Preparing AI Chatbots to Respond to Patient Distress and Suicidality in High-Risk Healthcare Settings

**Supplementary Results Information File**

## A. Additional Technical Details of Safety Router

**(i) Fine-tuning Method.** The Safety Router was developed using OpenAI's managed supervised fine-turning workflow using a data set of simulated user prompt/response pairs. We evaluated fine-tuned variants based on both gpt-4o-2024-08-06 and gpt-4o-mini-2024-07-18. This was not a LoRA-based implementation and was not a prompt-only classifier operating on an unmodified frozen base model. Because fine-tuning was performed through OpenAI's managed service, the lower-level training implementation was abstracted by the platform and not directly exposed to the authors. We will therefore describe the method as supervised fine-tuning through OpenAI's fine-tuning service rather than as a custom full-parameter or LoRA implementation.

For fine-tuning hyperparameters, we began with the OpenAI service's recommended default starting configuration available at the time. The platform resolved the batch size to 1 for these training jobs, with learning-rate multipliers of 1.8 for GPT-4o-mini runs and 2.0 for GPT-4o runs. The only hyperparameter we intentionally varied manually was n_epochs. Candidate models were trained with epoch counts of 5, 3, and 2 to assess the tradeoff between classification performance and overfitting risk. Because the labeled dataset was small, we specifically tested lower-epoch configurations after observing that some higher-epoch runs achieved very low training loss without corresponding improvement in validation behavior. The final selected Safety Router used gpt-4o-mini-2024-07-18, 2 epochs, batch size 1, and learning-rate multiplier 1.8.

**(ii) Labeling protocol.** The initial labeling dataset consisted of 200 simulated user prompts informed by early co-design work with peer recovery coaches that included statements with varying degrees of risk. To reduce burden on expert reviewers and encourage more focused efforts on a comprehensive review of the prompt/response pairs, the dataset was divided into smaller subsets of 50 with no overlap between reviewers. As such, item-level inter-rater reliability could not be calculated which is a limitation of our protocol. One set was overwhelmingly weighted by one reviewer as Tier 2, suggesting that more clarity was needed to determine if an emergency response was needed. We determined that including this label set without adjustment could over-weight the model toward triage behavior even when a message may warrant an emergency response. Specifically, if over-applied, this behavior could create a recursive or obstructive conversation pattern in which the chatbot repeatedly asks the user to "tell me more" rather than routing to the appropriate conversational flow or, critically, to emergency messaging when warranted. For this reason, initial fine-tuning experiments were conducted using the full labeled dataset excluding this highly triage-weighted submission. We then broadened testing by training additional models on selected subsets of the data to evaluate whether smaller, more internally consistent training sets improved validation behavior.

**iii) Safety Router prompt.** The full Safety Router prompt is shown below. The underlying model was fine-tuned to perform tier-based classification. TRUE corresponds to Tier 1, FALSE corresponds to Tier 3, and TRIAGE corresponds to Tier 2.

# Role and Objective

Acts as a safety screening layer for a chatbot assisting patients in a 12-week opioid treatment program. Evaluates every user message and determines if emergency intervention is needed.

# Instructions

- For each user message, select ONE outcome: TRUE, FALSE, or TRIAGE.

- Judge only by the user's message content.

- If ambiguous, default to TRIAGE.

- Return only: TRUE, FALSE, or TRIAGE. No other outputs.

## Outcome Definitions

- **TRUE**: Message shows explicit, immediate risk (e.g., suicidal/homicidal thoughts, intent, or specific plans).

- **FALSE**: Message has no signs meeting TRUE or TRIAGE. No risk or need for clarification.

- **TRIAGE**: Message has unclear, ambiguous, or possible signs of distress (incl. passive mentions, depression, hopelessness, possible self-harm/harm to others). Use when more context is needed or if unsure about risk.

Decisions must be based solely on the content of the user message, ensuring safety and clarity in every response. The model should err on the side of caution when evaluating ambiguous inputs and prioritize issuing a TRIAGE response when potential risks are hinted at but not explicitly clear. This is especially true if themes of depression, hopelessness, or self harm or harm to others are expressed--even passively voiced. The only permissible outputs for this model are TRUE, FALSE, or TRIAGE.

**(iv) Train/validation/test sets.** The training set contained 146 response pairs and the validation set contained 25 response pairs, corresponding to an approximately 85/15 train/validation split. No larger independent held-out test set was available. The validation set was used to compare candidate fine-tuned models against the corresponding base models across the three outputs: Tier 1, Tier 2, and Tier 3. The base GPT-4o and GPT-4o-mini models each correctly classified 20/25 validation response pairs. The final selected GPT-4o-mini fine-tuned model correctly classified 23/25 response pairs, improving observed validation accuracy from 80% to 92% on this small validation set.

**(v) To guard against model drift**: We conducted regression testing before deployment of GPT5.0 with a focused evaluation of a small test set of 33 simulated patient messages spanning a range of message types and risk levels. The same 4 expert reviewers evaluated all 33 messages in a blinded forced-choice task, selecting their preferred response from three candidate models (GPT-5.0, GPT-5.1, GPT-5.2), with instructions to prioritize safety and accuracy and include justification for their choice. Reviewers reached majority agreement on 30 of 33 items (91%),

with a clear overall preference for GPT-5.0 across message themes. GPT-5.0 was the preferred model across nearly all message themes, including those related to safety-critical content such as Emergency Handling and Non-Stigmatizing Language, with GPT-5.1 selected least frequently overall.

**B. Chatbot Safety Disclaimer (developed by our clinical team)**: "The chatbot is NOT a human. The chatbot is NOT an emergency resource. If you are having thoughts of suicide, self-harm or of harming others, please call 988 or text HOME to 741741 to receive support from a trained crisis counselor through a nationwide hotline. If this is an acute emergency requiring immediate medical attention or emergency personnel, please instead call 911 or go to the nearest emergency room." Users were required to acknowledge that they read and understood the safety disclaimer before proceeding ("please reply '1' to indicate that you read and understand the information above and wish to proceed").

**C. Preliminary Data on Safety Validation**

Of the n=55 participants enrolled in the parent study who interacted with Suzy, staff flagged 11 transcripts during the 12-week study that included 13 potentially risky user generated messages. Of the 12/13 messages, Suzy correctly responded with the risk disclaimer statement. Five of the 11 transcripts were deemed to contain messages suggesting suicide risk (e.g., "I hate my life") and prompted a staff call to administer the C-SSRS, whereas six were determined to need to no further risk assessment (e.g., "I have a lot of stress right now"). Of the 5 that indicated potential risk, the C-SSRS revealed low risk and participants were provided with the written mental health referral resources.

In a post hoc audit of 1,130 user messages with router trace data (90.5% of total user messages received), the Safety Router classified 5.13% as Tier 1 (safety risk), 3.54% as Tier 2 (triage/ambiguous), and 91.33% as Tier 3 (no risk). Of the staff-flagged 13 user messages, the router correctly identified 12, yielding 92.3% accuracy, 92.3% sensitivity, 92.3% specificity, 12.2% positive predictive value, and 99.9% negative predictive value. The one missed staff-flagged row did not prompt C-SSRS follow-up; all 7 rows that did were classified Tier 1 (n=6) or Tier 2 (n=1).