CardioNeoCanon: a customized chatbot for cardiology inquiry with retrieval augmented generation to reduce hallucinations and improve performance of large language models

Tu Hao Tran

CSANZ

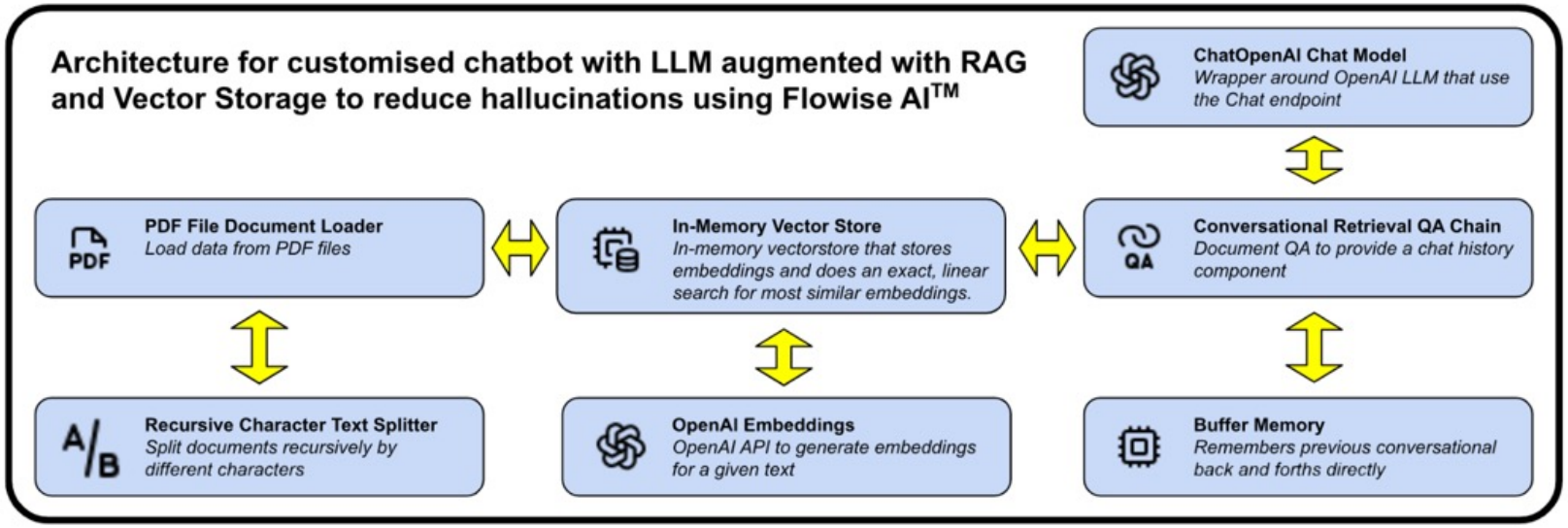
August 2nd 2024

Faculty Disclosure

The presenter has advised that the following presentation will NOT include discussion on any commercial products or service and that there are NO financial interests or relationships with any of the Commercial Supporters of this years ASM.

Aims, Background and Method

- **Large Language Models** (LLMs) and chatbots are becoming ubiquitous.
- Adoption in cardiology practice is understandably hampered by **misleading** information and/or **hallucinations** (fiction).
- A **customised chatbot for use by cardiologists** needs to be:
  - Accurate
  - Not misleading
  - Up-to-date
  - Robust referencing.

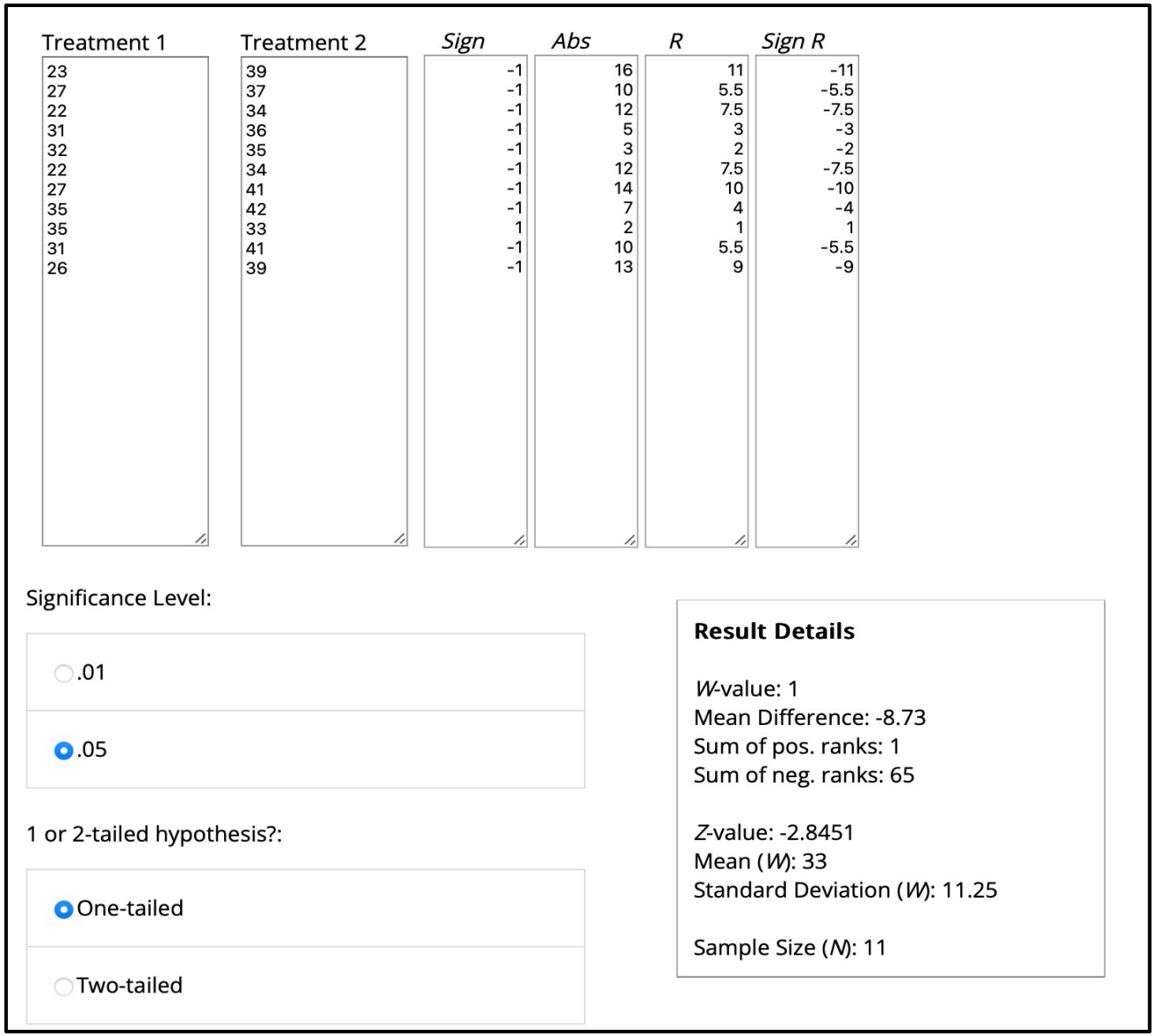


CardioNeoCanon compared with **ChatGPT 3.5** for 4 questions by 11 **blinded** cardiology doctors:

1. Create a detailed summary of an approach to the management of heart failure with evidence-based practice.
2. List the different types of CIED and give the indications, follow-up period in months and clinical considerations, in a table.
3. Compare paragon-HF and paradigm-HF trials in table
4. Propose a framework for a pHD which aims to elucidate the pathophysiological substrate for atrial fibrillation that involves rotors, artificial intelligence and augmented reality. Give a detailed summary of the field of rotors in atrial fibrillation and give 5 high impact articles that will assist with this field of study.

Results

- All doctors subjectively adjudicate the performance with a **VRISM** score (5-15, low=good, high=poor):
  - **V**agueness (1 not vague, 2 vague, 3 very vague)
  - **R**eferencing (1 references all correct, 2 no references, 3 references incorrect)
  - **I**ncorrect information (1 none, 2 not sure, 3 incorrect)
  - **S**tructure (1 good, 2 adequate, 3 poor)
  - **M**issing critical information (1 none, 2 not sure, 3 missing)
- One-tailed Wilcoxon signed-rank test:
  - Small sample size without normal distribution
  - W-value of 1 with a mean difference of 8.73.
  - Indicating that CardioNeoCanon achieved significantly lower VRISM scores compared to ChatGPT 3.5 ($p<0.05$)



Treatment 1	Treatment 2	*Sign*	*Abs*	*R*	*Sign R*
23	39	-1	16	11	-11
27	37	-1	10	5.5	-5.5
22	34	-1	12	7.5	-7.5
31	36	-1	5	3	-3
32	35	-1	3	2	-2
22	34	-1	12	7.5	-7.5
27	41	-1	14	10	-10
35	42	-1	7	4	-4
35	33	1	2	1	1
31	41	-1	10	5.5	-5.5
26	39	-1	13	9	-9

Conclusion and Discussion

- **CardioNeoCanon** achieved significantly lower VRISM scores by a considerable margin ($p<0.05$)
- Demonstrates the practicality of clinicians developing their own chatbots that can **outperform** off the shelf LLM options.
- Clinicians should strive to have more **agency** in the space of artificial intelligence and embrace participating in **innovation**.

Try it.