-level F1 of 0.931 relative to expert annotations. Even so, agreement drops against crowd annotations (article-level F1 = 0.530). Low human-human agreement (article-wise F1 = 0.49; α = 0.07) highlights the inherent difficulty of identifying polarizing rhetoric and suggests closer alignment between LLM and expert interpretations. Category-level analyses further reveal lower recall for higher-severity inflammatory language among LLMs, indicating potential false-positive aversion under safety alignment. Rather than framing these discrepancies as model failure, we interpret them as evidence that evaluator agents require fine-grained, severity-aware auditing. As in-progress and future work, we are expanding the dataset, testing additional annotation configurations, and refining prompt calibration to improve ground-truth reliability. We also plan to examine how LLM evaluators respond to different levels of polarizing language. Overall, we contribute a human-centered evaluation framework and outline directions for severity-aware meta-evaluation of LLM-based auditing pipelines.

*These authors contributed equally to this work.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; Information extraction; • **Human-centered computing** → **Collaborative and social computing**;

Keywords

LLM Evaluation, News, Bias, AI Auditing, Polarizing Language , Inflammatory Detection, Misinformation, Evaluator Agents

ACM Reference Format:

Prerana Khatiwada, Kathleen Higgins, Ashrey Mahesh, Varun Pappu, Benjamin E. Bagozzi, and Matthew Louis Mauriello. 2026. Severity-Dependent Bias in LLM Evaluators: A Span-Level Audit of Polarizing Language Detection in Everyday News. In *Human-centered Evaluation and Auditing of Language Models (HEAL Workshop @ CHI 2026), April 15, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 15 pages.

1 Background and Introduction

Fact-checking does not scale [1, 19]. Millions of news articles are published annually, while misinformation continues to erode public trust in journalism and democratic discourse [4, 34, 41]. Although explicit false claims and misinformation can often be verified through claim-level fact-checking [49], much contemporary online manipulation operates implicitly, through framing, exaggeration, emotionally charged rhetoric, and subtle provocation [29, 31, 39]. This form of implicit misinformation is partially true at the sentence level, yet misleading at the narrative level, making it resistant to traditional verification pipelines.

One scalable proxy for implicit manipulation is *polarizing language* [13]. Since polarizing language (persuasive propaganda and inflammatory rhetoric) co-occurs with misinformation [10], automated detection of linguistic patterns of rhetorical distortion may offer a scalable proxy approach that decreases the time of claim-by-claim fact-checking [21, 44, 50]. These signals range from structured persuasion techniques to emotionally provocative discourse that escalates tension or division [14, 16]. Given this need for scalable detection, LLMs are increasingly deployed as evaluator agents in moderation [22, 30], misinformation detection [7, 8], and auditing pipelines [23]. Recent work further demonstrates their use in large-scale annotation workflows. For example, Wang et al. show that ensembles of LLMs can produce high-agreement labels for election-related harmful social media content through a "wisdom-of-the-crowd" aggregation strategy [52]. Similarly, prior work has explored LLM-assisted labeling for multimodal election misinformation datasets using consensus-based aggregation and human validation [28]. While these approaches offer broad accessibility and strong general-purpose reasoning, emerging research highlights systematic limitations in using LLMs as scalable evaluators: safety-alignment mechanisms, often operationalized through reinforcement learning from human feedback (RLHF) and related post-training techniques, may induce conservative behaviors such as false-positive aversion (FPA) or over-refusal [9, 12]. In high-uncertainty settings, these mechanisms can discourage models from engaging with or labeling high-severity charged content. A critical question, therefore, emerges: ***Can safety-aligned LLMs reliably detect high-severity rhetorical manipulation in everyday news articles?***

This question is particularly urgent given the dual pressures shaping commercial LLM deployment. On one hand, LLMs demonstrate strong zero-shot and few-shot performance in classification, stance detection, and content moderation tasks [17, 33, 37]. On the other hand, their conservative alignment objectives, designed to minimize harmful or incorrect accusations, may create blind spots in tasks that require identifying subtle but high-severity rhetorical manipulation.

This tension becomes particularly salient in the context of existing computational approaches to propaganda and persuasion detection. Prior work has primarily focused on short-form content such as tweets, comments, headlines, or isolated claims and sentences [25, 35, 48]. Most systems primarily operationalize propaganda as a binary classification problem [38] or rely on source-level reputation signals rather than span-level rhetorical analysis. However, inflammatory rhetoric often operates differently from structured propaganda. Rather than persuading through argumentation alone, inflammatory language provokes emotional arousal, escalates hostility, and fosters division, often without explicit slurs or overt policy violations [24, 47]. Existing detection frameworks frequently subsume such rhetoric under hate speech or toxicity, overlooking more subtle forms of manipulative provocation (e.g., [2, 46]).

Despite its central role in political discourse and digital polarization, the fine-grained detection of inflammatory language in full-length news articles remains underdeveloped. To address this gap, we introduce a structured, human-centered taxonomy of polarizing language grounded in political communication and misinformation scholarship (e.g., [10]). Our taxonomy distinguishes between *Persuasive Propaganda* (exaggeration, slogans, bandwagon appeals, oversimplification, doubt) and *Inflammatory Language* (name-calling, demonization, scapegoating). This framework enables standardized span-level annotation across both human and machine evaluators and supports fine-grained analysis beyond binary propaganda detection. We operationalize this taxonomy through a human-in-the-loop annotation pipeline applied to full-length news articles. We collect both expert and crowd (Amazon Mechanical Turk) annotations to examine model performance alongside variability in human judgment. We then evaluate a multi-annotator LLM committee with adjudication against these human baselines using a span-level comparison framework.

Our pilot analyses reveal three interrelated patterns that complicate simple narratives of model performance. First, LLM annotations align closely with expert judgments (article F1 = 0.707; category F1 = 0.931 under a free-form span setting), suggesting moderate expert-level reliability in early evaluations. Second, agreement drops relative to crowd annotations (article F1 = 0.530), and expert–crowd agreement is itself low (article F1 = 0.494; $\alpha = 0.065$ binary), highlighting the inherent subjectivity of rhetorical classification. Third, initial category-level analyses suggest uneven detection: while overall span overlap is strong, LLMs exhibit comparatively lower recall on higher-severity inflammatory categories, consistent with a conservative labeling tendency.

These findings are preliminary and derived from a pilot dataset of 12 selected news articles; ongoing work extends this framework to a substantially larger corpus and additional annotation configurations. Rather than framing discrepancies as model failure, we position them as evidence that evaluator agents require severity-aware,

benchmark-explicit auditing before large-scale deployment. By examining not only where LLMs agree with human annotators but also how agreement depends on the chosen benchmark and severity threshold, this work contributes to emerging efforts in human-centered auditing of AI systems and ongoing discussions within the CHI community about the responsible deployment of AI tools to mitigate misinformation while preserving transparency and human oversight. Our work addresses the following research question:

(RQ1) How reliably and transparently do AI models detect and explain inflammatory and polarizing language in full-length news articles compared to human judgment?

Our expected and preliminary contributions are fivefold: (1) a structured taxonomy of polarizing language grounded in communication theory, (2) a span-level meta-evaluation framework for auditing evaluator LLMs, (3) empirical evidence of category-level detection disparities consistent with false-positive aversion, (4) a human-in-the-loop annotation platform for rhetorical analysis, and (5) a publicly released dataset and evaluation pipeline to support future research in AI auditing and misinformation detection. By centering both expert judgment and real-world annotation variability, this work advances a more realistic model of AI-assisted misinformation detection: one that recognizes rhetorical subtlety, human disagreement, and the need for severity-aware auditing in evaluator agents.

2 Empirical Study Design

Here, we describe our evolving study design, including the construction of the news dataset, development of the human annotation platform, gold-standard aggregation procedures, and the multi-agent LLM evaluation architecture. These components are part of ongoing experiments aimed at refining span-level comparisons between human and model annotations.

2.1 Dataset Construction

We constructed our evaluation dataset from a subset of 6,336 Microsoft PENS[1] news articles integrated with AllSides Media Bias[2] metadata through a multi-stage labeling and filtering pipeline. The corpus spans a politically diverse range of outlets, including mainstream sources (The New York Times, BBC, CNN) and partisan publishers (Infowars, Breitbart, Daily Kos). To construct our dataset, we aggregated the raw PENS corpus. We used GPT-4 to classify each article into one of six predefined categories (health, legal arguments, crime, science, environmental issues, or politics). GPT-4 inferred the article's news source: in many cases, the outlet name appeared explicitly in the body, often in a copyright or attribution line at the end (e.g., "© [Outlet Name]"), which made source extraction reliable. Entries with missing or indeterminate sources ("no label") were removed to maintain dataset quality. We then matched extracted sources to canonical entries in the AllSides dataset to merge corresponding political bias ratings (left, center, right).

To ensure article comparability and reduce annotation burden, we further filtered the dataset to ensure (1) all articles were between 300 and 700 words and (2) the dataset contained a balance of articles across political representation. This process yielded 551 candidate articles. For the initial validation, we randomly selected 12 full-length news articles for annotation. These articles contained 34–38 paragraph units after segmentation and formatting normalization. Paragraph-level segmentation was adopted to enable fine-grained span-level comparison between human and model annotations [18] and to reduce annotator cognitive fatigue.

2.2 Human Annotation Platform and Procedure

We developed a custom web-based annotation tool using React and Firebase and deployed it as a Human Intelligence Task (HIT) on Amazon Mechanical Turk (MTurk) [3]. A screenshot of the annotation tool is provided in Appendix B. The platform is deployed via Vercel for scalable access and iterative experimentation. The interface allowed annotators to read full-length news articles and highlight spans of 4–25 words corresponding to rhetorical strategies. Crowd workers were provided access to the full article for contextual reference, but content in paragraph-level units allowed annotators to focus on localized rhetorical signals without being overwhelmed by the full article at once [11]. For each highlighted span, annotators assigned one of three primary labels: **i) Inflammatory Language (IL)**, **ii) Persuasive Propaganda (PP)**, and **iii) No Polarizing Language (NPL)**, along with their respective subcategories. Each label was accompanied by explicit definitions and illustrative examples (e.g., neutral, positive, negative framing distinctions), developed through close linguistic analysis of subtle political framing patterns in consultation with a political science expert and three research team members. See Figure 2 for the complete label schema.

We recruited multiple batches totaling 36 MTurk participants. We required Turkers to hold a Master's qualification and maintain a high approval rating. No additional screening or exclusion criteria were applied based on demographic or other participant data. Each participant annotated one article. With three independent annotators per article, we obtained three annotations for each of the 12 articles. Before accessing the annotation task, participants were required to complete onboarding instructions and view a short voice-dubbed tutorial video embedded in the annotation platform.

2.3 Aggregation and Gold Standard Dataset

To construct a structured human reference dataset, raw span annotations were conservatively clustered prior to evaluation. Spans were merged only when they referred to the same article and paragraph index, when one normalized span contained the other, and when they shared at least two overlapping non-stopword tokens. To ensure stability, spans required support from at least two annotators to be retained. Paragraphs without qualifying spans defaulted to *No Polarizing Language (NPL)*. This conservative clustering procedure produced the gold standard dataset used for subsequent evaluation. See Algorithm 1 for details on span-level agreement computation and the construction of the gold standard dataset.

Initial inter-annotator agreement was assessed separately on the raw pre-consolidation annotations using Krippendorff's α, a reliability coefficient [32] appropriate for multiple annotators and nominal categories. Agreement was computed at the paragraph level ($n = 38$ units) under both a three-class taxonomy (*NPL*, *Persuasive*, *Inflammatory*) and a binary collapse (*Polarizing* vs. *NPL*).

[1] https://www.kaggle.com/datasets/divyapatel4/microsoft-pens-personalized-news-headlines

[2] https://www.allsides.com/

[3] https://www.mturk.com/

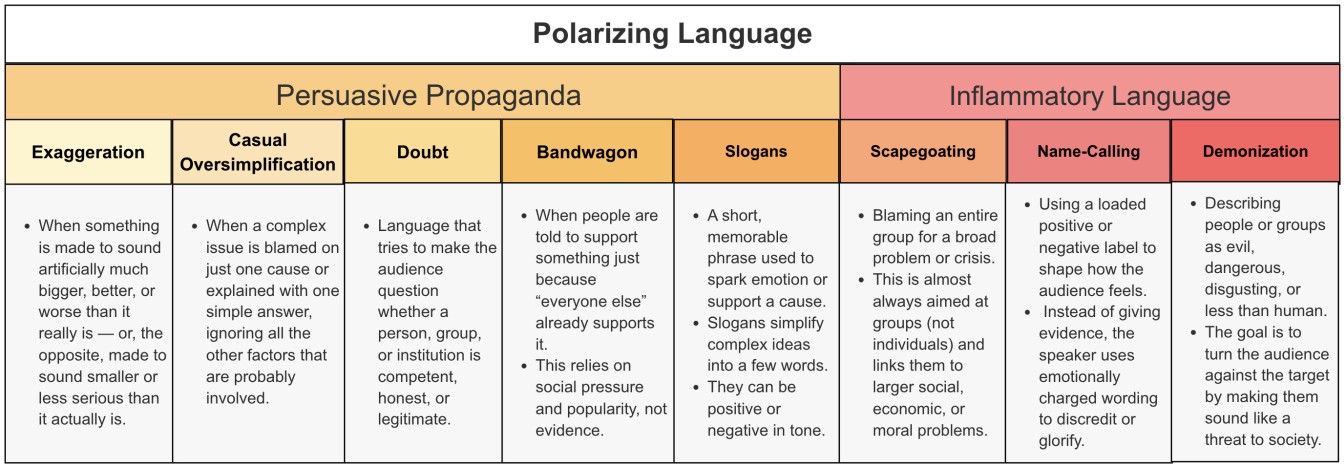

Figure 2: The taxonomy of polarizing language was developed from prior literature on misinformation and propaganda [10, 21, 44], and refined through consultation with an expert political scientist and three research team members. ***(See Appendix E for complete definitions, detailed explanations of each category and subcategory, and illustrative examples.)***

Algorithm 1 MTurk Span Aggregation and Gold Standard Construction

Require: Paragraphs $p_1, \ldots, p_m$; annotations A where each entry is (i, t, s, c): paragraph i, Turker t, span s, label $c \in \{\text{IL}, \text{PP}\}$
Ensure: Gold standard G

$G \leftarrow \emptyset$
for $i \leftarrow 1$ **to** m **do**
  Let A_i be all annotations in paragraph p_i
  Cluster spans in A_i using $\textsc{Match}(\cdot, \cdot)$
  for each cluster C **do**
    $\text{support}(C) \leftarrow$ number of unique Turkers who contributed a span to C
    if $\text{support}(C) \geq 2$ **then**
      $s \leftarrow$ representative span from C
      $c \leftarrow$ majority label in C
      Add (i, s, c) to G
    end if
  end for
  if no span was added to G for paragraph p_i **then**
    Add paragraph label (i, NPL) to G
  end if
end for
return G

Match rule. $\textsc{Match}(s_1, s_2) = 1$ iff one normalized span contains the other and they share ≥ 2 overlapping non-stopword tokens.

Krippendorff's α was computed as $\alpha = 1 - \frac{D_o}{D_e}$, where D_o denotes observed disagreement and D_e expected disagreement based on the empirical label distribution. Agreement was calculated separately for expert and MTurk annotator groups. We additionally report total coder-label counts to contextualize category prevalence, as skewed distributions influence expected disagreement.

2.4 Multi-Agent Evaluation Architecture

To evaluate LLM performance as annotators, we implemented a committee-of-agents architecture. This approach, increasingly common in recent LLM literature, aggregates multiple independent model judgments to improve robustness and reduce variance (e.g., [40]). In this setup, LLMs serve strictly as evaluators rather than as content generators. The architecture consisted of three independent annotator agents drawn from two prominent commercial LLM families (two GPT-family and one Gemini-family), followed by a GPT-family adjudicator agent that reconciled disagreements and produced the final output. Model selection was pragmatic rather than theoretically driven. As this study represents preliminary testing, we intentionally selected models from accessible and stable families that were straightforward to deploy, integrate, and replicate within our evaluation pipeline.

Each annotator received identical system instructions containing: (1) a formalized rhetorical codebook defining Inflammatory Language (IL), Persuasive Propaganda (PP), and No Polarizing Language (NPL); (2) strict JSON schema constraints specifying required fields, allowed category and subcategory enums, and paragraph-index alignment; (3) extraction constraints limiting spans to 4–25 words; and (4) a conservative labeling directive requiring NPL when uncertainty or ambiguity was present.

Specifically, annotators were instructed: "*Extract exact spans (4–25 words), no ellipses or paraphrasing. Annotate per paragraph. Be conservative: label only Persuasive Propaganda or Inflammatory Language when the language is explicit and clearly matches a definition; if unsure, choose No Polarizing Language. Return only valid JSON, with no additional explanation.*"

To introduce controlled variation in interpretive emphasis, we assigned each annotator a distinct disciplinary persona through explicit role framing. Annotator A was instructed: *"You are **Annotator A**, a political communication scholar. Strictly follow the codebook and JSON schema. Be conservative: if unsure, choose No Polarizing Language."* **Annotator B** was instructed: *"You are Annotator B, a linguistics and discourse analyst. Your strength is correct subcategory selection. Be conservative: avoid over-labeling; if unsure, choose No Polarizing Language."* **Annotator C** was instructed: *"You are Annotator C, a high-precision media psychology expert. Be conservative: only label when rhetoric is explicit; if unsure, choose No Polarizing Language."* Thus, we enforced a conservative labeling scheme across LLM roles that emphasized precision and deliberate annotation. The complete system prompts for all annotator and adjudicator roles are

included in Appendix A. All outputs were required to follow a predefined JSON schema. Minor structural deviations (e.g., missing fields, inconsistent casing, misplaced subcategories) were automatically normalized through repair procedures including enum standardization, category inference, NPL canonicalization, paragraph-index recovery, and metadata completion.

The adjudicator agent was framed through the following system instruction: *"You are the Adjudicator, a methods-oriented political scientist overseeing three annotators.* It received the structured JSON outputs of all three agents and was explicitly constrained to select only from spans proposed by the annotators. It was prohibited from inventing new spans and permitted to merge only exact duplicates (identical category, subcategory, span text, and paragraph index). This constraint ensured that the final annotation reflected structured consensus rather than generative synthesis.

To ensure consistent paragraph-level coverage, we implemented two configurable aggregation policies during system development: *Exact-One* and *Minimum-One*. Under the *Exact-One* policy, exactly one annotation is retained per paragraph. If multiple polarizing spans are predicted within a paragraph, only the most specific span (operationalized as the longest token span after normalization) is retained. If no polarizing span is identified, a canonical *No Polarizing Language (NPL)* placeholder is assigned. This policy enforces a strict one-to-one correspondence between human and model annotations at the paragraph level. Under the *Minimum-One* policy, all predicted polarizing spans within a paragraph are retained. However, each paragraph is guaranteed to contain at least one annotation; if no polarizing span is identified, an *NPL* label is assigned. Unlike *Exact-One*, this policy does not artificially restrict the number of annotations per paragraph and therefore allows one-to-many comparisons between model and human annotations. Although earlier iterations of our system used the *Exact-One* policy to simplify one-to-one evaluation, we ultimately adopted the *Minimum-One* policy for all reported analyses. This choice reflects our methodological preference for preserving the full expressive range of span-level predictions rather than constraining outputs to a single annotation per paragraph. The *Minimum-One* policy better reflects the natural distribution of rhetorical signals within paragraphs while maintaining paragraph-level coverage via the NPL fallback rule.

2.5 Evaluation Procedure

We conducted paragraph-indexed span matching between human and LLM annotations using a structured alignment procedure. During preprocessing, each article was segmented and explicitly indexed at the paragraph level. A human and LLM span were treated as a match only when (i) they referred to the same article (normalized title match), (ii) they came from the same paragraph index, (iii) one span's normalized text (lowercased, trimmed) was a contiguous substring of the other (to allow boundary expansions/contractions), and (iv) they shared at least two overlapping non-stopword tokens. To avoid inflated scores from duplicate alignments, we used greedy one-to-one matching, allowing each human span to be matched to at most one LLM span (and vice versa).

We report span-level precision, recall, and F1, category-level recall, and exact category and subcategory agreement on matched spans. See Algorithm 2 for the detailed evaluation plan. Importantly, our paragraph-level enforcement policy mitigates inflated performance by assigning the majority-class label "No Polarizing Language" (NPL). Because each paragraph is required to receive an explicit annotation, conservative overuse of NPL directly reduces recall for Inflammatory Language (IL) and Persuasive Propaganda (PP) categories. This design enables us to detect severity-dependent suppression patterns and systematic under-detection of higher-intensity rhetorical language.

3 Preliminary Findings

This section reports preliminary quantitative findings on span-level overlap between human and LLM annotations. All comparisons are conducted on the same set of 12 underlying news articles; both Turker and Expert annotations were collected on this shared article subset rather than on separate corpora. In addition to aggregate alignment measures, we analyze distributional patterns and detection skew to explore potential severity-dependent differences in annotation behavior. We report two levels of evaluation metrics. **Article-level metrics** assess whether the LLM and human annotators marked any overlapping span within the same paragraph, and compute precision, recall, and F1 based on paragraph-level overlap. **Category-level metrics** are computed only on overlapping spans and evaluate whether category and subcategory labels match, again using precision, recall, and F1. Throughout the results, we refer to two annotation datasets. The **Expert HIT** denotes our in-house annotation study conducted with trained annotators. The **Turker HIT** refers to our MTurk annotation study conducted with crowd workers. For clarity, we use the terms "Expert" and "Turker" going forward, while retaining the original identifiers when referencing specific experimental runs.

3.1 Aggregation Policy Effects

We compare two paragraph-level enforcement policies. As shown in Table 1, *Exact-One* produces higher alignment scores across both datasets. Under *Minimum-One*, alignment decreases at the article level due to increased mismatches in span overlap. Because annotation counts are constrained under *Exact-One*, precision, recall, and

Algorithm 2 Span-Level Evaluation: LLM vs Human Annotations

Require: LLM annotations L, Human annotations H
Ensure: Precision, Recall, F1
Normalize all spans (lowercase, trim whitespace)
$TP \leftarrow 0$, $FP \leftarrow 0$, $FN \leftarrow 0$
for each span $\ell \in L$ **do**
  if there exists $h \in H$ such that same article and paragraph index, substring containment, and ≥ 2 overlapping non-stopword tokens **then**
    $TP \leftarrow TP + 1$
    Mark h as matched
  else
    $FP \leftarrow FP + 1$
  end if
end for
for each unmatched span $h \in H$ **do**
  $FN \leftarrow FN + 1$
end for
$Precision \leftarrow \frac{TP}{TP+FP}$
$Recall \leftarrow \frac{TP}{TP+FN}$
$F1 \leftarrow \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$
return Precision, Recall, F1

Table 1: Comparison of paragraph-level aggregation policies across Turker (Hit) and Expert (Hit) datasets.

	Turker		Expert	
	Exact-One	Min-One	Exact-One	Min-One
Article Precision	0.579	0.500	0.842	0.659
Article Recall	0.579	0.564	0.842	0.763
Article F1	0.579	0.530	0.842	0.707
Category F1	1.000	0.909	0.969	0.931
Human Annotations	38	39	38	38
LLM Annotations	38	44	38	38

F1 are identical under that policy. The reduction under *Minimum-One* reflects structural differences in span overlap rather than a fundamental shift in annotation behavior.

3.2 LLM Alignment: Turker vs. Expert Annotations (Minimum-One)

Table 2 compares LLM alignment under the *Minimum-One* policy against Turker (crowd) and Expert (in-house) annotations. Under the *Minimum-One* policy, LLM alignment is substantially higher with Expert annotations than with Turker annotations at the article level (F1 difference = +0.177). Category-level alignment is high in both cases, with only a modest improvement relative to Experts (+0.022), suggesting that once span overlap occurs, label agreement is relatively stable. The larger gap at the article level indicates that disagreement is primarily about *whether* a paragraph is polarizing, rather than *how* it is categorized once identified. This pattern suggests that LLMs more closely approximate expert-level detection thresholds than crowd-level judgments. However, this does not necessarily imply expert-level semantic equivalence; instead, it may reflect shared conservatism or similar boundary-setting behavior. Importantly, the LLM committee produces the same number of annotations across both datasets (44), while human annotation volume differs slightly. Thus, differences in alignment are driven by qualitative span overlap rather than output quantity. Our findings highlight that model performance is highly sensitive to the choice of human benchmark and underscore the instability of "ground truth" in subjective rhetorical tasks.

To further examine the nature of these discrepancies, we analyze the raw subcategory confusion patterns between LLM predictions and pooled MTurk annotations. Figure 3 compares LLM-predicted subcategories against pooled MTurk annotations, ordered by increasing severity from top to bottom. A clear pattern of severity compression emerges: higher-severity categories (e.g., Demonization) are more frequently assigned by humans but are often mapped by the LLM to lower-severity categories (e.g., Exaggeration). As severity increases, agreement systematically declines.

3.3 Human–Human Agreement and Inter-Annotator Reliability

Agreement between Experts and Turkers is low at the paragraph (article) level (F1 = 0.494; see Table 3), indicating substantial disagreement about whether a paragraph is polarizing at all. However, at the category level, agreement on overlapping spans is perfect (F1 = 1.000), suggesting that disagreement is primarily about detection thresholds rather than category interpretation. Binary Krippendorff's α between Expert and Turker labels across shared paragraphs is 0.065, further reinforcing minimal agreement beyond chance at the binary presence/absence level. This pattern shows that the dominant source of variability lies in whether polarizing language is identified at all, rather than in how it is categorized once identified.

Table 2: LLM alignment under the Minimum-One policy: Turker vs. Expert datasets.

	Turker	Expert
Article Precision	0.500	0.659
Article Recall	0.564	0.763
Article F1	0.530	0.707
Category F1	0.909	0.931
Human Annotations	39	38
LLM Annotations	44	44

Table 3: Agreement between Turker and Expert annotations (Min. One).

Metric	Value
Article Precision	0.500
Article Recall	0.487
Article F1	0.494
Category F1	1.000
Turker Annotations	39
Expert Annotations	38

Within-group reliability is also modest (Table 4). Expert annotators demonstrate higher internal consistency than Turkers (binary α = 0.356 vs. 0.262), though both fall within ranges commonly associated with subjective interpretive tasks. The 3-class reliability scores are lower for both groups, reflecting additional ambiguity introduced by severity distinctions. Collectively, these findings suggest that polarizing language detection is a threshold-sensitive, interpretive task with limited stability across annotator populations. The instability of human agreement complicates claims of a singular

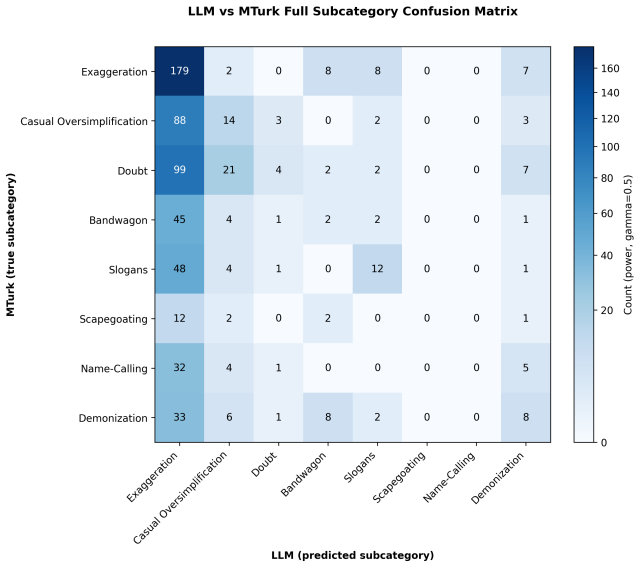

Figure 3: LLM vs. Raw MTurk subcategory confusion matrix. Rows represent pooled MTurk annotations (ordered by increasing severity from top to bottom), and columns represent LLM predictions. Darker cells indicate higher agreement.

Table 4: Krippendorff's α within annotation groups.

	3-Class α	Binary α
Expert	0.246	0.356
Turker	0.126	0.262

Table 5: Raw coder-label counts (pre-consolidation).

	Expert	Turker
NPL	84	46
Persuasive	53	78
Inflammatory	7	22
Binary NPL	84	46
Binary Polarizing	60	100

"ground truth" and underscores the importance of specifying which human benchmark LLM systems are expected to approximate.

4 Thematic analysis of Turkers rationales

Turker rationales demonstrate both thoughtful engagement and substantial variability in how polarizing language is interpreted. Many relied on what might be termed journalistic objectivity heuristics, equating balanced tone, factual detail, and attribution with the absence of polarization (e.g., "neutral, fact-based, and balanced manner," "sticks to verifiable details," "fairly represents both Democratic and Republican perspectives"). It suggests an implicit "straight-news = non-polarizing" rule, which may overlook subtle rhetorical manipulation embedded in framing, selection, or insinuation.

At the same time, when spans were flagged, decisions frequently hinged on salient surface cues such as slogans, dramatic metaphors, or promotional exaggeration (e.g., "'*Whatever it takes to win' looks like a slogan,*" "'*latest legal whiplash'… dramatizes the legal situation*," "*calling a 146% increase… 'staggering*'…"). While this indicates sensitivity to overt persuasion signals, it also suggests the possibility of overgeneralization, in which ordinary emphasis or genre conventions may be conflated with manipulation. High-severity accusations were treated more cautiously, with several workers indicating reluctance to assign strong labels unless evidence was explicit ("*I didn't think there was much polarizing language… I tried to find anything…*"). Ambiguity surrounding quoted speech further complicated judgments, as rhetoric embedded in legal complaints or political statements was sometimes viewed as reporting rather than endorsement ("*the most vivid… language comes directly from their legal complaint*"). These patterns mirror the severity-dependent detection profile observed in LLM outputs: lower-severity rhetorical devices are more readily identified, whereas higher-severity categories are underassigned due to uncertainty.

5 Discussion

Here, we discuss general observations from our findings, with particular attention to how such systems could be thoughtfully deployed in real-world news-reading environments, the central aim of this work.

5.1 LLMs as Assistive Detectors

As stated in the motivation and introduction, detecting polarizing language is inherently difficult. Under real-world conditions: limited time, cognitive load, and unfamiliarity with content, non-expert readers are unlikely to identify such language consistently. Our MTurk study was intentionally designed to simulate this reality: annotators had 15 minutes, no prior domain training beyond brief label instructions, and no opportunity for iterative calibration or discussion. Under these constraints, disagreement is not a methodological flaw but a realistic baseline that reflects the variability of non-expert judgments in applied settings.

The moderate agreement between Turkers and expert annotators (Article-level F1 = 0.494, see Section 3.3) shows that polarizing language is not universally recognized in the same way. Experts and non-experts appear to operate with meaningfully different mental models of what constitutes polarization. This divergence is critical: if the goal of deployment is to assist everyday readers, then expert-only benchmarks risk overlooking the population we intend to support. Against this backdrop, the LLM alignment patterns become more interpretable. Under the *Minimum-One* setting, LLMs align more closely with Experts (F1 = 0.707) than with Turkers (F1 = 0.530), see Section 3.1. When experts are treated as the gold standard, this suggests that contemporary LLMs can approximate expert-level annotation patterns more reliably than untrained human readers under time pressure. In other words, LLMs may already function as assistive tools that elevate non-expert detection performance toward expert norms. In deployment contexts aimed at assisting everyday readers, this raises an important design question: ***Should systems reflect expert norms, crowd intuitions, or some calibrated middle ground?***

5.1.1 ***Severity-Dependent Detection and False-Positive Aversion.*** At the same time, LLMs do not fully replicate expert nuance. The severity-ordered confusion patterns demonstrate a systematic underdetection of higher-severity polarizing categories (see Section 3.2). LLM recall decreases for higher-severity inflammatory language, even when lower-severity persuasive content is reliably flagged. This could mean models are more comfortable labeling low-severity rhetorical devices (e.g., exaggeration) than high-severity forms. This pattern also suggests a form of false-positive aversion: commercial LLMs are optimized to avoid confidently assigning high-severity labels when uncertainty exists [3, 45]. Such aversion is consistent with broader alignment training objectives, where models are rewarded for caution and hedging [43]. In contrast, expert annotators are more willing to make categorical high-severity judgments. Crucially, this is not simply a performance deficit; it is a structural property of alignment training. Conservative labeling may reduce over-flagging and preserve user trust, yet systematic under-detection of high-severity rhetoric risks attenuating precisely the forms of polarization that are most socially consequential. Thus, while LLMs appear "good enough" relative to expert benchmarks for low-to-moderate severity detection, their performance profile is not uniform across the severity spectrum.

5.1.2 ***Implications for Assistive Deployment.*** Overall, ongoing work suggests that LLMs are not replacements for expert judgment, but they may serve as practical bridges between expert standards and real-world reading conditions. If scaled datasets reproduce these patterns, deployment may be viable, particularly in assistive, decision-support roles rather than as autonomous arbiters. In practice, when a reader encounters a news article, an assistive deployment model could surface polarizing spans and provide lightweight explanations or category labels. Rather than replacing human judgment, such scaffolding would surface otherwise subtle rhetorical

cues, helping readers notice patterns they might have missed under time pressure.

Over time, repeated exposure to flagged examples and structured labels may support learning by reinforcing pattern recognition and category differentiation. Research in the learning sciences shows that spaced repetition and retrieval practice improve long-term retention and conceptual discrimination [6, 20]. Similarly, media literacy interventions suggest that guided reflection and repeated exposure to tasks can strengthen critical evaluation skills [15, 42]. In this sense, LLMs may function less as final judges and more as media literacy companions, augmenting attention, scaffolding reflection, and gradually strengthening users' independent detection skills. A natural extension of this work would be to experimentally evaluate such augmentation directly, for example, by comparing annotators operating independently versus those provided with initial LLM-flagged candidate spans or feedback. While such designs introduce challenges in maintaining annotator independence and avoiding anchoring effects, they offer a promising avenue for assessing whether AI-assisted annotation can meaningfully improve human sensitivity to high-severity rhetorical manipulation.

Future work should explore calibration strategies that retain LLM conservatism while improving sensitivity to high-severity polarization, alongside interface designs that transparently communicate uncertainty. Rather than asking whether LLMs are perfect detectors, a more pragmatic framing may be: ***Do they provide structured, benchmark-aware support that meaningfully improves detection relative to unaided reading conditions?*** Our findings suggest a cautious yes, albeit with clear and systematic limitations.

6 Final Position and Future Directions

We argue that LLM evaluator systems require category-level auditing rather than relying solely on aggregate metrics. High overall F1 scores can mask systematic underdetection in high-severity categories, particularly when safety alignment discourages confident assignment of controversial or sensitive labels. We also note that our pilot dataset (12 Microsoft PENS articles) contains relatively few overtly polarizing instances, with a strong skew toward No Polarizing Language. This class imbalance likely inflates majority agreement and suppresses recall for inflammatory content. Future work will address this limitation through more balanced sampling and targeted resampling of articles with higher concentrations of polarizing language, enabling more precise separation of safety-induced conservatism from dataset imbalance effects. Even within naturally occurring news distributions, our findings suggest that aggregate metrics alone are insufficient for evaluating evaluator reliability, as safety alignment may introduce subtle domain-specific blind spots in rhetorical analysis tasks.

As LLMs are increasingly positioned as auditing agents [53], recent work has shifted toward adaptive, model-centric evaluation frameworks. For example, FACT-AUDIT proposes dynamic multi-agent assessment of fact-checking capabilities beyond static classification metrics [36]. Our work aligns with and extends this trajectory. Rather than evaluating claim correctness, we audit span-level rhetorical detection within the full news article, with particular attention to severity-dependent suppression. In doing so, we contribute to a broader research agenda of meta-evaluating evaluator agents, examining not only what models predict but also where and under what thresholds their predictions systematically diverge.

Future work extends in four directions. First, we are scaling from the pilot subset to a 551-article corpus, enabling robust testing across diverse sources and political orientations. This expansion includes a large-scale MTurk baseline annotation phase, followed by deployment of an LLM verification workflow in which participants review and accept or reject model-generated spans. A structured disagreement step will capture corrective labels, generating data for retraining and systematic failure analysis. Second, we are exploring few-shot prompting strategies to reduce model conservatism and examine whether observed failures are driven by safety-induced bias rather than underlying capability limitations.

Third, establishing reliable ground truth remains challenging. Initial expert and MTurk annotations showed low inter-annotator agreement (Table 4), reflecting disagreement about what constitutes polarizing language. This challenge mirrors findings from prior human–AI collaborative annotation studies, where subjective tasks such as political ideology classification exhibit substantial interpretive variability across annotators [51]. Although we mitigated this using a two-of-three agreement threshold, ongoing work focuses on guided interfaces to improve non-expert consistency toward expert-level judgment. Strengthening human ground truth is a prerequisite for meaningful automated evaluation. Fourth, we envision packaging this work into a configurable annotation platform that supports evolving schemas and research goals, foregrounding flexibility rather than fixed workflows. Researchers could adapt annotation schemas, switch between multiple labeling goals, modify interaction modes, and refine workflows as their research questions evolve. The contribution would not be "**another annotation tool**," but rather a framework explicitly designed for configurability and iterative research settings.

Fifth, and most critically, our long-term goal is to integrate these findings into a Personal Informatics (PI) real-time news reading tool that generates LLM-driven, in-situ interventions as proposed in earlier work [26, 27]. Building on prior efforts to develop modular, browser-based infrastructures for misinformation analysis and intervention [27], we extend this paradigm to integrate calibrated evaluators within interactive reading environments. Rather than using LLMs solely for post-hoc auditing, we aim to embed detection models directly into the news consumption interface, highlighting polarizing language and providing contextual explanations as users engage with full articles. Prior work has shown that LLM-generated explanations can significantly influence human judgment and confidence during news annotation tasks [51], showing both the potential and the responsibility of designing explanation interfaces that support reflective reasoning rather than uncritical reliance. The objective of such a system will not be to censor content, but to support reflective consumption, promote critical reading, and increase awareness of rhetorical framing strategies. By combining validated detection pipelines with human-centered interface design, the PI tool will deliver timely, explainable feedback that helps readers recognize inflammatory and persuasive techniques in everyday news. If successful, this research will advance real-time polarizing language detection and help address the growing disparity between misinformation prevalence and scalable verification capacity [5]. By combining structured human annotation, evaluator-agent auditing,

guided interface design, and in-situ interventions, we aim to support the responsible integration of AI evaluators into digital information systems while preserving human agency and interpretive judgment.

7 Conclusion

Our pilot findings highlight a central tension in deploying safety-aligned LLMs to evaluate polarizing language. Although the overall performance of the models appeared reasonable, and aggregate alignment with expert annotations appears moderate, category-level analyses reveal uneven performance across severity levels. In particular, higher-severity inflammatory rhetoric is comparatively under-detected, suggesting a conservative labeling tendency consistent with false-positive aversion. Rather than indicating wholesale model failure, this pattern reflects a structural trade-off: the same guardrails that promote responsible behavior may attenuate detection sensitivity in evaluative contexts where identifying high-severity rhetoric is essential. At the same time, human annotation itself is unstable. Low agreement between expert and crowd annotators demonstrates that polarizing language detection is inherently threshold-sensitive and interpretive. These results complicate simplistic benchmark comparisons and show that model evaluation depends critically on which human standard is chosen. As LLMs are increasingly integrated into information systems for misinformation detection and other high-stakes decisions, this limitation becomes especially important.

Importantly, improvements in annotator consistency emerged not from model changes, but from redesigning the human annotation workflow, reminding us that evaluator auditing is as much a human-centered design problem as a modeling one. Robust automated systems depend on robust human ground truth. Beyond polarizing language detection, our findings invite broader reflection on how evaluator LLMs should be benchmarked, calibrated, and responsibly deployed in socially sensitive domains.

While safety alignment promotes responsible behavior, it may also introduce blind spots in high-severity detection tasks. Continued refinement of both human annotation workflows and evaluator-agent auditing frameworks is essential for building reliable, scalable systems for analyzing polarizing language in everyday news. We aim to prompt discussion within CHI and HEAL on building evaluator systems that are transparent, calibrated to context, and positioned as support, not authority.

Acknowledgments

We thank Qile Wang for his guidance and support throughout the project. We also thank Aarush Goyal for contributing as an expert annotator during one of the article annotation rounds. We are grateful to Dr. Kathleen McCoy and Dr. Amo Tong for their guidance as members of the dissertation committee, of which this project forms a part. We also thank Dustin Stark for his early analytical contributions to parts of this project.

References

[1] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances* 7, 36 (2021), eabf4393.

[2] Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. 2024. From online hate speech to offline hate crime: The role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–14.

[3] Debangshu Banerjee and Aditya Gopalan. 2024. Towards Reliable Alignment: Uncertainty-aware RLHF. arXiv:2410.23726 [cs.AI] doi:10.48550/arXiv.2410.23726

[4] W Lance Bennett and Steven Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication* 33, 2 (2018), 122–139.

[5] Arezo Bodaghi, Ketra A Schmitt, Pierre Watine, and Benjamin CM Fung. 2023. A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems* 11, 4 (2023), 5119–5145.

[6] Shana K Carpenter, Steven C Pan, and Andrew C Butler. 2022. The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology* 1, 9 (2022), 496–511.

[7] Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI magazine* 45, 3 (2024), 354–368.

[8] Mengyang Chen, Lingwei Wei, Han Cao, Wei Zhou, and Songlin Hu. 2023. Explore the potential of llms in misinformation detection: An empirical study. *arXiv preprint arXiv:2311.12699* (2023).

[9] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An Over-Refusal Benchmark for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 11515–11542. https://proceedings.mlr.press/v267/cui25a.html

[10] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1377–1414. doi:10.18653/v1/2020.semeval-1.186

[11] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5636–5646. doi:10.18653/v1/D19-1565

[12] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback: AD Lindström et al. *Ethics and Information Technology* 27, 2 (2025), 28.

[13] William Donohue and Mark Hamilton. 2022. A framework for understanding polarizing language. In *The Routledge handbook of language and persuasion*. Routledge, 207–223.

[14] Dennis J Downey. 2022. Polarization and persuasion: Engaging sociology in the moral universe of a divided democracy. *Sociological Perspectives* 65, 6 (2022), 1029–1051.

[15] Mira Feuerstein. 1999. Media literacy in support of critical thinking. *Journal of Educational Media* 24, 1 (1999), 43–54.

[16] Alan Fortuna. 2019. *Polarization: Rhetorical strategies in the Tea Party network*. Vol. 33. Walter de Gruyter GmbH & Co KG.

[17] Sina Gholamian, Gianfranco Romani, Bartosz Rudnikowicz, and Stavroula Skylaki. 2024. Llm-based robust product classification in commerce and compliance. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. 26–36.

[18] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. FALTE: A Toolkit for Fine-grained Annotation for Long Text Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Wanxiang Che and Ekaterina Shutova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 351–358. doi:10.18653/v1/2022.emnlp-demos.35

[19] Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. (2018).

[20] Robin F Hopkins, Keith B Lyle, Jeff L Hieb, and Patricia AS Ralston. 2016. Spaced retrieval practice increases college students' short-and long-term retention of mathematics knowledge. *Educational Psychology Review* 28, 4 (2016), 853–873.

[21] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14571–14589. doi:10.18653/v1/2023.acl-long.815

[22] Tao Huang. 2025. Content moderation by LLM: from accuracy to legitimacy. *Artificial Intelligence Review* 58, 10 (2025), 320.

[23] VM Iruku. 2025. LLM-Powered Self-Auditing Framework for Healthcare Data Pipelines: Continuous Validation Lifecycle. *European Journal of Computer Science and Information Technology* 13, 50 (2025), 82–100.

[24] Robin Jeshion. 2025. Slurs, Articulations, and Inflammatory Language. *Oxford Studies in Philosophy of Language* 4 (2025).

[25] Ansgar Kellner, Christian Wressnegger, and Konrad Rieck. 2020. What's all that noise: analysis and detection of propaganda on Twitter. In *Proceedings of the 13th European workshop on Systems Security*. 25–30.

[26] Prerana Khatiwada, Luke Halko, Nabiha Syed, Ashrey Mahesh, Aneseh Alvanpour, and Matthew Louis Mauriello. 2025. Spotting Online News: A Mixed Method Study of Online News Engagement and Perceptions on Misinformation Interventions. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–30.

[27] PRERANA KHATIWADA, IAN MUMMA, LUKE HALKO, ANESEH ALVANPOUR, and MATTHEW LOUIS MAURIELLO. 2022. Toward Browser-based Interventions to Tackle Misinformation Online. (2022).

[28] Prerana Khatiwada, Qile Wang, Kenneth E Barner, and Matthew Louis Mauriello. 2025. Towards a Multi-modal Multi-Label Election-Context Repository for Classifying Misinformation. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*, Vol. 26.

[29] Nathaniel Klemp. 2010. When rhetoric turns manipulative: disentangling persuasion and manipulation. In *Manipulating democracy*. Routledge, 77–104.

[30] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[31] Viktoriia Kozlova and Alona Posna. 2024. Pragmatic Aspect of English Fake News Discourse. *Acta Humanitatis* 2, 2 (2024), 92–123.

[32] Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication methods and measures* 5, 2 (2011), 93–112.

[33] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 865–878.

[34] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[35] Jinfen Li, Zhihao Ye, and Lu Xiao. 2019. Detection of propaganda using logistic regression. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. 119–124.

[36] Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See Kiong Ng, and Tat-Seng Chua. 2025. Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 360–381.

[37] Junxia Ma, Changjiang Wang, Hanwen Xing, Dongming Zhao, and Yazhou Zhang. 2024. Chain of stance: Stance detection with large language models. In *CCF International conference on natural language processing and chinese computing*. Springer, 82–94.

[38] Muhammad Shahid Iqbal Malik, Tahir Imran, and Jamjoom Mona Mamdouh. 2023. How to detect propaganda from social media? Exploitation of semantic and fine-tuned language models. *PeerJ Computer Science* 9 (2023), e1248.

[39] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist* 65, 2 (2021), 180–212.

[40] Gaurav Negi, MA Waskow, and Paul Buitelaar. 2026. Large Language Models as Automatic Annotators and Annotation Adjudicators for Fine-Grained Opinion Analysis. *arXiv preprint arXiv:2601.16800* (2026).

[41] Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review* (2020).

[42] Javier Pagán-Castaño, Esther Pagán-Castaño, Luis Callarisa-Fiol, and Javier Sánchez-García. 2025. The Strengthening critical thinking and its impact on new media literacy. *ESIC Market* 56, 1 (2025), e348–e348.

[43] Licheng Pan, Yongqi Tong, Xin Zhang, Xiaolu Zhang, Jun Zhou, and Zhixuan Chu. 2025. Understanding and Mitigating Overrefusal in LLMs from an Unveiling Perspective of Safety Decision Boundary. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 21057–21075. doi:10.18653/v1/2025.emnlp-main.1065

[44] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 231–240. doi:10.18653/v1/P18-1022

[45] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5377–5400. doi:10.18653/v1/2024.naacl-long.301

[46] Keith Somerville. 2011. Violence, hate speech and inflammatory broadcasting in Kenya: The problems of definition and identification. *Ecquid Novi: African Journalism Studies* 32, 1 (2011), 82–101.

[47] Marshall Soules. 2015. *Media, persuasion and propaganda*. Edinburgh University Press.

[48] Inna Stetsenko and Yuri Gordienko. 2024. Propaganda in Twitter Texts. In *Mathematical Modeling and Simulation of Systems: Selected Papers of 18th International Conference, MODS, November 13-15, 2023, Chernihiv, Ukraine*. Springer Nature, 200.

[49] Joseph E Uscinski and Ryden W Butler. 2013. The epistemology of fact checking. *Critical Review* 25, 2 (2013), 162–180.

[50] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. doi:10.1126/science.aap9559

[51] Qile Wang, Prerana Khatiwada, Avinash Chouhan, Ashrey Mahesh, Joy Mwaria, Duy Duc Tran, Kenneth E Barner, and Matthew Louis Mauriello. 2026. " The explanation makes sense": An Empirical Study on LLM Performance in News Classification and its Influence on Judgment in Human-AI Collaborative Annotation. *arXiv preprint arXiv:2602.19690* (2026).

[52] Qile Wang, Prerana Khatiwada, Carolina Coimbra Vieira, Benjamin E Bagozzi, Kenneth E Barner, and Matthew Louis Mauriello. 2026. Wisdom of the LLM Crowd: A Large Scale Benchmark of Multi-Label US Election-Related Harmful Social Media Content. *arXiv preprint arXiv:2602.11962* (2026).

[53] Fangbing Xiong, Quanhong Han, and Chengning Zhang. 2025. Design AI Agent for Auditing: Applying Large Language Models (LLMs) and Retrieval Augmented Generations (RAG) to Audit Workflows. *Journal of Emerging Technologies in Accounting* (2025), 1–10.

A Appendix. LLM Annotation Protocol and Consensus Pipeline

LLM-Based Multi-Annotator Configuration

Models and Roles

- **Annotator A** — OpenAI (`-openai-model`, default: gpt-5.1)
- **Annotator B** — Gemini (`-gemini-model`, default: gemini-3-pro-preview)
- **Annotator C** — OpenAI (`-openai-model`, default: gpt-5.1)
- **Adjudicator** — OpenAI (`-adjudicator-model`; defaults to gpt-5.1 if unspecified)

Model-Role Prompts

Annotator A "You are Annotator A, a political communication scholar. Strictly follow the codebook and JSON schema. Be conservative: if unsure, choose 'No Polarizing Language.'"

Annotator B "You are Annotator B, a discourse analyst. Your strength is correct subcategory selection. Be conservative: avoid over-labeling; if unsure, choose 'No Polarizing Language.'"

Annotator C "You are Annotator C, a high-precision media psychology expert. Be conservative: only label when explicit; if unsure, choose 'No Polarizing Language.'"

Adjudicator System Prompt (excerpt) "You are the Adjudicator, a methods-oriented political scientist overseeing three annotators."

Adjudicator constraints:

- Produce one final set of annotations.
- Only select spans from Annotator A/B/C outputs.
- Merge exact duplicate spans only.
- Do not invent new spans.
- Be conservative; use 'No Polarizing Language' if unsure.

Consensus Workflow

1. Annotators A, B, and C independently annotate each article and output structured JSON.
2. The Adjudicator receives all three JSON outputs.
3. The Adjudicator produces a single final consolidated JSON.
4. Paragraph-level policy enforcement is applied:
   - **Exact-One Policy**: Exactly one annotation per paragraph.
   - **Min-One Policy**: Retain all polarizing spans per paragraph, but ensure at least one annotation per paragraph.

B Annotation Interface and Verification Workflow

The interface designs (see Figures 4, 5, and 6) reflect our broader methodological goal: to approximate real-world annotation conditions while progressively guiding non-expert annotators toward expert-aligned judgments. The MTurk condition intentionally mirrors realistic constraints (limited time, no prior training, single-pass review) to capture how everyday readers interpret polarizing language. The verification and correction workflows then introduce lightweight definitional scaffolding and structured feedback. Rather than replacing human judgment, this design nudges annotators toward greater consistency with the expert gold standard by making category definitions explicit and encouraging deliberate reconsideration when disagreement occurs. In this way, the system serves both as a measurement instrument and as a calibration mechanism, showing gaps between lay and expert interpretations while incrementally moving annotations toward gold-standard alignment.

C Appendix: Full Corpus Exploratory Data Analysis (N=551)

To contextualize our pilot evaluation results, we provide descriptive statistics of the full 551-article corpus, including ideological balance, annotation density, and subcategory distributions. See Figures 7–9 for the corresponding distributions.

D Appendix D: Multi-HIT Agreement Progression

To support transparency around iterative deployment, here we report the full trajectory of agreement between LLM annotations and human annotations across multiple HIT batches (Figure 10).

As shown in Figure 10, earlier HITs are comparatively unstable, with low and fluctuating precision and Recall. These early results reflect exploratory deployment conditions (e.g., prompt refinement, adjudication tuning, and adjustments to our tool interface and workflow). Accordingly, we interpret these initial metrics as part of a calibration phase rather than a stable estimate of model–human agreement.

Across subsequent HITs, alignment strengthens, with later batches (January 2026) showing markedly higher consistency (Precision ≈ 0.66, Recall ≈ 0.76, F1 ≈ 0.71). Reporting the full multi-HIT trajectory helps distinguish early experimental instability from later stabilized performance.

E Appendix D: Polarizing Language Taxonomy

For completeness, we provide the full taxonomy of polarizing language used in this work, including detailed definitions and illustrative examples for each subcategory (Table 6). This appendix serves as a reference for the labeling schema underlying both the expert and crowd annotation studies.

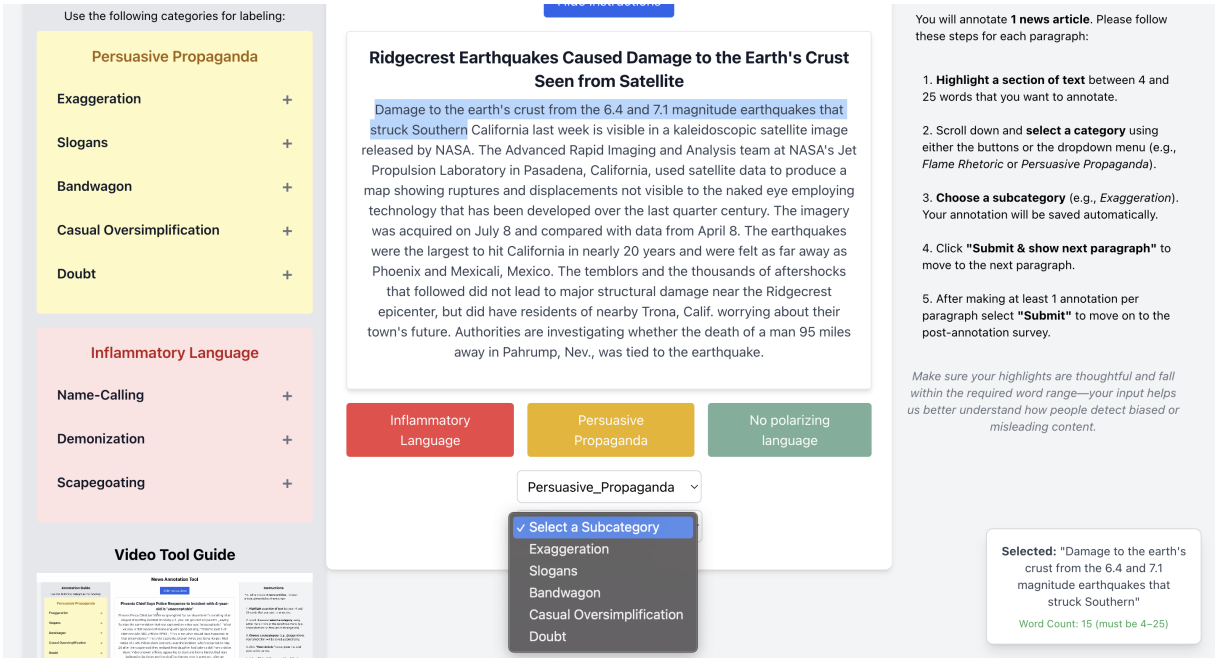

Figure 4: Human Annotation Interface (Span Selection Mode). In the fully manual annotation condition, annotators highlight spans (4–25 words), select a primary category (e.g., Persuasive Propaganda or Inflammatory Language), and choose a subcategory. This interface was used to construct the expert gold standard and collect real-world MTurk annotations.

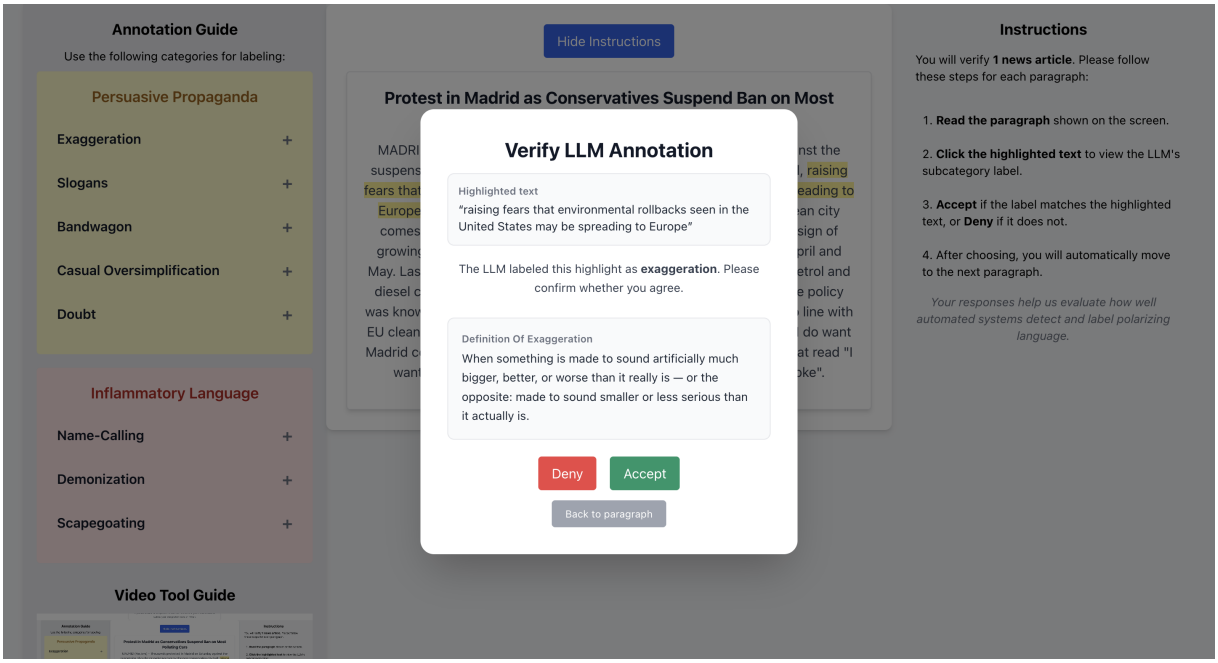

Figure 5: LLM Annotation Verification Interface. Annotators are shown highlighted spans pre-labeled by the LLM (e.g., Exaggeration) and are asked to either accept or deny the classification. The interface provides the highlighted text, the model's predicted subcategory, and a concise definition to support consistent human verification.

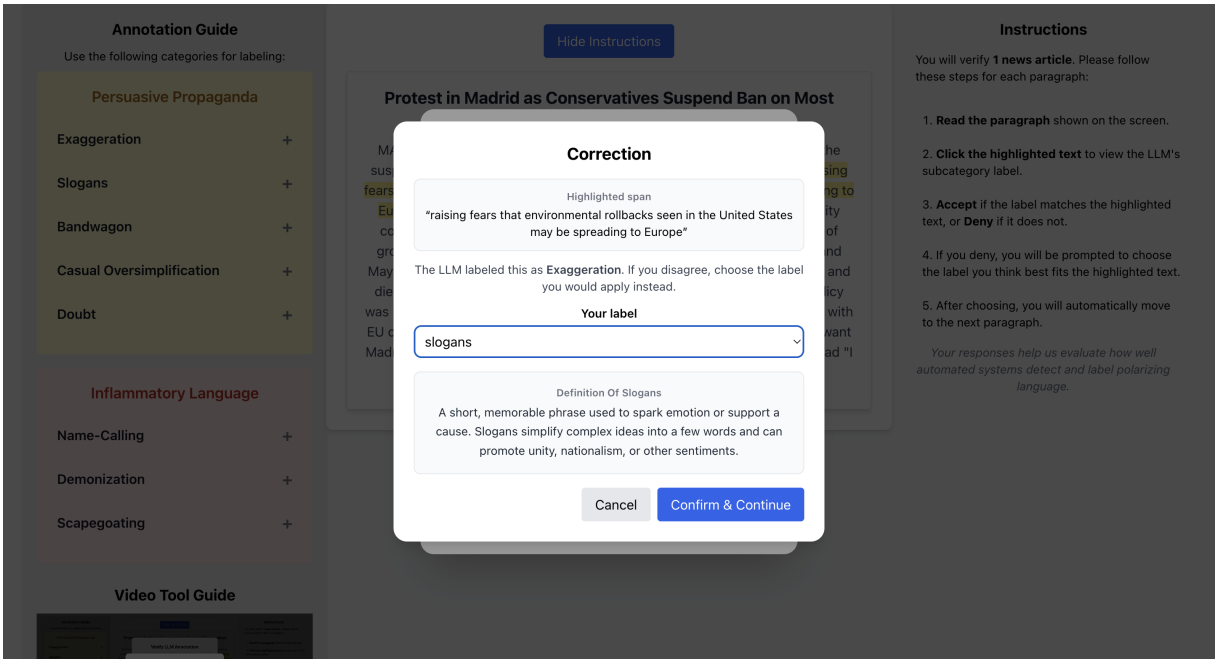

Figure 6: Correction Workflow After Disagreement. If annotators deny the LLM's label, they are prompted to select an alternative subcategory. Definitions are displayed inline to reduce ambiguity and improve annotation reliability. This step enables structured adjudication and refinement of model-generated labels.

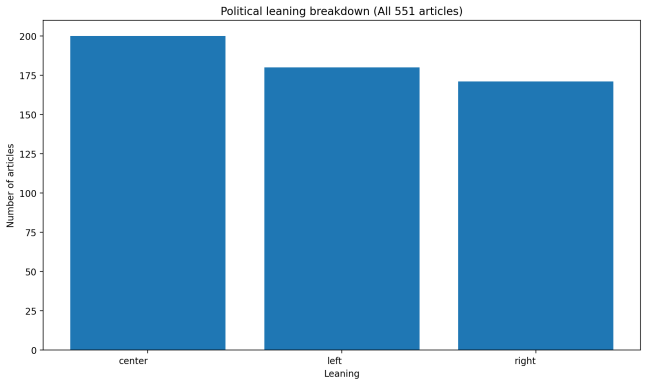

Figure 7: Political leaning distribution across the 551-article corpus. The dataset is relatively balanced across left, center, and right sources.

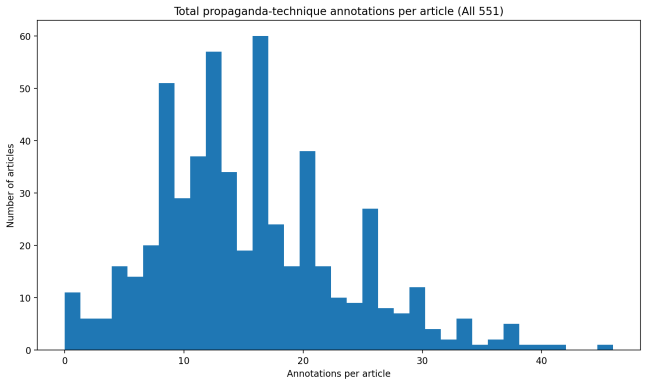

Figure 8: Distribution of total propaganda-technique annotations per article. The distribution is right-skewed, with most articles containing moderate counts and a smaller number of high-density outliers.

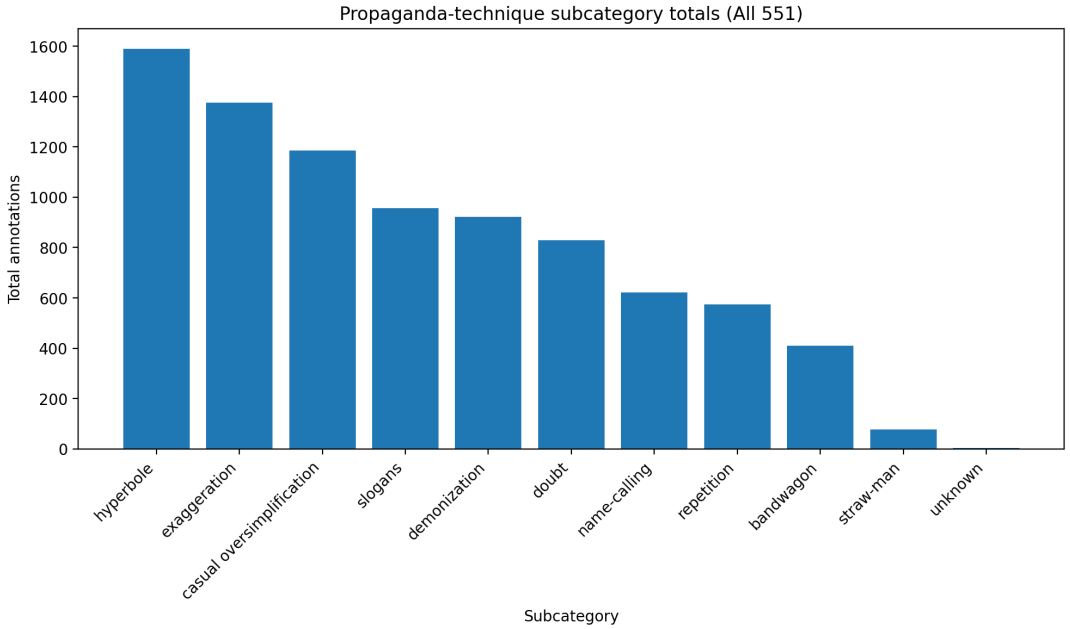

Figure 9: Total counts of propaganda subcategories across the corpus. Lower-severity techniques dominate the distribution, while high-severity categories occur less frequently.

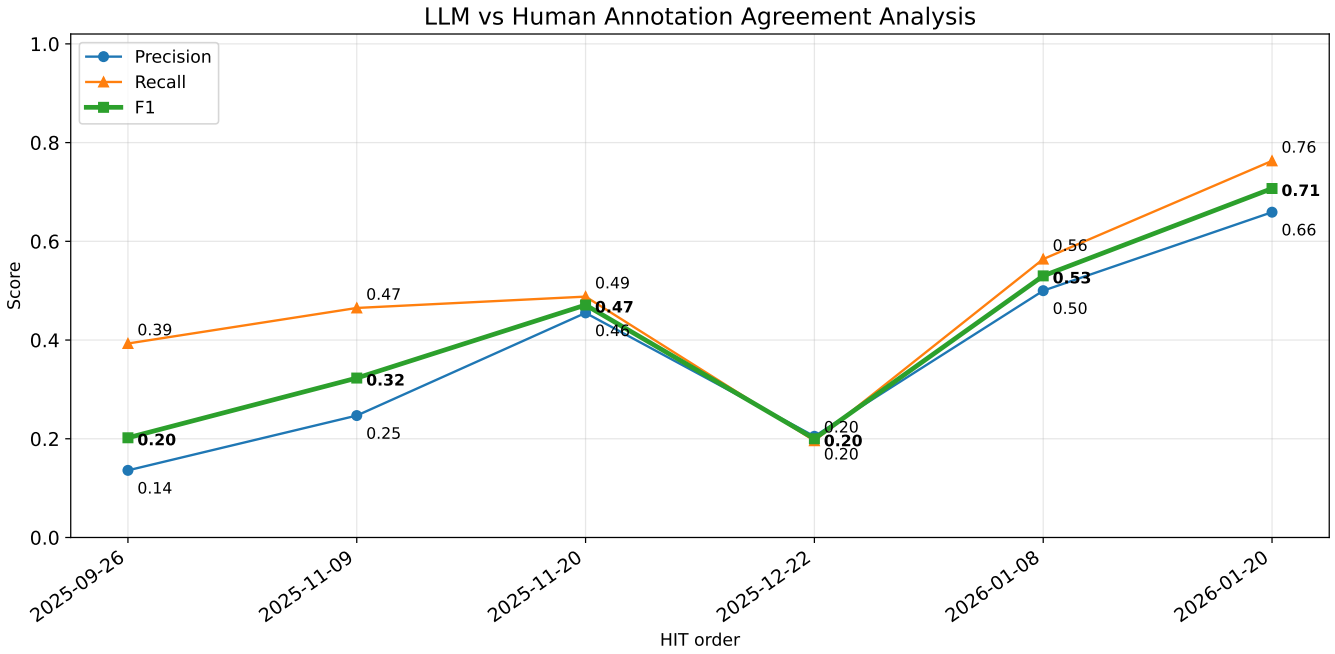

Figure 10: Agreement between LLM and human annotations across multiple HIT batches. Early HITs show instability during calibration (prompt refinement, adjudication tuning, and workflow/interface adjustments). Later iterations exhibit progressively stronger alignment, culminating in improved precision, recall, and F1.

Table 6: Polarizing language taxonomy with full definitions and examples.

Category	Definition	Examples
Exaggeration	When something is made to sound artificially much bigger, better, or worse than it really is — or made to sound smaller or less serious than it actually is.	"A local protest ignited waves of outrage and sent shockwaves through the nation." "This minor disagreement has become a national catastrophe." "The present scandal is nothing — just political theater."
Slogans	A short, memorable phrase used to spark emotion or support a cause; simplifies complex ideas into a few words.	"Make America Great Again" "No Justice, No Peace" "We Are the 99%"
Bandwagon	Encouraging support because many others supposedly support it; relies on social pressure rather than evidence.	"Most Americans back this plan." "Every true Republican supports this cause." "No serious economist still believes raising taxes is a good idea."
Casual Oversimplification	Explaining complex issues using a single cause or simple answer while ignoring broader factors.	"The media is the only reason the nation is divided." "Inflation rose solely because of the president's policies." "Crime is up because of progressive prosecutors."
Doubt	Language that makes the audience question whether a person, group, or institution is competent, honest, or legitimate.	"Is he really ready to be the Mayor?" "Is this leader even capable of running the country?" "Some experts question whether the agency's data can be trusted."
Name-Calling	Using loaded labels to shape audience perception instead of providing evidence.	"Radical extremists have demanded sweeping reform." "Big-money interests continue to profit." "The oft-labeled terrorist sympathizers took to the streets."
Demonization	Describing people or groups as evil, dangerous, corrupt, or less than human to portray them as a threat.	"The nation's bureaucrats are bleeding taxpayers dry." "Migrants are parasites stealing American jobs." "These politicians are eating away at the heart of this nation."
Scapegoating	Blaming an entire group for a broad societal problem or crisis.	"Greedy landlords are driving rising rents." "Teachers' unions are the reason kids are failing." "Homelessness rises because officials refuse to enforce laws."