

Utilizing LLM to Identify Inconsistencies in Clinical Notes

Sonanya B. Woldegebriel¹ , Yaa A. Kumah-Crystal, MD, MPH²

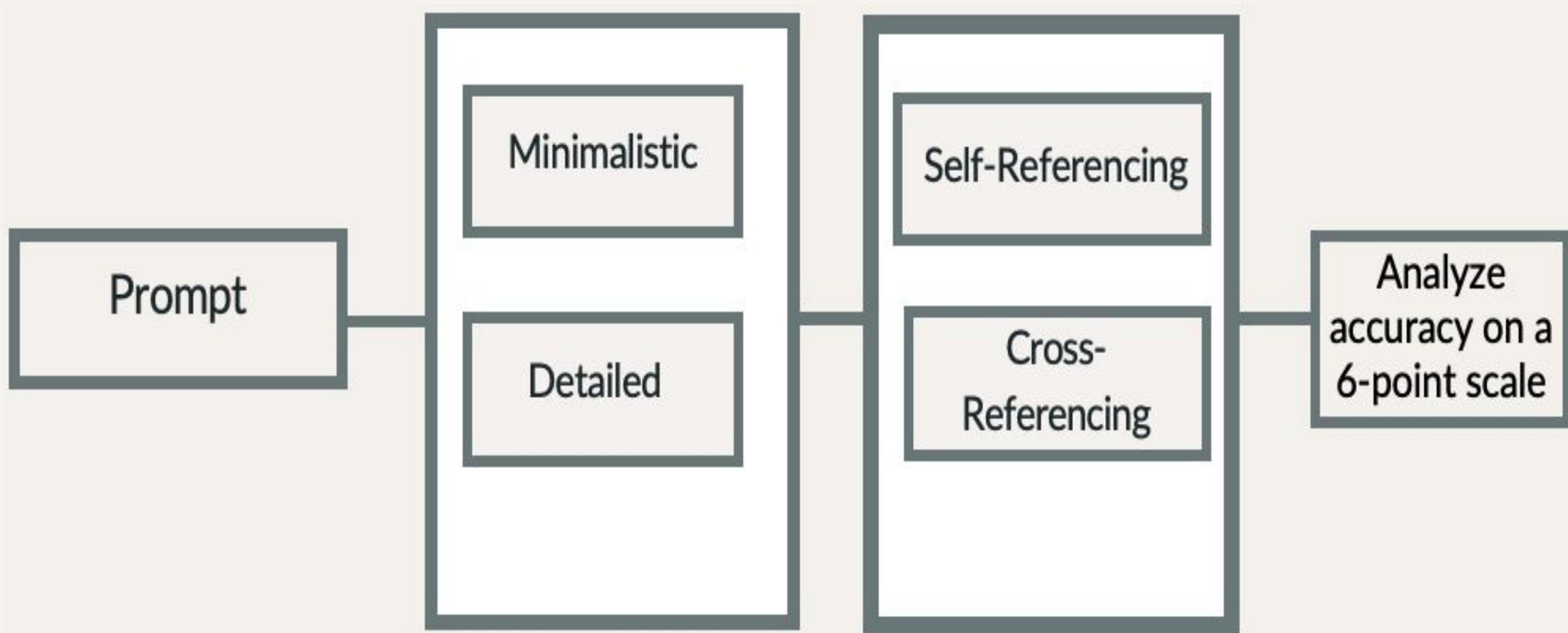
¹Hillsboro High School, ²Vanderbilt Department of Biomedical Informatics, Vanderbilt University Medical Center

INTRO

- According to an independent study published by the NIH in 2020, 20% of patients find mistakes when reading a clinical note, with 40% stating the errors are serious.

Methods

- Free versions of 4 LLMs: ChatGPT, Claude, Copilot, and Gemini.
- 6 EHR inconsistency types described by QNOTE and PDQI-9.
- Tested performance using Minimalistic prompts and Detailed prompts.
- Identified inconsistencies through self-referencing and cross-referencing clinical notes.



RESULTS

Self: 100% vs 88%

Cross: 93% vs 93%

Overall: 95% vs 92%

Model	ChatGPT		Claude		Copilot		Gemini	
Prompt	Minimal	Detailed	Minimal	Detailed	Minimal	Detailed	Minimal	Detailed
Laterization	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Temporal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Objective Data	Yes	No	Yes	Yes	Yes	Yes	Yes	No
Clinical State	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Demographic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Logic/Plan Gap	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SCORE	6/6	5/6	6/6	6/6	6/6	6/6	6/6	4/6

Model	ChatGPT In Claude		ChatGPT in Copilot		Gemini in Claude		Gemini in Copilot		Copilot in Claude		Claude in Copilot	
Prompt	M	D	M	D	M	D	M	D	M	D	M	D
Laterization	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Temporal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Objective Data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clinical State	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Demographic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Logic/Plan Gap	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SCORE	6/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6	5/6	6/6	5/6

Model	ChatGPT in Gemini		Claude in ChatGPT		Claude in Gemini		Copilot in ChatGPT		Copilot in Gemini		Gemini in ChatGPT	
Prompt	M	D	M	D	M	D	M	D	M	D	M	D
Laterization	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Temporal	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Objective Data	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clinical State	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Demographic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Logic/Plan Gap	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SCORE	6/6	6/6	4/6	6/6	6/6	6/6	6/6	3/6	6/6	6/6	6/6	3/6

Discussion

- Our results suggest that the Detailed prompting strategy did not improve detection. Some LLMs performed better at identify inconsistencies in cross-referencing than self-referencing

Future work

Future work would include the use of synthetic rather than real EHR clinical notes, additional LLMs, increase the number of trials, and test anonymized EHR clinical notes.

Conclusion

- LLM performance varied between self-referencing and cross-referencing conditions.
 - One possible explanation is that differences in model design or intended use may influence how LLMs identify inconsistencies.

VANDERBILT UNIVERSITY
MEDICAL CENTER

Department of
Biomedical Informatics
Sonanya Woldegebriel
woldegebrielsonanya@gmail.com