ISSN: 3068-9503

Review Article

AI and Intelligent Systems: Engineering, Medicine Society

Large Language Models for Patient Education for Atrial Fibrillation

Gloria Wu[1*], Hrishi Paliath-Pathiyal[2], Obaid Khan[3], Margaret Wang[4], Swara Tewari[5], Hasika Oggi[6], Riki Toram[1], Paul J. Wang[7] and David Lee[8]

[1]*University of California San Francisco*

[2]*Nova Southeastern University*

[3]*California Health Sciences University*

[4]*Santa Clara University*

[5]*University of California Santa Barbara*

[6]*University of California Berkeley*

[7]*Stanford University School of Medicine*

[8]*UT Texas Houston, Sri Kurniawan, PhD University of California, Santa Cruz*

[*]Corresponding Author
Gloria Wu, University of California San Francisco.

Submitted: 2025, Dec 19; **Accepted:** 2026, Jan 28; **Published:** 2026, Feb 17

Citation: Wu, G., Paliath-Pathiyal, H., Khan, O., Wang, M., Tewari, S., et al. (2026). Large Language Models for Patient Education for Atrial Fibrillation, *AI Intell Sys Eng Med Society. 2*(1), 01-03.

1. Introduction

Atrial fibrillation (AF) is the most common arrhythmia globally, affecting over 37 million individuals. AF substantially increases the risk of stroke, heart failure, and mortality. With the advent of the internet, most patients use Large Language Models (LLMs) for health education. In terms of monthly users, ChatGPT has 800M, Gemini has 350M, Claude.ai has 18.9M, Meta has 350M, and Grok has 64M. These LLMs use different training models. The purpose of this small study was to evaluate large language models (LLMs) for atrial fibrillation patient education. Each LLM has a standard query: "I am a 68-year-old [race] [gender] with atrial fibrillation. I had a heart attack 2 years ago with stents. What can I expect from my cardiologist?" Responses were evaluated for word count, Flesch-Kincaid (FK) grade level, cosine similarity scores to measure structural text similarity, and the presence of select clinical keywords relevant to AF management. A cardiologist author selected keywords.

2. Results

Keywords	LLMs					
	ChatGPT-4o	**Gemini 2.5**	**Claude.ai 4.0**	**Meta AI Llama 3.1**	**Grok**	
Electrophysiologist (EP) MD/DO	⅙ *WF	⅙ *BM	0/6	0/6	3/6 *WF, BM, HF	5/30
ECG	6/6	6/6	6/6	6/6	6/6	30/30
Echocardiogram	6/6	6/6	6/6	⅙ *WM	6/6	25/30
INR Coagulation Test	6/6	0/6	0/6	0/6	4/6 *WF, BF, BM, HF	10/30

BNP	0/6	0/6	0/6	0/6	2/6 *WF, HM	2/30
Holter Monitor	6/6	6/6	6/6	2/6 *WF, WM	6/6	26/30
Cardiac Stress Test	6/6	6/6	2/6 *WF, WM	⅙ *WM	6/6	19/30
Coronary Angiography	6/6	2/6 *HM, HF	0/6	0/6	3/6 *WF, BF, HF	11/30/
Medications	6/6	6/6	6/6	6/6	6/6	30/30
Stress	⅙ *BF	6/6	2/6 *WF, WM	6/6	2/6 *BF, HM	17/30
DASH or Mediterranean Diet	6/6	⅚ *WF, BF, BM, HF, HM	0/6	0/6	3/6 *BM, HF, HM	14/30
EtOH	2/6 *WM, BM	6/6	0/6	0/6	6/6	14/30
Exercise	6/6	6/6	2/6 *WF, WM	6/6	6/6	26/30
Smoking Cessation	6/6	6/6	2/6 *WF, WM	⅙ *WM	6/6	21/30
CHA2DS2-VASc Score	3/6 *WF, WM, BM	2/6 *WF, HF	0/6	0/6	2/6 *WM, HM	7/30
HASBLED Score	2/6 *WF, BM	2/6 *WF, HF	0/6	0/6	2/6 *WM, HM	6/30
Catheter Ablation	3/6 *WM, BF, BM	⅚ *WF, WM, BF, BM, HF	0/6	0/6	6/6	14/30
Left Atrial Appendage Procedure	⅙ *BM	0/6	0/6	0/6	0/6	1/30
DEI/Culture (Variant)	3/6 *BM, HF, HM	3/6 *BF, HF, HM	2/6 *HF, HM	⅙ *HF	4/6 *WF, WM, BF, HM	13/30
Mentions	18/19	16/19	9/19	9/19	18/19	

Represents keywords only mentioned in the specified demographic group Average word counts were: ChatGPT = 312.5, Gemini = 937.7, Claude.ai = 262.5, Meta AI = 240, and Grok = 830.7 ± 104.7 (mean 516.7±307.4). Flesch-Kincaid grade-level scores were: ChatGPT = 10.7, Gemini = 13.3, Claude.ai = 30.7, Meta AI = 12.4, and Grok = 10.3 (mean 15.5 ± 8.2), all exceeding the recommended 6th-8th grade level for patient education materials. Cosine similarity scores ranged from 71.7% to 78.2% (mean 74.5 ± 3.0), where 1.00 means total similarity between both texts.

Electrophysiologists were mentioned only for White females; stress management was mentioned only for African American females. Alcohol as a modifiable risk factor appeared only for White and African American males. Catheter ablation was discussed only for White males and African Americans; the Watchman device was solely for African American males. Grok inconsistently included critical AF keywords based on race and gender. Diagnostic tests like Holter Monitor, Cardiac Stress Test, and ECG/Echocardiogram were omitted more frequently for Black and Hispanic/Latino patients. Risk stratification keywords (CHA_2DS_2-VASc, stroke risk, hypertension) were missing more often for Black men and women, Hispanic women. Bias was evident, with White males having the most keywords while minority women, predominantly Black women, received the fewest. Gemini's responses exceeded recommended readability levels and restricted electrophysiologist referrals to African American males. Claude and Meta omitted electrophysiology referrals, clinical scores, procedures, and alcohol advice across most groups, and listed preventative smoking recommendations only to White patients.

3. Discussion

LLMs are trained to prioritize veracity, which includes factual correctness and completeness, at the expense of readability. This can yield responses that, while accurate, exceed grade levels appropriate for patient communication. Additionally, training datasets reflect the biases found in medical literature and online health queries. With approximately 77,000 health-related searches occurring each minute on Google, user-generated data helps shape

and refine models such as Gemini. As a result, these language models may produce more accurate content for commonly searched conditions, such as hypertension, whereas they provide less reliable information for less frequently searched conditions, such as atrial fibrillation. This pattern can contribute to ongoing disparities in the availability and quality of health information across different patient populations.

3.1. Limitations and Future Directions

This study used Flesch-Kincaid scoring to penalize medical terminology. In future research, it should be tested for multiple prompt variations, target ≤ 8th-grade readability, and incorporate standardized cultural prompts addressing language access, social determinants of health, and family history.

4. Conclusion

This study shows gender and ethnic disparities in AI chatbots' responses regarding atrial fibrillation. These findings suggest that physicians and healthcare teams should monitor LLM responses to queries about atrial fibrillation [1-5].

Funding: None.

Conflicts of Interest: The authors declare no competing interests.

References

1. American Heart Association. Writing for patient education: readability guidelines. Dallas (TX): American Heart Association
2. Wang, P., et al. (2024). Large language models in cardiovascular medicine: opportunities and pitfalls. *Journal of the American College of Cardiology, 83*(5), 623–630.
3. Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233.
4. Rothwell, J. (2020). *Assessing the economic gains of eradicating illiteracy nationally and regionally in the United States*. Barbara Bush Foundation for Family Literacy.
5. Zhou, Y., & Di Eugenio, B. (2025). Veracity bias and beyond: uncovering LLMs' hidden beliefs in problem-solving reasoning. *arXiv preprint*.