captain™ evals

We are excited to announce Captain's leading accuracy on the Open RAG Benchmark.

Methods

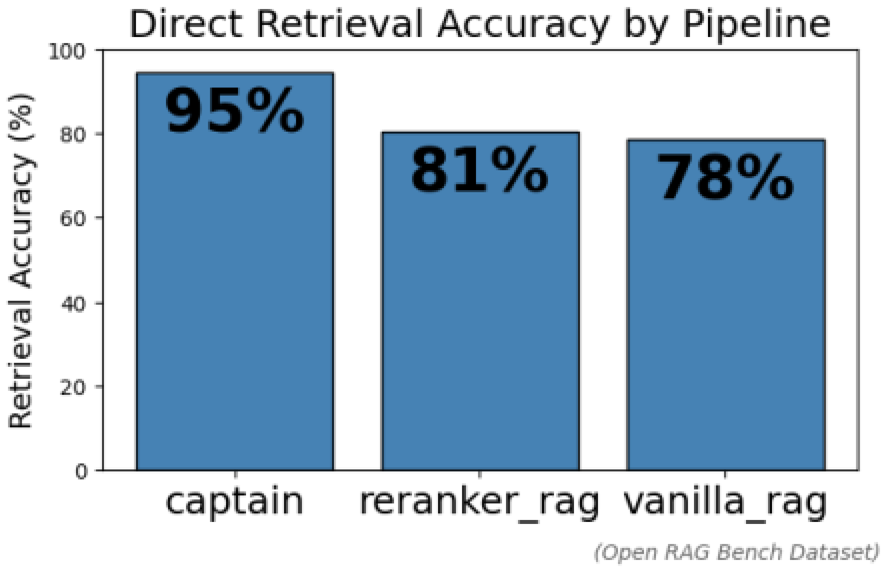
Captain achieves frontier accuracy through a weighted combination of advanced techniques in natural language processing and retrieval.

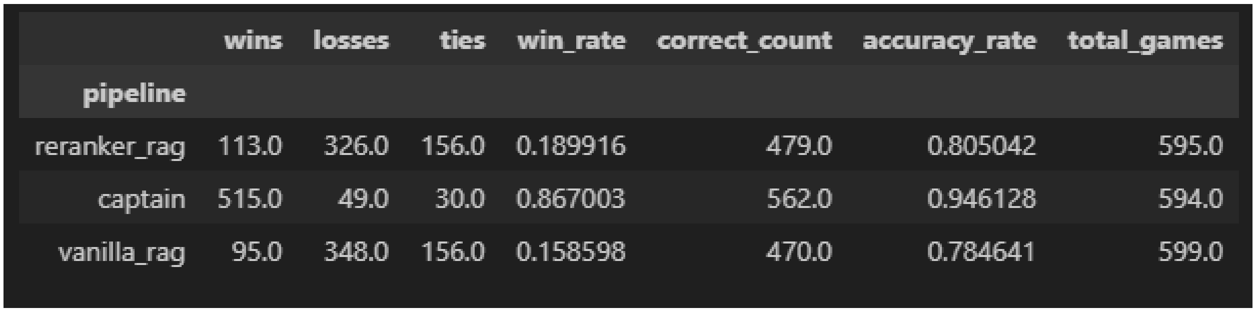
Key innovations include:

- normalizing vector embeddings to ensure consistent semantic representations across modalities,
- using a multi-pronged approach with four different indexing strategies:
    - multimodal embeddings
    - contextualized embeddings
    - BM25
    - Full-text search
- employing sophisticated chunking algorithms that leverage cosine distance conditionals to identify optimal document boundaries.

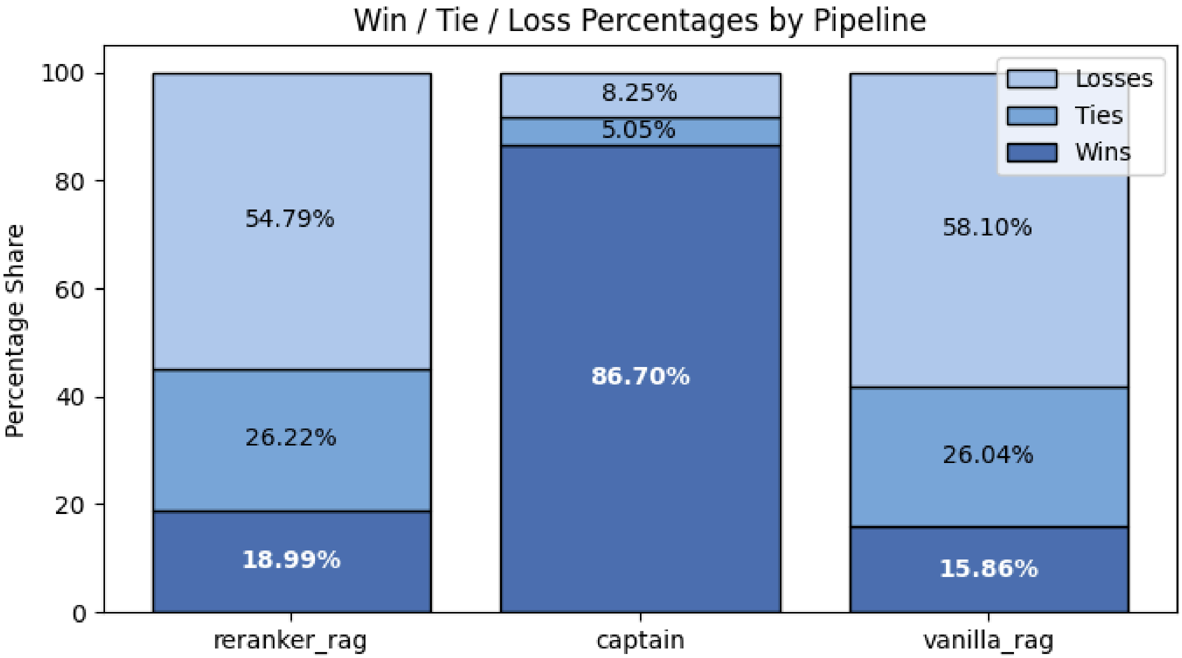
Results

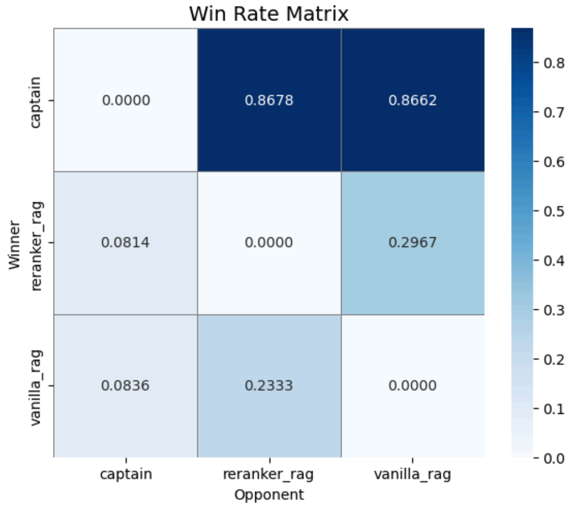
This comprehensive approach allows Captain to achieve a generalized accuracy of 95% on the Open RAG Benchmark (with an LLM-as-judge).

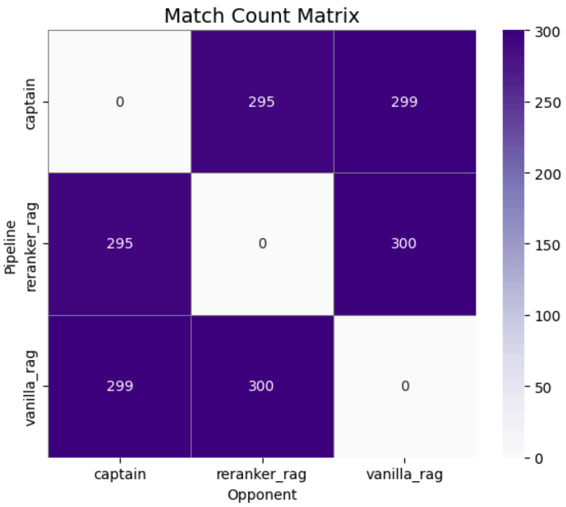




pipeline	wins	losses	ties	win_rate	correct_count	accuracy_rate	total_games
reranker_rag	113.0	326.0	156.0	0.189916	479.0	0.805042	595.0
captain	515.0	49.0	30.0	0.867003	562.0	0.946128	594.0
vanilla_rag	95.0	348.0	156.0	0.158598	470.0	0.784641	599.0

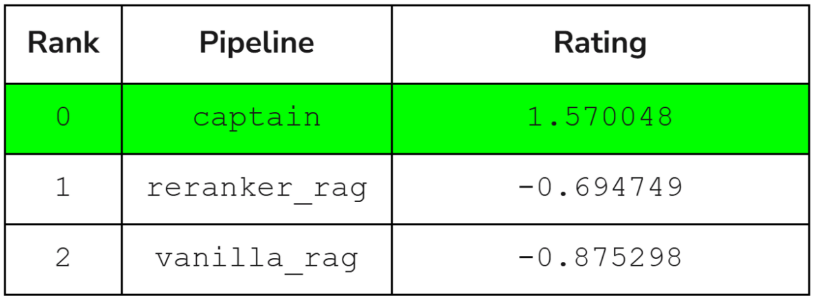






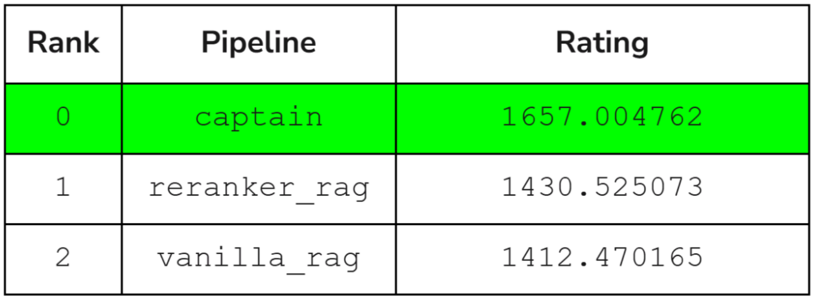
Elo Comparison

Final Bradley–Terry Leaderboard

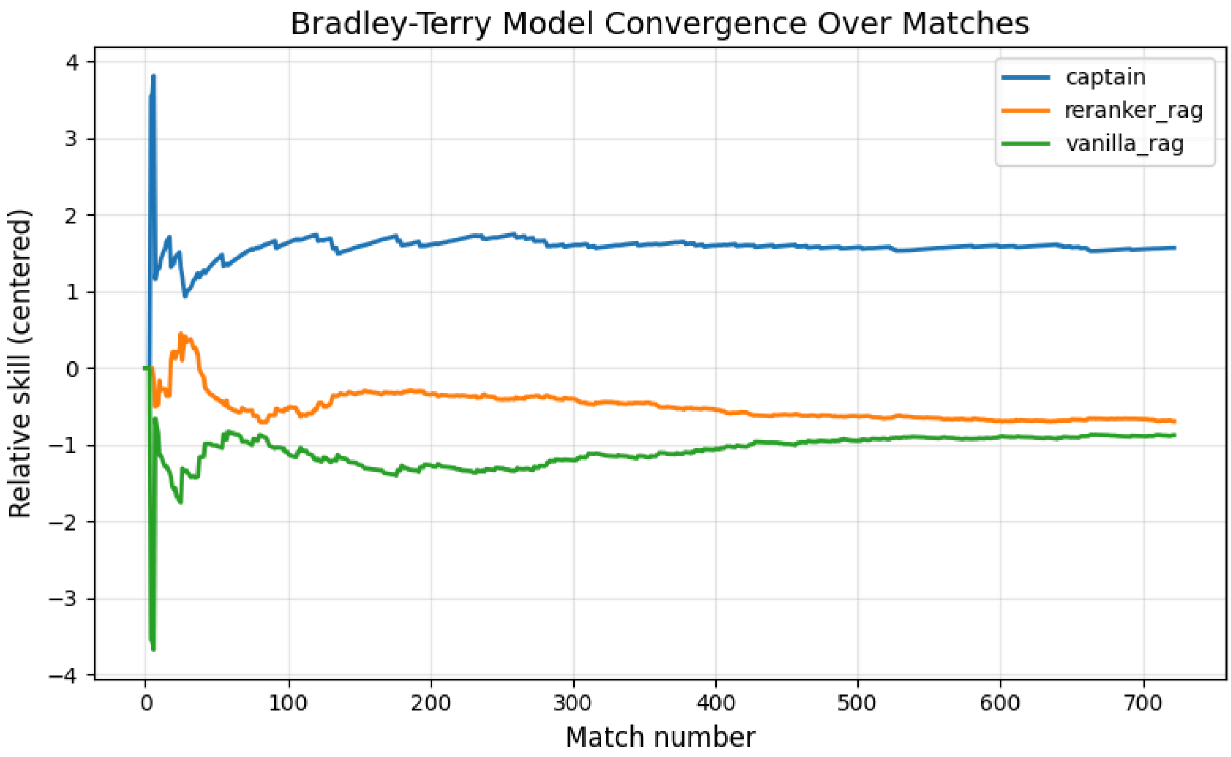


Rank	Pipeline	Rating
0	captain	1.570048
1	reranker_rag	-0.694749
2	vanilla_rag	-0.875298

Final Bradley–Terry Elo-Scaled Leaderboard



Rank	Pipeline	Rating
0	captain	1657.004762
1	reranker_rag	1430.525073
2	vanilla_rag	1412.470165


Bradley-Terry Model Convergence Over Matches
captain
reranker_rag
vanilla_rag
Relative skill (centered)
Match number