Comparative feedback across models

This are the results from the feedback given by the scientists comparing RAG fusion v Agentic RAG (with Claude 3) and RAG fusion v Agentic RAG (with Claude 3.7): <u>GenAI comparative testing</u>

Clarity and Structure

- Agentic RAG is repeatedly praised for better formatting and clearer summaries, especially in initial responses
- **RAG Fusion**, however, is also credited with better structural organization and concise output in some cases, particularly when using Claude 3.7.

Technical Content and Detail ${\mathscr O}$

- Agentic RAG delivers more detailed technical answers, often pulling quantitative data and VITIC tables, but tends to be verbose and lacks critical references and mechanistic insights.
- **RAG Fusion** provides mechanistic context and is more concise, especially with Claude 3.7, but sometimes omits supporting assay data or VITIC content, and may fail to address the core question.

Consistency and Repeatability ${\mathscr O}$

• Multiple entries report repeatable UI/timeout errors affecting both systems (e.g., failures after 1-3 minutes)

Relevance and Responsiveness ${\mathscr O}$

- Both models occasionally drift from the question (e.g., Fusion discussing carcinogenicity when not asked).
- Claude 3.7 integration shows notable improvement in relevance and conciseness for both systems, though verbosity remains a concern in some Agentic responses.

Key patterns

Dimension	Agentic RAG	RAG fusion
Formatting	Clear, readable summaries	Structured but sometimes lacking supporting detail
Detail	High (e.g., VITIC data, numbers)	Moderate; better mechanistic focus, sometimes lacking VITIC
References	Often missing	Sometimes included
Verbosity	Often too verbose	More concise, especially with Claude 3.7
Relevance to Question	Generally good, with occasional drift	Occasionally off-topic (e.g., unnecessary carcinogenicity)
Repeatability	UI/timeout issues reported	UI/timeout issues reported

Conclusion

Agentic RAG excels in data richness and technical depth, making it valuable for expert audiences, but struggles with verbosity, lack of references, and sometimes incomplete mechanistic context.

RAG Fusion, especially when powered by Claude 3.7, is more balanced and concise, delivering structured insights with occasional gaps in supporting data.

Neither model is consistently superior, but **Claude 3.7 integration marks a noticeable step forward**, particularly for RAG Fusion in terms of relevance and conciseness.