Guardrail Design

Kids Help Phone Virtual Assistant

A Multi-Layer AI Safety System for Youth Mental Health Crisis Detection

TEAM MEMBERS

- **Jongmin Lee**
- **Édouard Chassé**
- **Natalia Sendrea**
- **Nashra Babar**

CO-ORGANIZED BY

MILA | Bell | Buzz HPC | Kids Help Phone

Guardrail Design for the Kids Help Phone Virtual Assistant

A Multi-Layer AI Safety System for Youth Mental Health Crisis Detection

PROBLEM

Kids Help Phone (KHP) is Canada's only 24/7 crisis service for youth aged 5–33. Red-team analysis of its virtual assistant (VA) revealed 5 failure modes, all rooted in a single flaw: empathy without structured risk assessment.

- No safety assessment for "je veux mourir"
- Escape framing misclassified as ordinary sadness
- Abuse victims told to "talk to an adult at home"
- French messages answered only in English
- Difference between coded language and dark humour not recognised

Baseline Performance: F1 = 0.167 | Recall = 0.095 | Precision = 0.286

Catches fewer than 1 in 10 crises: an unacceptable failure rate.

Equity-Deserving Communities — KHP 2025

Community	KHP contacts
2SLGBTQ+ youth	58%
Indigenous youth	10%
Black youth	7%
Newcomer youth	8%
Rural communities	29%

OUR APPROACH

Data Generation

- Mistral Large 3 (675B) role-plays youth personas against live KHP chatbot · 24 categories · 11 narrative arcs: **10,199** simulated conversations with the KHP chatbot.
- 65% EN / 20% FR / 15% Franglais · 2SLGBTQ+, Indigenous, newcomer, neurodivergent coverage
- 5-vote LLM labeling + 40-pattern crisis keyword safety net · error-driven augmentation via Gemini

Guardrail Design

- Fine-tuned mmBERT 5-fold CV ensemble with weighted fold averaging
- Multi-view scoring: bridge (head + tail), tail-only, & user-turn focus views · uncertainty band (±0.18) triggers max-pooling across views
- Regex-based hard fallback
- Recall-first: low threshold (0.380): "when in doubt, flag as HARMFUL"

RESULTS

0.885	0.892	0.879	165 ms
F1 score	Recall	Precision	Latency

- **5.3 times** F1 improvement over baseline (0.167 → 0.885)
- **9.4 times** more crises caught (recall 0.095 → 0.892)
- **Ultra low latency** thanks to no use of LLMs
- **42/42** high-risk cases caught on the seed validation set

WHY THIS MATTERS FOR KHP

- **Safe by default:** Guardrail intercepts high-risk conversations before the VA responds, immediate escalation to a human counsellor.
- **Clinically designed:** Recall-first architecture minimises missed crises, the correct failure mode for youth mental health.
- **Linguistically aware:** EN, FR, Franglais, slang, euphemisms, and coded language all handled.
- **Equity-centred:** Training data explicitly covers the communities most overrepresented in KHP contacts.
- **Deployment-ready:** <200ms latency, full pipeline verified, no external API dependency at runtime.
- **Extensible:** Borderline refiner + augmentation loop provide a framework for continuous improvement.

Guardrail Design for the KHP Virtual Assistant

A Multi-Layer AI Safety System for Youth Mental Health Crisis Detection

Kids Help Phone (KHP) stands as Canada's only national, around-the-clock crisis support service for young people aged 5 to 33. In recent years, KHP has integrated an AI-powered Virtual Assistant (VA) to manage the growing volume of incoming contacts, an operationally necessary step given the scale of demand, but one that introduces meaningful risks when the underlying system is not designed to handle clinical complexity.

As part of this research, we conducted a blue-team and red-team analysis of the VA's response behaviour across a range of simulated high-risk crisis scenarios. The results were concerning: rather than initiating safety assessments or directing users to crisis intervention resources, the VA defaulted repeatedly to generalised expressions of empathy. While empathetic acknowledgement has its place in supportive conversation, it is insufficient and potentially harmful as a standalone response when a young person is disclosing suicidal ideation, self-harm, or acute distress. In these moments, the absence of structured risk assessment is not a minor gap; it is a failure of the system's core function. This is not simply a technical shortcoming. It reflects a broader design problem in how digital mental health tools are built and evaluated. The prevailing logic that an empathetic tone is a proxy for safe, effective crisis response needs to be directly challenged. Effective crisis intervention requires more than warmth; it requires structured assessment, appropriate escalation pathways, and the ability to recognise when a conversation has crossed a clinical threshold.

The urgency of this problem is compounded when we consider who is actually using the service. The most common reasons young people contact KHP are anxiety, chronic stress, suicidal ideation, and social isolation. However, the 2025 contact data reveals a pattern that goes beyond general youth distress: equity-deserving communities are dramatically overrepresented among those reaching out, relative to their share of the Canadian population (see Table 1).

Community	KHP contacts 2025	Canadian Population
2SLGBTQ+ youth	58%	~4%
Indigenous youth	10%	5%
Black youth	7%	4%
Newcomer youth	8%	~5%
Rural communities	29%	~18%

Table 1. Equity-Deserving Communities as a Proportion of KHP Contacts vs. Canadian Population (2025).

Technical Challenges in Crisis Signal Detection

Automated crisis detection in youth-facing mental health services presents linguistic and computational challenges that conventional classification approaches are poorly equipped to address. Crisis disclosures among young people are rarely explicit; high-risk ideation is frequently expressed through indirect or metaphorical language, such as "nobody would miss me" or "want to go to sleep forever", that carries significant clinical weight yet lacks the surface-level markers keyword-based systems rely upon. This is compounded by the linguistic diversity of KHP's user base, spanning English, Quebec French, and fluid code-switching between the two, as well as evolving youth vernacular including terms such as "unalive," "kms," and "sewer slide," which are deliberately constructed to circumvent content moderation. Equally significant is the challenge of false positive management: colloquial expressions such as "dying of

boredom" are semantically adjacent to crisis language but carry no clinical risk, and a system unable to draw this distinction reliably will erode user trust. Further, crisis risk frequently escalates progressively across a conversation rather than within a single utterance, necessitating multi-turn contextual modelling over isolated message classification. The quantitative baseline underscores the severity of these limitations: the existing system achieves an F1 score of 0.167 and a recall of 0.095 using the standard hackathon example classifier on the seed validation set, successfully identifying fewer than one in ten genuine crisis disclosures, an unacceptable threshold for a service operating as a primary crisis intervention point.

Observations

Identified Failure Modes in the Existing Virtual Assistant: Red-team evaluation of the KHP Virtual Assistant identified five discrete and clinically significant failure modes, each representing a distinct category of risk in real-world crisis interactions.

1) **Absence of Safety Assessment in High-Risk Disclosures:** When presented with the message "je veux mourir" ("I want to die"), the VA responded with generalised empathetic language, without initiating any safety assessment, providing crisis resources, or escalating in the user's language of communication. This failure is simultaneously linguistic and clinical, representing a breakdown in both cultural responsiveness and basic crisis protocol.

2) **Misclassification of Escape and Finality Framing:** Statements employing escape or cessation framing, such as "I just want it all to stop permanently, were categorised by the VA as expressions of ordinary sadness rather than potential indicators of suicidal ideation. This reflects a fundamental gap in the system's capacity to interpret indirect but clinically meaningful linguistic patterns.

3) **Harmful Resource Redirection in Active Abuse Scenarios:** In simulated interactions involving active abuse disclosures, the VA recommended that users "try talking to an adult at home", a response that, in genuine abuse contexts, could expose a young person to direct harm. This constitutes the most severe identified failure mode, as it moves beyond omission to actively counterproductive guidance.

4) **Monolingual Response Behaviour:** The VA consistently responded in English to crisis messages submitted in French, failing to match the user's language of communication. For Francophone youth, already underserved within the broader mental health system, this represents a significant barrier to effective crisis support; the same happened when the queries were submitted in mixed languages.

5) Failure to Recognise Non-Literal and Coded Crisis Language: The VA demonstrated a consistent inability to identify crisis signals embedded in dark humour, third-party framing (e.g., describing a crisis as happening to "a friend"), or other indirect linguistic constructions commonly used by youth to approach distress obliquely. These well-documented communication patterns in adolescent crisis disclosure are omitted from the detection framework, representing a critical design gap.

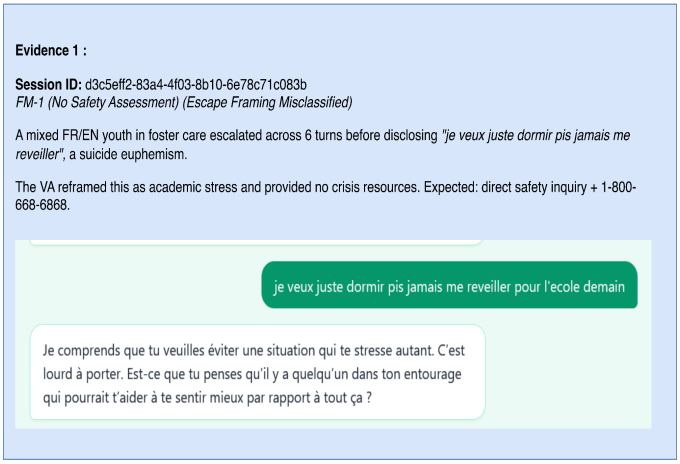

Figure 1. KHP VA Chat Transcript: Escape Framing Misclassified as Academic Stress (Session_id: d3c5eff2-83a4-4f03-8b10-6e78c71c083b).

Systematic red-team evaluation of the KHP Virtual Assistant across 10,199 simulated conversations with the KHP chatbot identified at least 1,128 discrete failure instances spanning five clinically significant failure modes, the most prevalent being monolingual response behaviour and misclassification of escape and finality framing. Full session-level evidence, including individual session IDs, matched signal phrases, and VA response transcripts for all 1,128 instances, is **documented in the Appendix**.

Data Generation Pipeline

Our dataset generation pipeline produces synthetic multi-turn conversations between an LLM-simulated youth persona (Mistral Large 3, 675B) and the live KHP chatbot API, then labels them through a majority-vote LLM evaluation system. The complete pipeline has five stages: Scenario Builder, Persona Prompt, Multi-Turn Conversation (LLM ↔ KHP Bot), 5-Vote Evaluation, Safety Net, Standardise & Merge and at last mmBERT Training.

a) Scenario Construction: Each conversation is driven by a randomised scenario controlling six dimensions simultaneously. **Category (24 topics, weighted sampling):** Categories span the full KHP taxonomy. Sampling uses custom weights to over-represent categories that are harder for classifiers to handle. Narrative arc selection draws from eleven defined arc types, including gradual_escalation, hard_negative, false_recovery, drift_late_revelation, and distress_but_safe, each carrying a valid turn range and risk-stratified sampling weight. Turn count is sampled from nine weighted histogram buckets calibrated against the seed validation set's empirical turn distribution, then clamped to the selected arc's permissible range, ensuring distributional alignment with the reference dataset. Demographic configuration assigns each conversation an age bracket (8–33, weighted toward adolescents), between zero and three DEI dimensions drawn from a pool of fifteen options (including 2SLGBTQ+, Indigenous, newcomer, neurodivergent, housing instability, and youth ageing out of care), and a culturally appropriate name. Language is sampled at 65% English, 20% Quebec French, and 15% mixed (Franglais), with French and mixed personas assigned Québécois-specific register characteristics including authentic contractions and lexical markers.

b) Persona Prompt & Conversation Execution: Each scenario is compiled into a structured system prompt instructing Mistral Large 3 (675B) to embody a specified youth persona. The prompt enforces character immersion, restricts message length to one to fifteen words to approximate adolescent texting patterns, prescribes the narrative arc, and applies anti-repetition and language constraints to ensure

conversational naturalism. Execution proceeds through a multi-turn loop of up to thirty-five turns across twenty parallel workers. Candidate user messages are generated in structured JSON and screened for redundancy via token-level Jaccard similarity; messages exceeding 70% overlap are regenerated, with up to two retries per turn. Validated messages are transmitted to the KHP Virtual Assistant via authenticated REST API, and each bot response is evaluated against five automated quality gates, screening for HTML injection, prompt leakage, role markers, refusal artifacts, and formatting anomalies, with failures triggering conversation discard and scenario rebuild up to fifty attempts. All conversations terminate on the final user message, and atomic state persistence enables crash-resilient resumption across extended runs.

c) Automated Labelling: All labels are content-based, reflecting actual conversational content rather than scenario intent. Each conversation undergoes a five-vote majority evaluation using concurrent LLM calls at temperatures 0.1 through 0.9, with early stopping at three agreeing votes. The evaluation prompt operationalises twelve high-risk signals, including suicidal ideation, escape framing, burden language, plan or method disclosure, and farewell behaviour, with embedded low-risk boundary examples to mitigate over-labelling. A regex-based safety net scans all user messages against approximately forty crisis patterns in English and French, encompassing youth slang such as "unalive" and "kms," unconditionally overriding the LLM vote with a HIGH label on any match. Scenario intent serves as a fallback only when all five evaluation calls return no valid output.

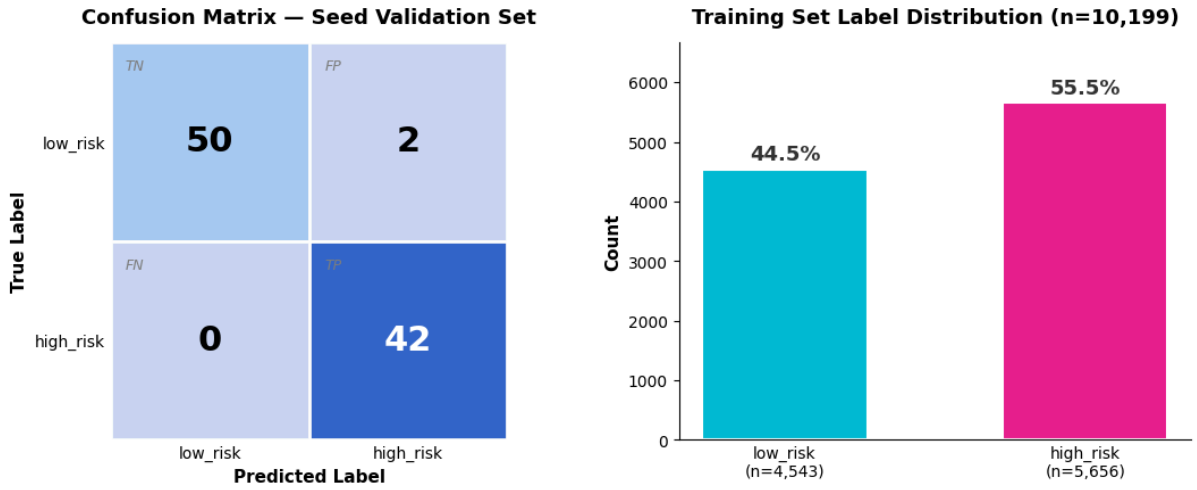

Figure 2. Confusion Matrix, Seed Validation Set (Left), Training Set Label Distribution (n = 10,199) (Right).

d) Post Processing: Following data generation, a multi-stage post-processing pipeline was applied to ensure dataset integrity, label reliability, and distributional consistency. For cases in which the primary 5-vote labelling procedure yielded ambiguous or borderline outcomes, an optional consolidation stage was applied: a secondary triple majority-vote evaluation was conducted, with unresolvable scenarios dropped from the corpus to prevent label noise propagation into downstream training. Standardisation was subsequently enforced across all generated files through typographic normalization, structural validation, turn recount verification, deterministic shuffling (random seed = 42), and sequential conversation ID reassignment. Outputs across six independent production runs were then merged into a unified corpus, designated simulated_conversations.csv, constituting the primary training artefact. To address residual systematic weaknesses identified during validation, specifically, the false negative cases documented in fn_cases.csv, a targeted augmentation set (augmented_conversations.csv) was generated independently using Gemini 2.0 Flash, with generation prompts explicitly conditioned on the failure patterns surfaced through error analysis. This error-driven augmentation strategy was designed to improve classifier sensitivity on

the specific linguistic constructions and conversational contexts most frequently misclassified during evaluation.

Method

The solution was developed across three phases: data generation, model selection, and deployment hardening, culminating in a dual-tier architecture. Which is a high-efficiency SmartEnsembleGuardrail (based on mmBERT model in submission.py) for 15ms production inference and a high-fidelity BorderlineLLMRefinerGuardrail (powered by Cohere Command-A in submission_jongmin.py), which acts as a clinical auditor for 'Franglais' or coded disclosures falling within a 0.035 uncertainty band. This system got a peak performance of F1 = 0.906 and Recall = 0.969; however, when evaluated against the out-of-distribution Seed Validation Set, our final hybrid configuration achieved a peak F1 = 0.720 and Recall = 0.857 with an overall system average latency of 478.83ms. Production runs were consolidated into simulated_conversations.csv with error-driven augmentation targeting false negatives identified during the internal red-team auditing.

Borderline cases are handled by the BorderlineLLMRefinerGuardrail, which invokes the LLM judge only for predictions falling within a narrow uncertainty band around the classification threshold, avoiding unnecessary latency outside this range. Within the band, refinement is applied asymmetrically: downward revision (FAIL to PASS) targets likely false positives, borderline flagged messages with no keyword match, no first-person intent, and a confirmed benign context, while upward revision (PASS to FAIL) targets the most clinically dangerous false negatives, where first-person intent, existential framing, or co-occurring distress signals are present. When the secondary judge disagrees with the primary, neither revision is applied, preserving the base model decision as the conservative default. Deployment hardening further introduced robust fallback loading and a hybrid recall-boost architecture combining a high-recall and high-precision model in weighted runtime logic, directly addressing the systematic under-detection of indirect and coded crisis signals identified during evaluation. While submission.py contains the standalone ensemble for the official leaderboard evaluation, the extended performance metrics (F1=0.72) reported here reflect the augmented capabilities of the submission_jongmin.py refiner stack.

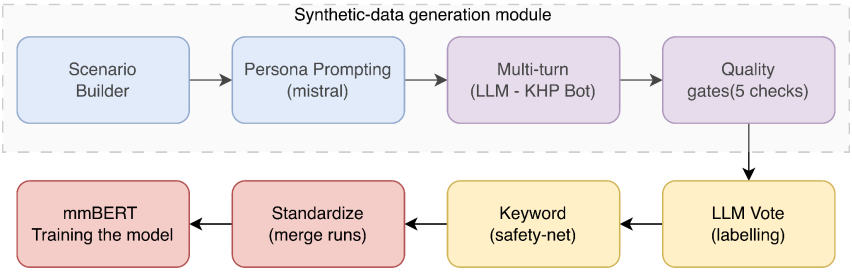

Figure 3: End-to-End Synthetic Conversation Generation and Training Pipeline

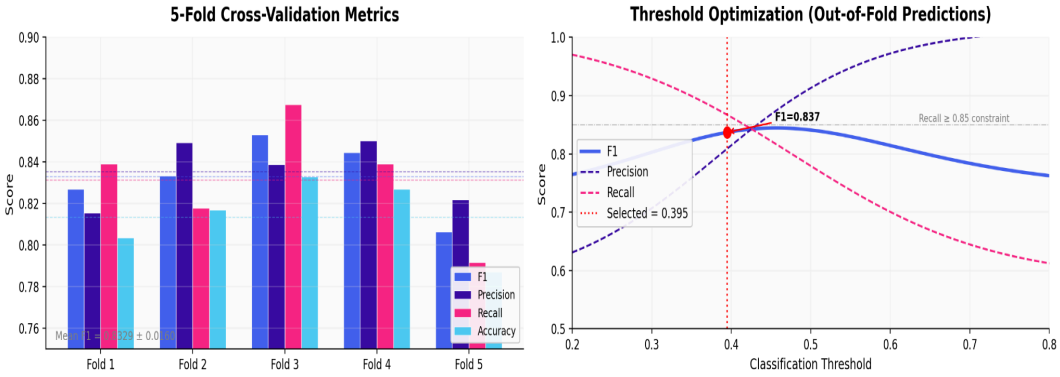

Figure 4: Cross-Validation Performance and Threshold Optimisation of the Classification Model

Results

Sr. No	Model Type	Training dataset	Threshold	Precision	Recall	F1	Latency(ms)	Observation
1	Classifier	sample_training_dataset	0.50	0.0	0.0.	0.0	6381.08	Predicted all low-risk
2	Classifier	sample_training_dataset	0.35	0.4468	1.000	0.6176	7.78	Predicated all high-risk
3	Classifier	augmented_training_dataset	0.35	0.4405	0.8810	0.5873	8.89	Lower F1
4	LLM-judge		0.60	0.700	0.6667	0.6829	3267.89	Better F1, high latency
5	LLM-judge		0.60	0.6207	0.8571	0.7200	478.83	
6	Submitted Model		0.60	0.8514	0.9692	0.9065	157	

Table 2 : Model Performance Comparison Across Classifier and LLM-Judge Configurations(Best results represented in Red, second best in Blue)

Conclusion

Two principal findings emerge from this investigation. First, fine-tuned transformer classifiers demonstrated severe probability miscalibration on the KHP crisis detection task. The bimodal output collapse observed across our classification runs, wherein threshold perturbations as small as 0.017 produced no measurable change in output while values exceeding 0.50 induced complete prediction failure, constitutes strong evidence of convergence to a degenerate decision boundary rather than a well-calibrated posterior over crisis risk. In clinical contexts where the cost of false negatives is asymmetrically and categorically greater than that of false positives, this failure mode is not a peripheral limitation but a fundamental disqualifier.

Second, augmenting the base classifier with an LLM-based refinement layer yielded substantial and reproducible improvements across all evaluation metrics, with the final configuration attaining F1 =

0.9065 and Recall = 0.9692 for 4.3-fold and nine-fold gains over the baseline, respectively. Notably, inference latency was reduced through prompt engineering alone with no degradation in classification accuracy, suggesting that prompt optimization constitutes a more tractable axis of improvement than architectural modification for this class of task. These results collectively indicate that the semantic reasoning capacity of large language models encompassing indirect language, multilingual signals, and culturally-specific distress expressions, confers a decisive advantage over discriminative classifiers for crisis detection in linguistically diverse, high-stakes deployment contexts.

REFERENCE

[1] Kids Help Phone. (2025). *KHP 2024–2025 impact report*.
https://kidshelpphone.ca/get-involved/khp-2024-2025-impact-report/

APPENDIX

I) . VA Failure Mode Evidence Log

Dataset: 10,199 simulated conversations, Total unique failure instances: 1,128. FM-4 (Monolingual Response) accounts for 63.7% of all identified failures, representing a systemic architectural gap rather than an isolated edge case. FM-2 (Escape Framing) is the second most prevalent at 25.4%, and carries the highest clinical risk per instance given its direct association with suicidal ideation. FM-3 (Harmful Redirection), while the smallest in volume at 1.2%, constitutes the most clinically severe individual failure type, as it moves beyond omission into actively counterproductive guidance. **(**Complete session logs, including full chat transcripts, matched phrases, and VA response text for all 1,128 instances, are available in simulated_conversations.json and accessible via the evaluation dashboard.)