Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Quantitative Behavioral Coding in a Partially Nested Randomized Control Trial with Racially Diverse Older Adults
Principal Investigator:	Yung-Wei Lin, Ph.D., Counselor Education, College of Education, University of Iowa (co-PI) Raghuraman Mudumbai, Ph.D., Electrical and Computer Engineering, College of Engineering, University of Iowa (co-PI)
Prepared by (IIAI):	Avinash Reddy Mudireddy
Other investigators:	Weiyu Xu, Ph.D., Electrical and Computer Engineering, University of Iowa Laura Shannonhouse, Ph.D., Counseling and Psychological Services, Georgia State University (DHHS grant PI) Erika LeBlanc, Ph.D., Counselor Education, University of Mississippi (DHHS grant co-PI) Matthew Fullen, Ph.D., Counselor Education, Virginia Tech (DHHS grant co-PI) Yusen Zhai, Ph.D., Counseling, University of Alabama at Birmingham
Date:	10/08/2026

Were specific aims fulfilled:	Partially. A transcript-based prototype was developed and evaluated for eight prioritized behavioral factors. Expansion to the full manual and the archive of more than 8,000 calls was not completed.
Readiness for extramural proposal?	No
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	The prototype performed better than a lexical baseline for selected factors but has not demonstrated reliable detection of rare suicide-related language or sufficient accuracy for unsupervised coding at archive scale. No paper has been submitted. The findings and limitations are detailed below.

Brief summary of accomplished results:

This pilot evaluated whether AI-based analysis of phone-call transcripts could reproduce human annotations of behavioral risk and protective factors in older adults. The factors included sadness, loneliness, religion or spirituality, social support, and suicide-related language.

The original 2023 proposal envisioned automated coding of audio recordings, beginning with 14 items, expanding to approximately 70, and ultimately processing more than 8,000 calls. The completed pilot had a narrower scope: transcript-based classification of eight investigator-prioritized factors.

The study included 236 phone calls from 118 older adults. Each person has two calls, dose 1 and dose 7. Two human coders annotated every call independently. Transcript text was available for 233 calls. Each of the eight factors was classified as expressed or not expressed based on the older adult's statements.

We compared several approaches using a participant-level held-out test set. The best-performing approach prompted an instruction-tuned large language model (LLM) to return short classification codes, achieving a macro F1 of 0.656 across the eight factors. A TF-IDF logistic-regression baseline, which scores a call from how often its words appear, achieved 0.557, while a later LLM approach incorporating investigator-provided coding rules achieved 0.567. Religion or spirituality and significant life events showed comparatively strong results. Suicide-related factors were too rare to establish reliable detection performance.

Research report:

Aims (provided by PI):

The following material is reproduced from the investigators' 2023 proposal and describes the original research plan. The subsequent Modified aims section distinguishes that plan from the work completed in this pilot.

Available Data: The data are from a research project supported by a DHHS grant. In this DHHS grant project, mental health care/treatment was provided to the older adults (n > 700) through phone calls across eight weeks. In total, more than 8,000 audio recordings of the care/treatment phone calls were collected and transcribed into texts. Also, the behavioral coding team has already developed a rigorous quantitative coding protocol with two sections: (1) the aging adult section aiming to identify risk/protective factors and measure emotional change/improvement of the older adults, (2) the caregiver section aiming to evaluate caregivers' skills and treatment effectiveness. Totally, more than 140 items (approximately 70 items for each section) are in the coding manual, and each item is coded with either dichotomous scale (yes/no) or Likert scales. Twenty-nine recordings have been coded with the aging adult section (each recording was coded by two independent coders) for coding training to optimize the coding protocol/manual and achieve acceptable inter-coder agreement (> 0.7). Thus, the available data include: (1) the 8,000 audio recordings and corresponding transcribed texts, and (2) the coding results on the 29 recordings (approximately 150 recordings will be coded by the behavioral coding team by the end of Fall 2023).

Aims & AI/ML Relevance: (1) We plan to first develop a prototype of AI coding with a subset of data. The behavioral coding team has already identified a subset (14 items with high intercoder consistency) from the aging adult section in the coding protocol for the AI/ML team to start building the prototype. The coding results by the coding team will be used for verifying the AI coding fidelity. (2) We then will expand the prototype to the full coding protocol (70 items). (3) We hope to optimize the AI coding model and use it to code all the rest of the audio recordings/transcribed texts. Please see the conceptual description of the input (X) and output (Y) variables in the next section.

Use of AI/ML in Research: We propose to develop a prototype of a machine learning model that can automate the coding protocol. Specifically, we propose to build a deep neural network classifier for each of the 14 items (variables) from the previously mentioned subset of the aging adult section. The input to the model is an audio recording (potentially also a transcript) and the output is a predicted label corresponding to a coding result from the protocol. A preliminary design for the architecture of this classifier is as follows. The input audio will be pre-processed into a standard sampling rate and a single channel. A Mel Spectrogram will be generated which will be processed through several convolutional

layers followed by a pooling layer. The output of these layers will be a small number of features (from the pooling layer) for each Mel frequency band. These features will be processed through several neural network layers typical for standard classifiers (e.g. fully connected layers with ReLU activations), ending with a soft max layer from which a label prediction is generated. We will need to experiment with a couple of possible loss functions for Likert scale coded labels and of course with hyperparameter tuning etc. Our hope is that the ML model predictions will achieve a level of accuracy comparable to the consistency between human coders. The training data for the ML model consists of the 150 recordings for which manual coding labels are available. A fraction of these 150 recordings will need to be withheld for testing. We will explore methods for data augmentation to make maximum use of this manually coded dataset. Some possible augmentations are: linear equalization and filtering of the recordings, adding background noise, echoes and other typical audio effects. In addition, we will continue optimizing this ML model in hope to code all the >8,000 audio recordings/transcribed texts, and we hope the finalized AI coding program will expedite and enhance the reliability of the training and treatment evaluation process. The development of this ML technique can also further our research to other areas and populations, such as mental health care/treatment to general adults, adolescents, and young children.

Current Support: The available data in this proposed project are from a large-scale study supported by a 1.3-million-dollar grant (Department of Health and Human Services, grant # 90INNU0021-01-00).

Modified aims:

The completed pilot differed from the 2023 proposal in three areas: input modality, coding scope, and availability of labeled data.

First, the implemented models analyzed transcripts rather than audio recordings; the proposed spectrogram-based audio classifier was not developed. A preliminary 2024 text-only exercise classified separate written passages for suicidal intent using word lists, sentiment and emotion tools, and an earlier language model (Mistral/Mixtral). That exploratory exercise was discontinued and is not part of the reported evaluation.

Second, the investigators prioritized eight items from the older-adult coding manual rather than the 14 originally proposed audio items. On July 17, 2026, they confirmed these eight factors and a binary output format (expressed/not expressed). The previous three-category scheme additionally recorded whether a volunteer responded to what the older adult expressed. That response distinction was excluded from the revised target. One exploratory evaluation of 20 factors is reported separately; it is not the primary analysis.

Third, the available labeled dataset comprised 236 independently double-coded calls, dose 1 and dose 7 for each of 118 participants, rather than the anticipated 150 recordings or the complete archive of more than 8,000 calls. The parent study delivered care by phone over eight weeks; the modeling files identify these two calls as dose 1 and dose 7.

Modified aim 1. Develop a transcript-based prototype for the eight priority factors and compare its predictions with agreed human labels and a lexical baseline. Status: completed.

Modified aim 2. Evaluate whether the prototype remains effective when expanded from 8 to 20 factors. A single exploratory comparison showed only a small improvement over the lexical baseline. Following the investigators' decision to retain eight factors, expansion toward 70 items was discontinued.

Modified aim 3. Apply a validated model to the remaining call archive. Status: not initiated. Performance on rare factors, particularly suicide-related content, was insufficient to justify automated coding of previously unlabeled calls.

Data:

Dataset unit and sample. Each observation is one phone call with an older adult. The dataset contains 236 calls from 118 participants, and every participant has exactly two calls: dose 1 and dose 7.

Human annotations. Every call was independently coded by two trained annotators, assigned in pairs to subsets of calls. No individual coder reviewed the complete dataset. When both coders agreed on a factor, that consensus label served as the evaluation reference. Disputed factor-call pairs were excluded from the primary score but retained in the dataset.

Transcript availability. Transcripts were available for 233 of 236 calls; three were unavailable because the text was missing or its link was inaccessible. Transcript lengths ranged from approximately 500 to 14,000 words, with a typical length of approximately 3,100 words.

Data partitioning. Participants, rather than individual calls, were assigned to mutually exclusive development, validation, and test subsets to prevent leakage across calls from the same person. The development subset comprised 164 calls from 82 participants (162 transcripts); the validation subset, 34 calls from 17 participants (all transcribed); and the held-out test subset, 38 calls from 19 participants (37 transcripts). These corresponded to approximately 70%, 15%, and 15% of participants, respectively. All reported final test metrics use the held-out subset.

Target factors. The eight investigator-selected factors were coded as expressed when supported by the older adult's statements and not expressed otherwise:

Hints of suicide. Indirect statements suggesting a wish not to be here, such as not knowing how to go on, wanting to go to sleep and not wake up, or saying things would be better if one were not here, without the direct wording used for the next factor.

Direct talk of suicide. Explicit references to wanting to die, wanting to kill oneself, or acknowledging suicidal thoughts when asked.

Hopeless or helpless. Statements reflecting despair, limited perceived control, difficulty seeking help, passivity, or a negative outlook.

Sadness. A low, passing feeling tied to loss, rejection, pain, or disappointment. The coding rule describes this as temporary or transitory sadness.

Religion or spirituality. References to faith, prayer, religious practices or texts, meditation, spirituality, gratitude to God, nature in a spiritual context, or an absence of religious belief.

Social support. References to past, present, or anticipated support from family, partners, friends, or staff, including both positive and negative experiences.

Loneliness or isolation. Statements indicating disconnection, lack of belonging, or absence of meaningful relationships or community.

Significant life events. Retrospective references to consequential personal events, including marriage, divorce, childbirth, bereavement, injury, relocation, graduation, or a first car.

Class prevalence among coder-agreed calls varied substantially: direct talk of suicide, 3/236 (1.3%); hints of suicide, 9/236 (3.8%); hopeless or helpless, 10/211 (4.7%); loneliness or isolation, 44/203 (21.7%); sadness, 49/175 (28.0%); religion or spirituality, 94/211 (44.5%); significant life events, 118/196 (60.2%); and social support, 206/213 (96.7%). This imbalance is important: a model can achieve high accuracy on highly prevalent or rare factors by predicting the majority class without detecting the minority class.

Inter-rater reliability was evaluated using raw agreement and Cohen's kappa (κ), which adjusts agreement for chance. As shown in Table 1, high raw agreement does not necessarily indicate strong reliability when one class predominates. Both metrics are therefore reported, particularly to contextualize the rare suicide-related categories.

Table 1. Inter-rater reliability by factor: raw agreement and Cohen's kappa (κ).

Theme	Raw agreement	Kappa
Hints of suicide	1.00	1.00
Direct talk of suicide	1.00	1.00
Hopeless or helpless	0.89	0.39
Sadness	0.74	0.43
Religion or spirituality	0.89	0.79
Social support	0.90	0.33
Loneliness or isolation	0.86	0.63
Significant life events	0.83	0.65

AI/ML Approach:

All evaluated methods used transcript text; none implemented the proposed audio-based architecture. Each method generated factor-level predictions, which were evaluated on the held-out test subset using only factor-call pairs for which the two human coders agreed.

The primary metric was macro F1, which summarizes precision and recall while assigning equal weight to each of the eight factors. Unlike raw accuracy, macro F1 is less dominated by common labels in an imbalanced dataset. This distinction is particularly relevant to suicide-related categories, for which predicting not expressed on nearly every call can yield high accuracy without meaningful detection.

Method 1: TF-IDF logistic regression. This is the simple comparison method. Later methods are worth using only if they beat it. TF-IDF means term frequency–inverse document frequency. Term frequency is how often a word appears in that call. Inverse document frequency up-weights words that are rare across calls and down-weights words that appear in almost every call. The transcript was lowercased and split into single words and two-word phrases. A phrase had to occur in at least two development calls to be kept. Phrases that occurred in more than 95 percent of those calls were dropped. At most 50,000 phrases were retained.

Those phrase weights were the input to a separate logistic regression for each factor. Logistic regression is a standard statistical classifier: it combines the weights into a probability and then chooses expressed or not expressed. Class weights were balanced, so the rare answer counted as much as the common one during fitting. Each factor was fit only on development calls for which the two coders agreed on that

factor. Agreement on the other factors was not required. If a factor had only one answer in the development calls, the regression was replaced by a rule that always predicts that answer. The same TF-IDF setup was also fit to the original three-way labels. The score reported here collapses those labels to expressed versus not expressed. Macro F1 was 0.560 on the check set and 0.557 on the held-out test set.

Method 2: asking a language model, without teaching it our labels. The model was Google's instruction-tuned Gemma 4 E4B. Instruction-tuned means the model had already been trained by its developers to follow a written request. E4B is the model size used here. The weights were stored in a compressed 4-bit form, called NF4 quantization, so the model could run on one graphics processor. None of those weights were updated on this study's calls.

The request cast the model as a coding assistant, listed the eight factors with a short definition of each, and required the answer as JSON: one field per factor and no extra prose. JSON is a rigid text format, so an invalid answer can be detected instead of interpreted loosely. The model was not sampled at random. It returned its single most likely continuation, and that continuation was limited to 512 new tokens. A token is a piece of text, often a word or part of a word. If the answer was not valid JSON, or used a code outside the allowed set, the model was shown the error and asked once to repair it.

The whole request, instructions plus transcript, had to fit in 8,192 tokens. After the instructions and a small safety margin, 7,225 tokens remained for the transcript. On the held-out test set, 24 of 37 transcripts fit in full. The other 13 were shortened by keeping the beginning and the end and omitting the middle.

Three wordings of the request were compared. The best one asked for three short codes. N meant not expressed. ENR meant the older adult expressed the factor and the volunteer did not respond meaningfully. ER meant the older adult expressed it and the volunteer did respond meaningfully. For the score in this report, ENR and ER were both counted as expressed, so the three-way answer could be compared with the later yes-or-no rules. That short-code request scored 0.696 on the check set and 0.656 on the held-out test set.

A second request asked for the full three-way wording instead of N, ENR, or ER. It scored 0.610 on the check set and 0.631 on the test set. A third request added two already-coded development calls as examples, each shortened to at most 1,200 tokens. It scored 0.645 on the check set and 0.472 on the test set. The examples helped on the check set and failed to carry over to new participants, so they were not kept.

Method 3: the same Gemma model, asked to follow the investigators' written rules. This run used the same 4-bit Gemma 4 E4B model, the same single-best-answer decoding, the same 512-token answer limit, and the same one-time repair of an invalid answer. The request changed to match the July 17, 2026 decision. For every factor it included both a description of the theme and the written rule for when to mark it. The only allowed answers were Expressed and NOT expressed. The request told the model not to use the volunteer-response codes from Method 2.

The written rules made the instructions longer, so the transcript budget fell from 7,225 tokens to 6,690. A first pass ran out of graphics-processor memory on some long calls. Those calls were rerun after the processor memory was cleared. If a call still failed, its transcript window was shortened again. On the test set, 27 calls were scored at the 6,690-token budget and 10 at 2,594 tokens. After the reruns, all 34 check-set transcripts and all 37 available test transcripts had an answer. Macro F1 was 0.651 on the check set and 0.567 on the test set. The test number includes the rerun calls. It is lower than Method 2 even though this request matches the coding manual the investigators want to use going forward.

Method 4: Gemma as a reader only, then a separate classifier. The model was not asked to name a factor, and its weights were not updated. The same 4-bit Gemma 4 E4B model read each transcript, and the final-layer vector for every token was averaged into one numeric summary of the call. That summary is called an embedding. A call that fit in 4,096 tokens was summarized in one pass. A longer call was split into chunks of at most 4,096 tokens, with 512 tokens of overlap between chunks; each chunk was summarized, and those summaries were averaged. Of the 233 transcripts, 194 were one pass and 39 were two to four chunks.

The embedding was then the input to one classifier per factor, trained only on calls where the two coders agreed on that factor. The numbers were first rescaled to a common range. Two classifiers were fit. One was logistic regression with balanced class weights, the same kind of classifier as Method 1 but using the embedding instead of TF-IDF. The other was a small neural network with a single hidden layer of 128 units. On the held-out test set the neural network scored 0.568 and the logistic regression scored 0.544. On the check set those scores were 0.526 and 0.558. Table 2 reports the neural network, because it was the better of the two on the test set. Neither beat Method 2, and both stayed close to TF-IDF.

Method 5: a small update of Gemma on our agreed labels. The update is called LoRA, for low-rank adaptation. A thin set of new weights is trained, and the original 4-bit model stays frozen. Those new weights had rank 8 and a scaling of 16, and they were attached to the attention and feed-forward layers. Only the text portion of Gemma was loaded. The model was then trained to write the same short codes as Method 2, N, ENR, and ER.

The training calls were stricter than in Method 1. A development call was used only when the two coders agreed on all eight factors at once. There were 74 such calls. Each example was the transcript followed by the eight correct codes, and the training loss was applied only to those codes, not to the transcript. Memory limits kept the input at 1,024 tokens, with the transcript itself capped at 768 tokens. That is much shorter than the 7,225-token transcript window in Method 2. The reported run used 80 update steps, a learning rate of 0.0002, and updates accumulated over 8 calls before each step.

At test time the updated model had to return the same short-code JSON. It produced a usable answer for 51 of the 71 check-set and test-set calls combined. The held-out macro F1 of 0.429 uses only the test calls that returned a valid answer, so it is not a score on the full test set. The untuned short-code request in Method 2, which saw more of each transcript and did not have to learn from 74 examples, scored 0.656. Further adaptation is not justified until many more fully agreed calls are available.

Exploratory 20-factor evaluation. The Method 2 short-code request and the Method 1 TF-IDF logistic regression were repeated with 20 factors instead of 8, under the same participant split. Short-code Gemma scored 0.639 on the check set and 0.603 on the held-out test set. TF-IDF scored 0.595 on that test set. The advantage over TF-IDF shrank from about a tenth of a point on eight factors to less than a hundredth on twenty. A LoRA update was also attempted for the 20 factors. Only 10 development calls had coder agreement on all 20 factors at once, and the updated model did not return usable answers, so that update is not scored. After the July 17, 2026 decision to keep eight factors, the 20-factor expansion was stopped.

Experimental methods, validation approach:

All methods used the same participant-level partition. Development data supported model construction, validation data supported interim comparisons, and the held-out test set was reserved for final evaluation. Calls from the same participant never appeared in different subsets, and the two annotations for a call remained associated with that call.

The primary reference labels were the coders' agreed annotations; disputed factor-call pairs were not included in primary scoring. For the investigator-guided method, predictions were also scored separately against each coder. Among named coders, the average score on the held-out test set ranged from 0.44 to 0.57, so the result is not explained by one coder alone.

Macro F1 was the primary model-selection metric. Additional measures included accuracy, balanced accuracy, and model-versus-reference Cohen's kappa, with the number of scored calls reported for each factor. Discrepancies between high accuracy and low macro F1 were interpreted in the context of class imbalance rather than as evidence of successful minority-class detection.

Table 1 provides the human inter-rater reliability reference. Cohen's kappa and macro F1 are different statistics, so their numerical values should not be interpreted as a direct ranking of human and AI performance. They are presented together to contextualize the difficulty of individual factors.

Validation scores were not treated as final estimates because the validation subset contained only 34 calls. For example, under investigator-guided prompting, loneliness or isolation achieved a factor-level macro F1 of 0.943 on validation but only 0.408 on the held-out test subset.

Results:

On the held-out test set, direct short-code Gemma prompting achieved the highest eight-factor macro F1 (0.656), ahead of TF-IDF logistic regression (0.557). Investigator-guided prompting achieved 0.567, the Gemma embedding with a neural network achieved 0.568, and the LoRA update achieved 0.429. Table 2 presents the full method comparison, including the exploratory 20-factor result.

The investigator-guided approach declined from 0.651 on validation to 0.567 on the held-out test set. This difference was not attributable to missing model outputs: all 37 available test transcripts were scored. The small test sample limits the precision of performance estimates; nevertheless, 0.567 is the appropriate reported test result rather than the higher validation score.

Table 2. Validation and held-out test macro F1 by method. Scores cover eight factors unless the 20-factor evaluation is specified; evaluation includes only coder-agreed factor-call pairs.

Method	Check set	Final set
Gemma, short codes, then collapsed to expressed or not	0.696	0.656
Gemma, asked to write the full theme wording	0.610	0.631
Gemma, short codes, on 20 themes (set aside)	0.639	0.603
Gemma embedding, then a neural network	0.526	0.568
Gemma, investigators’ written yes-or-no rules	0.651	0.567
TF-IDF logistic regression	0.560	0.557
Gemma LoRA update	0.492	0.429

Table 3. Factor-level held-out results: human inter-rater kappa, model macro F1, and number of scored calls. Factor-call pairs with coder disagreement are excluded from model scoring.

Theme	Human kappa	Word counting	Short-code Gemma	Written-rules Gemma	Calls scored
Hints of suicide	1.00	0.49	0.74	0.48	37
Direct talk of suicide	1.00	0.49	1.00	0.49	37
Hopeless or helpless	0.39	0.49	0.44	0.42	30
Sadness	0.43	0.64	0.40	0.71	24
Religion or spirituality	0.79	0.61	0.78	0.81	33
Social support	0.33	0.49	0.49	0.49	34
Loneliness or isolation	0.63	0.44	0.64	0.41	29
Significant life events	0.65	0.80	0.76	0.72	31

Religion or spirituality showed the clearest positive result. Human inter-rater kappa was 0.79 (Table 1), while short-code and investigator-guided Gemma achieved factor-level macro F1 scores of 0.78 and 0.81, respectively, compared with 0.61 for the lexical baseline (Table 3). These are different metrics and are not directly equivalent.

Significant life events also showed comparatively strong performance across methods (macro F1: 0.72–0.80). The lexical baseline (0.80) outperformed short-code Gemma (0.76) and investigator-guided Gemma (0.72), indicating that the more complex model offered no advantage for this factor.

For sadness, investigator-guided Gemma achieved the strongest factor-level result (0.71), exceeding the lexical baseline (0.64) and short-code Gemma (0.40). Human agreement was relatively limited ($\kappa = 0.43$; raw agreement = 0.74), underscoring the difficulty of this coding decision. The written-rule prompt was more effective for this factor.

Loneliness or isolation exhibited the opposite pattern: short-code Gemma achieved 0.64, compared with 0.44 for the lexical baseline and 0.41 for investigator-guided Gemma. Thus, the written rules improved some factors, such as sadness, while reducing performance on others, motivating factor-specific evaluation of prompt design.

Social support remained unresolved despite high apparent accuracy. Because 206 of 213 coder-agreed calls (96.7%) were labeled expressed, majority-class predictions produced accuracy near 0.97 but factor-level macro F1 of approximately 0.49 across the three methods. Human kappa was 0.33, highlighting the difficulty of distinguishing the uncommon not-expressed cases.

Suicide-related performance should not be interpreted as validated detection, including the short-code score of 1.00 for direct talk of suicide. Across the full agreed dataset, only three calls expressed direct talk of suicide and nine expressed hints of suicide. In the 37-transcript test subset, the agreed labels contained one expressed call for direct talk of suicide and two expressed calls for hints of suicide. The investigator-guided results (0.49 and 0.48, respectively) reflected poor minority-class detection. Hopeless or helpless

was similarly sparse (10 expressed calls overall; human $\kappa = 0.39$), with model macro F1 between 0.42 and 0.49.

Scope of supported conclusions. These findings do not establish that AI outperforms human coders or that the remaining archive can be coded reliably without human review. They support the narrower finding that short-code LLM prompting exceeded the lexical baseline on the eight-factor aggregate, with comparatively promising factor-level results for religion or spirituality, significant life events, and investigator-guided classification of sadness.

Two limitations are central to interpretation. First, the held-out test set contained only 37 transcripts. Second, the human labels were originally assigned using the earlier three-category manual and subsequently collapsed to binary labels; the calls were not recoded under the July 2026 definitions. Consequently, the investigator-guided binary approach was evaluated against labels created before its decision rules were adopted.

Ideas/aims for future extramural project:

The pilot supports a follow-up methodological study focused on annotation quality and model evaluation, rather than immediate autonomous coding of the full archive. Religion or spirituality and significant life events showed promising detection results, as did sadness when investigator-provided rules were included. Reliable identification of rare suicide-related language remains unestablished.

A follow-up project could pursue four aims:

Future aim 1: improve reference labels and rare-class coverage. Adjudicate disagreements on sadness, hopelessness, loneliness, and significant life events through a third independent review, and obtain more positive examples of suicide-related factors. The current totals of three direct-suicide and nine indirect-suicide expressed calls are inadequate for a robust detection claim.

Future aim 2: standardize the coding protocol. Finalize the written factor definitions and decision rules before additional human annotation. Collect new expressed/not-expressed labels directly rather than retrospectively collapsing the previous three-category coding scheme.

Future aim 3: conduct a pre-specified evaluation. Repeat the lexical-baseline and LLM comparisons using adjudicated labels and a participant-level held-out test set. Define the primary endpoint in advance, using macro F1 or recall at a fixed false-positive rate, and report results alongside human kappa. Defer further model adaptation until the number of agreed training examples substantially exceeds that used in the 0.429 experiment.

Future aim 4: evaluate controlled deployment. Only after the third aim meets an investigator-defined performance threshold should a fixed model be applied to uncoded archive calls. The workflow should include human review of every suicide-related flag. This would resume the original archive-scale aim under appropriate validation conditions.

The funding areas identified in the 2023 proposal—social disconnection, suicide risk in later life, and mental health services research—remain relevant. Given the present evidence, a focused methodological proposal addressing measurement reliability, adjudication, and pre-specified comparison against human annotations is more appropriate than an efficacy trial that assumes a validated automated coder.

Publications resulting from project:

The research team has conducted two presentations in national conferences based on this pilot project, including one content session in Association for Assessment and Research in Counseling Annual Conference and one poster session in American Association of Suicidology Annual Conference. The manuscript is currently under development. The research team is also planning to pursue external grant funding for follow-up studies.