Addressing Domain Shift in Health Coaching Note Analysis through Factorized Synthetic Data Generation

Michael Tänzer[1,2], **Iva Bojic**[1], **Ashwini Lawate**[1], **Andy Hau Yan Ho**,[1]
Andy W. H. Khong[1]
[1] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[2] Imperial College London, Imperial Global Singapore, Singapore
{cs-michael.tanzer, iva.bojic, ashwini.lawate, andyhyho, andy.khong}@ntu.edu.sg

Abstract

Automatic extraction of behavioral goals from health coaching notes is essential for scalable monitoring of coaching programs, yet training data is scarce and exhibits substantial domain shift across programs. We collect and annotate 157 notes from a coaching program and show that models trained on the only existing public corpus, SMARTSpan (173 notes), suffer a drop of up to 30 percentage points in exact-match F1 when transferred to our data. To address this, we propose a factorized synthetic data generation pipeline that decomposes note variation into three largely independent axes (health coach documentation structure, patient goal content, and patient persona), extracts empirical priors from a small in-domain seed set, and samples from them to generate diverse synthetic notes with embedded goal-span labels validated via cycle-consistency filtering. In low-resource experiments with only 57 in-domain training notes, our approach outperforms rephrasing and back-translation baselines on both exact-match and partial-match F1. Ablation analysis demonstrates that augmentation must target the in-domain distribution to be effective, and a human evaluation confirms that synthetic notes are structurally faithful, with detection driven by surface artifacts rather than content or organizational flaws. All code and generated data will be published at the following GitHub repository: cl4health-factorized-augmentation.

Keywords: health coaching, synthetic data generation, domain adaptation, goal span extraction, low-resource NLP, data augmentation

1. Introduction

Health coaching is a person-centered intervention that supports prevention, self-management, and sustained behavior change in individuals at risk of chronic disease (Wolever et al., 2013; Loughnane et al., 2025), with demonstrated benefits for physical activity, cardiovascular risk, and quality of life (Olsen and Nesbitt, 2010; Kivelä et al., 2014). Central to behavioral coaching is the formulation of Specific, Measurable, Attainable, Relevant, and Time-bound (SMART) goals, which support clients in maintaining momentum between sessions and adhering to agreed actions (Doran, 1981; Wallace et al., 2018; White et al., 2020; Bahrami et al., 2022). Health coaches record these goals in free-text session notes (Gupta et al., 2021; Bojic et al., 2025), making automatic extraction of goal spans a prerequisite for scalable monitoring and evaluation (Flocke and Stange, 2004; Bowman et al., 2015; Zhou et al., 2024; Loughnane et al., 2025).

Training span extraction models for this task is hampered by the lack of suitable annotated data. Conversational and counseling corpora emphasize emotional support or dialogue structure rather than measurable behavior change (Malhotra et al., 2022; Xu et al., 2025; Qi et al., 2025), and resources that do include goal-related content are often small, narrowly scoped, or limited to short text-based interactions (Gupta et al., 2020, 2021; Zhou et al., 2024). The closest publicly available resource, SMARTSpan (Bojic et al., 2025), provides 173 annotated notes from a single randomized controlled trial targeting cardiovascular risk reduction through statin adherence and lifestyle modification among patients with hyperlipidemia. To assess transferability, we collect and annotate a new corpus of 157 notes from 56 clients seen by three health coaches in a separate coaching program. Retraining the SpanQualifier-based (Huang et al., 2023) benchmark models from SMARTSpan (Bojic et al., 2025) and evaluating them on our corpus reveals a sharp performance drop (Table 1), and a t-distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten and Hinton 2008) projection of note embeddings confirms that the two corpora occupy distinct regions of the representation space (Figure 1).

Given the cost of full-scale annotation, a natural alternative is to augment limited in-domain data with synthetic examples. However, surface-level augmentation methods such as paraphrasing or back-translation are insufficient because variation in health coaching notes is not purely lexical—such variation arises from the interaction of health coach-specific documentation *structure*, patient-specific *goal content*, and *persona* context. We propose a corpus-grounded generation pipeline that factorizes these axes, extracts empirical priors from a small set of in-domain notes, and samples from them to produce diverse synthetic notes with em-

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 26–40
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

bedded goal-span labels. A cycle-consistency filter ensures label quality before the synthetic data is used for training.

The contributions of this work are as follows:

1. **Factorized Augmentation Pipeline:** We propose a corpus-grounded generation pipeline that decomposes health coaching note variation into structure, goal content, and persona axes before extracting empirical priors from a small in-domain corpus, and sampling from them to produce diverse synthetic notes with embedded goal-span labels validated via cycle-consistency filtering.
2. **Comprehensive Low-Resource Evaluation:** We conduct systematic experiments across multiple training regimes and ablation settings, demonstrating that factorized generation consistently outperforms general-purpose augmentation baselines and recovers strong span extraction performance from limited in-domain data.
3. **In-Domain Enrichment:** We demonstrate that factorized generation also improves performance when applied within an established corpus, indicating that the pipeline provides value beyond cross-domain adaptation by enriching the training signal even in well-resourced single-domain settings.

2. Method

2.1. Data

Our dataset comprises 157 health coaching sessions collected from 56 unique clients and delivered by three professional health coaches. All notes are written in English. Sessions were conducted as part of a structured health coaching program involving education, medication adherence support, and behavioral goal setting, with clients participating in repeated follow ups over time (mean = 2.82 ± 1.73 sessions per client). Among the 55 clients with available demographic information, 28 (50.9%) were male. The majority were of Chinese ethnicity (n = 43, 78.2%), followed by Malay (n = 5, 9.1%), Indian (n = 3, 5.5%), and other ethnic backgrounds (n = 4, 7.3%). Participants were middle-aged to older adults, ranging from 22 to 77 years (mean = 57.3 ± 10.3), reflecting a population typically targeted for long-term lifestyle modification and chronic disease prevention interventions. Table 2 summarizes key characteristics of our dataset in comparison with `SMARTSpan`, highlighting differences in note length, goal density, and overall structure.

Our 157 notes were annotated for goal spans following guidelines aligned with those of Bojic et al. (2025) including, for instance, removing leading function words such as *to* from span boundaries. The notes were randomly partitioned into three equal subsets and distributed among three annotators such that each note was independently labeled by exactly two annotators. The annotators then met jointly to review all notes on which any disagreement had been identified (47 of 157 notes, affecting 175 of 509 goal-span comparisons) and resolved for each case through discussion until full consensus was reached. Inter-annotator agreement was measured on the dual-annotated notes prior to adjudication: word-level Cohen's $\kappa = 0.94$ overall (0.92 - 0.96 across annotator pairs), and goal-level $\mathrm{F1} = 0.79$ (0.77 - 0.81), indicating substantial agreement on both token-level span boundaries and goal identification. The adjudicated labels serve as the gold standard for all experiments.

2.2. Data Generation

Health coaching notes exhibit variation along three largely independent axes: the *structure and style* imposed by the health coach's documentation habits, the *semantic content* determined by the patient's agreed goals, and the *persona* reflecting the patient's demographic, personal preference, and contextual profile. Surface-level augmentation methods such as paraphrasing and back-translation operate only on the lexical form of existing notes and cannot introduce new patient personas, vary the structural patterns that characterise different coaches, or alter the distribution of goal attributes.

We therefore propose a corpus-grounded generation pipeline that factorizes these axes, extracts empirical priors from the actual corpus, and samples from them to produce diverse, labeled synthetic notes. The pipeline proceeds in two phases: a *corpus analysis* phase that extracts distributional priors along each axis of variation and a *generation* phase that samples from those priors to produce synthetic notes with embedded goal-span labels. An overview of the above process is illustrated in Figure 2.

2.2.1. Corpus Analysis

The corpus analysis phase performs three passes over the real notes, one per axis of variation, to extract the distributional priors that later guide synthetic note generation. All extraction calls use Gemini 3.0 Pro with temperature 0 to ensure deterministic outputs; the full prompts and output schemas are provided in Appendix B.

Goal Attribute Enrichment. Each annotated goal span within the corpus is processed individually to extract a structured set of attributes grounded

	SMARTSpan dataset		Our dataset	
Model	EM F1	PM F1	EM F1	PM F1
BERT-large-cased	90.21 ± 2.29	94.35 ± 2.80	62.39 ± 2.15	74.79 ± 3.41
DeBERTa-v3-large	97.56 ± 2.03	98.27 ± 1.34	71.85 ± 2.38	81.78 ± 3.39

Table 1: Cross-domain evaluation of models trained on SMARTSpan and tested on both datasets. Exact-Match F1 (EM F1) requires predicted and gold spans to match exactly; Partial-Match F1 (PM F1) awards credit proportional to token overlap. Results report mean ± standard deviation over five random data splits.

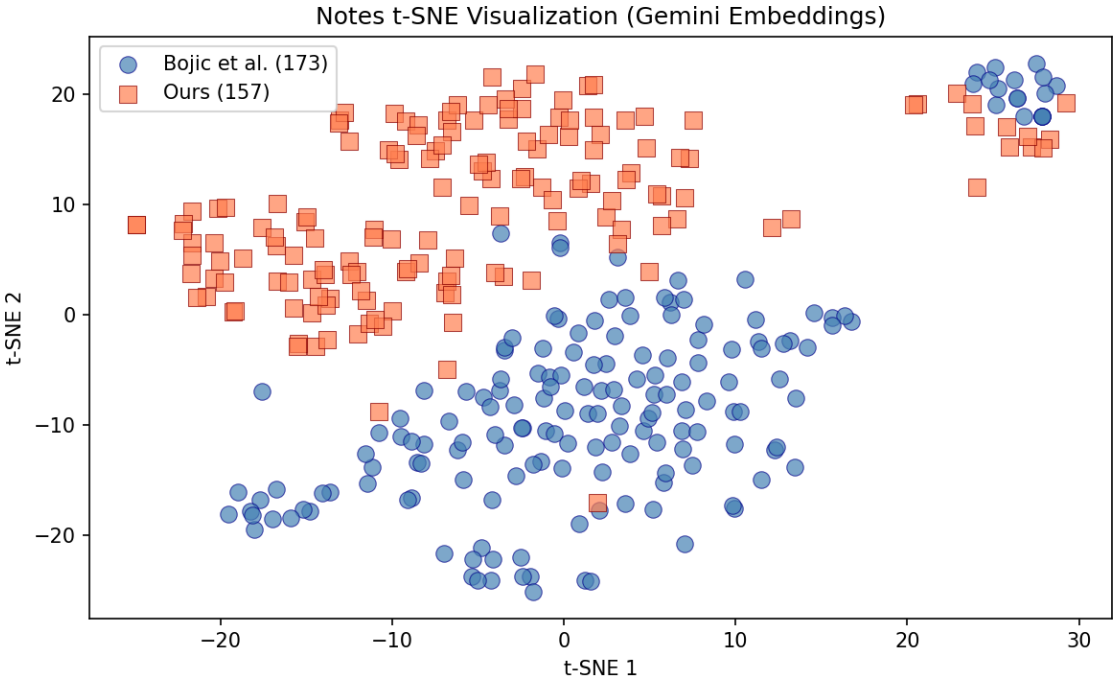

Figure 1: t-SNE visualization of embeddings comparing the SMARTSpan dataset (Bojic et al., 2025) ($n = 173$) and our collected dataset ($n = 157$). The distinct clustering highlights a significant domain shift that explains the performance degradation of the baseline models.

	SMARTSpan	Our dataset
# notes	173	157
# health coaches	3	3
# goals	266	331
Goals per note (avg.)	1.5	2.1
Max goals per note	5	7
Goals per note (90th pct.)	3	4
Zero-goal notes	46 (26.6%)	41 (26.1%)
Words per note (avg.)	161	469
Characters per note (avg.)	1,052	2,678

Table 2: Dataset summary statistics for the SMARTSpan and our datasets.

Goal Type	Description
Physical activity	Exercise or movement goals
Dietary behavior	Nutrition and eating goals
Med. adherence	Taking prescribed medication
Substance use	Reducing substance use
Self-monitoring	Tracking health behaviors

Table 3: Goal type categories, following the taxonomy of Hessler et al. (2019).

in the behavioral goal taxonomy of Hessler et al. (2019). The whole note is used as context for each goal's extraction. Each span is assigned a goal type from the five categories listed in Table 3, together with the specific activity or behavior referenced and, where present, optional attributes: location, time of day, frequency (value and unit), duration (value and unit), timeframe (type and text), conditionality, and patient confidence (type and value). Notes containing no goals receive a separate classification by reason (Table 4). Corpus-wide frequency distributions over all attribute values form the prior from which goal configurations are sampled during generation.

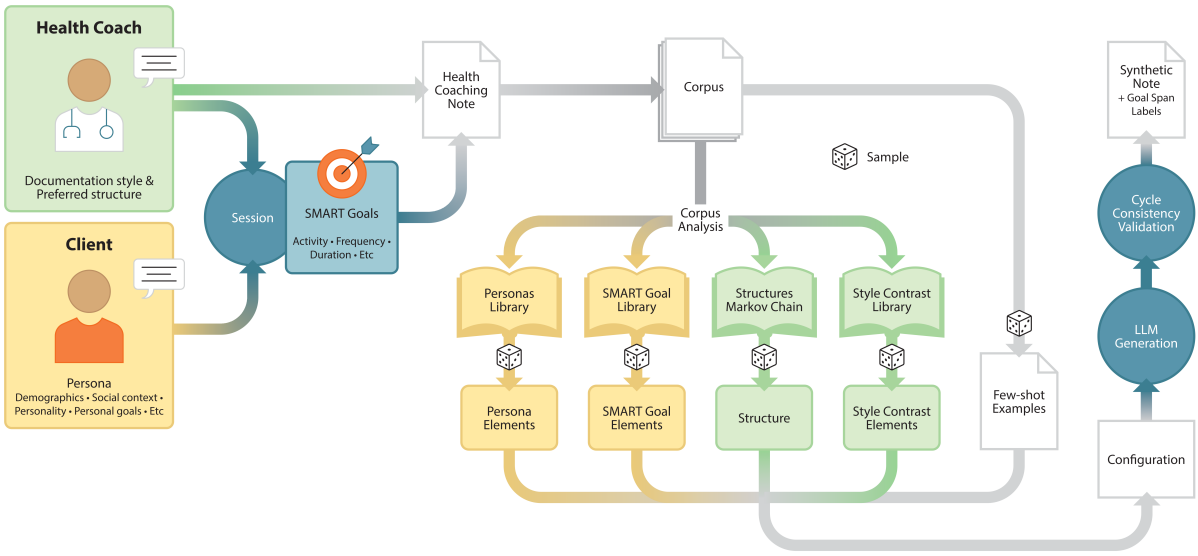

Figure 2: Overview of the factorized synthetic data generation pipeline. The process begins with a corpus analysis phase that decomposes real health coaching notes into independent axes of variation: client persona, SMART goal content, documentation structure, and stylistic contrast. Empirical priors are extracted from the corpus to populate respective libraries. During the generation phase, elements are independently sampled from these libraries to assemble a target configuration. A Large Language Model uses this configuration, alongside few-shot examples drawn from the corpus, to produce a synthetic note with embedded goal span labels. Finally, a cycle-consistency validation step filters the output to ensure label quality.

Reason	Description
No-show	Patient did not attend the session
Declined	Patient declined to set goals
Review-only	Session focused on reviewing goals
Other	Other administrative reasons

Table 4: Zero-goal note classifications.

Structural Analysis. To model the health coach structure axis, each note is decomposed into an ordered sequence of functional sections. Section types are drawn from a closed vocabulary of nine categories defined in consultation with certified health coaches based on the session checklist used in clinical practice: weekly goals review, long-term goals review, app usage review, journaling review, goal setting, barriers discussion, patient information, generative moment (collaborative exploration of values or new directions), and other. Each section is additionally labeled with its format (numbered list, unordered list, or embedded narrative), narrative voice (third person, first person, coach-we, passive, or mixed), and dominant tense pattern (primarily past, primarily future, or mixed). Post-processing removes consecutive duplicate section types.

From the resulting sequences, we construct a first-order Markov chain over section types to model how health coaches habitually organise their notes. Let $\mathcal{S} = \{s_1, \ldots, s_9\}$ denote the nine section types and let END denote a terminal state. The chain comprises an initial-state distribution $\pi_0(s) = P(S_1 = s)$, a transition matrix $P(s' \mid s)$ for $s, s' \in \mathcal{S} \cup \{\text{END}\}$, and an empirical sequence-length distribution $P_L(\ell)$ estimated from the observed number of sections per note. Transition probabilities are smoothed using interpolated Kneser-Ney smoothing (Kneser and Ney, 1995) with a discount factor $d = 0.75$. At generation time, the target sequence length ℓ is first drawn from P_L, then the chain is forward-stepped for exactly ℓ transitions starting from a section type sampled from π_0.

Persona Extraction. To model the patient axis beyond goal content, each note is processed to extract a structured set of persona elements derived from commonly used definitions of sociodemographic factors (National Library of Medicine, 2024) and Social Determinants of Health as defined by the World Health Organization (World Health Organization, 2019). Social demographics are extracted as structured fields: age group, gender, ethnicity, occupation, education level, marital status, and income level. Four categories of contextual elements are extracted as free-text lists using snake_case normalization to facilitate deduplication and frequency counting: *social context* (e.g. `lives_with_partner`), *economic context* (e.g. `late_shifts`), *physical environment* (e.g.

`access_to_gym`), and *individual characteristics* (e.g. `routine_dependent`). Results are aggregated into a persona library recording unique element sets and corpus-wide frequency distributions. The compositional structure of this library allows new patient descriptions to be assembled element-by-element rather than copied from any real record, supporting both diversity and privacy preservation.

2.2.2. Style Contrast Library

Health coach-level stylistic variation is captured through a library of paired real and synthetic notes ranked by stylistic divergence. A diverse subset of real notes is selected from the corpus via stratified round-robin sampling over *(health coach, goal count bucket)* strata, maximizing coverage across coaches and session types. For each selected note, a synthetic twin is generated by Gemini 3.0 Pro (T=1.0) using a configuration constructed directly from that note's own extracted attributes (goals, structure, persona) rather than sampled from corpus distributions. This grounds each twin to its real counterpart, so that embedding distance between the pair predominantly reflects stylistic divergence rather than content divergence. We note that information lost during attribute extraction introduces some residual content variation; however, this information bottleneck is consistent across all pairs, so the resulting noise is approximately uniform and does not systematically bias distance-weighted sampling toward content-divergent pairs.

Both notes are embedded using Gemini Embedding (Lee et al., 2025) and cosine distance is computed per pair. During generation, a style contrast pair is selected from the library with probability proportional to its cosine distance, providing the generation model with a concrete positive (real) and negative (synthetic) stylistic anchor.

2.2.3. Synthetic Note Generation

Each synthetic note is produced through a structured sampling and prompting procedure.

Configuration Sampling. For each note to generate, a full configuration is assembled by sampling independently from the corpus-derived priors along each axis.

Structure. The target number of sections ℓ is drawn from the empirical length distribution P_L. An initial section type is sampled from π_0, and the chain is forward-stepped for ℓ transitions using the Kneser-Ney smoothed transition matrix.

Goals. The number of goals is sampled from the corpus distribution of goal counts. For each goal, attributes are sampled sequentially: goal type from the marginal corpus distribution, activity from the type-conditioned activity list, and each optional attribute a (frequency, duration, timeframe, conditionality, confidence) sampled with null probability

$$P(a = \varnothing) = \hat{p}_{\varnothing}^{(a)} = \frac{|\{g \in \mathcal{G} : a_g \text{ is absent}\}|}{|\mathcal{G}|}, \quad (1)$$

where $\mathcal{G}$ is the set of all annotated goals in the corpus. Conditional on $a \neq \varnothing$, the value is drawn from the empirical distribution of non-null values for that attribute, preserving the realistic sparsity observed in the corpus.

Persona. Social demographic fields are sampled from corpus distributions, each with a per-field null probability estimated analogously to Equation 1. List-type persona categories are populated by first drawing a count from the empirical count distribution for that category, then sampling that many elements without replacement from the category-specific element library.

Few-shot examples. Five few-shot examples are drawn from the corpus using a tiered filter relaxation strategy that prioritizes stylistic similarity while ensuring availability: selection first attempts to match on health coach identity, then relaxes successively to note format, goal count, zero-goal reason, and finally random selection.

Style contrast. A single style contrast pair is injected into the prompt, selected from the precomputed library with probability proportional to the cosine distance between the real note and the synthetic note.

Prompt Construction and Generation. The sampled configuration is assembled into a structured prompt and passed to Gemini 3.0 Pro (T=1.0). Separate templates are used for goal-bearing and zero-goal notes; the latter omits persona and structure sections, which are largely absent from no-show or administrative records. The model returns the full note text together with goal spans as exact substrings in a single structured output call, eliminating the need for post-hoc span alignment.

2.2.4. Cycle-Consistency Validation

Generated notes are filtered for label quality using a cycle-consistency criterion inspired by the forward-backward consistency checks common in unsupervised learning (Zhu et al., 2017). A separate Gemini 3.0 Pro call independently re-extracts goal spans from each generated note with no access to the generation labels. Re-extracted spans are compared against the generated labels using token-level F1: for each generated goal span g_i and its best-matching re-extracted span $\hat{g}_i$, we compute

$$\mathrm{F1}(g_i, \hat{g}_i) = \frac{2\,|\,\mathrm{tok}(g_i) \cap \mathrm{tok}(\hat{g}_i)\,|}{|\mathrm{tok}(g_i)| + |\mathrm{tok}(\hat{g}_i)|}, \quad (2)$$

where $\mathrm{tok}(\cdot)$ denotes the token set of a span. The note-level score is the mean F1 across all generated goals.

To account for noise in the re-extraction model itself rather than genuine label errors, up to three re-extraction trials are performed at $T{=}0.7$ and the best score across trials is retained. A note is accepted only if this best score reaches a threshold of 1.0, i.e. at least one re-extraction trial recovers all generated labels exactly at the token level, without adding or removing goals or text in any of the goals. This strict criterion reflects the fact that training downstream models on partially incorrect span labels introduces systematic noise into the learned representations.

Under this protocol, approximately 20% of generated notes are rejected. Failed notes are saved with their validation metadata and can be rechecked under revised configurations without discarding already-accepted examples. The generation loop continues until a target number of accepted notes is reached.

2.3. Experimental Setting

Data Splits. From our 157-note corpus we randomly sample 57 notes for training and reserve the remaining 100 as a held-out test set. The 57-note training set is deliberately small to simulate the realistic low-resource scenario in which a new health coaching program (or an existing program undergoing severe distribution shift after, for instance, onboarding several new health coaches) can only afford limited annotation effort. The full `SMARTSpan` corpus (173 notes) serves as the out-of-domain training resource. For experiments evaluated on the `SMARTSpan` test set (Table 6), we follow a 60/20/20 train/validation/test split over the 173 notes, repeated five times with different random partitions. Synthetic notes whose source examples overlap with the test fold are excluded from training and validation to prevent data leakage.

Augmentation Conditions. We compare three augmentation strategies applied to the 57 in-domain notes:

- **Rephrasing:** each note is rewritten by Gemini 3.0 Pro with instructions to paraphrase the text while jointly paraphrasing and extracting the goal spans.
- **Backtranslation:** each note is translated from English to a language selected at random from a list comprised of all languages that appear in more than 1% of the available text on the internet (Statista, 2025), and back using Gemini 3.0 Pro, with a post-processing step that realigns goal-span boundaries to the backtranslated text via fuzzy string matching.
- **Ours (factorized generation):** synthetic notes are generated following the pipeline described in Section 2.2, using the 57 in-domain notes as the analysis corpus.

Each method produces 200 synthetic notes per condition. When augmentation is applied to the `SMARTSpan` corpus, the same procedure is used with `SMARTSpan` as the analysis corpus. Figure 3 provides an overview of the dataset composition across all experimental conditions.

Training Configurations. Table 5 reports results under two regimes. In the *low-resource* regime, only the 57 in-domain notes (with or without augmentation) are used for training. In the *combined* regime, the full `SMARTSpan` corpus is concatenated with the 57 in-domain notes before augmentation is applied to the in-domain subset. An additional *cross-domain baseline* trains exclusively on `SMARTSpan` and evaluates on our test set, quantifying the domain shift reported in Introduction.

Models and Training. Following Bojic et al. (2025), we fine-tune two pre-trained encoders: `BERT-large-cased` (Devlin et al., 2019) and `DeBERTa-v3-base` (He et al., 2020) with a token classification head for BIO-tagged goal span extraction. All models were trained for 100 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017), with a learning rate of 3×10^{-5}, a training batch size of 32, an evaluation batch size of 1, and gradient accumulation over 4 steps. Input sequences were truncated to a maximum length of 512 tokens, and experiments were conducted with a fixed random seed of 30.

Evaluation Metrics. We report two span-level F1 scores: *Exact-Match F1* (EM F1), which requires predicted and gold spans to match exactly, and *Partial-Match F1* (PM F1), which awards credit proportional to the token overlap between predicted and gold spans Li et al. (2022). Both metrics are computed at the note level and averaged across the test set.

Statistical Protocol. All experiments are repeated over five random splits to account for variance due to data partitioning. We report the mean and standard deviation of each metric across the five runs as *mean* ± *std*.

Human Distinguishability Test. To assess the perceptual realism of the generated notes, we conducted a two-alternative forced-choice experiment in which human evaluators were asked to distinguish real notes from synthetic ones. In each trial, one genuine note sampled from our corpus and

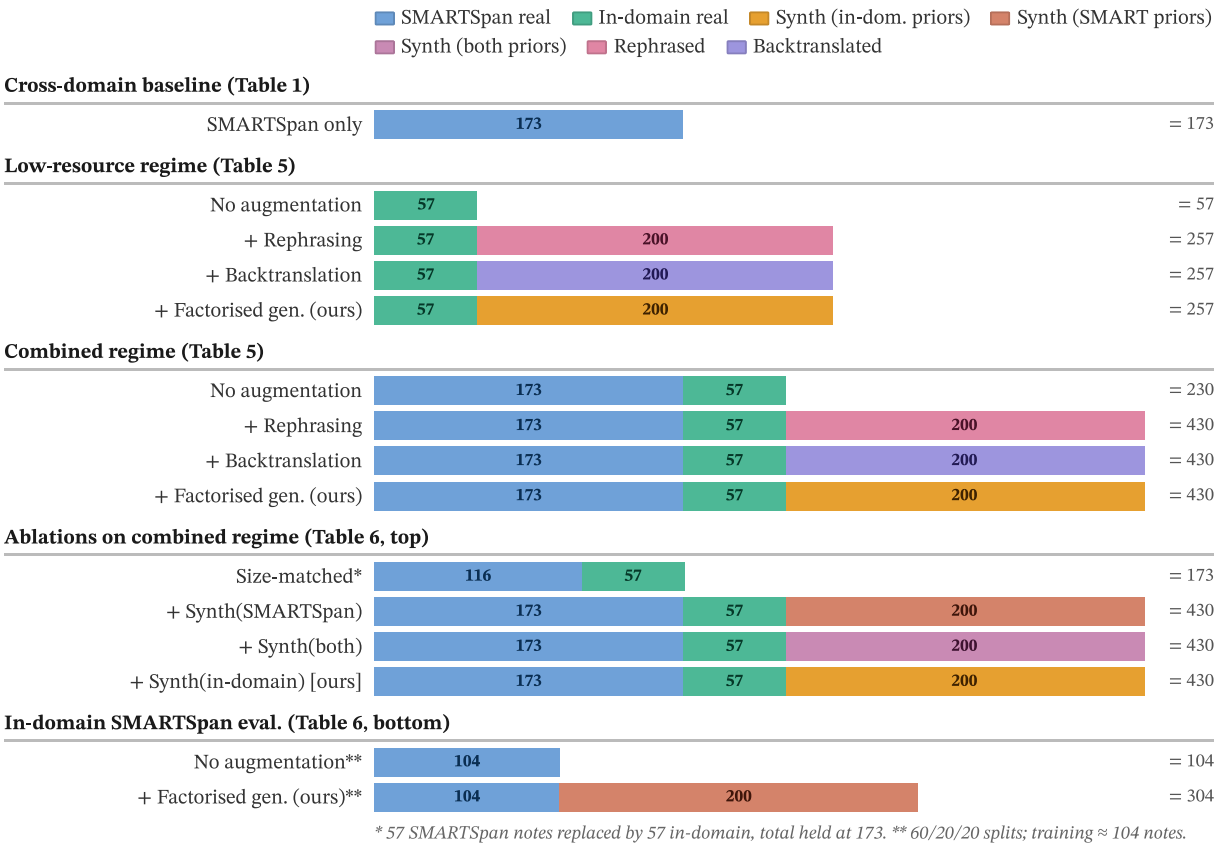

Figure 3: Composition of training sets across all experimental conditions. Each bar shows the number and source of notes used for training. Synth(in-domain) and Synth(S) denote synthetic notes generated from in-domain and SMARTSpan priors, respectively.

one synthetic note produced by the factorized generation pipeline were presented side by side in random order, and the evaluator was asked to select the note they believed to be real. Two evaluators, both familiar with health coaching documentation, completed a total of 50 trials (25 each). We report detection accuracy as the proportion of trials in which the evaluator correctly identified the real note, with chance performance at 50%.

3. Results

Domain shift confirms the need for adaptation. Training exclusively on SMARTSpan and evaluating on our corpus yields EM F1 scores of 62.39 (BERT) and 71.85 (DeBERTa), well below the in-domain performance reported by Bojic et al. (2025). This cross-domain gap, consistent with the distributional divergence visualized in Figure 1, motivates all subsequent experiments.

Factorized generation is the strongest augmentation in low-resource settings. With only 57 in-domain notes and no external data, our pipeline lifts EM F1 from 75.28 to 80.85 (+5.57) for BERT and from 82.79 to 87.30 (+4.51) for DeBERTa (Table 5, low-resource group). Both surface-level baselines also improve over the unaugmented condition, but neither matches factorized generation: backtranslation reaches 79.24 and 80.49, while rephrasing reaches 78.02 and 83.39 for BERT and DeBERTa, respectively. The advantage is most pronounced on PM F1, where our method achieves 90.01 and 96.08 compared to 87.16 and 91.60 without augmentation, suggesting that the factorized sampling introduces greater diversity in span boundaries.

Augmentation benefits persist when out-of-domain data is available. Adding SMARTSpan to the 57 in-domain notes (Table 5, combined group) raises unaugmented EM F1 to 79.99 (BERT) and 84.13 (DeBERTa), confirming that out-of-domain data provides a useful prior despite the domain shift. Applying factorized generation to the in-domain subset further improves performance to 82.38 and 85.01, again outperforming both backtranslation (82.29, 84.89) and rephrasing (78.99, 82.62). Notably, rephrasing slightly degrades performance relative to the unaugmented combined baseline for both models, likely because lexical variation alone cannot compensate for the additional noise introduced by imperfect span realignment.

Training data	BERT EM F1	BERT PM F1	DeBERTa EM F1	DeBERTa PM F1
Cross-domain baseline				
SMARTSpan only	62.39±2.15	74.79±3.41	71.85±2.38	81.78±3.39
Low-resource (57 in-domain notes)				
No augmentation	75.28±3.26	87.16±0.91	82.79±3.81	91.60±3.83
+ Rephrasing	78.02±1.65	85.66±1.77	83.39±4.36	91.24±1.84
+ Backtranslation	79.24±1.89	87.16±2.24	80.49±2.34	90.52±2.64
+ Factorised generation (ours)	**80.85±1.85**	**90.01±1.38**	**87.30±1.27**	**96.08±1.48**
Combined (SMARTSpan + 57 in-domain notes)				
No augmentation	79.99±2.90	90.41±2.25	84.13±0.88	93.56±0.79
+ Rephrasing	78.99±1.91	87.89±2.14	82.62±2.64	92.62±1.52
+ Backtranslation	82.29±1.91	90.77±1.81	84.89±2.52	**94.41±2.10**
+ Factorised generation (ours)	**82.38±1.33**	**91.45±1.74**	**85.01±1.78**	94.14±1.43

Table 5: Goal span extraction performance on our held-out test set ($n = 100$). All models are evaluated on the same 100 in-domain test notes. Results report mean ± standard deviation over five random data splits. Best result per column within each group is reported in **bold**.

Configuration	BERT EM F1	BERT PM F1	DeBERTa EM F1	DeBERTa PM F1
Ablations (tested on our held-out set)				
SMARTSpan + 57 ID (size-matched)*	81.31±3.81	90.94±1.10	84.76±2.34	94.08±1.40
SMARTSpan + Synth(S)† + 57 ID	77.86±3.86	89.63±2.05	82.13±2.62	93.12±1.33
SMARTSpan + 57 ID + Synth(both)‡	80.88±1.44	91.14±1.19	84.49±1.44	93.82±2.14
SMARTSpan + 57 ID + Synth(ID) *[ours]*	**82.38±1.33**	**91.45±1.74**	**85.01±1.78**	**94.14±1.43**
In-domain SMARTSpan (tested on SMARTSpan)				
No augmentation	90.21±2.29	94.35±2.80	**97.56±2.03**	98.27±1.34
+ Factorised generation (ours)	**91.44±3.00**	**95.70±2.22**	97.08±2.01	**98.43±1.50**

* 57 SMARTSpan notes replaced with 57 ID notes (total training set size held constant).
† Synth(S) = factorised generation applied to SMARTSpan notes only. ‡ Synth(ID) = applied to the 57 ID notes; Synth(both) = applied to both corpora.

Table 6: Ablation studies and in-domain (ID) SMARTSpan evaluation. *Top:* variations on the combined training regime, tested on our held-out set ($n = 100$). *Bottom:* models trained and tested on SMARTSpan using five random 60/20/20 splits, with synthetic examples whose source IDs overlap the test fold excluded from training. Results report mean ± standard deviation. Best result per column within each group is reported in **bold**. ID = in-domain.

Augmenting out-of-domain data does not help. The ablation study in Table 6 reveals that applying factorized generation to the SMARTSpan corpus rather than the in-domain notes ("Synth(S) + 57 in-domain") hurts EM F1 by 4.52 (BERT) and 2.88 (DeBERTa) compared to augmenting the in-domain data. Augmenting both corpora simultaneously ("Synth(both)") recovers some of this loss but still under-performs in-domain-only augmentation, suggesting that synthetic SMARTSpan notes amplify out-of-domain patterns that compete with the target distribution. The replacement ablation, which swaps 57 SMARTSpan notes for in-domain examples to control for training set size, performs comparably to the unaugmented combined condition, confirming that the improvement from augmentation is not merely a dataset-size effect.

Factorized generation also improves in-domain performance on SMARTSpan. When both training and evaluation are conducted within SMARTSpan (Table 6, bottom), adding synthetic data generated from the same corpus lifts EM F1 from 90.21 to 91.44 for BERT. For DeBERTa, performance remains near saturation, with EM F1 at 97.08 vs. 97.56 without augmentation. This demonstrates that the pipeline provides value beyond cross-domain adaptation: it also enriches the training in better-resourced, single-domain settings.

Synthetic notes are difficult for humans to distinguish from real ones. Across 50 trials, evaluators correctly identified the real note with an overall accuracy of 72%. Notably, the evaluators showed little to no preference in terms of reported quality between genuine notes (average quality score: 4.08/5) and synthetic quality (average quality score:

4.06/5). While above chance, this detection rate was reported as largely attributable to a small set of surface-level artifacts rather than to global stylistic differences. In post-evaluation debriefing, both evaluators independently reported relying on the same telltale cues to identify *real* notes: references to named clinics, general practitioners, or community resource centers specific to the local healthcare system; mention of exact appointment dates; the presence of spelling errors and locally conventional abbreviations that the language model did not reproduce; frequent use of clinical shorthand such as *pt*, *hx*, and abbreviated day names; and idiosyncratic formatting habits of individual health coaches. Conversely, the use of em dashes was noted as a recurring indicator of synthetic origin. Notably, evaluators reported that when these surface markers were absent, the two notes were perceptually indistinguishable, suggesting that the factorized pipeline captures the higher-level structure, tone, and goal content of real notes with high fidelity.

4. Discussion

The advantage of factorized generation stems from an inductive bias that mirrors how notes are actually produced: by sampling structure, goal content, and persona independently, the pipeline generates plausible configurations never observed in the seed corpus, a combinatorial diversity that lexical methods like paraphrasing and backtranslation cannot achieve. The disproportionate PM F1 gains suggest this diversity is especially beneficial at span boundaries, where the surrounding context varies more widely across synthetic notes than across paraphrases of the same original.

Critically, the ablation results show that the pipeline acts as a distribution-faithful amplifier rather than a general-purpose data multiplier: augmenting the in-domain seed set enriches the target distribution, whereas augmenting `SMARTSpan` amplifies competing out-of-domain patterns. Practitioners should therefore direct augmentation exclusively at the corpus whose distribution they wish to reinforce, even when out-of-domain data is available for joint training. The human evaluation further confirms that the remaining realism gap is largely cosmetic, driven by missing local abbreviations, absence of spelling errors, and formatting artifacts rather than structural or content flaws, and could likely be narrowed by injecting realistic surface noise during post-processing.

5. Conclusion

We presented a factorized synthetic data generation pipeline that decomposes health coaching note variation into structure, goal content, and persona axes, extracts empirical priors from a small in-domain corpus, and samples from them to produce diverse, label-validated training examples. In low-resource adaptation experiments the pipeline consistently outperformed rephrasing and backtranslation baselines, and ablation analysis showed that targeting augmentation at in-domain data is critical, augmenting out-of-domain data amplifies the wrong distribution. A human evaluation confirmed that synthetic notes are structurally and tonally faithful, with detection driven by surface artifacts rather than content or organizational flaws.

Future work will explore relaxing the cycle-consistency threshold with confidence-weighted training to retain a broader set of synthetic examples, extending the pipeline to multilingual coaching programs, and applying the factorization principle to other forms of structured clinical documentation.

6. Limitations

Several limitations should be noted. First, the pipeline has been evaluated on a single downstream task (goal extraction) with two encoder architectures; its benefit for other information extraction objectives or generative models remains untested. Second, both generation and cycle-consistency validation depend on a single proprietary model (Gemini 3.0 Pro), meaning that systematic biases in that model propagate into the synthetic corpus; whether open-source alternatives achieve comparable factorised generation quality remains to be tested. The strict cycle-consistency threshold mitigates label noise but may preferentially accept simpler notes, subtly skewing the training distribution toward shorter or fewer-goal examples. Third, our new corpus originates from a single coaching program with a predominantly Chinese Singaporean demographic; generalizability to programs with different populations, languages, or documentation conventions has not been established. Finally, the 512-token input limit approaches the mean length of our notes (469 words), and longer notes may lose information at the tail, a constraint that would be alleviated by long-context encoder models.

7. Ethics Statement

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).

9. Bibliographical References

Zeynab Bahrami, Atena Heidari, and Jacquelyn Cranney. 2022. Applying smart goal intervention leads to greater goal attainment, need satisfaction and positive affect. *International Journal of Mental Health Promotion*, 24(6).

Iva Bojic, Qi Chwen Ong, Stephanie Hilary Xinyi Ma, Lin Ai, Zheng Liu, Ziwei Gong, Julia Hirschberg, Andy Hau Yan Ho, and Andy WH Khong. 2025. 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.

Julia Bowman, Lise Mogensen, Elisabeth Marsland, and Natasha Lannin. 2015. The development, content validity and inter-rater reliability of the smart-goal evaluation method: A standardised method for evaluating clinical goals. *Australian occupational therapy journal*, 62(6):420–427.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the ACL: human language technologies, volume 1*, pages 4171–4186.

George T Doran. 1981. There's a smart way to write managements's goals and objectives. *Management review*, 70(11).

Susan A Flocke and Kurt C Stange. 2004. Direct observation and patient recall of health behavior advice. *Preventive medicine*, 38(3):343–349.

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*, pages 276–289.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Danielle M Hessler, Lawrence Fisher, Vicky Bowyer, L Miriam Dickinson, Bonnie T Jortberg, Bethany Kwan, Douglas H Fernald, Matt Simpson, and W Perry Dickinson. 2019. Self-management support for chronic disease in primary care: frequency of patient self-management problems and patient reported priorities, and alignment with ultimate behavior goal selection. *BMC family practice*, 20(1):120.

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 (CIKM)*, pages 874–884.

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.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Croía Loughnane, Justin Laiti, Róisín O'Donovan, and Pádraic J Dunne. 2025. Systematic review exploring human, ai, and hybrid health coaching in digital health interventions: trends, engagement, and lifestyle outcomes. *Frontiers in Digital Health*, 7:1536416.

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.

National Library of Medicine. 2024. Medical subject headings (mesh) descriptor: Sociodemographic factors. Accessed: 2026-02-19.

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.

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*.

Statista. 2025. Most common languages used on the internet as of january 2025, by share of websites. Accessed: 2026-02-19.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Anne M Wallace, Matthew T Bogard, and Susan M Zbikowski. 2018. Intrapersonal variation in goal setting and achievement in health coaching: cross-sectional retrospective analysis. *Journal of Medical Internet Research*, 20(1):e32.

Nicole D White, Vicki Bautista, Thomas Lenz, and Amy Cosimano. 2020. Using the smart-est goals in lifestyle medicine prescription. *American journal of lifestyle medicine*, 14(3):271–273.

Ruth Q Wolever, Leigh Ann Simmons, Gary A Sforzo, Diana Dill, Miranda Kaye, Elizabeth M Bechard, Mary Elaine Southard, Mary Kennedy, Justine Vosloo, and Nancy Yang. 2013. A systematic review of the literature on health and wellness coaching: defining a key behavioral intervention in healthcare. *Global advances in health and medicine*, 2(4):38–57.

World Health Organization. 2019. Self-care interventions for health: global evidence and recommendations. Technical report. Accessed: 2026-02-19.

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*, pages 5367–5378.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, and Nikolaos Agadakos. 2024. Modeling low-resource health coaching dialogues via neuro-symbolic goal summarization and text-units-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11498–11509.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A. Example Health Coaching Notes

Figures 4 and 5 present two real and two synthetically generated health coaching notes, respectively. Goal spans are highlighted in green. The synthetic notes were produced by the factorised generation pipeline, sampling structure, goal attributes, and persona independently from the empirical priors extracted during corpus analysis (Section 2.2.1).

B. Corpus Analysis Prompts and Schemas

All corpus analysis calls (Section 2.2.1) use Gemini 3.0 Pro with temperature 0 and Pydantic-enforced structured generation. Figures 6–8 reproduce the system prompt, user prompt template, and output schema for each analysis axis.

29 Dec 2025 (1st session). Current goals: Exercise: Jogging 30 minutes every Saturday around 5pm at the park near his house. 6/10 confident. Medication: To take medication regularly. Morning and night. 9/10 confident. Wellness Vision: He hopes to maintain his current levels of health, in terms of blood pressure, cholesterol and blood glucose. He does not want to develop complications or stroke due to the previous mention conditions. He is afraid that he will suffer and have his quality of life affected. 3 month goal To lose weight. Notes: Deals with company operations. Works normal working hours. Married with 2 children. One of his son is attending health coaching sessions. He is doubtful about the effectiveness of health coaching as he does not believe one would have the luxury of time to make lifestyle changes. He has high blood pressure and cholesterol. There are currently within target as he is compliant with medication. He had previously attempted lifestyle changes like diet and exercise. While they did reduce his cholesterol levels, it was not enough. He therefore agreed to start on medication. His original concerns were that once a person starts medication it is for life. He understands the importance of keeping his blood cholesterol and pressure in check. He does not want to get a heart attack or stroke. Getting a serious illness is fine as long as he passes away straight. He expressed concern about not being able to provide for his family and the suffering one has to go through after getting a stroke. Some changes to his lifestyles were eating only one meal a day and exercising on weekends. As he is now on medication, he admitted that he has not been that consistent with his healthy habits as medication has been keeping his health in check. He was previously maintaining the healthier lifestyle for one to 2 years. While the results were not immediate, he was able to pursue on, and the improvements motivated him to continue being consistent. Improvements were seen in blood test results and reduce clothing sizes. He managed to lose 4kg. He was also previously a smoker and stopped smoking when he developed high blood pressure. What was helpful in today's session was the realization about the accomplishments he had made in the past regarding his lifestyles. That it can be accomplished and how much he has changed.

Session 3 (08/07/2025): (1) Exercise: Continue with muscle strengthening exercises (leg lifts and plants) daily (7/10), (2) Diet: Eat 2 meals a day for 5 days at home (7/10), (3) Diet: No fried chicken (KFC/Tenderbest) for the month, even if there is a good offer (10/10). Preferred name: Sok Leng Wellness vision: Free of new chronic illness, managing weight, being independent, and maintaining physical health to enjoy the things I like. 3m goal: 77kg - 74kg Highlight of the month: Her pipe burst at home, she took 2 weeks to resolve it because she needed to source for plumbers who can fix it at a reasonable price. Due to that, it affected her health goal of cooking at home. Another highlight was a temporary role as an invigilator/usher for an international math olympiad where she was tasked to care for children in Pri 1 - 3. She moved around a lot during those times and felt that her activity levels increased. GR 1: Muscle strenghtening (60%) completion. She was glad that she started the exercises and that had helped with the muscle aches. She is motivated by her parent's experience (where they require support moving around), and would like to work on preventive measures to support mobility. The pipe incident at home did not affect her ability to exercise as she believed that it was a small matter and did not allow it to affect her. GR2: Diet (60%) completion. Her pipe burst incident affected her goal of making home cooked food but she was able to take healthier, home-cooked meals when she was visiting her parents. She shared that there were days where she had KFC fried chicken for her meals because of a bundle deal that was hard to resist. She also shared her appreciation for a deep-fried chinese bun that was sold a few MRT stops from her place, which she buys when there is a sale. Other: In the later part of July, she will be travelling to Shenzhen (till 26th), followed by a university trip to Guizhou (27th onwards). It was a learning trip that she was keen to experience. Action plan: Recommend food videos. Continue to work around adjusting diets - making some progress on this. In discussion, she shared that she feel less aches and pains after trips as compared to the past - this may be due to the terrain, but if she raises it again, perhaps focus on the enabling factors or what was different.

Figure 4: Two real health coaching notes from our corpus. Goal spans are highlighted in green. Top: 2 goals. Bottom: 3 goals.

Session 3 (14/08/2025). Patient is a young adult male, financially stable but finds his current work challenging due to constantly chasing payments and meeting high demands. He is reflective about his health and life direction. Known case of familial hypercholesterolemia, currently managing with medication and herbal supplements. Lives in an area with good access to food stalls and amenities, and has running gear available at home. Wellness Vision: To feel energetic and in control of my body, reducing reliance on long-term medication. 3-month goals: 1. Improve lipid profile naturally 2. Establish a consistent running routine 3. Clean up diet significantly. We discussed his medication adherence; he is compliant but expressed a strong desire to eventually stop taking statins by proving he can manage his cholesterol through lifestyle changes alone. He understands this requires significant dietary discipline. Barriers: Work stress often leads to emotional eating or skipping workouts. He notes that when he is chasing payments, his anxiety spikes, making him reach for comfort foods. Technical issues: He had trouble syncing his wearable device earlier this week but resolved it; energy levels have been fluctuating with work stress. He feels drained by 6 PM most days. Barriers regarding exercise were discussed: previously, rain and fatigue were issues, but he has access to a treadmill now. Goals set for this period: Diet: To consume No chips and no ice cream; substitute with fruits and tea (Confidence 8/10). Diet: No fried chicken (KFC/Tenderbest) 3 per week. Exercise: To do a run on the treadmill / pool walking 1 per day (Confidence 6/10). Regarding diet, I noted that he tends to eat out at hawker centers often. He has been trying to manage his carbohydrate intake by asking for less rice, but admits the gravy makes it hard to gauge calories.

15 Oct 2025 (Session 4). Goals: Physical Activity: Meet 5k steps by jogging and brisk walking. He plans to do this 2 times per day for a total duration of 8 hours during his night shifts. 10/10 confident. This is a strategy to combat the fatigue he feels during the long remote shifts. Diet: Abstain from fried food (e.g. fried chicken, potato chips) in the morning for the next 3 months. 8/10 confident. He noted that after his night shift ends, he often craves greasy breakfast foods with his wife, but wants to break this cycle. Generative moment: The client reflected on his recent interactions with his elderly father, who has limited mobility. Seeing his father struggle to walk has been a wake-up call for him. He realized that his sedentary job and previous lack of activity were leading him down a similar path. He expressed a strong desire to not be reliant on medication like his father is. This emotional connection to his family's health history seems to be a major driver for his current motivation. Longterm goals review: His wellness vision is to maintain a healthy weight and good energy levels so he can enjoy cycling on the weekends without feeling exhausted. He wants to reduce his blood pressure medication dosage within the next year. Patient information: Client will continue to navigate the challenges of working night shifts from home. He will likely face peer pressure from colleagues who suggest ordering fast food during virtual meetings, but he intends to have meal prep ready. He will also be managing the care of his elderly parents on his days off. Other points: He originally joined the program because a nurse at his annual screening flagged his rising blood pressure and weight. He admitted that working from home made him complacent, often sitting for 10-12 hours a day. He is currently experiencing significant fatigue, which he attributes to the shift work and poor diet. App usage review: He will use the app to track his steps during his shift. He will also watch the educational videos on managing shift work sleep patterns to help him get better rest.

Figure 5: Two synthetically generated health coaching notes. Goal spans are highlighted in green. Top: 3 goals. Bottom: 2 goals. Both notes were produced by the factorised generation pipeline.

Goal Attribute Enrichment

PROMPT

System:

```
You are an expert at analyzing health coaching goals. Your task is to extract detailed attributes
from SMART goals (Specific, Measurable, Achievable, Relevant, Time-bound).

Extract attributes precisely from the given goal text. If information is not present, return null
for that field. Be accurate and don't infer information that isn't explicitly stated.
```

User template (formatted per goal span, with full note as context):

```
Goal to analyze: "{goal_text}"

Full context from coaching note:
{note_context}

Extract the following attributes from this goal:
1. goal_type: Type of health behavior goal (physical_activity, dietary_behavior,
medication_adherence, substance_use, self_monitoring)
2. goal_activity: The main activity or behavior (e.g., 'jogging', 'cooking', 'taking medication')
3. location: Where the activity takes place (e.g., 'at home', 'gym', 'park')
4. time: Time of day or specific time (e.g., 'morning', 'evening', 'after work')
5. frequency: How often (value, unit, and full text)
6. duration: How long per session (value, unit, and full text)
7. timeframe: When/for how long (type and full text)
8. conditionality: Any conditional aspects
9. confidence: How confidence is expressed (type, value if numeric, and full text)
10. zero_goal_reason: If "Goal to analyze" is empty (""), select one of:
    no_show | not_discussed_declined | not_discussed_review_only | other

Return null for any attribute not present in the text, EXCEPT for zero_goal_reason which must
always have a value.
```

OUTPUT SCHEMA

Field	Type	Description
goal_type	`Optional[Literal]`	*physical_activity* | *dietary_behavior* | *medication_adherence* | *substance_use* | *self_monitoring*
goal_activity	`Optional[str]`	Main activity or behavior
location	`Optional[str]`	Where the activity takes place
time	`Optional[str]`	Time of day or specific time
frequency_value	`Optional[int]`	Numeric frequency value
frequency_unit	`Optional[Literal]`	*week* | *day* | *month* | *year* | *null* | *other*
frequency_text	`Optional[str]`	Full frequency text as written
duration_value	`Optional[int]`	Numeric duration value
duration_unit	`Optional[Literal]`	*hours* | *minutes* | *days* | *weeks* | *months* | *null* | *other*
duration_text	`Optional[str]`	Full duration text as written
timeframe_type	`Optional[Literal]`	*specific_day* | *date_range* | *ongoing* | *other*
timeframe_text	`Optional[str]`	Full timeframe text
conditionality	`Optional[str]`	Conditional aspects
confidence_type	`Optional[Literal]`	*numeric_percent* | *numeric_scale* | *none*
confidence_value	`Optional[float]`	Numeric confidence value
confidence_text	`Optional[str]`	Full confidence text
zero_goal_reason	`Optional[Literal]`	*no_show* | *not_discussed_declined* | *not_discussed_review_only* | *other*

Goal types follow the taxonomy of Hessler et al. (2019). Schema enforced via Pydantic structured generation.

Figure 6: Prompt and output schema for goal attribute enrichment (Section 2.2.1). Goal types follow the taxonomy of Hessler et al. (2019).

Structural Analysis

PROMPT

System:

```
You are an expert at analyzing the structure and formatting of health coaching notes. Your task
is to identify the structural patterns used in coaching notes.

Extract structural attributes precisely from the given note text. Be accurate and classify based
on what is actually present in the text.

IMPORTANT: Be thorough and granular. Most coaching notes contain many distinct sections
(typically 4-10). Each time the note shifts to a different topic, that is a separate section. Do
NOT lump multiple topics into a single section. For example, if a note discusses medication, then
food, then exercise as separate paragraphs or blocks, each of those is its own section — even if
they lack explicit headers.
```

User template (formatted per note):

```
Coaching note to analyze:
{note_text}

Analyze the structural pattern of this coaching note. Identify EVERY distinct section — a new
section begins whenever the note shifts to a different topic or purpose, whether or not there is
an explicit header.

For each section, extract:
  * "type": weekly_goals_review, longterm_goals_review, app_usage_review, journalling_review,
goal_setting, barriers_discussion, patient_information, generative_moment, or other_points
  * "format": numbered_list, embedded_narrative, or unordered_list
  * "voice": third_person, first_person, coach_we, passive, or mixed
  * "tense_pattern": mixed_past_future, primarily_future, or primarily_past
  * "section_summary": Optional short summary (only for app_usage_review, journalling_review,
longterm_goals_review, weekly_goals_review, other_points)

Important guidelines:
- Be GRANULAR: if the note discusses medication adherence, then diet, then exercise as separate
blocks, those are THREE separate sections — not one combined section.
- The same section type CAN appear multiple times.
- Typical coaching notes have 4-10 sections. If you find only 1-2 sections in a long note, re-
read it more carefully.
- Return empty list ONLY if the note has no clear sections at all.
```

OUTPUT SCHEMA (LIST[SECTION])

Field	Type	Description
type	`Literal`	Section type from checklist *weekly_goals_review* | *longterm_goals_review* | *app_usage_review* | *journalling_review* | *goal_setting* | *barriers_discussion* | *patient_information* | *generative_moment* | *other_points*
format	`Literal`	*numbered_list* | *embedded_narrative* | *unordered_list*
voice	`Literal`	*third_person* | *first_person* | *coach_we* | *passive* | *mixed*
tense_pattern	`Literal`	*mixed_past_future* | *primarily_future* | *primarily_past*
section_summary	`Optional[str]`	Short summary (only for review/other_points types; null otherwise)

Section types derived from the health coaching session checklist used in clinical practice.

Figure 7: Prompt and output schema for structural analysis (Section 2.2.1). Section types are derived from the health coaching session checklist used in clinical practice.

Persona Extraction

PROMPT

System:

```
You are an expert at analyzing health coaching notes to identify persona elements. Your task is
to extract individual persona characteristics present in the note.

Extract persona elements based on what is stated or reasonably inferred from the note. Be
thoughtful in your inferences but avoid being overly creative — stick to what the text supports.
```

User template (formatted per note):

```
Coaching note to analyze:
{note_text}

Extract persona elements present in this note across the following categories:

1. Social demographics:
   - age_group: young_adult (18-35), middle_aged (36-60), senior (60+), or unknown
   - occupation: Current occupation or employment status (free-form)
   - education_level: Education level (free-form)
   - gender: Gender identity (free-form)
   - ethnicity: Ethnicity or racial background (free-form)
   - income_level: Income level or socioeconomic status (free-form)
   - marital_status: Marital status (free-form)

2. Social context: Social relationships, social factors, and social barriers that influence
health behaviors (e.g., lives_with_partner, social_eater, family_dinners, social_triggers)

3. Economic context: Economic and financial factors, employment-related circumstances, and
economic barriers (e.g., cost_concerns, financial_stress, late_shifts, work_demands)

4. Physical environment: Physical environment, living situation, and physical barriers (e.g.,
lives_in_urban_area, access_to_gym, limited_kitchen_space, travel_disruptions)

5. Individual characteristics and behaviours: Behavioral patterns, health literacy, psychological
state, and individual-level barriers (e.g., all_or_nothing_thinker, routine_dependent,
tech_comfortable, time_constraints)

Guidelines:
- Use snake_case naming for all extracted elements
- Return empty lists for categories where no elements are present
- Return null for demographics not available
- When barriers are mentioned, categorize them into the appropriate context category based on
their nature
```

OUTPUT SCHEMA

Field	Type	Description
social_demographics (SocialDemographics)		
age_group	`Optional[Literal]`	*young_adult | middle_aged | senior | unknown*
occupation	`Optional[str]`	Occupation or employment status
education_level	`Optional[str]`	Education level
gender	`Optional[str]`	Gender identity
ethnicity	`Optional[str]`	Ethnicity or racial background
income_level	`Optional[str]`	Income level or socioeconomic status
marital_status	`Optional[str]`	Marital status
Context and behavioural factors (all List[str], snake_case)		
social_context	`List[str]`	Social relationships and social barriers influencing health behaviours
economic_context	`List[str]`	Economic/financial factors and employment-related barriers
physical_environment	`List[str]`	Physical environment, living situation, and environmental barriers
individual_characteristics_and_behaviours	`List[str]`	Behavioural patterns, health literacy, psychological state, individual barriers

Demographic categories informed by MeSH Sociodemographic Factors and WHO self-care frameworks.

Figure 8: Prompt and output schema for persona extraction (Section 2.2.1). Demographic categories informed by MeSH Sociodemographic Factors and WHO self-care frameworks.