CoachLah: A Singlish–English Parallel Corpus of Health Coaching Conversations with Behavior Goal Annotations

Iva Bojic[1], **Mathieu Ravaut**[1], **Stephanie Hilary Xinyi Ma**[1], **Doreen Tan**[2], **Andy Hau Yan Ho**[1], **Andy W. H. Khong**[1]
[1]Nanyang Technological University, Singapore
[2]National University of Singapore, Singapore
{iva.bojic, mathieu.ravaut, hilary.ma, andy.ho, andy.khong}@ntu.edu.sg
doreen.tan@nus.edu.sg

Abstract

Health coaching (HC) aims to promote sustainable behavior change through goal-oriented dialogue, but research in this area is limited by the scarcity of authentic, transcript-based corpora. Existing datasets are small, English-only, and Western-centric, overlooking cultural and linguistic factors that shape real-world HC interactions. We introduce `CoachLah`, the first Singlish–English parallel corpus of HC conversations collected from a randomized controlled trial in Singapore. The dataset comprises 36,852 utterances transcribed from almost 160 hours of recorded HC sessions with 51 clients and 4 professional health coaches. Each dialogue is speaker-labeled, transcribed in Singlish, and aligned with high-quality English translations to preserve linguistic and cultural nuances. All sessions include HC summaries written by health coaches after each HC session, from which behavioral goals were manually annotated. To demonstrate the dataset's utility, we benchmark two downstream tasks: (i) Singlish-to-English translation using fine-tuned open-weight models (e.g., `Gemma-2-9B-it`) with Low-Rank Adaptation, and (ii) behavioral goal extraction from unstructured HC summaries using span-based modeling (e.g., `DeBERTa-v3-base`). Together, these contributions establish the first culturally grounded benchmark for low-resource, goal-oriented dialogue research in HC. Both the code and the dataset are available at: `https://github.com/IvaBojic/CoachLah`.

Keywords: multilingual dialogue modeling, behavioral data annotation, low-resource languages, cultural adaptation

1. Introduction

Cardiovascular disease (CVD) remains a leading cause of mortality in Singapore, accounting for 31% of all deaths in 2023 (Ministry of Health, Singapore, 2024). Its prevalence continues to rise, driven strongly by modifiable metabolic risk factors such as hypertension, hyperlipidemia, and diabetes affecting one in five Singaporeans (National Heart Centre Singapore, 2023). Nearly half of Singapore's disease burden is attributable to lifestyle-related risk factors (Tan et al., 2023), highlighting the urgent need for scalable interventions that promote sustainable behavior change. Health coaching (HC)—a client-centered, goal-oriented process often grounded in motivational interviewing—has been shown to improve health outcomes, reduce cardiovascular risk, and enhance quality of life (Olsen and Nesbitt, 2010; Kivelä et al., 2014).

Despite its effectiveness, HC remains costly and resource-intensive, limiting its scalability (Zhou et al., 2024a; Unick et al., 2024). While AI-driven HC systems powered by large language models (LLMs) have shown promising results (Bojic et al., 2025a; Ong et al., 2024; Mamykina et al., 2024), progress is constrained by the scarcity of publicly available transcript-based HC datasets. The existing corpora are small, predominantly English-only, and reflect Western communication norms, overlooking cultural and linguistic factors that shape coaching interactions in non-English-speaking communities (Zhou et al., 2024b; Xu et al., 2025). This gap underscores the need for realistic and culturally grounded HC dialogues that can serve as benchmarks for tasks such as HC session summarization, dialogue modeling, and behavior goal extraction.

Understanding Singlish HC conversations adds another layer of complexity for NLP. Beyond conversational English disfluencies such as fillers and repetitions (Shriberg, 2001; Zhang, 2020), Singapore-English (Singlish) further introduces elliptical structures with frequent omission of verbs, subjects, and auxiliaries (Platt, 1993; Leimgruber, 2011). It is also heavily code-switched, incorporating Malay, Hokkien, Cantonese, and Tamil terms (Gupta, 1994; Lim, 2004). These features make utterances culturally rich yet challenging for models to interpret, affecting tasks such as intent recognition and goal tracking. Figure 1 illustrates typical phenomena that complicate automatic processing. Combined with the absence of large, publicly available corpora, these factors position Singlish as a low-resource variety whose study is essential for building robust, culturally competent HC systems (Liu et al., 2022).

In this paper, we introduce `CoachLah`, a large-scale Singlish–English parallel corpus of HC conversations. The dataset comprises nearly 160 hours of speaker-labeled transcripts from 51 clients coached by four HC professionals, parallel English translations for every conversational turn, and sessions enriched with detailed HC summaries and structured behavior goal annotations.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 35–49
11–16 May 2026.

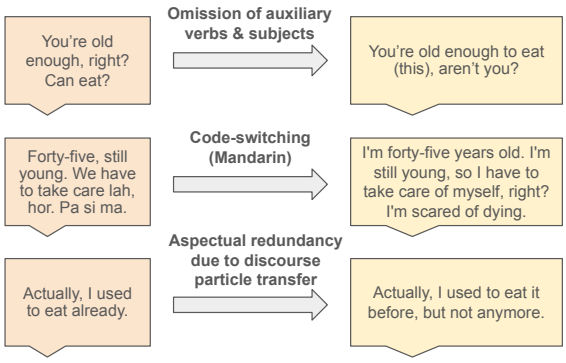

Figure 1: Examples of challenges in Singlish-to-English translation, including *ellipsis*, *particles*, *code-switching*, and *aspectual redundancy*.

To demonstrate its utility, we present two downstream benchmarks: (i) *Singlish-to-English translation*, establishing baseline results for conversational normalization and improving accessibility of code-switched dialogue, and (ii) *behavior goal extraction*, enabling structured tracking and longitudinal analysis of client goals across coaching sessions. These contributions provide the first reproducible benchmark for HC research in a low-resource, code-switched setting, paving the way for culturally adaptive and data-driven HC systems.

2. Related Work

Linguistic and Cultural Diversity. Most conversational corpora are English-only and Western-centric, such as HOPE (Malhotra et al., 2022), MEMO (Srivastava et al., 2022), and MentalChat-16K (Xu et al., 2025)—these corpora overlook the indirect, face-saving, and high-context communication styles that characterize Asian settings (Gudykunst, 2003; Ting-Toomey, 2017). Cultural differences have implications for HC practice, influencing perceptions of health and wellbeing, approaches to navigating the HC relationship and patterns of help-seeking behavior (Gopalkrishnan, 2018). A few Asian resources exist, including KokoroChat in Japanese (Qi et al., 2025), KMI in Korean (Kim et al., 2025), EmoInHindi in Hindi (Singh et al., 2022), and MEDIC in Chinese (Zhu et al., 2023), but all remain monolingual. Parallel or bilingual corpora are even rarer: BiMISC covers English–Dutch counseling (Sun et al., 2024). This scarcity of multilingual or code-switched conversational data—especially from Southeast Asia—limits cross-cultural modeling and perpetuates Western norms in dialogue systems. In contrast, the `CoachLah` dataset introduces code-switching across English, Malay, Hokkien, and Tamil, paired with standard English translations, offering the first culturally grounded resource of its kind for Asian HC.

Behavior Change and Goal-Oriented Tasks. Most conversational datasets do not reflect the iterative, goal-driven nature of HC. Resources in psychotherapy (Gunal et al., 2025) and motivational interviewing (MI) (Wu et al., 2022; Cohen et al., 2024; Kim et al., 2025) center on therapist behaviors, reflective strategies, or session summaries rather than measurable progress over time. Likewise, broader counseling and emotional support datasets (Qi et al., 2025; Xu et al., 2025) emphasize affective support, empathy, or alliance-building rather than behavior change. Only a few HC datasets exist, most notably Gupta et al.'s SMS-based corpus and its extensions (Gupta et al., 2020, 2021) and Zhou et al.'s follow-up goal-tracking dataset (Zhou et al., 2024a). These corpora consist of SMS message exchanges in English, cover relatively small participant populations, and focus on weekly goal summaries. While previous work has introduced tools like *SMARTMiner* (Bojic et al., 2025b) to extract goals from HC notes, these rely on post-session summaries written by health coaches. By contrast, `CoachLah` captures the raw, unstructured negotiation of these goals through full-length HC session transcripts, preserving the longitudinal progress and explicit behavior goal structure as they emerge in natural dialogue.

Data Authenticity. Existing conversational datasets vary widely in how closely they reflect real-world interactions. Several psychotherapy (Malhotra et al., 2022) and MI corpora (Wu et al., 2022; Cohen et al., 2024) are transcribed from publicly available counseling videos, offering ecological realism but limited scale and privacy constraints. Others rely on anonymized online platforms, such as the Chinese Client Reactions corpus (Li et al., 2023), or simulated role-play setups (Qi et al., 2025; Zhu et al., 2023) that trade naturalism for control. Recent work also introduces synthetic corpora generated by LLMs (Kim et al., 2025) that improve scalability but risk stylistic drift from real practice. Only a few datasets originate from structured HC interventions (Gupta et al., 2020; Zhou et al., 2024a) but are in a short-form and text-only. `CoachLah` advances data authenticity by drawing from HC sessions conducted within a randomized controlled trial (RCT), preserving natural turn structure and culturally embedded communication.

3. `CoachLah` Dataset

Figure 2 illustrates how the `CoachLah` dataset was created. Audio was captured via an online HC platform in controlled, quiet environments without post-processing (Huang et al., 2025; Liu et al., 2014). The pipeline begins with automatic speaker diarization (e.g., (Liu et al., 2021)), followed by Singlish transcription and English translation.

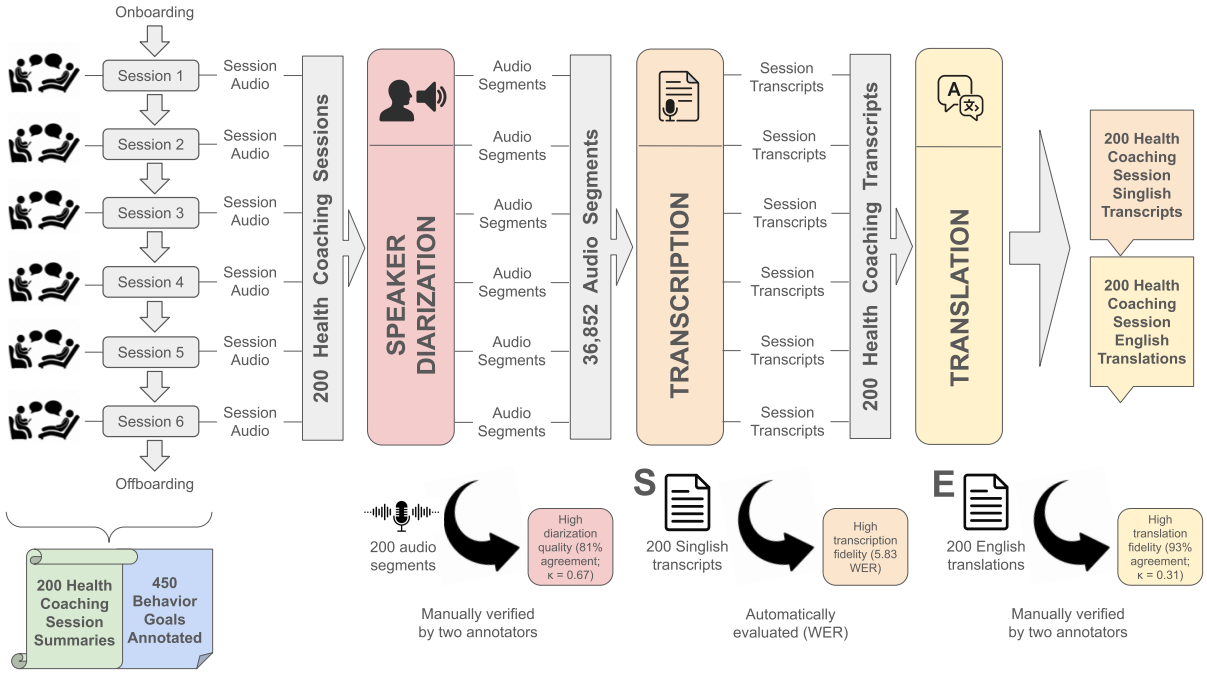

Figure 2: Overview of the CoachLah data processing pipeline.

To ensure reliability, two annotators independently verified each stage. In tandem, HC session summaries were retrieved from the platform, and behavior goals were manually annotated to align with the corresponding dialogues. The resulting dataset comprises 200 sessions of authentic coach-client interactions. These parallel Singlish–English transcripts enable fine-grained analysis of code-switching and translation quality. Furthermore, the inclusion of structured HC summaries and behavior goals for every HC session supports contextual, goal-oriented dialogue modeling (Table 1).

Metrics	Dataset
# HC sessions	200
# Utterances	36,852
# Unique tokens	13,431
Avg. utterances per HC session	184.26
Avg. tokens per utterance	42.39

Table 1: Dataset statistics for the HC sessions.

3.1. Data Collection

Data were sourced from the intervention arm of a multi-center RCT evaluating a mobile health (mHealth) application integrated with human HC to improve lipid-lowering medication adherence. Participants were recruited on a rolling basis from two public polyclinics and provided informed consent. At the time of collection, recruitment had been active for more than a year and a half. The study protocol was approved by the National Healthcare Group Domain Specific Review Board (no. 2023/00438), and participants were compensated for their time.

HC was delivered at monthly intervals over a six-month period, focusing on education, adherence support, and goal setting. The final dataset consists of 200 audio-recorded sessions conducted by 4 trained health coaches with 51 clients identified as non-adherent to therapy. The longitudinal breakdown of these sessions includes: 43 first sessions (22%), 44 second sessions (22%), 39 third sessions (19%), 39 fourth sessions (19%), 23 fifth sessions (12%), and 12 sixth sessions (6%).

3.2. Sociodemographic Characteristics

Among the 51 participants included in the CoachLah dataset, 28 (55%) were male. The majority of participants were of Chinese ethnicity (n=38, 75%), followed by Malay (n=8, 16%), Indian (n=2, 4%), and other ethnicities (n=3, 6%). These distributions broadly reflect the demographic composition of Singapore's resident population, which is predominantly Chinese (74%), with smaller proportions of Malay (13%) and Indian (9%) (Singapore Department of Statistics, 2021). Participants were middle-aged to older adults, ranging from 22 to 81 years of age (mean = 59, SD = 10), reflecting the target population for primary and secondary cardiovascular disease (CVD) prevention interventions, where long-term medication adherence and lifestyle modification support are particularly relevant.

3.3. Evaluation Subset of HC Sessions for Quality Assessment

To assess the quality of our `CoachLah` dataset, we extracted a balanced evaluation subset of 200 audio segments (0.7 hours in total) from six HC sessions. The segments had an average duration of 11.7 seconds (ranging from 2.1 to 30.1 seconds) and were selected to ensure demographic diversity across health coach–client dyads. This evaluation subset of HC sessions was consistently used to assess speaker diarization performance, transcription accuracy, and translation quality. Table 2 summarizes the evaluation dyads by gender and ethnicity of both participants and reports the number of diarized audio segments per dyad.

		Health Coaches		
		F / W	**F / CH**	**M / CH**
Clients	**F / CH**		26	48
	M / CH	31	22	
	M / MY			25
	M / IN		48	

Table 2: Health coach–client dyads in the evaluation subset (number of audio segments). F = female, M = male; ethnicity: W = White, CH = Chinese, MY = Malay, IN = Indian.

3.4. Speaker Diarization

Methods. Our HC session recordings consisted of continuous audio files containing both the health coach and the client speech segments. As a first step, we automatically separated the two speakers using the pretrained `pyannote/speaker-diarization-3.1` model (Bredin et al., 2020). Since conversational speech often contains vocal fillers (e.g., "um", "mm-hmm"), these were not treated as turn boundaries to avoid over-segmentation. Instead, only speaker-attributed segments longer than one second were retained. Consecutive diarization turns assigned to the same speaker were merged to prevent fragmentation caused by brief hesitations, pauses, or back channels. Overlapping speech was not explicitly resolved and remained labeled according to the diarization output.

Quality Assessment. To evaluate diarization quality, two annotators independently reviewed all 200 segments from the evaluation subset of HC sessions: one Singaporean native speaker and one non-native annotator with several years of residence in Singapore and daily exposure to Singlish. Segment-level evaluation was performed, and each segment was assigned to one of six categories (Broux et al., 2018; McKnight et al., 2022):

- *Correct* — speaker label and segment boundaries are accurate;
- *Minor error* — small, non-impactful deviations such as inclusion of brief fillers or back channels from the other speaker;
- *Wrong speaker label* — segment attributed to the incorrect speaker;
- *Missed split (under-segmentation)* — two distinct turns were merged into a single segment due to a missing boundary;
- *Extra split (over-segmentation)* — a single turn was incorrectly divided into multiple segments by spurious boundaries;
- *Overlap* — both speakers talk simultaneously but the segment is labeled with only one speaker.

Inter-annotator agreement was measured using Cohen's kappa (κ) (Landis and Koch, 1977).

Error analysis. Across all 200 segments, the first annotator assigned 171 and the second 178 cases to the *Correct* or *Minor error* categories, indicating that the vast majority of diarized segments were accurate or only minimally affected by boundary noise. The annotators agreed on 161 out of 200 segments (80.5% raw agreement), yielding Cohen's $\kappa = 0.67$ (SE = 0.04, 95% CI [0.58, 0.76]), reflecting *substantial agreement* between annotators. Taken together, these results suggest that diarization quality was high overall, with only a small proportion of segments exhibiting errors likely to meaningfully affect downstream analyses.

3.5. Transcription

Methods. Following diarization, speaker-attributed segments were transcribed using a fine-tuned Automatic Speech Recognition (ASR) model specifically adapted for this task. Since Singlish is underrepresented in general-purpose ASR systems, we curated a bespoke dataset, NSCP16, from the Singapore National Speech Corpus (NSC) (Koh et al., 2019), which combines non-conversational speech (phonetically-balanced scripts and random sentences; Parts 1–2) with conversational and expressive speech (natural dialogues, stylized debates and emotions, and scenario-based interactions; Parts 3, 5, and 6).

An overview of the dataset splits, including the number of samples, duration, and utterance length statistics, is provided in Table 3. We fine-tuned `Whisper-medium` (Radford et al., 2023) on `NSCP16_train`. Training was conducted for five epochs with a batch size of 16 using the AdamW optimizer (learning rate 10^{-5}, weight decay 10^{-2}) and a ReduceLROnPlateau scheduler (reduction factor 0.5, patience 2, threshold 0.01), with the final model selected based on lowest validation loss.

Name	Samples	Hours	Avg. (s)	Min (s)	Max (s)
train	2,048,000	2,944.1	5.2	0.1	30.1
valid	50,000	73.4	5.3	0.8	29.1
test	10,000	19.1	6.9	1.0	26.1

Table 3: Overview of the NSCP16 transcription dataset derived from NSC Parts 1–6.

Quality Assessment. On the held-out NSCP16_test set, the fine-tuned Whisper-medium[1] substantially reduced transcription errors, with WER dropping to 6.63 — a 69% relative improvement over the off-the-shelf model. To further assess the quality of the automatically transcribed CoachLah dataset, we conducted a quality assurance procedure on 200 segments sampled from the evaluation subset of HC sessions. These segments were manually transcribed by an annotator who is not a native Singaporean speaker but has lived in Singapore for several years and is deeply familiar with Singlish through extensive daily exposure, and were subsequently verified and corrected where necessary by a Singaporean native speaker.

Error analysis. Table 4 summarizes transcription performance on the evaluation subset of sessions. Our fine-tuned Whisper-medium achieved the lowest normalized sentence-level WER of 5.83, outperforming all pre-trained ASR baselines and surpassing the MERaLiON-AudioLLM (He et al., 2024) fusion model—based on Whisper-large-v2 (Radford et al., 2023)—by 2%. This performance is comparable to human transcription accuracy for conversational English (5.1–6.8 WER) (Saon et al., 2017). Across the 200 evaluated segments, 79 (40%) were transcribed perfectly (WER = 0), while 111 segments (56%) achieved WER ≤ 5, thus falling below the lower bound of the human WER range.

A qualitative inspection of the transcription errors reveals three primary patterns. First, phonetic ambiguity within the Singaporean accent led to occasional lexical substitutions, such as mistranscribing *"wife"* as *"wifi"* or *"cough"* as *"call"*. Second, the model occasionally struggled with local abbreviations and domain-specific terminology; for instance, the term *"MC"* (Medical Certificate) was mistranscribed as *"i want to see"* in one context. Third, several errors involved the omission or insertion of minor discourse markers (e.g., *"you know"*, *"okay"*) or numeric discrepancies in colloquial speech (e.g., *"two eighty two"* vs. *"two eight"*). These errors generally affect discourse markers rather than core semantic content, resulting in minimal impact on downstream tasks such as translation and behavioral goal extraction.

[1]The fine-tuned checkpoint is publicly available at: https://huggingface.co/ivabojic/whisper-medium-sing2eng-transcribe

Model name	WER ↓	Diff (%)
Whisper-medium (fine-tuned)	**5.83**	—
Whisper-small (pre-trained)	15.75	63
Whisper-medium (pre-trained)	12.62	54
Whisper-large-v3 (pre-trained)	12.02	52
SeamlessM4T-medium (pre-trained)	48.07	88
SpeechT5 (pre-trained)	67.33	91
MERaLiON-AudioLLM	<u>5.94</u>	2

Table 4: Normalized WER comparison of the fine-tuned Whisper-medium with pre-trained ASR models and MERaLiON-AudioLLM. A lower WER indicates better performance (↓). The best WER is highlighted and the second best underlined.

3.6. Translation

Methods. To obtain high-quality Singlish-to-English translations of the HC transcripts, we employed GPT-4o mini (OpenAI, 2024). The initial HC training dataset consisted of GPT-generated translations for 5,000 original audio transcriptions, each longer than two seconds. To enrich lexical and syntactic diversity, we applied three additional rephrasing prompts to each original transcript, resulting in four translations per segment and expanding the training set to 20,000 samples in total. For evaluation, we constructed HC validation and test sets containing 2,000 samples each, generated using a single prompt to ensure consistency and comparability. The characteristics of the translation datasets are summarized in Table 5. The exact prompts used for translation and rephrasing are provided in Appendix A.

Name	Samples	Total hours	Utt. Dur. (sec.)		
			Avg.	Min.	Max.
train	20,000	94.9	17.1	2.0	378.4
valid	2,000	9.2	16.6	2.0	463.0
test	2,000	9.0	16.2	2.0	172.7

Table 5: Overview of HC translation splits. Utterance durations are reported in seconds.

Quality Assessment. To assess translation fidelity, all 200 translations in the evaluation subset of HC sessions were independently reviewed by two annotators: one Singaporean native speaker and one non-native annotator with several years of residence in Singapore and daily exposure to Singlish. Each translation was rated on a binary scale (1 = acceptable, 0 = unsatisfactory). For every case marked as 0, the annotators provided a short comment describing the issue (e.g., hallucination, over-smoothing, or omission). Disagreements were resolved through discussion and joint review of the corresponding source transcript. Inter-annotator agreement was measured using Cohen's kappa (κ) (Landis and Koch, 1977).

Error analysis. Both annotators independently rated 186 out of 200 translations (93%) as acceptable, indicating that `GPT-4o mini` produced overall high-quality, fluent, and faithful Standard English outputs for the vast majority of Singlish inputs. Inter-annotator consistency was *moderate*, with agreement on 182 translations (91% raw agreement) and a Cohen's $\kappa = 0.31$ (SE = 0.12, 95% CI [0.06, 0.54]). A qualitative review of the 18 cases of inter-annotator disagreement and the 14 translations rated as unacceptable by both annotators shows that most issues arose from *pragmatic shifts* rather than semantic failures.

In several instances, `GPT-4o mini` converted tentative coaching questions into more assertive statements or introduced minor hallucinated phrases to improve sentence flow. Some disagreements reflected different annotation preferences, with one annotator favoring preservation of Singlish discourse markers and the other accepting stronger grammatical standardization. Overall, the identified errors were infrequent and typically involved subtle meaning adjustments rather than severe mistranslations, and none resulted in the loss of the core behavioral goal, indicating that the translation pipeline remains reliable for downstream goal extraction tasks.

3.7. HC Session Summaries and Goal Annotations

Methods. Health coach interactions were documented as free-text summaries. We retrieved all HC summaries via an SQL query. Three trained annotators independently reviewed the 200 summaries, marking spans that expressed specific behavioral goals. Following this independent review, all labeling discrepancies were resolved through a consensus-based approach to reach a final, unified set of gold-standard annotations.

Results. From 200 HC summaries, 450 goals were extracted. The median number of goals per HC summary was 2 [IQR: 1–3], with a range of 0–7. Most HC summaries contained two or three goals (56%), and 81% of HC summaries contained at least one goal (Figure 3). Contributions were skewed toward a single health coach, who authored 139/200 (70%) HC summaries and contributed 324/450 goals (72%). Figure 4 illustrates an example of the end-to-end annotation process, demonstrating how a snippet from an HC conversation is summarized by health coaches and subsequently used to extract behavioral goals from the HC summary. These HC summaries serve as the basis for the manual identification and annotation of 450 behavioral goals across the dataset, capturing the actionable health-related objectives discussed during the HC sessions.

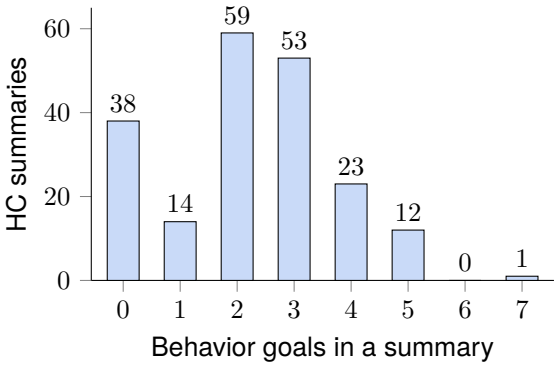

Figure 3: Distribution of behavior goals per HC session summary (N=200).

4. Evaluation Experiments

To demonstrate the usability and research potential of the `CoachLah` dataset, we conducted a series of evaluation experiments across two downstream tasks. These experiments assess how well existing open-source models can adapt to the linguistic, cultural, and conversational characteristics of HC dialogues. Each task highlights a different dimension of the dataset's applicability: (1) translation of conversational Singlish into standard English and (2) extraction of behavioral goals from HC session summaries. These evaluations provide empirical evidence of the dataset's utility for both language technology research and AI-driven HC applications.

4.1. Tasks

4.1.1. Singlish-to-English Translation

This task evaluates the usability of the `CoachLah` dataset for developing translation systems that convert conversational Singlish utterances into fluent standard English. Accurate translation is essential for downstream applications such as cross-lingual goal extraction, empathy modeling, and multilingual HC dialogue analysis. The task benchmarks open-source LLMs and investigates their ability to adapt to the colloquial and code-switched characteristics of Singaporean English.

4.1.2. Extraction of Behavioral Goals from HC Summaries

This task evaluates the usability of the `CoachLah` dataset for structured information extraction, focusing on the identification of behavioral goals within free-text HC session summaries. Each HC summary records a health coach's reflections on client progress, challenges, and goals. Automatically extracting these goal spans enables downstream applications such as behavior goal tracking and behavior change analytics.

We frame behavioral goal extraction as a multi-span reading comprehension problem, where each HC summary serves as the context paired with a fixed question Q: *"What are the behavior goals mentioned in the HC summary?"*. Although the original dataset includes annotated goal spans, it does not contain question–answer pairs. To enable span-based supervision, we recast the dataset into a QA format using this fixed question. The model is trained to predict all non-overlapping goal spans.

4.2. Experimental Setup

4.2.1. Translation Models

We fine-tuned six state-of-the-art open-source large language models (LLMs)—`Llama-3.1-8B-instruct` (Dubey et al., 2024), `Llama-3.2-3B-instruct` (AI, 2024), `Gemma-2-9B-it` (Team et al., 2024), `Phi-3.5-mini-instruct` (Abdin et al., 2024), `DeepSeek-R1-Distill-Llama-8B`, and `DeepSeek-R1-Distill-Qwen-7B` (Deepseek-AI, 2025)—alongside the multilingual `NLLB-200-3.3B` encoder-decoder model (Costa-Jussà et al., 2022). All models were fine-tuned using a standardized Singlish-to-English translation prompt (see Appendix A). Comprehensive model specifications, including parameter counts and training cut-off dates, are detailed in Appendix B.

Parameter-efficient low-rank adaptation (LoRA) (Hu et al., 2021) was applied with rank $r = 128$ and scaling factor $\alpha = 16$. Each model was fine-tuned for one epoch on a single NVIDIA A40 (48GB) GPU using a batch size of 8, a learning rate of 10^{-3}, AdamW 8-bit optimization, linear learning rate scheduling, gradient checkpointing with mixed precision, and weight decay of 0.01. Decoding was performed using greedy generation with temperature set to 0 and a maximum output length of 256 tokens. All experiments were implemented using the HuggingFace Transformers library.

4.2.2. Goal Extraction Model

For behavioral goal extraction, we fine-tuned a span-extractive model using the *SpanQualifier* framework (Huang et al., 2023), which computes span-level representations and jointly optimizes span classification and boundary regression objectives. The starting checkpoint was the publicly available `DeBERTa-v3-base` model[2] previously fine-tuned on the general-domain *MultiSpanQA* dataset (Li et al., 2022), providing a strong initialization for this task.

Following the configuration in Huang et al. (2023), for fine-tuning, we used the AdamW optimizer with a learning rate of 3×10^{-5} and weight decay of 0.01, together with linear learning rate scheduling and a 5% warm-up ratio. The maximum sequence length was set to 512 tokens, and training was performed for up to 100 epochs. We used a training batch size of 32 with gradient accumulation of 4. Early stopping was applied after five epochs without improvement on the validation set.

Given the limited size of the `CoachLah` dataset (200 annotated HC summaries containing 450 behavior goal spans), we adopted a five-fold cross-validation strategy to ensure robust evaluation and mitigate sampling bias. Each fold followed a 70%/15%/15% `train`/`validation`/`test` split, yielding 140 summaries for training and 30 summaries each for validation and testing. As summarized in Table 6, the number of goals per `test` split varied across folds, reflecting natural differences in goal density across HC session summaries.

[2]The fine-tuned checkpoint is publicly available at: `https://huggingface.co/ivabojic/deberta-v3-base_MultiSpanQA`

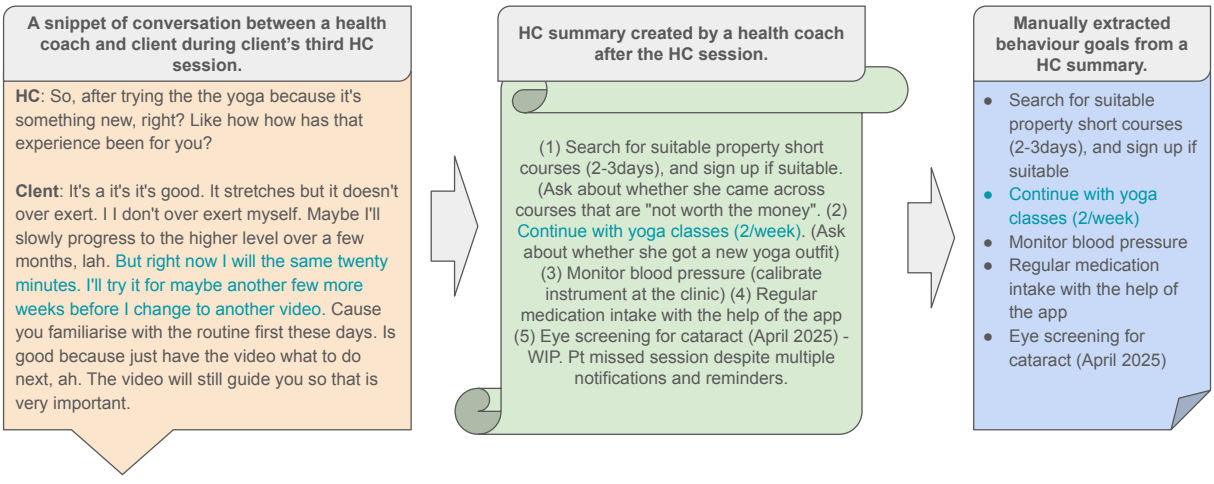

Figure 4: Example of the goal extraction process from Singlish transcript of a HC session.

Goals	0	1	2	3	4	5	6+	∑ Goals
Split 1	3	2	12	7	6	0	0	71
Split 2	7	3	6	7	4	2	1	69
Split 3	3	4	7	11	3	2	0	73
Split 4	4	1	8	10	7	0	0	75
Split 5	6	3	9	5	2	5	0	69

Table 6: Distribution of goals per HC summary across five cross-validation `test` splits.

4.3. Results

4.3.1. Translation

To assess the impact of supervised fine-tuning, we compared each model's zero-shot and fine-tuned performance on the held-out `HC_test` set using sentence-level BLEU (Papineni et al., 2002). Results are summarized in Table 7. Among all evaluated models, `Gemma-2-9B-it` achieved the highest BLEU score, followed closely by `Llama-3.1-8B-instruct`. These results indicate that compact instruction-tuned LLMs can effectively learn the linguistic patterns of conversational Singlish with minimal fine-tuning. In contrast, the multilingual `NLLB-200-3.3B`, despite being designed for low-resource translation, achieved lower performance, suggesting that adaptation to conversational style is more important than broad language coverage for this task.

4.3.2. Goal Extraction

Table 8 summarizes the performance of the baseline and fine-tuned models across five cross-validation splits using exact match (EM) and partial match (PM) F1 metrics following Li et al. (2022). The fine-tuned model consistently outperformed the baseline across all splits. On average, fine-tuning on the `CoachLah` dataset increased EM F1 from 2.40 to 83.47 and PM F1 from 16.37 to 93.22, demonstrating the substantial benefit of domain-specific HC data for training span-based goal extraction models.

4.4. Error Analysis

4.4.1. Translation

To better understand the limitations of the fine-tuned `Gemma-2-9B-it` model, we analyzed 237 outputs with sentence-level BLEU scores below 0.10 using `GPT-5` to cluster error types and assign severity levels (Chiang and Lee, 2023; Liu et al., 2023). Seven error categories were identified: (1) *Addition*, (2) *Grammar/Fluency*, (3) *Lexical mismatch*, (4) *Omission*, (5) *Semantic drift*, (6) *Reordering*, and (7) *Overcompression*.

The majority of errors were *Lexical mismatches* (156/237, 66%), typically rated as moderate. *Addition* (28/237, 12%) and *Overcompression* (17/237, 7%) errors were generally major, introducing or removing substantial content. *Omission* (22/237, 9%) and *Semantic drift* (8/237, 3%) were frequently critical, causing severe meaning distortion. *Grammar/Fluency* and *Reordering* errors were rare (combined <3%). Example translations illustrating several error categories are provided in Appendix C.

4.4.2. Goal Extraction

To examine typical failure modes in behavioral goal extraction, we identified five error categories:

- *Boundary / segmentation errors*: Multi-clause goals are truncated, merged, or segmented differently from the gold annotation, resulting in incomplete or misaligned span predictions.
- *Overgeneralization*: Broad or underspecified health intentions are predicted as goals despite lacking a clearly actionable component.
- *Vision confusion*: Longer-term wellness aspirations or broader health visions are incorrectly extracted as short-term behavioral goals.
- *Omission*: Gold goals present in the note are not extracted by the model.
- *Hallucination*: The model predicts a goal span that is not supported by the source note or does not correspond to an annotated goal.

Model	Zero-shot BLEU	Fine-tuned BLEU	Diff (%)
Llama-3.1-8B-instruct	<u>0.3494</u>	<u>0.4379</u>	25
Llama-3.2-3B-instruct	0.2861	0.4227	48
Gemma-2-9B-it	**0.3910**	**0.4448**	14
Phi-3.5-mini-instruct	0.1852	0.4344	135
DeepSeek-R1-Distill-Llama-8B	0.2998	0.4215	41
DeepSeek-R1-Distill-Qwen-7B	0.2713	0.3977	47
NLLB-200-3.3B	0.2953	0.4125	40

Table 7: Zero-shot vs. fine-tuned BLEU scores of open-source LLMs on the `HC_test` set. Higher BLEU indicates better translation quality (↑). The last column shows relative improvement from the zero-shot baseline. The best BLEU score is highlighted, while the second best is underlined.

Split	Baseline (MultiSpanQA)		Fine-tuned (CoachLah)	
	EM F1 ↑	PM F1 ↑	EM F1 ↑	PM F1 ↑
Split 1	1.52	13.58	87.25	92.96
Split 2	0.00	17.24	86.09	94.58
Split 3	2.78	19.70	85.71	95.49
Split 4	1.53	11.50	83.12	92.67
Split 5	6.15	19.83	75.18	90.38
Mean ± SD	2.40 ± 2.32	16.37 ± 3.72	83.47 ± 4.71	93.22 ± 1.90

Table 8: Goal extraction performance of baseline and fine-tuned models across five splits. Higher scores indicate better performance (↑).

We analyzed *Split 5*, the weakest-performing fold, to examine these failure modes before and after fine-tuning. Before fine-tuning, the baseline model produced correct extractions for only 1 of 30 notes (3%) and was dominated by semantic errors, most notably *overgeneralization* in 12 notes (40%) and *vision confusion* in 7 notes (23%), where broad wellness intentions or long-term aspirations were predicted instead of concrete weekly goals.

After fine-tuning, correct extraction increased to 18 of 30 notes (60%), while both *overgeneralization* and *vision confusion* were eliminated. The main remaining error type was *boundary / segmentation errors*, observed in 7 notes (23%), indicating that the model generally identified the correct goal content but sometimes split or merged multi-part goals differently from the gold annotation. Representative examples comparing baseline and fine-tuned extraction errors are provided in Appendix D.

5. Conclusions and Future Work

In this paper, we introduced CoachLah, the first large-scale, culturally grounded parallel corpus of Health Coaching (HC) conversations designed to bridge the gap in non-Western, low-resource dialogue modeling. Comprising 200 real-world sessions and over 36,000 utterances, the dataset captures the linguistic complexity of Singaporean English, including extensive code-switching, ellipsis, and local discourse markers across a demographically diverse client pool (75% Chinese, 16% Malay, 4% Indian, and 6% other ethnicities). By providing parallel Singlish-to-Standard English translations and expertly annotated behavioral goals, CoachLah establishes a robust foundation for building culturally competent AI HC systems.

Our benchmark experiments demonstrate that domain-specific fine-tuning effectively addresses the challenges of this low-resource variety. We achieved a competitive WER of 5.83 for Singlish ASR, established strong translation baselines using instruction-tuned LLMs such as Gemma-2-9B-it, and demonstrated massive performance gains in behavioral goal extraction, improving exact match F1 scores from 2.40 to 83.47. Ultimately, CoachLah provides a vital, reproducible resource for developing scalable, culturally adaptive mHealth interventions capable of addressing lifestyle-related diseases in diverse, multilingual populations.

The release of the CoachLah dataset opens several promising directions for future research in healthcare NLP and dialogue systems. First, although our translation baselines achieve high fluency, further work is needed to reduce lexical mismatches and omissions in highly code-switched segments, potentially by incorporating multilingual lexicons covering Malay, Hokkien, and Tamil. Second, the longitudinal structure of the dataset offers opportunities to study temporal dialogue modeling, including behavioral goal tracking, changes in client motivation, and the evolving coaching relationship across sessions. Third, the parallel transcripts could support the development of end-to-end generative AI health coaches that adapt their language style to the user's sociolinguistic context. Finally, future data collection should expand representation of minority ethnic groups and broader chronic health conditions to improve the robustness and inclusivity of AI-driven HC systems.

6. Limitations

The CoachLah dataset was developed through a largely automated pipeline encompassing speaker diarization, transcription, and translation, with each stage undergoing quality control by two independent reviewers, one with high proficiency in Singlish and English and a computer science background, and the other a native Singlish English speaker, psychologist, and professional health coach who authored many of the original summaries. Evaluation results showed that diarization errors were minor, transcription exhibited error patterns comparable to those of human transcribers, and translation outputs maintained high semantic fidelity across the evaluation subset. Together, these findings indicate that the automated pipeline is sufficiently reliable to support downstream analysis, benchmarking, and model development on the CoachLah dataset.

The conversations in the dataset were collected within a clinical trial targeting clients with suboptimal adherence to statin therapy and elevated cardiovascular risk, which may limit generalizability to other health domains or clinical contexts. As the dataset primarily represents Singapore's multilingual population, language use and communication styles may differ substantially from those in other cultural settings. Finally, because participant recruitment was ongoing at the time of data collection, not all clients had completed the full six-month HC program, and therefore the dataset provides only a partial snapshot of longitudinal goal progression rather than a complete representation of behavioral change trajectories across all participants.

7. Ethical Considerations

The randomized controlled trial from which the health coaching session transcripts and notes were derived received ethics approval from the National Healthcare Group Domain Specific Review Board, Singapore (no. 2023/00438). All participants provided written informed consent prior to enrollment.

8. Acknowledgments

This work was supported by CArdiovascular DiseasE National Collaborative Enterprise (CADENCE) National Clinical Translational Program (MOH-001277-01).

Bibliographical References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open models. Accessed: February 3, 2025.

Iva Bojic, Qi Chwen Ong, Sakura Ito, Jintana Liu, Ashwini Lawate, Malar Palaiyan, Elizabeth Nair, May Lwin, Yin Leng Theng, Michael Chia, et al. 2025a. Ai-empowered health coaching for university students: A mixed-method process evaluation. *Computers in Biology and Medicine*, 194:110271.

Iva Bojic, Qi Chwen Ong, Stephanie Hilary Xinyi Ma, Lin Ai, Zheng Liu, Ziwei Gong, Julia Hirschberg, Andy Hau Yan Ho, and Andy W. H. Khong. 2025b. Smartminer: Extracting and evaluating smart goals from low-resource health coaching notes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16288–16305.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Pierre-Alexandre Broux, David Doukhan, Simon Petitrenaud, Sylvain Meignier, and Jean Carrive. 2018. Computer-assisted speaker diarization: How to evaluate human corrections. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Deepseek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Narayan Gopalkrishnan. 2018. Cultural diversity and mental health: Considerations for policy and practice. *Frontiers in public health*, 6:179.

William B Gudykunst. 2003. *Cross-cultural and intercultural communication*. Sage.

Aylin Gunal, Bowen Yi, John Piette, Rada Mihalcea, and Verónica Pérez-Rosas. 2025. Examining spanish counseling with midas: a motivational interviewing dataset in spanish. *arXiv preprint arXiv:2502.08458*.

Anthea Fraser Gupta. 1994. *The step-tongue: Children's English in Singapore*, volume 101. Multilingual Matters.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2021. Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue,*.

Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F Chen, and Ai Ti Aw. 2024. Meralion-audiollm: Technical report. *arXiv preprint arXiv:2412.09818*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Gongping Huang, Jesper R. Jensen, Jingdong Chen, Jacob Benesty, Mads G. Christensen, Akihiko Sugiyama, Gary Elko, and Tomas Gaensler. 2025. Advances in microphone array processing and multichannel speech enhancement. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zixian Huang, Jiaying Zhou, Chenxu Niu, and Gong Cheng. 2023. Spans, not tokens: a span-centric model for multi-span reading comprehension. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 874–884.

Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. 2025. Kmi: A dataset of korean motivational interviewing dialogues for psychotherapy. *arXiv preprint arXiv:2502.05651*.

Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.

Jia Xin Koh, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and YY Tan. 2019. Building the singapore english national speech corpus. *Malay*, 20(25.0):19–3.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jakob RE Leimgruber. 2011. Singapore english. *Language and Linguistics Compass*, 5(1):47–62.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. *arXiv preprint arXiv:2306.15334*.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. Multispanqa: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 1250–1260.

Lisa Lim. 2004. *Singapore English: A grammatical description*, volume 33. John Benjamins Publishing.

Benxu Liu, Vaninirappuputhenpurayil Gopalan Reju, and Andy W. H. Khong. 2014. A linear source recovery method for underdetermined mixtures of uncorrelated ar-model signals without sparseness. *IEEE Transactions on Signal Processing*, 62(19):4947–4958.

Hexin Liu, Leibny Paola García Perera, Xinyi Zhang, Justin Dauwels, Andy W.H. Khong, Sanjeev Khudanpur, and Suzy J. Styles. 2021. End-to-End Language Diarization for Bilingual Code-Switching Speech. In *Interspeech 2021*, pages 1489–1493.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.

Lena Mamykina, Elliot Mitchell, Pooja Desai, and David Albers. 2024. Intelligent decision support in personal health: Personalized health coaching in type 2 diabetes. In *Human Computer Interaction in Healthcare: The Role of Cognition*, pages 413–438. Springer.

Simon W McKnight, Aidan OT Hogg, Vincent W Neo, and Patrick A Naylor. 2022. Studying human-based speaker diarization and comparing to state-of-the-art systems. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–8. IEEE.

Ministry of Health, Singapore. 2024. Principal causes of death. https://www.moh.gov.sg/resources-statistics/singapore-health-facts/principal-causes-of-death. Accessed 2025-02-02.

National Heart Centre Singapore. 2023. The burden and risk of cardiovascular disease. Accessed: 2025-02-02.

Jeanette M Olsen and Bonnie J Nesbitt. 2010. Health coaching to improve healthy lifestyle behaviors: an integrative review. *American journal of health promotion*, 25(1):e1–e12.

Qi Chwen Ong, Chin-Siang Ang, Davidson Zun Yin Chee, Ashwini Lawate, Frederick Sundram, Mayank Dalakoti, Leonardo Pasalic, Daniel To, Tatiana Erlikh Fox, Iva Bojic, et al. 2024. Advancing health coaching: A comparative study of large language model and health coaches. *Artificial Intelligence in Medicine*, 157:103004.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-02-12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

John Talbot Platt. 1993. *Dynamics of a contact continuum: Singaporean English*. Clarendon Press.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. Kokorochat: A japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. *arXiv preprint arXiv:2506.01357*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. Int. Conf. Machine Learn.*, pages 28492–28518.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al. 2017. English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*.

Elizabeth Shriberg. 2001. To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the international phonetic association*, 31(1):153–169.

Singapore Department of Statistics. 2021. Census of population 2020 statistical release 1: Demographic characteristics. Technical report, Ministry of Trade and Industry, Singapore. Accessed: 2025-10-03.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoinhindi: A multi-label emotion and intensity annotated dataset in hindi for emotion recognition in dialogues. *arXiv preprint arXiv:2205.13908*.

Aseem Srivastava, Tharun Suresh, Sarah P Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos TP Dobber, and Jos A Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with llms: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621.

Vanessa Tan, Julian Lim, Katika Akksilp, Wai Leng Chow, Stefan Ma, and Cynthia Chen. 2023. The societal cost of modifiable risk factors in singapore. *BMC public health*, 23(1):1285.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Unsloth Team. 2024. Unsloth: Efficient fine-tuning of llms. Accessed: Feb. 3, 2025.

Stella Ting-Toomey. 2017. Conflict face-negotiation theory: Tracking its evolutionary journey. In *Conflict management and intercultural communication*, pages 123–143. Routledge.

Jessica L Unick, Christine A Pellegrini, Shira I Dunsiger, Kathryn E Demos, J Graham Thomas, Dale S Bond, Robert H Lee, Jennifer Webster, and Rena R Wing. 2024. An adaptive telephone coaching intervention for patients in an online weight loss program: A randomized clinical trial. *JAMA Network Open*, 7(6):e2414587–e2414587.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5367–5378.

Hong Zhang. 2020. *The Distribution Of Disfluencies In Spontaneous Speech: Empirical Observations And Theoretical Implications*. Ph.D. thesis, University of Pennsylvania.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, and Nikolaos Agadakos. 2024a. Modeling low-resource health coaching dialogues via neuro-symbolic goal summarization and text-units-text generation. *arXiv preprint arXiv:2404.10268*.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. 2024b. Towards enhancing health coaching dialogue in low-resource settings. *arXiv preprint arXiv:2404.08888*.

Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6054–6062.

Appendices

A. Translation, Rephrasing, and Fine-Tuning Prompts

To construct the Singlish-to-English translation dataset, we used `GPT-4o mini` to generate the target Standard English translations. During this dataset creation phase, we employed four distinct instruction prompts. Prompt 1 was used to generate the base translations, while Prompts 2 through 4 were used together to expand the diversity of the training set by introducing controlled lexical and syntactic variation. Subsequently, during the actual fine-tuning of our open-source models (e.g., `Gemma-2-9B-it`), we used only Prompt 1. This strategy ensured consistency between the fine-tuning and evaluation phases. The four prompts used to query `GPT-4o mini` are listed below:

1. *"Please rephrase the following text minimally, adjusting only as needed to make it follow official English structure while keeping the original meaning intact and considering that it was spoken Singlish."*

2. *"Please make minimal adjustments to the following text to ensure it follows standard English grammar and structure, while preserving the original meaning. Keep in mind that the text was originally spoken in Singlish."*

3. *"Adjust the following text as needed to match standard English grammar and structure, ensuring the original meaning remains unchanged. Note that the text was originally conveyed in Singlish."*

4. *"Please adjust the following text to standard English grammar and structure while preserving its original meaning. Note that the text was initially expressed in Singlish."*

B. Evaluated LLM Models

Table 9 lists the pre-trained large language models used in the translation experiments together with their parameter sizes, release dates, and reported pre-training knowledge cutoff dates. All models contain fewer than 10B parameters and were fine-tuned using `Unsloth` (Team, 2024), an optimized library that reduces GPU memory consumption and training time.

Table 9: Pre-trained LLMs used in the translation experiments, including parameter sizes, release dates, and reported pre-training knowledge cutoff dates.

Model name	Parameters	Release date	Pre-training cutoff
Llama-3.1-8B-instruct (Dubey et al., 2024)	8B	23/7/24	Dec. 2023
Llama-3.2-3B-instruct (AI, 2024)	3B	25/9/24	Dec. 2023
Gemma-2-9B-it (Team et al., 2024)	9B	27/6/24	—
Phi-3.5-mini-instruct (Abdin et al., 2024)	3.8B	20/8/24	Oct. 2023
DeepSeek-R1-Distill-Llama-8B (Deepseek-AI, 2025)	8B	20/1/25	post Oct. 2023
DeepSeek-R1-Distill-Qwen-7B (Deepseek-AI, 2025)	7B	20/1/25	post Oct. 2023
NLLB-200-3.3B (Costa-Jussà et al., 2022)	3.3B	6/7/22	—

C. Translation Error Examples

To illustrate typical translation errors observed in the qualitative analysis, Figure 5 presents representative examples produced by the fine-tuned `Gemma-2-9B-it` model. The examples highlight several error categories identified in the analysis, including *addition*, *lexical mismatch*, and *omission*. Addition errors introduce content that is not present in the original Singlish utterance, often expanding short conversational expressions into longer sentences. Lexical mismatches occur when the translation substitutes words with semantically related but contextually incorrect alternatives. Omission errors remove important tokens or phrases from the source utterance, sometimes resulting in incomplete or empty translations. While such errors are relatively infrequent, they illustrate the types of deviations that may arise when translating informal conversational Singlish into standardized English.

D. Goal Extraction Error Examples

Figure 6 illustrates typical goal extraction errors from the baseline `MultiSpanQA` and fine-tuned `CoachLah` models. The examples highlight three key scenarios: a baseline boundary error truncating a multi-clause goal (resolved by `CoachLah`), a rare hallucination where the fine-tuned model predicts a goal when none exists, and a severe baseline omission of a dietary instruction that `CoachLah` correctly extracts. Together, these examples show that fine-tuning substantially improves both span completeness and adherence to the behavioral goal definition. Ultimately, while the baseline struggles with semantic errors and omissions, the fine-tuned model proves highly accurate, with its rare errors largely confined to minor boundary shifts or occasional spurious predictions. This performance confirms that the adapted pipeline is reliable.

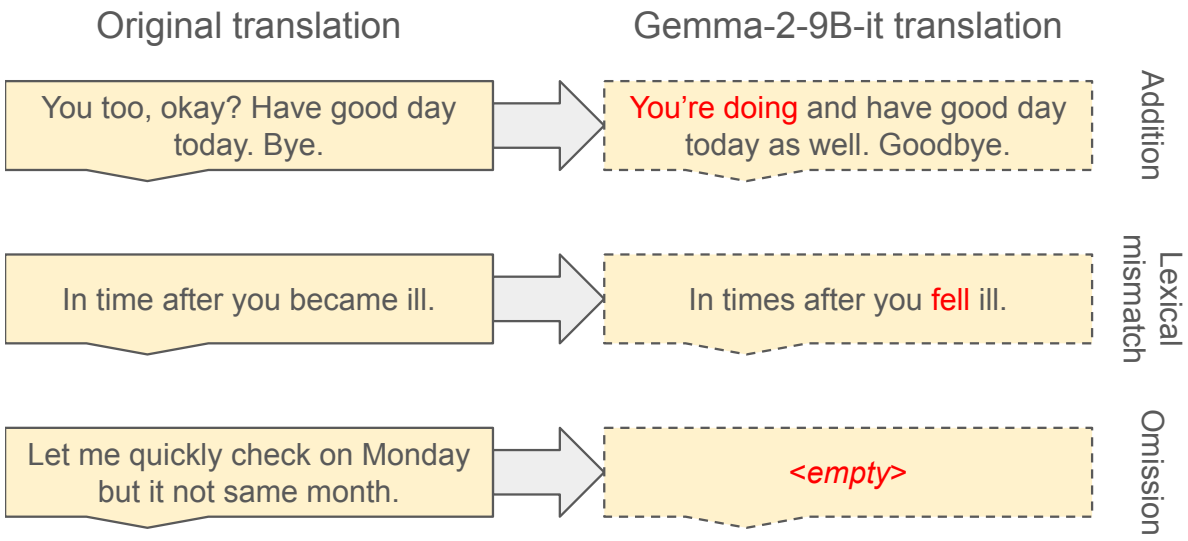

Figure 5: Representative translation errors produced by the fine-tuned `Gemma-2-9B-it` model, illustrating the categories of *addition*, *lexical mismatch*, and *omission*. Each example shows the original reference translation and the corresponding model output highlighting the error type.

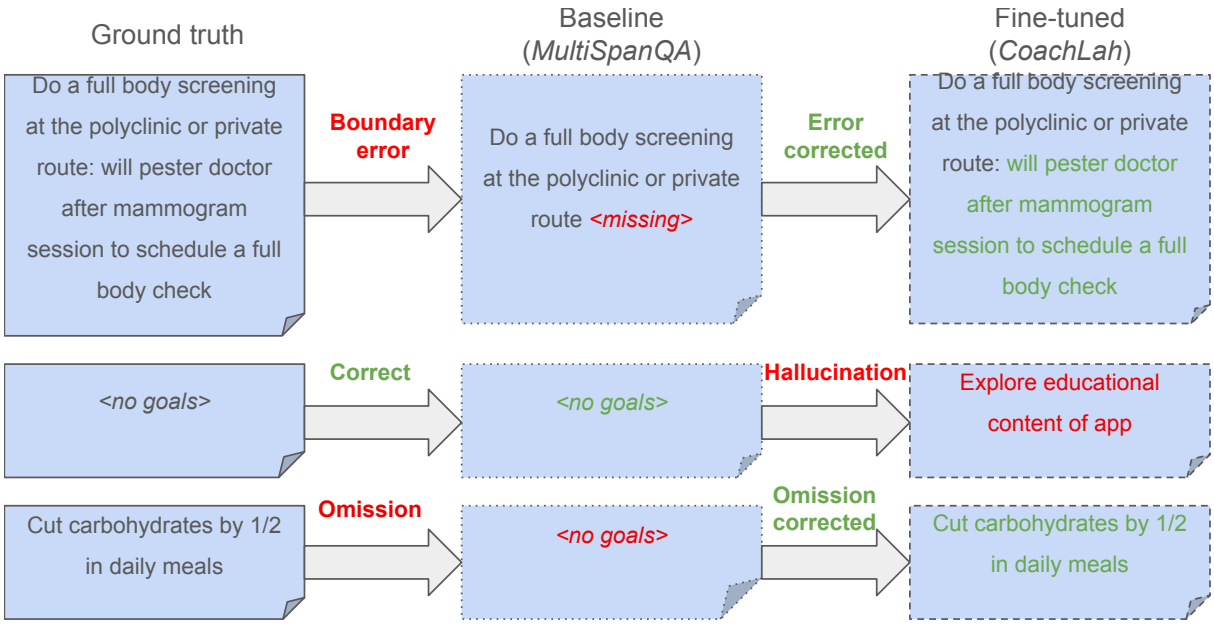

Figure 6: Representative goal extraction errors before and after fine-tuning. The examples illustrate a boundary error corrected by the fine-tuned model, a rare hallucination introduced by the fine-tuned model, and a severe omission by the baseline that the fine-tuned model successfully resolved.