Agentic AI vs RAG in Complex HTA Writing Tasks: A Randomized Preference Study

Pharos Labs

A. Wiehe and F. Woeste.
Pharos Labs, Germany

Introduction

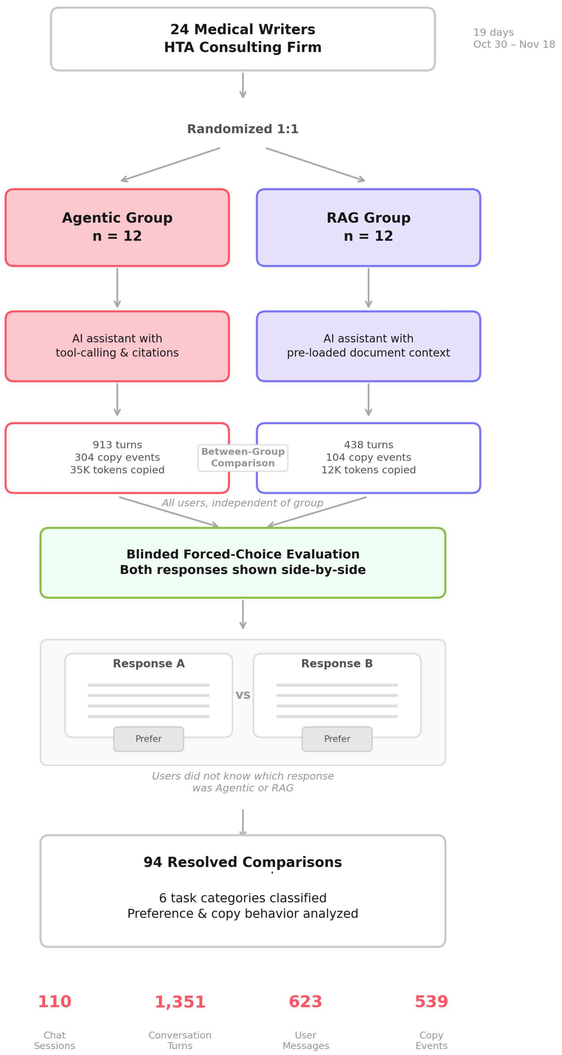
AI writing assistants for Health Technology Assessment (HTA) typically use Retrieval-Augmented Generation (RAG): documents are pre-retrieved and injected into the model's context before generating a response. This single-pass approach is effective for straightforward tasks but constrains the model to only the content retrieved upfront.

Agentic architectures offer an alternative: the model autonomously decides when and what to retrieve through iterative tool calls, enabling multi-step reasoning. However, whether this flexibility translates into user-preferred outputs for professional medical writing is unknown.

Objective

To compare user preference and content adoption for an agentic AI assistant versus conventional RAG across different HTA writing task types.

Methods



Results

Head-to-Head Preference

60% Agentic

40% RAG

n=94 | p=0.08 | Cohen's h=0.19

2.8× more content adopted

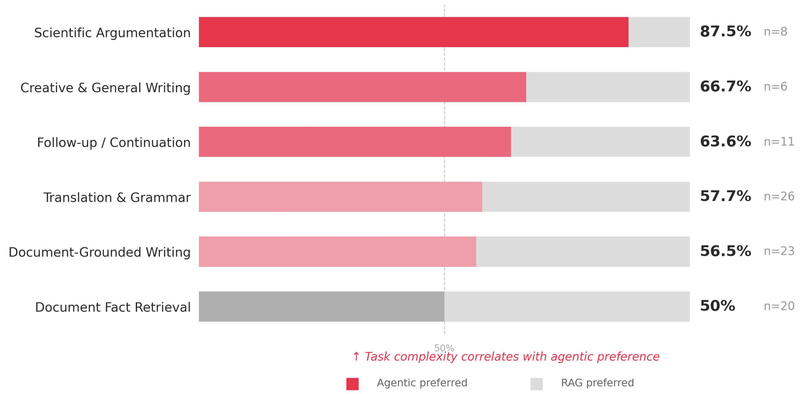
35K vs 12K tokens

2.9× more copy events

304 vs 104 copy events

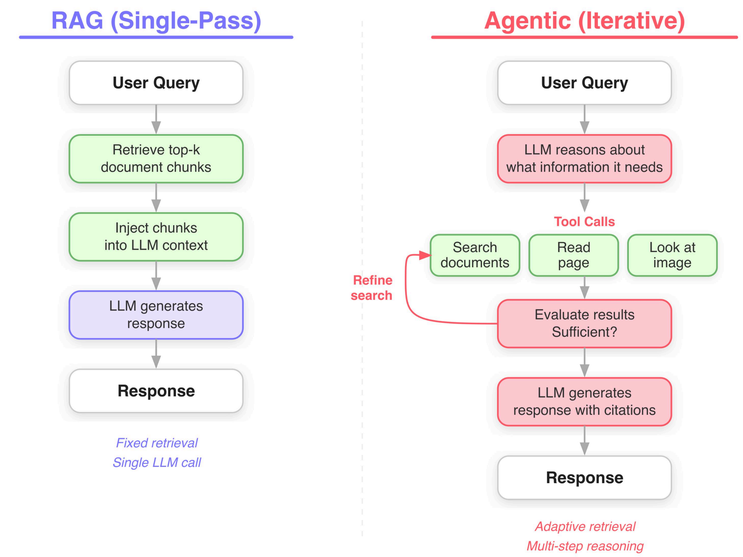
Between-group comparison; groups had comparable baseline activity.

Preference by Task Complexity



Limitations

- Single consulting firm, single domain (German HTA)
- Small n per task category (6–26 comparisons)
- 19-day study period
- Manual prompt classification (single rater)
- Copy behavior is between-groups, not within-subject



Conclusions

Key Finding Medical writers adopted 2.8× more content from agentic AI outputs than RAG, with preference strongly correlated with task complexity.
Copy behavior data shows 2.8× more content adopted from agentic outputs (between-group), directionally consistent with the 60% head-to-head preference.
Implication For complex HTA tasks requiring multi-step reasoning, agentic architectures may better serve medical writers than single-pass RAG.

References

1. Lewis, P. et al (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.
2. Schick, T. et al (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. NeurIPS 2023
3. Zheng, L. et al (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track.

Disclosure

Author(s) of this presentation have the following to disclose concerning possible financial or personal relationships with commercial entities that may have a direct or indirect interest in the subject matter of this presentation:

Anton Wiehe:
Co-founder and employee – PHAROS Labs GmbH

Florian Woeste:
Co-founder and employee – PHAROS Labs GmbH

Contact

Anton Orell Wiehe, Head of AI
anton.wiehe@pharos-labs.com
linkedin.com/in/antonwiehe
pharos-labs.com

Find out more in a demo!

PHAROS Labs GmbH · MD Florian Woeste & Timm Volmer
Alter Teichweg 25a · D–22081 Hamburg · Amtsgericht Hamburg HRB 151603