Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

ToxiSight: Leveraging Moderator Expertise Through Behavioral Measurement in Gaming Toxicity Annotation

Zachary Yang[†,‡,◇*], Vicki Chen[†], Domenico Tullo[†], Reihaneh Rabbany[‡, ◇]
[†] Ubisoft La Forge
[‡] McGill
[◇] Mila

Abstract

Content moderation systems commonly treat human annotators as interchangeable label sources, resolving disagreements through majority voting or expert arbitration. We present **ToxiSight**, an annotation platform that reframes this assumption: rather than extracting consensus, the system supports moderator reasoning by treating hesitation, revision, and disagreement as signals revealing where content is genuinely ambiguous and where taxonomic guidelines fail. ToxiSight integrates gaming-specific contextual widgets with behavioral telemetry, capturing the cognitive processes underlying toxicity validation decisions. Through deployment with 10 professional moderators across 60,000 lines of gaming chat, we demonstrate that behavioral patterns expose systematic category failures invisible to traditional inter-annotator metrics. The Controversial category shows 72% revision rates with fast processing times, indicating immediate recognition of definitional breakdown, while Threats (Life-Threatening) exhibits 75% revisions with slow processing, signaling genuine interpretive complexity. Completion rates improved from 60% to 95%, and moderators reported reduced decision stress when permitted to express uncertainty. This case study demonstrates that trustworthy toxicity detection requires annotation systems designed around the irreducible complexity of human judgment, not against it.

Content Warning: This paper contains examples of real-world toxicity.

Keywords: Content Moderation, Toxicity Annotation, Sociotechnical Systems, Human-AI Collaboration, Behavioral Analytics, Gaming

The dominant paradigm for toxicity annotation treats disagreement between human annotators as measurement error to be minimized, and hesitation as inefficiency to be optimized away. This framing, embedded in most annotation infrastructure, embodies what Birhane [1, 2] identifies as a Cartesian worldview: an attempt to impose categorical certainty onto the active, fluid, messy, and context-dependent nature of human communication. The Turing Institute's "Doing AI Differently" article similarly argues that trustworthy AI requires centering human expertise throughout system design rather than treating human oversight as a final-stage quality check [3].

These critiques carry practical weight for online gaming moderation. Gaming chat unfolds through rapid exchanges, domain-specific slang that evolves weekly, obfuscated language (leetspeak, deliberate misspellings), and social dynamics shaped by team roles, competitive outcomes, and shared gameplay history [4–7]. Whether "noob" constitutes playful banter or targeted harassment cannot be determined from surface text alone. Surveys confirm that over 80% of online multiplayer gamers experience harassment [8, 9], yet annotation tools remain designed for generic workflows, providing little support for the interpretive complexity gaming moderators navigate daily.

We introduce **ToxiSight**, an annotation platform that treats validation behavior (i.e., reaction times, revision patterns, widget usage, and uncertainty expression) as a rich information source for diagnosing both model failures and taxonomy inadequacies. It instruments

* zachary.yang@mail.mcgill.ca

the annotation process itself to understand *how* professional moderators reason about gaming toxicity when validating outputs from ToxBuster [10], a contextual toxicity detection model trained on gaming chat.

Our central argument: *trustworthy moderation systems must measure and leverage moderator expertise rather than treating annotators as interchangeable label sources.* Behavioral signals reveal where content is genuinely complex (slow processing, low revision, high widget usage) versus where guidelines structurally fail (fast processing, high revision, minimal context consultation). This distinction is invisible to aggregate accuracy metrics yet critical for continuous system improvement.

Through empirical work with 10 professional gaming moderators and 5 managers across 60,000 annotations, we demonstrate three findings: (1) moderator-centered design reduces cognitive burden, improving completion rates from 60% to 95% and SUS scores from 52.3 to 78.9; (2) four recurring behavioral signatures diagnose category quality ranging from effective boundaries (Hate: 20% revision) to taxonomic collapse (Controversial: 72% revision despite fast decisions); and (3) human deliberation is diagnostic, not supplementary — moderator hesitation and revision reveal systematic failures invisible to model-only evaluation.

1. Related Works

1.1. Gaming Toxicity and Content Moderation

Gaming toxicity exhibits linguistic and social patterns that distinguish it from general social media toxicity. Content moderation in gaming must account for evolving slang, obfuscated language, and social dynamics shaped by team roles and competitive outcomes [4, 6, 7, 10, 11]. Studies of cyberbullying patterns in League of Legends [5] and domain-specific slang in DOTA 2 [12] demonstrate that toxicity is tightly coupled to game mechanics and community terminology. Research further shows that toxicity is adaptive: players modulate behavior based on perceived norms and social consequences [13].

Foundational work in content moderation scholarship emphasizes that moderation is a *sociotechnical interpretive practice* entangled with ambiguity, shifting norms, and contextual dependence [14–16]. Roberts [14] documents the psychological toll and interpretive complexity of commercial content moderation, revealing how platform policies interact with annotator discretion, cultural background, and organizational constraints. Recent work shows that LLM-based moderation performance drifts over time as language norms evolve [17], underscoring the need for human oversight systems capable of diagnosing when automated systems fail.

1.2. Annotation Platforms, Disagreement, and Human-AI Collaboration

General-purpose annotation systems such as POTATO [18], Label Studio [19], and Prodigy [20] prioritize schema flexibility but provide limited behavioral logging or domain-specific decision support. Research demonstrates that annotation UX directly affects label quality: contextual snippets, structured rationales, and attention cues reduce cognitive burden and improve consistency [21].

Birhane's [1] critique of "automating ambiguity" provides the theoretical core of ToxiSight's design. Birhane argues that machine learning systems inherit a Cartesian worldview presuming that disagreement can be resolved, ambiguity eliminated, and human judgment rendered deterministic. Disagreement-oriented frameworks such as CrowdTruth [22] challenge this assumption, demonstrating that annotator variation often reflects genuine ambiguity rather than error. The data-centric AI movement [23] similarly emphasizes capturing richer metadata — confidence scores, rationales, hesitation patterns — to diagnose noise and refine schemas.

Tools such as Errudite [24] and DARMA [25] connect annotation behavior to model performance, but focus on model debugging rather than domain-tailored behavioral UX. The PaTAT framework [26] supports iterative schema refinement through pattern-based annotation, but does not target gaming's contextual challenges or instrument behavioral telemetry for model feedback. Human-AI collaboration systems for annotation must also balance model suggestions against human agency: research demonstrates that naive model outputs can anchor annotators, reducing label diversity and introducing automation bias [27, 28].

Cognitive psychology grounds our behavioral metrics. Response time reflects cognitive effort and decision uncertainty [29]; hesitation and revision patterns correlate with task difficulty [30, 31]. Basile et al. [32] and Pavlick and Kwiatkowski [33] demonstrate that many NLP labels encode subjective judgments, and that systems should surface rather than collapse ambiguity.

ToxiSight synthesizes these traditions by providing domain-specific contextual scaffolds for gaming toxicity while instrumenting the cognitive processes — reaction time, revision, uncertainty expression, contextual tool usage — that reveal where content is genuinely complex versus where taxonomies structurally fail.

2. Methodology

Our goal is to improve annotation data quality for gaming toxicity detection by designing platforms that support rather than constrain moderator expertise. Our guiding question: *What can moderator reasoning processes (i.e., hesitation, revision, contextual tool usage) reveal about where content is genuinely ambiguous and where taxonomic guidelines fail?*

This question requires engaging both **front-line moderators** (who perform daily toxicity validation) and **player safety managers** (who set organizational policy and define category boundaries). ToxiSight operates within a validation framework where moderators review outputs from ToxBuster [10]; moderators validate, correct, or refine these pre-generated predictions.

Participants. All participants were professional employees of GAMING_COMPANY (anonymized), recruited through internal channels. **Front-line moderators ($N = 10$)** had 0–2 years of experience ($M = 0.35$ years), reflecting the high turnover typical of safety roles. **Player safety managers ($N = 5$)** are senior staff responsible for defining toxicity categories and setting organizational policy. All participants were co-located in the same regional office and provided informed consent with the right to withdraw.

2.1. Three-Phase Requirements Analysis

Phase 1 (Observational Workflow Analysis). We conducted non-intrusive observations by joining moderation team calls and shadowing daily workflows, documenting tool-switching patterns, decision bottlenecks, and expressed frustrations. Observational notes were synthesized into workflow diagrams identifying fragmentation points.

Phase 2 (Semi-Structured Interviews). We conducted 1-hour semi-structured interviews with all 10 front-line moderators, probing what makes content difficult to categorize, when moderators feel uncertain, and what information would support better decisions. Interviews were audio-recorded (with consent), transcribed, and analyzed using reflexive thematic analysis [34]. Separately, structured interviews with the 5 player safety managers addressed organizational requirements, category rationale, and known taxonomy limitations.

Phase 3 (Iterative Platform Evaluation). We conducted hands-on sessions comparing three interface conditions: (1) baseline spreadsheet workflows; (2) ToxiSight v1 (unified interface without contextual widgets); and (3) ToxiSight full system. Moderators completed

annotation tasks under think-aloud protocols, followed by SUS questionnaires and qualitative interviews about perceived utility.

Surveillance transparency. Behavioral tracking carries meaningful risks of being perceived as performance surveillance, particularly in emotionally taxing safety work. Before Phase 3, we explicitly clarified that behavioral logging was intended solely to improve platform design and understand annotation difficulty patterns, not to monitor individual performance. All metrics were aggregated and anonymized before analysis; raw interaction data were accessible only to the research team; participation was voluntary with the right to withdraw at any time. No participants raised surveillance concerns during sessions, which we attribute to this upfront transparency.

2.2. Requirements Identified

Thematic analysis of interview transcripts and observational notes converged on four critical requirements that directly shaped ToxiSight's design:

R1 – Unified Workflow Integration. Current moderator workflows require constant switching between annotation interfaces, chat context viewers, translation services, and terminology databases. This fragmentation creates cognitive overhead that disrupts decision-making flow: *"By the time I find the chat history and translate the line, I've forgotten what the original prediction was."* Beyond inconvenience, tool-switching breaks the interpretive thread that gaming toxicity assessment requires. Context gathered in one tool must be mentally held while acting in another, compounding error and fatigue.

R2 – Gaming-Aware Contextual Support. Gaming toxicity exhibits contextual dependencies absent from social media platforms. The same phrase can constitute legitimate gaming terminology or targeted harassment depending on participant relationships, team dynamics, game context, and cultural background. Moderators repeatedly described situations where surface-text classification was impossible without knowing who was speaking to whom and why: *"'Noob' isn't automatically toxic – it depends who's saying it, whether they're on the same team, and if it's the tenth time they've said it."* Moderators need integrated access to conversational context, gaming-specific terminology disambiguation, and cultural nuance. This information was scattered across external tools.

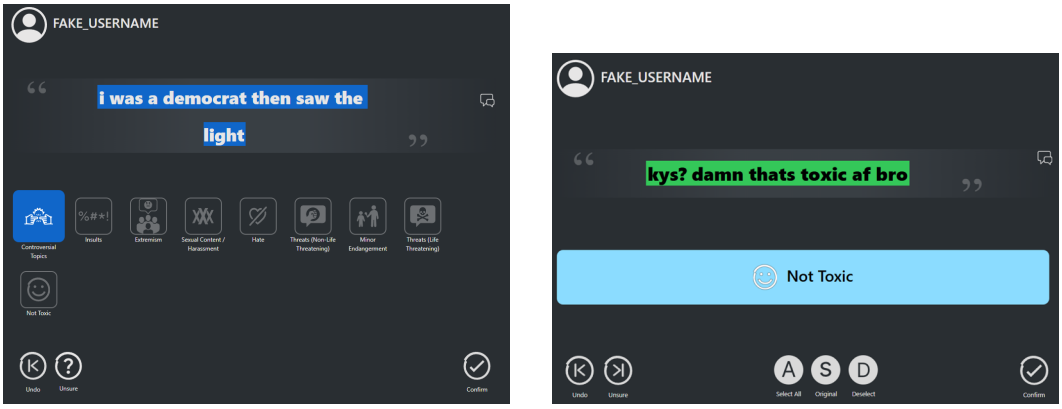
R3 – Legitimized Uncertainty Expression. Current binary accept/reject paradigms force categorical decisions on inherently subjective content, creating stress and quality degradation when moderators encounter genuinely ambiguous cases. Gaming toxicity often involves subtle contextual factors – rhetorical versus literal threats, banter versus harassment. Moderators reported that being forced to choose without recourse produced both inaccurate labels and emotional burden. Managers confirmed that understanding *where* moderators are uncertain would be more valuable for taxonomy refinement than the labels themselves.

R4 – Behavioral Signals as Systematic Diagnostic Data. Traditional inter-annotator agreement metrics (Cohen's kappa, Fleiss' kappa) reveal only *that* disagreement exists, not *why*. When moderators hesitate, revise decisions, or struggle with specific categories, these behavioral signals simultaneously reveal two distinct things: content that is genuinely difficult for humans to interpret, and model predictions that systematically misalign with expert judgment.

3. ToxiSight Platform Design

3.1. Unified Interface with Behavioral Logging (R1, R4)

ToxiSight consolidates previously fragmented tools into a single platform. The validation workflow comprises two sequential phases. In **Category Selection** (Fig. 1, left), moderators see ToxBuster's initial prediction alongside the highlighted toxic span, and may confirm,



(a) Category selection: ToxBuster's prediction shown with highlighted toxic span in blue.

(b) Span adjustment: blue = model prediction, green = moderator adjustment.

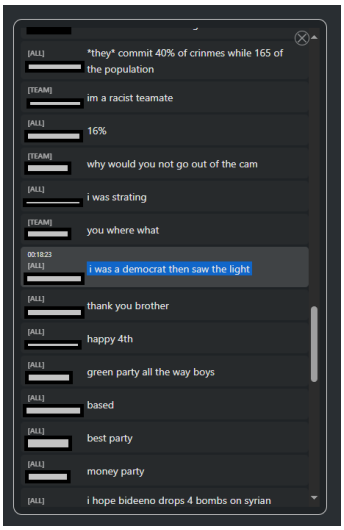
Figure 1. ToxiSight core validation interface. Two sequential phases decompose the moderator's decision into category selection and span boundary confirmation.

navigate to an alternative category, or enter "Unsure" mode for ambiguous cases. In **Span Adjustment** (Fig. 1, right), moderators validate or refine the highlighted toxic spans, with color coding distinguishing model predictions (blue) from human adjustments (green).

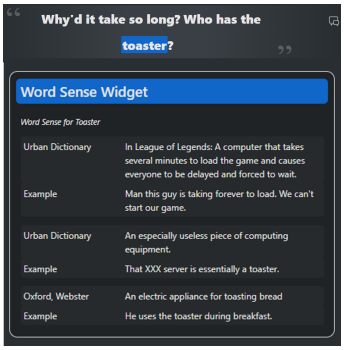
The behavioral logging infrastructure captures reaction time (precise timestamps from content presentation to confirmation), revision probability (whether the initial category prediction was changed), widget invocations (which contextual tools were accessed and in what order), and navigation patterns (category exploration sequences). Confirmations under 0.33 seconds are excluded as precluding genuine cognitive processing; sessions auto-terminate after 5 minutes to prevent measurement artifacts from breaks or task abandonment.

3.2. **Gaming-Specific Contextual Scaffolds (R2)**

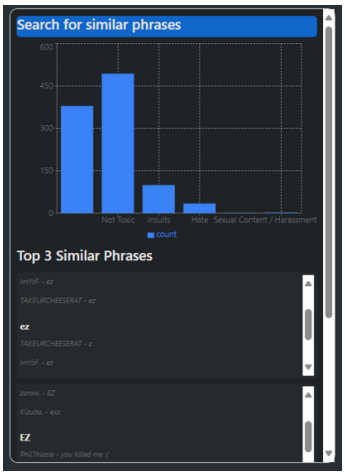
ToxiSight integrates six specialized widgets providing domain-aware decision support (Fig. 2). **Category Definitions** provide context-sensitive access with gaming-specific examples distinguishing legitimate terminology from genuine threats. **Chat History** (hotkey `W`) displays complete conversational context with current-line highlighting, participant relationships, and scrollable context of ± 10 surrounding messages. **Word-Sense Disambiguation** addresses gaming slang polysemy: the widget generates embeddings from Webster, Oxford, and Urban Dictionary, ranks word senses by cosine similarity to conversational context, and displays disambiguated definitions (e.g., "toaster" as kitchen appliance, low-spec computer in gaming slang, or specific weapon). **Training Data Attribution** uses embedding-based similarity retrieval (cosine threshold > 0.5) to display how comparable phrases were previously classified, helping moderators assess category boundary consistency. **LLM Dual Rationale** generates *contrasting* explanations (i.e., "Why this might be toxic" versus "Why this might NOT be toxic") using GPT-4o-mini with ± 5 conversational lines as context, explicitly framed as advisory to preserve human agency. **Multilingual and Cultural Context** provides literal translation, natural interpretation accounting for slang, and cultural context explanation, addressing gaming's inherently multilingual communities.



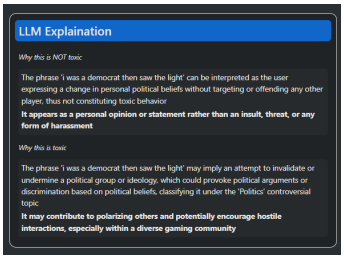
(a) Chat history: conversational context with participant relationships.



(b) Word-sense disambiguation: ranked polysemous definitions.

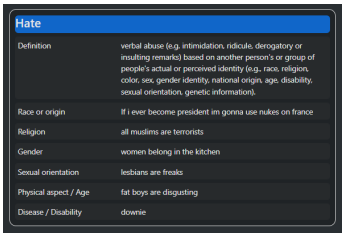


(c) Training data attribution: precedent distribution for similar phrases.



(d) LLM dual rationale: contrasting toxic vs. non-toxic explanations.

(e) Multilingual support: literal, interpretive, and cultural context.



(f) Category definitions with gaming-specific examples.

Figure 2. ToxiSight's six contextual widgets operationalize the principle that gaming toxicity depends on conversational dynamics, linguistic ambiguity, cultural norms, and categorical precedent simultaneously. Following Birhane [1], context is treated as irreducibly multi-layered.

3.3. **Uncertainty Expression and Behavioral Analysis (R3, R4)**

For genuinely ambiguous content, moderators may enter **Unsure mode**, enabling multi-category selection with confidence sliders (e.g., Controversial 60% + Hate 40%). This directly challenges the assumption that every item has one correct label, legitimizing uncertainty as a diagnostic signal revealing content that resists categorical judgment, category boundaries that are poorly defined, and model predictions that fall in ambiguous regions. Moderators reported that Unsure mode reduced decision stress: *"It feels better to say 'I'm not sure' than to force a wrong label."*

From logged interaction traces, we derive four behavioral measures: **Reaction Time (RT)**, grounded in choice reaction time paradigms from cognitive psychology [29], reflecting cognitive effort and interpretive complexity; **Revision Probability**, quantifying prediction stability; **Widget Usage Patterns**, correlated with RT and revision; and **Uncertainty Expression Frequency**, the proportion of items triggering Unsure mode.

By mapping the joint distribution of RT and revision probability across toxicity categories, we identify four recurring behavioral signatures: **Obvious Keep** (fast RT, low revision) indicates effective category design and reliable model performance; **Nuanced Keep** (slow RT, low revision) indicates inherent content complexity with trustworthy model predictions; **Obvious Change** (fast RT, high revision) indicates immediate recognition of model or taxonomy failure; and **Nuanced Change** (slow RT, high revision) indicates fundamental category ambiguity or definitional overlap. These signatures are descriptive rather than predictive; quadrant assignments rely on median splits of RT and revision probability, and their validity for downstream model improvement requires further validation.

4. Results

We evaluated ToxiSight across three interface conditions: (1) baseline spreadsheet workflows; (2) ToxiSight v1 (unified interface, no widgets); and (3) ToxiSight full system. This progression isolates the contributions of workflow consolidation versus contextual support. The primary deployment engaged 10 front-line moderators across 60,000 lines of gaming chat drawn from ToxBuster's validation set, under naturalistic conditions with no artificial time pressure.

Evaluation combined System Usability Scale (SUS) administered after each interface condition, behavioral logging of RT and revision probability per item aggregated by category, and post-deployment semi-structured interviews (30–45 minutes) analyzed through thematic analysis.

4.1. Usability and Workflow Efficiency

Across a two-week deployment period, each moderator was assigned a subsample of the dataset calibrated to their committed hours. Throughput was estimated at 100 lines/hour based on prior labeling sessions with this moderator cohort, ensuring completion targets were proportional to individual availability rather than uniform across the team. Completion rates improved from 60% (spreadsheet baseline) to 78% (ToxiSight v1) to 95% (ToxiSight full). The 58% relative improvement reflects elimination of tool-switching overhead, reduced cognitive load from integrated contextual support, and a more predictable workflow rhythm (RT variance decreased 23%). SUS scores progressed from 52.3 (below acceptable threshold) to 67.1 (borderline) to 78.9 ("Good" range), quantitatively demonstrating that contextual widgets reduce rather than increase cognitive burden.

Interview themes were consistent across participants. Moderators described the baseline workflow as "fatiguing" and "disorienting" due to constant tool-switching. ToxiSight v1 improved focus but moderators still felt "information-starved" for complex cases. The full system was described as "giving me what I need, when I need it" (P7). Several moderators noted that Unsure mode reduced stress.

Widget usage. Chat History was invoked for 23% of items, primarily for Controversial and Threats categories where participant relationships and conversational context are critical. Multilingual Translation was invoked for 18% of items and nearly half (48%) of items flagged as ambiguous via Unsure mode, confirming that linguistic ambiguity is a major driver of interpretive difficulty. Widget usage was highest for items in the slow, high-revision quadrant, confirming that contextual scaffolds are most valuable precisely where content is most difficult. Training Data Attribution was most frequently used when moderators initially disagreed with ToxBuster's prediction but reconsidered after reviewing precedent. LLM Dual Rationale was used sparingly, primarily for Nuanced Change cases.

Table 1. Category-level behavioral profiles. High revision rates combined with high RT variance signal taxonomic problems; low revision with low variance signal effective categories. Categories are ordered by revision probability to highlight the Obvious Change and Nuanced Change signatures.

Category	RT (sec ± SD)	Revision Prob. (± SD)	n
Threats (LT)	11.54 ± 28.58	$75\% \pm 43\%$	4,271
Controversial	7.62 ± 25.12	$72\% \pm 45\%$	49,791
Minor Endangerment	14.90 ± 31.48	$58\% \pm 49\%$	19,706
Threats (Non-LT)	14.26 ± 31.53	$58\% \pm 49\%$	1,298
Extremism	10.38 ± 28.10	$38\% \pm 48\%$	1,835
Insults	9.87 ± 28.57	$30\% \pm 46\%$	69,636
Hate	7.60 ± 23.73	$20\% \pm 40\%$	21,190
Sexual Content	10.73 ± 28.98	$20\% \pm 40\%$	10,278
Not Toxic	8.12 ± 25.66	$3\% \pm 17\%$	19,706

4.2. **Behavioral Signatures of Category Quality**

Table 1 provides category-level behavioral profiles. Figure 3 visualizes categories along mean RT and revision probability axes, revealing the four signatures.

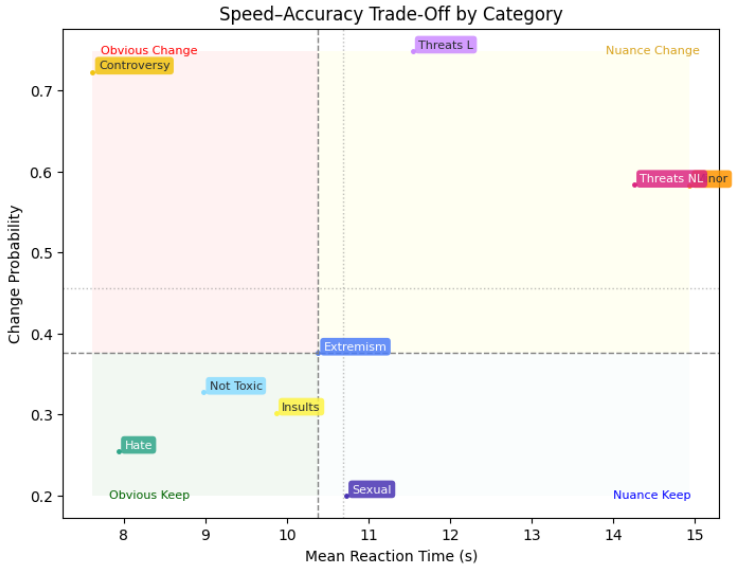


Figure 3. Behavioral signatures of category reliability. Categories plotted by mean reaction time (cognitive effort) and revision probability (prediction instability). Dashed lines denote median splits. Four interpretable regions emerge: **obvious keep** (fast, stable), **nuanced keep** (slow, stable), **obvious change** (fast, unstable), and **nuanced change** (slow, unstable). This structure highlights where model–moderator alignment is strong, where taxonomic revisions are needed, and where content is genuinely ambiguous.

Obvious Keep. Not Toxic (RT = 8.12s, revision = 3%) and Hate (RT = 7.60s, revision = 20%) cluster in the low-RT, low-revision region. Moderators described these as "straightforward almost every time" (P2). These categories serve as behavioral benchmarks for successful human-AI alignment.

Nuanced Keep. Sexual Content (RT = 10.73s, revision = 20%) shows elevated RT due to genuine interpretive complexity – moderators must assess cultural norms, implied meanings, and context – but revision remains low, confirming that ToxBuster's predictions

align with expert judgment even for difficult cases: *"It takes longer, but the model usually gets it right"* (P6).

Obvious Change. Controversial (RT = 7.62s, revision = 72%) exemplifies this signature. Fast processing paired with high revision indicates *immediate recognition of failure*, not content simplicity. Moderators consistently described Controversial as "too broad" (P4), "catching everything the model wasn't confident about" (P9), or "not a real category" (P1). This behavioral signature reveals taxonomic collapse: a catch-all category whose boundaries are incoherent.

Nuanced Change. Threats (Life-Threatening) (RT = 11.54s, revision = 75%) shows slow processing with high revision, indicating fundamental ambiguity. Moderators reported genuine difficulty distinguishing rhetorical threats ("I'll destroy you" in competitive banter), strategic threats ("focus their healer"), and literal harm ("I know where you live"). This signature signals categories where guidelines require refinement, not where moderators lack expertise.

Logistic regression. To assess whether behavioral patterns generalize beyond descriptive trends, we modeled revision probability using logistic regression with category identity and RT as predictors. Model fit was strong (LR $\chi^2 = 47{,}681$, $p < .001$; pseudo-$R^2 = 0.20$). RT shows a statistically significant effect: each additional second increases revision odds by 0.2% (OR = 1.002, 95% CI [1.001, 1.003], $p < .001$). While modest at the unit level, this effect accumulates across thousands of annotations, reflecting how cognitive uncertainty compounds. Category coefficients align with behavioral signatures: Not Toxic (OR = 0.01, $P(\text{change}) = 1.2\%$) and Sexual Content (OR = 0.10, 8.7%) anchor low-revision predictions; Threats LT (OR = 1.13, 53.1%) anchors high-revision predictions. Full per-category coefficients, odds ratios, and confidence intervals are provided in supplementary materials.

4.3. **Category Transition Analysis**

Analysis of transitions from ToxBuster's initial predictions to final human-validated labels reveals where models systematically misalign with expert judgment. Sankey diagrams illustrating the full distribution of these transitions are provided in supplementary materials. Not Toxic and Hate retain model assignments in over 90% of cases, confirming effective boundaries.

In contrast, Minor Endangerment retains only 42% of predictions; 33% are reassigned to Not Toxic and 8% to Insults. Managers confirmed that Minor Endangerment's definition ("intention or mentioning of where a child is or will be in a potentially harmful situation") is difficult to operationalize in fast-paced gaming chat, where accusations of "grooming" or "pedophilia" are sometimes used as generic insults rather than literal accusations.

Threats (Life-Threatening) retains only 25% of predictions; 49% are downgraded to Threats (Non-Life Threatening), 9% to Hate, and 9% to Not Toxic. Moderators described difficulty distinguishing severity levels in short, informal messages: *"'I'll kill you' could be a joke, a strategic call, or a genuine threat* (P7). Cross-category confusions such as Controversial to Hate are invisible to aggregate accuracy metrics but surface clearly through transition analysis combined with behavioral profiling.

Across all categories, fast decisions correspond to high agreement (Not Toxic), while slow decisions reflect either nuanced consensus (Sexual Content) or systematic confusion (Threats LT). The joint distribution of RT and agreement – not either measure alone – distinguishes *legitimate complexity* from *category breakdown*, a distinction critical for continuous improvement: legitimate complexity requires contextual support, whereas category breakdown requires definitional revision.

5. Conclusion

ToxiSight's central contribution is methodological: human deliberation is a primary diagnostic data source, not a supplementary check after automation fails. By instrumenting *how* moderators reason (i.e., measuring hesitation, revision, contextual tool usage, and uncertainty expression) ToxiSight reveals distinctions invisible to aggregate inter-annotator agreement or model accuracy metrics. Three design principles emerge. First, **legitimize uncertainty**: binary annotation paradigms force categorical decisions on inherently ambiguous content; multi-category confidence scoring reduces decision stress and surfaces diagnostic signals about category clarity. Second, **integrate multi-layered context**: gaming toxicity depends simultaneously on conversational dynamics, linguistic ambiguity, cultural norms, and precedent; a unified interface reduces cognitive overhead without sacrificing interpretive fidelity. Third, **treat human deliberation as diagnostic data**: rather than penalize slow processing or disagreement, measure *where* and *why* moderators hesitate. Behavioral signals expose systematic model and taxonomy failures that accuracy metrics cannot. These principles generalize beyond gaming; the moderator-centered methodology provides a template for annotation platforms across any domain requiring expert human interpretation, from medical text annotation to legal document review.

Ethics Statement

Behavioral tracking carries risks of being perceived as performance surveillance, particularly in the emotionally taxing context of safety work. We explicitly clarified that all logging was for platform improvement and aggregate analysis, not individual performance monitoring. All metrics were aggregated and anonymized before analysis; raw data were accessible only to the research team. LLM-generated rationales are framed as advisory, presented in dual format to encourage critical evaluation, and never required for task completion. Gaming data from GAMING_COMPANY is not publicly available.

Limitations

Our 10 moderators are all from a single company, regional office, and cultural context, with limited experience (M = 0.35 years). Whether behavioral patterns differ for senior moderators or across organizational cultures remains unexplored. LLM-generated rationales and translations (GPT-4o-mini) inherit biases and hallucination risks; dual-format rationales and advisory framing mitigate but do not eliminate anchoring effects. The four behavioral signatures are descriptive, not predictive; quadrant assignments rely on median splits whose threshold sensitivity has not been validated against downstream outcomes. RT is an imperfect proxy for cognitive load – delays may reflect distractions or interface friction – partially mitigated by excluding sub-0.33s confirmations and analyzing distributions rather than raw values.

Acknowledgements

This section can be left blank during double-blind review.

References

[1] A. Birhane. "The Impossibility of Automating Ambiguity". In: *Artificial Life* 27 (May 2021), pp. 1–18. DOI: 10.1162/artl_a_00336.

[2] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao. "The Values Encoded in Machine Learning Research". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022. DOI: 10.1145/3531146.3533083.

[3] D. Hemment, C. Kommers, et al. *Doing AI Differently: Rethinking the Foundations of AI via the Humanities*. Tech. rep. The Alan Turing Institute, 2025. URL: https://www.turing.ac.uk/news/publications/doing-ai-differently.

[4] B. van Aken, J. Risch, R. Krestel, and A. Löser. "Challenges for Toxic Comment Classification: An In-Depth Error Analysis". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Ed. by D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 33–42. DOI: 10.18653/v1/W18-5105. URL: https://aclanthology.org/W18-5105/.

[5] J. Blackburn and H. Kwak. "STFU NOOB! Predicting crowdsourced decisions on toxic behavior in online games". In: Apr. 2014, pp. 877–888. DOI: 10.1145/2566486.2567987.

[6] H. Kwak, J. Blackburn, and S. Han. "Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games". In: 22 (Apr. 2015). DOI: 10.1145/2702123.2702529.

[7] B. Gambäck and U. K. Sikdar. "Using Convolutional Neural Networks to Classify Hate-Speech". In: *Proceedings of the First Workshop on Abusive Language Online*. Ed. by Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 85–90. DOI: 10.18653/v1/W17-3013. URL: https://aclanthology.org/W17-3013/.

[8] ADL. *Online hate and harassment: The American Experience 2021*. 2021. URL: https://www.adl.org/online-hate-2021.

[9] ADL. "Hate and Harassment in Online Games". In: *Anti-Defamation League* (2022). URL: https://www.adl.org/sites/default/files/documents/2022-12/Hate-and-Harassment-in-Online-Games-120622-v2.pdf.

[10] Z. Yang, N. Grenon-Godbout, and R. Rabbany. "Towards Detecting Contextual Real-Time Toxicity for In-Game Chat". In: *Findings of the Association for Computational Linguistics: EM dNLP 2023*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, 2023, pp. 9894–9906. DOI: 10.18653/v1/2023.findings-emnlp.663. URL: https://aclanthology.org/2023.findings-emnlp.663/.

[11] J. H. Park and P. Fung. "One-step and Two-step Classification for Abusive Language Detection on Twitter". In: *Proceedings of the First Workshop on Abusive Language Online*. Ed. by Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 41–45. DOI: 10.18653/v1/W17-3006. URL: https://aclanthology.org/W17-3006/.

[12] H. Weld, G. Huang, J. Lee, T. Zhang, K. Wang, X. Guo, S. Long, J. Poon, and S. C. Han. *CONDA: a CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection*. 2021. arXiv: 2106.06213 [cs.CL].

[13] J. Fox and W. Y. Tang. "Sexism in online video games: The role of conformity to masculine norms and social dominance orientation". In: *Comput. Hum. Behav.* 33 (Apr. 2014), 314–320. ISSN: 0747-5632. DOI: 10.1016/j.chb.2013.07.014. URL: https://doi.org/10.1016/j.chb.2013.07.014.

[14] S. T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019.

[15] R. Gorwa, R. Binns, and C. Katzenbach. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance". In: *Big Data & Society* 7.1 (2020), p. 2053951719897945. DOI: 10.1177/2053951719897945. eprint: https://doi.org/10.1177/2053951719897945. URL: https://doi.org/10.1177/2053951719897945.

[16] R. Caplan and T. Gillespie. "Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy". In: *Social Media + Society* 6.2 (2020), p. 2056305120936636. DOI: 10.1177/2056305120936636. eprint: https://doi.org/10.1177/2056305120936636. URL: https://doi.org/10.1177/2056305120936636.

[17] X. Liu and et al. "Longitudinal Monitoring of LLM Content Moderation of Social Issues". In: *arXiv preprint arXiv:2510.01255* (2025).

[18] J. Pei, A. Ananthasubramaniam, X. Wang, N. Zhou, A. Dedeloudis, J. Sargent, and D. Jurgens. "POTATO: The Portable Text Annotation Tool". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Abu Dhabi,

UAE: Association for Computational Linguistics, 2022, pp. 327–337. DOI: 10.18653/v1/2022.emnlp-demos.33. URL: https://aclanthology.org/2022.emnlp-demos.33/.
[19] HumanSignal. *Label Studio: Open Source Data Labeling*. 2019. URL: https://labelstud.io.
[20] Explosion AI. *Prodigy: An annotation tool for AI, machine learning & NLP*. 2017. URL: https://prodi.gy.
[21] A. Manzoor, B. Wallace, and M. Lease. "Designing Annotation Interfaces for Subjective NLP Tasks". In: *AAAI HCOMP*. 2023.
[22] L. Aroyo and C. Welty. "CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement". In: *Semantic Web* 10.1 (2019), pp. 1–24.
[23] A. Ng. "The Data-Centric AI Movement". In: *Harvard Data Science Review* 4.1 (2022).
[24] T. Wu, M. T. Ribeiro, and J. Heer. "Errudite: Scalable, Reproducible, and Testable Error Analysis for NLP Models". In: *ACL*. 2019.
[25] M. Lee and et al. "DARMA: Model-in-the-Loop Annotation for Dataset Development". In: *ACL*. 2023.
[26] S. Gebreegziabher, Z. Zhang, X. Tang, Y. Meng, E. Glassman, and T. Li. "PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis". In: Apr. 2023, pp. 1–19. DOI: 10.1145/3544548.3581352.
[27] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan. "Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, 2022, pp. 1–19.
[28] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445717. URL: https://doi.org/10.1145/3411764.3445717.
[29] R. Ratcliff and G. McKoon. "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks". In: *Neural Computation* 20.4 (Apr. 2008), pp. 873–922. ISSN: 0899-7667. DOI: 10.1162/neco.2008.12-06-420. eprint: https://direct.mit.edu/neco/article-pdf/20/4/873/817277/neco.2008.12-06-420.pdf. URL: https://doi.org/10.1162/neco.2008.12-06-420.
[30] A. Ye and A. Zhang. "Confidence Contours: Uncertainty-Aware Annotation". In: *AAAI HCOMP*. 2023.
[31] A. E. M. Méndez, M. Cartwright, J. P. Bello, and O. Nov. "Eliciting Confidence for Improving Crowdsourced Audio Annotations". In: *Proceedings of the ACM on Human-Computer Interaction*. Vol. 6. CSCW1. ACM, 2022, pp. 1–29.
[32] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, and A. Uma. "We Need to Consider Disagreement in Evaluation". In: *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Ed. by K. Church, M. Liberman, and V. Kordoni. Online: Association for Computational Linguistics, Aug. 2021, pp. 15–21. DOI: 10.18653/v1/2021.bppf-1.3. URL: https://aclanthology.org/2021.bppf-1.3/.
[33] E. Pavlick and T. Kwiatkowski. "Inconsistency and Subjectivity in Natural Language Annotation". In: *Transactions of the ACL* 7 (2019), pp. 677–694.
[34] V. Braun and V. Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3 (Jan. 2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.