Evaluating the use of GPT-4 in Health Economics and Market Access Projects.

Johnson&Johnson
MedTech

24/9/2024

Agenda

- An AI-enhanced SLR case study
- An Internal JnJ AI-value brief POC
- Our Thoughts

Assessing Generative AI's capability in Systematic Literature Reviews (SLRs), a case study.

Nikolaos Takatzoglou

External PhD Candidate
Erasmus School of Health Policy & Management

EMEA HEMA WC & Healing Lead & HEMA Lead Greece & Cyprus, Johnson & Johnson

Maureen Rutten-van Mölken

Prof. of Economic Evaluation of Innovations for Health
Erasmus School of Health Policy & Management

Mike Kukushkin

Former External Consultant of JNJ

Cindy Tong

Director, EMEA RWE & Global VBHC Analytics

Johnson & Johnson

Ken Redekop

Associate Professor of Health Technology Assessment
Erasmus School of Health Policy & Management

Gautam Iyer

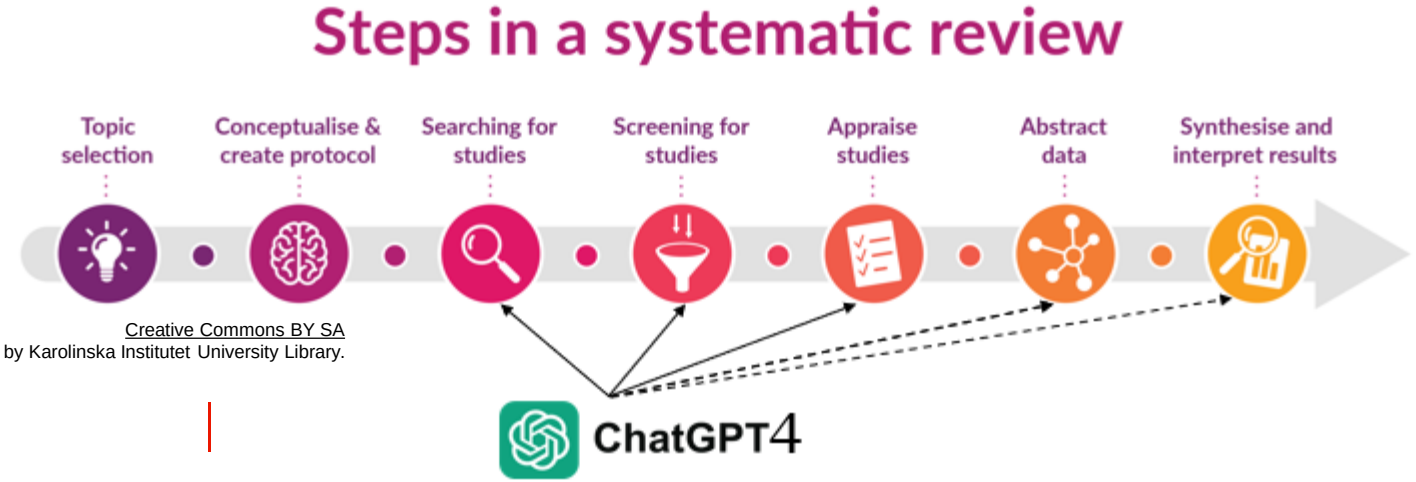
HEMA Intern

Johnson & Johnson

Two-Fold Research Scope

- Identify publications that predict the HTA outcomes and corresponding drivers; these publications will be used as features in an HTA ML prediction model we are developing.

- Can Generative-AI, such as GPT4, help with SLR?



Search Strategy

1. Humans chose databases
(Pubmed, Scopus, Embase, Arxiv, iHTA, ISPOR)

2. Humans devised search terms

3. **GPT4 suggested additional search terms**

↓

humans accepted many of the additions

- Human(s)
- Robot
- Human + Robot

Screening Phase

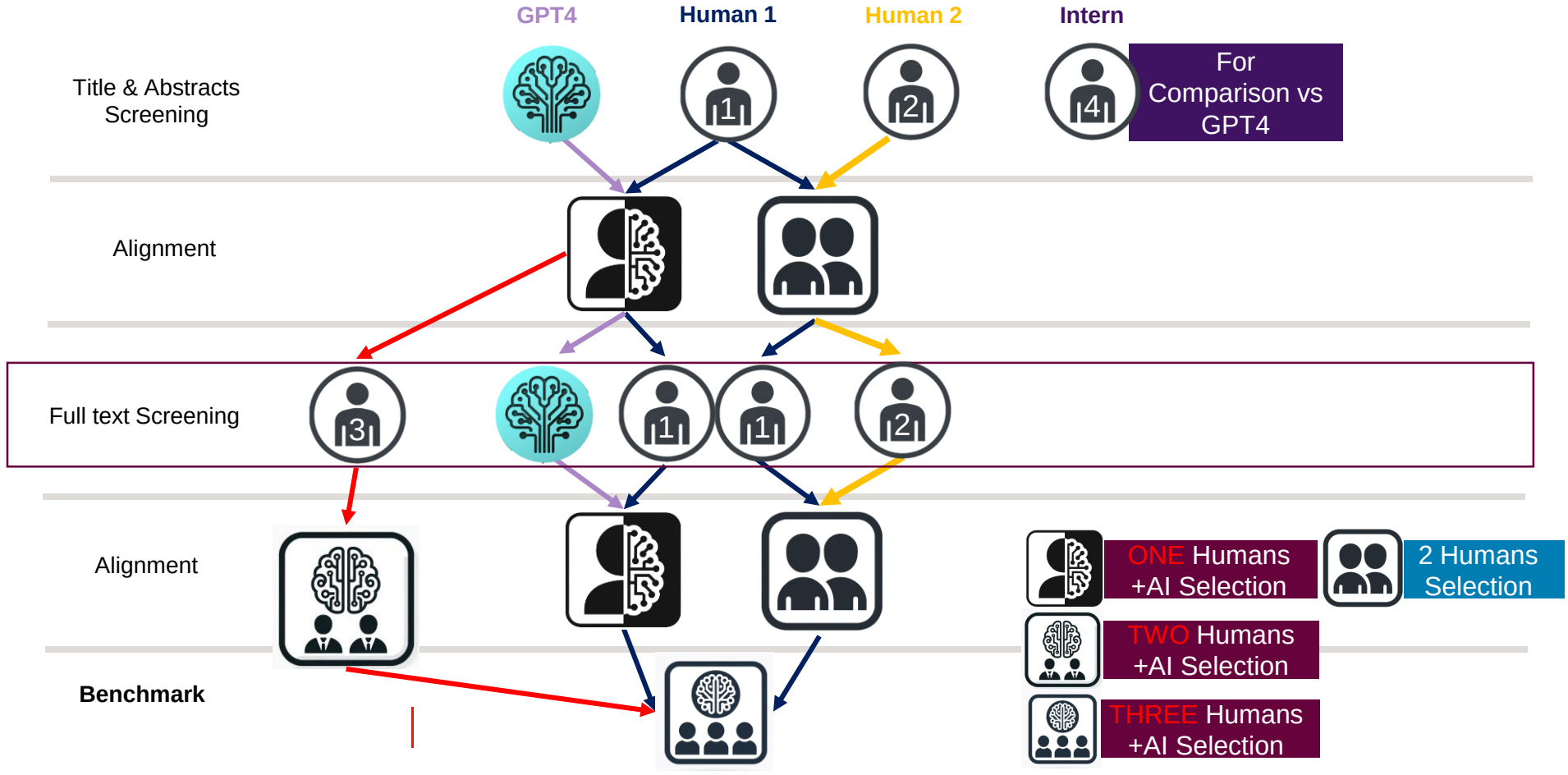
1. 2406 Title & Abstracts (T&A) were retrieved **manually**

2. 2406 T&A =
   972 pages split in 1500 words chunks =
   440 prompts (iterations) using the prompting "Secret Sauce"

Human(s)

Robot

Human + Robot


GPT4
Human 1
Human 2
Intern
Title & Abstracts Screening
For Comparison vs GPT4
Alignment
Full text Screening
Alignment
Benchmark
ONE Humans +AI Selection
2 Humans Selection
TWO Humans +AI Selection
THREE Humans +AI Selection

Instructions for AI

inspired by Tree of Thoughts, Chain of thought and self-Consistency methods[1]

Four-PERSONAS: HTA Expert, Librarian (SLR expert), Statistician, DS/ML Expert

I. PLANNING PHASE

1. Each persona reads
   I. the input [1500-word chunk]
   II. "Chief Scientist's" 3 Inclusion and 2 exclusion criteria

To be included studies:

- Studies that have used statistics or Machine Learning to
  - Predict HTA decisions.
  - Identify features/drivers of HTA decisions
  - Compare HTA decisions of different HTA bodies

To NOT be included studies / Irrelevant studies:

- studies that discuss about /report HTA decision(s) but do not focus on showing the prediction or the drivers of that HTA decision
- studies that discuss about the HTA outcome of a specific intervention; this is too narrow of a scope to be included in our SLR

II. EXECUTION PHASE

1. Yao, Shunyu, et al. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." arXiv:2305.10601, 2023

Instructions for AI

inspired by Tree of Thoughts, Chain of thought and self-Consistency methods[1]

Four-PERSONAS: HTA Expert, Librarian (SLR expert), Statistician, DS/ML Expert

I. PLANNING PHASE

1. Each persona reads
   I. the input [1500-word chunk]
   II. "Chief Scientist's" 3 Inclusion and 2 exclusion criteria
2. Devises a plan on how to assess the T&A for inclusion/exclusion
3. Critique each others and own's work
4. Based on critique devise a final combined plan

II. EXECUTION PHASE

The 4-personas, acting upon their final plan, develop:

I. Potential inclusion reasons for each T&A
II. Potential exclusion reasons for each T&A

Instructions for AI

inspired by Tree of Thoughts, Chain of thought and self-Consistency methods

3 polymaths independently assess the **inclusion** and **exclusion** arguments, and give an 1-5 ranking based on the following categories:

(1) Totally Irrelevant = fits >1 exclusion criteria perfectly

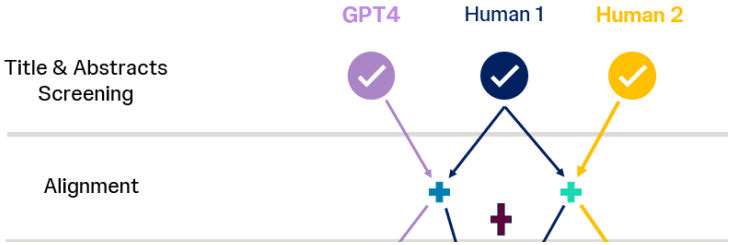
(2) Marginally Relevant = fits >1 inclusion criteria but vaguely

(3) Ambiguously Relevant = Probably meets 0 inclusion and exclusion criteria

(4) Generally Relevant = Meets >1 inclusion & >1 exclusion criteria

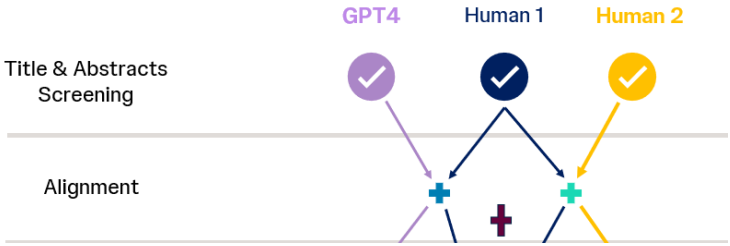
(5) Precisely Relevant = Meets >1 inclusion criteria & 0 exclusion criteria

Results – GPT4 Title & Abstracts phase



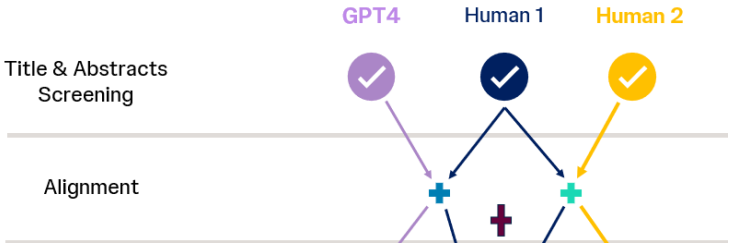
Agreement before Alignment	Human1	Human2
GPT4	91%	92%
Human1	-	95%

Results – GPT4 Title & Abstracts phase



	AI: Yes	AI: No	Human convinced by AI
Human 1: Yes	2%	4.5%	60% (out of 4.5%)
Human 1: No	4.9%	88.6%	12.8% (out of 4.9%)

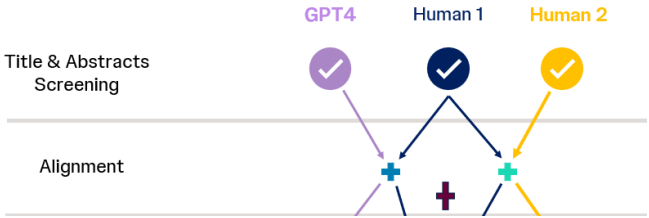
Results – GPT4 Title & Abstracts phase



After Alignment	Accuracy*	Sensitivity*
AI	92.6%	44.9%
Human1	96.1%	68.9%
Human2	98.4%	78.7%

*Benchmark = 2 humans + AI selection after alignment was considered the ideal selection

Results – GPT4 Title & Abstracts phase



After Alignment	Accuracy*	Sensitivity*
AI+Human1	98.4%	77.2%
2 humans	99.3%	93.7%
2 Humans + AI	100%	100%

*Benchmark = 2 humans + AI selection after alignment was considered the ideal selection

Results – GPT4 Full-Text

After Alignment	Accuracy*	Sensitivity*
AI+1 Human / AI+2 Humans	99%	72.5%
2 humans	99.8%	96.1%
3 Humans + AI	100%	100%

*Benchmark = 3 humans + AI selection after alignment was considered the ideal selection

Results – GPT4 Full-Text

After Alignment	Accuracy*	Sensitivity*
AI	58.1%	67.7%
Human2	85.4%	81.6%
Human3	85.6%	89.5%
Human1	88.9%	98.7%

*Benchmark = 2 humans + AI selection after alignment was considered the ideal selection

Key observations

- The SLR topic was too broad as shown by low sensitivity for both Human and AI.
- AI manages to convince Human in some cases.
  - Sticks more to the inclusion and exclusion criteria
  - Helps with some missed articles by Humans
  - Deciphers poorly written abstracts
- Better refined inclusion/exclusion criteria helps AI performance.
- Full-text GPT4 outcomes are much worse compared to T&A
- Can Generative-AI, such as GPT4, help with SLR?
  Yes, based on preliminary results it can help, but not replace a human and still needs more work.

A JnJ Case Study: AI-value brief POC

J&J MedTech

Assessing Generative AI's capability in AI-value brief POC

Nikolaos Takatzoglou
EMEA HEMA WC & Healing Lead
HEMA Lead Greece & Cyprus
EMEA MedTech HEMA

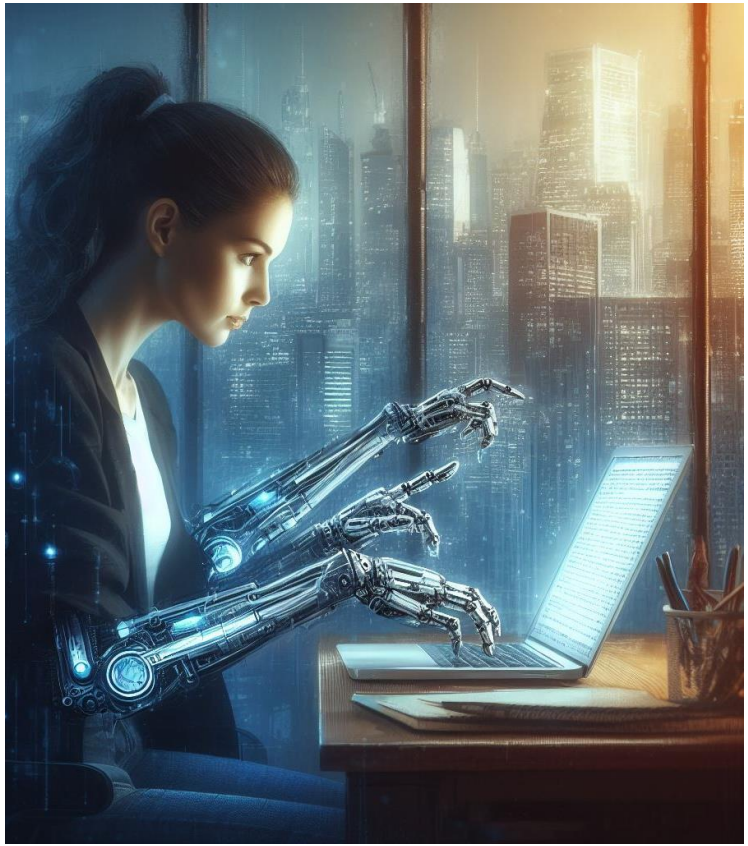
Polina, Vrouchou
Associate Director, Global Health Technology Assessment
MedTech HEMA

Naj Gunja
Associate Director, Wound Closure and Healing
MedTech HEMA

Cindy Tong
Director, EMEA RWE & Global VBHC Analytics
MedTech HEMA

Project Scope

- Proof-of-concept (POC) value brief for JnJ Product
- Semi-automatic process with ChatGPT4
- Final deliverable: 10-20 pages Value Brief for Internal Use, based on NICE 1000 pages evaluation document



Our Vision

- An End-to-End automated system "Auto-GPT Draft Value Brief Creator" with “push a button”

- Final editing will be performed by human with help of AI

POC considered successful

Scarce hallucinations and good accuracy

- Controlled by our prompting methodology
- Process for input/output evaluation

FIRST DRAFT:

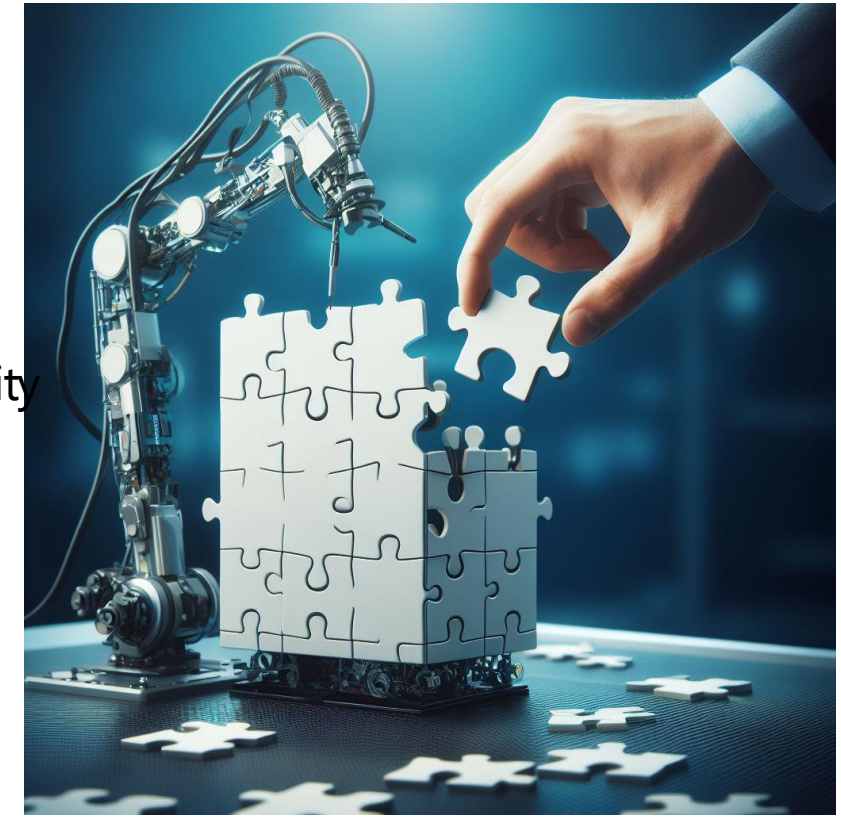
- First draft required ~20 less human working hours, vs 100% in-house draft
- Moderate quality vs Human Draft due to lack of flow, caused by input word count limit.

FINAL DRAFT:

- Additional 8-24 hours is expected for final human editing
- Expecting better quality vs human version

Key learnings
Some of them...

- Human review critical for accuracy
- Fewer pages but more relevant = higher quality
- Human touch for final version necessary
- It's feasible and unavoidable



Our Thoughts

Our Thoughts

- The use of Large Language Models can help with summarizing evidence but also identifying them.
- Prompting is very important.
- LLMs are not capable to replace a human, yet.