



Universiteit
Leiden
The Netherlands

Measuring the practice and implementation of shared-decision making (OPTION12) how open-sourced smaller LLMs (OS-sLLMs) perform on this task

Wit, T.; Han, L.; Heipon, C.; Lindevelt, D.; Stiggelbout, A.; Verberne, S.

Citation

Wit, T., Han, L., Heipon, C., Lindevelt, D., Stiggelbout, A., & Verberne, S. (2026). *Measuring the practice and implementation of shared-decision making (OPTION12): how open-sourced smaller LLMs (OS-sLLMs) perform on this task*. Dartmouth: ISDM. Retrieved from <https://hdl.handle.net/1887/4299407>

Version: Accepted Manuscript

License: [Creative Commons CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/)

Downloaded from: <https://hdl.handle.net/1887/4299407>

Note: To cite this publication please use the final published version (if applicable).

Lead author and co-authors	<u>Wit, Tamara</u> ¹ ; <u>Han, Lifeng</u> ^{1,2} ; Heipon, Carly ¹ ; Lindevelt, David ² ; Stiggelbout, Anne ¹ ; Verberne, Suzan ² (co-first: TW and LH) corresponding: {t.wit, l.han}@lumc.nl
Affiliations	¹ Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands. ² Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands.

Title and abstract	Title: <Measuring the practice and implementation of shared-decision making (OPTION12): how open-sourced smaller LLMs (OS-sLLMs) perform on this task> Background: Shared-decision making (SDM) is an important practice in clinical consultations in which patients and their loved ones jointly discuss and decide on treatment options with clinicians. This SDM process is often coded by the OPTION12 instrument, an observer-based tool, consisting of 12 items, rated on a 5-point Likert scale. This coding is done by human coders, which costs much time and very often disagreement happens. With the development of LLMs, we explore the capability of open-sourced privacy-preserving smaller LLMs on performing this task. It has a potential to automate or partially automate, with humans in the loop, the coding task. Methods: For human coding, we annotated the transcripts of clinical consultations using OPTION12 on 26 melanoma interviews that include 3 roles: caregivers, patients and doctors. Two human coders resolved their disagreements after independent coding. To measure SDM using OS-sLLMs, we divide the data into development and testing sets. We designed the full investigation framework as shown in the figure attached. It includes 1) a pilot study on development set (11 interviews) for prompt-refinement and best performing sLLM selection as a judge-sLLM; and 2) deploying the finetuned prompts and judge-sLLM with development examples on the testing set (15 interviews) and asking the judge-sLLM to resolve the disagreement on other OS-sLLMs' scoring. Firstly, for prompt refinement, we use chain-of-thoughts (CoTs), LLM-assisted prompting, human-in-the-loop with sample output (few-shot) feedback. For OS-sLLMs, we use both 1) general domain models Llama, Gemma, and Mistral, and 2) medical domain models Meditron and Medllama. Secondly, for judge-sLLM selection, at the system level, we measure the overall correlation between each sLLM and human coding using Spearman and Pearson correlation
----------------------------------	---

	scores. At the segment-level, we also look into the most agreed-upon items and the disagreed-upon items (from the 12 items) for qualitative and quantitative studies. Finally, we deploy the OS-sLLMs on the testing set and use the judge-sLLM to resolve disagreements. Preliminary results: Five OS-sLLMs show the following findings: 1) 3 general domain OS-sLLMs perform better than the 2 medical domain ones that both generate hallucinations and not following prompts precisely, indicating further developments needed for medical OS-sLLMs. 2) Mistral7b outperformed the other two Gemma3:12b and Llama3.1:8b by 4 consensus with human coding, vs 3 items. 3) The overall correlation with human coding on these 12 items is (0.83, 0.80, 0.64) using Pearson correlation, and (0.81, 0.78, 0.61) using Spearman rank correlation from the three models (gemma3:12b, llama3.1:8b, mistral7b). 4) For the items that OS-sLLMs agree with human coders, OS-sLLMs can pick up the same sentences as the humans in some cases, but in other cases, they can also pick up even better quotes than humans. Discussion: We report the first research findings using OS-sLLMs on measuring SDM with the OPION12 in melanoma patients' consultation transcripts. The performance of such OS-sLLMs are promising and valuable to our task. It has agreements with human coders on certain items and we are looking into the disagreement to see if we can have different inputs from OS-sLLMs, or if we can fine-tune OS-sLLMs to achieve human level performances on the rest of the items. In the long term view, we expect fine-tuned OS-sLLMs can achieve human level performance and be deployed into the coding task with humans in the loop for verification and quality control.
Conference	ISDM Dartmouth Abstract deadline: 8 dec 2025 (deadline may be extended)

