Using large language models to evaluate the offer of options in clinical encounters by focusing on an item of the Observer OPTION-5 measure of shared decision-making

Sai Prabhakar Pandi Selvaraj, Renata West Yen, Rachel Forcino, Glyn Elwyn

Submitted to: JMIR AI
on: February 26, 2024

Table of Contents

Using large language models to evaluate the offer of options in clinical encounters by focusing on an item of the Observer OPTION-5 measure of shared decision-making

Sai Prabhakar Pandi Selvaraj[1] MS, BTECH; Renata West Yen[2] MPH, PhD; Rachel Forcino[3] BA, MSc, PhD; Glyn Elwyn[2] BA, MB BCh, MSc, PhD

[1]Generative AI Tempus AI, Inc. Redwood Shores US
[2]The Dartmouth Institute for Health Policy & Clinical Practice Dartmouth College New Hampshire US
[3]University of Kansas School of Medicine University of Kansas School of Medicine Kansas City US

Corresponding Author:
Sai Prabhakar Pandi Selvaraj MS, BTECH
Generative AI
Tempus AI, Inc.
Suite 450
255 Shoreline Drive,
Redwood Shores
US

Abstract

Methods: We used a dataset of 287 clinical encounter transcripts of women diagnosed with early breast talking with their surgeon to discuss treatments. Each transcript had been previously scored by two researchers using OO5 (0 to 4 scale). We set up two rules-based baselines, one random and one using trigger words, and classified option talk instances using GPT-3.5 Turbo, GPT-4, and PaLM 2. To develop and compare the performance of these models, we randomly selected 16 transcripts for additional human annotation focusing on option talk instances (binary). To assess performance, we calculated Spearman correlations (rS) between the researcher-generated scores for item 1 for the remaining 271 transcripts and the item 1 instances predicted by the LLMs.

Results: We observed high levels of correlation between the LLMs and researcher-generated scores. GPT-3.5 Turbo with a few-shot example had an rS=0.60 (P<.001) with the mean of the two scorers. Other LLMs had slightly lower correlation levels.

Discussion: The LLMs, particularly GPT-3.5 Turbo with few-shot examples, demonstrated superior performance in identifying option talk instances compared to baseline models. GPT-3.5 Turbo demonstrated the best performance, achieving higher precision and recall.

Conclusions: Further improvements in score correlations may be possible through improvements in and better understanding of LLMs.

(JMIR Preprints 26/02/2024:57790)
DOI: https://doi.org/10.2196/preprints.57790

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Using large language models to evaluate the offer of options in clinical encounters by focusing on an item of the Observer OPTION-5 measure of shared decision-making

Sai P. Selvaraj[a1], Renata W. Yen[b], Rachel Forcino[c], Glyn Elwyn[b]

[a]Corresponding author
Tempus AI, Inc.
Generative AI,
600 W Chicago Ave #510,
Chicago IL 60654 USA
Sai Prabhakar Pandi Selvaraj, Senior Machine Learning Scientist, aps.prabhakar@gmail.com
+14128059136

[b]The Dartmouth Institute for Health Policy & Clinical Practice,
Dartmouth College
New Hampshire NH 03756 USA
Glyn Elwyn, BA, MB BCh, MSc, PhD, Professor, glynelwyn@gmail.com
Renata Yen, MPH, PhD, Research Scientist, Renata.West.Yen@dartmouth.edu

[c]University of Kansas School of Medicine
Department of Population Health
3901 Rainbow Boulevard
Kansas City, KS 66160 USA
Rachel Forcino, BA MSc PhD, Assistant Professor, rforcino@kumc.edu

Abstract

Introduction: Human assessment of clinical encounter recordings using observer-based measures of shared decision-making, such as Observer OPTION-5 (OO5), is expensive. In this study, we aimed to assess the potential of using large language models (LLMs) to automate the rating of the OO5 item focused on offering options (item 1).

Methods: We used a dataset of 287 clinical encounter transcripts of women diagnosed with early breast talking with their surgeon to discuss treatments. Each transcript had been previously scored by two researchers using OO5 (0 to 4 scale). We set up two rules-based baselines, one random and one using trigger words, and classified option talk instances using GPT-3.5 Turbo, GPT-4, and PaLM 2[2]. To develop and compare the performance of these models, we randomly selected 16 transcripts for additional human annotation focusing on option talk instances (binary). To assess performance, we calculated Spearman correlations (r_s) between the researcher-generated scores for item 1 for the remaining 271 transcripts and the item 1 instances predicted by the LLMs.

Results: We observed high levels of correlation between the LLMs and researcher-generated scores. GPT-3.5 Turbo with a few-shot example had an r_s=0.60 (P<.001) with the mean of the

1 Work done independently outside Tempus AI, Inc.
2 Codes used in the work will be release upon acceptance.

two scorers. Other LLMs had slightly lower correlation levels.
Discussion: The LLMs, particularly GPT-3.5 Turbo with few-shot examples, demonstrated superior performance in identifying option talk instances compared to baseline models. GPT-3.5 Turbo demonstrated the best performance, achieving higher precision and recall.
Conclusions: Further improvements in score correlations may be possible through improvements in and better understanding of LLMs.

Keywords: Generative AI; LLM; GPT; Option talk; Shared Decision-Making

Introduction

Shared decision-making (SDM) leads to improved outcomes that include patients having a better knowledge of treatment options, lower utilization rates, lower costs, and improved patient-centered health outcomes [1–4]. Policy initiatives in the US, such as the Merit-based Incentive Payment System (MIPS) and Medicare Access and CHIP Reauthorization Act of 2015 (MACRA), support the SDM approach, and the Centers for Medicare and Medicaid Services incentivize SDM with evidence-based patient decision aids in a range of clinical contexts [5].

However, despite the move to incentivize SDM, there is agreement that better measures are required [6]. Patient-reported experience measures (PREMs) of SDM have been developed [7,8] but are not widely implemented, and response rates are often low [9]. Moreover, these PREMs exhibit desirability bias and strong ceiling effects [10]. Observer-based measures (OMs) are more reliable indicators that SDM has been accomplished. Observer-based assessments also overcome the limitations of PREMs: PREMs tend to show high SDM with low variation between clinicians, while OMs—which are closer to “ground truth” by being based directly on the recorded clinical encounter—typically find low levels of SDM accompanied by high levels of variation at the clinician level [11]. However, OMs are labor-intensive, time-consuming, and expensive because they rely on the training and maintenance of calibrated human judges who make assessments of recordings or transcripts. Moreover, the use of OMs has been so far restricted to research studies [12,13].

Observer OPTION-5 (OO5), a widely used OM, based on the collaborative deliberation model [14], has demonstrated good validity in prior research [13,15–19]. The recommended assessment method advocates the use of two independent raters to judge SDM performance, which is time-consuming and expensive. There is an opportunity, therefore, to reduce the training burden and cost of using OMs by automating the approach, using the developments in natural language processing (NLP) and artificial intelligence (AI). We therefore wish to automate the assessment of SDM, using natural language processing (NLP) methods to substitute for human evaluators. Automation would: (1) make research into SDM more feasible, enable the analysis of large samples of clinical encounters; (2) make OM easier to use as an outcome measure in trials; and (3) potentially create the opportunity to give feedback to practitioners about their accomplished levels of SDM.

Large Language Models (LLMs) [20] are a significant development in the field of artificial intelligence (AI). These models, characterized by large neural network architectures, have redefined prior benchmarks for natural language processing (NLP). LLMs have capabilities that go beyond general tasks and are being used for many applications in healthcare, including diagnostics, clinical decision support, and medical literature interpretation [21–23]. Recent research is exploring the potential of LLMs to replace and potentially supplant existing machine

learning models across various domains [24,25]. Although not fully understood, the phenomenon of emergent behavior appears when large models exhibit abilities they weren't explicitly trained to achieve. For example, language models originally trained on broad datasets, demonstrate unexpected proficiency in extracting nuanced medical information from unstructured clinical notes, showcasing the potential for unforeseen NLP capabilities in healthcare applications.

OpenAI's Generative Pre-trained Transformer (GPT) series is a well-known model [26]. GPT-3.5 underwent training on a 45GB corpus of text. It has capabilities that range from complex translation tasks to the generation of computer codes [20,27–30]. PaLM (Pathway-based Language Model), a 540 B parameter model from Google, was developed explicitly as a generalist few-shot learner [31]. An updated version, PaLM2, is available [32]. The LLMs are capable of performing task execution without prior demonstrations or training, so-called zero-shot tasks [33], and using few-shot methods [20], task execution using only a few example demonstrations [31].

Given the rapid development in the capabilities of these models, we wanted to assess whether an LLM could be used to detect specific instances of speech acts within transcripts of clinical encounters recorded in healthcare settings and then compare the model's detection level to that of human assessors who had previously done the same the task using the Observer OPTION-5 item measure. To simplify the task, we focused on the first item of the measure that asks the assessors to identify if a clinician indicates to a patient that options exist that need to be considered.

Methods

Dataset Utilized

We used a sample of encounter transcripts derived from recordings obtained during a three-arm randomized trial based in four cancer centers [34]. The trial compared the impact of different versions of conversation aids designed to facilitate discussing treatment options that were assessed in this analysis [34]. We recorded encounters between patients and their surgeons discussing the surgical management of recently diagnosed early-stage breast cancer. Breast surgeons in the intervention arms had been trained in the use of a conversation aid (Option Grid) that compared two surgical options for early breast cancer. These conversations were focused on comparing breast-conserving surgery with radiation versus removal of the breast (i.e., mastectomy). Other therapies were also sometimes discussed, including chemotherapy, radiation, and genetic testing. Patients knew of their breast cancer diagnosis prior to the appointment. We obtained ethical approval to undertake the analysis (IRB WMM STUDY00030157 R).

Transcript Preparation

The conversations were transcribed as separate speaker turns by human transcribers. We used a natural language processing library application called spaCy to split the speaker turns into individual lines, based on punctuations inserted by the transcribers.

Detecting Instances of Option Talk (Item 1 of Observer OPTION-5 Measure)

We developed an automated system to detect instances within clinician-patient conversations where options or alternatives were discussed or where clinicians offered options to patients: we refer to these as “option talk instances”. Examples of option talk instances are provided in (Table 1). These option talk instances would correspond to a positive score being given to the first item of the Observer OPTION-5 measure [13]:

Item 1 (Observer OPTION-5) For the health issue being discussed, the clinician draws attention to or confirms that alternate treatment or management options exist or that the need for a decision exists. If the patient rather than the clinician draws attention to the availability of options, the clinician responds by agreeing that the options need deliberation.

Table 1. Examples of clinicians using option talk instances.

Clinician A: “You can remove just the tumor, which is called a lumpectomy - that’s what this looks like here versus doing mastectomy where you remove the whole breast …”
Clinician B: Line 1: “We’ll talk about what the options are now. With that, you can make a decision.” Line 2: Patient B: “Ok, let's do it” Line 3: “You have three options for the treatment …”

Option Talk Instance: Human Annotation

From the 287 trial transcripts (see (Table A1) in (Multimedia Appendix 1)), we randomly selected 16 conversations that would be evaluated by two researchers (RWY and GE) for the existence of option talk instances. We used a random number generator to select conversations from the different trial sites (5 out of 110 from site 1, 3 out of 46 from site 2, 2 out of 8 from site 3, and 5 out of 123 from site 4) to approximately correspond to the conversations recorded at each trial site. From the sixteen transcripts, 9 were from an intervention arm (Option Grid or Picture Option Grid), and 7 were from the usual care arm [34]. RWY and GE annotated the 16 transcripts to identify option talk instances, talk segments where clinicians describe the existence of more than one option. From the 16 conversations, we used two for experimenting with the LLMs, and the remaining 14 were used to test the automated evaluations.

Observer OPTION-5 Scores

All 287 encounter transcripts had been previously evaluated by trained researchers using the Observer OPTION-5 measure, where each item is scored between 0 and 4 [35]. A score of 4 indicated the highest possible score and was interpreted as a positive instance of option talk.

Automated Evaluations of the Selected Transcripts

To automate the identification of option talk instances, we used large language models (LLMs) to detect instances where clinicians state that a treatment option exists. LLMs may complete tasks without training data. We achieve this by including a description of the task in addition to the conversation that needs to be analyzed as input to the LLM models (see (Table A2) in the (Multimedia Appendix 1)). We benchmarked different versions of LLMs, from OpenAI and Google,

against: 1) baselines like keyword-based models using trigger words, see below for more details, and, 2) by comparing them to the annotated instances of option talk identified by human annotators (GE and RWY).

We instructed the LLMs to identify transcript lines that were positive for option talk instances. Lines not identified as positive were considered negative for option talk instances. We use three metrics, precision, recall, and F1 score, to evaluate the classification performance of the LLM [36]. Precision is the ratio of correctly identified option talk instances (true positive, TP) to all option talk instances identified by the classifier, (whether true or false positive (FP)), where Precision = True Positive / (True Positive + False Positive). Recall (R) refers to the ability to find all of the positive option talk instances, where Recall = True Positive / (True Positive + False Negative). Recall is, therefore, the ratio of correctly identified option talk instances within the transcript to the total of those that are true positives and false negatives. The F1 score represents a balanced measure of precision and recall, it is the harmonic mean of precision and recall. Higher values for precision, recall, and F1 score indicate better performance for the task. The F1 score is crucial in imbalanced datasets, as here, where option talk instances are infrequent. Where TP is the True Positive, FP is the False Positive, and FN is the False Negative, these metrics are represented by the following formulae: Precision=TP/(TP+FP), Recall=TP/(TP+FN), F1 Score=((P-1 + R-1)/2)-1. In assessing the classification model performance, we measured the instances of option talk by a clinician. To illustrate, let us consider an example conversation transcript T, with n transcript lines arranged in chronological order: $T=[l_1,l_2,l_3,...l_n]$, where $l_1,l_2,l_3,$ and l_n represent the first, second, third, and the last lines of the transcript respectively. In this example, let's assume the human annotation process identifies 7 ground truth option talk instances $[l_{G5},l_{G6},l_{G9},l_{G13},l_{G15},l_{G19}]$, and let's assume the model predicts these 4 lines $[l_{P7},l_{P11},l_{P12},l_{22}]$ as positive for option talk instances. Note that the lines $\{l_6,l_{G6},l_{P6}\}$ all refer to the 6th line in the transcript T.

In situations where an option talk instance extends across multiple lines in the transcript, we grouped the option talk instances based on their proximity to each other. We use a proximity threshold of 1, i.e., if two option talk instances on l_6 and l_8 are within 1 transcript line of each other, we classify them as the same cluster. For example, in (Table 1), Lines 1 and 3 from Clinician B are option talk instances, and Line 2 is not. Since Lines 1 and 3 are within 1 transcript line of each other, we will cluster them together as {Line 1, Line 2, Line 3} one option talk instance cluster.

In the above example, therefore, the ground truth instances $[l_{G5},l_{G6},l_{G9},l_{G13},l_{G15},l_{G19}]$ are classified as three clusters $[\{l_{G5},l_{G6},l_{G7},l_{G8}\},\{l_{G13},l_{G14},l_{G15}\},\{l_{G19}\}]$. Similarly, the predicted instances $[l_{P2},l_{P7},l_{P11},l_{P12},l_{P22}]$ are classified in four clusters $[\{l_{P2}\},\{l_{P7}\},\{l_{P11},l_{P12}\},\{l_{P22}\}]$. Clustering enables the model to account for the variability in the lengths of clinicians' option talk instances. We also considered two option talk instance clusters from ground truth and prediction to be a match (correct predictions) if they either:

1. Share at least one overlapping option talk line between them, or
2. If the cluster ranges are contiguous, even if they do not have an overlapping option talk instance between them, for example, $\{l_{P11},l_{P12}\}$ and $\{l_{G13},l_{G14},l_{G15}\}$ are contiguous clusters.

We adopted this method because the precise position of the option talk instance within the length of the transcript is not important, and its exact pinpointing by a human annotator would be subjective. This allows the ground truth and classifier prediction clusters to marginally differ, yet be considered a positive match. In our example, comparing the ground truth clusters $[\{l_{G5},l_{G6},l_{G7},l_{G8}\},\{l_{G13},l_{G14},l_{G15}\},\{l_{G19}\}]$ to the predicted clusters $[\{l_{P2}\},\{l_{P7}\},\{l_{P11},l_{P12}\},\{l_{P22}\}]$ we find that:

1. Clusters $\{l_{G5}, l_{G6}, l_{G7}, l_{G8}\}$ and $\{l_{P7}\}$ are a match as they both contain line l_7. Therefore $\{l_{P7}\}$ is a True Positive prediction.
2. Similarly, clusters $\{l_{G13}, l_{G14}, l_{G15}\}$ and $\{l_{P11}, l_{P12}\}$ are also a match as they form a contiguous range. Hence $\{l_{P11}, l_{P12}\}$ is also a True Positive prediction.
3. Clusters $\{l_{P2}\}$ and $\{l_{P22}\}$ are not a match with any of the ground truth clusters. Hence they are False Positive predictions.
4. Similarly, $\{l_{G19}\}$ is not a match with any of the predicted clusters. Hence by not identifying l_{G19} the model has a False Negative assessment.

In the above hypothetical example, #True Positive = 2, #False Positive = 2, and #False Negative = 1. Hence, using the formulae, Precision = 0.5, Recall = 0.667, and F1 = 0.572. These modifications enable a robust assessment of the model's effectiveness in identifying option talk instances, accommodating variations in the length and position of such talk.

Establishing Baselines

To evaluate the usefulness of the machine learning models, we established two baselines for the option talk identification task. The random baseline involves randomly predicting 2.9% (which is the percentage of annotated option talk instances in our dataset) of the transcript lines as positive for option talk instances to determine whether the models perform better than chance. The trigger word baseline identifies pre-specified words ('option', 'choice', 'decision', 'either') that clinicians might use when talking to patients about options and classifies a line as a positive if the trigger words (or their plural forms) are present. These two baselines were used to compare the machine learning model's performance against these rule-based approaches.

Option Talk Instance Identification by LLMs

We used LLMs available from commercial vendors OpenAI and Google. From OpenAI we used GPT-3.5 Turbo and GPT-4. The GPT-3.5 Turbo we used had two versions released on March 1, 2023, and June 13, 2023. Similarly, GPT-4 is a 1.76 trillion parameter model with two stable versions released on March 14, 2023, and June 13, 2023. From Google, we used PaLM 2's version "text-bison@001" released on June 7, 2023. Although both OpenAI and Google have not disclosed what changes are made between the two versions, we assume that the main change is that the models were trained on updated datasets to give access to new knowledge. We did not experiment with other available variations of GPT and PALM LLMs.

LLMs with emergent behaviors, when given a natural language description of the task, enable tasks to be completed without a training phase [37], using task prompts. We designed task prompts using trial-and-error with domain experts' inputs on the two conversations that were not used during testing. The task prompt instructs the LLM to identify the most relevant five option talk instances and then to structure and explain its outputs in a programmatically parseable manner. To improve performance, we provide a system-level prompt ("You are a medical reviewer.") and furnish two few-shot examples that demonstrate the task with a made-up conversation and output [Citation error]. More details are provided in (Table A2) in the (Multimedia Appendix 1). The outputs/predictions that don't conform to the format are ignored, for example, if the output is <13> rather than <Line ID: 13>.

Given that LLM model performance decreases as the context length increases, we divided the transcripts into sections of 80 transcript lines, formatted as shown in (Table A2) in the (Multimedia Appendix 1). The outputs were then aggregated to obtain predictions for each encounter.

Correlation with Human-generated Observer OPTION-5 Scores

Using the 271 encounter transcripts that were not used during the development stages, we correlated the model predictions of option talk instances with the previously obtained researcher-assessed Observer OPTION-5 scores. After counting the option talk instances predicted by the model for each transcript, we followed standard practice and removed outlier conversations (>3 standard deviations) from this analysis, considering the number of transcript lines and the number of predicted option talk instances. Given the non-normal distribution of the data, we calculated a Spearman correlation using the remaining transcription scores.

Cost Analysis

We also performed cost analysis on the models we have used in our study to provide cost insights for adopting them in similar applications. LLMs are deployed on advanced acceleration hardware. The cost of running the LLMs depends on the size of the model, the number of tokens in the user's input sentence, and the number of tokens in the sentences generated from the model, where tokens are a predefined combination of characters. The cost per 1000 tokens in the input and the output are $0.03 and $0.06 for GPT-4, and ≈$0.001 and ≈$0.002 for GPT-3.5 turbo and PaLM 2. We report the average cost per conversation for option talk instance prediction for each of the models used in our work. We also compare the cost for these predictions with the cost required for the manual OO5 scoring using the trained researchers.

Results

Performance comparisons of the two baselines, LLM variants, with and without using few-shot examples, are shown in (Table 2). Text Word Baseline obtained higher scores than the Random Baseline. All the LLM models perform better than the random baseline and the trigger-word baselines by a good margin. Overall we got the best performance from GPT-3.5 Turbo (03/01) with few-shot examples. Among models with no examples, PaLM 2 obtained higher precision but lower recall compared to the others. Before adding few-shot examples, GPT-4 performed better than GPT-3.5 Turbo as expected, because GPT-4 is a larger model. However, after adding few-shot examples GPT-3.5 Turbo performed even better, yet GPT-4's performance remained unchanged. Few-shot examples improve performance most of the time, but there are cases when examples have the opposite effect. We found that a more recent LLM version of the LLM model did not outperform the previous model. GPT-3.5 Turbo (June 2023) had a significant decrease in performance compared to the March 2023 version with few-shot examples. Although PaLM 2 lacks recall, it performed better in terms of the F1 score without the few-shot examples. We show some examples of errors made by the best model GPT-3.5 Turbo (03/01) with a few-shot examples model in (Table 3).

Table 2. Option Talk Instance Identification: Comparative Model Performance (Sample N=14 test conversations).

Model Used	*Precision*	*Recall*	*F1*
Random Baseline	0.01	0.05	0.001
Trigger Word Baseline	0.15	0.67	0.21
GPT-3.5 Turbo (03/01) + Zero-shot	0.12	0.82	0.20
GPT-3.5 Turbo (06/13) + Zero-shot	0.12	0.82	0.20

GPT-4 (03/14) + Zero-shot	0.17	0.87	0.28
GPT-4 (06/13) + Zero-shot	0.16	0.85	0.27
PaLM 2 (06/07) + Zero-shot	0.21	0.69	0.29
GPT-3.5 Turbo (03/01) + Few-shot	*0.24*	*0.90*	*0.38*
GPT-3.5 Turbo (06/13) + Few-shot	0.16	0.87	0.28
GPT-4 (03/14) + Few-shot	0.17	0.77	0.27
GPT-4 (06/13) + Few-shot	0.17	0.82	0.28
PaLM 2 (06/07) + Few-shot	0.12	0.70	0.17

Table 3. Comparing Human Annotation with Model Errors: GPT-3.5 Turbo Few Shot (03/14/2023).

Excerpt from transcript	*Annotator*	*Model Prediction*	*Error Type*
Clinician reading from the conversation aid: "Will my lymph nodes be removed? I sample the lymph nodes with either surgery, so when you're asleep, I go in underneath the arm, and I take one or two lymph nodes to see if the cancer has spread there. I do that with both." *Clinician reading from the conversation aid: "Will I need chemotherapy?* Well, we already talked about that. For you, you're definitely going to need it because you have the HER2".	Negative	Positive	False Positive (Does not contain an option talk instance)
"If there is cancer, and if positive for cancer in the sentinel node or the clipped node, then the standard treatment would be to do an axillary dissection. That's the surgery. Axillary dissection would be standard, and the radiation therapy, this would be ... standard would be radiation therapy."	Negative	Positive	False Positive
"Yes, we'll go through these options here, but at the end, I'm going to recommend what I think we need to do, okay? So, this goes through the two different options talking about lumpectomy, just taking out the lump versus a mastectomy, removing the whole breast."	Positive	Negative	False Negative
"I can give you my impression and tell you	Positive	Negative	False

what I think, and you can decide whatever you want, so that's fine."			Negative

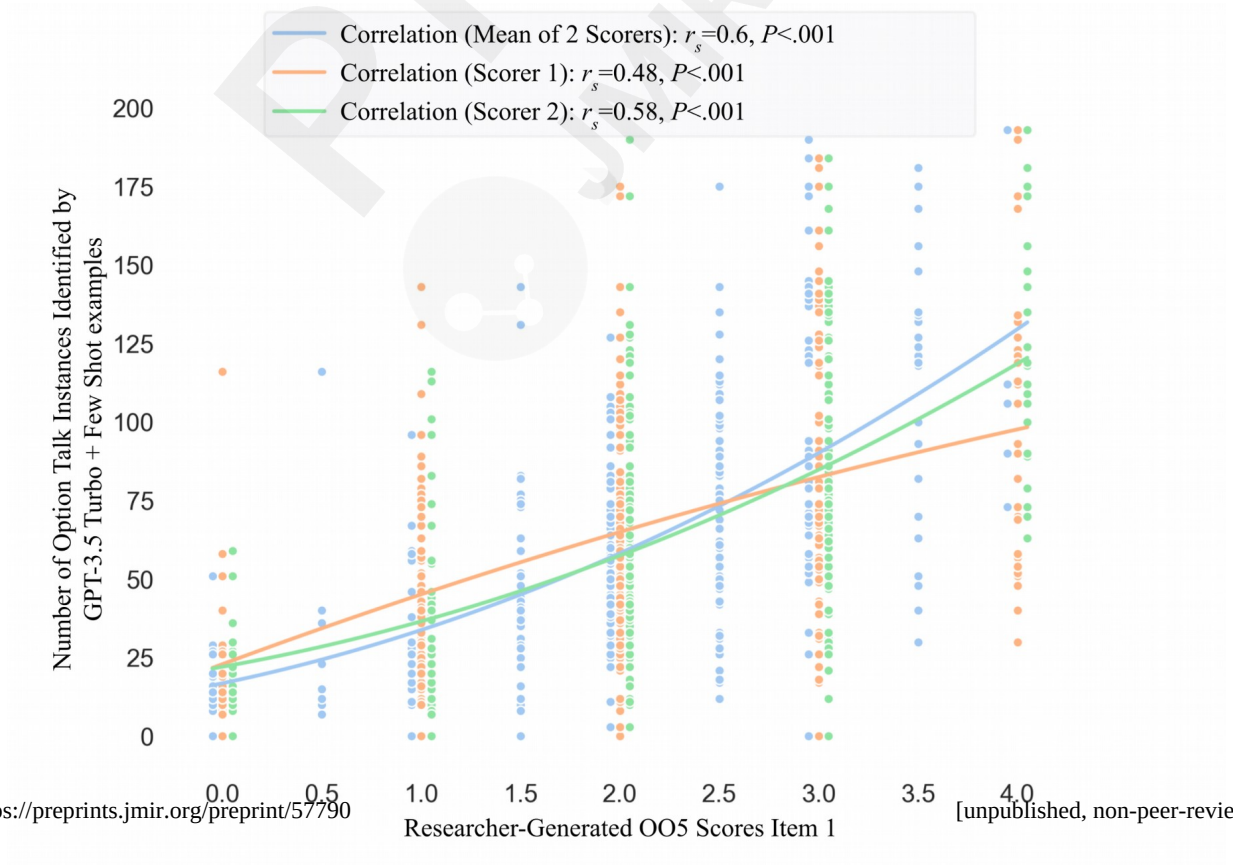
Correlation with Human-generated Observer OPTION-5 Scores:

The results of the Spearman correlation (r_s) are shown in Figure 1, which shows data from the best performing model, GPT-3.5 Turbo (03/01) with Few-shot examples, against the researcher-generated OO5 scores on 266 transcripts, after removing 5 outlier transcripts. The correlation coefficients were 0.47 and 0.57 (P<.001) between the model-predicted option talk instances and the individually generated OO5 Item 1 score. The correlation coefficient between the model and the averaged OO5 scores of the two researchers was 0.60 (P<.001).

Cost Analysis:

Our dataset contains conversations with an average duration of 24 minutes. For the option talk instance prediction, on average, it costs $0.017 per conversation with GPT-3.5 Turbo and $0.7 for GPT-4 API calls with few-shot examples. For the manual coding process, we paid the researchers an equivalent of $25 per hour in 2024, and the time they required to code a conversation is approximately equal to the duration of the conversation, hence the average cost per conversation is $10 per conversation. Since in our work, we used two researchers for double scoring we spent $20 per conversation to obtain manual scores.

Figure 1. Spearman Correlation between researcher-generated Observer OPTION-5 Scores and GPT-3.5 Turbo (03/01/2023) + Few-shot predictions.



Discussion

Principal Findings

The LLMs, particularly GPT-3.5 Turbo with few-shot examples, outperformed the baseline random and trigger-word models when identifying option talk instances. When comparing the performance of a GPT-3.5, GPT-4, and PaLM 2, GPT-3.5 Turbo (03/01) with few-shot examples demonstrated the best overall performance on the task provided, achieving higher precision and recall. Additionally, the correlation analysis between LLM predictions and researcher-generated OPTION-5 scores showed moderate to strong correlations, indicating alignment between model predictions and human scoring using item 1 of the OO5 measure. The cost of using an LLM such as GPT-3.5 turbo for this specific task was relatively low but does not represent the true cost of collecting and preparing the data for analysis.

Strengths and Weaknesses of the Method

The LLMs, especially GPT-3.5 Turbo with few-shot examples, outperformed the baseline models that we had developed, suggesting the potential utility of LLMs in automating elements of speech in clinical encounters, and specifically the identification of option talk instances in conversations. The correlation analysis demonstrated a significant association between the LLM predictions and researcher-generated assessments, reinforcing the model's utility. Occasionally, LLMs, including PaLM 2 and GPT-4, produced outputs that deviated from instructions, highlighting the importance of careful construction of prompts and evaluating the model predictions. The reduced performance of more recent updates in the LLM points to the need to explore and assess the behavior of each version as they are made available.

Results in Context

Our findings contribute to the evolving landscape of natural language processing in healthcare contexts, and their application to a wide set of tasks [25]. The application of LLMs in analyzing healthcare conversations to detect discrete elements of talk between clinicians and patients is unexplored new ground [38,39]. Our demonstration that LLMs can detect speech acts such as option talk instances points to the potential to use this technology to measure approaches such as shared decision-making in healthcare contexts [40]. The trade-off between precision and recall, as observed in PaLM 2, underscores the importance of considering multiple factors when evaluating model capabilities. The unexpected performance of the Text Word Baseline underscores the influence of domain-specific speech patterns, adding depth to our understanding of baseline comparisons.

Implications

This study took one item of an existing OM for shared decision-making and showed that an LLM can identify speech elements within transcripts that correlate to a modest degree with a high score given by researchers. If the same could be achieved for the next four items of Observer OPTION-5, we could evaluate the degree of correlation for overall scores, and begin to consider whether the assessment of shared decision-making in clinical encounters could be automated if audio recordings

and transcripts were available. Such a development would point the way to other evaluations of communication within clinical encounters REF.

Collecting the recordings and having prior evaluation scores generated by two independent researchers for the Observer OPTION-5 measure required access to research data and processes that spanned several years. Undertaking this prior work is necessary before collaborating with a researcher capable of developing baseline models, and applying the capability of LLMs. Similarly, we can also get multiple outputs from LLM by using different samples or different models and prompt methods. Using data from patient encounters needs strategies to safely manage personal health information (PHI), ensuring safe anonymization, or high-security platforms that protect privacy in accordance with the relevant policies. Similar compliance with LLM API providers will be required using contracts such as Business Associate Agreements (BAAs).

Further progress will require sustained interdisciplinary work across technology services, which will require attention to security and data use agreements. Our work could lead to several areas of future research, including the task of fine-tuning the LLMs, customizing their capability for healthcare contexts, and investigating the optimal use of few-shot examples, with a focus on tailoring examples to individual LLMs for improved performance.

Conclusion

Our study lays some of the groundwork for leveraging LLMs to enhance SDM measurement in clinical encounters and the possibility of using such data for improving future patient-physician communication processes.

Acknowledgments

We thank Padhraig Ryan for his comments on this article.

Conflicts of Interest

Glyn Elwyn’s academic interests are focused on shared decision-making and coproduction. He owns copyright in measures of shared decision-making (collaboRATE) and care integration (integRATE), a measure of experience of care in serious illness (consideRATE), a measure of goal setting (coopeRATE), a measure of clinician willingness to do shared decision-making (incorpoRATE), an observer measure of shared decision-making (Observer OPTION-5 and Observer OPTION-12). He is the Founder and Director of &think LLC, which owns the registered trademark for Option GridsTM patient decision aids. He is an adviser to EBSCO Publishing, abridge, and Fora Health.

Abbreviations

OO5: Observer OPTION-5
NLP: Natural Language Processing
LLM: Large Language Model
SDM: Shared Decision-Making

GPT: Generative pre-trained transformer
PaLM: Pathway-based Language Model
TP: True Positive
FP: False Positive
FN: False Negative
R: Recall
F1: F1 Score
r_s: Spearman Correlation
OMs: Observer-Based Measures
PREMs: Patient-Reported Experience Measures

References

1. Clayman ML, Bylund CL, Chewning B, Makoul G. The Impact of Patient Participation in Health Decisions Within Medical Encounters. *Med Decis Making*. 2016;36(4):427-452.
2. Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making and patient outcomes. *Med Decis Making*. 2015;35(1):114-131.
3. Stacey D, Légaré F, Lewis K, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev*. 2017;4:CD001431.
4. Durand MA, Carpenter L, Dolan H, et al. Do interventions designed to support shared decision-making reduce health inequalities? A systematic review and meta-analysis. *PLoS One*. 2014;9(4):e94670.
5. Do payment programs incentivize shared decision making in US healthcare? *Patient Educ Couns*. 2023;113:107798.
6. Gärtner FR, Bomhof-Roordink H, Smith IP, Scholl I, Stiggelbout AM, Pieterse AH. The quality of instruments to assess the process of shared decision making: A systematic review. *PLoS One*. 2018;13(2):e0191747.
7. Dyer N, Sorra JS, Smith SA, Cleary PD, Hays RD. Psychometric properties of the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Clinician and Group Adult Visit Survey. *Med Care*. 2012;50 Suppl(Suppl):S28-S34.
8. Zill JM, Christalle E, Müller E, Härter M, Dirmaier J, Scholl I. Measurement of physician-patient communication--a systematic review. *PLoS One*. 2014;9(12):e112637.
9. Bergeson SC, Gray J, Ehrmantraut LA, Laibson T, Hays RD. Comparing Web-based with Mail Survey Administration of the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Clinician and Group Survey. *Prim Health Care Res Dev*. 2013;3. doi:10.4172/2167-1079.1000132
10. Male L, Noble A, Atkinson J, Marson T. Measuring patient experience: a systematic review to evaluate psychometric properties of patient reported experience measures (PREMs) for emergency care service provision. *Int J Qual Health Care*. 2017;29(3):314-326.
11. Couët N, Desroches S, Robitaille H, et al. Assessments of the extent to which health-care providers involve patients in decision making: a systematic review of studies using the OPTION instrument. *Health Expect*. 2015;18(4):542-561.
12. Elwyn G, Hutchings H, Edwards A, et al. The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. *Health Expect*. 2005;8(1):34-42.
13. Elwyn G, Tsulukidze M, Edwards A, Légaré F, Newcombe R. Using a “talk” model of shared decision making to propose an observation-based measure: Observer OPTION 5 Item. *Patient Educ Couns*. 2013;93(2):265-271.
14. Elwyn G, Lloyd A, May C, et al. Collaborative deliberation: A model for patient care. *Patient Educ Couns*. 2014;97(doi: 10.1016/j.pec.2014.07.027. [Epub ahead of print]):158-184.

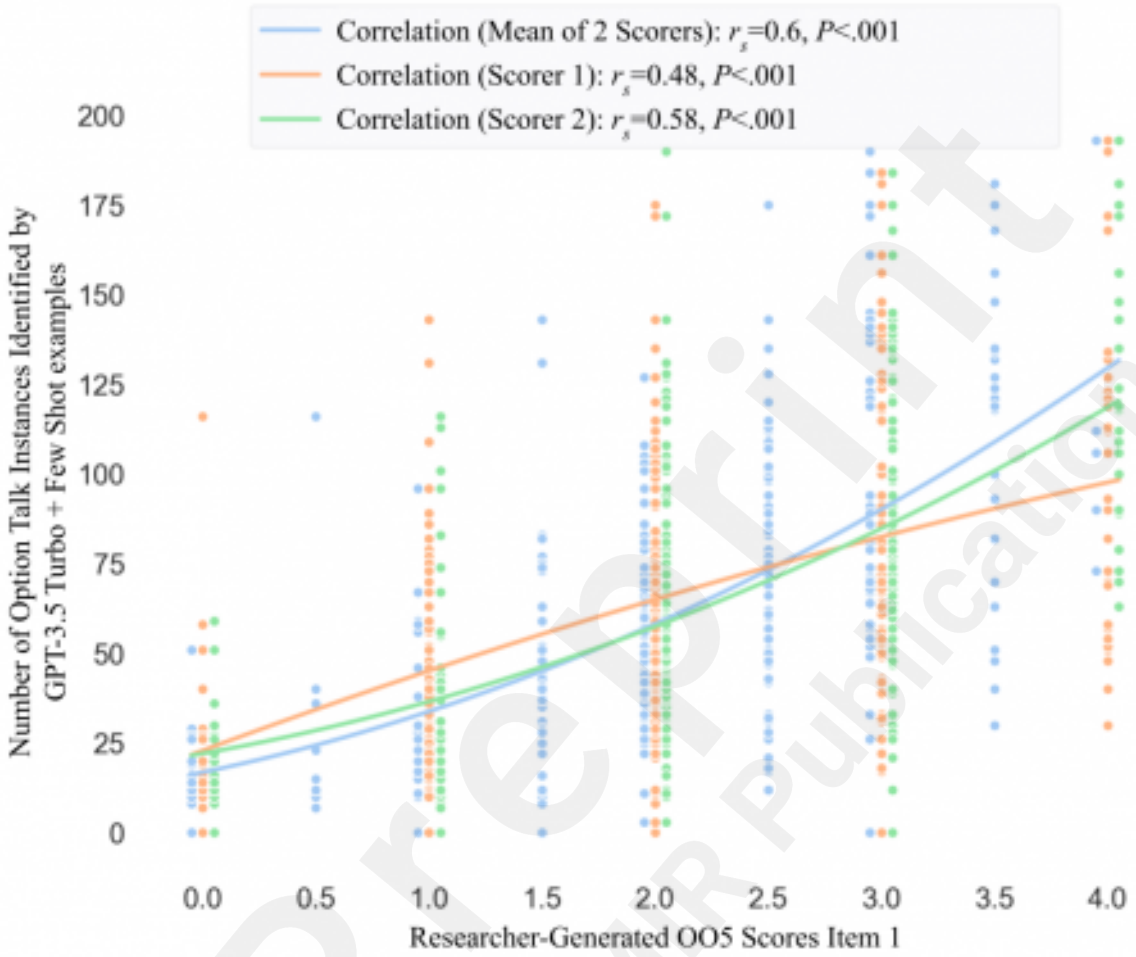
15. Barr PJ, O'Malley AJ, Tsulukidze M, Gionfriddo MR, Montori V, Elwyn G. The psychometric properties of Observer OPTION5, an observer measure of shared decision making. *Patient Educ Couns*. 2015;98(8):970-976.
16. Kölker M, Topp J, Elwyn G, Härter M, Scholl I. Psychometric properties of the German version of Observer OPTION5. *BMC Health Serv Res*. 2018;18(1):74.
17. Dillon EC, Stults CD, Wilson C, et al. An evaluation of two interventions to enhance patient-physician communication using the observer OPTION5 measure of shared decision making. *Patient Educ Couns*. 2017;100(10):1910-1917.
18. Vortel MA, Adam S, Port-Thompson AV, Friedman JM, Grande SW, Birch PH. Comparing the ability of OPTION(12) and OPTION(5) to assess shared decision-making in genetic counselling. *Patient Educ Couns*. 2016;99(10):1717-1723.
19. Stubenrouch FE, Pieterse AH, Falkenberg R, et al. OPTION(5) versus OPTION(12) instruments to appreciate the extent to which healthcare providers involve patients in decision-making. *Patient Educ Couns*. 2016;99(6):1062-1068.
20. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901.
21. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the Bar Exam. Published online March 15, 2023. doi:10.2139/ssrn.4389233
22. Shea YF, Lee CMY, Ip WCT, Luk DWA, Wong SSW. Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis. *JAMA Netw Open*. 2023;6(8):e2325000.
23. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. 2023;388(13):1233-1239.
24. Ramprasad S, Ferracane E. Generating more faithful and consistent SOAP notes using attribute-specific parameters. *Machine Learning for*. Published online 2023. https://proceedings.mlr.press/v219/ramprasad23a.html
25. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus*. 2023;15(5):e39305.
26. Zhou C, Li Q, Li C, et al. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *arXiv [csAI]*. Published online February 18, 2023. http://arxiv.org/abs/2302.09419
27. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
28. Tomašev N, Harris N, Baur S, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat Protoc*. 2021;16(6):2765-2787.
29. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. *arXiv [csLG]*. Published online August 16, 2021. http://arxiv.org/abs/2108.07258
30. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *NATO Adv Sci Inst Ser E Appl Sci*. 2021;11(14):6421.
31. Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling Language Modeling with Pathways. *arXiv [csCL]*. Published online April 5, 2022. http://arxiv.org/abs/2204.02311
32. Anil R, Dai AM, Firat O, et al. PaLM 2 Technical Report. Published online May 17, 2023. Accessed January 8, 2024. http://arxiv.org/abs/2305.10403
33. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst*. 2022;35:22199-22213.
34. Durand MA, Yen RW, O'Malley AJ, et al. What matters most: Randomized controlled trial of breast cancer surgery conversation aids across socioeconomic strata. *Cancer*. 2021;127(3):422-436.

35. Durand MA, Yen RW, O'Malley AJ, et al. What matters most: protocol for a randomized controlled trial of breast cancer surgery encounter decision aids across socioeconomic strata. *BMC Public Health*. 2018;18(1):241.
36. Powers DMW. Evaluation: from precision, recall and f-factor to roc. *Informedness, Markedness & Correlation (Tech Rep)*.
37. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. *Nat Hum Behav*. 2023;7(9):1526-1541.
38. Van Veen D, Van Uden C, Blankemeier L, et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *Research Square*. doi:10.21203/rs.3.rs-3483777/v1
39. Yao Z, Schloss BJ, Selvaraj SP. Improving summarization with human edits. *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Published online 2023. https://aclanthology.org/2023.emnlp-main.158.pdf
40. Yim WW, Ben Abacha A, Adams G, Snider N, Yetisgen M. Overview of the MEDIQA-sum task at ImageCLEF 2023: Summarization and classification of doctor-patient conversations⋆. Accessed January 21, 2024. https://ceur-ws.org/Vol-3497/paper-109.pdf

Supplementary Files

Spearman Correlation between researcher-generated Observer OPTION-5 Scores and GPT-3.5 Turbo (03/01/2023) + Few-shot predictions.

Figure 1 Spearman Correlations between researcher-generated Observer OPTION-5 Scores and GPT-3.5 Turbo predictions



Multimedia Appendixes

Dataset description and llm implementation details.
URL: http://asset.jmir.pub/assets/435693b86cdb113eaa4cd2c3c2c792fd.docx