SMARTMiner: Extracting and Evaluating SMART Goals from Low-Resource Health Coaching Notes

Iva Bojic[1] **Qi Chwen Ong**[1] **Stephanie Hilary Xinyi Ma**[1] **Lin Ai**[2] **Zheng Liu**[3]
Ziwei Gong[2] **Julia Hirschberg**[2] **Andy Hau Yan Ho**[1] **Andy W. H. Khong**[1]
[1] Nanyang Technological University
[2]Department of Computer Science, Columbia University
[3]Khoury College of Computer Science, Northeastern University
{iva.bojic, qichwen.ong, stephaniehilary.ma, andyhyho, andykhong}@ntu.edu.sg
{lin.ai, sara.ziweigong, julia}@columbia.edu
liu.zheng9@northeastern.edu

Abstract

We present *SMARTMiner*, a framework for extracting and evaluating specific, measurable, attainable, relevant, time-bound (SMART) goals from unstructured health coaching (HC) notes. Developed in response to challenges observed during a clinical trial, the *SMARTMiner* achieves two tasks: (i) extracting behavior change goal spans and (ii) categorizing their SMARTness. We also introduce *SMARTSpan*, the first publicly available dataset of 173 HC notes annotated with 266 goals and SMART attributes. *SMARTMiner* incorporates an extractive goal retriever with a component-wise SMARTness classifier. Experiment results show that extractive models significantly outperformed their generative counterparts in low-resource settings, and that two-stage fine-tuning substantially boosted performance. The SMARTness classifier achieved up to 0.91 SMART F1 score, while the full *SMARTMiner* maintained high end-to-end accuracy. This work bridges healthcare, behavioral science, and natural language processing to support health coaches and clients with structured goal tracking—paving way for automated weekly goal reviews between human-led HC sessions. Both the code and the dataset are available at: https://github.com/IvaBojic/SMARTMiner.

1 Introduction

Health coaching (HC) is a person-centered intervention designed to facilitate sustainable change in healthy behavior and support self-management of chronic diseases. Recent systematic reviews demonstrated that HC can significantly enhance physical activity, reduce pain, and improve psychological outcomes such as self-efficacy and quality of life among individuals with chronic conditions (Weiss et al., 2025).

A cornerstone of HC is the co-creation of actionable short-term goals that drive long-term behavior change, which in turn forms the foundation of its effectiveness as a behavioral intervention (Wallace et al., 2018). Overall, by providing focus and measurable targets, the research suggests that specific and actionable goals tend to improve health behavior outcomes more than unclear or generic behavioral goals (Bahrami et al., 2022).

Specific, measurable, attainable, relevant, and time-bound (SMART) goals (Figure 1) often yield better short-term results and can help sustain behavior change (Doran, 1981; White et al., 2020), as seen in improved exercise levels, weight loss, and self-management behaviors in various studies (Wolever et al., 2010). However, setting and evaluating SMART goals are laborious and complex (Bowman et al., 2015). Some goals may not be fully SMART as they could lack one or more SMARTness components.

This may stem from patient's varying levels of readiness to engage in health-related behaviors (Prochaska and Velicer, 1997). Additionally, patients often struggle to recall goals set during HC sessions (Flocke and Stange, 2004), underscoring the need for improved goal documentation and consistent support between the HC sessions. This is especially important given that HC sessions are typically scheduled biweekly or monthly—intervals that exceed the intended time frame for most goals. As a result, patients often face a multi-week gap in feedback and reinforcement between sessions.

In response, we present a *SMARTMiner* framework that automatically extracts multiple (SMART) goals set during HC sessions from unstructured session notes. Since these goals are embedded within free-text narratives and cannot be audited at scale, we address two key challenges:

- **(i)** ***Goal extraction*** – identifying multiple behavior change goal spans within unstructured HC notes; and
- **(ii)** ***SMARTness classification*** – determining which SMART attributes each extracted goal satisfies and where it falls short.

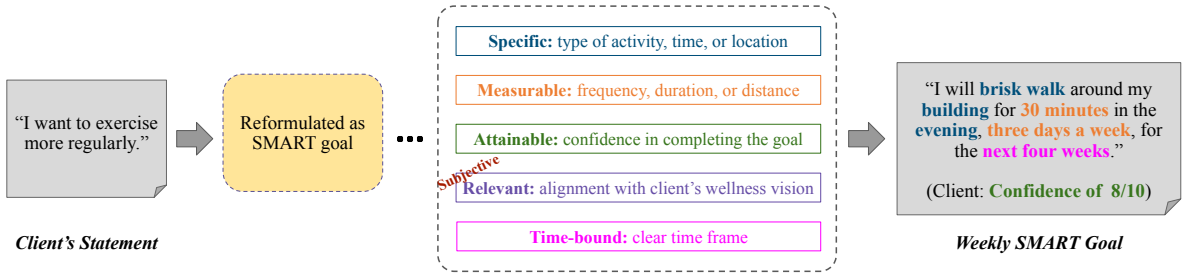

Figure 1: Reformulation of generic behavior goal into a SMART goal.

Our contributions are three-fold:

1. ***SMARTSpan* dataset** – the first publicly available dataset comprising 173 HC notes with 266 unique goal spans, exhaustively annotated for both goal boundaries and SMART attributes.
2. ***SMARTMiner* framework** – a span extractor that locates every potential goal and an attribute-level classifier that flags missing SMART components, yielding actionable, interpretable feedback.
3. **Comprehensive evaluation and analysis** – five-fold in-domain and cross-domain experiments with diverse extractive and generative baselines reveal how low-resource, domain-specific data degrade state-of-the-art large language models (LLMs); a qualitative error taxonomy (e.g., hallucination) pinpoints safety-critical failure modes for clinical deployment.

By combining healthcare, behavioral science, and natural language processing (NLP), the proposed *SMARTMiner* framework enables low-touch (SMART) goal tracking: it helps health coaches refine goals as required and allows clients to review their (SMART) goals between HC sessions.

2 Related Work

2.1 Clinical Notes

Information extraction (IE) from unstructured clinical notes has evolved from rule-based heuristics to transformer-based models capable of capturing a wide range of clinically relevant signals (Pai et al., 2024). In particular, *ClinicalBERT* detects medication gaps (Gobbel et al., 2022), while other models extract lifestyle and social factors (Zhou et al., 2019; Romanowski et al., 2023) or patient goals (Gupta et al., 2021). Beyond tagging, span-based models handle action items (Mullenbach et al., 2021), and temporal models align events chronologically (Miller et al., 2023). Prompt-driven LLMs such as GPT-4 now match or surpass supervised baselines with minimal data (Ramachandran et al., 2023), signaling a shift toward adaptable, low-resource IE solutions.

2.2 HC Conversations

Extracting SMART goals from HC conversations presents challenges at the intersection of clinical NLP and behavioral health (Chen and Hirschberg, 2024). Early approaches utilized rule-based systems and sequence labeling to identify SMART components within HC dialogues (Gupta et al., 2019, 2020b). Subsequent methods incorporated dialogue act modeling and transformer-based architectures to enhance goal extraction accuracy (Gupta et al., 2020a,b; Mullenbach et al., 2021). As opposed to the exploitation of sparsity in systems (Khong and Naylor, 2005), recent studies proposed modularized and neuro-symbolic approaches to enhance goal summarization in low-resource settings, where labeled data is sparse and HC dialogues vary in format and structure (Zhou et al., 2022, 2024). However, the vagueness of conversational language calls for models that identify discontinuous spans and align to goal criteria, motivating span-based frameworks that are grounded in real-world HC interventions.

2.3 Multi-Span Reading Comprehension

Extracting multiple behavior-change goal spans from HC notes can be formulated as a multi-span reading comprehension (MSRC) task. The existing MSRC methods fall into three categories:

- *Extractive* methods select answer spans from the input without generating new tokens and are further divided into *token-based* and *span-based*. Token-based models predict answers through token-level outputs, where each token individually influences span selection (Hu et al., 2019; Yang et al., 2020; Segal et al.,

2020; Li et al., 2022; Luo et al., 2024). In contrast, span-based models explicitly score or classify candidate spans as wholes, considering span-level representations (Huang et al., 2023a; Zhang et al., 2023b, 2024).

- *Generative* methods, on the other hand, produce answers by generating tokens, often extending pre-trained generative models with fine-tuning strategies (Ai et al., 2024) or prompt engineering to adapt to the multi-span scenario (Mallick et al., 2023; Zhang et al., 2023a; Huang et al., 2023b).

- *Hybrid* methods leverage the advantages of both paradigms, either through data augmentation techniques (Lee et al., 2023) or via unified frameworks (Lin et al., 2024).

Most datasets used to evaluate models for MSRC focus on questions requiring the extraction of multiple discontinuous answer spans from text. Prominent benchmarks include *MultiSpanQA* (Li et al., 2022), which contains over 6,500 multi-span questions initially and around 19,000 in the expanded set; *QUOREF*, which comprises more than 24,000 questions requiring coreference resolution (Dasigi et al., 2019); and *DROP*, which includes approximately 96,000 questions involving arithmetic or reasoning over multiple spans (Dua et al., 2019). More specialized benchmark on the healthcare domain, *MASH-QA*, consists of approximately 35,000 question–answer pairs with long, multi-sentence answers (Zhu et al., 2020).

While existing MSRC models perform well on large, clean, and publicly available datasets (e.g., Wikipedia-based or curated medical content), such datasets are expensive and time-consuming to curate in clinical practice, including HC. As a result, span-centric models remain underexplored in low-resource, domain-specific settings. To address this gap, we introduce *SMARTSpan*, a curated dataset of HC notes with annotated behavioral goals, designed to support span-based goal extraction in practical, small-scale scenarios.

3 *The SMARTSpan* Dataset

SMARTSpan is a new dataset comprising 173 annotated HC notes collected from a randomized controlled trial (RCT), designed to evaluate multi-span extraction in low-resource, domain-specific scenarios. Each HC note summarizes real-world goal-setting conversations between clients and health coaches, offering a unique testbed for behavior change modeling. Unlike structured MSRC datasets—where answers in the form of multiple discontinuous spans are extracted in response to specific questions—*SMARTSpan* presents challenges rooted in the complexity of extracting goals that are often diffuse or repeated across different sections of an HC note (details in Appendix A).

3.1 Data Collection

The *SMARTSpan* dataset originates from an RCT aimed at the prevention of cardiovascular disease through a multicomponent digital behavioral intervention, focusing on improving patient's adherence to statin therapy and promoting healthy behavioral change to reduce low-density lipoprotein (LDL) cholesterol levels.

One key component is human-led HC, where intervention participants receive six monthly coaching sessions via a mobile app. Clients are encouraged to set weekly SMART goals during each HC session, which are reviewed in the next session. Although the sessions were conducted in English, some conversations were interjected with Chinese or other dialects. Language detection models (Liu et al., 2022) were not required since after each session, health coaches document key observations and any SMART goals set with clients as unstructured free-text notes, without any given standardized template or guidance, on a web-based platform (details in Appendix A).

At the time of dataset creation, a multiracial cohort of approximately 130 patients with hyperlipidemia had been enrolled in the ongoing RCT, with over 60 in the intervention arm receiving support from three health coaches. More than 180 HC sessions were conducted. Using a custom SQL query on the platform's backend, we retrieved 173 HC notes as the *SMARTSpan* dataset.

Anonymization. Protecting client privacy is essential when handling real-world HC notes with sensitive information. We adhered to an *iterative anonymization process* that combined manual redaction with selective automated support via `GPT-4o`. All direct personal identifiers (e.g., names, IDs) were manually removed before any HC note was processed using `GPT-4o`. The model was used solely to assist with redacting broader contextual details such as references to family members, geographic locations, and temporal markers, while preserving narrative coherence. Following this, two

# of answer spans or goals annotated	0	1	2	3	4–5	6–8	9–12	13–21
#Spans from *MultiSpanQA*	–	–	3,791	1,414	915	337	71	8
#Goals from *SMARTSpan*	46	46	37	32	12	–	–	–

Table 1: Comparison of the number of answer spans per question in *MultiSpanQA* (Li et al., 2022) and the number of goals per HC note in *SMARTSpan* dataset.

human annotators independently reviewed each of the 173 notes to verify full anonymization and ensure that medical metrics (e.g., cholesterol, body weight) and personal attributes (e.g., age, occupation) were appropriately generalized. Multiple review cycles ensured both privacy preservation and data utility for downstream modeling. As a final step before dissemination, the fully anonymized dataset underwent institutional review and was approved for research use and public release.

3.2 Data Annotation and Exploration

After anonymization, the dataset creation process was conducted in two phases.

Goal Annotation. In the first phase, two annotators manually reviewed 173 HC notes and identified goal statements within each HC note, marking any text that reflected specific behavioral objectives discussed or set during HC sessions. As shown in Table 1, the distribution of goals per note ranges from 0 to 5, depending on the depth and focus of the HC session. Notably, 27% of the HC notes contain no goals, and another 27% include only one. Unlike datasets such as *MultiSpanQA*—where each sample has at least one span and most have two or more (58% with 2, 22% with 3), *SMARTSpan* intentionally retains goal-absent HC notes to reflect the variability of real-world HC and to enable models to detect when no goal is present. This sparsity represents a core challenge in adapting existing MSRC techniques to HC scenarios, where relevant spans are infrequent, diverse, and sometimes entirely absent. Handling such cases is essential for reliable deployment in practical HC workflows.

SMARTness Annotation. In the second phase, each extracted goal was annotated for three core SMART components: specific (S), measurable (M), and attainable (A). We excluded relevant (R) due to its subjectivity and assumed time-bound (T) was implicitly satisfied, as goals were intended to be achieved before the next HC session. Each of the three assessed components was annotated as a binary label (0 or 1). The final multiclass label was derived deterministically: *SMART* if all three were true (i.e., S=1, M=1, A=1), *Partially SMART* if two were true, and *Not SMART* if one or none were true. Based on this scheme, out of 266 annotated goals, 113 (42.5%) were *SMART*, 77 (28.9%) were *Partially SMART*, and 76 (28.6%) were *Not SMART*.

goals per HC note	0	1	2	3	4	5	$\sum$ Goals
split_1	5	9	3	7	0	1	41
split_2	4	7	4	8	2	0	47
split_3	4	7	8	2	4	0	45
split_4	7	6	5	4	2	1	41
split_5	12	5	5	3	0	0	24

Table 2: Distribution of the number of goals per HC note and the total number of goals in the *SMARTSpan* test sets across five splits.

To create a labeled dataset for supervised classification, two annotators with prior HC training independently rated the extracted goals. Disagreements were resolved through discussion. Inter-annotator agreement (IAA) was assessed using Cohen's kappa coefficient (Cohen, 1960) for each of the three SMART components across 266 annotated behavior goals. IAA was moderate for S ($\kappa = 0.574$, $z = 9.36$, $p < 0.001$), substantial for M ($\kappa = 0.812$, $z = 13.3$, $p < 0.001$), and near perfect for A ($\kappa = 1.00$, $z = 16.3$, $p < 0.001$). These results indicate statistically significant IAA beyond chance across all three components, with the highest consistency observed in component A. Additional details on annotation process are provided in Appendix B.

Cross-Validation Setup. Given the limited size of *SMARTSpan* (173 annotated HC notes with a total of 266 goals), we adopted a five-fold cross-validation for robust model evaluation, each follows a 70%/15%/15% (train/valid/test) split. Each test and validation sets in splits contain 25 HC notes, with the remaining 123 HC notes used for training. As shown in Table 2, the number of goals per test split varies from 24 to 47 across splits, reflecting differences in goal density and highlighting the importance of evaluating model robustness.

4 Methods

We propose a *SMARTMiner* framework for multi-span behavioral SMART goal extraction from unstructured HC notes, as shown in Figure 2. The *goal extraction* module formulates goal identification as a span-based question answering task, enabling the extraction of multiple goal mentions from free-text HC notes. The *SMARTness classi-*

fication module subsequently evaluates these extracted goals along three key dimensions to determine their alignment with the SMART criteria.

4.1 Goal Extraction Module

Several extractive models were implemented within the goal extraction module by formulating the extraction task as an MSRC problem. Each HC note is treated as the context $C = \{c_1, c_2, \ldots, c_n\}$, and paired with a fixed question Q: *"What are the goals mentioned in the text?"* Although the original dataset includes annotated goal spans, it does not contain varying questions. To enable span-based supervision, we recast the dataset into a QA format using this fixed question. The model is then trained to extract all non-overlapping spans $P = \{p_1, p_2, \ldots, p_t\}$ that correspond to goals such that

$$P = M(C, Q). \tag{1}$$

Here, M denotes the fine-tuned extractive model. For generative models, we fine-tuned them using a prompt template with *instruction*, *input* and *output* (see Appendix C for the exact format).

4.2 SMARTness Classification Module

Classification Objective. Given an extracted goal span p_i, the classifier independently predicts binary values for each component

$$v_i^{(S)},\ v_i^{(M)},\ v_i^{(A)} = \text{Classifier}(p_i), \tag{2}$$

where $v_i^{(d)} \in \{0, 1\}$ for each dimension $d \in \{\text{S}, \text{M}, \text{A}\}$. These binary predictions are obtained via sigmoid-activated heads and trained using binary cross-entropy loss for each dimension. The final multiclass label is assigned post-hoc via rule-based aggregation over the binary predictions for each component. Specifically, we define

$$y_i = f\left(v_i^{(S)} + v_i^{(M)} + v_i^{(A)}\right), \tag{3}$$

where $v_i^{(d)} \in \{0, 1\}$ denotes the binary prediction for SMART dimension $d \in \{\text{S}, \text{M}, \text{A}\}$ and $f(\cdot)$ is a deterministic mapping from component count to structured label. A sum of 3 indicates a *SMART* goal; 2, *Partially SMART*; and 0 or 1, *Not SMART*.

Model Architecture. The classifier is implemented as a transformer-based model. We employed a pre-trained encoder to obtain contextualized representations of the input goal. Specifically, the embedding of the `[CLS]` token is extracted and passed through a shared processing stack consisting of dropout, a linear projection layer, and ReLU activation to generate a latent representation. This representation is then fed into three independent binary classification heads—implemented as linear layers with sigmoid activation—to estimate the probabilities $P^{(S)}$, $P^{(M)}$, and $P^{(A)}$ that the goal is specific, measurable, and attainable, respectively.

Training and Loss. We trained the classifier using binary cross-entropy (BCE) loss independently for each SMART component. The overall training objective is defined as

$$\mathcal{L} = \frac{1}{3} \sum_{d} \text{BCE}(v^{(d)}, y^{(d)}), \tag{4}$$

where $v^{(d)}$ is the predicted probability and $y^{(d)}$ the ground truth label for SMART dimension $d \in \{\text{S}, \text{M}, \text{A}\}$. This setup enables the model to learn each structural dimension independently while supporting interpretable component-level diagnostics.

5 Experiment Setting

Datasets. To evaluate the effectiveness of our proposed *SMARTMiner* framework for multi-span behavioral SMART goal extraction in a low-resource setting, we used the *SMARTSpan* dataset described in Section 3. As no other existing datasets annotates SMART goal spans, we also evaluate on *MultiSpanQA* (Li et al., 2022), a widely used MSRC benchmark whose "find-all-spans" setup mirrors our extraction task. This positions our model against established baselines while making clear the domain shift from open-domain question answering to the HC notes.

Evaluation Metrics. For goal extraction, we adopt both exact match (EM) and partial match (PM) precision, recall, and F1 as the primary evaluation metrics, following Li et al. (2022). For SMARTness classification, we report accuracy, macro-averaged F1, and the class-wise F1 score for the SMART label. These are computed over the three predicted classes: SMART, Partially SMART, and Not SMART.

Goal Extraction Module. We implemented and evaluated both extractive and generative approaches as the goal extraction modules. Since prior research suggests that span-centric methods outperform token-centric ones in multi-span settings Huang et al. (2023a); Zhang et al. (2023b,

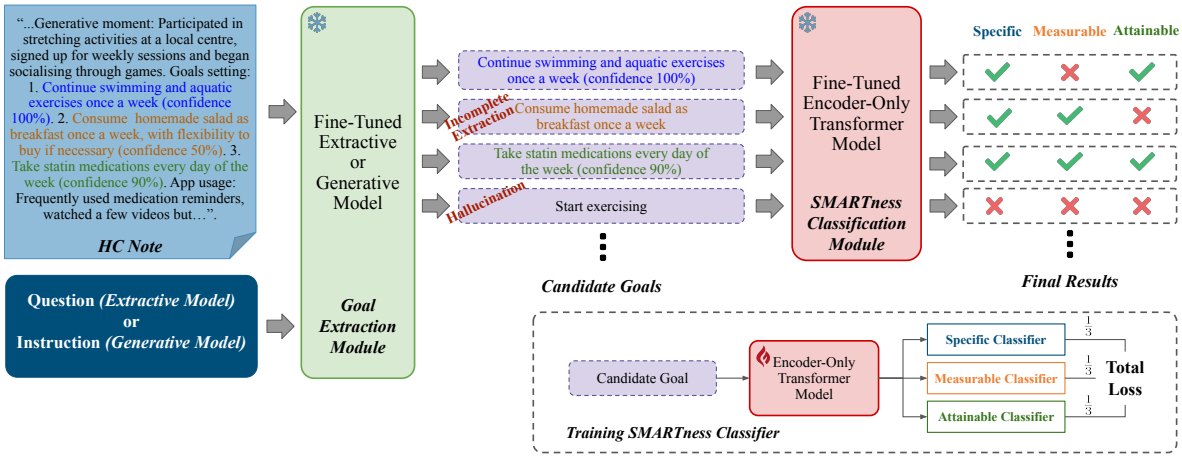

Figure 2: The overall architecture of our proposed *SMARTMiner* framework.

2024), we selected two representative extractive strategies: *SpanQualifier* (Huang et al., 2023a), which scores candidate spans, and the *Contrastive Span Selector (CSS)* (Zhang et al., 2023b), which ranks spans using contrastive learning with positive and negative contextual cues. All models were fine-tuned using a single NVIDIA A40 GPU, with the exception of CSS on *MultiSpanQA*, which was fine-tuned using 8 NVIDIA A40 GPUs.

For *SpanQualifier*, we adopted the same configuration as reported in (Huang et al., 2023a), with the exception of setting the same random seed to ensure consistency across all experiments. Given initially low performance when fine-tuning directly on the in-domain *SMARTSpan* dataset, we also experimented with a two-stage fine-tuning strategy: first pretraining on the *MultiSpanQA* dataset, followed by continued fine-tuning on *SMARTSpan*. We evaluated this approach using two pretrained language models: `BERT-base-uncased` (Devlin et al., 2019) and `DeBERTa-v3-base` (He et al., 2020).

For *CSS*, we adopted the same configuration as reported in (Zhang et al., 2023b), except for setting the random seed to 30 for consistency with other experiments and increasing the number of training epochs to 20. We evaluated this approach using two pretrained language models: `BERT-base-uncased` and `RoBERTa-base` (Liu et al., 2019).

To evaluate the performance of generative language models in low-resource, domain-specific goal extraction tasks, we fine-tuned decoder-only and encoder-decoder architectures (see Appendix D for the full list) using LoRA-based parameter-efficient adaptation (Hu et al., 2021). Fine-tuning was performed using the `Unsloth` library (Unsloth, 2024), which is optimized for reduced memory usage and faster training. We applied LoRA with a rank of 8 and an alpha scaling factor of 32, enabling efficient adaptation of LLMs while maintaining performance. Each pretrained model was fine-tuned on a single GPU for 20 epochs, using a batch size of 4, a weight decay of 0.01, and a learning rate of 1×10^{-4} for encoder–decoder models and 3×10^{-5} for decoder-only models. We employed AdamW with 8-bit optimization, linear learning rate scheduling, mixed-precision training, and gradient checkpointing. Early stopping was used with a patience of 5 valid steps to prevent overfitting. For further details on the learning rate grid search, see Appendix E.

We also adapted the Question-Attended Span Extraction (QASE) framework (Ai et al., 2024) to enable span-level supervision in generative models. QASE is a lightweight question-aware decoder-side supervision approach originally proposed to enhance span alignment in generative settings. While we followed the original hyperparameter settings proposed by the authors, we introduced two key modifications. First, we set the batch size to 4 across all experiments to fit within our GPU memory constraints. Second, given the smaller size of our target dataset, we increased the number of training epochs to 20. For the larger *MultiSpanQA* dataset, however, we retained the original configuration of training for 3 epochs.

Finally, we evaluated a zero-shot, schema-based prompting approach using `GPT-4.1` (OpenAI, 2023) as a non-fine-tuned baseline. The model was guided by structured extraction instructions, using the same instruction template as that for the

fine-tuned generative models, with a single modification: the final sentence was changed from "*Format your response as a numbered list.*" to "*Always respond in JSON format.*", which ensures compatibility with schema-based function calling. The model was deployed via OpenAI's function-calling API to extract weekly SMART goals from unstructured HC notes. SMART goal extraction was treated as a semantic parsing task, with outputs constrained to a predefined JSON schema to enforce structural compliance (see Appendix F).

SMARTness Classification Module. We fine-tuned three transformer-based models: `deberta-v3-base`, `deberta-v3-large`, and `RoBERTa-large`. Each model was trained using a batch size of 4, a maximum sequence length of 64, and a learning rate of 2×10^{-5}. Optimization was performed using AdamW with a weight decay of 0.01. Training proceeded for up to 20 epochs with early stopping based on validation loss, using a patience threshold of 5. Random seed was fixed to ensure stability and reproducibility.

6 Results and Discussion

Table 3 compares the performance of different models on the *SMARTSpan* and *MultiSpanQA* datasets for goal extraction, while Table 4 summarizes the results of three encoder-based models fine-tuned on *SMARTSpan* for SMARTness classification. In both tables, results on *SMARTSpan* are reported as mean and standard deviation, as all models were evaluated across five test splits as described in Section 3.2. Full split-wise results for all models are provided in Appendix G and Appendix H.

6.1 Performance of Extractive Models on Goal Extraction Task

Extractive models consistently outperform generative models on the goal extraction task across both datasets (Table 3). The strongest performance on *SMARTSpan* is achieved by the `DeBERTa-v3-base` model, fine-tuned sequentially on *MultiSpanQA* and *SMARTSpan* using the *SpanQualifier* framework—establishing a strong span-based extraction baseline. However, when extractive models using *SpanQualifier* are trained solely on *SMARTSpan*, performance degrades substantially. For instance, `DeBERTa-v3-base` fine-tuned only on *SMARTSpan* yields the lowest performance among all evaluated models, highlighting that this framework struggles to generalize in low-resource settings without prior exposure to a larger dataset.

This significant performance reduction highlights the sensitivity of extractive models to limited training data. The *SMARTSpan* training split comprises only 123 examples, which appears to be insufficient for the model to learn effective span representations without prior exposure to larger datasets like *MultiSpanQA* (see Appendix I). Despite this, the two-stage fine-tuning process mitigates the performance gap. As shown in Appendix K, models first exposed to *MultiSpanQA* can successfully identify goal-relevant regions even in loosely structured HC notes, demonstrating the importance of pretraining on richly supervised multi-span data before adapting to low-resource, domain-specific datasets such as *SMARTSpan*.

6.2 Performance of Generative Models on Goal Extraction Task

Generative LLMs consistently underperform span-extractive baselines on *SMARTSpan*. For instance, `mistral-7b-instruct-v0.3` achieves only 48.04 EM and 73.49 PM F1 score, whereas a 20 times smaller CSS extractor achieves 75.23 EM and 85.56 PM F1 score. Their larger split-to-split standard deviations provide further evidence of their instability under limited training data. In contrast, the same generative models match or surpass extractive systems on *MultiSpanQA* (up to 87.10 PM F1 score), highlighting a strong sensitivity to domain shift from open-domain MSRC to low-resource, HC notes. Until tighter span-faithfulness constraints emerge, extractive or hybrid pipelines remain the safer choice for clinical goal extraction.

Manual inspection further pinpoints two recurrent failure modes in generative outputs:

- (i) *hallucination* – insertion of content absent from the note (e.g., random HTML tags, stray characters, and fabricated text as shown in Appendices K and L); and
- (ii) *null extraction* – Span-grounding methods such as QASE (Ai et al., 2024) reduce these errors on *MultiSpanQA*, yet prove ineffective on *SMARTSpan*, leaving responsibility-sensitive failures largely unaddressed.

6.3 Analysis of the SMARTness Classifiers

As shown in Table 4, we evaluated three transformer-based models trained on *SMARTSpan* for SMARTness classification:

Model	#Params	*SMARTSpan*						*MultiSpanQA*					
		EM (mean ± sd)			**PM** (mean ± sd)			**EM**			**PM**		
		P↑	**R↑**	**F↑**	**P↑**	**R↑**	**F↑**	**P↑**	**R↑**	**F↑**	**P↑**	**R↑**	**F↑**
Extractive models (SpanQualifier)													
DeBERTa-v3-base$_{SMARTSpan}$		0.28 (0.17)	20.94 (10.36)	0.56 (0.35)	22.20 (3.12)	75.79 (4.49)	34.26 (3.39)	–	–	–	–	–	–
DeBERTa-v3-base$_{MultiSpanQA_SMARTSpan}$	180M	**85.24** (4.89)	84.46 (5.00)	**84.83** (4.83)	**93.69** (4.48)	90.72 (3.70)	**92.14** (3.57)	–	–	–	–	–	–
DeBERTa-v3-base$_{MultiSpanQA}$		12.54 (5.00)	13.16 (6.27)	12.81 (5.60)	39.55 (5.85)	27.44 (7.86)	32.11 (7.09)	**76.56**	73.31	**74.90**	88.49	83.37	85.86
BERT-base-uncased$_{SMARTSpan}$		0.34 (0.13)	24.43 (6.86)	0.68 (0.24)	23.35 (1.96)	79.42 (5.75)	36.07 (2.86)	–	–	–	–	–	–
BERT-base-uncased$_{MultiSpanQA_SMARTSpan}$	110M	78.96 (1.41)	74.78 (2.04)	76.80 (1.56)	89.37 (3.31)	84.24 (2.02)	86.72 (2.47)	–	–	–	–	–	–
BERT-base-uncased$_{MultiSpanQA}$		9.47 (4.98)	10.58 (6.02)	9.92 (5.32)	35.73 (5.92)	22.07 (5.42)	26.79 (4.48)	66.99	68.81	67.89	80.07	78.17	79.11
Extractive models (CSS)													
RoBERTa-base	125M	65.80 (10.22)	**88.12** (3.28)	74.84 (6.72)	79.24 (8.56)	**92.80** (2.97)	85.14 (4.58)	74.93	69.91	72.33	85.95	77.51	81.51
BERT-base-uncased	110M	68.00 (7.65)	84.85 (3.99)	75.23 (5.45)	82.46 (6.86)	89.18 (4.50)	85.56 (5.04)	69.93	61.22	65.29	81.82	70.26	75.60
Generative models (LoRA fine-tuning)													
phi-4	14B	29.55(19.59)	19.96(19.34)	23.46(19.86)	30.80(22.05)	21.38(22.18)	24.82(22.55)	73.82	70.96	72.36	88.74	83.75	86.17
Mistral-Nemo-Instruct-2407	13B	46.65(18.22)	43.55(23.16)	44.61(21.15)	66.30(28.20)	62.28(33.79)	63.68(31.43)	17.79	**74.31**	28.70	37.10	**90.92**	52.70
gemma-2-9b-it	9B	30.41(5.26)	53.31(5.10)	38.60(5.43)	54.29(6.78)	85.34(3.13)	66.12(5.45)	72.72	73.10	72.91	87.52	85.99	86.75
Meta-Llama-3.1-8B	8B	25.60(12.03)	14.96(9.50)	18.79(10.74)	25.60(12.03)	14.96(9.50)	18.79(10.74)	54.68	63.84	58.91	72.76	79.07	75.78
mistral-7b-instruct-v0.3	7B	46.95(16.78)	50.03(21.60)	48.04(19.55)	71.37(28.77)	76.82(34.51)	73.49(31.84)	74.20	74.05	74.12	88.31	84.62	86.43
Llama-3.2-3B	3B	26.11(6.92)	37.77(11.39)	30.32(7.77)	46.45(8.26)	60.90(12.33)	51.92(7.70)	54.00	72.00	61.72	70.53	86.75	77.80
flan-t5-large	750M	44.38 (10.44)	37.80 (6.09)	40.57 (7.45)	66.54 (10.88)	78.18 (4.20)	71.54 (7.67)	74.03	71.90	72.95	88.01	85.40	86.68
bart-large	406M	34.80 (10.91)	23.65 (8.32)	28.14 (9.46)	69.79 (12.15)	55.71 (6.34)	61.70 (8.36)	48.92	47.56	48.23	65.30	59.93	62.50
Generative models (QASE)													
flan-t5-large	750M	35.84 (4.92)	28.89 (3.02)	31.89 (3.44)	70.35 (8.00)	50.01 (1.57)	58.34 (3.88)	75.59	70.17	72.78	**91.48**	83.12	**87.10**
Generative models (LLM Schema)													
GPT-4.1	–	39.89 (8.52)	34.92 (9.28)	37.14 (8.84)	84.57 (6.31)	67.42 (6.02)	74.88 (5.40)	–	–	–	–	–	–

Table 3: Performance evaluation on *SMARTSpan* and *MultiSpanQA* datasets for all Goal Extraction models, sorted by parameter size (descending). All metrics are higher-the-better (↑) and numbers in parentheses correspond to standard deviation. The best result per column is in **bold**, second-best is underlined.

Model	#Params	Accuracy (↑)	Macro F1 (↑)	SMART F1 (↑)
deberta-v3-large	435M	**0.86 ± 0.02**	**0.85 ± 0.03**	**0.91 ± 0.03**
RoBERTa-large	355M	0.70 ± 0.26	0.67 ± 0.30	0.70 ± 0.40
deberta-v3-base	180M	0.83 ± 0.05	0.81 ± 0.06	0.90 ± 0.03

Table 4: Evaluation on *SMARTSpan* dataset for all SMARTness Classification models, sorted by parameter size (descending). All metrics are higher-the-better (↑). Best results per column are in **bold**.

DeBERTa-v3-large, RoBERTa-large, and DeBERTa-v3-base. DeBERTa-v3-large achieved the highest scores on all evaluation metrics. Its accuracy reached 0.86, with a macro-average F1 of 0.85 and a SMART F1 score of 0.91. DeBERTa-v3-base is a smaller model but performed nearly as well in all categories. Its SMART F1 was only modestly lower, which suggests that it may serve as a practical alternative when computational resources are limited.

To better understand the remaining mistakes, we categorized the errors into three main types. The first is *boundary confusion*, where the model struggles to distinguish between categories such as SMART and Partially SMART. This often occurs when a goal is nearly completed but lacks a minor element, such as a specific time frame. The second type is *overclassification*, where incomplete goals are incorrectly labeled as SMART. This usually occurs when vague wording or loosely defined measures are interpreted as sufficient. The third type is *underclassification*, which refers to clearly defined SMART goals being labeled as less specific because key attributes are implied rather than being stated directly. These observations suggest that classification performance could be improved by including more training examples that reflect subtle variations in goal phrasing. It may also help to better define how each SMART components should be recognized during training.

6.4 Performance Evaluation of the *SMARTMiner* Framework

Finally, we built and evaluated an end-to-end *SMARTMiner* framework by combining the best-performing goal extraction model (DeBERTa-v3-base) with a SMARTness classification model based on DeBERTa-v3-large. Inference across all five *SMARTSpan* test splits yielded 198 golden goals, 76% of which were unique. The goal extractor module correctly extracted 185 gold goals (recall 93.4%), with 13 omissions and 18 hallucinations (precision 91.2%). Among the omitted goals, 5 (38.5%) were labeled as Not SMART, 6 (46.2%) as Partially SMART, and 2 (15.4%) as SMART—suggesting that extraction failures were more likely for goals with lower SMARTness. Conversely, among hallucinated

label \ pred	Not SMART	Partially SMART	SMART
Not SMART	**34 (75.6%)**	10 (22.2%)	1 (2.2%)
Partially SMART	5 (9.8%)	**31 (60.8%)**	15 (29.4%)
SMART	1 (1.1%)	3 (3.4%)	**85 (95.5%)**

Table 5: Performance evaluation of the *SMARTMiner* framework across all five *SMARTSpan* test splits. Correct predictions are shown in **bold**.

goals, a majority were labeled as SMART (10, 55.6%), followed by Partially SMART (7, 38.9%) and Not SMART (1, 5.6%), indicating that the extractor tends to generate plausible SMART goals even when they are annotated.

Table 5 reports the SMARTness confusion matrix over matched extractions only ($n = 185$). Overall matched-only accuracy is 81.1%, with the SMART category showing the strongest performance (95.5%). Most errors occur between the two adjacent categories Partially SMART and SMART (29.4%), with the *SMARTMiner* framework upgrading goals by marking them as measurable in 10 (66.7%) such cases, even when the gold standard did not. This pattern suggests that the classifier module of *SMARTMiner* is more sensitive to detecting measurability than to correctly leaving goals in the Partially SMART category when measurability is absent. Considering omissions as errors, the end-to-end framework accuracy is 75.8%.

7 Conclusions and Future Work

This paper presents the *SMARTMiner* framework to support health coaches in identifying and categorizing client goals—particularly those that may require reformulation—while making all extracted goals accessible to clients through a mobile app interface. The framework was evaluated in both synthetic (i.e., *MultiSpanQA*) and real-world (i.e., *SMARTSpan*) datasets to assess its generalizability and practical utility.

Our experiments reveal that extractive span-based models, such as *SpanQualifier*, consistently outperform fine-tuned generative models in identifying multiple goals per HC note. This performance gap was particularly pronounced in *SMARTSpan*—a real-world, low-resource dataset where precision and span fidelity are crucial. These results emphasize the importance of architecture selection and data-aware fine-tuning strategies, especially in settings with limited annotated data.

Our findings are aligned with prior work showing that extractive models (e.g., BERT-based) often outperform generative models in zero-shot and few-shot setups. For example, Huang et al. (Huang et al., 2023a) demonstrated that their DeBERTa-based extractive model outperformed several GPT-based baselines, including GPT-3.5-turbo-0301, on *MultiSpanQA*. Similarly, Zhang et al. (Zhang et al., 2024) found that a simple BERT-base extractor surpassed few-shot prompting with ChatGPT-3.5.

Looking ahead, we aim to extend *SMARTMiner* into *GoalGuardian*—a fully autonomous system designed to conduct weekly goal review sessions in between monthly HC appointments. This system would initiate structured check-ins, evaluate client progress, reinforce accountability, and generate timely summaries for health coaches. Ultimately, it seeks to sustain client engagement and foster behavior change even in the absence of direct human involvement.

Limitations

As stated in Section 3, for anonymization of our data, we employed ChatGPT only for *minimal, word-level rewrite*; every rewrite suggestion was reviewed and, where necessary, corrected by human annotators to ensure that no personally identifiable information remained and that the note structure, wording, and formatting were preserved exactly, avoiding hallucinated edits or template drift. This intensive human-in-the-loop pipeline is feasible for the present corpus of 173 HC notes but does not scale linearly. For larger releases we will introduce an additional annotation layer that *grades* each LLM-generated rewrite for fidelity and formatting compliance before acceptance, allowing us to maintain real-world note realism while keeping human effort tractable.

Given the nature of the collected data (i.e., HC notes), we focused only on labeling the three core SMART components: specific (S), measurable (M), and attainable (A). The time-bound (T) element was implicitly defined by the structure of the intervention, where clients set weekly goals or targets to be completed before the next HC session. However, since this temporal aspect was rarely stated explicitly in the HC notes, it could not be annotated as a gold label or reliably captured by our classifier. Similarly, the relevant (R) dimension was excluded from annotation, as health coaches served as the first filter—only documenting goals that were deemed relevant within the scope of the study. As a result, R was not labeled in the dataset.

All HC notes in the *SMARTSpan* dataset are written in English and were collected over a one-year period from a specific population with elevated LDL levels, participating in an intervention to improve adherence to statin therapy. The dataset reflects the language, behaviors, and health system context of this particular group. Anyone using the dataset should be mindful of these contextual limitations when applying the models or generalizing findings to other populations or settings.

Ethical Considerations

The randomized controlled trial from which the health coaching session notes were derived received ethics approval from the National Healthcare Group Domain Specific Review Board in Singapore (no. 2023/00438). All participants provided written informed consent prior to enrollment. The trial is conducted across multiple healthcare institutions in Singapore, including National University Hospital, National University Polyclinics, and National Healthcare Group Polyclinics.

Acknowledgments

This work was supported by CArdiovascular DiseasE National Collaborative Enterprise (CADENCE) National Clinical Translational Program (MOH-001277-01). ZG is supported by the NSF DBI-2229929 (The NSF AI Institute for Artificial and Natural Intelligence).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Lin Ai, Zheng Hui, Zizhou Liu, and Julia Hirschberg. 2024. Enhancing pre-trained generative language models with question attended span extraction on machine reading comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10046–10063.

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).

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.

Yu-Wen Chen and Julia Hirschberg. 2024. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain SOAP notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 1–9, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5925–5932.

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).

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1*, pages 2368–2378.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Glenn T Gobbel, Michael E Matheny, Ruth R Reeves, Julia M Akeroyd, Alexander Turchin, Christie M Ballantyne, Laura A Petersen, and Salim S Virani. 2022. Leveraging structured and unstructured electronic health record data to detect reasons for suboptimal statin therapy use in patients with atherosclerotic cardiovascular disease. *American Journal of Preventive Cardiology*, 9:100300.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020a. 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, 1st virtual meeting. Association for Computational Linguistics.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2019. Modeling health coaching dialogues for behavioral goal extraction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1188–1190.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2020b. Goal summarization for human-human health coaching dialogues. In *The Thirty-Third International FLAIRS Conference*.

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, Singapore and Online. Association for Computational Linguistics.

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

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*.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1596–1606.

Zixian Huang, Jiaying Zhou, Chenxu Niu, and Gong Cheng. 2023a. 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.

Zixian Huang, Jiaying Zhou, Gengyang Xiao, and Gong Cheng. 2023b. Enhancing in-context learning with answer feedback for multi-span question answering. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 744–756. Springer.

Andy W. H. Khong and Patrick A. Naylor. 2005. Selective-tap adaptive algorithms in the solution of non-uniqueness problem for stereophonic acoustic echo cancellation. *IEEE Signal Processing Letters*, 12(4):269–272.

Seongyun Lee, Hyunjae Kim, and Jaewoo Kang. 2023. Liquid: a framework for list question answering dataset generation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, pages 13014–13024.

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.

Jiayi Lin, Chenyang Zhang, Haibo Tong, Dongyu Zhang, Qingqing Hong, Bingxuan Hou, and Junli Wang. 2024. Correct after answer: Enhancing multi-span question answering with post-processing method. In *Findings of the ACL: EMNLP 2024*, pages 2701–2717.

Hexin Liu, Leibny Paola Garcia Perera Perera, Andy W. H. Khong; Eng Siong Chng; Suzy J. Styles, and Sanjeev Khudanpur. 2022. Efficient self-supervised learning representations for spoken language identification. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1296–1307.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyi Luo, Yingying Zhang, and Shuyun Luo. 2024. A token-based transition-aware joint framework for multi-span question answering. *Information Processing & Management*, 61(3):103678.

Prabir Mallick, Tapas Nayak, and Indrajit Bhattacharya. 2023. Adapting pre-trained generative models for extractive question answering. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 128–137.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open models. Accessed: 17-Sep-25.

Timothy Miller, Steven Bethard, Dmitriy Dligach, and Guergana Savova. 2023. End-to-end clinical temporal information extraction with multi-head attention. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 313–319, Toronto, Canada. Association for Computational Linguistics.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. https://openai.com/research/gpt-4. Accessed: 17-Sep-25.

Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. 2024. A survey on open information extraction from rule-based model to large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9586–9608, Miami, Florida, USA. Association for Computational Linguistics.

James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48.

Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nic Dobbins, Ozlem Uzuner, and Meliha Yetisgen. 2023. Prompt-based extraction of social determinants of health using few-shot learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 385–393, Toronto, Canada. Association for Computational Linguistics.

Brian Romanowski, Asma Ben Abacha, and Yadan Fan. 2023. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *Journal of the American Medical Informatics Association*, 30(8):1448–1455.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3074–3080.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Unsloth. 2024. Unsloth: Efficient fine-tuning of llms. Accessed: 17-Sep-25.

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.

Jocelyn M Weiss, Bala Munipalli, Miranda P Kaye, Katherine Smith, Eli Shur, Sebastian Harenberg, Rachel Garofalo, Arya B Mohabbat, Arden Robinson, Stefan N Paul, and 1 others. 2025. Compendium of health and wellness coaching: 2023 addendum. *Journal of integrative and complementary medicine*.

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.

RQ Wolever, M Dreusicke, J Fikkan, TV Hawkins, S Yeung, J Wakefield, L Duda, P Flowers, C Cook, and E Skinner. 2010. Integrative health coaching for patients with type 2 diabetes. *The Diabetes Educator*, 36(4):629–639.

Junjie Yang, Zhuosheng Zhang, and Hai Zhao. 2020. Multi-span style extraction for generative reading comprehension. *arXiv preprint arXiv:2009.07382*.

Chen Zhang, Jiuheng Lin, Xiao Liu, Yuxuan Lai, Yansong Feng, and Dongyan Zhao. 2023a. How many answers should i give? an empirical study of multi-answer reading comprehension. In *Findings of the ACL: ACL 2023*, pages 5811–5827.

Penghui Zhang, Guanming Xiong, and Wen Zhao. 2023b. Css: Contrastive span selector for multi-span question answering. In *Pacific Rim International Conference on Artificial Intelligence*, pages 225–236. Springer.

Yingying Zhang, Zhiyi Luo, and Zuohua Ding. 2024. A simple and effective span interaction modeling method for enhancing multiple span question answering. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 188–200. Springer.

Xin Zhou, Yanshan Wang, Sunghwan Sohn, Terry M Therneau, Hongfang Liu, and David S Knopman. 2019. Automatic extraction and assessment of lifestyle exposures for alzheimer's disease using natural language processing. *International journal of medical informatics*, 130:103943.

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, Torino, Italia. ELRA and ICCL.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. 2022. Towards enhancing health coaching dialogue in low-resource settings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 694–706, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the ACL: EMNLP 2020*, pages 3840–3849.

A Example of HC Note

Figure 3 shows an example of a real-world HC note inputted into the HC web-based platform. The note includes SMART goals, lifestyle observations (e.g., travel), and structured follow-up on weekly SMART goal performance with patient-reported adherence and perceived success rates.

HC note

```
1. Swimming and aquarobics for 30 minutes once a week.
2. Homemade salad for breakfast once a week, buy salad as
alternative if busy.
3. Take medications every day of the week.

Client recently returned from a trip with a friend that
included food outings and leisure travel.

Goals review:
1. Aquatic exercises were done once a week due to work
commitments, perceived success 40-50%.
2. Had breakfast three times a week, usually bread and
coffee, perceived success 100%.
3. Took medications 6 out of 7 days, found pillbox and
app reminders helpful especially after late shifts,
perceived success 100%.

Generative moment: Participated in stretching activities
at a local centre, signed up for weekly sessions and began
socialising through games.

Goals setting:
1. Continue swimming and aquatic exercises once a week
(confidence 100%).
2. Consume  homemade salad as breakfast once a week,
with flexibility to buy if necessary (confidence 50%).
3. Take statin medications every day of the week
(confidence 90%).

App usage: Frequently used medication reminders, watched
a few videos but cited lack of time.

Other concerns: Currently adjusting antihypertensive
medication under physician's guidance due to low heart rate
and elevated BP readings. Client was instructed to monitor
and record BP and heart rate daily using the app diary and
dictation tool.
```

Extracted goals

The following goals were manually extracted from the above-mentioned HC note as golden labels

```
1. Continue swimming and aquatic exercises once a week
(confidence 100%)
2. Consume homemade salad as breakfast once a week, with
flexibility to buy if necessary (confidence 50%)
3. Take statin medications every day of the week
(confidence 90%)
```

From this example, we can see that goal-related statements often appear in multiple locations within a single HC note, varying in specificity and finality. The initial mention of swimming—*"Swimming and aquarobics for 30 minutes once a week"*—serves as a general intention. Later, a more definitive formulation—*"Continue swimming and aquatic exercises once a week (confidence 100%)"*—reflects a more completed formulation of the goal.

Moreover, some statements may resemble goals in language or structure but actually describe past behavior, such as *"Aquatic exercises were done once a week."* These retrospective statements must be excluded from annotations, as they do not represent forward-looking intentions and could mislead models during training. Accurately separating past reflections from future commitments is essential for developing systems that support behavior change interventions, ensuring only actionable goals are surfaced for downstream use.

SMARTness of extractive goals

Following our core SMARTness framework, the extracted goals are manually evaluated for specificity, measurability, and attainability. A label of `[1,1,1]` indicates that the goal meets all three core criteria and it is labeled as SMART.

```
1. [1,0,1]
2. [1,1,1]
3. [1,1,1]
```

In the example above, the first goal is *specific* (it mentions swimming and aquatic exercises) and *attainable* (confidence of 100%), but not *measurable*, as it does not specify the duration of swimming or aquatic activities. The second goal is *specific* (consume homemade salad), *measurable* (once a week), and *attainable* (confidence assessed). Finally, the third goal is *specific* (statin medications), *measurable* (every day of the week), and *attainable* (confidence of 90%). Some other examples of Partially SMART and Not SMART goals in the dataset are

```
1. Complete 5km walk two times a week from midweek onwards.
[1,1,0]
2. Exercise daily for at least 1 hour. [0,1,0]
3. Exercise daily for 30 minutes (confidence level 8/10).
[0,1,1]
4. Reduce portion sizes and adjust ingredients to
healthier options. [0,0,0]
```

Goal 1 is Partially SMART as it is *specific* (5km walk) and *measurable* (twice a week), but lacks an *attainability* marker such as confidence or feasibility. Goal 2 is not SMART as it is *measurable* (daily, 1 hour), but lacks *specificity* (type of exercise) and does not include *attainability*. Goal 3 is Partially SMART as it is *measurable* (daily, 30 minutes) and *attainable* (confidence of 8/10), but not *specific* about the type of activity. Goal 4 is Not SMART as it lacks all three SMART components—there is no clear target behavior (not specific), no quantifiable element (not measurable), and no indication of feasibility (not attainable).

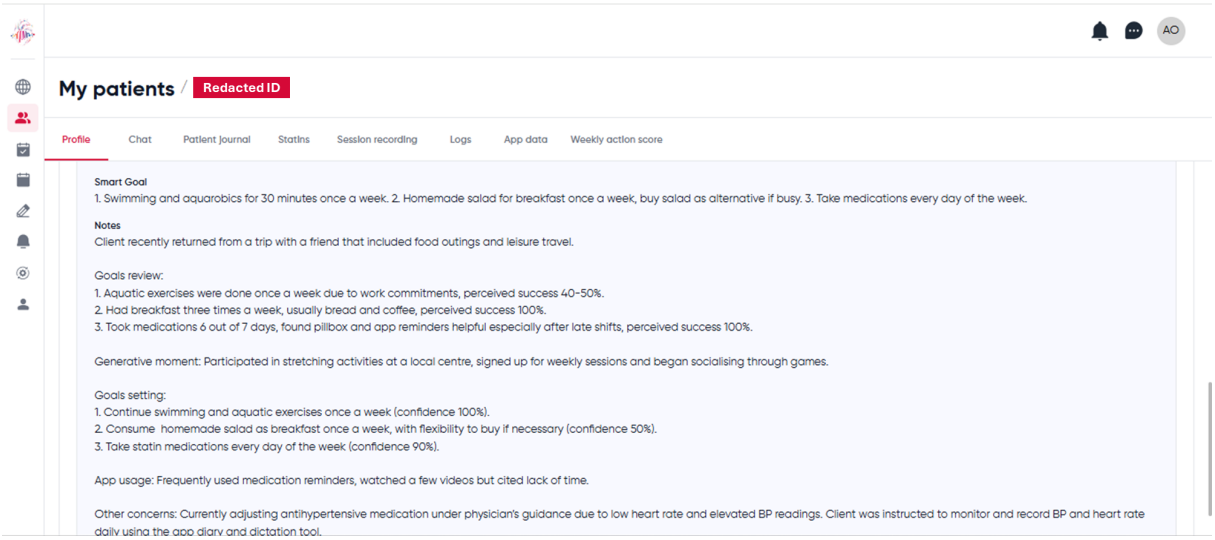

Figure 3: Example of a HC note recorded in the HC web-based platform.

B Annotation of *SMARTSpan* Dataset

The annotation was performed by two academic researchers who hold a doctoral degree (PhD and MBBS) and have received prior training in HC. One annotator also has clinical experience in delivering HC interventions. Prior to the annotation task, both annotators jointly reviewed relevant literature and established a shared understanding of the definitions and criteria for the specific (S), measurable (M) and attainable (A) components.

To minimize potential biases arising from practice effects or subjective interpretation, both annotators were first presented with illustrative examples of well- and poorly formulated SMART goals. Following this, a structured calibration exercise involving 10 sample goals—sourced independently of the main dataset—was conducted to align the annotators' understanding of the classification rubric and ensure consistency in rating. Notably, no reimbursement was provided for their effort, as both annotators are members of the core research team.

C Extraction Prompt for LoRA Fine-Tuning of Generative Models

The prompt below defines the task used for fine-tuning generative models to extract goals from unstructured HC notes. It clearly frames the task as extractive rather than generative, instructing the model to copy exact span text without paraphrasing. The instruction emphasizes exclusion of vague categories and long-term intentions, and focuses only on short-term, concrete weekly goals. The model is further guided to format the output as a numbered list.

```
Below is an instruction that describes a task, paired with an input
that provides further context. Write a response that appropriately
completes the request.

### Instruction:
You are an expert assistant that extracts only SMART weekly goals
from health coaching session notes. Extract the exact parts of
text, don't rephrase the text! This is an NLU task, and not an
NLG task! Only include goals that are: Specific, Measurable,
Attainable, Relevant, and Time-bound (SMART). Do not include
vague or broad categories like 'Exercise', 'Medication', or
'Diet' unless they are written as specific SMART goals. Ignore
6-month, long-term, or vague intentions. Focus only on short-term,
concrete weekly SMART goals that the patient committed to.
Format your response as a numbered list.

### Input: {input}

### Output: {output}
```

D Overview of Generative Models used for Fine-Tuning

Table 6 summarizes key details for the generative models used in our evaluation, including their release dates, pre-training cutoff dates, and the number of fine-tuned parameters relative to total model size. These models vary widely in scale and recency, which may influence their ability to extract structured SMART goals under zero-shot or fine-tuned conditions.

Model	Train params	Release	Cutoff
`phi-4` (Abdin et al., 2024)	33M out of 14B	2024-12-13	Jun 2024
`Mistral-Nemo-Instruct-2407`	29M out of 12B	2024-07-18	Apr 2024
`gemma-2-9b-it` (Team et al., 2024)	27M out of 9B	2024-06-27	–
`Meta-Llama-3.1-8B` (Dubey et al., 2024)	21M out of 8B	2024-07-23	Dec 2023
`mistral-7b-instruct-v0.3`	21M out of 7B	2024/05/22	–
`Llama-3.2-3B` (Meta, 2024)	12M out of 3B	2024-09-25	Dec 2023

Table 6: Release and pre-training cutoff dates for used generative models.

E Learning Rate Grid Search

To better understand the sensitivity of our models to hyperparameters, we conducted a grid search over four learning rates: 1×10^{-5}, 3×10^{-5}, 5×10^{-5}, and 1×10^{-4}. This analysis was performed only for *SMARTSpan* split_1, and results are reported using the Partial Match F1 (PM_F1) metric.

Based on the results in Table 7, we adopted a learning rate of 1×10^{-4} for encoder–decoder models (Flan-T5 and BART), which consistently achieved their strongest performance at this setting. For decoder-only models, we selected 3×10^{-5} as a balanced choice: it produced the best results for both Mistral variants and remained competitive across Llama and Gemma models, ensuring stable and robust training across architectures.

Model	Learning Rate			
	1e-5	3e-5	5e-5	1e-4
phi-4	14.08	14.08	**18.67**	14.08
Mistral-Nemo-Instruct-2407	82.77	**86.59**	60.67	22.35
gemma-2-9b-it	**82.08**	68.79	66.99	75.34
Meta-Llama-3.1-8B	14.08	14.08	**16.87**	14.08
mistral-7b-instruct-v0.3	82.27	**93.16**	80.78	69.73
Llama-3.2-3B	52.65	53.01	**54.87**	44.83
flan-t5-large	30.38	37.14	62.86	**72.71**
bart-large	20.28	49.87	56.81	**74.43**

Table 7: PM_F1 scores across learning rates on *SMARTSpan* Split 1. The best-performing learning rate for each model is highlighted in bold.

F Zero-Shot SMART GPT-4.1 Goal Extraction Prompt and Schema

To support zero-shot structured extraction of weekly SMART goals, we used OpenAI's GPT-4.1 with the function-calling API. The model was instructed via a detailed system prompt (i.e., instruction) and constrained to return outputs in a predefined JSON schema. Below we provide both the prompt and the schema used.

Instruction

```
You are an expert assistant that extracts only SMART weekly goals from health coaching session notes. Extract the exact parts of text, don't rephrase the text! This is an NLU task, and not an NLG task! Only include goals that are: Specific, Measurable, Attainable, Relevant, and Time-bound (SMART). Do not include vague or broad categories like "Exercise", "Medication", or "Diet" unless they are written as specific SMART goals. Ignore 6-month, long-term, or vague intentions. Focus on short-term, concrete weekly SMART goals that the patient committed to. Respond in JSON format.
```

Function Schema

```
[
  {
    "type": "function",
    "function": {
      "name": "extract_weekly_smart_goals",
      "description": "Extract only weekly SMART goals.
                      Ignore long-term or monthly goals.",
      "parameters": {
        "type": "object",
        "properties": {
          "goals": {
            "type": "array",
            "items": { "type": "string" },
            "description": "List of weekly SMART goals" }
        },
        "required": ["goals"] }
    }
  }
]
```

G Performance of the Goal Extraction Models Across *SMARTSpan* Splits

Table 8 reports the performance of `DeBERTa-v3-base` and `BERT-base-uncased` on five test splits of the *SMARTSpan* dataset under three fine-tuning settings using the *SpanQualifier* framework: (*i*) trained only on *SMARTSpan*, (*ii*) first on *MultiSpanQA* then further fine-tuned on *SMARTSpan*, and (*iii*) trained only on *MultiSpanQA*. Results clearly show that models fine-tuned sequentially on *MultiSpanQA* and *SMARTSpan* achieve the highest scores across all splits, with `DeBERTa-v3-base` reaching up to **90.11** EM F1 and **97.41** PM F1, and `BERT-base-uncased` up to **78.65** EM F1 and **89.62** PM F1. In contrast, models trained solely on *SMARTSpan* perform poorly, with EM F1 scores close to zero despite relatively higher PM recall. Models fine-tuned only on *MultiSpanQA* exhibit modest performance improvements over *SMARTSpan*-only training, but still fall short of the multi-stage fine-tuning setup.

Model	EM			PM		
	P↑	R↑	F↑	P↑	R↑	F↑
DeBERTa-v3-base$_{SMARTSpan_1}$	0.26	19.57	0.52	21.74	81.29	34.31
DeBERTa-v3-base$_{SMARTSpan_2}$	0.35	21.57	0.70	27.84	78.34	41.08
DeBERTa-v3-base$_{SMARTSpan_3}$	0.12	12.24	0.23	21.05	73.05	32.69
DeBERTa-v3-base$_{SMARTSpan_4}$	0.11	10.42	0.22	20.83	68.63	31.96
DeBERTa-v3-base$_{SMARTSpan_5}$	0.58	38.89	1.14	19.56	77.64	31.24
DeBERTa-v3-base$_{MultiSpanQA_SMARTSpan_1}$	91.11	89.13	90.11	100.00	94.96	97.41
DeBERTa-v3-base$_{MultiSpanQA_SMARTSpan_2}$	81.13	84.31	82.69	88.03	90.55	89.27
DeBERTa-v3-base$_{MultiSpanQA_SMARTSpan_3}$	89.58	87.76	88.66	97.81	91.69	94.65
DeBERTa-v3-base$_{MultiSpanQA_SMARTSpan_4}$	78.26	75.00	76.60	91.44	83.91	87.51
DeBERTa-v3-base$_{MultiSpanQA_SMARTSpan_5}$	86.11	86.11	86.11	91.18	92.50	91.84
DeBERTa-v3-base$_{MultiSpanQA_1}$	11.11	10.87	10.99	40.08	23.36	29.52
DeBERTa-v3-base$_{MultiSpanQA_2}$	9.09	7.84	8.42	36.68	21.36	27.00
DeBERTa-v3-base$_{MultiSpanQA_3}$	6.00	6.12	6.06	32.94	18.91	24.03
DeBERTa-v3-base$_{MultiSpanQA_4}$	16.98	18.75	17.82	50.30	38.77	43.79
DeBERTa-v3-base$_{MultiSpanQA_5}$	19.51	22.22	20.78	37.75	34.82	36.22
BERT-base-uncased$_{SMARTSpan_1}$	0.19	15.22	0.38	20.33	75.72	32.05
BERT-base-uncased$_{SMARTSpan_2}$	0.42	25.49	0.83	26.13	87.35	40.23
BERT-base-uncased$_{SMARTSpan_3}$	0.33	24.49	0.65	23.59	82.91	36.73
BERT-base-uncased$_{SMARTSpan_4}$	0.24	20.83	0.48	24.43	80.42	37.48
BERT-base-uncased$_{SMARTSpan_5}$	0.54	36.11	1.06	22.26	70.72	33.86
BERT-base-uncased$_{MultiSpanQA_SMARTSpan_1}$	81.40	76.09	78.65	93.92	85.70	89.62
BERT-base-uncased$_{MultiSpanQA_SMARTSpan_2}$	78.00	76.47	77.23	86.92	84.24	85.56
BERT-base-uncased$_{MultiSpanQA_SMARTSpan_3}$	78.72	75.51	77.08	91.61	87.11	89.30
BERT-base-uncased$_{MultiSpanQA_SMARTSpan_4}$	77.27	70.83	73.91	89.82	82.63	86.08
BERT-base-uncased$_{MultiSpanQA_SMARTSpan_5}$	79.41	75.00	77.14	84.58	81.53	83.03
BERT-base-uncased$_{MultiSpanQA_1}$	11.36	10.87	11.11	43.19	20.36	27.68
BERT-base-uncased$_{MultiSpanQA_2}$	3.92	3.92	3.92	29.19	18.33	22.52
BERT-base-uncased$_{MultiSpanQA_3}$	3.17	4.08	3.57	33.20	15.18	20.83
BERT-base-uncased$_{MultiSpanQA_4}$	14.89	14.58	14.74	42.41	26.43	32.57
BERT-base-uncased$_{MultiSpanQA_5}$	14.00	19.44	16.28	30.65	30.04	30.34

Table 8: Extractive models (`DeBERTa-v3-base` and `BERT-base-uncased`) performance across five *SMARTSpan* test splits using *SpanQualifier*.

Model	EM			PM		
	P↑	R↑	F↑	P↑	R↑	F↑
RoBERTa-base$_{split_1}$	78.00	84.78	81.25	89.81	87.77	88.78
RoBERTa-base$_{split_2}$	52.38	86.27	65.19	66.71	93.20	77.76
RoBERTa-base$_{split_3}$	69.70	93.88	80.00	84.10	96.20	89.75
RoBERTa-base$_{split_4}$	55.13	89.58	68.25	71.88	95.18	81.90
RoBERTa-base$_{split_5}$	73.81	86.11	79.49	83.69	91.67	87.50
BERT-base-uncased$_{split_1}$	76.92	86.96	81.63	86.79	89.94	88.34
BERT-base-uncased$_{split_2}$	53.75	84.31	65.65	71.53	90.63	79.95
BERT-base-uncased$_{split_3}$	69.84	89.80	78.57	88.22	94.82	91.40
BERT-base-uncased$_{split_4}$	69.49	85.42	76.64	88.52	89.49	89.00
BERT-base-uncased$_{split_5}$	70.00	77.78	73.68	77.26	81.01	79.09

Table 9: Extractive models (`BERT-base-uncased` and `RoBERTa-base`) performance across five *SMARTSpan* test splits using CSS.

Table 9 reports the performance of `BERT-base-uncased` and `RoBERTa-base` across five test splits of the *SMARTSpan* dataset, using the CSS fine-tuning framework. Both models achieve strong and stable PM F1 performance, averaging above **85**. EM F1 scores for both models also vary within a similar range—approximately 65 to 81—suggesting comparable sensitivity to training data splits. While `RoBERTa-base` achieves a slightly higher peak PM F1 score (**89.75**) and EM F1 score (**81.25**), `BERT-base-uncased` performs competitively and exhibits a more balanced performance across metrics. These results suggest that both models are well-suited for multi-span extraction under CSS, with `RoBERTa-base` offering marginally higher peaks and `BERT-base-uncased` offering stable returns.

Table 10 illustrates substantial variability in *SMARTSpan* performance among generative models fine-tuned using LoRA. Notably, `mistral-7b-instruct-v0.3` emerges as the strongest overall, reaching up to **65.98** EM F1 and **95.51** PM F1 across splits. In contrast, larger models like `Mistral-Nemo-Instruct-2407` achieve a comparable peak EM F1 of **66.67**, but fall short on PM F1, peaking at only **90.42**, suggesting less consistent partial span recovery. Meanwhile, models such as `phi-4` and `Meta-Llama-3.1-8B` display high variance across splits—EM F1 fluctuates from as low as **10.53** to as high as **62.69**—indicating instability in learning span fidelity. Similarly, `gemma-2-9b-it` achieves strong partial match performance (up to **74.23** PM F1), yet underperforms on exact matches, underscoring persistent challenges in precise span generation. Overall, smaller models with stable fine-tuning regimes consistently outperform their larger counterparts on this task.

Model	EM			PM		
	P↑	R↑	F↑	P↑	R↑	F↑
phi-4$_{split_1}$	20.00	10.87	14.08	20.00	10.87	14.08
phi-4$_{split_2}$	16.00	7.84	10.53	16.00	7.84	10.53
phi-4$_{split_3}$	16.00	8.16	10.81	16.00	8.16	10.81
phi-4$_{split_4}$	28.00	14.58	19.18	28.00	14.58	19.18
phi-4$_{split_5}$	67.74	58.33	62.69	74.02	65.47	69.48
Mistral-Nemo-Instruct-2407$_{split_1}$	52.17	52.17	52.17	88.11	85.11	86.59
Mistral-Nemo-Instruct-2407$_{split_2}$	37.14	25.49	30.23	54.21	37.15	44.09
Mistral-Nemo-Instruct-2407$_{split_3}$	16.00	8.16	10.81	16.00	8.16	10.81
Mistral-Nemo-Instruct-2407$_{split_4}$	63.83	62.50	63.16	86.69	86.29	86.49
Mistral-Nemo-Instruct-2407$_{split_5}$	64.10	69.44	66.67	86.50	94.71	90.42
gemma-2-9b-it$_{split_1}$	28.95	47.83	36.07	58.24	84.00	68.79
gemma-2-9b-it$_{split_2}$	31.33	50.98	38.81	57.55	81.05	67.31
gemma-2-9b-it$_{split_3}$	22.22	48.98	30.57	47.92	84.71	61.21
gemma-2-9b-it$_{split_4}$	38.67	60.42	47.15	62.89	90.58	74.23
gemma-2-9b-it$_{split_5}$	30.88	58.33	40.38	44.87	86.38	59.06
Meta-Llama-3.1-8B$_{split_1}$	20.00	10.87	14.08	20.00	10.87	14.08
Meta-Llama-3.1-8B$_{split_2}$	16.00	7.84	10.53	16.00	7.84	10.53
Meta-Llama-3.1-8B$_{split_3}$	16.00	8.16	10.81	16.00	8.16	10.81
Meta-Llama-3.1-8B$_{split_4}$	28.00	14.58	19.18	28.00	14.58	19.18
Meta-Llama-3.1-8B$_{split_5}$	48.00	33.33	39.34	48.00	33.33	39.34
mistral-7b-instruct-v0.3$_{split_1}$	55.10	58.70	56.84	91.20	95.21	93.16
mistral-7b-instruct-v0.3$_{split_2}$	16.00	7.84	10.53	16.00	7.84	10.53
mistral-7b-instruct-v0.3$_{split_3}$	44.83	53.06	48.60	79.17	94.14	86.01
mistral-7b-instruct-v0.3$_{split_4}$	65.31	66.67	65.98	96.04	94.99	95.51
mistral-7b-instruct-v0.3$_{split_5}$	53.49	63.89	58.23	74.43	91.93	82.26
Llama-3.2-3B$_{split_1}$	17.39	34.78	23.19	43.89	66.92	53.01
Llama-3.2-3B$_{split_2}$	29.21	50.98	37.14	44.10	71.07	54.43
Llama-3.2-3B$_{split_3}$	19.15	18.37	18.75	40.20	36.94	38.50
Llama-3.2-3B$_{split_4}$	36.00	37.50	36.73	62.70	61.98	62.33
Llama-3.2-3B$_{split_5}$	28.81	47.22	35.79	41.34	67.59	51.31
flan-t5-large$_{split_1}$	51.35	41.30	45.78	72.03	76.99	74.43
flan-t5-large$_{split_2}$	55.56	39.22	45.98	75.92	78.65	77.26
flan-t5-large$_{split_3}$	25.49	26.53	26.00	50.33	70.86	58.86
flan-t5-large$_{split_4}$	47.37	37.50	41.86	77.45	82.81	80.04
flan-t5-large$_{split_5}$	42.11	44.44	43.24	56.97	81.59	67.09
bart-large$_{split_1}$	40.62	28.26	33.33	83.10	64.63	72.71
bart-large$_{split_2}$	15.62	9.80	12.05	49.75	52.96	51.30
bart-large$_{split_3}$	30.00	18.37	22.78	67.84	48.88	56.82
bart-large$_{split_4}$	45.45	31.25	37.04	81.82	61.84	70.44
bart-large$_{split_5}$	42.31	30.56	35.48	66.44	50.24	57.22

Table 10: Generative models performance across five *SMARTSpan* test splits using LoRA fine-tuning.

Table 11 presents the performance of the `flan-t5-large` model across five splits using the QASE framework. While PM F1 scores remain relatively stable, ranging from **51.21** to **62.93**, EM F1 scores are considerably lower–between **26.67** and **37.04**. This pattern highlights the model's ability to capture semantically appropriate spans, even when exact matches are missed. However, it also underscores the challenge of aligning predicted spans precisely with gold-standard boundaries.

Model	EM			PM		
	P↑	R↑	F↑	P↑	R↑	F↑
flan-t5-large$_{split_1}$	42.86	32.61	37.04	78.53	52.50	62.93
flan-t5-large$_{split_2}$	30.77	23.53	26.67	70.88	50.00	58.63
flan-t5-large$_{split_3}$	35.00	28.57	31.46	73.68	50.64	60.03
flan-t5-large$_{split_4}$	40.00	29.17	33.73	73.52	49.12	58.89
flan-t5-large$_{split_5}$	30.56	30.56	30.56	55.14	47.80	51.21

Table 11: `flan-t5-large` model performance across five *SMARTSpan* test splits using QASE.

Model	EM			PM		
	P↑	R↑	F↑	P↑	R↑	F↑
GPT-4.1$_{split_1}$	48.72	41.30	44.71	96.55	75.83	84.94
GPT-4.1$_{split_2}$	33.33	27.45	30.11	84.46	61.26	71.01
GPT-4.1$_{split_3}$	27.03	20.41	23.26	83.01	60.13	69.74
GPT-4.1$_{split_4}$	42.00	43.75	42.86	79.60	71.79	75.49
GPT-4.1$_{split_5}$	48.39	41.67	44.78	79.22	68.07	73.22

Table 12: `GPT-4.1` models performance across five *SMARTSpan* test splits using LLM Schema.

Table 12 reports the performance of `GPT-4.1` across five *SMARTSpan* test splits. The model achieves strong PM F1 scores, ranging from **69.74** to **84.94**, indicating high semantic alignment with relevant spans. However, EM F1 scores are substantially lower, varying between **23.26** and **44.78**, highlighting challenges in predicting exact span boundaries. These results suggest that while `GPT-4.1` is highly effective at identifying relevant content in zero-shot settings, it falls short in producing precise extractions. In contrast, smaller generative models fine-tuned with the QASE framework show improved EM performance under supervision, despite lower PM scores. This suggests that while `GPT-4.1` excels at general semantic retrieval, QASE-enhanced models offer better span-level precision when trained with task-specific data.

H Performance of the SMARTness Classification Models Across *SMARTSpan* Splits

Table 13 shows that across the five test splits, both `deberta-v3-base` and `deberta-v3-large` exhibit consistently strong performance in SMARTness classification, with macro F1 scores exceeding **0.740** and SMART F1 scores ranging from **0.864** to **0.955**. `deberta-v3-large` achieves the highest overall scores, reaching up to **0.878** accuracy, **0.881** macro F1, and a peak SMART F1 of **0.955**, demonstrating that increased model capacity leads to greater robustness and precision.

In contrast, while `RoBERTa-large` performs competitively on most splits, it shows instability on split_ 3, where classification performance collapses entirely (SMART F1 = **0.000**). This divergence highlights the importance of evaluating models across multiple test partitions and supports the reliability of DeBERTa-based architectures—particularly the large variant—for structured SMARTness prediction in goal-oriented health coaching contexts.

I Output for `DeBERTa-v3-base` fine-tuned only on *SMARTSpan* using *SpanQualifier*

The list below shows the final inference output of `DeBERTa-v3-base` trained with the *SpanQualifier* framework on the *SMARTSpan* dataset using only 123 training examples. This model achieved an average EM F1 of **0.56** and average PM F1 of **34.26**. As shown below, the model returns a large set of overlapping span candidates with minimal filtering or boundary control.

This output illustrates *SpanQualifier's* reliance on large-scale supervision. In low-resource conditions, it fails to discriminate meaningful spans from irrelevant ones and defaults to producing dense n-gram windows. Without upstream training (e.g., on *MultiSpanQA*) or parameter-efficient adaptation (e.g., LoRA), the model lacks a coherent span representation and performs poorly on *SMARTSpan*.

```
"24": [
    ") exercise : switch from casual walking to brisk walking",
    "exercise : switch from casual walking to brisk walking ,",
   ": switch from casual walking to brisk walking , maintaining",
   "switch from casual walking to brisk walking , maintaining a",
    "from casual walking to brisk walking , maintaining a daily",
    "casual walking to brisk walking , maintaining a daily goal",
    "walking to brisk walking , maintaining a daily goal of",
    "to brisk walking , maintaining a daily goal of 10",
    "brisk walking , maintaining a daily goal of 10 ,",
    "walking , maintaining a daily goal of 10 , 000",
    ", maintaining a daily goal of 10 , 000 steps",
    "maintaining a daily goal of 10 , 000 steps (",
    "a daily goal of 10 , 000 steps ( confidence",
    "of 10 , 000 steps ( confidence 7 / 10",
    "10 , 000 steps ( confidence 7 / 10 )",
    ", 000 steps ( confidence 7 / 10 ) .",
    "( confidence 7 / 10 ) . ( 2 )",
    "confidence 7 / 10 ) . ( 2 ) diet",
    "7 / 10 ) . ( 2 ) diet :",
    "/ 10 ) . ( 2 ) diet : reduce",
    "10 ) . ( 2 ) diet : reduce the",
    ...
    "activity . future sessions may explore strategies to
    support consistent medication intake and improving sleep"
    ]
```

Model	Accuracy↑	Macro F1↑	SMART F1↑
deberta-v3-base$_{split_1}$	0.780	0.744	0.900
deberta-v3-base$_{split_2}$	0.809	0.786	0.864
deberta-v3-base$_{split_3}$	0.822	0.782	0.913
deberta-v3-base$_{split_4}$	0.902	0.897	0.927
deberta-v3-base$_{split_5}$	0.833	0.836	0.870
deberta-v3-large$_{split_1}$	0.854	0.842	0.900
deberta-v3-large$_{split_2}$	0.830	0.809	0.864
deberta-v3-large$_{split_3}$	0.867	0.834	0.955
deberta-v3-large$_{split_4}$	0.878	0.881	0.895
deberta-v3-large$_{split_5}$	0.875	0.874	0.909
RoBERTa-large$_{split_1}$	0.780	0.743	0.923
RoBERTa-large$_{split_2}$	0.787	0.745	0.851
RoBERTa-large$_{split_3}$	0.244	0.131	0.000
RoBERTa-large$_{split_4}$	0.878	0.860	0.913
RoBERTa-large$_{split_5}$	0.833	0.849	0.833

Table 13: Encoder-only transformer models (`deberta-v3-base`, `deberta-v3-large` and `RoBERTa-base`) performance across five *SMARTSpan* test splits.

J Output for `DeBERTa-v3-base` after Sequential Fine-Tuning on *MultiSpanQA* and *SMARTSpan* using *SpanQualifier*

The list below presents the final output of `DeBERTa-v3-base` fine-tuned using the *SpanQualifier* framework, first on *MultiSpanQA* and subsequently on *SMARTSpan*. Unlike the low-resource setting where the model was trained only on *SMARTSpan*, this model produces a concise and accurate set of spans, closely aligned with the gold annotations. The final predictions are well-formed and exhibit clear span boundaries, indicating that prior training on a large-scale multi-span dataset (i.e., *MultiSpanQA*) successfully bootstraps the model's ability to extract meaningful multi-span goals from *SMARTSpan*.

This result underscores the importance of task-aligned pretraining and sufficient supervision. By first training on a diverse, high-resource dataset and then adapting to the smaller in-domain *SMARTSpan*, the model effectively generalizes and avoids overgeneration.

```
"24": [
  "switch from casual walking to brisk walking, maintaining
  a daily goal of 10, 000 steps (confidence 7/10)",
  "reduce the overall quantity of food consumed by 25 %,
  confidence 7/10",
  "reduce blood sugar"
  ]
```

K Output for `Mistral-Nemo-Instruct-2407` Fine-Tuned on *MultiSpanQA*

Examples of HTML tags:

```
"m20r1ol7e07go0pqexqq": [
  "</div >",
  "France",
  "Galician",
  "Germany",
  "Italy",
  "Portugal",
  "Spain",
  "UK",
  "the Americas"
],

"1ove66khudx8x16s6i81": [
  "</div>",
  "<br />",
  "<div class=\"column col-3 col-last\">",
  "south of Vancouver , British Columbia",
  "the Hell on Wheels set",
  "the University Women 's Club of Vancouver"
]
```

Example of stray characters:

```
"v9ohyt3b06g4y03cqysk": [
  ")",
  "CANOE",
  "OCEAN",
  "activity",
  "agreeableness",
  "assertiveness",
  "conscientiousness",
  "excitement seeking",
  "extraversion",
  "gregariousness",
  "neuroticism",
  "openness to experience",
  "positive emotions",
  "warmth"
],

"hblv00fzqtwavy96o582": [
  ":",
  "Don Felder",
  "Don Henley",
  "Glenn Frey",
  "Joe Walsh",
  "the Eagles",
  "the band"
]
```

Other examples:

```
"503kdfmuelcn8in1eggi": [
  "### Input:",
 "Ansel Elgort ( born March 13, 1994 ) is an American actor and
  musician. He is best known for his leading role in the 2014
 film The Fault in Our Stars, and for playing the title character
  in the 2017 film Baby Driver. He has also appeared in the
  films Divergent ( 2014 ), The Divergent Series: Insurgent
  ( 2015 ), and The Divergent Series: Allegiant ( 2016 ), and
  in the 2017 film The Black Balloon. Elgort was born in New
  York City, the son of Arthur Elgort, a fashion photographer,
  and Grethe Barrett Holby, an artist. His father is Jewish,
  while his mother is of Norwegian and English descent. He
  was named after his great - grandfather, the Swedish - born
 artist Andrew Olsson. He grew up in Manhattan and attended the
 Professional Children 's School. He began acting at age 9, and
 began taking acting classes at the Stagedoor Manor summer camp.
  He made his stage debut in the 2002 production of the musical
  Guys and Dolls at the Kennedy Center in Washington, D.C. He
  appeared in a short film, The Scapegoat, in 2002, and in the
  film Daddy 's Home in",
  "Edgar Wright",
  "Eiza González",
  "Jamie Foxx",
  "Jon Bernthal",
  "Jon Hamm",
  "Kevin Spacey",
  "Lily James"
]
```

L Output for `Llama-3.2-3B` Fine-Tuned on *SMARTSpan*

Example of made up content:

```
Input: "final session. at the request of the client, the session
 was short and focused on summarizing past experience and
 suggestions for improvement. client reflects a desire to reduce
 cholesterol levels and medication intake and had hoped for an
 intervention with immediate results. suggests that health
 coaching sessions could be more directive and structured. finds
 the educational materials useful but did not access them, stating
 they felt like \"common sense\" and expressing a general
preference for not using the phone frequently. describes managing
 diet and exercise in a strict way and felt demoralized when
 cholesterol levels did not improve. hc reflected that small
 changes were still observed, such as greater variety in diet.
client shared that \"resistance is futile\" and accepts the need
 for long-term medication. hc affirmed that lifestyle may be
 helping prevent further worsening despite lack of measurable
improvement. client reported bugs in the app: receiving multiple
 medication reminders one hour after the original, and not
 receiving notifications while overseas even though login was
 possible. hc closed the session with well wishes and
 encouragement to continue prioritizing health."

Output: [
  "explore ways to improve diet and exercise",
  "explore ways to improve health literacy",
  "explore ways to improve medication adherence",
  "explore ways to increase medication adherence",
  "explore ways to reduce cholesterol levels"
]
```