Return to Submission Overview

ID: 1435

A Lightweight Change Talk Classifier for Enhancing AI-driven Motivational Interviewing

Akshaye Shenoi, MSc; Tianze Li, BComp, Singapore-ETH Centre; Tobias Kowatsch, PhD, University of St. Gallen; Florian von Wangenheim, PhD, ETH Zurich; Jacqueline L. Mair, PhD, Singapore-ETH Centre

Please rank the top three most appropriate categories (Primary indicates best fit)

Primary

Machine Learning and Artificial Intelligence

Secondary

Behavioral Health (e.g., sleep, diabetes, cancer)

Tertiary

Web-based interventions

Context: Motivational Interviewing (MI) is a collaborative client-centred counselling approach aimed at eliciting the client's own reasons for changing behaviour. A core component of this approach is the counsellor's ability to identify Change Talk (CT)—client language that signals commitment or readiness to change. Recent advancements in language models have paved the way for scalable and accessible MI counselling via AI conversational agents. However, these agents are not explicitly trained

to identify CT and rely on simple prompting techniques to apply appropriate MI strategies. Detecting CT as a first step before generating a response—while also limiting processing latencies—could potentially enhance the agent's strategy selection, in turn enabling high-fidelity and responsive AI-driven MI counselling at scale.

Methods: We developed a classification pipeline using a Sentence Transformer (ST) backbone with a multilayer-perceptron projection head to categorize client utterances as change, sustain, or neutral talk. In particular, the MPNet ST was selected to optimize for low-latency and computationally efficient inference. Full finetuning was performed on an open MI dataset consisting of 241 client utterances. Furthermore, we systematically varied the context window size—i.e., number of previous client and counsellor utterances provided to the model—to assess impact on accuracy. Performance was measured using 5-fold cross-validation, with the mean macro F1-score and the average latency per classification serving as the evaluation metrics. We benchmarked the ST classifier against medium-sized, commodity-GPU capable LLMs (e.g., Gemma 4 26B) that were prompted with instructions to identify CT. The computational requirements of both approaches were also compared.

Results: The best-performing ST classifier attained a macro F1-score of 0.740 and an average latency of 10.56 ms using a 3-utterance context window. In comparison, the LLM outperformed the ST classifier on classification performance, recording a macro F1-score of 0.845 using a 4-utterance context window. However, the LLM incurred an average latency of 1288 ms per inference. Varying the context window size did not have a significant impact on performance across both approaches. The ST classifier operated on a 0.1 billion parameter architecture whereas the LLM used 35 billion parameters.

Implications: Foremost, our findings demonstrated the feasibility of using lightweight Sentence Transformers to classify client CT utterances in MI sessions. When compared against mid-sized LLMs, we observed a clear trade-off between classification accuracy and inference latency: despite higher macro F1-scores, the longer latencies incurred by LLMs might potentially pose as a bottleneck in real time sessions. Ultimately, integrating lightweight CT classifiers into AI counselling agents may improve MI strategy selection thereby enabling high-fidelity yet responsive MI counselling.

© 2026 ISRII