Research Square

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

Incremental Value of Structured Metadata for Text Based Risk Triage in a Role Play Derived University Counselling Corpus

Danny Leza

danny.leza@unza.ac.zm

University of Zambia

Jackson Phiri

University of Zambia

Research Article

Keywords: mental health informatics; natural language processing; risk triage; feature ablation; algorithmic fairness; calibration; African universities

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11120039/v1

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

Additional Declarations: No competing interests reported.

Incremental Value of Structured Metadata for Text Based Risk Triage in a Role Play Derived University Counselling Corpus

Danny Leza, Jackson Phiri

Department of Computer Science, University of Zambia, Lusaka, Zambia
Corresponding author: Danny Leza, danny.leza@unza.ac.zm

Abstract Automated triage systems for student mental-health support may combine free-text messages with demographic and intake fields, but the incremental value of structured data requires empirical testing. This study compared text-only, structured-only and combined logistic-regression classifiers using a role-play-derived university counselling corpus. The original corpus contained 673 scenarios and the expanded corpus 1,672 scenarios; complete-case analyses used 654 and 1,509 scenarios, respectively. Under five-fold shuffled cross-validation, text-only escalation ROC-AUC was 0.705 (original) and 0.805 (expanded), compared with 0.687 and 0.799 for combined models and 0.633 and 0.752 for structured-only models. Paired bootstrap and repeated-split analyses found no demonstrated advantage from adding structured variables, while structured-only performance remained lower than text-only performance. Problem-category indicators carried most of the structured signal; demographic fields contributed little. At a 0.5 threshold, text-only sensitivity was 62.8% and 66.8%, while Critical-class recall in the four-class model was 25.0% and 64.9%. Achieving 95% escalation sensitivity required flagging 87% and 78% of cases. Structured metadata therefore provided no consistent incremental benefit once message text was available. However, participant-level grouping was unavailable, the corpus was authored through role-play, and no model achieved performance suitable for autonomous safety-critical triage. Independent validation on authentic counselling text is required before operational use.

Keywords mental health informatics; natural language processing; risk triage; feature ablation; algorithmic fairness; calibration; African universities

1 Introduction

University students experience a substantial burden of common mental disorders, while treatment uptake remains low in many settings. The World Mental Health International College Student initiative documented high levels of mental disorders among first-year students across countries [1], and South African evidence shows that many students with mental-health problems do not receive treatment [2]. These pressures are particularly consequential in African universities, where counselling services operate within broader shortages of specialist mental-health personnel and uneven access to care.

Natural language processing (NLP) can identify clinically relevant patterns in counselling conversations and other text, but published systems differ widely in their populations, labels and validation procedures [3]. Suicide-risk studies have shown that language contains useful predictive signal [4], [5], while recent counselling-specific work demonstrates that fine-grained risk assessment from text is technically feasible [6]. Such evidence supports research on triage assistance, but it does not establish that a classifier is safe for autonomous decision-making.

A recurring design choice is whether to add demographic and intake metadata to text features. More variables can improve discrimination when they contain independent information, yet sensitive attributes can also reproduce inequities or act as proxies for social disadvantage [7], [8]. The relevant question is therefore empirical: does the structured information improve out-of-sample performance after the model has access to the student's own words? This paper reports a controlled feature-ablation study in an African university context. It is framed throughout as a methodological study of an AI-co-designed, role-play-derived scenario corpus — not as evidence that any tested model is ready for real triage decisions.

2 Related work

2.1 Mental health classification from language

Systematic reviews describe rapid growth in NLP for mental-health interventions, alongside limited linguistic diversity, population bias and incomplete reproducibility [3]. Earlier social-media studies established expert-labelled risk assessment from online posts [4], while shared-task research showed that preprocessing and evaluation decisions materially affect reported performance [5]. PsyGUARD extends this line to psychological counselling and provides a detailed suicide-ideation taxonomy [6]. These studies motivate text-based modelling but also show why dataset origin and label construction must be disclosed.

2.2 Demographic features and fairness

Clinical prediction systems may reproduce inequity when protected characteristics or their proxies encode unequal access, documentation or treatment [7]. Fairness cannot be reduced to deleting a sensitive variable, because correlated text and contextual variables may still carry demographic information [8]. Nevertheless, data minimisation in the predictive model is valuable when a sensitive field adds no demonstrated benefit there, even though the same fields may still be needed elsewhere in a system for representativeness and fairness auditing. A feature-ablation design provides direct evidence about incremental predictive value and supports a more proportionate decision about what belongs in the model itself [9].

2.3 Validation in small structured datasets

Cross-validation estimates can be noisy in small samples [10] and optimistic when dependent observations are separated across training and test data [11]. Leakage may occur during preprocessing, label construction or splitting [12]. We therefore refit text transformations inside each fold, exclude near-deterministic label-derived tags, report the grouping structure of the evaluation explicitly (Section 4.4), and supplement a single-split analysis with a repeated-split check to characterise sensitivity to the fold assignment itself.

3 Objectives and research questions

The study evaluates whether structured demographic and problem-category features improve risk classification beyond free text, whether structured features are useful on their own, whether that structured signal is driven by demographics or by problem categories, and whether model ranking is stable to the choice of fold assignment and to a substantially larger, though not directly comparable, corpus version. The two corpus versions differ in more than size (Section 4.1), so any difference between them is reported as an association, not a controlled effect of sample size. Consistent with the corpus's authored, role-play origin, the study is scoped as a methodological feature-ablation exercise, not as a validation of clinical deployment readiness. The research questions are:

RQ1: Does adding structured fields improve escalation ROC-AUC relative to text alone, and is any such difference distinguishable from evaluation noise under both case-resampling and fold-assignment perturbation?

RQ2: How do text-only, structured-only and combined models compare on four-class risk-level macro-F1, and on class-specific recall for the Critical risk category?

RQ3: Is the structured-only signal attributable to demographic fields, problem-category fields, or both, and does the text-only model's ranking hold across the two corpus versions?

4 Materials and methods

4.1 Study design and data

This retrospective computational study used short first-person counselling scenarios developed for the broader UNZA doctoral research programme by a fixed team of 12 students and 10 counsellors, recruited and compensated by the researcher for the scenarios they produced. In role-play, each student voiced an affected person presenting a mental-health-related issue — drawn from problems the student had observed among peers rather than necessarily their own — while a counsellor responded in their professional capacity; the exchange was documented into structured records comprising message text, a four-level risk label (Low, Moderate, High or Critical), problem-category

indicators, year of study, gender and residence. Neither the risk labels nor the problem categories were assigned arbitrarily. Problem categories were fixed during the first part of the role-play workshop itself, informed by an earlier GHQ-12 mental-health screening survey of the student body that identified which issues were actually prevalent; that same first, three-day workshop then produced slightly above 100 scenarios (107 scenarios) and a subsequent five-day workshop grew the corpus to the 673-scenario original corpus analysed here. Risk_Level was derived from the GHQ-12 survey analysis — the same four-band distress framework used to score the survey informed how role-played presentations were assigned a risk band — with each scenario's assigned level then validated by the counsellor who conducted that role-play. The binary escalation target combined High and Critical as escalation cases and Low and Moderate as non-escalation cases.

Data collection did not stop after the second workshop: the same team continued producing scenarios in an ongoing, less formally scheduled arrangement, and the shared worksheet has kept growing since. The expanded corpus analysed here (1,672 scenarios: 672 matched to the original corpus plus 1,000 further scenarios from this continued collection) is a snapshot of that growth, not a second discrete, controlled batch — which further supports treating the original-versus-expanded comparison as an association rather than a controlled sample-size effect (Section 3, Section 6). Complete-case comparisons used n=654 and n=1,509 because the combined and structured pipelines required non-missing year-of-study and residence fields. Of the 673 original-corpus scenarios, 672 carry through into the expanded corpus's matched-original batch; the excluded case (original identifier 123) had no usable message content — its Student_Message field held only a two-character placeholder (";.") — and was dropped as a data-quality exclusion. The expanded corpus also introduced broader gender and residence category sets, a further batch difference; expanded-corpus estimates remain internal performance on the combined authored corpus, not evidence of transfer to authentic counselling records.

Generative AI assisted specific, identifiable parts of corpus construction beyond the role-play itself. Drafts of the Ideal_AI_Response field were reviewed and, where flagged as unsafe or irrelevant, rewritten with AI assistance; two documented corrections are a response that had hallucinated an unrelated medical event and a response to a disclosure of intimate-partner violence that inappropriately placed responsibility on the survivor. The augmented message variants mentioned above were generated by an automated paraphrasing step and screened with an automated quality filter. A subsequent curation pass added a Quality_Score/Quality_Notes rating and a controlled Notes_for_AI action-vocabulary tag set. None of this documented AI involvement concerns the Student_Message, Risk_Level or Problem_Category fields used as model inputs and targets in this study; those originate from the human role-play and its labelling.

Because the counsellor who validated a scenario's risk level is the same person who had just conducted that role-play, this validation was not blinded to the demographic fields, the specific wording used, or the conversational context established during the exchange; whether more than one rater independently assessed any scenario, and what a disagreement-resolution procedure (if any) looked like, is not established. This is treated as a limitation (Section 7) rather than an open question, since the answer — unblinded, single-rater validation by construction of the workflow — follows directly from the process just described.

4.2 Data quality and leakage controls

Duplicate text was removed before modelling. The problem-category tags "Suicidal" and "Crisis" were excluded because they were assigned through the same rating process as risk level and were near-deterministic with the outcome. Retaining them would allow the predictor set to reproduce the label rather than learn a generalisable relationship. Perfectly collinear category pairs, including GBV and Trauma, were merged into single indicators. All preprocessing that learned corpus statistics was fitted on training folds only.

4.3 Feature sets and classifiers

The text-only pipeline used TF-IDF representations of unigrams and bigrams, minimum document frequency of two, English stop-word removal and sublinear term-frequency scaling. The structured-only pipeline used numeric year of study, one-hot encoded gender and residence, and cleaned problem-category indicators; this set was further decomposed into a demographic-only subset (year, gender, residence) and a problem-category-only subset (Table 7) to attribute the source of any structured signal. Numeric features were standardised. The combined pipeline concatenated the sparse text matrix and transformed structured features. Each pipeline used L2-regularised logistic regression (C=1.0, class_weight="balanced", solver=lbfgs, max_iter=3000), fixed in advance with no hyperparameter search, using scikit-learn 1.8.0. Multiclass classification used the same feature definitions and macro-F1 as the primary summary metric, with class-specific precision/recall/F1 reported for the Risk_Level task (Table 6) and the arg-max multiclass decision rule used throughout — distinct from the 0.5 probability threshold used for the binary escalation metrics in Table 4.

4.4 Evaluation and statistical reporting

Models were evaluated with five-fold cross-validation, fold assignment shuffled (random_state=42 for the primary/canonical split reported in Tables 1–2 and 6–8). Folds were originally allocated using GroupKFold with scenario identifier as the group, intended to prevent a scenario from contributing to both the training and test fold in the same split. However, every Scenario_ID in both complete-case corpora occurs exactly once; grouping by scenario identifier therefore provided no protection beyond an ordinary shuffled K-fold. No participant, counsellor, role-play session or author identifier was available in the data as supplied to test for latent non-independence between nominally

distinct scenarios. Given that only 22 people (12 students, 10 counsellors) produced all 673–1,672+ scenarios (Section 4.1), this is treated as a likely rather than merely hypothetical source of optimism in the reported estimates; it is reported as an open limitation (Section 7), not corrected for, pending the confirmation requested in Section 4.1.

Two complementary summaries of the same primary split are reported for every model: the mean±SD of ROC-AUC (or macro-F1) computed separately on each of the five held-out folds, and the pooled ROC-AUC/macro-F1 computed once on out-of-fold predictions concatenated across all five folds. Both are generated by the same code path from the same fold assignment and model fits, so any difference between them (typically ≤0.003 in this corpus) reflects only the two ways of summarising five correlated fold results, not a different run. All escalation-target figures in Tables 1, 3 and 7, and all four-class figures in Tables 2 and 6, follow this convention. Because five correlated folds give limited information, the fold SD is reported as a descriptive measure of fold dispersion, not a confidence interval [10], and Table 3's bootstrap interval targets sampling variation in the case set rather than fold-assignment variation. To characterise sensitivity to the fold assignment itself, a separate check repeated the entire five-fold procedure 20 times with different random seeds and summarised the resulting distribution of pooled AUC differences (Section 5). Neither analysis constitutes a formal equivalence or non-inferiority test; no non-inferiority margin was pre-specified, so a confidence interval spanning zero is reported only as an absence of demonstrated superiority, not as proof that two models perform equivalently.

Safety-relevant performance for the binary escalation target was summarised at the conventional 0.5 probability threshold (sensitivity, specificity, PPV, NPV, confusion-matrix counts) and, because that threshold has no established clinical basis in this setting, additionally at the thresholds required to reach 80%, 90% and 95% sensitivity, each reported with the resulting specificity and the proportion of all cases that threshold would flag positive (Table 4). Threshold-independent PR-AUC and a calibration check (Brier score; logistic calibration slope/intercept from regressing the outcome on the logit of the predicted probability) were also reported. Two baselines were evaluated for context: a fold-respecting majority-class classifier, and a 27-term crisis-keyword rule (full lexicon in Table 8's note), applied by case-insensitive substring match without any tuning to this corpus and reported as exploratory. Missing-data associations (Section 5) were summarised with Cramér's V (categorical associations) and Cohen's d (message-length difference) alongside chi-square p-values.

Ethical approval was granted by the University of Zambia Natural and Applied Sciences Research Ethics Committee (NASREC; approval number NASREC: 2026-APR-012; 4 May 2026). All methods were performed in accordance with the World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants and applicable NASREC guidelines and regulations. The approved protocol covered the GHQ-12 survey and counsellor-student role-play; all procedures occurred after approval, and all participants received study information and provided written informed consent before participation.

5 Results

Text-only models produced the highest escalation ROC-AUC in both complete-case datasets. Adding structured variables reduced mean ROC-AUC in both corpora.

Table 1 Escalation classification using five-fold shuffled cross-validation for the primary split

Model	Original n=654 (fold-mean±SD / pooled)	Expanded n=1,509 (fold-mean±SD / pooled)
Text only	0.705±0.035 / 0.704	0.805±0.021 / 0.803
Structured only	0.633±0.013 / 0.631	0.752±0.018 / 0.753
Text and structured	0.687±0.029 / 0.691	0.799±0.013 / 0.799

Fold-mean±SD and pooled out-of-fold AUC from the identical primary split and model fits (Section 4.4); the two summaries agree to within 0.003 in every row.

The four-class task was more difficult, with the same ordering.

Table 2 Four-class risk-level classification for the primary split

Model	Original n=654 (fold-mean±SD / pooled)	Expanded n=1,509 (fold-mean±SD / pooled)
Text only	0.389±0.054 / 0.391	0.460±0.033 / 0.463
Structured only	0.243±0.028 / 0.247	0.423±0.015 / 0.426
Text and structured	0.350±0.042 / 0.354	0.462±0.016 / 0.465

Macro-F1, fold-mean±SD and pooled out-of-fold value from the identical primary split.

A text-only sensitivity analysis retained scenarios with incomplete structured fields. ROC-AUC was 0.696±0.048 for all 673 original scenarios and 0.793±0.009 for all 1,672 expanded scenarios, the same broad pattern without complete-case restriction.

5.1 Uncertainty in model differences

Two independent checks were used: a paired bootstrap over cases (fixed fold assignment) and a repeated check that instead varied the fold assignment itself (20 different random seeds, same models refit each time).

Table 3 Uncertainty in AUC differences for the escalation target

Comparison	Case-bootstrap 95% CI	20-repeat fold-assignment range
Combined − Text (n=654)	−0.013 [−0.050, +0.023]	−0.012 [−0.027, +0.002]
Combined − Text (n=1,509)	−0.003 [−0.017, +0.009]	−0.003 [−0.007, +0.002]
Structured − Text (n=654)	−0.073 [−0.126, −0.021]	−0.068 [−0.087, −0.048]
Structured − Text (n=1,509)	−0.050 [−0.075, −0.025]	−0.039 [−0.048, −0.028]

Case-bootstrap: 3,000 resamples on the primary split's out-of-fold predictions. Fold-assignment range: 2.5th–97.5th percentile of the pooled-AUC difference across 20 independent five-fold splits (different random seeds, full refit each time).

Both checks agree: the combined-versus-text-only difference was not distinguishable from noise under either source of evaluation variability, while the structured-only deficit was consistently negative under both. This is evidence against a demonstrated advantage for the combined model, not evidence that the two models are equivalent — no non-inferiority margin was defined in advance, and the case-bootstrap interval for the original corpus includes AUC differences (up to +0.023) that some applications might consider meaningful.

5.2 Safety critical performance of the text only model

Point estimates of discrimination do not describe operational safety. Table 4 reports threshold-based and calibration metrics; because a 0.5 cutoff has no established clinical basis here, operating points for 80/90/95% sensitivity are also reported.

Table 4 Safety-critical metrics for the text-only escalation model

Metric (threshold = 0.5)	Original n=654	Expanded n=1,509
Sensitivity / Specificity	0.628 / 0.675	0.668 / 0.799
PPV / NPV	0.603 / 0.698	0.646 / 0.814
PR-AUC	0.632	0.755
Brier / cal. slope / intercept	0.224 / 1.77 / −0.16	0.170 / 1.51 / −0.48

At 0.5, sensitivity of 0.63–0.67 means 33–37% of escalation cases are missed; slopes above 1 indicate under-confident probabilities. These estimates indicate miscalibration and motivate evaluating recalibration in an independent sample — recalibration is not demonstrated here.

Table 5 Operating points for target sensitivity using the text-only model

Target sensitivity	Orig. threshold / actual sens. / spec. / % flagged	Exp. threshold / actual sens. / spec. / % flagged
80%	0.444 / 0.80 / 0.48 / 64%	0.399 / 0.80 / 0.52 / 60%
90%	0.403 / 0.90 / 0.31 / 78%	0.348 / 0.90 / 0.38 / 72%
95%	0.359 / 0.95 / 0.19 / 87%	0.314 / 0.95 / 0.32 / 78%

To catch 95% of true escalation cases, 77–87% of all cases would need to be flagged positive at this threshold — a concrete illustration of the sensitivity/workload trade-off at default settings, not a claim that this specific threshold was clinically validated.

5.3 Performance for the highest risk category

Because a missed Critical case is the most consequential error type, Table 6 reports precision, recall and F1 separately for each Risk_Level class from the text-only model's out-of-fold, arg-max multiclass predictions — a different decision rule from the 0.5 binary threshold used in Table 4, reported separately to avoid conflating the two.

Table 6 Per-class metrics for the text-only risk-level model using out-of-fold arg-max predictions

Class	Orig. n Prec/Rec/F1	Support	Exp. n Prec/Rec/F1	Support
Critical	0.368 / 0.250 / 0.298	28	0.626 / 0.649 / 0.637	111
High	0.549 / 0.585 / 0.566	260	0.431 / 0.482 / 0.455	425
Moderate	0.565 / 0.548 / 0.556	303	0.473 / 0.327 / 0.387	667
Low	0.141 / 0.143 / 0.142	63	0.311 / 0.464 / 0.372	306

Precision/Recall/F1 pooled across the five primary-split folds; macro-F1 = 0.391 (original) and 0.463 (expanded), matching Table 2's pooled column by construction.

In the original corpus, Critical-class recall was only 0.250 — three in four Critical scenarios were misclassified into a lower risk category. The expanded corpus improved this to 0.649, plausibly reflecting its larger Critical-class sample (111 versus 28), but two-fifths of Critical cases remain misclassified even there. This is, in our view, the single most important safety-relevant number in this study.

5.4 Decomposition of structured feature signals

Table 7 decomposes the structured feature set into its demographic and problem-category components, with bootstrap 95% CIs (same procedure as Table 3) against the text-only model, to identify which subset carries the modest structured-only signal.

Table 7 Feature-set decomposition for pooled escalation ROC-AUC in the primary split

Feature set	Orig. n=654 AUC [95% CI vs. text]	Exp. n=1,509 AUC [95% CI vs. text]
Demographics only	0.564 [−0.197,−0.081]	0.566 [−0.270,−0.204]
Categories only	0.608 [−0.149,−0.042]	0.696 [−0.137,−0.077]
Demographics + categories	0.631 [−0.126,−0.021]	0.753 [−0.075,−0.025]
Text + demographics	0.700 [−0.030,+0.022]	0.802 [−0.007,+0.005]
Text + categories	0.685 [−0.052,+0.014]	0.801 [−0.013,+0.010]
Text + demographics + categories	0.691 [−0.050,+0.023]	0.799 [−0.017,+0.009]
Text only (reference)	0.704	0.803

Demographic fields carried little signal alone (AUC 0.564–0.566, CI clearly excluding zero but with the largest negative gap to text-only of any row) and did not improve on text-only when added to it (CI spans zero). Problem-category indicators carried more of the structured-only signal, especially in the expanded corpus (0.696 alone), which is expected because problem categories are themselves derived from message content and so partially redundant with text rather than independent information. This supports attributing the previously reported "structured-only" effect predominantly to problem categories rather than demographics.

5.5 Baseline comparison

A fold-respecting majority-class classifier and a 27-term crisis-keyword rule (suicide, suicidal, kill myself, end my life, end it all, self-harm, self harm, cutting myself, overdose, rape, raped, assault, assaulted, abuse, abused, die, dying, don't want to live, no reason to live, hopeless, worthless, gun, pills, hang myself, can't go on, want to disappear, harm myself; case-insensitive substring match, no tuning to this corpus) were evaluated for context.

Table 8 Baseline comparison for the escalation target

Model	Original n=654	Expanded n=1,509
Majority-class (accuracy / macro-F1)	0.560 / 0.359	0.645 / 0.392
Keyword rule (ROC-AUC / sens. / macro-F1)	0.555 / 0.177 / 0.501	0.562 / 0.153 / 0.525
TF-IDF text-only (ROC-AUC)	0.704	0.803

The keyword rule's high specificity but low sensitivity (0.15–0.18) indicates escalation in this corpus is not simply detectable from a small crisis-term list; the TF-IDF model's advantage over both baselines is consistent with it learning broader linguistic patterns, though this corpus's authored origin means the finding may not transfer to less formulaic, authentic text. Because the lexicon was not selected by inspecting this corpus, it is best read as an exploratory reference point rather than a tuned baseline.

5.6 Missing data patterns in the expanded corpus

The complete-case restriction excluded 163 of 1,672 expanded-corpus scenarios (9.7%), non-randomly: missingness differed by authoring batch (1.0% in the original 672-scenario batch vs. 15.6% in the additional 1,000-scenario batch; χ^2 $p=1.75\times10^{-22}$, Cramér's V=0.24, a small-to-medium association) and by Risk_Level (0.9% Critical to 15.9% Low; χ^2 $p=1.00\times10^{-7}$, Cramér's V=0.15, a small association). Excluded scenarios also had shorter messages (mean 19.1 vs. 30.9 words; Cohen's d=0.57, a medium effect). Because Critical and High cases were least likely to be excluded, the complete-case sample is, if anything, slightly enriched for higher-risk cases relative to the full expanded corpus.

6 Discussion

The principal finding is absence of a demonstrated incremental benefit from the structured fields, not proof that none exists. Free text alone matched or exceeded the combined pipeline under both the case-bootstrap and the fold-assignment robustness check, and the feature-set decomposition shows the modest structured-only signal is carried mostly by problem-category fields, which are themselves largely text-derived,

rather than by demographic fields, which performed close to chance on their own (Table 7). Neither the problem categories nor the risk-level framework were arbitrary: both trace back to an earlier GHQ-12 survey of the student body (Section 4.1), which grounds the category and risk definitions in independent survey evidence even though the scenario text itself remains authored role-play. That grounding is at the level of the framework, not the individual case, however: each scenario's specific risk level was validated by a single, unblinded counsellor rather than an independent rater (Section 7).

This supports leaving demographic fields out of the prediction model in this corpus. It does not, by itself, support removing them from data collection generally: the same fields may still be needed to test subgroup performance, measure representativeness, or audit for disparate false-negative rates — none of which was possible here without adequately sized subgroups. Nor does the text model's higher discrimination establish that it is fair, since language can encode demographic identity and social context indirectly [7], [8].

A second, arguably more consequential finding is that no tested model or default threshold reaches a sensitivity adequate for autonomous safety-critical use. At 0.5, the recommended text-only model missed roughly a third of escalation cases overall and, in the original corpus, three in four Critical-risk cases; reaching 95% sensitivity required flagging 77–87% of all cases positive, a workload that would likely be unworkable in practice without further model or system improvement. We agree with the framing that this study's primary contribution is methodological — an ablation and evaluation procedure that shows where this AI-co-designed corpus is and is not informative — rather than a demonstration that any tested configuration, at any threshold examined, is ready to inform real triage decisions. Within this role-play corpus specifically, the text representation provided most of the internally measured predictive signal; whether that reflects a general property of counselling text, or partly reflects that role-play messages were composed to express their intended label, cannot be distinguished with the data available here.

The higher scores in the expanded corpus should not be read as proof that authored augmentation improves real-world clinical performance, nor attributed to sample size alone: the two corpus versions differ in authoring batch, missingness pattern, category sets and Critical-class prevalence, so their comparison is associational, not a controlled sample-size effect. External validation must use authentic, temporally later and institutionally distinct data. Reporting should follow prediction-model guidance such as TRIPOD+AI, including intended use, participant flow, missingness, evaluation procedures and limitations [13]; Section 4.4 and this section's missing-data discussion are a step toward that standard, though full TRIPOD+AI and PROBAST+AI compliance additionally requires the provenance and independence detail flagged in Section 4.1. The unresolved provenance and independence questions are therefore treated as material limitations that require resolution in future validation work.

7 Limitations

I. Risk_Level validation was unblinded and single-rater by construction: the same counsellor who conducted a role-play also validated its risk level, with full knowledge of the demographic fields and conversational context, and no second rater or documented disagreement-resolution procedure (Section 4.1). This could inflate agreement between Risk_Level and the message content in ways a blinded or multi-rater protocol would not.

II. With only 12 students and 10 counsellors producing all scenarios, it is likely, not merely possible, that individual participants each contributed many scenarios; cross-validation could not be grouped by author because no such identifier was available in the data supplied, so reported performance is likely optimistic due to within-author or within-session dependence not captured by the scenario-level shuffled split (Section 4.1, Section 4.4). This is the study's most significant open internal-validity risk.

III. Problem-category features are themselves largely text-derived rather than independent of the message (Table 7), and, like Risk_Level, were fixed through the same workshop process rather than by a separate blinded procedure.

IV. The complete-case comparison excluded 163 expanded-corpus scenarios non-randomly, associated with authoring batch and Risk_Level (Section 5); the structured and combined pipelines were not re-evaluated with explicit missing-category indicators or fold-fitted imputation, so the size of this effect on the central ablation result is not established.

V. No threshold examined (including the 80/90/95%-sensitivity operating points in Table 5) was set in consultation with counsellors or validated as clinically appropriate; they illustrate a trade-off rather than recommend an operating point.

VI. Calibration slope and intercept are reported without confidence intervals or a calibration plot, and were estimated on pooled out-of-fold predictions from the balanced logistic model rather than an independent sample; recalibration is motivated, not demonstrated.

VII. Only linear TF-IDF models were evaluated; the results do not establish that structured fields cannot help calibrated transformer, temporal or multimodal models.

VIII. Five-fold internal cross-validation from one institution, even with the added repeated-split check, cannot estimate performance under temporal, linguistic or institutional shift; no subgroup fairness or calibration analysis by demographic group was available.

8 Conclusion

Within this authored, role-play-derived university counselling corpus, free-text features provided most of the internally measured predictive signal. Structured demographic and problem-category fields showed no demonstrated incremental benefit for escalation ROC-AUC or four-class macro-F1 once text was included, under both a case-resampling and a fold-assignment robustness check, and a feature-set decomposition attributes what structured-only signal exists mainly to problem categories rather than demographics. At the same time, no tested model or examined threshold reached a sensitivity, and in particular a Critical-class recall, adequate for autonomous safety-critical use. The result supports leaving demographic fields out of this prediction model, without implying they should not be collected for other purposes, and is best read as a methodological feature-ablation study of this corpus rather than as evidence that text-only modelling is ready for real university counselling triage. Resolving the open provenance and case-independence questions in Section 4.1, authentic-text external validation, counsellor-informed threshold and calibration work, and subgroup error analysis are required before this line of work can inform counselling triage.

9 Declarations

Ethics approval and consent to participate. Ethical approval was granted by the University of Zambia Natural and Applied Sciences Research Ethics Committee (NASREC; Institutional Review Board registration number 00006465), under approval number NASREC: 2026-APR-012, dated 4 May 2026. All methods were performed in accordance with the World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants and the applicable guidelines and regulations of the University of Zambia Natural and Applied Sciences Research Ethics Committee. The approved protocol covered the GHQ-12 survey and the counsellor-student role-play activities used in this study. All role-play activities and associated data collection were conducted after ethical approval had been granted. All participants received study information and provided written informed consent before participation.

Clinical trial number: not applicable.

Data availability. The scenario corpus is not publicly available because it contains sensitive mental-health content and potentially identifying contextual information. A de-identified data dictionary, analysis code and aggregate fold results are available from the corresponding author on reasonable request, subject to ethics, institutional and data-governance approval.

Funding. No external funding was received for this work. The research was self-funded.

Competing interests. The author declares no competing financial or non-financial interests.

Author contributions. Danny Leza conceived the study, curated the data, designed and conducted the analyses, interpreted the findings, and drafted and revised the manuscript. The author read and approved the final manuscript.

References

[1] R. P. Auerbach et al., "WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders," Journal of Abnormal Psychology, vol. 127, no. 7, pp. 623–638, 2018, https://doi.org/10.1037/abn0000362.

[2] J. Bantjes, M. J. Kessler, X. Hunt, D. J. Stein, and R. C. Kessler, "Treatment rates and barriers to mental health service utilisation among university students in South Africa," International Journal of Mental Health Systems, vol. 17, art. 38, 2023, https://doi.org/10.1186/s13033-023-00605-7.

[3] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: A systematic review and research framework," Translational Psychiatry, vol. 13, art. 309, 2023, https://doi.org/10.1038/s41398-023-02592-2.

[4] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology, 2018, pp. 25–36, https://doi.org/10.18653/v1/W18-0603.

[5] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead, "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts," in Proc. Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 24–33, https://doi.org/10.18653/v1/W19-3003.

[6] H. Qiu, L. Ma, and Z. Lan, "PsyGUARD: An automated system for suicide detection and risk assessment in psychological counseling," in Proc. 2024 Conf. Empirical Methods in Natural Language Processing, 2024, pp. 4581–4607, https://doi.org/10.18653/v1/2024.emnlp-main.264.

[7] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019, https://doi.org/10.1126/science.aax2342.

[8] H. Suresh and J. V. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in Proc. 1st ACM Conf. Equity and Access in Algorithms, Mechanisms, and Optimization, 2021, art. 17, https://doi.org/10.1145/3465416.3483305.

[9] Republic of Zambia, Data Protection Act, No. 3 of 2021, Government Printer, Lusaka, Zambia, 2021.

[10] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," NeuroImage, vol. 180, pp. 68–77, 2018, https://doi.org/10.1016/j.neuroimage.2017.06.061.

[11] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913–929, 2017, https://doi.org/10.1111/ecog.02881.

[12] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, art. 100804, 2023, https://doi.org/10.1016/j.patter.2023.100804.

[13] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, e078378, 2024, https://doi.org/10.1136/bmj-2023-078378.