

To Flag or Not to Flag? Detecting Concerning Content in Situational Judgment Test Responses

Susha Suresh, Cole Walsh, Rodica Ivan, Colleen Robb

Acuity Insights Inc., Toronto, ON, Canada

Abstract

This study compares approaches to automating the detection of concerning content in open-response situational judgment tests (SJTs). Such assessments are often used in higher education admissions to assess applicants' personal and professional skills, and the open-response format provides rich insights into test takers' approach and rationale by allowing them to craft their own responses. This flexibility also comes with the risk of test takers including concerning content, such as discriminatory remarks, unethical proposals, or unprofessional language. It is often desirable to flag this type of content separately from regular scoring processes for review by decision-makers, however flagging is typically manual, time consuming, and difficult to scale. We compare the effectiveness of fine-tuned BERT-based encoders (BERT and Toxic-BERT, a domain-adapted variant pre-trained on toxicity detection) against zero-shot and fine-tuned large language models (Qwen2.5-7B and Ministral3-8B). Notably, despite the LLMs receiving nuanced scenario context absent from BERT's training data, the fine-tuned BERT model achieved the strongest performance in identifying concerning language, indicating that fine-tuned encoders can be preferable to LLMs in some settings. While the results of our experiments are promising, performance remains short of what autonomous decision-making would require in a high-stakes admissions setting. We discuss implications for operational use based on these findings.

1 Background

Open-response formats are increasingly used in high-stakes educational assessment, particularly in admissions, because they reveal applicants' reasoning, judgment, and professional dispositions (Dore et al., 2017). This same openness, however, allows test takers to introduce concerning content such as discriminatory remarks, unethical proposals, or

unprofessional language, which test administrators may wish to identify and incorporate into evaluations. Identifying these instances often involves multiple layers of human review, however, a process that is slow, subjective, and difficult to scale as assessment volumes grow.

Automated methods offer one avenue to help prioritize responses for review while supporting more consistent identification of content that warrants additional attention. Despite its importance, however, automated detection of concerning content in educational assessment remains comparatively underexplored. Most prior work targets toxicity in informal or adversarial online environments (Waseem et al., 2017; Kumar et al., 2024) rather than the professional, reflective writing typical of assessments, and educational contexts introduce fundamentally different constraints, risks, and challenges (Pilicita and Barra, 2025). For instance, applicants may express controversial opinions as part of legitimate intellectual exploration, making the boundary between concerning and acceptable content more nuanced than in typical toxicity detection. Two further barriers are specific to this setting. First, data availability is limited: unlike social media platforms with publicly available toxicity datasets (Jigsaw, 2020; Mathew et al., 2021), admissions data is proprietary and protected to ensure applicant privacy. Second, concerning content is inherently rare, making it difficult to collect a sufficient sample of positive examples for traditional supervised learning.

The detection of toxic, offensive, or hateful content in text has received substantial attention in the broader NLP research field. While early approaches relied on lexicon-based methods and traditional machine learning (Machová et al., 2022), transformer-based models, particularly BERT (Devlin et al., 2019), have dramatically improved performance on toxicity detection benchmarks. Hanu and Unitary team (2020) introduced Toxic-BERT,

a BERT-based model fine-tuned on the Jigsaw Toxic Comment Classification dataset. The model achieves strong performance on detecting various forms of toxic content including threats, obscenity, and identity-based attacks. However, these models are typically trained on social media content, which differs substantially from the professional, reflective writing found in assessment contexts. More recently, Large Language Models (LLMs) have been employed with specific prompts to identify offensive language (Thomas et al., 2025) and more nuanced expressions of offensive language such as problematic stances that may be subtly rather than overtly toxic, which could also be helpful in educational assessments.

Given the data scarcity described above, transfer learning offers another promising path: by fine-tuning models pre-trained on related tasks (e.g., toxicity detection), researchers can leverage pre-trained representations and achieve reasonable performance even with limited domain-specific data. The same scarcity also motivates the use of LLMs, which can perform competitively in low-resource settings through zero-shot or few-shot prompting, leveraging knowledge acquired during pre-training without requiring task-specific labeled examples. These two approaches have different trade-offs, and it remains unclear which is more effective for a specialized task with very few positive examples. Recent work shows that instruction-tuning quality and data diversity significantly impact LLM performance (Renze Lou, 2025; Liu, 2024) and that models fine-tuned on smaller, high-quality datasets can outperform those trained on larger but less curated corpora (Chunting Zhou et al., 2023), suggesting that more data or larger models do not guarantee better results. This motivates a systematic comparison of fine-tuned encoders against zero-shot and fine-tuned LLMs under the same data-scarce conditions to determine which best supports automated detection of concerning content in educational assessment.

2 Aims

In this study, we investigate the effectiveness of fine-tuned transformer-based encoders and zero-shot and fine-tuned LLMs for identifying concerning content in open-response SJTs. We address four primary research questions:

RQ1 (Transfer learning efficacy): Can models pre-trained on general toxic content detection (e.g.,

Toxic-BERT) transfer effectively to the more nuanced task of identifying concerning responses in educational assessments, where “concerning” encompasses not only explicit toxicity but also unprofessional conduct, ethical violations, and discriminatory reasoning?

RQ2 (Zero-shot LLMs with contextual enrichment): Can zero-shot LLMs with enriched context (scenario prompt, questions, and explicit flagging instructions) achieve competitive performance on concerning content detection despite receiving no task-specific training? Critically, do they match or exceed fine-tuned BERT despite being substantially larger?

RQ3 (Fine-tuned models comparison): How does fine-tuning impact both encoders and LLMs? Can fine-tuned LLMs exceed fine-tuned BERT’s performance, and does the additional computational cost of larger models justify performance gains?

RQ4 (Error characterization): What types of concerning content do automated approaches reliably detect versus miss, and what are the implications for human-in-the-loop collaboration in the review process?

3 Sample

3.1 Casper SJT

In this study we used data collected from the Casper SJT (Dore et al., 2017). In Casper, scenarios are delivered either as text-based prompts (brief written descriptions of a situation) or as video-based prompts featuring scripted situations enacted by trained actors. Following each prompt, test takers answer associated questions within a set time limit, either by typing their answers or by recording video responses. For this study, we focused on text-based scenarios with typed responses to avoid variability due to format differences. Below is an example of a text-based scenario included on the Casper SJT:

Think of a time when you had to make a sacrifice in order to accomplish a goal.

Test takers would answer two questions related to this scenario:

1. Briefly describe the situation and the sacrifice you made to accomplish the goal.
2. Imagine that despite the sacrifice made, you failed to accomplish the goal. How would

you have felt in this situation? Explain your reasoning.

Each response is subsequently evaluated by a trained rater and assigned a score on a 1–9 Likert scale. Human raters receive annual training and scenario-scoring guidelines that help them contextualize the relevant competencies, evaluate how effectively responses address the questions, and when to apply concerning flags when content warrants reporting to admissions programs.

3.2 Concerning Flag Process

Raters are instructed to raise a concerning flag when an applicant’s response contains hateful or offensive content, or demonstrates prejudice, stereotyping, or discrimination. Raters are also instructed to flag responses that dismiss or stigmatize personal characteristics or mental health, use these to justify unfair treatment, propose unethical courses of action, or display overtly unprofessional conduct. When raising a flag, raters must provide a written justification. For instance:

Applicant Response: “Its not expected because Usually its women who take the responsibility of changing diapers and hence at many places changing table is in women’s bathrrom”

Rater comment: “Sexist comment. This applicant said ‘women who take the responsibility of changing diapers’. I think this is sexist comment because of gender stereotyping.”

Once a rater raises a flag, subject matter experts (SMEs) verify whether the concern is valid. Notably, only 11.96% of concerning flags raised by raters are ultimately verified as genuinely concerning, highlighting both the subjectivity of initial flagging decisions and the substantial burden on human reviewers.

3.3 Task Formulation

This two-stage process of initial rater flagging followed by SME verification forms the basis of our labeling scheme. We formulated concerning flag prediction as a binary classification task:

1. **Concerning (1):** Responses where a concerning flag was raised AND verified as concerning after review

Table 1: Dataset labeling scheme based on flag outcome.

Flag Outcome	Label	Count
Flagged + Verified	Concerning	312
Flagged + Not Verified	Hard Negative	2,296
Non-Flagged	Non-Concerning	2,848
Total Available		5,456

2. **Non-Concerning (0):** Responses where either no flag was raised, or a flag was raised but not verified

This labeling prioritized precision by requiring human verification, reducing noise from initial flags. We incorporated the rejected flags as hard negatives during training, hypothesizing that exposure to “near-miss” cases would improve the model’s ability to distinguish truly concerning content from superficially similar but acceptable responses. In machine-learning terminology, concerning responses constitute the positive class; non-concerning responses the negative class.

3.4 Dataset Descriptive Statistics

The dataset spans 2016–2025. Concerning content is exceptionally rare (<1% of responses): across this period, verified-concerning responses number only 312, against millions of non-concerning responses. We retained all 312 verified positives and all 2,296 rejected flags (responses flagged by raters but not verified by SMEs), and drew a random sample of 2,848 never-flagged responses from the much larger pool, yielding a working set of 5,456 responses across three categories (Table 1). Initial experiments on the natural 9:1 imbalance revealed that models achieved high precision (>0.85) but very low recall (<0.25), missing most concerning content. To address this, we undersampled the data to a 3:1 ratio, pairing the 312 positives with 936 sampled negatives to form a balanced set of 1,248 examples (Altalhan et al., 2025). We then split this balanced set using stratified sampling into training (798 / 64%), validation (200 / 16%), and test (250 / 20%).

4 Model Architecture, Training, and Evaluation

We selected open-weight models throughout to preserve applicant privacy, avoiding the transmission of responses to third-party APIs and keeping all models deployable on local infrastructure. We focused on BERT and Toxic-BERT as encoders, and

selected Qwen2.5-7B and Ministral3-8B as representative, strong-performing open-weight LLMs at a comparable parameter scale, drawn from distinct model families.

Encoder Architecture and Training. We implemented a transfer learning approach, fine-tuning BERT and Toxic-BERT for this domain-specific task. Toxic-BERT is a BERT-base model pre-trained on the Jigsaw Toxic Comment Classification dataset, providing relevant language context for identifying harmful content. The base encoder has 12 transformer layers (768 hidden dimensions, 12 attention heads; $\sim 109\text{M}$ parameters). We used the final-layer [CLS] representation (Sun et al., 2019), applied dropout (0.1), and projected to two output classes (non-concerning, concerning). Inputs were tokenized with the BERT WordPiece tokenizer and truncated/padded to 512 tokens; responses average ~ 150 words (~ 180 tokens), so truncation was rarely triggered.

Loss Functions. To address class imbalance, we compared two loss functions: standard Binary Cross Entropy (BCE) and Focal Loss (Lin et al., 2020). Focal Loss down-weights well-classified examples and focuses learning on difficult cases, defined as:

$$FL(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t is the predicted probability of the correct class. We set the focusing parameter $\gamma = 2.0$ to reduce the contribution of easy examples, and the class weighting parameter $\alpha = 0.25$ to further balance the loss contribution between classes.

LLM Zero-Shot and Fine-Tuned Baselines. To assess whether zero-shot reasoning with open-weight LLMs can complement fine-tuned encoder performance, we evaluated two instruction-tuned models: Qwen2.5-7B-Instruct (Qwen Team, 2025) and Ministral3-8B (Mistral Team, 2026). We prompted each model with zero-shot classification, providing the scenario context, questions, and applicant responses along with explicit inclusion and exclusion criteria for concerning content derived from our guidelines. The criteria encompassed eight categories aligned with our flagging rubric: sexist, racist, homophobic, transphobic, unethical, unprofessional, mental health stigma, and inflammatory statements. Critically, we instructed the model that having a personal stance on sensitive topics is not itself concerning, only responses containing prejudice, discrimination, or unprofessional

content should be flagged. The complete prompt template is provided in Figure 1.

Binary labels were extracted via rule-based parsing: if the output contained “Non-Concerning” (or variants with spaces/hyphens), the response was classified as non-concerning; otherwise, if “Concerning” appeared, it was classified as concerning.

LLM Training Configuration. To investigate whether domain-specific fine-tuning improves over zero-shot performance, we employed QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) for parameter-efficient fine-tuning of Qwen2.5-7B and Ministral3-8B. We fine-tuned LLMs for 3 epochs using a stratified train/test split, preserving the original class distribution in the test set while downsampling the majority class to a 3:1 ratio during training to address the extreme class imbalance.

Evaluation. Given the class imbalance inherent in our task, accuracy alone is an insufficient metric. We report per-class precision, recall, and F1 scores, with particular attention to the concerning class given its importance for the downstream use case. During validation, we additionally evaluated performance across multiple classification thresholds (0.1 to 0.9) to identify the optimal operating point. This allowed us to tune the precision-recall trade-off based on deployment requirements (e.g., prioritizing recall to minimize missed concerning content at the cost of more false positives requiring human review).

We address our four research questions as follows: (RQ1) a systematic comparison of BERT variants (standard BERT and domain-adapted Toxic-BERT), with and without fine-tuning, across loss functions (BCE, focal) and tuned thresholds; (RQ2) evaluation of zero-shot LLMs with enriched contextual input on the same test set; (RQ3) a comparison of fine-tuned LLMs against fine-tuned BERT; and (RQ4) qualitative analysis of error patterns across all configurations, focusing on stance-based concerns and vocabulary bias.

5 Experiment Results and Analysis

Model Performance Comparison.

In Table 2, we present comparative performance of all model configurations across encoders and LLMs.

Encoder Baselines vs. Fine-Tuned Models. Domain-specific pre-training via Toxic-BERT provided modest transfer learning benefits compared

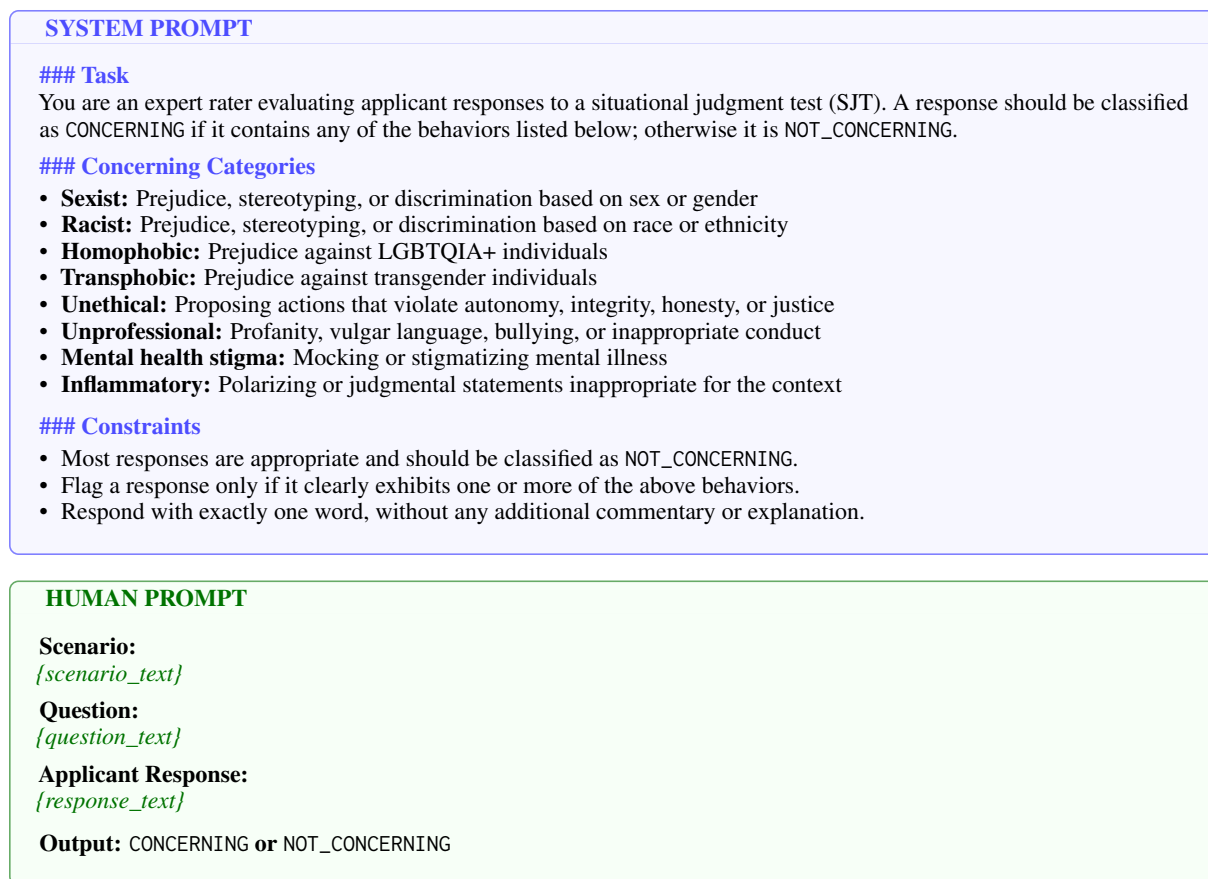


Figure 1: Prompt structure for LLM-based concerning content detection. The system prompt defines the task, the flagging categories, and constraints; the human prompt supplies the scenario, question, and applicant response.

to baseline BERT (0.092 vs. 0.000 F1 concerning), but did not substantially improve fine-tuned performance. Among the fine-tuned encoders, the choice of loss function demonstrated significant impact. Focal Loss configurations achieved higher precision (0.65 for Toxic-BERT) but reduced recall (0.23), while BCE Loss maintained higher recall (0.40 for BERT). Notably, BERT with Focal Loss achieved the best encoder performance (0.460 F1_{conc}), suggesting that general-purpose pre-training combined with minority-class-focused loss functions may be advantageous. This precision-recall trade-off has important implications, as discussed in Section 5.1.

Zero-Shot LLMs. Zero-shot performance was strongly model-dependent: Qwen2.5-7B achieved 0.395 F1_{conc}, approaching fine-tuned BERT’s 0.460 despite receiving no domain-specific training, while Ministral3-8B lagged well behind at 0.190. Because the two LLMs are of comparable scale and both instruction-tuned, this gap between them is unlikely to stem from capacity; it more plausibly reflects differences in instruction-

tuning data and alignment across the two model families. Relative to BERT, Qwen’s competitive zero-shot performance can be attributed to three factors: (1) substantially larger model capacity (7B vs. 110M parameters), enabling richer language understanding; (2) instruction-tuning during pre-training, which aligns the model to follow classification directives; and (3) enriched prompting that provides task-relevant context including the scenario description, question text, and inclusion/exclusion criteria, allowing the model to leverage its pre-trained knowledge for effective zero-shot reasoning. This last factor reflects an architectural distinction. Unlike instruction-tuned LLMs, BERT’s encoder cannot interpret classification instructions provided in natural language: where LLMs generate responses conditioned on prompt instructions, BERT produces fixed-dimensional embeddings mapped to class labels through a learned linear layer. Consequently, BERT must learn the concerning/non-concerning distinction implicitly from labeled examples, whereas LLMs can leverage explicit definitions (e.g., "sexist," "unethical")

at inference time. This explains why enriched prompting benefits LLMs but not encoder-based models.

Fine-Tuned LLMs. Qwen2.5-7B fine-tuned achieved very high precision (0.89) but extremely low recall (0.22) for imbalanced data settings, flagging only the most confident cases while missing substantial portions of concerning content. Balancing to 3:1 shifted the recall (0.53) and precision (0.28), leaving $F1_{\text{conc}}$ essentially unchanged (0.356 $\rightarrow$ 0.362). Ministral3-8B (3:1) behaved similarly, reaching the highest recall in our evaluation (0.58) at low precision (0.24) for an $F1_{\text{conc}}$ of 0.339. In all configurations, fine-tuned LLMs remained below fine-tuned BERT on the concerning class (best $F1_{\text{conc}}$ 0.460). This result is not surprising given that the BERT is inherently designed for discriminative prediction and may benefit from stronger inductive biases for binary classification under limited data conditions. In contrast, the LLMs were originally trained for next-token prediction and subsequently adapted to classification, which may make them less efficient in this setting. Additionally, fine-tuning did not improve Qwen over its own zero-shot result (0.395), suggesting that fine-tuning on a small set of data diminished model pre-training competence that had made the model effective zero-shot.

5.1 Error Analysis

We conducted a detailed qualitative analysis of a subset of the false negatives produced by the best encoder, BERT with focal loss. Qualitative coding of these responses revealed distinct categories of concerning content that the model failed to detect.

False Positive Analysis. The false positives shared a common characteristic: test takers were engaging thoughtfully with sensitive social topics while using relevant vocabulary associated with concerning content. For example, one response discussing workplace gender bias was flagged due to terms like “sexist” and “dismissing female applicants,” despite demonstrating critical thinking and ethical reasoning. Similarly, a response critically analyzing the harms of “slut shaming” was incorrectly classified as concerning. These errors revealed a vocabulary bias inherited from the pre-training, where the model learned to associate certain terms with toxicity regardless of contextual usage. Within an assessment meant to elicit responses in various nuanced situations, this creates a problematic pattern where applicants demonstrat-

ing sophisticated engagement with social issues may be incorrectly flagged.

False Negative Analysis. The most prevalent category, *ethical relativism* (n=8), comprised responses normalizing academic dishonesty, deception, or rule-bending. These responses typically used formal, polite language while expressing problematic ethical positions, such as: “I believe if it benefits yourself or others then it is fair. If it’s not stealing money or any crime then it’s fair game.” *Dismissive responses* (n=6) presented a detection challenge due to their brevity. Responses such as “nooooo,” “nope cause fu** them” lacked sufficient textual signal for the model to assess concerning patterns. *Discriminatory undertones* (n=5) represented particularly subtle errors where test takers expressed bias through seemingly reasonable arguments. One response advocating for relocating a homeless shelter stated concerns about “break-ins and theft,” reflecting stigmatizing assumptions without using explicitly discriminatory language.

6 Discussion and Conclusion

Our experiments yield four principal findings to address our research questions. First, the benefit of domain-specific pre-training was confined to the zero-shot setting and did not persist after fine-tuning. Without domain fine-tuning, Toxic-BERT’s toxicity pre-training gave it a smaller edge over baseline BERT on the concerning class (0.092 vs. 0.000 $F1_{\text{conc}}$). Once both encoders were fine-tuned, however, this advantage disappeared: general-purpose BERT with Focal Loss achieved the strongest performance overall (0.460 $F1_{\text{conc}}$), exceeding both fine-tuned Toxic-BERT configurations (0.340–0.350). This suggests that while toxicity pre-training provides a useful prior in the absence of task-specific data, it offers no lasting benefit. Second, zero-shot LLMs with enriched context did not outperform fine-tuned BERT on the concerning class (0.395 F1 for Qwen2.5-7B vs. 0.460 for BERT) despite being roughly 64 times larger. Third, fine-tuning the LLMs failed to close this gap: even with instruction tuning and domain-specific fine-tuning, the best LLM configuration reached only 0.362 F1 on the concerning class, still lower than fine-tuned BERT. Fourth, and most critically, current models exhibit vocabulary bias over reasoning assessment. Models are triggered by sensitive vocabulary regardless of context while missing concerning reasoning expressed in neutral

Table 2: Performance comparison across model configurations. $F1_{\text{conc}}$, Precision, and Recall are for the concerning class. Optimal thresholds determined via validation set evaluation (0.1–0.9). Baselines are applied without domain fine-tuning. Best values in **bold**.

Model	Loss	Test F1	$F1_{\text{conc}}$	Prec.	Recall	Best Ep.	Opt. Th.
<i>Baseline encoders (no domain fine-tuning)</i>							
BERT	–	0.485	0.000	0.00	0.00	–	–
Toxic-BERT	–	0.532	0.092	1.00	0.05	–	–
<i>Fine-tuned encoders</i>							
BERT	BCE	0.640	0.440	0.49	0.40	10	0.55
BERT	Focal	0.650	0.460	0.53	0.40	17	0.20
Toxic-BERT	BCE	0.600	0.350	0.48	0.28	13	0.40
Toxic-BERT	Focal	0.610	0.340	0.65	0.23	9	0.40
<i>Baseline LLMs (no domain fine-tuning)</i>							
Qwen2.5-7B (zero-shot)	–	0.682	0.395	0.38	0.42	–	–
Ministral3-8B (zero-shot)	–	0.579	0.190	0.22	0.17	–	–
<i>Fine-tuned LLMs: Effect of Class Balancing</i>							
Qwen2.5-7B (imbalanced, 9:1)	–	0.668	0.356	0.89	0.22	–	–
Qwen2.5-7B (balanced, 3:1)	–	0.659	0.362	0.28	0.53	–	–
Ministral3-8B (balanced, 3:1)	–	0.642	0.339	0.24	0.58	–	–

language. This represents a fundamental limitation for educational applications where the goal is assessing attitudes rather than detecting toxic vocabulary.

6.1 Deployment Implications

We envision this tool serving two complementary functions in operational deployment, though given the persistence of false positives, human review of all flags remains essential in any high-stakes context. First, as a **triage tool**, the model can prioritize responses for verification review. Given that 88% of human-raised flags in our dataset were ultimately rejected, running the tool on flagged responses before human verification could allow the review team to prioritize high-confidence cases and reduce time spent on likely false alarms. Responses where both the rater and model agree on concern would receive immediate attention, while rater-only flags with low model confidence could be batched for lower priority review. Second, this tool can support **rater quality assurance and training** by identifying potentially concerning responses that human raters may have missed, enabling targeted feedback and corrective actions. While empirically imperfect, our results reflect the reality of deploying machine learning in complex, real-world settings. A fully-manual process is itself imperfect as raters miss concerning content, apply inconsistent standards, and raise flags that are often rejected.

This tool adds value by complementing human judgment by catching what raters may miss and prioritizing what they flag. A system combining human intuition with model-based screening will likely outperform either approach alone.

6.1.1 Implications for Educational Safety.

The false negatives consist of responses expressing concerning attitudes, including normalization of dishonesty, discriminatory views, and disengagement, which would incorrectly not be flagged for SME follow-up. Conversely, the false positives represent responses engaging constructively with difficult topics which might incorrectly receive flags, but these can be filtered during human review.

From a safety perspective, false negatives carry greater risk, as concerning attitudes may go unaddressed. This tension reflects a fundamental challenge in automated detection of concerning content in educational settings.

6.2 Limitations

Three key limitations constrain our approach: (1) models struggle with subtle concerns in professional language; (2) limited training data (312 verified flags) restricts learning, especially for rare categories; (3) evaluation was limited to a single assessment.

6.3 Future Work

First, synthetic data augmentation using a LLM could address training data scarcity, particularly for underrepresented categories like ethical relativism and implicit bias. Second, larger open-weight instruction-tuned models that can be fine-tuned on-premise (e.g., Mistral-Small 24B, GPT-OSS-20B, or GPT-OSS-120B) may improve stability and performance beyond current approaches while preserving applicant privacy. Closed, API-only models such as GPT-5 or Claude Opus 4.8 cannot be fine-tuned locally, but their stronger zero-shot reasoning may make them useful in the enriched-context setting explored here. Third, multi-task learning that jointly predicts concern categories alongside binary flags would provide interpretable output for human reviewers, while explainability methods such as attention visualization could support appropriate oversight. Finally, this approach could extend to other constructed-response assessments, including low-stakes, ongoing, and in-process evaluations used to monitor student progress.

References

- Manahel Altalhan, Abdulmohsen Algarni, and Monia Turki-Hadj Alouane. 2025. [Imbalanced data problem in machine learning: A review](#). *IEEE Access*.
- Puxin Xu Chunting Zhou, Pengfei Liu and 1 others. 2023. [Lima: Less is more for alignment](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Advances in Neural Information Processing Systems*, 36.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.
- Kelly L. Dore, Harold I. Reiter, Sharyn Kreuger, and Geoffrey R. Norman. 2017. [CASPer, an online pre-interview screen for personal/professional characteristics: prediction of national licensure scores](#). *Advances in Health Sciences Education: Theory and Practice*, 22(2):327–336.
- Laura Hanu and Unitary team. 2020. [Detoxify](#). <https://github.com/unitaryai/detoxify>.
- Jigsaw. 2020. Jigsaw multilingual toxic comment classification. <https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification>.
- Ashish Kumar, Nilesh Ware, and Sachin Gupta. 2024. [Leveraging transfer learning: fine-tuning methodology for enhanced text classification using BERT](#). In *2024 IEEE Pune Section International Conference (PuneCon)*, pages 1–5, Pune. IEEE.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. [Focal loss for dense object detection](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.
- Tianyi Liu. 2024. [Data diversity matters for robust instruction tuning](#). *arXiv preprint arXiv:2311.14736*.
- Kristína Machová, Marián Mach, and Kamil Adamišin. 2022. [Machine learning and lexicon approach to texts processing in the detection of degrees of toxicity in online discussions](#). *Sensors*, 22(17):6468.
- 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.
- Mistral Team. 2026. [Ministral 8b](#).
- Anabel Pilicita and Enrique Barra. 2025. [LLMs in education: evaluation GPT and BERT models in student comment classification](#). *Multimodal Technologies and Interaction*, 9(5):44.
- Qwen Team. 2025. Qwen2.5: A pretrained language model for generation and understanding. <https://arxiv.org/pdf/2412.15115>.
- Wenpeng Yin Renze Lou, Kai Zhang. 2025. [Large language model instruction following: A survey of progresses and challenges](#). *Computational Linguistics*, 50(3).
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. [How to fine-tune BERT for text classification?](#) In *Proceedings of the 18th China National Conference on Computational Linguistics (CCL)*, pages 194–206.
- Kurt Thomas, Patrick Gage Kelley, David Tao, Sarah Meiklejohn, Owen Vallis, Shunwen Tan, Blaž Bratanič, Felipe Tiengo Ferreira, Vijay Kumar Eranti, and Elie Bursztein. 2025. [Supporting human raters with the detection of harmful content using large language models](#). In *Proceedings of the 2025 IEEE Symposium on Security and Privacy (S&P)*, pages 2772–2789. IEEE.
- Zeeraak Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. [Understanding abuse: a typology of abusive language detection subtasks](#). In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.