Methodological Reporting of Studies Evaluating the Performance of Conversational Agents on Unstructured Data Analysis in Healthcare: Preliminary findings of a rapid review

Oliver T. Nguyen, MSHI[1]; Arsalan Ahmad, MS[1]; Mahdi Fayazi, MS[1]; Douglas A. Wiegmann, PhD[1]

[1]University of Wisconsin at Madison

Introduction

- Conversational agents based on large language models (LLMs) are increasingly used to analyze unstructured data
- The methodological details of evaluations of LLM-based conversational agents are often inconsistently reported, creating challenges in interpretation
- **Study Aim**: Assess the extent that key methodological elements are reported in empirical studies evaluating conversational agents in healthcare

Methods

- Developed a conceptual model of six methodological elements that can influence LLM-based conversational agents' performance

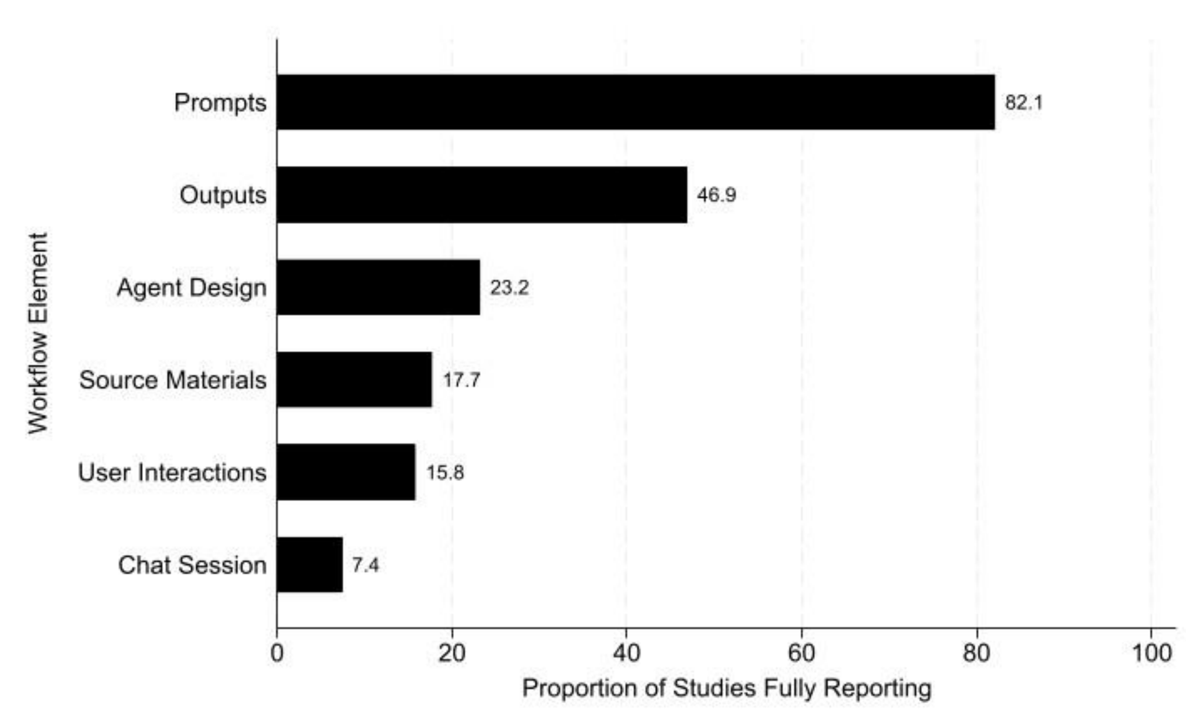
Model Parameter	Description
Chat session structure	Individual interactions between users and agents
Source materials	Documents or data provided to agents to support analysis
Prompt	Textual input provided by users within the agent's interface
Agent design	Agent configuration and operational settings
Output formatting	Agents' response to a given prompt and requested format (or lack thereof)
User characteristics	Users' actions in the workflow (e.g., familiarity with source materials and output)

- Developed **search strategy** that contained search terms on LLM-based conversational agents
- Ran search strategy for papers published October 31, 2024 or in prior year
- **Inclusion criteria**: Studies that are peer-reviewed, reported empirical research, written in English, used a publicly-accessible LLM-based conversational agent, occurred in healthcare context, and reported quantitative findings on agent's performance
- Data abstraction included agent name, agent version, and data type (images, videos, audio, clinical records, other data sources)
- Tabulated proportion of studies that reported each methodological element

Results

- Search strategy identified 11,491 studies, and 475 studies were included in this review
- ChatGPT was the most commonly studied LLM-based conversational agent
- Most studies analyzed clinical records or cases with fewer analyzing images, videos, and audio files
- Studies' methodological reporting ranged from reporting only two elements to all six methodological elements
- On average, the studies reported 1.93 (SD: 1.18) elements
- Across all included studies, the reporting of individual elements ranged from 7.4% (chat sessions) to 82.1% (prompts)

Methodological Element	Proportion of Studies Fully Reporting N (%)
Chat session structure	110 (23.2)
Source materials	35 (7.4)
Prompt	84 (17.7)
Agent design	390 (82.1)
Output formatting	223 (46.9)
User characteristics	75 (15.8)



Implications

- Comprehensive reporting of all methodological parameters was uncommon
- There remains a need to standardize the design and reporting of studies that evaluate LLM-based conversational agents' performance
- We developed a reporting checklist based off the conceptual model we used for this study
- Challenges in assessing bias in studies will persist until standard reporting is achieved

Methodological Element	Description	Reported?
Agent design	Name of LLM-based conversational agent(s)	
	Version number or access date	
	Settings enabled/disabled	
Chat session	Number of chat sessions used per test case	
	Whether chats were terminated and refreshed prior to next analysis	
Source materials	Description of content and formatting of source materials	
	Method that agent(s) received source materials	
Prompts	Description or copy of prompts used	
Output	Description of desired output formatting	
Users	Description of user characteristics (e.g., whether user is representative of typical individual that would complete this analysis) user's actions upon receiving the output (e.g., continuing chat, re-running analysis to troubleshoot issues)	

Future Research

- Evaluate how well LLM-based conversational agents perform on analyzing videos, audios, and images
- Assess to what extent the current conceptual model on relevant methodological parameters for <u>unstructured</u> data analysis extend to <u>structured</u> data analysis
- Assess the performance of a diverse range of LLM-based conversational agents (e.g., Copilot, Gemini)

Acknowledgements

This work was supported by the Wisconsin Advanced Opportunity Fellowship and the Office of Naval Research (ONR) Grant N00014-23-1-2554 (PI: Wiegmann). The funders had no role in the design or analysis. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Office of Naval Research.