ScholarMind — Semantic Query Extraction Benchmark

1 Introduction

This report summarizes our transition from a traditional keyword extraction approach to a semantic retrieval design for scientific documents. Unlike keyword-based methods, embeddings capture the **semantic meaning** of queries, enabling retrieval of relevant documents even when exact words do not match.

2 Dataset Preparation

We used the **SciFact** dataset from the BEIR benchmark for testing semantic retrieval. The dataset includes scientific claims (queries) and relevant abstracts (documents).

2.1 Sample Query and Document

- **Sample Query:** *CRISPR gene editing in humans*
- **Sample Document:** First document in the corpus: "CRISPR-Cas systems allow precise genome editing and have been applied to human cells for therapeutic purposes..."

3 Benchmarking Models

We benchmarked several sentence embedding models to evaluate their performance for semantic retrieval:

Model	Dimension	Query Latency (ms)	Vector Norm	Query–Doc Similarity
SciBERT	768	129.99	17.600000	0.5054
SPECTER	768	383.98	21.959999	0.6242
MPNet	768	588.03	1.000000	0.1254
MiniLM	384	67.85	1.000000	0.1329

3.1 Chosen Model

4 Choosing SPECTER

From the benchmarking, **SPECTER** demonstrated the highest semantic similarity with relevant scientific documents. It is thus selected for ScholarMind's semantic retrieval

🔥 Benchmarking SciBERT ...

🔥 Benchmarking SPECTER ...

🔥 Benchmarking MPNet ...

🔥 Benchmarking MiniLM ...

	Model	Dim	Latency (ms)	Vector Norm	Query-Doc Similarity
0	SciBERT	768	129.99	17.600000	0.5054
1	SPECTER	768	383.98	21.959999	0.6242
2	MPNet	768	588.03	1.000000	0.1254
3	MiniLM	384	67.85	1.000000	0.1329

Figure 1: Comparison of embedding models on query-document similarity metrics.

engine because:

- Excels in capturing semantic intent of scientific queries.
- Produces embeddings of manageable dimension (768).
- Integrates easily with SentenceTransformers for semantic search.

5 Example: Semantic Similarity

We tested two queries:

1. Direct: "CRISPR gene editing in humans"
2. Paraphrased: "How scientists use CRISPR to modify human DNA for therapeutic purposes"

Semantic Similarity Score: 0.85

This confirms that SPECTER captures the same intent even when queries are phrased differently.

6 Implementation Notes

- We used `SentenceTransformer` with the SPECTER model.
- Query embeddings are cached to reduce repeated computation using `functools.lru_cache`.
- Cosine similarity was used to measure semantic similarity between query and document embeddings.

```python
# Example queries
queries = [
    "CRISPR gene editing in humans",  # Direct
    "How scientists use CRISPR to modify human DNA for therapeutic purposes"  # Paraphrased
]

# Encode both queries
vectors = embedding_model.encode(queries, convert_to_numpy=True)
v1, v2 = vectors[0], vectors[1]

# Compute cosine similarity
sim_score = round(cosine_similarity([v1], [v2])[0][0], 4)

print("\n 🔹 Semantic Similarity Score:")
print("•", sim_score)

if sim_score > 0.6:
    print("The queries share the same semantic intent!")
else:
    print("⚠ Semantic intent is weak – consider tuning models")
```

```
🔹 Semantic Similarity Score:
• 0.8312
The queries share the same semantic intent!
```

Figure 2: Cosine similarity between direct and paraphrased queries showing semantic match.

7 Conclusion

SPECTER is the most suitable embedding model for ScholarMind's semantic query extraction engine. It ensures relevant scientific documents are retrieved based on meaning rather than exact keyword matches.