

Annotating Extremist Severity in Brazilian Portuguese: Human–LLM Divergence Across Labels, Evidence, and Interpretation*

Ricardo Cabral Penteado
Universidade de São Paulo (USP)
ricardo.penteado@usp.br

Barbara McGillivray
King’s College London
barbara.mcgillivray@kcl.ac.uk

Abstract

This study examines how large language models annotate implicit extremist discourse in Brazilian Portuguese, and where their decisions diverge from human judgement. Using 1,843 risk-enriched segments, we compare three trained human annotators with GPT-4o, Gemini 2.5 Pro, and Claude 4 Sonnet under the same schema-constrained protocol. Each annotation includes a three-level severity label, an extractive evidence span, discourse-pragmatic metadata, and confidence, while human disagreements are adjudicated to produce a single adjudicated reference. We evaluate models against the adjudicated reference using accuracy, macro- F_1 , and Cohen’s κ , complement this with a multi-reference evaluation that preserves individual human variation, and additionally examine evidence-span overlap and whether shared evidence leads to the same severity judgement. LLMs agree with one another more than humans do, yet align only moderately with the adjudicated reference. They provide longer and cleaner evidence spans, but overlap with human rationales does not predict label correctness. Errors concentrate in the implicit class, where models often recognise similar cues but assign different severity levels. These findings identify agreement, confidence, and rationale overlap as unreliable proxies for annotation validity, and motivate evidence- and interpretation-level diagnostics for harmful-content annotation.

1 Introduction

Deciding whether online content crosses a violent-extremist policy line is a high-stakes operational judgement, not merely a labelling exercise. False negatives may leave explicit, actionable extremist content in circulation, with direct safety consequences, while false positives may escalate implicit

or non-actionable content to the same threshold, producing wrongful removal, suspension, or referral and affecting legitimate speech. These asymmetric costs are especially consequential for safety-oriented Natural Language Processing (NLP) systems in under-resourced online-harms settings such as Brazilian Portuguese, where extremist communication often relies less on overt hostility, profanity, or explicit threats than on ideological framing, collapse narratives, subcultural signalling, and the aestheticisation of harm (Torregrosa et al., 2023; ElSherief et al., 2021). The central challenge is therefore interpretive, since annotation requires determining how coded, indirect, and culturally situated constructions participate in the normalisation, justification, or legitimisation of extremist violence. This asymmetry structures the empirical analysis in Section 4.5, which shows that GPT-4o and Gemini escalate implicit content into the explicit class more often than they downgrade explicit content, whereas Claude displays the reverse pattern.

This challenge is most visible in the grey zone of implicit extremism, where content does not directly incite, threaten, or instruct violence, but still mobilises ideological cues, symbolic references, or discursive frames relevant to extremist interpretation. This zone overlaps with what policy and industry work describe as *borderline content*, namely material that approaches but does not clearly cross terrorist and violent-extremist policy lines while still potentially contributing to radicalisation (Saltman and Hunt, 2023). Such content is difficult to annotate and moderate because severity depends on pragmatic context, subcultural knowledge, and the relation between evidence and interpretation rather than on isolated words alone. Disagreement among annotators may therefore reflect not only noise or individual error, but also instability in the construct being annotated (Plank, 2022; Uma et al., 2021; Basile et al., 2021). At the same time, large language models (LLMs) are increas-

***Content warning.** This paper discusses violent extremist discourse and reports aggregate statistics over such material. Illustrative examples are synthetic paraphrases and no source text is reproduced verbatim.

ingly used as annotators, judges, or adjudication aids in harmful-content research (Tan et al., 2024). However, inter-model agreement does not by itself establish annotation validity, especially when the target construct is culturally specific, ambiguous, and severity-based. We therefore evaluate human and LLM annotation at three levels—label, evidence, and interpretation¹—detailed in Section 3.

We apply this framework to 1,843 Brazilian Portuguese segments, independently annotated by three human annotators and three frontier LLMs under an identical schema-constrained protocol (Section 3).

The results show a dissociation between reliability and validity. The paper contributes a three-level framework for evaluating human and LLM annotation of Brazilian Portuguese extremist discourse across labels, evidence, and interpretation. It also shows that inter-model agreement and evidence-span overlap are limited proxies for annotation validity, since models may agree with one another, or attend to similar evidence as humans, while still assigning different severity labels. Finally, it identifies interpretation-level disagreement as a diagnostic category, especially in the grey zone of implicit extremism, where models often detect relevant cues but fail to stabilise their pragmatic, ideological, or severity-related function in the same way as human annotators. By focusing on an under-resourced language and identifying operational limitations in widely deployed commercial models, the study contributes to debates on how online harms are conceptualised, measured, and audited.

2 Related Work

Human label variation and perspectivism. Recent work in NLP treats annotator disagreement as informative rather than as noise, challenging the assumption of a single ground truth in subjective tasks and documenting irreducible variation in textual inference (Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019). This perspectivist view has been extended across data, modelling, and evaluation (Plank, 2022; Uma et al., 2021), including abusive-language tasks where multiple annotator standpoints may be valid (Basile et al., 2021; Akhtar et al., 2020; Frenda et al., 2024; Weber-Genzel et al., 2024). In contentious domains, final

¹We use *interpretation* throughout in a narrow, operational sense defined in Section 3.5: not model interpretability or explainability, and not the broader linguistic or hermeneutic sense.

labels are also shaped by adjudication and governance rather than by direct access to an objective category (Krippendorff, 2018; Sap et al., 2019; Davidson et al., 2019). We adopt this stance in interpreting the moderate human agreement observed in our data.

Extractive rationales and evidence evaluation.

Rationale-based annotation links categorical decisions to input spans that justify them, from early work on rationales (Zaidan et al., 2007) to neural rationale extraction, benchmark evaluation, and hate-speech detection (Lei et al., 2016; DeYoung et al., 2020; Mathew et al., 2021). However, post-hoc attribution does not guarantee faithfulness (Jacovi and Goldberg, 2020), and span overlap measures textual similarity rather than interpretive alignment. This distinction between formal evidence overlap and interpretive validity is central to our analysis.

LLMs as annotators and implicit harm.

LLMs are increasingly used for zero- and few-shot annotation of harmful language (Brown et al., 2020; Ouyang et al., 2022; Tan et al., 2024), but empirical work reports instability under prompt variation, inconsistent detection of implicit harm, calibration drift, and bias in LLM-produced labels (Huang et al., 2023; Roy et al., 2023; Dönmez et al., 2024; Das et al., 2024). Implicit and coded harm is itself a recognised challenge, with work on implicit hate speech and coded hate symbols showing that harmful meaning may depend on indirect or culturally specific signals rather than explicit abuse (ElSherief et al., 2021; Qian et al., 2019). Existing evaluations nevertheless remain largely label-centred and focus mainly on constructs such as sentiment, toxicity, or binary hate speech. The present study addresses a different problem by jointly examining labels, evidence, and interpretation in implicit extremist discourse, where severity depends on discourse-pragmatic function rather than lexical content alone.

3 Methods

3.1 Corpus

The corpus contains 1,843 Brazilian Portuguese segments drawn from a restricted academic repository of extremist-aligned primary sources, comprising manually curated extremist manifestos and canonical texts associated with recognised extremist ideological currents, together with compiled propaganda posts and instructional material circulated

within extremist online subcultures, held under institutional access controls (see Ethical Considerations). Segments correspond to paragraph-level excerpts from selected texts and multi-sentence excerpts from compiled posts, following natural document boundaries rather than a fixed token window. In the final analysed set, segments have an estimated median length of 54 tokens (IQR: 29–127), with only 38 segments (2.1%) exceeding 500 estimated tokens. Risk-enrichment concentrated segments exhibiting collapse-oriented narratives, mobilisation cues, subcultural signalling, aestheticisation of violence, or misanthropic framing; Appendix B reports these same categories as recomputed, model-generated discourse flags over the final segment set, independent of the original curation step. Segments are labelled under the three-level ordinal severity scheme defined in Section 3.2 (0 = non-extremist, 1 = implicit/context-dependent, 2 = explicit). The reference-label distribution is 338 segments in class 0 (18.3%), 590 in class 1 (32.0%), and 915 in class 2 (49.6%). All comparisons are restricted to the common base case of segments for which all three human annotators and all three models returned schema-valid annotations.

3.2 Human Annotation Procedure and Severity Scheme

The corpus was annotated by three trained annotators with a background in linguistics, all native speakers of Brazilian Portuguese. Before annotation, they completed a calibration session based on the shared codebook. Each annotator then labelled every segment independently through a structured interface, assigning an ordinal severity label, one or more extractive evidence spans, a dominant speech-act type, a primary target type, a confidence level, and an optional note. The interface used closed value sets for the categorical fields, ensuring that the human and model annotation layers were directly comparable.

The severity scheme distinguishes three mutually exclusive classes. Annotators assess the segment’s linguistic expression of risk of violent mobilisation—not offensiveness, ideological affiliation, or legal status—at the discourse level rather than over isolated keywords, since offensive vocabulary alone does not imply extremist severity and extremist framing may occur without overt abuse. Toxicity is therefore treated as an independent property.

Class 0 covers non-extremist content. Class 1

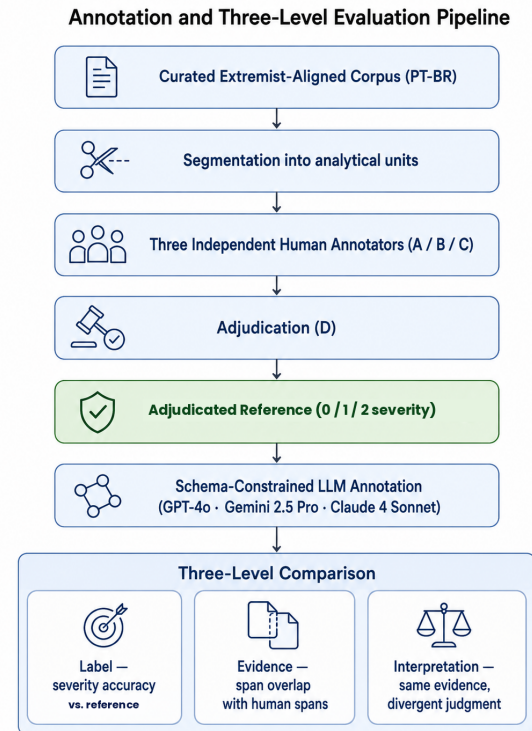


Figure 1: Annotation and three-level evaluation pipeline. Segments from a curated extremist-aligned Brazilian Portuguese corpus are labelled independently by three human annotators (A/B/C); an adjudicator (D) produces the adjudicated reference labels under the ordinal severity scheme (0/1/2). The same segments are annotated by three LLMs under an identical schema-constrained protocol. Human and model annotations are then compared at three levels: label, evidence, and interpretation.

covers implicit or context-dependent extremist content, where severity depends on ideological framing, subcultural signalling, symbolic reference, or pragmatic inference. Borderline cases between class 0 and class 1 are assigned to class 1 when the segment participates in extremist meaning-making without directly threatening, instructing, or inciting violence. Class 2 covers explicit extremist content. Any actionable element, such as a threat, operational suggestion, direct incitement, or instruction, assigns the segment to class 2 regardless of tone, irony, or meme-based phrasing. The ordering reflects actionability and imminence of violence risk, not degree of ideological commitment: an explicit, actionable threat is treated as higher-severity than an ideologically extreme but non-actionable statement, consistent with the risk-of-mobilisation framing above rather than with a scale of extremism per se.

Each annotation links the severity label to an

extractive evidence span and to two discourse-pragmatic variables. The evidence span is copied verbatim from the segment and provides the textual basis for the decision. The speech-act variable records the dominant communicative function, such as threat, incitement, instruction, glorification, ideological framing, or none. The target variable records the entity most directly implicated, such as an individual, group, institution, infrastructure, or abstract target. This structure makes each decision auditable and supports the three-level comparison reported in Section 4. Worked examples for the three classes, including evidence spans and pragmatic variables, are provided in Appendix A.

The adjudicated reference was produced through a documented resolution procedure rather than by treating majority vote as ground truth. Full agreement, i.e. identical labels across the three annotators, holds for 732 segments (39.7%); majority agreement (two of three) holds for 1,013 segments (55.0%); and three-way disagreement holds for 98 segments (5.3%). All non-unanimous segments were resolved by a fourth annotator, who assigned the final label after inspecting the conflicting annotations. Original annotations, adjudication decisions, and timestamps were retained. Against this adjudicated reference, annotator A reaches accuracy 0.826, annotator B 0.656, and annotator C 0.711 (Table 1). Lower accuracy against the adjudicated reference reflects distance from the adjudicator’s decision on contested items, not lower annotation quality; consistent with the perspectivist stance adopted here (Section 1), Annotators B and C’s divergences from the adjudicator are treated as legitimate alternative readings rather than as errors intrinsic to their judgement. We interpret the adjudicated reference as a governed outcome of an explicit annotation procedure, not as a ground truth. We use *reference label* as shorthand for an item’s value under the adjudicated reference throughout the remainder of the paper.

Even under a closed-variable codebook with an explicit severity scheme and fixed category definitions (Section 3.2), the underlying construct admits no fully determinate decision procedure: legal traditions, platform policies, and cultural context draw the boundary between ideological commitment and actionable mobilisation differently, and trained annotators operating under the same schema and calibration can still legitimately disagree on which side of that boundary a borderline segment falls. Multi-reference evaluation, as used in machine translation

where the space of acceptable outputs is effectively unbounded (Papineni et al., 2002; Lin and Och, 2004), offers one way to probe the practical limits of this residual interpretive diversity, evaluating a model’s judgement against the union of plausible expert readings rather than requiring convergence on a single label. With only three ordinal classes, however, this criterion is comparatively weak on its own (Section 4.2). More importantly, perspectival disagreement does not remove the need for an operational decision in a moderation setting: potentially harmful content may still require a determination about whether it crosses an action threshold. We therefore retain the adjudicated reference as a single, documented, and auditable decision outcome, analogous to the role of an escalation or review layer in content moderation, while treating the underlying annotator disagreements as legitimate alternative readings rather than as noise. Standard classification metrics (accuracy, macro- F_1 , Cohen’s κ) are reported against this adjudicated reference (Krippendorff, 2018), and multi-reference evaluation is used alongside them to preserve information about residual interpretive diversity.

3.3 LLM Annotation Protocol

The same 1,843 segments were submitted independently to GPT-4o (gpt-4o), Gemini 2.5 Pro (gemini-2.5-pro), and Claude 4 Sonnet (claude-4-sonnet) on 5 February 2026, using a single schema-constrained prompt that reproduced the human codebook categories. The prompt required a severity label, an extractive evidence span copied verbatim from the segment, a speech-act category, a target type, and a confidence score, returned as one structured object. Decoding was deterministic (temperature = 0). Outputs that failed schema validation were excluded: of 1,966 segments with complete human annotation, 123 (6.3%) were dropped because at least one model output failed schema validation, leaving the common base case of 1,843 items analysed throughout.

3.4 Three levels of comparison

We compare human and model annotation at three levels. At the *label* level, we report accuracy, macro-averaged F_1 , Cohen’s κ against the adjudicated reference, and pairwise agreement within and across the human and model groups. At the *evidence* level, we compare extractive spans in terms of length, segment coverage, literal containment,

token-level F_1 , and Jaccard overlap against the union of the three human spans.

At the *interpretation* level,² we test whether shared evidence yields a shared severity label: when a model’s evidence overlaps with the human reference, we ask whether the severity judgement also agrees. “Interpretation-level disagreement” refers specifically to divergent severity on shared evidence. Speech-act and target-type agreement against the human majority are reported as secondary metadata that help locate mapping failures. For completeness, Appendix B reports the prevalence of model-only auxiliary discourse flags— toxicity–abuse overlap, aestheticisation, subcultural signalling, imminence or planning, and misanthropy. These categories are descriptive only (see Limitations).

3.5 Evidence-label typology

To relate evidence selection to label correctness, we cross two binary distinctions for each model and item, namely whether the model label matches the adjudicated reference and whether its evidence span has high overlap with the human reference union. We define shared evidence as $F_1 \geq 0.60$, a conservative threshold that requires lexical grounding while allowing minor boundary differences. This yields four cells crossing correct versus wrong labels with shared versus different evidence. The wrong-with-shared-evidence cell is the key diagnostic case, since it identifies items where the model attends to the same textual evidence as human annotators but assigns a different severity label.

Span overlap does not predict label correctness. Treating overlap as a predictor of correctness yields near-chance discrimination for all three models (AUC 0.44–0.50: 0.47 [0.45, 0.50] for GPT-4o, 0.44 [0.41, 0.47] for Gemini, 0.50 [0.47, 0.52] for Claude; 95% bootstrap CIs from 2,000 resamples all overlapping 0.5), so this is not an artifact of the $F_1 \geq 0.60$ threshold used elsewhere. The same pattern holds at $F_1 \geq 0.50$ and $F_1 \geq 0.70$: accuracy conditional on shared evidence is never higher than accuracy conditional on different evidence, and the gap widens under stricter overlap, with the relative risk of an incorrect label given shared evidence

rising from 1.04–1.21 to 1.14–1.35 (Appendix C). Higher token-level agreement with human rationales therefore does not confer label validity, and may even track slightly worse severity agreement.

4 Results

4.1 Label-level reliability–validity dissociation

We use *validity* in a restricted sense, namely alignment with the adjudicated human-derived reference rather than correspondence with an external ground truth. Since this reference is produced through governed adjudication over only moderately agreeing annotators (Section 3.2), the contrast below concerns *inter-model reliability* versus *alignment with the adjudicated reference*.

Table 1 reports label-level performance against the adjudicated reference. Model accuracy ranges from 0.587 to 0.626, with Cohen’s κ between 0.35 and 0.38. Table 2 shows that model pairs agree substantially more with one another than human pairs do. Mean within-group κ is 0.31 for humans, with 95% CI [0.29, 0.34], and 0.60 for models, with 95% CI [0.58, 0.63]; mean raw agreement is 0.58 and 0.75, respectively. The non-overlapping intervals indicate that models form a more internally consistent group than the human annotators, yet their alignment with the adjudicated reference remains only moderate. Higher inter-model reliability therefore does not translate into higher validity, especially given that the reference itself is a governed reconstruction rather than a fixed truth.

We test paired differences in correctness against the adjudicated reference using McNemar’s test (McNemar, 1947). All three models perform significantly below annotators A and C ($p < 0.001$). Against annotator B, the gap is smaller. Claude is significantly below B ($\chi^2 = 18.1$, $p < 0.001$), Gemini is marginally below B ($\chi^2 = 5.4$, $p = 0.02$), and GPT-4o does not differ significantly from B ($\chi^2 = 3.6$, $p = 0.06$). Claude also performs significantly below GPT-4o ($\chi^2 = 10.3$, $p = 0.001$). The gap is therefore robust against annotators A and C but narrows relative to Annotator B, whose accuracy against the adjudicated reference (0.656) approaches the model range. This reframes LLMs less as a uniformly inferior class than as additional annotators with systematic interpretive profiles, consistent with the reliability–validity dissociation developed in Section 5.

²We use *interpretation* in a narrow operational sense: it refers to the mapping from identified textual evidence to the assigned severity label, rather than to model interpretability, explainability, or a broader linguistic or hermeneutic notion of interpretation.

System	Acc [95% CI]	Macro- F_1	κ
Annotator A	0.826 [0.809, 0.843]	0.796	0.714
Annotator B	0.656 [0.634, 0.677]	0.545	0.425
Annotator C	0.711 [0.691, 0.732]	0.687	0.543
GPT-4o	0.626 [0.603, 0.647]	0.567	0.379
Gemini 2.5 Pro	0.619 [0.596, 0.641]	0.564	0.369
Claude 4 Sonnet	0.587 [0.564, 0.609]	0.573	0.350

Table 1: Label-level performance against the adjudicated reference ($n = 1,843$). Human annotators are scored as individuals against the adjudicated reference. Bracketed values are 95% bootstrap confidence intervals from 2,000 resamples.

Pair	Type	Agr.	κ	95% CI
A vs. B	human	0.583	0.292	[.258,.325]
A vs. C	human	0.612	0.379	[.345,.413]
B vs. C	human	0.546	0.266	[.234,.300]
GPT-4o vs. Gemini	llm	0.814	0.668	[.637,.699]
GPT-4o vs. Claude	llm	0.703	0.535	[.504,.565]
Gemini vs. Claude	llm	0.746	0.601	[.572,.630]
Human mean	human	0.580	0.312	[.290,.335]
LLM mean	llm	0.754	0.601	[.578,.626]

Table 2: Pairwise agreement and Cohen’s κ within the human group and within the model group ($n = 1,843$). Bracketed values are 95% bootstrap confidence intervals from 2,000 item resamples. The human- and LLM-mean intervals do not overlap.

4.2 Multi-reference evaluation

To incorporate legitimate human variation directly into evaluation, we complement the adjudicated-reference results with a multi-reference analysis inspired by machine-translation evaluation. A model is first credited whenever its label matches any of the three individual human annotators, yielding accuracies of 0.828 for GPT-4o, 0.812 for Gemini 2.5 Pro, and 0.807 for Claude 4 Sonnet.

We additionally use a reference-count-matched comparison, holding out each human annotator in turn and evaluating both that annotator and each model against the same two remaining human references. Under this stricter comparison, model accuracy falls to 0.743, 0.728, and 0.718, respectively, compared with 0.733–0.798 for the held-out human annotators. Thus, even when models are explicitly credited for falling within the envelope of legitimate human disagreement, the reliability–validity dissociation remains. Full results are reported in Appendix E.

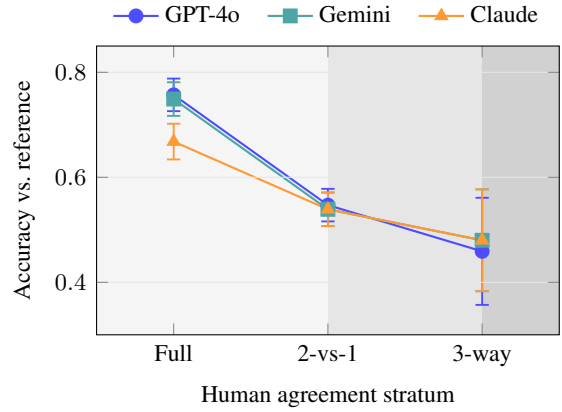


Figure 2: Accuracy by human agreement stratum ($n = 1,843$), with 95% bootstrap confidence intervals (2,000 item resamples) shown as error bars. Background shading darkens as human agreement—and the stability of the adjudicated reference—decreases; all three models degrade across the strata, and this within-model degradation is well outside bootstrap error. Between-model differences within the three-way-disagreement stratum ($n = 98$) are not distinguishable at this sample size: the intervals overlap substantially.

4.3 Performance by human agreement stratum

Model performance varies with the level of human agreement (Figure 2). On segments with full human agreement, accuracy ranges from 0.668 to 0.757 (bootstrap 95% CI); on segments with three-way human disagreement, accuracy falls to between 0.459 and 0.480, with confidence intervals that overlap across all three models given the small size of this stratum ($n = 98$). Model error concentrates where the human annotation is least stable, which are also the strata in which the reference label rests on adjudication; the apparent ranking among models in the three-way stratum should not be over-interpreted.

Inter-model consensus shows the same stratification. The three models are unanimous on 1,180 segments and reach accuracy 0.697 on this subset, against 0.530 on the 630 segments resolved by model majority. Unanimity is class-dependent, with accuracy among unanimous predictions reaching 0.782 for the explicit class, 0.600 for the non-extremist class, and 0.525 for the implicit class. Agreement among the three models is thus a weaker indicator of correctness for the implicit class than for the explicit class.

Full six-way concordance (all three annotators and all three models agreeing with the adjudicated reference) holds for 433 segments, of which 362

Source	Med. tok	Cov.	Cont.
Annotator A	13	0.41	0.93
Annotator B	7	0.20	0.98
Annotator C	2	0.08	0.87
GPT-4o	26	0.62	0.99
Gemini 2.5 Pro	24	0.59	1.00
Claude 4 Sonnet	20	0.48	1.00

Table 3: Evidence span statistics. Med. tok: median span length in tokens. Cov.: mean coverage of the segment. Cont.: mean literal containment rate.

are explicit and only 51 implicit, confirming that agreement concentrates in the anchored classes.

4.4 Evidence level: formal validity without label validity

The annotators and the models differ in evidence-selection behaviour (Table 3). Human spans are short, with median lengths of 13, 7, and 2 tokens for annotators A, B, and C, covering between 8% and 41% of the segment on average. Model spans are longer, with median lengths between 20 and 26 tokens and coverage between 48% and 62%. Model spans are also more often literally contained in the source, with containment rates exceeding 0.99 for all three models, against 0.87 to 0.98 for the annotators. The mean token-level F_1 of model spans against the human reference union is between 0.51 and 0.54. Models therefore produce spans that are formally cleaner than human spans without selecting the same regions of text. This is consistent with the span data itself: human annotators frequently marked single-token or short-phrase spans, and one annotator’s spans include multi-fragment, semicolon-delimited selections in roughly 10% of items, whereas the LLM prompt required one contiguous verbatim extraction — a difference in annotation granularity rather than in evidence quality.

Formal span validity does not transfer to label validity. For each model, a sizeable subset of items combines high span overlap with the human reference and an incorrect label, with 314 cases for GPT-4o, 324 for Gemini, and 281 for Claude (Table 4). These figures are comparable to the number of correct items with shared evidence, namely 470, 420, and 378, respectively. High overlap with human rationales is therefore not a reliable indicator that the model has assigned the human severity label.

Evidence-label cell	GPT-4o	Gemini	Claude
Correct, shared evid.	470 (25.5%)	420 (22.8%)	378 (20.5%)
Correct, diff. evid.	683 (37.1%)	721 (39.1%)	704 (38.2%)
Wrong, diff. evid.	376 (20.4%)	378 (20.5%)	480 (26.0%)
Wrong, shared evid.	314 (17.0%)	324 (17.6%)	281 (15.2%)

Table 4: Evidence-label typology by model ($n = 1,843$). Each cell shows the count and its percentage of all items for that model; the four cells sum to 100% per column. Shared evidence is span $F_1 \geq 0.60$ against the human reference union. Items for which no model span could be scored against the human reference (five for Gemini, eleven for Claude) carry no token-level overlap and are counted as different-evidence cells, split by label correctness.

4.5 Interpretation level: shared evidence, divergent judgment

The wrong-with-shared-evidence cell is the locus of interpretation-level disagreement, where the model attends to the same evidence as the human annotators and assigns a different severity label. Table 4 reports the four cells per model, and Table 5 breaks the wrong-with-shared-evidence cell down by reference class.

For GPT-4o and Gemini this cell concentrates in the implicit class: 189 of the 314 GPT-4o items and 180 of the 324 Gemini items have reference label 1. Within the high-overlap subset for the implicit class, the corresponding error rates are 71.9% (189 of 263) for GPT-4o and 68.2% (180 of 264) for Gemini. For Claude the cell concentrates in the explicit class (127 of 281 items), which is consistent with its directional tendency described below. The pattern holds across all three classes for every model but is most pronounced in the implicit region for two of the three, which is the region the human annotators find least stable. Where the models err with shared evidence, the failure is not one of evidence retrieval but of how the shared evidence is mapped to a severity judgment.

Appendix A gives qualitative cases where, even when the label is stable, the pragmatic mapping of shared evidence remains unstable across annotators and models, indicating that the divergence is interpretive rather than an artefact of token-level span matching.

Speech act, target type, and confidence Speech-act agreement is associated with higher label accuracy in the same direction for all three models, but the effect is statistically reliable only for Claude; it is marginal for Gemini and not signif-

Reference class	GPT-4o	Gemini	Claude
Class 0 (non-ext.)	63 (20.1%)	71 (21.9%)	65 (23.1%)
Class 1 (implicit)	189 (60.2%)	180 (55.6%)	89 (31.7%)
Class 2 (explicit)	62 (19.7%)	73 (22.5%)	127 (45.2%)
Total	314 (100.0%)	324 (100.0%)	281 (100.0%)

Table 5: Wrong-with-shared-evidence items by reference class. Percentages are computed within each model’s wrong-with-shared-evidence subset.

icant for GPT-4o. Target-type agreement shows the reverse profile, reaching significance for GPT-4o but not for Gemini or Claude. We therefore do not claim a uniform discourse-pragmatic effect: the direction is consistently positive across models and categories, while which category yields a statistically reliable signal is model-dependent (full two-proportion tests and confidence intervals in Appendix D).³

Reported confidence does not usefully separate correct from incorrect predictions, since the gap between the two means is at most about three percentage points for every model (GPT-4o 0.954 vs. 0.939; Gemini 0.989 vs. 0.978; Claude 0.915 vs. 0.883). Consistent with documented overconfidence in verbalised LLM confidence (Dönmez et al., 2024), it cannot be used to triage model annotations for review.

Directional Asymmetry at the Implicit–Explicit Boundary The label errors are not symmetric with respect to the implicit–explicit boundary, and the asymmetry differs across models (Figure 3). Among segments whose reference label is implicit or explicit, GPT-4o assigns the explicit label to 273 implicit segments and the implicit label to 108 explicit segments, yielding a ratio of 2.53 escalations per downgrade. Gemini shows the same direction, with 254 escalations against 122 downgrades and a ratio of 2.08. Claude reverses the pattern, assigning the implicit label to 311 explicit segments and the explicit label to 132 implicit segments, yielding a ratio of 2.36 downgrades per escalation.

The two directions carry different operational risks. Escalation across this boundary corresponds to a false-positive risk, in which non-actionable implicit content is labelled explicit; downgrading corresponds to a false-negative risk, in which actionable explicit content is labelled implicit. The

³These pairwise comparisons are secondary and exploratory, and p -values are not corrected for multiple comparisons.

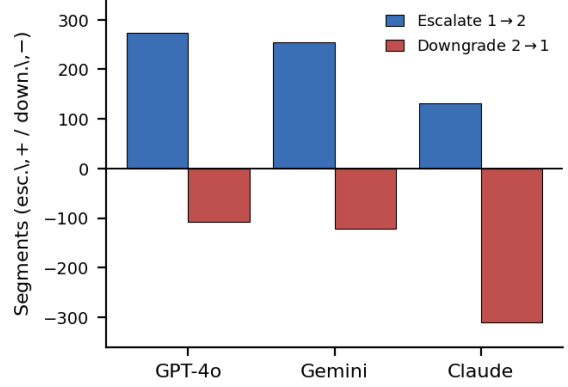


Figure 3: Directional errors at the implicit–explicit boundary ($n = 1,843$). Bars above zero are implicit segments labelled explicit (false-positive risk); bars below zero are explicit segments labelled implicit (false-negative risk). Counts are given in the text of Section 4.5.

two model families therefore fail in opposite ways at the same boundary.

5 Discussion

The results show that LLM-assisted annotation of implicit extremist discourse cannot be evaluated through label agreement alone. Across labels, evidence, and interpretation, we observe a reliability–validity dissociation. The models agree with one another more than the human annotators do, yet align only moderately with the adjudicated reference. Inter-model agreement is therefore a weak proxy for annotation validity, especially in the implicit class, where severity depends on pragmatic inference, ideological framing, subcultural signalling, and culturally situated interpretation rather than on stable cues such as threats, instructions, incitement, or operational suggestions.

The evidence-level results sharpen this point. Models produce longer and more literally contained spans than humans, but formal evidence validity does not guarantee label validity. In many cases, models select evidence that overlaps with the human reference while assigning a different severity label. These wrong-with-shared-evidence cases indicate failures of interpretation rather than retrieval. The model identifies a relevant cue, but maps it to a different communicative function or risk level. Rationale overlap should therefore not be treated as evidence that the assigned label is valid under the annotation scheme.

Importantly, this dissociation persists under

multi-reference evaluation: explicitly crediting model predictions that match individual human readings increases absolute agreement but does not eliminate the human–model gap under the reference-count-matched comparison (Section 4.2).

The models also encode different thresholds at the implicit–explicit boundary. GPT-4o and Gemini more often escalate implicit content into the explicit class, creating false-positive risk, while Claude more often downgrades explicit content, creating false-negative risk. Model choice is therefore not operationally neutral, since different LLMs generate different legal and ethical risks when used for moderation or triage. The speech-act, target-type, and confidence analyses further support this interpretation. Severity classification is partly discourse-pragmatic, while model-reported confidence is not a reliable triage signal. More useful diagnostics include human agreement strata, model disagreement, class boundaries, and wrong-with-shared-evidence cases.

These asymmetries carry different consequences in a moderation pipeline. Under-flagging explicit, actionable content (Claude’s dominant pattern) risks leaving material that meets platform or statutory removal thresholds in circulation; over-flagging implicit content (GPT-4o’s and Gemini’s dominant pattern) risks wrongful takedown or referral, reproducing the borderline-content tension introduced in Section 1. This mirrors a known platform-level risk, where variation in enforcement standards creates exploitable gaps that extremist actors adapt to (Mitts, 2025), and a documented model-level one, where different LLM providers draw the line between permitted and refused content differently and inconsistently over time (Dai et al., 2025). Selecting a single LLM as an annotator or triage tool is therefore an implicit choice of which error profile a pipeline tolerates; we do not quantify legal risk directly, and flag this as an operational consequence of the directional-asymmetry finding rather than a claim we test.

This tension extends beyond our corpus. Across the workshop’s decade-long history, higher agreement has often functioned as a proxy for progress on online harms, while the interpretive difficulty of borderline and culturally situated content has remained insufficiently resolved. Addressing this gap requires diagnostics that evaluate interpretation directly, rather than relying on further gains in agreement alone.

6 Conclusion

This paper examined human and LLM annotation divergence in Brazilian Portuguese implicit extremist discourse across severity labels, evidence spans, and interpretation. The findings show that LLMs can be internally consistent and select formally valid evidence while still failing to align with human-derived severity judgements. Inter-model agreement, evidence overlap, and model-reported confidence are therefore insufficient proxies for annotation validity in culturally situated harmful-content tasks. The main divergence lies in how models map shared evidence to severity within the implicit class, following the direction-specific pattern documented in Section 4.5, with distinct operational consequences for moderation pipelines.

The results support a multi-layer approach to LLM-assisted annotation. Evaluation should examine evidence selection and the interpretive relation between evidence and severity, rather than final-label agreement alone. LLMs may be useful as diagnostic and assistive annotators for surfacing unstable cases and model-specific thresholds, but they should be embedded in expert-in-the-loop workflows rather than treated as substitutes for human judgement. For harmful-content research, this means routinely reporting evidence- and interpretation-level diagnostics alongside label agreement, and treating model selection as an operational decision with class-specific risk profiles rather than as a single aggregate accuracy comparison.

Limitations

The adjudicated reference is the outcome of a governed annotation process, not an objective ground truth, and with full human agreement on 39.7% of segments and moderate human κ , validity against it should be read together with the per-stratum results and with the multi-reference results in Section 4.2, rather than as a single aggregate. The model-generated discourse categories reported in Appendix B have no human reference and are presented as interpretive traces, not as validated features. The corpus is risk-enriched and is not a representative sample of online Brazilian Portuguese, so the reported rates are not deployment estimates for general streams. Although compact, the dataset follows the logic of expert-grounded resources in legally and ethically sensitive harmful-content research, where evidential value depends on annota-

tion quality, adjudication, and governance rather than corpus size (Kums et al., 2025). Span overlap is computed on lexical containment and token-level F_1 , which privilege surface similarity and may undervalue paraphrastic or discourse-level justification. The high-overlap threshold of $F_1 \geq 0.60$ is one operationalisation among several; we therefore verified that the wrong-with-shared-evidence pattern persists at $F_1 \geq 0.50$ and $F_1 \geq 0.70$ (Appendix C), though the absolute size of the high-overlap subset is threshold-dependent and a token-level overlap metric remains a proxy for evidential agreement.

Ethical Considerations

The source materials are extremist-aligned primary documents held in a restricted academic repository under institutional access controls. We do not redistribute raw source text. Illustrative examples, where used, are synthetic paraphrases that preserve the annotation-relevant structure of corpus segments without reproducing them verbatim, following WOA reporting guidance; they avoid reproducing operational content, and group references, slurs, and identifying material are obfuscated. The annotation schema records discourse-level signals and extractive spans and does not prompt models to generate new harmful content. Segments may contain identifying references, and any released artefacts exclude personal data. The corpus and analysis are intended for methodological research on annotation and evaluation, not for automated enforcement or deployment without human oversight.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Authors’ Contributions RCP: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft. BMcG: Methodology, Supervision, Writing – Review & Editing.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

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

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yunlang Dai, Emma Lurie, Danaé Metaxa, and Sorelle A. Friedler. 2025. [Longitudinal monitoring of LLM content moderation of social issues](#). *Preprint*, arXiv:2510.01255.

Amit Das, Zheng Zhang, Fatemeh Jamshidi, Vinija Jain, Aman Chadha, Nilanjana Raychawdhary, Mary Sandage, Lauramarie Pope, Gerry Dozier, and Cheryl Seals. 2024. [Investigating annotator bias in large language models for hate speech detection](#). *Preprint*, arXiv:2406.11109.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Esra Dönmez, Thang Vu, and Agnieszka Falenska. 2024. Please note that I’m just an AI: Analysis of behavior patterns of LLMs in (non-)offensive speech identification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18340–18357, Miami, Florida, USA. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide

- Bernardi. 2024. [Perspectivist approaches to natural language processing: A survey](#). *Language Resources and Evaluation*. Published online; volume/issue to be confirmed.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297, Austin, TX, USA. ACM.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics.
- Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications.
- Vincent Kums, Florian Meyer, Luisa Emily Pivit, Uliana Vedenina, Jonas Wortmann, Melanie Siegel, and Dirk Labudde. 2025. A novel dataset for classifying German hate speech comments with criminal relevance. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 41–52. Association for Computational Linguistics.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.
- Chin-Yew Lin and Franz Josef Och. 2004. [ORANGE: a method for evaluating automatic evaluation metrics for machine translation](#). In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- Tamar Mitts. 2025. *Safe Havens for Hate: The Challenge of Moderating Online Extremism*. Princeton University Press, Princeton, NJ.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to decipher hate symbols. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: Strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.
- Erin Saltman and Micalie Hunt. 2023. [Borderline content: Understanding the gray zone](#). Insight, Global Internet Forum to Counter Terrorism (GIFCT). Accessed: 2026-06-26.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.
- Javier Torregrosa, Gema Bello-Orgaz, Eugenio Martínez-Cámara, Javier Del Ser, and David Camacho. 2023. A survey on extremism analysis using natural language processing: Definitions, literature review, trends and challenges. *Journal of Ambient Intelligence and Humanized Computing*, 14:9869–9905.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. [VariErr NLI: Separating annotation error from human label variation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2007. Using “annotator rationales” to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

A Class Examples

OFFENSIVE CONTENT WARNING: This appendix contains examples of extremist discourse. This is strictly for the purpose of enabling this research, and we have sought to minimise the number of examples where possible. Please be aware that this content could be offensive and cause distress.

The examples below are synthetic, corpus-inspired illustrations shown in Brazilian Portuguese with English glosses. Following WOAHP reporting guidance, we do not reproduce source segments verbatim: each example is a constructed paraphrase that preserves the annotation-relevant structure of the original (syntax, register, degree of aggression, and pragmatic function) while altering surface wording and obfuscating group references and any operational or threatening content. Group targets are replaced with a neutral placeholder.

Implicit extremism resists operationalisation precisely because the cues that mark it are culturally situated rather than lexical: collapse-oriented narratives that frame democratic institutions as terminally corrupt, and the aestheticisation of violence as renewal or purification, circulate in Brazilian Portuguese extremist discourse through registers and intertextual references that a schema built on explicit threat or incitement cannot capture. The examples below are chosen to make this culturally situated instability visible at the level of individual segments, not only in aggregate statistics.

Class 0: non-extremist. A segment is non-extremist when it does not normalise, encourage,

rationalise, or aestheticise violence, even if it discusses crisis, nihilism, or security events. A representative segment reports a data breach:

“Vazados quase 3 BILHÕES de documentos de Seguro Social”

“Nearly 3 *BILLION* Social Security documents leaked.”

The evidence span covers the report of the leak, the speech act is *none*, and the target is *institution*. All human annotators and all models assign label 0.

Class 1: implicit violent extremism. A segment is implicit when it normalises, aestheticises, or rationalises violence without explicit threat, instruction, or operational content. Typical signals include collapse-oriented reasoning, subcultural signalling, aestheticised destruction, and diffuse targets such as society or the system. This is the interpretive centre of the scheme and the region in which construct stability is lowest. A representative segment frames a hostile worldview through collapse-oriented reasoning without naming an action:

“[ALVO] é a percepção de que tudo o que previmos era absolutamente verdade. A democracia sendo totalmente corruptora; a criação de uma distopia anti-[GRUPO-ALVO]; pedofilia e degeneração se tornando mainstream; tudo isso. Estávamos certos [...]”

“[*TARGET*] is the realisation that everything we predicted was absolutely true. Democracy being utterly corrupting; the creation of an anti-[*TARGET GROUP*] dystopia; paedophilia and degeneracy becoming mainstream; all of it. We were right [...]”

The reference label is 1, and all six annotators assign label 1. Yet the annotation diverges *within* this agreed label: human spans range from the group-directed clause (“a criação de uma distopia anti-[GRUPO-ALVO]”) to a single abstract token (“democracia”), and the speech act is read as *ideological framing* by two annotators but as *incitement* by the third. The three models all read it as *rationalisation*, but split on the target between *group* and *diffuse*. The label is stable; its pragmatic interpretation is not. Because the corpus is risk-enriched rather than representative, this instability

rate should not be read as a deployment estimate for general content-moderation streams; and because this illustration is a constructed paraphrase rather than the verbatim segment, it demonstrates the annotated pattern rather than constituting independent evidence for it, a limitation of the qualitative component we flag alongside the quantitative typology in Table 5.

A second segment illustrates the same instability through an explicit rejection of electoral politics:

“Novamente – tudo o que importa é que você não pense fora da caixa em que eles te colocaram desde a escola primária. [...] Há uma solução, ela só não vem na forma de um voto [...]”

“Again – all that matters is that you do not think outside the box they placed you in since primary school. [...] There is a solution; it simply does not come in the form of a vote [...]”

The reference label is again 1, with full six-way agreement on the label. The evidence converges — two human annotators select the same clause (“*Há uma solução, ela só não vem na forma de um voto*”) — yet the interpretation does not: humans label the speech act *ideological framing* or *incitement* with an *abstract* target, while the three models label it *rationalisation* and again disagree on whether the target is *abstract*, *diffuse*, or *group*. The clause rejects the vote as the locus of change without specifying an alternative action, leaving open several legitimate readings, from ideological framing or implicit incitement to non-violent alternatives such as political protest or civil disobedience; the schema does not adjudicate between these readings, and this openness is itself the interpretation-level divergence analysed in Section 4.5.

Class 2: explicit violent extremism. A segment is explicit when it threatens, incites, instructs, or operationalises violence. A representative segment is a direct threat against a group:

“Todo [GRUPO-ALVO] será morto”

“Every [TARGET GROUP] will be killed”

The evidence span covers the full clause, the speech act is *threat*, and the target is *group*. All human annotators and all models assign label 2. Explicit segments are the class on which inter-model agreement is most predictive of the reference label, as discussed in Section 4.3.

B Auxiliary Discourse Flags

We additionally examine the auxiliary discourse flags produced by the models. These flags are not evaluated against a human reference; instead, they are used descriptively to characterise model-specific interpretive profiles. The prevalence of these categories differs across models (Table 6). GPT-4o marks toxicity–abuse overlap most often, Gemini marks subcultural signalling most often, and Claude marks aestheticisation more often than the other models. These differences suggest that the models foreground different discourse dimensions when mapping evidence to severity, but they should not be read as validated properties of the corpus.

Category (present)	GPT-4o	Gemini	Claude
Toxicity–abuse overlap	0.558	0.281	0.391
Aestheticisation	0.223	0.155	0.320
Subcultural signalling	0.251	0.436	0.389
Imminence / planning	0.314	0.286	0.307
Misanthropy	0.115	0.075	0.072

Table 6: Prevalence of selected model-generated discourse categories ($n = 1,843$). No human reference exists for these categories; they are reported only as descriptive traces of model-specific interpretive profiles.

C Threshold Robustness of the Evidence–Label Typology

We assess whether the four-cell evidence–label typology (Table 4) depends on the $F_1 \geq 0.60$ span-overlap threshold. Table 7 recomputes the partition at $F_1 \geq 0.50$, 0.60 , and 0.70 . Each row is computed over all 1,843 items; items with no scorable span are counted as different-evidence. As the threshold rises, fewer items qualify as shared, so the absolute wrong-with-shared count falls; but accuracy conditional on shared evidence is never higher than accuracy conditional on different evidence at any threshold, and the deficit grows as the criterion tightens. The 0.50 and 0.70 partitions are nested re-thresholdings of the same items rather than independent samples, so the across-threshold trend is a descriptive robustness check, not a hypothesis test.

D Discourse-Pragmatic Agreement and Label Accuracy

Table 8 reports, for each model, label accuracy on items where the model matches the human-majority speech act (or target type) versus items

Model	thr.	C/Sh	C/Df	W/Df	W/Sh	A Sh	A Df
GPT-4o	0.50	672	481	270	420	.615	.640
	0.60	470	683	376	314	.599	.645
	0.70	297	856	469	221	.573	.646
Gemini	0.50	619	522	269	433	.588	.660
	0.60	420	721	378	324	.565	.656
	0.70	249	892	479	223	.528	.651
Claude	0.50	571	511	346	415	.579	.596
	0.60	378	704	480	281	.574	.595
	0.70	195	887	596	165	.542	.598

Table 7: Robustness of the evidence $\times$ correctness partition to the span-overlap threshold ($n = 1,843$ per row). Cells count items by label correctness (C/W) and whether the model’s span shares high overlap with the human reference union (Sh/Df); items with no scorable span are counted as different-evidence. A|Sh and A|Df give label accuracy conditional on shared vs. different evidence. The primary threshold ($F_1 \geq 0.60$) is in bold. The threshold-free rank association between span F_1 and correctness is at or below chance (AUC = 0.47/0.44/0.50 for GPT-4o/Gemini/Claude; 95% bootstrap CI from 2,000 item resamples, computed over scorable items: [0.45, 0.50], [0.41, 0.47], [0.47, 0.52]).

where it does not. We test each contrast with a two-proportion χ^2 test with continuity correction on the disjoint matched/unmatched item sets; confidence intervals are Wald intervals for the difference in proportions. The accuracy difference is positive in every case, but reaches conventional significance only for Claude on speech act and for GPT-4o on target type. These are secondary, exploratory comparisons; p -values are not corrected for multiple comparisons.

Cat.	Model	Match	No-match	Δ [95% CI]	p
Speech	GPT-4o	.653	.634	+.020 [−.029, .068]	.46
	Gemini	.650	.606	+.044 [−.004, .093]	.08
	Claude	.628	.564	+.065 [.015, .114]	.01
Target	GPT-4o	.691	.628	+.062 [.007, .118]	.03
	Gemini	.699	.688	+.011 [−.045, .066]	.75
	Claude	.638	.632	+.007 [−.053, .066]	.88

Table 8: Label accuracy by discourse-pragmatic agreement with the human majority. Match / No-match give label accuracy on items where the model’s speech-act (or target-type) category does or does not match the human-majority category. Δ is the difference in accuracy with a Wald 95% confidence interval; p is from a two-proportion χ^2 test with continuity correction. The positive direction holds throughout, but is statistically reliable only for Claude (speech act) and GPT-4o (target type).

System	3-ref (any of A,B,C)	2-ref (matched)
Annotator A (held out)	—	0.798
Annotator C (held out)	—	0.761
Annotator B (held out)	—	0.733
GPT-4o	0.828	0.743
Gemini 2.5 Pro	0.812	0.728
Claude 4 Sonnet	0.807	0.718
Majority-class baseline	0.698	—

Table 9: Multi-reference accuracy ($n = 1,843$). The 3-ref column credits a label matching any of the three human annotators. The 2-ref column is the reference-count-matched comparison: each human annotator is held out and compared against the other two, and each model is compared against the same three annotator-pairs (mean reported). No model reaches the 2-ref self-consistency shown by any individual human annotator.

E Multi-reference evaluation

Table 9 reports the full multi-reference comparison used to test whether the main results depend on evaluating models against a single adjudicated reference. The 3-reference criterion credits a model whenever its label matches any of the three human annotators. The reference-count-matched 2-reference criterion instead evaluates each held-out human annotator and each model against the same pairs of remaining human references.