ISPOR 2024
May 5-8, 2024

Can Large Language Models Simulate HTA Committee Discussions?

Findings and Challenges from a Case Study in Neoadjuvant Treatment of Resectable Non-Small Cell Lung Cancer

Tim Reason[1], Sven Klijn[2], Andy Gimblett[2], Bill Malcolm[1]

[1]Estima Scientific; [2]Bristol Myers Squibb

A brief history of AI

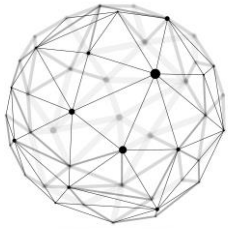
1956

The Dartmouth Conference

1956

The birth of the artificial neural network

1958



The rise of Deep Blue

1997

ImageNet: AI becomes useful

2012

AlphaGo: Neural nets defeat humans

2016

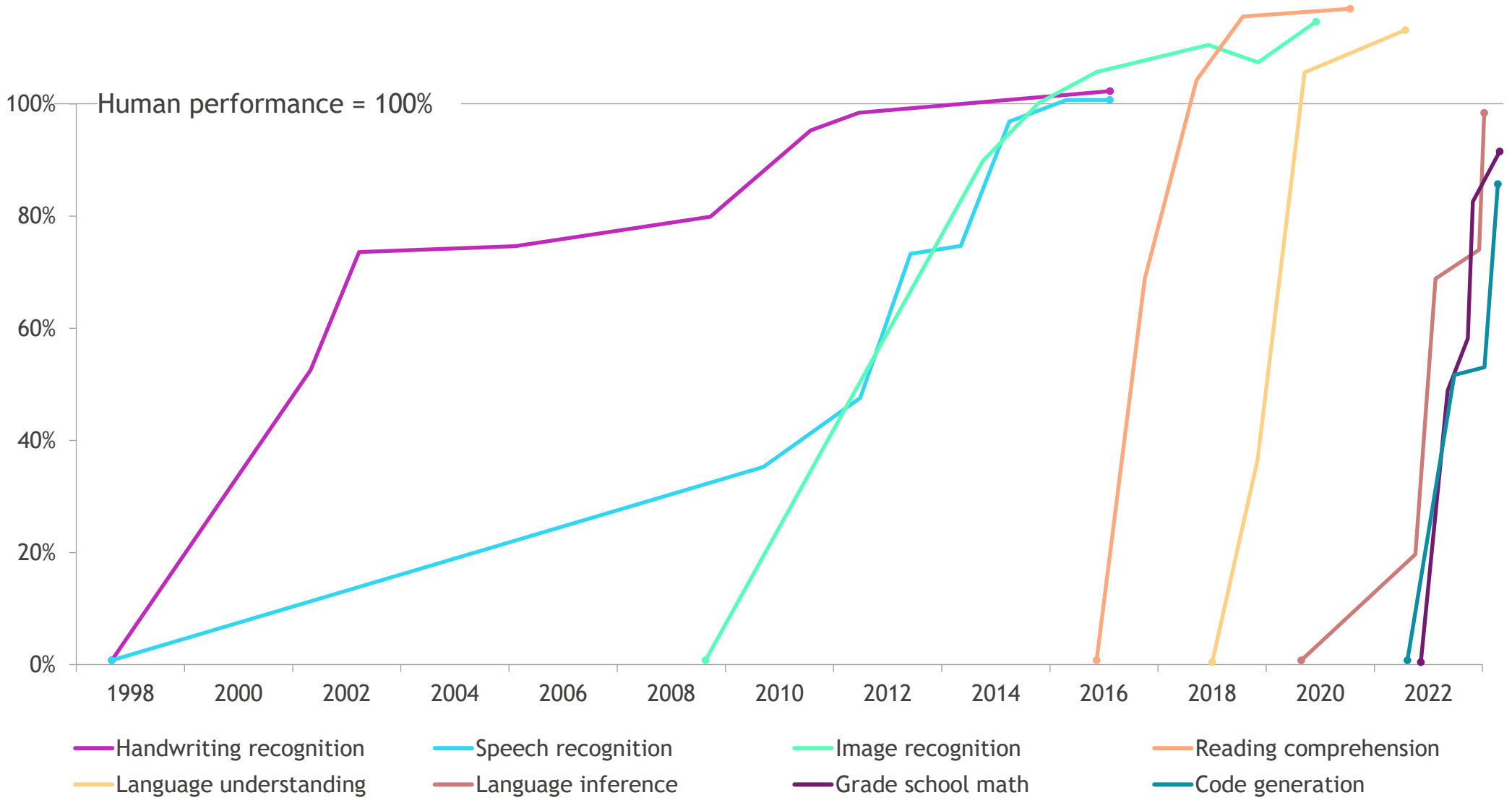
LLMs go mainstream

2022

2023

AI: Artificial Intelligence, LLM: Large Language Model

Speed of AI development is accelerating



AI: artificial intelligence | Source: Accenture (2024), Technology Vision 2024

Virtual committees: The concept

- **What were we trying to do?**
  - Simulate a NICE committee discussion
  - Use purely virtual committee members (LLMs)
  - Arbitrary case study: NICE appraisal of “Nivolumab with chemotherapy for neoadjuvant treatment of resectable non-small cell lung cancer”
- **Why do this?**
  - Access to experts and scientific advice is expensive and hard to come by
  - Test the abilities of AI and LLMs to provide robust discussion
- **How did we do it?**
  - LLMs (GPT-4) to simulate a fully AI-based HTA discussion
  - Individual characters (all LLMs) with character attributes and probability distributions
  - NICE ERG report [ID3757] as basis for discussion

ERG: Evidence Review Group, GPT: Generative Pre-trained Transformer, HTA: Health Technology Assessment, LLM: Large Language Model, | NICE: National Institute for Health and Care Excellence

Virtual committee members: Meet the team

Tia Walti

Role: *Chair*

Job: *Clinician*

Pharma fondness: *Scathing*

Technical HEOR: *Low*

Willingness to argue: *High*

Mattias Tigler

Role: *Health economist*

Job: *Academic*

Pharma fondness: *Low*

Technical HEOR: *High*

Willingness to argue: *High*

Pat Shent (lung cancer survivor)

Role: *Patient Rep*

Job: *Barrister*

Pharma fondness: *Optimistic*

Technical HEOR: *None*

Willingness to argue: *Low*

Corbyn Parrot

Role: *Pharma rep*

Job: *HTA Manager*

Pharma fondness: *High*

Technical HEOR: *Medium*

Willingness to argue: *Very high*

Clint Hermanson

Role: *Statistician*

Job: *Statistician (expert in CER)*

Pharma fondness: *Medium*

Technical HEOR: *High*

Willingness to argue: *Moderate*

Clare Ignition

Role: *Clinical expert*

Job: *Clinician*

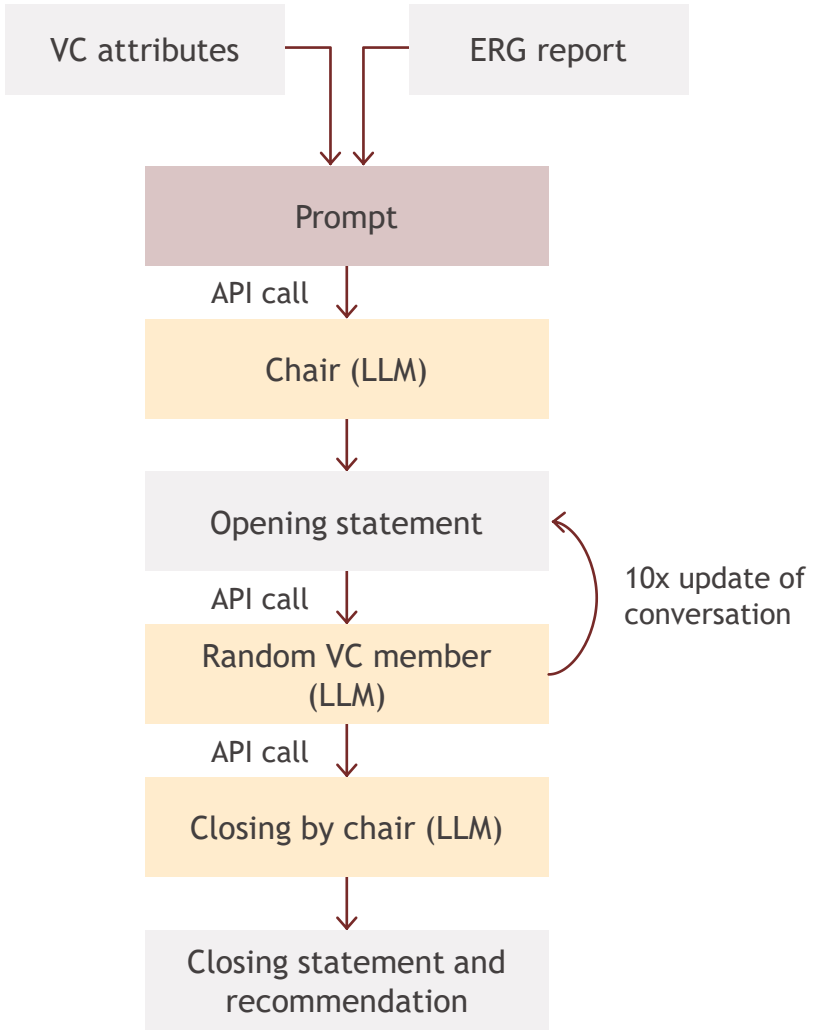
Pharma fondness: *Cautious*

Technical HEOR: *Low*

Willingness to argue: *Low*

CER: Cost-effectiveness Research, HEOR: Health Economics and Outcomes Research, HTA: Health Technology Assessment, Rep: Representative

Process overview



- GPT-4 Turbo was used to simulate each committee member
- The chair was asked to make an opening statement and summarize the evidence, based on the ERG report
- Virtual committee members were then randomly selected to add to the conversation
- This process was repeated 10 times before the virtual chair was asked to make a closing statement and recommendation

API: Application Programming Interface, ERG: Evidence Review Group, GPT: Generative Pre-trained Transformer, LLM: Large Language Model, VC: Virtual Committee

Virtual committee: The process

Define committee members and roles

- Choose names (arbitrarily)
- Assign roles, e.g.:
  - Chair
  - Patient representative
  - Clinician

Define attributes

- Set different attributes based on roles assigned
- Attributes:
  - HE/statistics knowledge
  - Fondness for pharma industry
  - Job outside committee
  - Willingness to argue

Read PDF + introduction by chair

- PDF for NICE ERG report read into Python and fed to LLM
- Request virtual chair to summarize for the rest of the virtual committee

Simulate probability of next person speaking

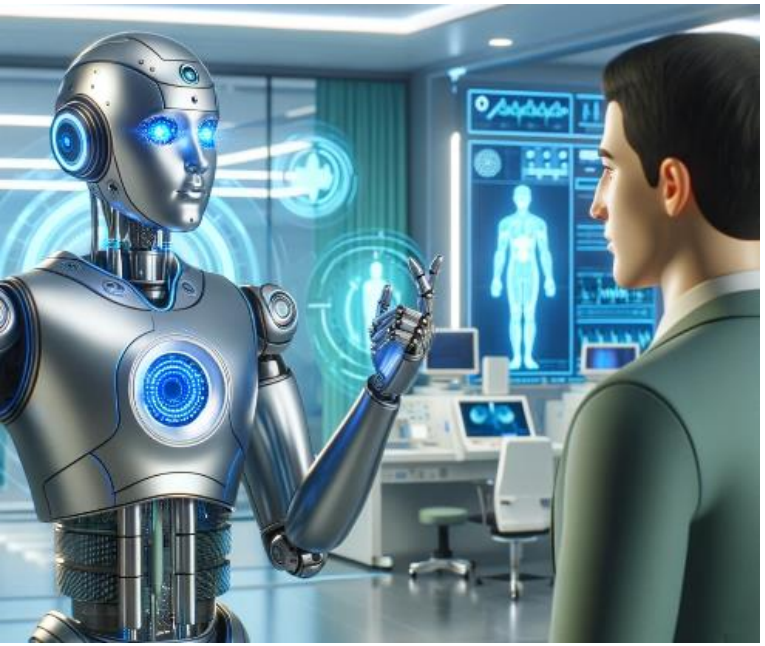
- Simulate from a Dirichlet distribution who speaks next
- Feed the entire conversation so far (starting with chair summary) and NICE ERG report into LLM to simulate next response from chosen speaker

Run conversation and print final output

- Run this “model” for 10 responses (flexible)
- Ask chair to summarize the entire discussion and make a recommendation
- Print conversation and recommendation

ERG: Evidence Review Group, HE: Health Economics, LLM: Large Language Model, NICE: National Institute for Health and Care Excellence, PDF: Portable Document Format

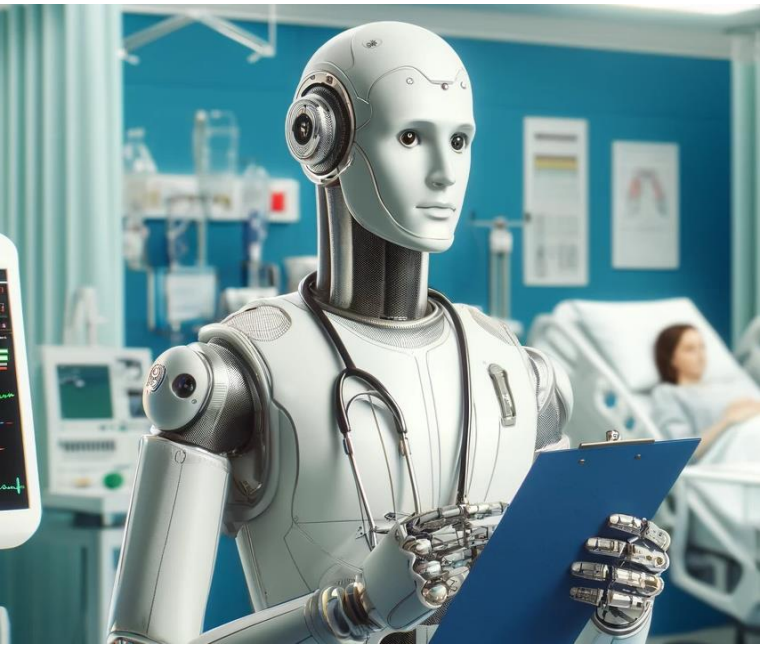
Common themes across human and virtual committees



- **Extrapolation**
  - Virtual chair was able to accurately summarize the ERG report
- **Cure assumption**
  - Virtual committee members argued from their respective perspectives
  - Evidence presented, the ERG report, was considered by the virtual committee
- **Generalizability**
  - Good flowing conversation, considering the objective and the evidence presented
  - Overly verbose at times
- **Re-treatment assumption**
  - Mentioned in the chair opening and closing statements
  - Not extensively discussed by the committee
- **The actual model**
  - Relatively easy to get virtual committee to discuss assumptions around the model
  - Relatively difficult to get them to extensively discuss the model itself and its results

ERG: Evidence Review Group

Common themes: Example output from virtual committee



Survival extrapolation

“Regarding the extrapolation models for progression, the use of these models must be justified with a careful validation against external data, where available. If the model’s long-term predictions are largely uncertain, this increases the risk of making an incorrect recommendation. We should consider the most conservative and plausible extrapolation models, especially in the absence of long-term data.”

Cure assumption

“Regarding the cure assumption, I acknowledge that it is a common approach in oncology modeling; however, it must still be grounded in evidence. Given the transformative implications of labeling patients as ‘cured’, we need more than precedent—we need data to support such a claim. This is especially critical for a treatment that could be widely adopted based on our recommendation.”

Generalizability: Applicability to UK

“On the matter of trial applicability, the CheckMate-816’s geographic and procedural disparities raise questions about the external validity of the results. Without appropriate adjustments or UK-specific data, our confidence in these results being reflective of the English clinical setting is limited.”

Cost-effectiveness estimates

“The base-case probabilistic analysis indicates that nivolumab + PDC is cost-effective compared with neoadjuvant CRT, adjuvant PDC, and surgery alone at a threshold of £20,000 per QALY gained. However, the decision uncertainty is high, particularly for the comparison with neoadjuvant CRT. The subgroup analyses further increase the decision uncertainty due to the immaturity of the CheckMate-816 data and the potential variation in effectiveness across subgroups and treatments.”

CRT: Chemoradiotherapy, PDC: Platinum-doublet Chemotherapy, QALY: Quality-Adjusted Life Year, UK: United Kingdom

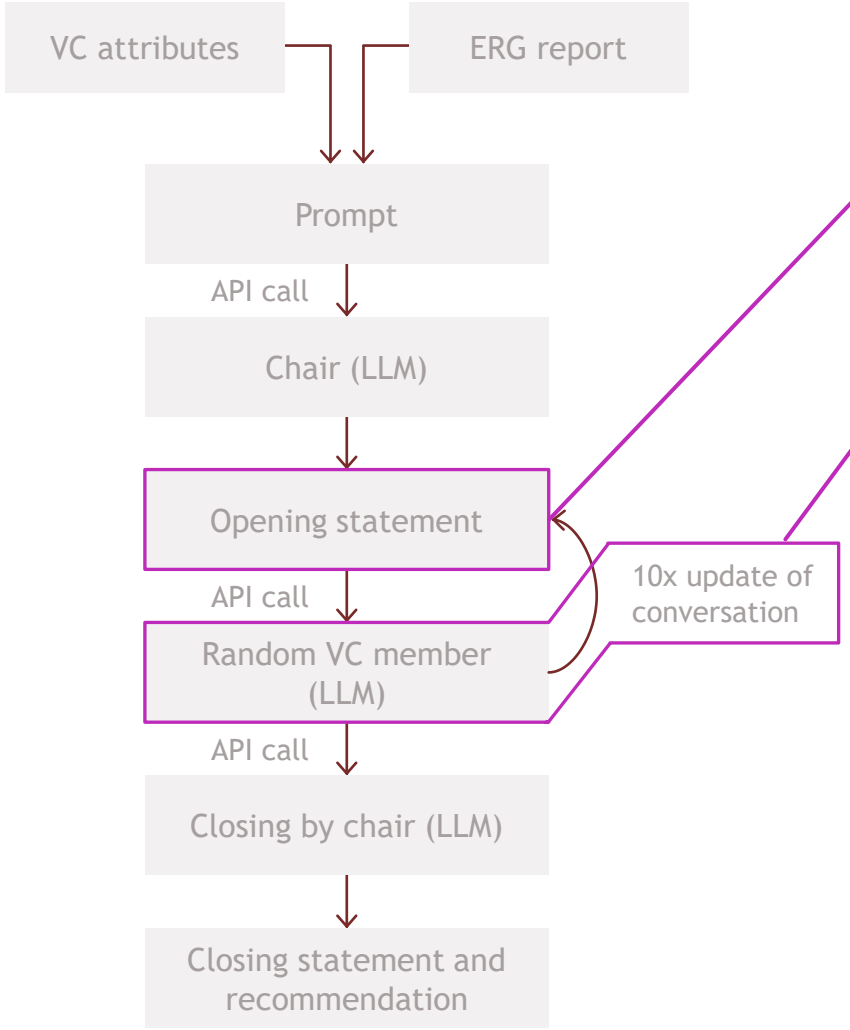
Human and virtual committees: Similar recommendation

- **Human committee**
  "Nivolumab with chemotherapy is recommended, within its marketing authorisation, as an option for the neoadjuvant treatment of resectable (tumours at least 4 cm or node positive) non-small-cell lung cancer (NSCLC) in adults. It is only recommended if the company provides it according to the commercial arrangement (see section 2)."

- **Virtual committee**
  "Due to current uncertainties and the need for rigorous methods, we issue a conditional recommendation for reimbursing nivolumab combined with chemotherapy for neoadjuvant treatment of resectable NSCLC. This is based on the condition that the manufacturer supplies further evidence and analyses to resolve concerns about its effectiveness in early-stage disease, representation of the UK population, and long-term results. We also require that the manufacturer conducts post-marketing surveillance in the UK to gather data on outcomes, health-related quality of life, and long-term survival."

NSCLC: Non-Small Cell Lung Cancer, UK: United Kingdom

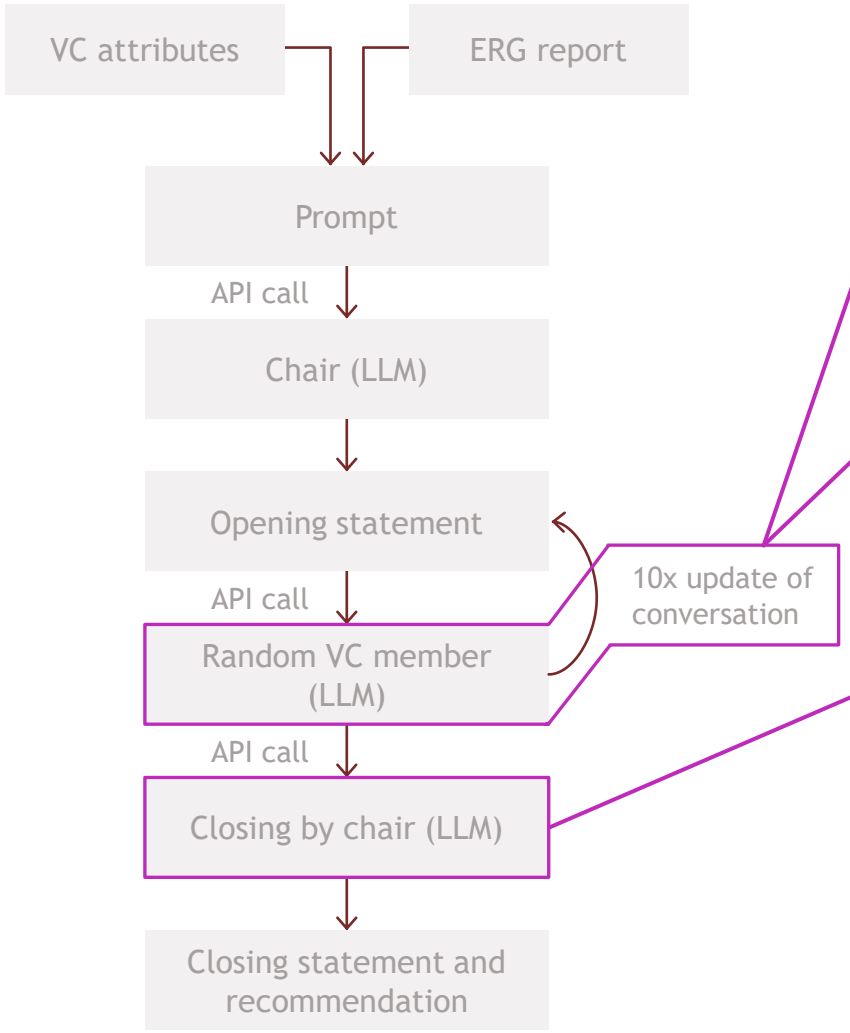
Results: What worked well?



- **Opening statement**
  - Chair was able to accurately summarize the ERG report
  - Provided a solid platform for the discussion
- **Realistic and accurate discussion**
  - Different virtual committee members were able to argue specific perspectives based both on the role assigned to them and the evidence presented (i.e., the ERG report)
  - Conversation flows well and makes sense in the context of the objective and the evidence presented
  - Themes discussed were similar in the human and virtual committee discussions, particularly with regards to summarizing the evidence and issues

API: Application Programming Interface, ERG: Evidence Review Group, LLM: Large Language Model, VC: Virtual Committee

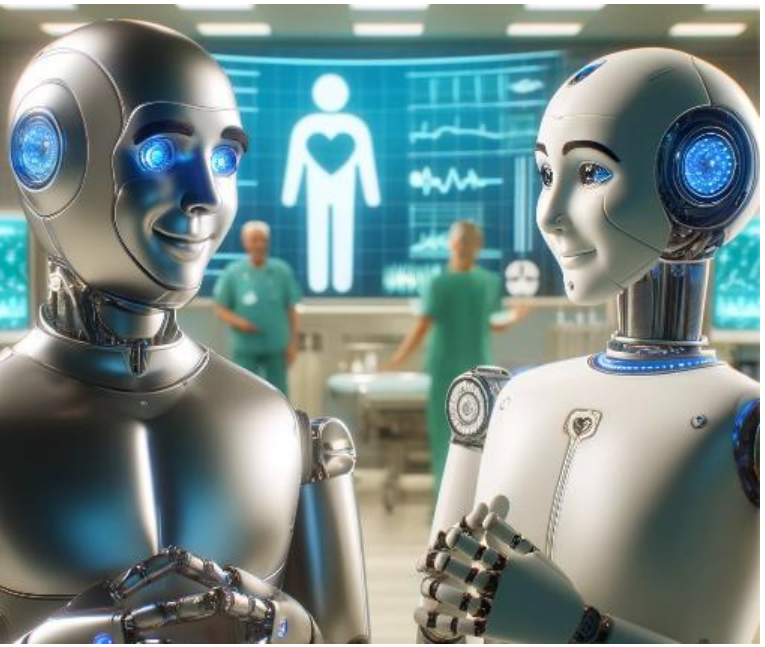
Results: What were the limitations?



- **Consensus**
  - Despite toggling willingness to argue to high or very high for several members and making “next speaker” random, it was very difficult to stop the LLMs reaching consensus even after several runs
- **Focus**
  - Despite the conversation flowing well, it was hard to get the committee to focus on issues beyond the primary ones (in this case cure, extrapolation and re-treatment)
- **Recommendation**
  - Final recommendation made was conditional based on the manufacturer submitting further evidence and analyses
  - Despite tweaks it was difficult to get the virtual committee to give a recommendation that did not involve further data collection

API: Application Programming Interface, ERG: Evidence Review Group, LLM: Large Language Model, VC: Virtual Committee

Potential directions for further development



- **More complex interactions**
  - Make probability of speaking conditional on prior speaker
  - Vary response lengths
- **Expanded context**
  - Increased token limits allow for providing additional documents
  - Retrieval-Augmented Generation (RAG) may be used to circumvent token limits
  - Additional context may be made selectively available to specific virtual committee members
- **Other settings**
  - LLMs have potential to simulate discussions for different HTA bodies, languages and settings in general

HTA: health technology assessment, LLM: large language model, RAG: retrieval-augmented generation

Thank you