Human v Test Runner feedback comparison

Based on Human comparative testing, [LGA-223] System testing - BE comparative for RagFusionCombinedV2 vs Agentic Multi - Production and [LGA-225] System testing - Claude 3.7 observations for RagFusionCombinedV2 and Agentic Multi - Production this is the result of comparing human testing with the test runner.

Claude 3

Consistencies:

- **Both sources agree** that:
  - Agentic has stronger depth and structure, especially for complex topics.
  - Fusion is more direct, but may be too superficial or incomplete.
  - Both have issues with answer consistency and context inclusion.

Discrepancies:

- The Test Runner scores Fusion higher on relevancy and recall, but human testers often prefer Agentic due to depth and nuance.
- The automated scores underrepresent structural and mechanistic insight, which human reviewers care about more (e.g., in mechanistic genotoxicity or DES analysis).
- Human feedback penalizes Fusion more heavily for omissions and misdirection (e.g., failure to mention mutagenicity or focusing on carcinogenicity instead).

Claude 3.7

Consistencies:

- Fusion is concise and relevant, but sometimes omits context or nuance (captured by high relevancy and low recall in test metrics, and echoed in human remarks).
- Agentic RAG provides depth and structure, which helps with technical accuracy, but hurts semantic overlap (visible in low BERT/Rouge scores and human complaints about verbosity).

Discrepancies:

- Human reviewers value Agentic's technical rigor (especially for genotoxicity and potency comparisons), but test runner penalizes it for semantic divergence from expected reference (low BERT/Rouge).
- Fusion's brevity gets high automated scores, but misses mechanistic or contextual breadth, flagged repeatedly by humans (e.g., lack of Vitic data, over-generalization).

Conclusion

Overall, there is a meaningful degree of alignment between human testing and test runner results, particularly around the strengths of RAG Fusion in producing concise, faithful, and directly relevant answers. Both evaluation methods consistently recognize Agentic RAG's deeper contextual coverage and richer technical detail, though this is often undervalued by automated metrics like BERT and Rouge.

Key discrepancies emerge where human reviewers prioritize mechanistic insight, regulatory relevance, and structured reasoning (elements that current automated scoring fails to fully capture). This indicates that while automated metrics provide useful signals, human evaluation remains essential for assessing scientific and domain-specific answer quality.