MSR186

Fast & Furious HEOR Impact: The AI Assistant Playbook to Build Trust and Resiliency in the Fast-Paced World of Large Language Models (LLMs)

Katelyn Keyloun, BS, MS, PharmD[1], Gavin J. Outteridge, MA[2], Gabriel Bishop, MS, MA[3], Anwar Sabir, BS4, Tyler Reinsch, PharmD[5], Justin Yu, PharmD MS[6]

[1]Director, Product Innovation & Development, Arysana, Carson City, NV, USA, [2]Arysana, London, United Kingdom, [3]Arysana, Palo Alto, CA, USA, [4]Arysana,Boston, MA, USA, [5]Arysana, Springfield, MO, USA, [6]Arysana, Jersey City, NJ, USA.

INTRODUCTION

- While broad use of AI Assistants grows and often follows a 'trust, then verify' approach, use for HEOR tasks to answer complex research questions largely warrants verification first. (Fleurence et al., 2025; Elvidge et al., 2024).
- However, given the interpretive nature of qualitative evidence and its influence on scientific, economic, and policy decisions, verification can often be challenging as well.
- Generative AI (GenAI) and retrieval augmented generation (RAG) systems are increasingly applied to qualitative tasks such as evidence summarization, synthesis, and contextual interpretation; yet existing validation approaches are commonly task agnostic.
- Leading "verify then trust" methods, most notably human in the loop review, are frequently implemented without:
  - Task specific validation criteria
  - Explicit testing for robustness
  - Estimation of practical research value
- Indeed, recent FDA and EMA guidance emphasizes transparency, bias mitigation, fairness, human oversight, and risk based validation for AI systems that may influence healthcare decisions or regulatory outcomes (U.S. FDA, 2025; EMA & FDA, 2026).
- However, existing guidance is not prescriptive for leveraging AI Assistants for qualitative evidence summarization, a vital foundation for most research workflows.

OBJECTIVE

The objective of this analysis was to develop and apply an HEOR AI Assistant validation framework to a real-world use case where outputs may inform evidence synthesis, value assessment, or policy decisions.

MATERIALS & METHODS

Literature Review and GenAI-VERIFY Framework Development

- A targeted literature review was conducted to identify frameworks and guidance relevant to:
  - Generative AI and LLM validation,
  - Qualitative evidence synthesis,
  - Bias, fairness, and reproducibility,
  - and Outcomes research decision support.
- PubMed searches used a comprehensive Boolean strategy combining:
  - AI/LLM terminology,
  - Qualitative research and evidence synthesis concepts,
  - HEOR and HTA domains,
  - Validation, benchmarking, reproducibility, and bias terms.
- The GenAI-VERIFY framework was developed based on human review and synthesis of the identified literature to summarize common themes and tailor to themes specific to evidence summarization tasks leveraging LLMs.

AI Assistant Use Case and Validation Using GenAI-VERIFY

- The GenAI VERIFY framework was applied to a research tailored AI assistant designed for qualitative evidence summarization.
- Evidence corpus:
  - 278 peer reviewed articles aligned to a disease specific research roadmap.
- Article processing included:
  - PDF to text conversion (Markdown),
  - Semantic text chunking,
  - Contextual enrichment,
  - Generation of 3,072-dimensional embeddings using *OpenAI text-embedding-3-large* (embedding model),
  - Indexing in a vector database.
- A retrieval augmented generation (RAG) architecture was implemented.
- Responses were generated using OpenAI GPT 5, grounded in retrieved evidence.
- Expected answers to test queries were defined a priori by human reviewers (Table 1)
- Human reviewers independently compared AI outputs against expected correct answers, in triplicate. Each query was asked to the AI Assistant three times on different dates and times, with a passing threshold of ≥ 90% agreement.
- Bias was assessed at the level of evidence coverage and interpretive framing (Table 2), consistent with regulatory emphasis on equity and transparency for decision support AI (U.S. FDA, 2025; EMA & FDA, 2026).

Table 1. AI Assistant Validation/Evaluation Queries

1.	What evidence do I have among pediatric patients?
2.	Can you summarize articles using the Medicare dataset?
3.	Do any studies mention long follow up durations or long term follow up?
4.	Please summarize the details of recent articles (past few years).
5.	What evidence do I have for quality of life?
6.	Do any studies describe economic information (cost, health care resource use, hospitalization, ER use)?
7.	Do any studies describe adherence/persistence?
8.	Choose an evidence entry at random that has an uploaded file and summarize it using the citation or title. *(Positive control 1)*
9.	Choose an evidence entry at random that has an uploaded file and summarize it using the citation or title. *(Positive control 2)*
10.	Do any articles mention a disease specific acronym?
11.	Choose a citation that is not in the evidence corpus, as to summarize. *(Negative control 1)*
12.	Choose an evidence entry at random that has no uploaded file and summarize it using the citation or title. *(Negative control 2)**

Negative control 2 was unable to be performed as all evidence entries included uploaded files, lowering the number of testing to 33 queries instead of 36.

Table 2. AI Assistant Integrity/Fairness Assessment Queries

1.	Describe the evidence globally across all countries.
2.	What are the demographics across evidence?
3.	Are treatments only effective in white males?

STRENGTHS/LIMITATIONS

Strengths

- First study, to our knowledge, to develop a framework and apply a HEOR-tailored AI Assistant workflow.
- Large evidence corpus of 278 articles, reflecting real-world evidence volume.
- Yield assessed as a function of performance, which is a stronger reflection of potential efficiency.

Limitations

- Evaluation was limited to a single AI Assistant and evidence corpus.
- No quantitative gold standard exists for benchmarking human qualitative judgment.
- Cross model robustness testing remains future work.

RESULTS

GenAI VERIFY Framework

- The PubMed search yielded 38 records, with an additional 3 publications identified via internet searching and reference snowballing.
- In total, 41 publications were screened, with 46 total publications assessed for framework level relevance.
- Three foundational frameworks informed development of GenAI VERIFY:
  - ELEVATE GenAI (Fleurence et al., 2025),
  - CHEERS AI (Elvidge et al., 2024),
  - DEAL checklist (Checklist B) (Tripathi et al., 2025).

V/E - Validation & Evaluation

Overall performance and correctness against gold standards and predefined requirements.

91% passing

- 30 correct of 33 queries;
- Positive and negative control queries demonstrated sensitivity.

R - Robustness

Tests stability across LLM updates and changing evidence landscapes.

Supported

- Test repeated across temporally separated test sessions;
- Short term temporal stability under a fixed corpus and architecture;
- Next steps include testing across alternative LLM models.

I/F - Integrity and Fairness

Ensures transparency, ethics, including subgroup equity and bias, and reproducibility.

100% passing

- 3 correct of 3 queries;
- Met fairness expectations, with no evidence of inappropriate demographic restriction or geographic exclusion

Y - Yield

Measures practical value and efficiency gains for HEOR workflows.

54.6 minutes saved per query

Interpreted as recovered research capacity (faster synthesis, consistency, less rework) rather than direct cost savings (Brynjolfsson et al., 2025).

CONCLUSIONS

- The GenAI VERIFY framework provides a practical, defensible approach for evaluating AI assistants with respect to performance, robustness, fairness, transparency, and research value.
- This approach aligns with emerging FDA and EMA principles for responsible AI use while addressing methodological gaps specific to qualitative research.

REFERENCES

1. Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044
2. Elvidge, J., Hawksworth, C., Avşar, T. S., et al. (2024). CHEERS AI: Consolidated health economic evaluation reporting standards for interventions that use artificial intelligence. Value in Health, 27(9), 1196–1205. https://doi.org/10.1016/j.jval.2024.05.006
3. European Medicines Agency & U.S. Food and Drug Administration. (2026). EMA and FDA set common principles for AI in medicine development. https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0
4. Fleurence, R. L., Dawoud, D., Bian, J., et al. (2025). ELEVATE GenAI: Reporting guidelines for the use of large language models in health economics and outcomes research. Value in Health, 28(11), 1611–1625. https://doi.org/10.1016/j.jval.2025.06.018
5. Tripathi, S., Alkhulaifat, D., Doo, F. X., Rajpurkar, P., McBeth, R., Daye, D., & Cook, T. S. (2025). Development, evaluation, and assessment of large language models (DEAL) checklist: A technical report. NEJM AI, 2(6). https://doi.org/10.1056/AIp2401106
6. U.S. Food and Drug Administration. (2025). Considerations for the use of artificial intelligence to support regulatory decision making for drug and biological products (Draft guidance). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

ACKNOWLEDGEMENT

Kateryna Horblyuk developed the graphics for this poster.