Patient Education and Counseling 142 (2026) 109362

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.journals.elsevier.com/patient-education-and-counseling

Automating the Observer OPTION-5 measure of shared decision making: Assessing validity by comparing large language models to human ratings

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

[a] Anterior, Inc., New York, USA
[b] The Dartmouth Institute for Health Policy & Clinical Practice, Dartmouth College, Lebanon, USA
[c] University of Kansas School of Medicine, Department of Population Health, Kansas City, USA

ARTICLE INFO

Keywords:
Generative AI
LLM
GPT
Option talk
Shared decision-making

ABSTRACT

Objectives: Observer-based measures of shared decision rely on human raters, it is resource-intensive, limiting routine assessment and improvement. Generative artificial intelligence could increase the speed and accuracy of observer-based evaluation while reducing the burden. This study aimed to assess the performance of large language models (LLMs) from Gemini, GPT, and LLaMA family of models in evaluating the extent of shared decision-making between clinicians and women considering surgery for early-stage breast cancer.

Methods: LLM-generated scores were compared with those of trained human raters from a randomized controlled trial using the 5-item Observer OPTION-5 measure. We analyzed 287 anonymized transcripts of breast cancer consultations. A series of prompts were tested across models, assessing correlations with human scores. We also evaluated the ability of LLMs to distinguish high versus low encounters and the impact of inter-rater agreement on performance.[1]

Results: The scores for Observer OPTION-5 items generated by the GPT-4o and Gemini-1.5-Pro-002 correlated with human ratings (Pearson $r \approx 0.6$, p-value<0.01), representing ≈ 75–$80\,\%$ of the correlation observed between human raters themselves ($r = 0.77$). Providing detailed descriptions and examples improved the models' performance. The results also confirm that the models could distinguish high- from low-scoring encounters, with an independent-samples t-test showing a large and significant separation between the two groups ($t > 10$, $p < 0.01$).

Conclusions: Based on the breast cancer surgery dataset we explored, LLMs can evaluate aspects of clinician-patient dialog using existing measures, providing the basis for the development and fine-tuning of prompts. Future work should focus on generalizability, larger datasets, and improving model performance.

Practice implications: The prospect of being able to automate the assessment of shared decision-making opens the door to rapid feedback as a means for reflective practice improvement.

1. Introduction

Shared decision-making (SDM) is a collaborative process where clinicians and patients share information and deliberate treatment options. SDM improves patient knowledge, lowers costs, and enhances outcomes [1–4]. US policy initiatives like the Merit-based Incentive Payment System (MIPS), Medicare Access and CHIP Reauthorization Act of 2015 (MACRA), and the Centers for Medicare and Medicaid support and incentivize SDM [5]. Similarly, the UK National Health Service (NHS) has embedded SDM into its care strategy [6], Canada established pan-Canadian initiatives through the Health Canada SDM frameworks [7], and the Netherlands has made SDM a cornerstone of oncology and chronic care guidelines [8]. However, there is agreement that measurement methods must be improved [9].

Patient-reported experience measures exist [10,11] but are biased, have low response rates, and are not widely implemented [9,12,13]. Observer-based measures (OMs), such as Observer OPTION-5 (OO5), provide more reliable assessments by analyzing recorded clinician–patient interactions and are not based on patients' memory as in patient-reported experience measures [10]. OMs also typically reveal

* Correspondence to: Dartmouth College, The Dartmouth Institute for Health Policy & Clinical Practice, 1 Medical Center Drive, WTRB, Level 5-504, Lebanon, NH 03756, USA
E-mail address: glynelwyn@gmail.com (G. Elwyn).

[1] Codes available here

https://doi.org/10.1016/j.pec.2025.109362
Received 22 July 2025; Received in revised form 15 September 2025; Accepted 20 September 2025
Available online 22 September 2025

lower SDM levels and significant differences in performance between clinicians [10]. However, OMs are resource-intensive, limiting their use to research [11,12]. Automating dialogue assessment has been suggested as a way to provide rapid feedback [14–18].

Observer OPTION-5 (OO5), a validated 5-item tool based on the collaborative deliberation model [13], demonstrated good validity in prior research [12,19–23], typically requires two independent raters. Advancements in natural language processing (the field of computer science focused on automated understanding and processing of human language) and artificial intelligence (AI) offer opportunities to automate SDM assessment, reducing training burdens and costs, enabling large-scale research and clinical trial use, and potentially offering practitioners direct feedback [16–18].

Large Language Models (LLMs–AI systems trained on vast text corpora that can generate and interpret languages) have redefined AI benchmarks, and are increasingly applied in healthcare (diagnostics, decision support, literature interpretation) [24–31]. They show emergent abilities, such as extracting nuanced information from clinical notes, highlighting their potential in healthcare [28,31,32].

OpenAI's GPT-3.5 and GPT-4 and Google's PaLM2 rank among the top-performing models in language understanding and generation [33,34]. These LLMs enable zero-shot learning (i.e., requiring no additional training data or examples) [31,35] and few-shot learning (where providing only limited examples can enhance task performance) [24,30,36–41]. These LLMs are therefore particularly useful when labeled training data is scarce or impractical to collect even in production systems [29–31,37,41,42].

While other models like BERT [43–45] (a widely used neural network architecture for language understanding) are used in medical text classification and generation [42,46,47], we opted for LLMs due to limited training data availability. Publicly available LLM models like LLaMA [48] and Mistral [49] were not considered, as their zero- and few-shot performance lagged behind the commercial LLMs used in this study.

Given the rapid advancements in LLM capabilities, our primary goal is to assess their potential for automatically rating clinical conversations using Observer OPTION-5 (OO5). We achieve this by using an LLM to detect specific speech acts in clinical transcripts and comparing their performance to human raters who previously used the OO5 measure to assess the same data.

2. Methods

2.1. Design

We conducted secondary analyses of an existing corpus of anonymized transcripts from audio recordings of conversations between breast surgeons and patients about treatment for early-stage breast cancer. Anonymization was done in three rounds: (1) the transcription company tagged patient-identifiable information using brackets (e.g., "You must be [Angela]."), (2) A trained staff member removed all bracketed content and reviewed each transcript for other identifiers, including clinician information, and (3) another staff member verified the removal of all identifiers.

This work builds on a prior proof-of-concept study where we developed an automated rating process for the first item of the Observer OPTION-5 measure [16]. We show all five items of the measure in Box 1. These secondary analyses were reviewed and approved by the Dartmouth College Institutional Review Board (STUDY00030157).

We used transcripts from a randomized trial conducted in four cancer centers. [50] The trial compared versions of a conversation aid for surgical decision making in early-stage breast cancer. Surgeons in the intervention arms were trained to use the Option Grid tool, which compared breast-conserving surgery with radiation versus mastectomy. Other therapies were sometimes discussed, including chemotherapy, radiation, and genetic testing. Patients already knew their breast cancer diagnosis before the appointment. We excluded encounters with interpreters for non-English communication to avoid added complexity.

2.2. Transcript preparation and scoring with OO5

We used spaCy to split speaker turns into individual line segments based on transcription punctuation. Two independent human raters, formally trained in Observer OPTION-5 (OO5), scored each encounter by listening to recordings and rating items 1–5 on a 0–4 scale (0 =no evidence, 4 =highest achievement) [19].

The transcripts were divided into contiguous segments of 120 lines. Each segment and its line numbers were input to the LLM, which generated scores for relevant OO5 items. We used 120-line segments because LLMs have limited input lengths and performance declines with longer inputs [51]. Segment-level predictions were aggregated into encounter-level scores. The LLM was prompted to score transcripts (and identify the corresponding line numbers) and identify corresponding line numbers using these rules:

Box 1
Items of the Observer OPTION-5 Measure [19].

Item	Statement Description
1	**Decision awareness:** 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.
2	**Team talk:** The clinician reassures the patient or reaffirms that the clinician will support the patient to become informed or deliberate about the options. If the patient states that they have sought or obtained information prior to the encounter, the clinician supports such a deliberation process.
3	**Option talk:** The clinician gives information or checks understanding about the options that are considered reasonable (this can include taking no action), to support the patient in comparing alternatives. If the patient requests clarification, the clinician supports the process.
4	**Preference elicitation:** The clinician makes an effort to elicit the patient's preferences in response to the options that have been described. If the patient declares their preference(s), the clinician is supportive
5	**Decision talk:** The clinician makes an effort to integrate the patient's elicited preferences as decisions are made. If the patient indicates how best to integrate their preferences as decisions are made, the clinician makes an effort to do so.

1. If all lines scored 0→transcript item score= 0
2. If any non-zero→transcript item score=average of non-zero scores.

For item-level scores overall and by clinician, we averaged the two rater scores and computed summed averages. Clinician-level scores were calculated by averaging the summed OO5 scores across encounters. Inter-rater agreement was assessed using overall and item-level correlations. Following OO5 conventions, LLM outputs (0–4) were rescaled to 0–20; combined item scores therefore ranged 0–100.

To generate item-level scores overall and by clinician, we first averaged the two rater scores, and we computed the summed average score. Similarly, we computed clinician-level scores by averaging the summed OO5 item scores for their encounters. To assess agreement between the two independent raters, we performed overall and item-level correlation analyses. Following OO5 scoring conventions, we rescaled the LLM output for each item score from 0 to 4–0–20. Thus, when combining the scores of the five OO5 items, we get a score range of 0–100, which we use in this article. Transcript scores were calculated as the sum of averaged item scores, and clinician-level scores as the average of their encounter-level sums.

For model development and evaluation, we randomly split the dataset into a validation set and a test set. The validation set (n = 40 encounters) was used to iteratively design and optimize prompts, allowing us to compare alternative prompt formulations and select the best-performing one. The remaining transcripts (n = 247 encounters) formed the held-out test set, which was used exclusively for the final evaluation of model performance. This separation ensured that the test results reflect out-of-sample performance, independent of the data used for prompt optimization.

2.3. Comparisons of LLM OO5 score prediction

We defined the task for the LLMs as the identification and scoring of clinician utterances in the clinical encounter transcripts that correspond to item statements in the OO5 manual. We compared how well LLMs from open-source and closed-source families performed this task and selected the following LLMs: Meta's LLAMA series, OpenAI's GPT series, and Google's Gemini series. During these evaluations, we selected privacy settings for the API provider so that the LLMs did not log or store any part of the data.

2.3.1. The design and optimization of prompts for the LLMs

Prompts for these models were designed to optimize the LLM's performance, and compared how well the scores correlated with OO5 scores provided by human raters. An LLM prompt typically has three parts, see Box 2.

The outline of our prompts can be seen in Appendix Table 1. We described the clinical setting of the encounter transcripts and described the task objective, namely, to find and score instances in the transcripts of OO5 items. The prompts also contained detailed descriptions of each OO5 item, supplemented by example phrases and/or statements illustrating the scoring spectrum. Instructions for scoring and formatting the output were included to ensure a consistent, uniform output format. Depending on the comparative design, we instructed the LLMs to identify and score relevant phrases/statements for each item or all-items simultaneously. Given our focus on assessing surgeons' communication about breast cancer surgery, the LLMs were instructed to exclude unrelated dialogue. Additionally, the LLMs were instructed to explain their scoring decisions because there is evidence that this strategy enhances performance [52]. Similarly, we also included examples of the task (taken from the OO5 User Manual, Appendix Table 1). We evaluated multiple prompts and models on the validation set to identify the best-performing configurations (see Box 3):

2.3.2. Analysis of the LLM performance/statistical analysis to differentiate performance

We evaluated the correlations between LLM-generated scores and the human rater scores at the level of OO5 items, the sum of OO5 items for each encounter, and at the OO5 score level for each clinician (the mean of their encounter OO5 scores).

To contextualize LLM performance, we also measured the level of the two raters' agreement by computing Pearson (r_p) and intraclass correlation (ICC) level scores. The best-performing prompt from the prompt optimization step on GPT-4o was selected for further evaluation on the test set with other LLMs. We report r_p and ICCs, consistent with prior OPTION-5 studies, to assess consistency and absolute agreement on continuous ratings.

Prior research has shown that inter-rater reliability in OPTION-5 scoring is modest, with ICC values often in the 0.6–0.7 range [19,53]. Similar levels of reliability (0.5–0.6) have also been reported in broader health conversation coding tasks [54,55]. Because the average of two human raters was used as the reference standard in our study, the maximum achievable correlation for any model is naturally constrained by the agreement between those raters. In this context, we interpret model–human correlations that achieve at least 70–80 % of the measured human–human agreement as strong evidence of alignment.

We examined the LLM's ability to distinguish between high- and low-performing conversations by dividing the test set into two groups:

1. **Low-Performing SDM Encounter:** Where OO5 sum scores < 50.
2. **High-Performing SDM Encounters:** Where OO5 sum scores ≥ 50.

This threshold was chosen a priori based on OPTION-5 guidance and prior validation studies, which conceptualize the midpoint of the scale (50/100) as distinguishing minimal from more consistent evidence of shared decision-making behaviors [12,19,23]. We therefore defined high/low groups using this interpretive benchmark rather than dataset-derived values such as the mean or median.

Clinician-level segregation was evaluated to determine whether LLMs could distinguish between high- and low-performing clinicians based on their average OO5 conversation scores. To assess this, we used an independent-samples *t*-test, which evaluates whether the means of two groups differ more than would be expected by chance. In our case, the two groups were high versus low-performing encounters (OO5 $\geq$ 50

Box 2
Standard LLM Prompt Design.

Part	Description
i)	A task description, with optional provision of detailed examples that provide details for scoring.
ii)	Data input
iii)	Statements that elicit the required results or prediction, e.g., "Output predictions:"

Box 3
Evaluation Using Multiple Prompt Versions.

1. **Baseline:** The task description only used definitions of each OO5 item.
2. **Detailed Descriptions**: The task description, in addition to definitions, included granular examples for each score in the OO5 items.
3. **Simultaneous Prediction**: The LLM was instructed to predict scores for all OO5 items simultaneously.
4. **Catch-All Category Addition**: Introducing a category for good communication practices not covered by OO5 or related to breast surgery, e.g., the clinician greeting the patient or covering OO5 items related to post-surgery treatments.

vs. < 50). The t-statistic quantifies the size of the difference relative to the variability within each group, with higher values indicating greater separation. We also explored grouping the 12 clinicians into high- and low-performing categories based on their summed OO5 scores.

2.3.3. Trend analysis of LLM performance and human-rater agreement

For this analysis, the test set was divided into two subsets based on rater agreement on the OO5 sum scores:

- **High Rater Agreement Test Set:** Encounter scores where the difference between the two rater OO5 scores was ≤ 10.
- **Low Rater Agreement Test Set:** Encounter scores where the difference between the two rater OO5 scores was greater than 10.

We used these subsets to assess whether the level of human-rater agreement affected the LLM performance and visualized the difference between LLM 005 Scores and Human Rater Scores vs the difference between the human rater scores.

Although we report multiple statistical comparisons (e.g., encounter-level, clinician-level, item-level, and subgroup analyses), these are all derived from a single fixed set of predictions per model generated on the test set after prompt optimization. In total, only six model outputs were evaluated, and the various analyses represent different perspectives on these same outputs rather than independent hypothesis tests. This design reduces the risk of Type I error inflation typically associated with multiple testing. Nevertheless, we interpret the findings as exploratory and emphasize the importance of consistent patterns across analyses rather than isolated *p*-values.

3. Results

3.1. Available transcripts

Our final set of encounter transcripts included 110 collected at center 1, 46 collected at center 2, 8 collected at center 3, and 123 collected at center 4 – a total of 287 conversations with 12 surgeons. We used a random number generator to select 40 transcripts to create a validation set of encounters. The remaining 247 conversations formed the test set of encounters. The conversations were transcribed as separate speaker turns. On average, each encounter transcript contained 488 lines (standard deviation of 334), with the longest transcript containing 1675. The mean human-rater summed OO5 score across the encounter transcripts was 54.15 (standard deviation = 25.77). Slightly less than a third (30 %) of the transcripts had an average OO5 sum score of 25 or lower, while 35 % had an average sum score of 75 or higher. For more detailed statistics, refer to Appendix Table 2. The levels of agreement between human raters were moderate to high, as shown by the correlation in Table 1. As expected, correlations for the overall sum score were higher than for individual items, reflecting the increased reliability of composite scores that aggregate across items. Because model–human correlations are bounded by the level of agreement between the two human raters (r = 0.77), we interpret performance relative to this ceiling. For reference, correlations of 0.54, 0.62, and 0.69 correspond to 70 %, 80 %, and 90 % of the human–human agreement, respectively. These thresholds provide context for evaluating the strength of model–human correlations reported below.

3.2. The optimization of prompts for the LLMs

The results of the prompt optimization experiments on the validation set (40 transcripts) using GPT-4o are summarized in Table 2. Prompts that included only the item's definitions performed poorly, yielding a low and non-significant correlation (r = 0.22, p = 0.17) when predicting one item at a time and (r = 0.46, p < 0.01) when predicting all five OO5 items simultaneously. Incorporating detailed descriptions and granular scoring examples in addition to the task description for each item of OO5 improved performance significantly, with correlations increasing to 0.45 (p < 0.01) for predicting one-item at a time and 0.50 (p < 0.01) for predicting all items simultaneously. Refer to Appendix Table 1 for the complete prompt with the task description, detailed description, and scoring examples.

The addition of the Catch-All Category (see Box 3) reduced false-positive errors, but had a negative impact on LLM performance. Sensitivity analysis on the addition of the Catch-All Category shows that the LLMs often misclassified instances of the five items into the catch-all category (in both cases of predicting for one item at a time and all items simultaneously). Sensitivity analysis on comparing predicting single items versus all items simultaneously shows that, when predicting one item at a time, the model frequently misclassifies instances of the other four items as positive for the predicted item; these errors are

Table 1
Correlations between two human raters (N = 287 encounters).

	Rater Correlation (Pearson r_p)	Rater Correlation (ICC)
Item 1	0.54	0.44
Item 2	0.53	0.34
Item 3	0.81	0.74
Item 4	0.73	0.61
Item 5	0.69	0.54
Overall Sum Score	0.77	0.77

Table 2
Correlation between LLM-predicted (GPT-4o) OO5 scores and mean Rater Scores (Validation dataset of randomly selected 40 transcripts).

Experiment	Pearson Correlation (r_p)	P-value
Predicting one item at-a-time	0.22	0.17
Predicting one item at-a-time + Detailed description	0.45	<0.01
Predicting all-items at once	0.46	<0.01
Predicting all-items at once + Detailed description	0.50	<0.01
Predicting all-items at once + Detailed description + Catch-all category	0.32	0.039

lowered when predicting all items simultaneously.

From the prompt optimization on the evaluation set, we find that predicting all five items simultaneously with detailed descriptions and examples performed best (Appendix Table 1). We used this 'best' prompt to test across different LLMs and conduct further analysis in the rest of the paper.

3.3. Analysis of LLM performance

In Table 3, we show the all-item encounter-level Pearson correlations of different LLMs on the test dataset with the optimized prompts. The top-performing models were Gemini-1.5-Pro-002 and GPT-4o, achieving r_p of 0.59 and 0.64, respectively. Given that the ceiling for model–human agreement is bounded by the observed human–human correlation (r_p=0.77), these results correspond to 77 % (Gemini-1.5-Pro-002) and 83 % (GPT-4o) of the maximum possible agreement. Both, therefore, meet our pre-specified threshold of strong performance (>70 % of the ceiling). At the clinician level, Gemini-1.5-Pro-002 recorded a r_p= 0.88, and GPT-4o recorded a correlation of 0.75. Both results were statistically significant. Other models, including those from the LLAMA family, exhibited lower correlations (r_p=0.21), and their results were not statistically significant.

3.3.1. Stratification by high and low OO5 scores, encounter and clinician levels

The *t*-test results for distinguishing High and Low-Performing SDM encounters (OO5 scores greater than 50/100 from encounters scoring lower than 50/100 are shown in Table 4. The best-performing models from Table 3, Gemini-1.5-Pro-002 and GPT-4o, recorded *t*-test statistics of 9.0 ($p < 0.01$) and 10.05 ($p = 0.039$), respectively. These results reflect a statistically significant and strong difference between the predicted scores of high- and low-performing conversations. Despite the limited number of clinicians, Gemini-1.5-Pro-002 and GPT-4o recorded moderate *t*-test values of 4.0 ($p < 0.01$) and 2.7 ($p = 0.02$), respectively **(**Table 4**)**. For other low-performing LLM models, clinician-level segregation results were not statistically significant, so we are not including the results here.

3.3.2. Item-level scores: correlation between LLM and human rater OO5 scores

Table 5 shows the item-level correlation for Gemini-1.5-Pro-002 and GPT-4o. Items 4 and 5 showed a high correlation with rater scores for both models (Gemini-1.5-Pro-002: ≈0.6, GPT-4o: ≈0.5). Conversely, item 1 demonstrated the lowest correlation, with values of ≈0.15 for both models.

Table 3
Correlation between LLM-predicted scores and mean human rater scores (Test dataset of 247 encounters).

Comparisons	Encounter level	
	Pearson Correlation r_p	P-value
GPT−4o	0.64	<0.01
Gemini−1.5-Pro−002	0.59	<0.01
Llama 405b	0.40	0.018
LLama 70b	0.33	<0.01
Gemini−1.5-flash−001	0.32	<0.01
Gpt−4o-mini	0.196	<0.01
Comparisons	**Clinician level**	
	Pearson Correlation r_p	**P-value**
Gemini−1.5-Pro−002	0.88	**<0.01**
GPT−4o	0.75	**<0.01**
LLama 70b	0.46	0.128
Llama 405b	0.44	0.279
Gemini−1.5-flash−001	0.36	0.256
Gpt−4o-mini	0.31	0.327

Table 4
Evaluating the ability of LLMs to differentiate high vs low OO5 scores at encounter and clinician levels.

Experiment	Encounter level		Clinician level	
	T-test (t)	P-value	T-test (t)	P-value
Gemini−1.5-pro−002	9.02	<0.01	4.02	<0.01
GPT 4o	10.05	0.039	2.71	0.022

3.3.3. Trend analysis of LLM performance and human-rater agreement

Table 6 compares the overall OO5 performance of the model in the test set stratified into high and low rater agreement subsets. In the high rater agreement set, where raters 1 and 2 exhibited higher agreement (r_p=0.98), the LLMs also showed a stronger correlation (0.69 for Gemini-1.5-Pro-002 and 0.75 for GPT-4o, tracking 71 % and 77 % of the ceiling of 0.98) with overall OO5 rater scores. Conversely, in the low rater agreement subset, where human agreement was lower (r_p=0.56), the LLMs' correlations dropped to 0.48 and 0.52 (tracking 86 % and 93 % of the ceiling of 0.56), respectively. Fig. 1 also visualizes the relationship between rater agreement and LLM performance. A positive correlation was observed between the score differences of rater 1 and rater 2 and the deviation of LLM-predicted scores from the mean rater's scores. This suggests that when raters disagreed significantly, LLM predictions also deviated more from human scores, but the LLM performance remains close to the ceiling.

Similarly, on item-level scores, we can see from Table 5 that models perform well on items with high inter-rater agreement – items 4 (Preference elicitation) and 5 (Decision talk) (0.73 and 0.69). Conversely, in items 1 (Decision Awareness) and 2 (Team talk where the inter-rater agreement is low (0.54 and 0.53), the models performed worse. However, item 3 presented a unique challenge; despite high human correlations, both models showed considerably worse performance for this item. We have included Bland–Altman analyses on bias in Appendix B.

4. Discussion and conclusion

4.1. Discussion

4.1.1. Principal findings

Scores for OO5 items generated by GPT-4o and Gemini-1.5-Pro-002 correlated strongly with human ratings of SDM in clinical encounters. Both models distinguished high- from low-scoring encounters. Incorporating detailed prompts further improved performance. Correlations were higher when human raters agreed and lower when they disagreed, consistent with prior evidence that OPTION-5 and related coding tasks show only modest inter-rater reliability. When compared against the ceiling set by human agreement in our dataset, model–human correlations reached 75–85 % of the maximum possible, which we interpret as strong performance given the inherent difficulty of dialogue-based coding. These results reinforce the potential for using LLMs to automate these assessments.

4.1.2. Strengths and weaknesses of the method

Strengths include the use of real-world clinical transcripts and trained human raters, with ethical approval for secondary analysis [50]. This is based on a prior work [16–18] on one item, in which we showed sufficient promise to evaluate on all five items. We deliberately optimized prompts and compared open- and closed-source models to ensure robustness and generalizability, and our conclusions are not specific to a single prompt design or model. Evaluating both open (e.g., LLAMA) and closed (e.g., GPT-4o, Gemini-1.5-Pro-002) models, to highlight options for institutions with resource and privacy constraints. It is important to note that, similar to our observation here, the inter-rater agreement in health communication coding is typically modest, reflecting the inherent difficulty of evaluating subtle dialogue behaviors in clinical conversations [19,53,56]. Model-human correlations are bounded by

Table 5
OO5 Scores: Correlation between the scores of the best-performing LLM and the correlation between the two human raters (N = 247 encounters).

	Pearson Correlations of Mean OO5 Human Rater Scores		Gemini-1.5-pro-002		GPT-4o	
	Pearson Correlation	P-value	Pearson Correlation	P-value	Pearson Correlation	P-value
Item 1 (Decision Awareness)	0.54	0.02	0.15	0.018	0.17	<0.01
Item 2 (Team talk)	0.53	<0.01	0.23	<0.01	0.37	<0.01
Item 3 (Option talk)	0.80	<0.01	0.39	<0.01	0.31	<0.01
Item 4 (Preference elicitation)	0.73	<0.01	0.63	<0.01	0.63	<0.01
Item 5 (Decision talk)	0.69	<0.01	0.52	<0.01	0.50	<0.01
Overall	0.77	<0.01	0.59	<0.01	0.64	<0.01

Table 6
Pearson correlation of best-performing LLMs' predicted scores and average rater scores on rater score consistency between the human raters. For Spearman correlation, refer to Appendix Table 3 (Test dataset of 247 encounters).

	Test set (247)		High Rater Agreement Test subset (n = 89)		Low Rater Agreement test subset (n = 158)	
	Pearson Correlation	P-value	Pearson Correlation	P-value	Pearson Correlation	P-value
Gemini−1.5-pro−002	0.59	<0.01	0.69	<0.01	0.49	<0.01
GPT−4o	0.64	<0.01	0.75	<0.01	0.53	<0.01
Rater correlation	0.77	<0.01	0.98	<0.01	0.56	<0.01

human-human agreement; our results show that LLMs are already approaching this ceiling. Limitations include reliance on one breast cancer trial in which surgeons were exposed to an SDM intervention, limiting generalisability. Broader confirmation in other clinical contexts, as well as with untrained clinicians, is needed. Preparing and anonymizing transcripts was also resource-intensive.

4.1.3. Results in context

The range of possible uses for generative AI to advance SDM is well-recognized [15,17,18,57–59], and pre-LLM AI to automate assessments of provider-patient interactions has been considered [42,46,60–62]. Similarly, the potential to automate clinical diagnoses such as dementia or depression based on transcripts [63], the use of digital scribes for automated medical documentation [31,64] has also been considered. However, we have not identified studies focused on using validated measures to assess different approaches in clinical encounters, such as agenda-setting or the adoption of SDM. Our study adds to this landscape by showing that LLM–human agreement can approach ≈ 80 % of the ceiling set by human–human agreement, suggesting that these models are already performing near the achievable practical upper bound.

Comparable challenges and performance have been reported in education (essay/classroom scoring; [65]), business (call center dialogue analysis; [66]), and law (argument mining; [67]). In these domains, correlations of 0.5–0.7 are often considered sufficient for feedback, suggesting similar potential in healthcare.

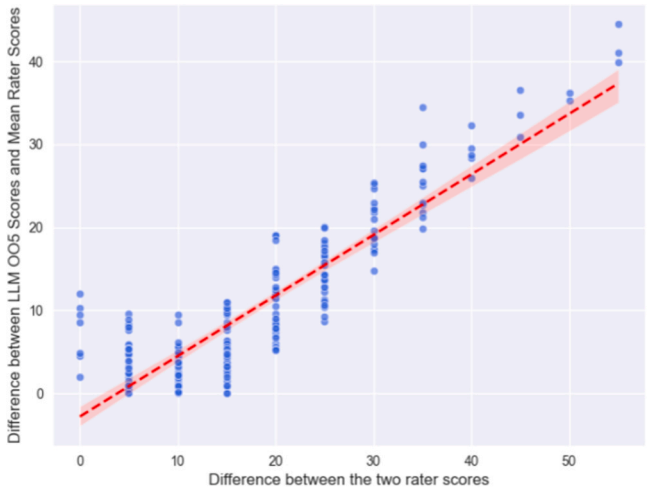

Fig. 1. Relationship between rater agreement and LLM-rater agreement visualized.

4.1.4. Implications

Human rating of SDM is resource-intensive, prone to lapses in concentration, and yields only modest reliability. This is especially important as assessing SDMs requires high levels of concentration to be consistent, and salient aspects of the dialogue are short and easy to miss. Automated assessments, if accurate, offer greater consistency. Our best-performing models already capture much of the human ceiling, suggesting that larger datasets and better prompts may eventually close or exceed this gap. In the future, only LLM or Hybrid approaches—where LLMs identify and score dialogue segments for human validation, might even be superior to human rating. A further hypothesis is that a reliable distinction between high- and low-performing encounters could enable actionable clinician feedback, as suggested in prior work [12,19], though this requires testing.

4.2. Conclusion

The findings highlight the potential of LLMs in evaluating aspects of dialogue between clinicians and patients where existing measures exist that provide the basis for the development of reliable prompts. Our conclusions are limited to a single RCT dataset in breast cancer surgery, and generalisability beyond this setting remains to be established. Expanding datasets and refining prompts will be key. As in education, business, and policy, automated dialogue assessment could provide scalable, consistent evaluations of communication quality.

4.3. Practice implications

Generative AI LLMs could provide a way to efficiently evaluate SDM performance. At present, our findings, though limited by one dataset, primarily support their use as a feasible and scalable research tool for automating observer-based assessments. The possibility of extending these methods to provide actionable feedback to clinicians remains theoretical and will require further study, but our results suggest this is a promising direction for future work.

Abbreviations

OO5	Observer OPTION-5
LLM	Large Language Model
SDM	Shared Decision-Making

GPT	Generative pre-trained transformer
PaLM	Pathway-based Language Model
r_p	Pearson Correlation
OM	Observer-Based Measure

CRediT authorship contribution statement

Rachel Forcino: Writing – original draft, Data curation. **Renata W. Yen:** Writing – original draft, Data curation. **Sai P. Selvaraj:** Writing – original draft, Methodology, Formal analysis, Conceptualization. **Glyn Elwyn:** Writing – original draft, Supervision, Project administration, Data curation, Conceptualization.

Declaration of Competing Interest

Glyn Elwyn's academic interests focus 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. Sai P. Selvaraj, Renata W. Yen, and Rachel C. Forcino have no disclosures.

Acknowledgments

We thank Christopher Jacobs for proofreading the article and the What Matters Most study team for their permission to undertake secondary analysis of the data.

Appendix A

Appendix Table 1
The best-performing prompt. The prompt instructs LLM to score all-time at once, with detailed descriptions for each of the items. The placeholders in the prompt, for e.g., [Item 1 OO5 Definitions] are taken directly from [12,16]

We are interested in categorizing and evaluating doctor's sentences in their conversation with their patients on shared decision making with the Observer OPTION 5 (OO5) five-item measure. We are interested in the following categories:

The Observer OPTION 5 Measure definitions:

Item 1 OO5: [Item 1 OO5 Definitions]
[Item 1 score 0–4 definition and examples]
Item 2 OO5: [Item 2 OO5 Definitions]
[Item 2 score 0–4 definition and examples]
Item 3 OO5: [Item 3 OO5 Definitions]
[Item 3 score 0–4 definition and examples]
Item 4 OO5: [Item 4 OO5 Definitions]
[Item 4 score 0–4 definition and examples]
Item 5 OO5: [Item 5 OO5 Definitions]
[Item 5 score 0–4 definition and examples]

Instructions for selecting sentence IDs:

[Instructions for selecting sentence IDs for each of the items]

Output data structure:

```
{
    "item 1": [
        {
            "sentence_id": int or list or "All" "sentence index(s) contains item 1",
            "score": int "item 1 score for the sentence id(s)",
            "explanation": str "short explanation of why the sentence is scored as such for item 1",
        },
        …
        ],
    …
}
```

Examples:-

Example output 1:

```
{
    "item 1": [
        {
            "sentence_id": 8,
            "score": 3,
            "explanation": "Doctor points out the two options for breast cancer with an aim of comparing them.",
        },
        …
}
```

Example output 2:

```
{
```

(*continued on next page*)

Appendix Table 1 (*continued*)

```
        "item 1": [{
            "sentence_id": "All",
            "score": 0,
            "explanation": "Doctor did not point out the medical options for the patient.",
        }
        …
    }
    Instructions for output format:
        [Instructions for output format]
    Output JSON:
```

Appendix Table 2
Detailed Statistics on the dataset

Measure	Value
Transcript-level statistics	
Mean number of lines	488
Standard deviation (lines)	334
Maximum transcript length	1675
OO5 sum scores (0–100)	
Mean	54.15
Standard deviation	25.77
Median	55.0
% $\leq$ 25	30 %
% $\geq$ 75	35 %
Range	100.00
Q1 (25th percentile)	52.5
Q3 (75th percentile)	75.00
Interquartile range (IQR)	52.5

Appendix Table 3
Spearman correlation of best-performing LLMs' predicted scores and average rater scores on rater score consistency between the human raters. (Test dataset of 247 encounters)

	Test set (247)		**High Rater Agreement Test subset (n = 89)**		**Low Rater Agreement test subset (n = 158)**	
	Spearman Correlation	P-value	Pearson Correlation	P-value	Pearson Correlation	P-value
Gemini−1.5-pro−002	0.54	< 0.01	0.64	< 0.01	0.45	< 0.01
GPT−4o	0.60	< 0.01	0.75	< 0.01	0.50	< 0.01
Rater correlation	0.74	< 0.01	0.97	< 0.01	0.54	< 0.01

Appendix B. Bias Detection

B.1 Methods

To complement correlation and ICC analyses, we conducted Bland–Altman analyses to examine systematic and proportional bias between raters and models. For overall Observer OPTION-5 scores (0–4 scale), we calculated mean differences (bias), 95 % limits of agreement, and tested for proportional bias using linear regression of differences against means. We applied this approach both to (1) the best-performing LLM compared to the human reference (mean of two raters) and (2) human Rater 1 compared to Rater 2.

B.2 Results

The Bland–Altman analysis revealed that the best-performing LLM (GPT-4o) exhibited a systematic positive bias of + 0.25 points across all five OO5 items (on the 0–4 scale, $p < 0.01$). Proportional bias was also present, with model deviations increasing at higher OO5 scores.

Comparisons between human Rater 1 and Rater 2 also showed a systematic bias of + 0.10 points ($p < 0.01$), alongside evidence of proportional bias. Given the restricted 0–4 scoring range, these biases represent approximately 6 % (LLM vs. humans) and 2.5 % (human vs. human) of the total item range.

B.3 Discussion

These analyses indicate that the LLM's systematic bias, while statistically significant, is modest in absolute terms and falls within the general magnitude of variability observed between trained human raters. Prior studies of OPTION-5 have similarly reported non-negligible inter-rater

variability, with mean rater differences of ≈ 0.3–0.4 points on the 0–4 scale and ICCs in the 0.6–0.7 range [19,53]. This suggests that both systematic and proportional bias are features of human-based OO5 scoring itself. In this context, the LLM's performance can be considered comparable to human variability, although future work may focus on calibration strategies to further reduce fixed and proportional bias.

References

[1] Clayman ML, Bylund CL, Chewning B, Makoul G. The impact of patient participation in health decisions within medical encounters. Med Decis Mak 2016; 36:427–52.

[2] Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making and patient outcomes. Med Decis Mak 2015;35:114–31.

[3] Stacey D, Lewis KB, Smith M, Carley M, Volk R, Douglas EE, Pacheco-Brousseau L, Finderup J, Gunderson J, Barry MJ, Bennett CL, Bravo P, Steffensen K, Gogovor A, Graham ID, Kelly SE, Légaré F, Sondergaard H, Thomson R, Trenaman L, Trevena L. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2024;(1):CD001431.

[4] Durand M-A, Carpenter L, Dolan H, Bravo P, Mann M, Bunn F, Elwyn G. Do interventions designed to support shared decision-making reduce health inequalities? A systematic review and meta-analysis. PLoS One 2014;9:e94670.

[5] Elwyn G, Sierpe A, Forcino R. Do payment programs incentivize shared decision making in US healthcare? Patient Educ Couns 2023;113:107798.

[6] Shared decision making. London: National Institute for Health and Care Excellence (NICE), 2021.

[7] Légaré F, Adekpedjou R, Stacey D, Turcotte S, Kryworuchko J, Graham ID, Lyddiatt A, Politi MC, Thomson R, Elwyn G, Donner-Banzhoff N. Interventions for increasing the use of shared decision making by healthcare professionals. Cochrane Database Syst Rev 2018;(7):CD006732.

[8] Stiggelbout AM, Pieterse AH, De Haes JCJM. Shared decision making: concepts, evidence, and practice. Patient Educ Couns 2015;98:1172–9.

[9] 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:e0191747.

[10] Couët N, Desroches S, Robitaille H, Vaillancourt H, Leblanc A, Turcotte S, Elwyn G, Légaré F. 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;18:542–61.

[11] Elwyn G, Hutchings H, Edwards A, Rapport F, Wensing M, Cheung W-Y, Grol R. The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. Health Expect 2005;8:34–42.

[12] 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;93:265–71.

[13] Elwyn G, Lloyd A, May C, van der Weijden T, Stiggelbout A, Edwards AGK, Frosch DL, Rapley T, Barr PJ, Walsh T, Grande SW, Montori V, Epstein R, van der Weijden A, Stiggelbout Trudy, Edwards AGK, Frosch DL, Rapley T, Barr PJ, Walsh T, Grande SW, Victor M, Epstein R. Collaborative deliberation: a model for patient care. Patient Educ Couns 2014;97:158–84.

[14] Antel R, Abbasgholizadeh-Rahimi S, Guadagno E, Harley JM, Poenaru D. The use of artificial intelligence and virtual reality in doctor-patient risk communication: a scoping review. Patient Educ Couns 2022;105:3038–50.

[15] Ryan P, Luz S, Albert P, Vogel C, Normand C, Elwyn G. Using artificial intelligence to assess clinicians' communication skills. BMJ 364 2019:l161.

[16] S.P. Selvaraj, R.W. Yen, R. Forcino, G. Elwyn, 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, Patient Educ. Couns. (n. d.).

[17] Selvaraj SP, Yen RW, Forcino R, Elwyn G. 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. JMIR Preprints 2024. http://preprints.jmir.org/preprint/57790.

[18] Selvaraj SP, Yen RW, Forcino R, Elwyn G. 192 Using large language models to evaluate the offer of options in clinical encounters by using a single item of Observer OPTION-5, a measure of shared decision-making. BMJ Evidenced Based Med 2024;29(Suppl 1):A90.

[19] Barr PJ, O'Malley AJ, Tsulukidze M, Gionfriddo MR, Montori V, Elwyn G. The psychometric properties of observer OPTION(5), an observer measure of shared decision making. Patient Educ Couns 2015;98:970–6.

[20] Kölker M, Topp J, Elwyn G, Härter M, Scholl I. Psychometric properties of the German version of observer OPTION5. BMC Health Serv Res 2018;18:74.

[21] Dillon EC, Stults CD, Wilson C, Chuang J, Meehan A, Li M, Elwyn G, Frosch DL, Yu E, Tai-Seale M. An evaluation of two interventions to enhance patient-physician communication using the observer OPTION5 measure of shared decision making. Patient Educ Couns 2017;100:1910–7.

[22] 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;99:1717–23.

[23] Stubenrouch FE, Pieterse AH, Falkenberg R, Santema TKB, Stiggelbout AM, van der Weijden T, Aarts JAWM, Ubbink DT. OPTION(5) versus OPTION(12) instruments to appreciate the extent to which healthcare providers involve patients in decision-making. Patient Educ Couns 2016;99:1062–8.

[24] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Others, language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.

[25] D.M. Katz, M.J. Bommarito, S. Gao, P. Arredondo, GPT-4 Passes the Bar Exam, (2023). https://doi.org/10.2139/ssrn.4389233.

[26] Shea Y-F, 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:e2325000.

[27] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for Medicine. N Engl J Med 2023;388:1233–9.

[28] Ramprasad S, Ferracane E, Selvaraj SP. Generating more faithful and consistent SOAP notes using attribute-specific parameters. Machine Learning for Healthcare Conference. PMLR; 2023. p. 631–49.

[29] Yao Z, Schloss BJ, Selvaraj SP. Improving summarization with human edits. 2023 Conf Empir Methods Nat Lang Process (EMNLP) 2023:2604–20.

[30] West M, Heilbroner SP, Selvaraj SP, Weiss H, Yazbek J, Halloran P, Koornwinder A, Altay G, Pojman N, Pedrosa J, Chiou VL, Saha A. Cohort builder in Tempus Lens: Querying a large oncology database with generative AI. J. Clin. Oncol. 2025;43: e13589.

[31] Selvaraj SP, Ganguly S, Nagpal K, Altay G, Lane H, Chiou VL, Mitchell K, Gomez GG, Bonet AC, Saha A. Patient records to timelines: A LLM-based approach for summarizing oncology patient history. J. Clin. Oncol 2025;43:e13688.

[32] Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus 2023;15:e39305.

[33] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P.S. Yu, L. Sun, A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, arXiv [cs.AI] (2023). ⟨http://arxiv.org/abs/2302.09419⟩.

[34] R. Anil, A.M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J.H. Clark, L.E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G.H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C.A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, Li, Music, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A.C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D.R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, Y. Wu, PaLM 2 Technical Report, arXiv [cs.AI], 2023. ⟨http://arxiv.org/abs/2305.10403⟩ (accessed January 8, 2024).

[35] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Adv Neural Inf Process Syst 2022;35:22199–213.

[36] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature 2023;620:172–80.

[37] Tomašev N, Harris N, Baur S, Mottram A, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Magliulo V, Meyer C, Ravuri S, Protsyuk I, Connell A, Hughes CO, Karthikesalingam A, Cornebise J, Montgomery H, Rees G, Laing C, Baker CR, Osborne TF, Reeves R, Hassabis D, King D, Suleyman M, Back T, Nielson C, Seneviratne MG, Ledsam JR, Mohamed S. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nat Protoc 2021;16:2765–87.

[38] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J.Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D.E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P.W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X.L. Li, X. Li, T. Ma, A. Malik, C.D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J.C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J.S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A.W. Thomas, F. Tramèr, R.E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S.M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the Opportunities and Risks of Foundation Models, arXiv [cs.LG] (2021). ⟨http://arxiv.org/abs/2108.07258⟩.

[39] Jin D, Pan E, Oufattole N, Weng W-H, 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;11:6421.

[40] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S. Others, palm: scaling language modeling with pathways. J Mach Learn Res 2023;24:1–113.
[41] Banerjee J, Taroni JN, Allaway RJ, Prasad DV, Guinney J, Greene C. Machine learning in rare disease. Nat Methods 2023;20:803–14.
[42] Patel D, Konam S, Prabhakar S. Weakly supervised medication regimen extraction from medical conversations. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2020. p. 178–93.
[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019: pp. 4171–4186..
[44] Lee J, Yoon W, Kim S, Kim D, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 2020;36:1234–40.
[45] Raffel C, Shazeer N, Robert A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. 2020;21:1–67.
[46] Selvaraj SP, Konam S. Medication regimen extraction from medical conversations. Explain AI Healthc Med 2020. ⟨https://link.springer.com/chapter/10.1007/978-3-030-53352-6_18⟩.
[47] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4:86.
[48] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P.S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E.M. Smith, R. Subramanian, X.E. Tan, B. Tang, R. Taylor, A. Williams, J.X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, arXiv [cs.CL] (2023). https://doi.org/10.48550/arXiv.2307.09288.
[49] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux M-A, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE. Arxiv [cs.CL. Mistral 7B 2023. ⟨http://arxiv.org/abs/2310.06825⟩.
[50] Durand M-A, Yen RW, O'Malley AJ, Schubbe D, Politi MC, Saunders CH, Dhage S, Rosenkranz K, Margenthaler J, Tosteson ANA, Crayton E, Jackson S, Bradley A, Walling L, Marx CM, Volk RJ, Sepucha K, Ozanne E, Percac-Lima S, Bergin E, Goodwin C, Miller C, Harris C, Barth Jr RJ, Aft R, Feldman S, Cyr AE, Angeles CV, Jiang S, Elwyn G. What matters most: randomized controlled trial of breast cancer surgery conversation aids across socioeconomic strata. Cancer 2021;127:422–36.
[51] Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, Liang P. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist 2024;12:157–73.
[52] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. Others, Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824–37.
[53] Bobak CA, Barr PJ, Malley O, Ubel AJ, Meurer PA, Montori WJ, Hayward VM, Krumholz RA, Spertus HM. J.A. Holmes-rovner, agreement between physician and observer OPTION(5) scores in the video recording of clinical encounters. Patient Educ Couns 2018;101:1601–7.
[54] Stortenbeker I, Stommel W, Van Der Vleuten C, Van Dulmen S, Essers G, Van Weert J, Pieterse AH. Developing a codebook for linguistic analysis of shared decision making in oncology: lessons learned. Patient Educ Couns 2022;105:3829–36.
[55] Mandhana D, Beattie A, Mcguire A, Grande SW, Joseph-Williams N, Cribb A, Entwistle V, Elwyn G. Unhurried conversations: development and evaluation of a novel measure of shared decision making. Patient Educ Couns 2024;107:1521–9.
[56] Scholl I, Loon MKoelewijn-van, Sepucha K, Elwyn G, Légaré F, Härter M, Dirmaier J. Measurement of shared decision making - a review of instruments. Z Evid Fortbild Qual Gesund 2011;105:313–24.
[57] Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, Ghadiri P, Grad R, Poenaru D, Gore G, Zomahoun HTV, Légaré F, Pluye P. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inf 2022;10:e36199.
[58] Elwyn G, Ryan P, Blumkin D, Weeks WB. Meet generative AI… your new shared decision-making assistant. BMJ Evid Based Med 2024;29:292–5.
[59] Zahidy MAL, Montori V, Ponce OJ. 005 Co-creating the future: AI for assessing and enabling shared decision making. 12th International Shared Decision Making Conference. BMJ Publishing Group Ltd; 2024. A2–A2.
[60] Wallace BC, Laws MB, Small K, Wilson IB, Trikalinos TA. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. Med Decis Mak 2014;34:503–12.
[61] Hoxha J, Chandar P, He Z, Cimino J, Hanauer D, Weng C. DREAM: classification scheme for dialog acts in clinical research query mediation. J Biomed Inf 2016;59:89–101.
[62] Park J, Kotzias D, Kuo P, Logan Iv RL, Merced K, Singh S, Tanana M, Karra Taniskidou E, Lafata JE, Atkins DC, Tai-Seale M, Imel ZE, Smyth P. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. J Am Med Inform Assoc 2019;26:1493–504.
[63] Mirheidari B, Blackburn D, Harkness K, Walker T, Venneri A, Reuber M, Christensen H. Toward the automation of diagnostic conversation analysis in patients with memory complaints. J Alzheimers Dis 2017;58:373–87.
[64] Wang J, Lavender M, Hoque E, Brophy P, Kautz H. A patient-centered digital scribe for automatic medical documentation. JAMIA Open 2021;4:ooab003.
[65] Shermis MD, Burstein J. Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York: Routledge; 2013.
[66] Mctear M. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Cham: Springer; 2020.
[67] Lawrence J, Reed C. Argument Mining: Understanding Arguments in Natural Language. San Rafael, CA: Morgan & Claypool; 2019.