Queryable Shared Reference Repository

Building an intelligent, privacy-preserving system for scientific research lab (VITEK)

Atyab Hakeem, Kishan Sathish Babu, Naga Kushal Ageeru, Pranav Kanth Anbarasan

Objective

Growing Volume

Research groups struggle to manage ever-increasing scientific literature.

Limited Search

Current reference managers lack intelligent, context-aware querying capabilities

Privacy Concerns

Cloud-based LLM’s raise data privacy issues and produce hallucinated outputs

Data Source

Sources & Formats:

- **Scientific Papers**:
  - PDF's and webpages
  - Variable layouts (journal / publisher differences)
- **Metadata Files:**
  - Formats: .bib (BibTeX) and reference documents
  - Enable citation and filtering

Documents	297
Avg Words	1,782
Vocabulary	39,144
Avg Tables	1
Avg Figures	7

Volume and Scale

Current Capacity

- 300 scientific paper

Scalability:

- Expandable to 10,000 papers

User Access:

- 1 - 3 concurrent users
- Max 10 lab members

Preprocessing

Goal:

- Standardize diverse research files into clean, metadata-linked text for querying and analysis

File Handling:

- PDF’s: PyMuPDF
- HTML: Beautiful Soup
- DOCX: Python-docx

Metadata Tagging:

- Parsed .ris via rispy
- Cleaned duplicated titles
- Fuzzy-matched filenames
- Saved as .csv for modeling

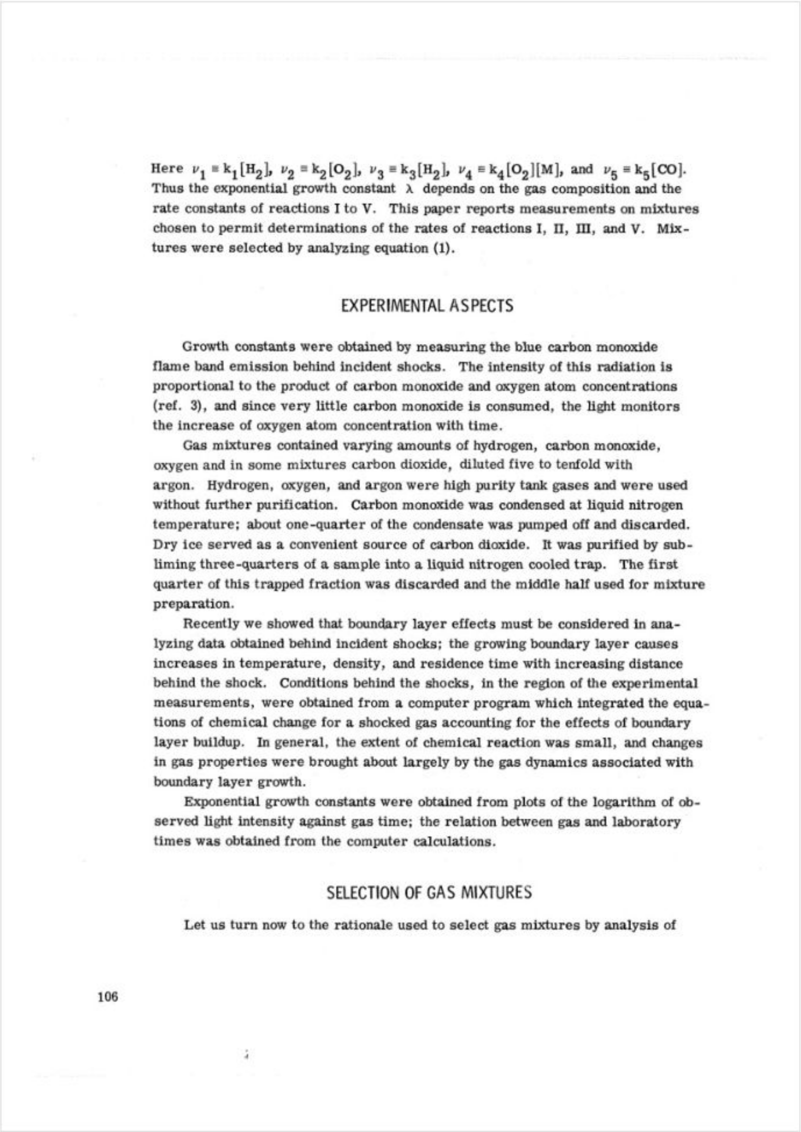
Preprocessing

SLM Based Updates

1. **IBM\Granite-Docling: 0.3B Params**

   Built for Image-Text to Text transcription. General purpose and can handle multiple formats and types of documents.

2. **Nougat (Neural Optical Understanding for Academic Documents)**
   a. • **Architecture**: Vision encoder-decoder model using Swin Transformer (visual encoder) + mBART (text decoder), processes PDF pages as images and generates structured markdown/LaTeX output
   b. • **Model Size**: ~250M parameters base model (~350M for large variant), relatively lightweight compared to modern LLMs while maintaining high accuracy on scientific documents



Here $\nu_1 = k_1[H_2]$, $\nu_2 = k_2[O_2]$, $\nu_3 = k_3[H_2]$, $\nu_4 = k_4[O_2][M]$, and $\nu_5 = k_5[CO]$. Thus the exponential growth constant λ depends on the gas composition and the rate constants of reactions I to V. This paper reports measurements on mixtures chosen to permit determinations of the rates of reactions I, II, III, and V. Mixtures were selected by analyzing equation (1).

EXPERIMENTAL ASPECTS

Growth constants were obtained by measuring the blue carbon monoxide flame band emission behind incident shocks. The intensity of this radiation is proportional to the product of carbon monoxide and oxygen atom concentrations (ref. 3), and since very little carbon monoxide is consumed, the light monitors the increase of oxygen atom concentration with time.

Gas mixtures contained varying amounts of hydrogen, carbon monoxide, oxygen and in some mixtures carbon dioxide, diluted five to tenfold with argon. Hydrogen, oxygen, and argon were high purity tank gases and were used without further purification. Carbon monoxide was condensed at liquid nitrogen temperature; about one-quarter of the condensate was pumped off and discarded. Dry ice served as a convenient source of carbon dioxide. It was purified by subliming three-quarters of a sample into a liquid nitrogen cooled trap. The first quarter of this trapped fraction was discarded and the middle half used for mixture preparation.

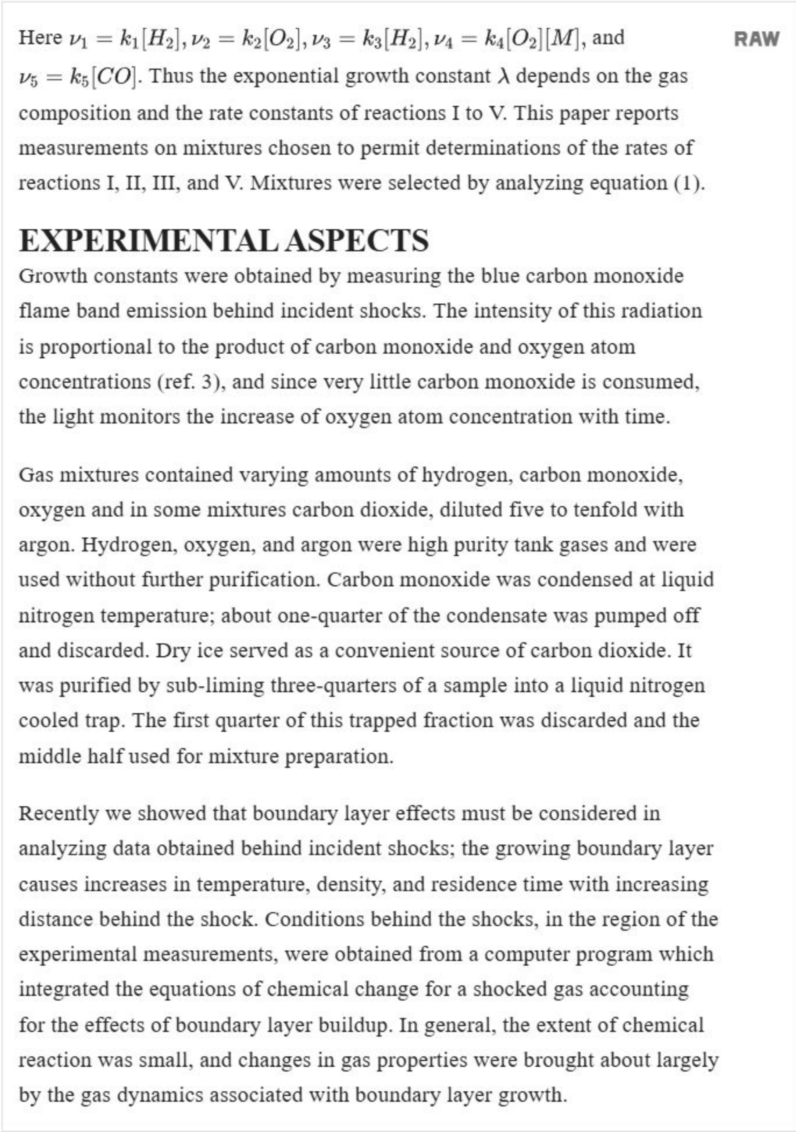
Recently we showed that boundary layer effects must be considered in analyzing data obtained behind incident shocks; the growing boundary layer causes increases in temperature, density, and residence time with increasing distance behind the shock. Conditions behind the shocks, in the region of the experimental measurements, were obtained from a computer program which integrated the equations of chemical change for a shocked gas accounting for the effects of boundary layer buildup. In general, the extent of chemical reaction was small, and changes in gas properties were brought about largely by the gas dynamics associated with boundary layer growth.

Exponential growth constants were obtained from plots of the logarithm of observed light intensity against gas time; the relation between gas and laboratory times was obtained from the computer calculations.

SELECTION OF GAS MIXTURES

Let us turn now to the rationale used to select gas mixtures by analysis of

106



RAW

Here $\nu_1 = k_1[H_2], \nu_2 = k_2[O_2], \nu_3 = k_3[H_2], \nu_4 = k_4[O_2][M]$, and $\nu_5 = k_5[CO]$. Thus the exponential growth constant λ depends on the gas composition and the rate constants of reactions I to V. This paper reports measurements on mixtures chosen to permit determinations of the rates of reactions I, II, III, and V. Mixtures were selected by analyzing equation (1).

EXPERIMENTAL ASPECTS

Growth constants were obtained by measuring the blue carbon monoxide flame band emission behind incident shocks. The intensity of this radiation is proportional to the product of carbon monoxide and oxygen atom concentrations (ref. 3), and since very little carbon monoxide is consumed, the light monitors the increase of oxygen atom concentration with time.

Gas mixtures contained varying amounts of hydrogen, carbon monoxide, oxygen and in some mixtures carbon dioxide, diluted five to tenfold with argon. Hydrogen, oxygen, and argon were high purity tank gases and were used without further purification. Carbon monoxide was condensed at liquid nitrogen temperature; about one-quarter of the condensate was pumped off and discarded. Dry ice served as a convenient source of carbon dioxide. It was purified by sub-liming three-quarters of a sample into a liquid nitrogen cooled trap. The first quarter of this trapped fraction was discarded and the middle half used for mixture preparation.

Recently we showed that boundary layer effects must be considered in analyzing data obtained behind incident shocks; the growing boundary layer causes increases in temperature, density, and residence time with increasing distance behind the shock. Conditions behind the shocks, in the region of the experimental measurements, were obtained from a computer program which integrated the equations of chemical change for a shocked gas accounting for the effects of boundary layer buildup. In general, the extent of chemical reaction was small, and changes in gas properties were brought about largely by the gas dynamics associated with boundary layer growth.

Technical Jargon Prevalence

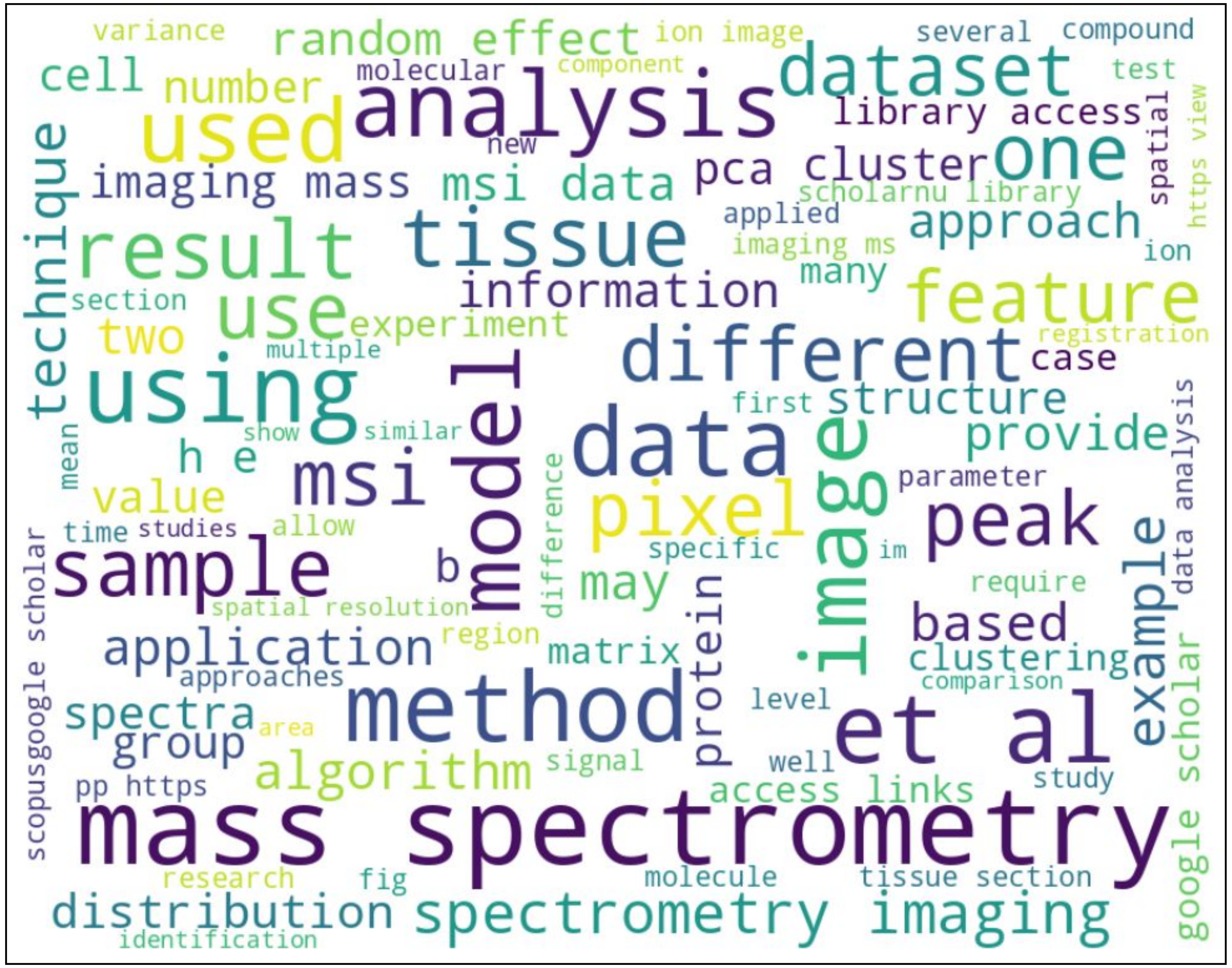


Figure 1: Jargon prevalence in corpus

Dataset Overview

- Computational imaging biology papers
- Formats: PDF, HTML, DOCX
- Metadata: Authors, publication dates, URL

Document Size Distribution Right Skewed

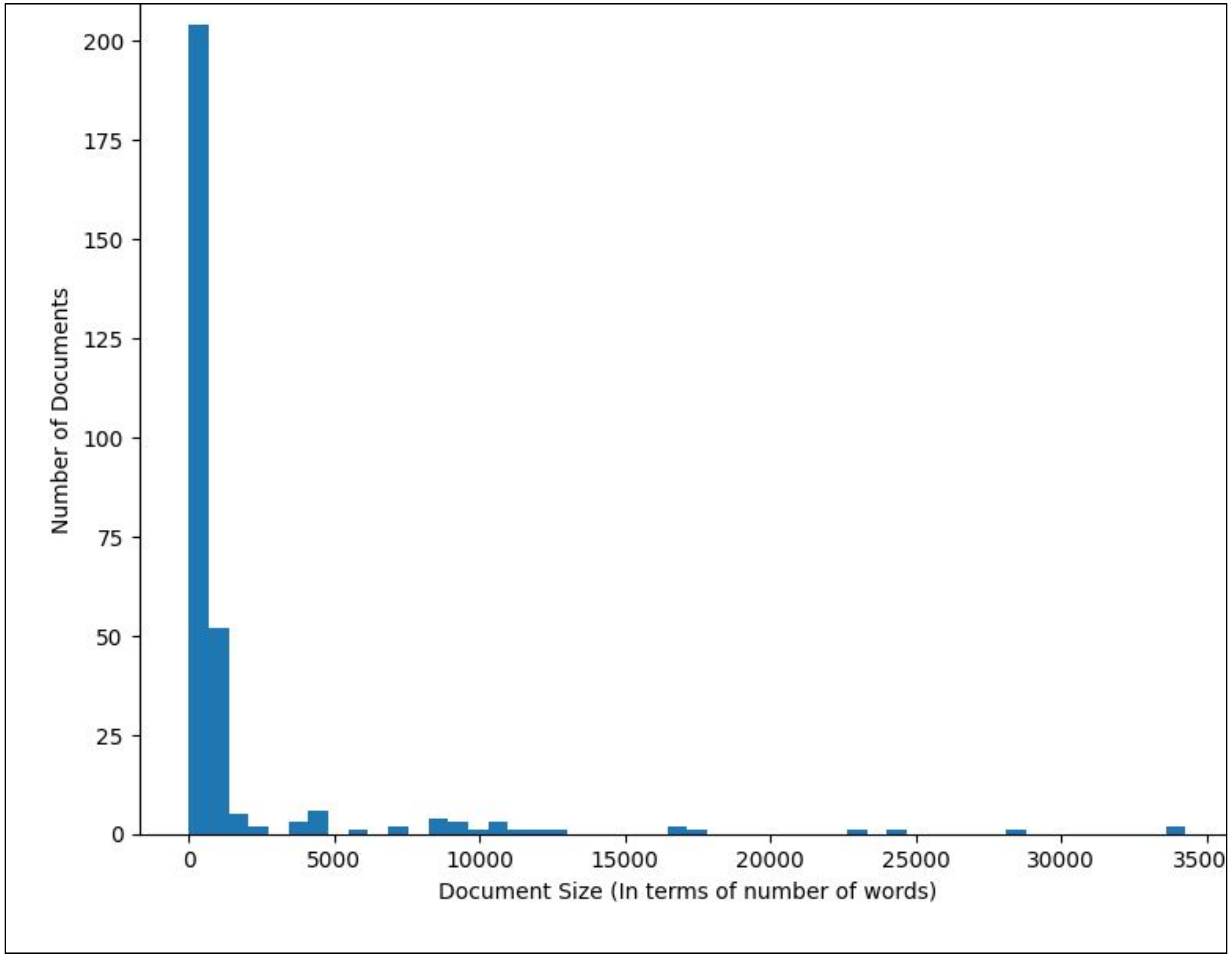


Figure 2: Document size distribution

Model Tokenizers Cause Minimal Splitting

Model Compatibility

- Token-to-word ratio: **~1.5**
- Technical jargon preserved in vocabulary
- Minimal Byte Pair Encoding splitting
- Decent performance without fine-tuning

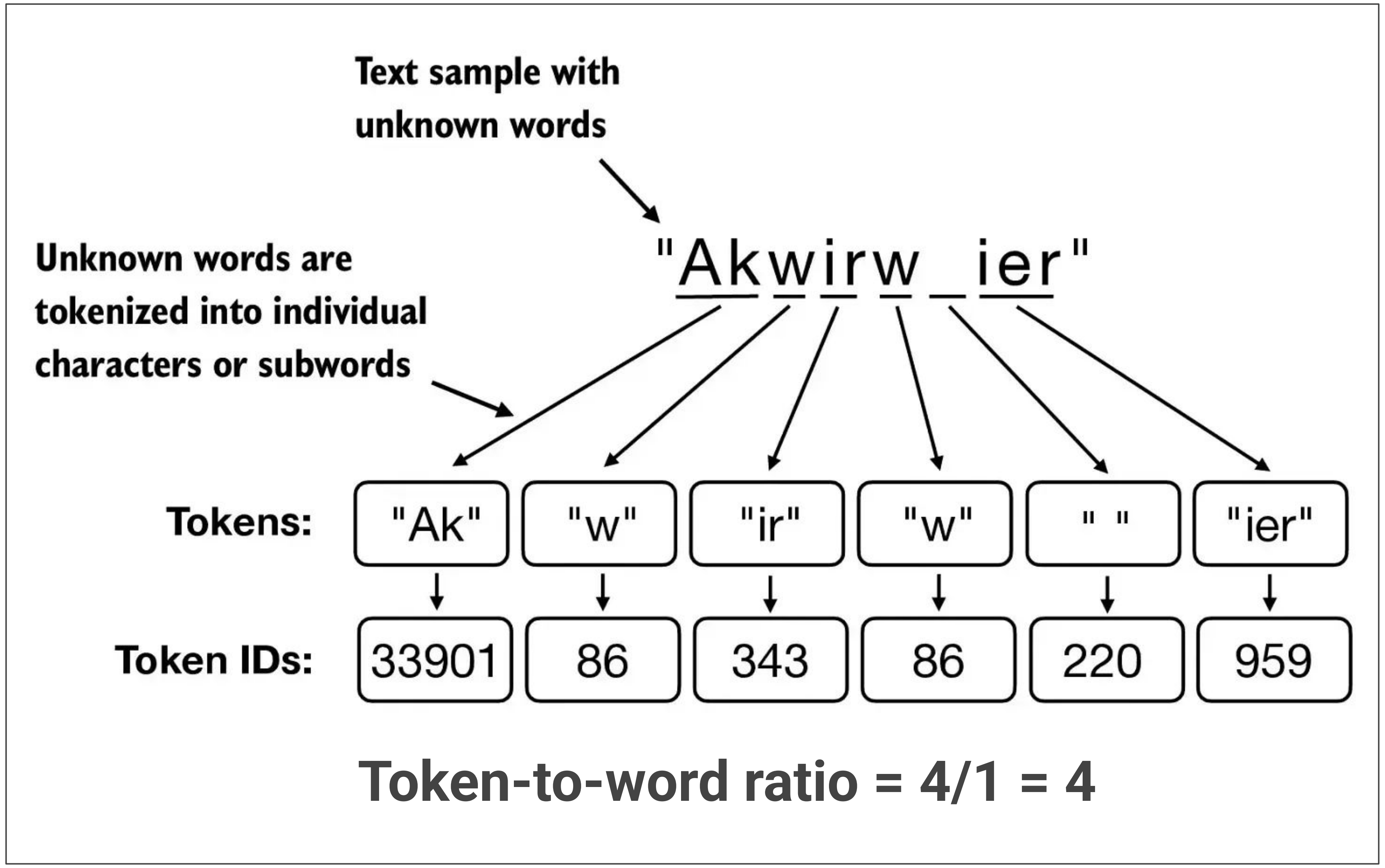


Figure 3: Byte Pair Tokenization Visualization [Source: Vizuara, "Understanding Byte Pair Encoding"]

Chunking & Embedding

Chunking

- Splitting text into smaller segments
- Improves
  - Precision
  - Information captured
  - Response Quality
- Recursive splitter uses priority

```
The degree to which the returns for
performance are superlinear.
```

Character Splitter; Chunk size = 25; Overlap = 0

```
The degree to which the returns for
performance are superlinear.
```

Character Splitter; Chunk size = 10; Overlap = 3

Embedding

- Conversion to vectors
- Similarity search to fetch similar text

```
The degree to which the r ➔ [[0.5..]..]
eturns for performance ar ➔ [[0.3..]..]
e superlinear. ➔ [[0.6..]..]
```

Source: chunkviz.up.railway.app

Chunking Evaluation Synthetic Data

- Model: Llama 3.3 70B
- Issues: Self-containment, Formatting, Out of vocabulary tokens
- ~800 Questions

Question	What is the typical outcome of a MALDI-imaging study?
Excerpt	A typical MALDI-imaging study results in a set of ions of interest…
Source	Alexandrov - 2012 - MALDI imaging mass spectrometry…

Hit Rate

Hit Rate = $1/n \sum_{i=1}^{n} \mathbb{1}(\text{recall}_i > \tau)$

$\text{recall}_i = |D_i \cap G_i| / |G_i|$

n = number of queries

D_i = retrieved documents for query i

G_i = ground truth documents for query i

τ = threshold

$\mathbb{1}(\cdot)$ = indicator function

Query: What is the typical outcome of a MALDI-imaging study?

Ground Truth:
A typical MALDI-imaging study results in a set of ions of interest

Retrieved Documents:
MALDI-imaging study results in a set of ions of interest. confocal microscopy imaging techniques. experimental outcomes vary significantly.

$\tau = 0.5$	$\tau = 0.9$
$\lvert D_i \cap G_i \rvert = 11$	$\lvert D_i \cap G_i \rvert = 11$
$\lvert G_i \rvert = 13$	$\lvert G_i \rvert = 13$
Recall = 0.846	Recall = 0.846
Hit Rate = 1	**Hit Rate = 0**

Mean Reciprocal Rank

$MRR = 1/n \sum_{i=1}^{n} 1/rank_i$

$rank_i$ = rank of the first relevant document

relevant document = $recall_i > \tau$

$recall_i = |D_i \cap G_i| / |G_i|$

τ = threshold

Query: What is the typical outcome of a MALDI-imaging study?

Ground Truth:
A typical MALDI-imaging study results in a set of ions of interest

Retrieved Documents:

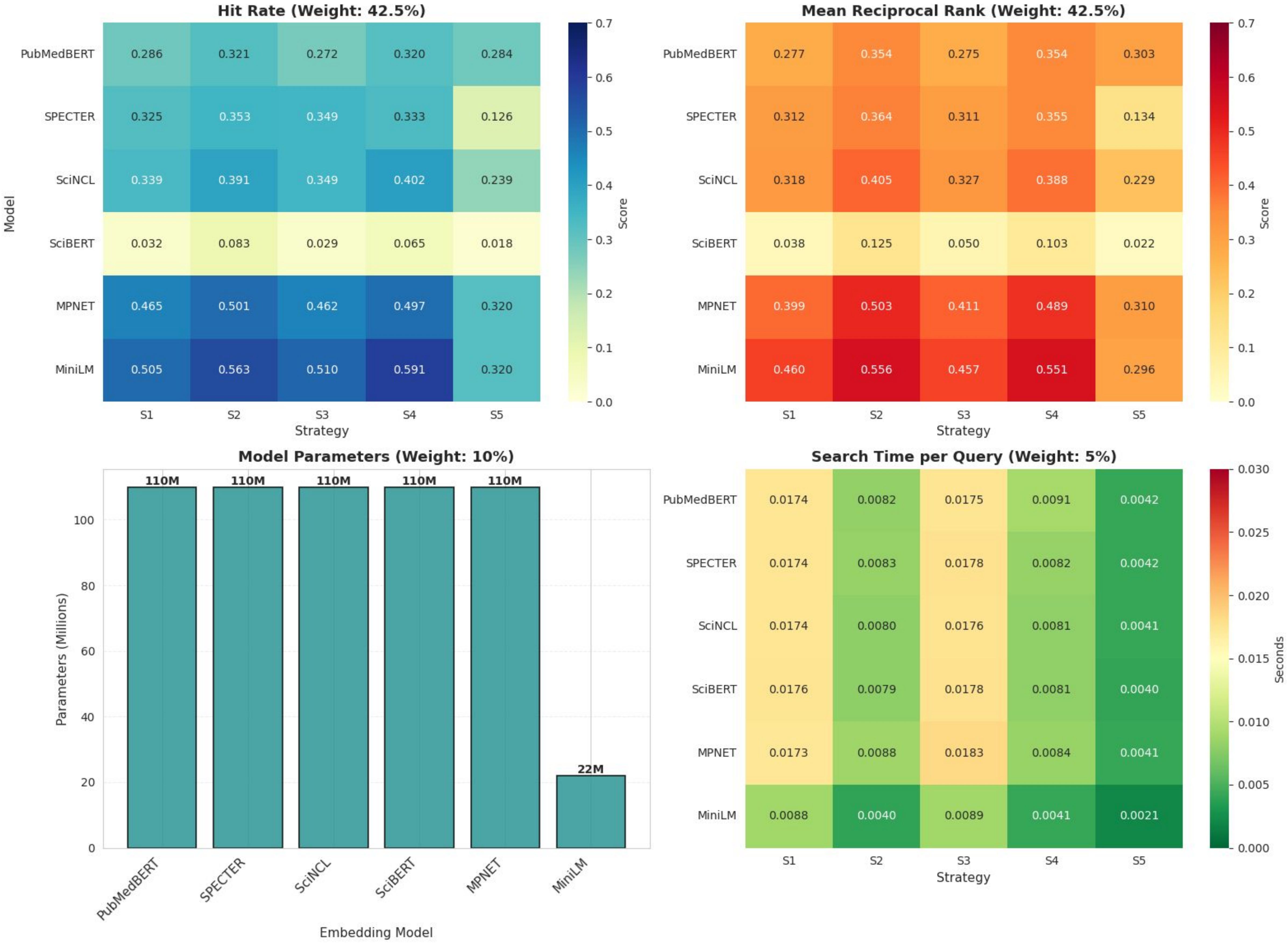
1. confocal microscopy imaging techniques.
2. MALDI-imaging study results in a set of ions of interest
3. experimental outcomes vary significantly.

$\tau = 0.5$	$\tau = 0.9$
$\lvert D_i \cap G_i \rvert = 11$; $\lvert G_i \rvert = 13$	$\lvert D_i \cap G_i \rvert = 11$; $\lvert G_i \rvert = 13$
Rank = 2	Rank = 2
Recall = 0.846	Recall = 0.846
MRR = 0.5	**MRR = 0**

Chunking Evaluation Metrics Key Takeaways

- Higher-level metrics for text overlap (less granular)
- Use Recall as base (ground truth focussed)
- Normalized length effect
- Hit Rate evaluates retrieved content
- Mean Reciprocal Rank evaluates ranking

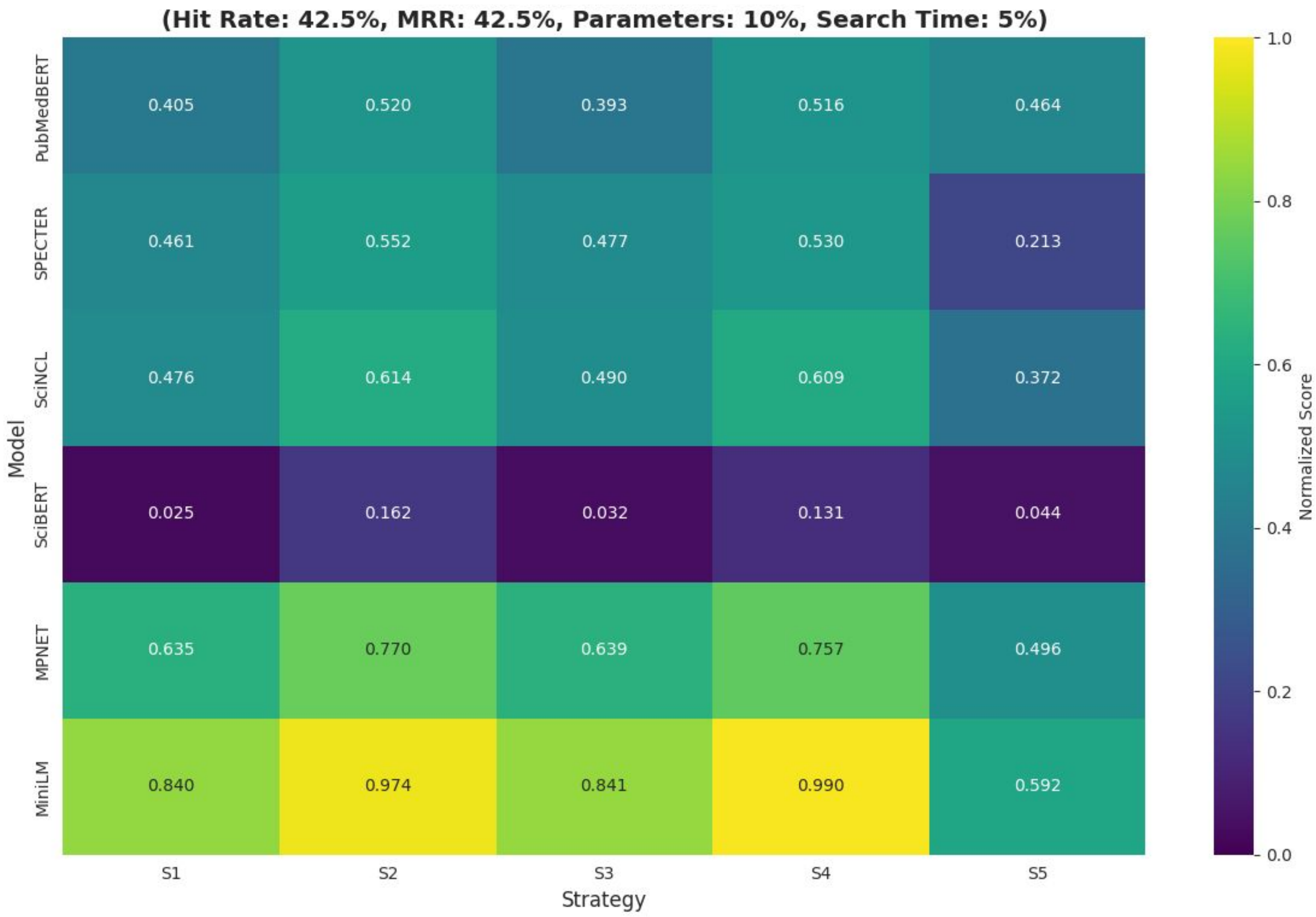
Parsing + MiniLM Has Highest Quality Metrics & Lowest Cost



Strategy	Type	Chunk Size	Chunk Overlap
S1	Recursive	256	0
S2	Recursive	512	0
S3	Recursive	256	25
S4	Recursive	512	50
S5	Semantic	**Breakpoint**: Percentile	**Threshold**: 90

Figure 3: Comparison of the quality metrics & performance metrics across embedding models & chunking strategies

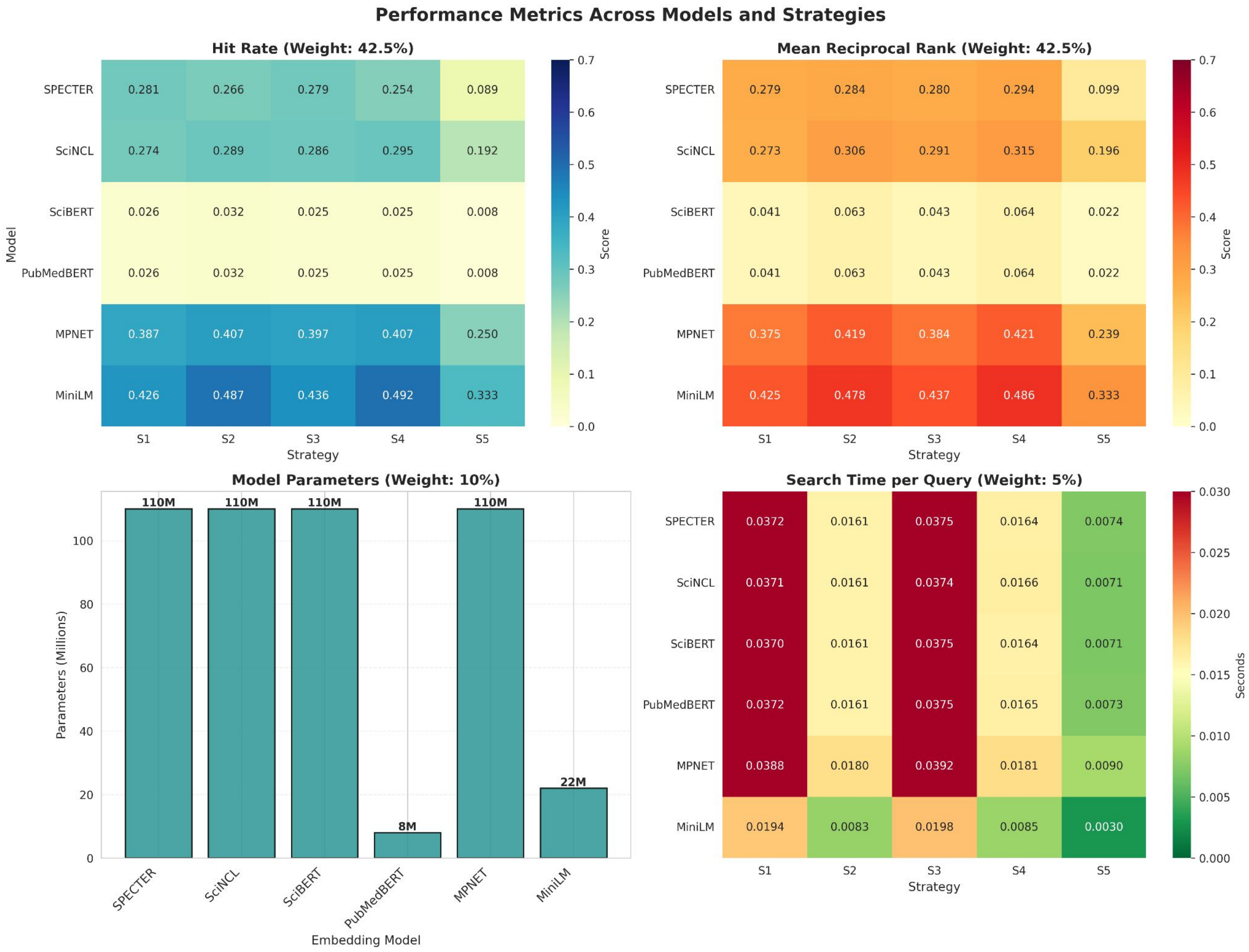
Parsing + MiniLM + Strategy 4 Has Optimal Quality-Cost Tradeoff



Strategy	Type	Chunk Size	Chunk Overlap
S1	Recursive	256	0
S2	Recursive	512	0
S3	Recursive	256	25
S4	Recursive	512	50
S5	Semantic	**Breakpoint**: Percentile	**Threshold**: 90

Figure 4: Weighted normalized scores across embedding models & chunking strategies

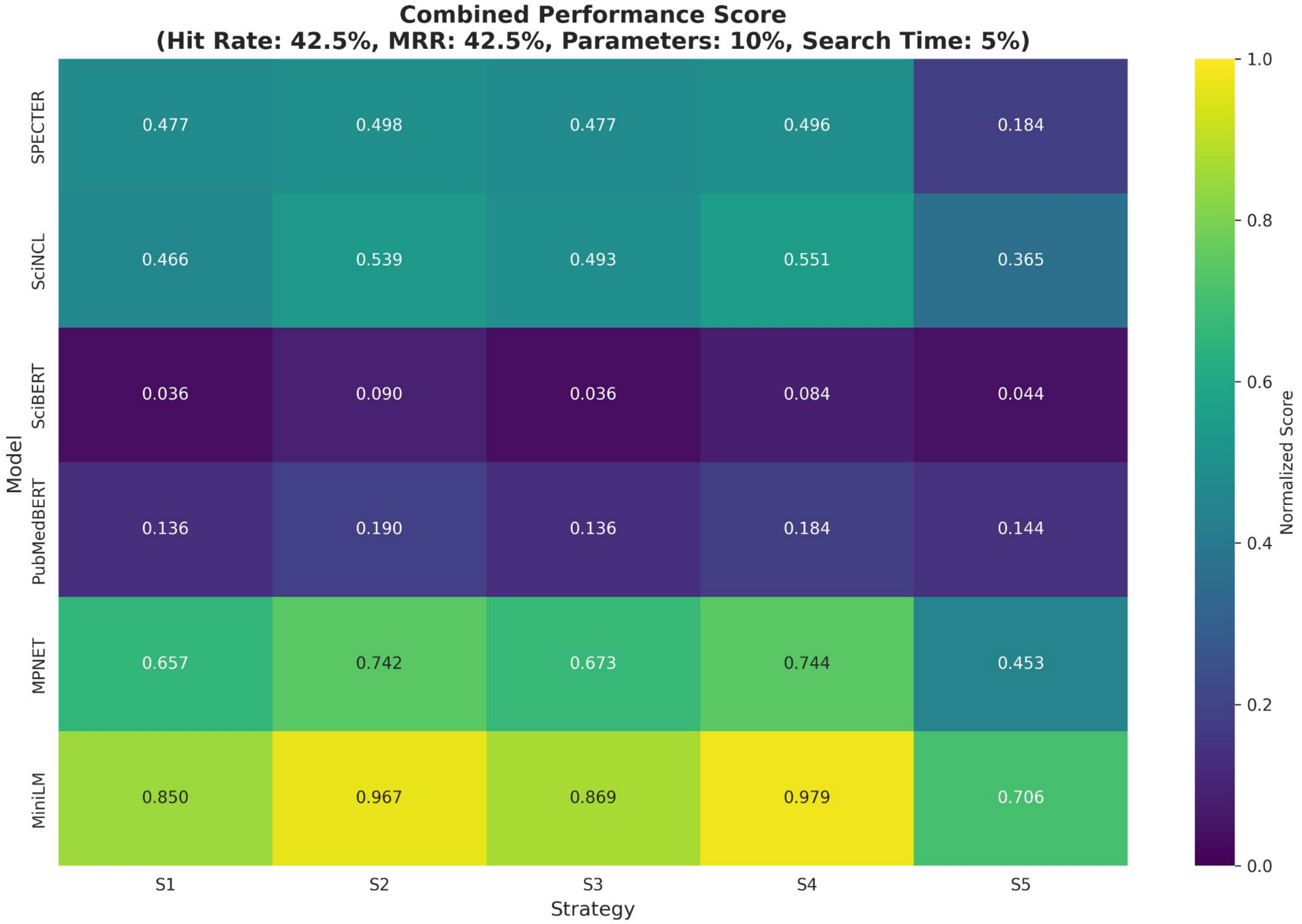
Docling - MiniLM Has Highest Quality Metrics & Lowest Cost



Strategy	**Type**	**Chunk Size**	**Chunk Overlap**
S1	Recursive	256	0
S2	Recursive	512	0
S3	Recursive	256	25
S4	Recursive	512	50
S5	Semantic	**Breakpoint**: Percentile	**Threshold**: 90

Figure 5: Comparison of the quality metrics & performance metrics across embedding models & chunking strategies

Docling - MiniLM + Strategy 4 Has Optimal Quality-Cost Tradeoff



Strategy	Type	Chunk Size	Chunk Overlap
S1	Recursive	256	0
S2	Recursive	512	0
S3	Recursive	256	25
S4	Recursive	512	50
S5	Semantic	**Breakpoint**: Percentile	**Threshold**: 90

Figure 6: Weighted normalized scores across embedding models & chunking strategies

Chunking, Embedding - Implications & Solution

- Strategy - 4 (512 tokens, recursive splitting, 50 chunk overlap) ➔ Context-completeness and continuity favored
- ~59.1% Hit Rate ➔ 59.1% of queries fetched relevant content
- ~55.1% MRR ➔ Relevant content between rank 1 & 2
- Actual data needed for better evaluation
- Filters to improve performance

Hallucination

Retrieval Based Queries

- Output not grounded in retrieved documents

Query: Where did Mr. Bob live?
Documents: Mr. Bob loves to travel. He lives in USA.

Hallucination: Mr. Bob lives in Massachusetts.

Fact Based Queries

- Information the model hasn't seen (low confidence.)

Query: What's Adam Kalai's birthday?

Hallucination: Adam Kalai's birthday is March 7th.

Brainstorming Queries

- No shackles

Query: Can moss predict internet outages?

????: Track their growth rate pattern variations.

Hallucination Evaluation Synthetic Data

- Model: Llama 3.3 70B
- Issues: Formatting, Out of vocabulary tokens
- ~1800 Questions

Question	What is the color of the nucleus in H&E-stained slides?
Excerpt	The hematoxylin\u2013eosin (H&E) stain has stood the test of time…
Correct Answer	Basophilic
Wrong Answer	Eosinophilic
Source	Chan - 2014 - The Wonderful Colors of the Hematoxylin…

MiniCheck RoBERTa Has Largest Efficiency

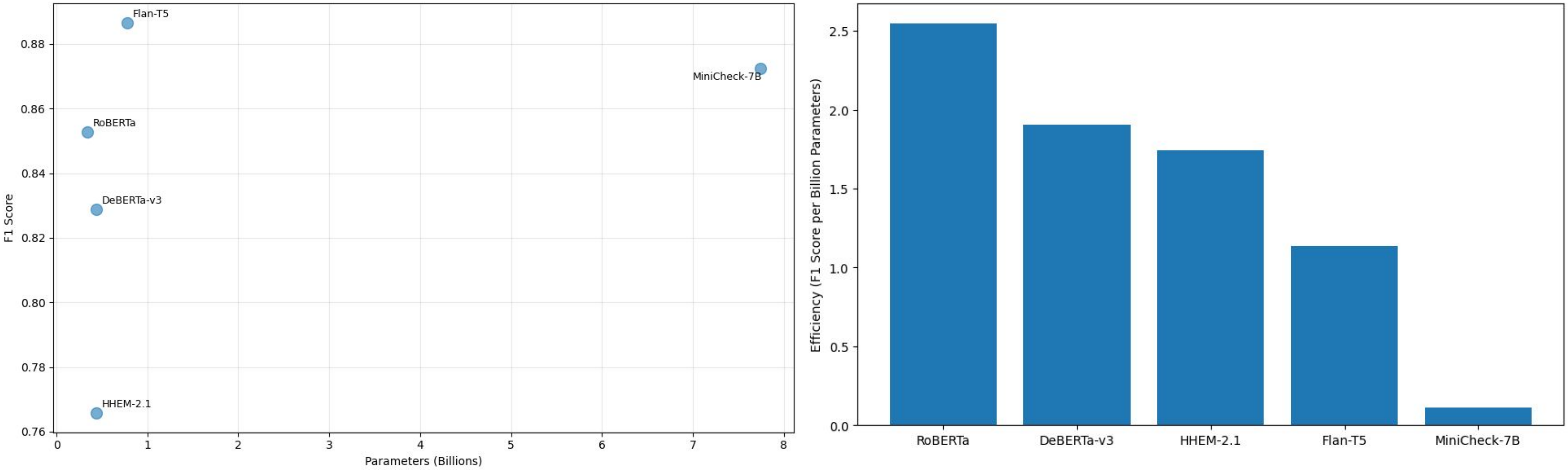


Figure 7: Efficiency comparison of the hallucination detection models

Optimal F-1 Score 85.3% At Threshold 26.3%

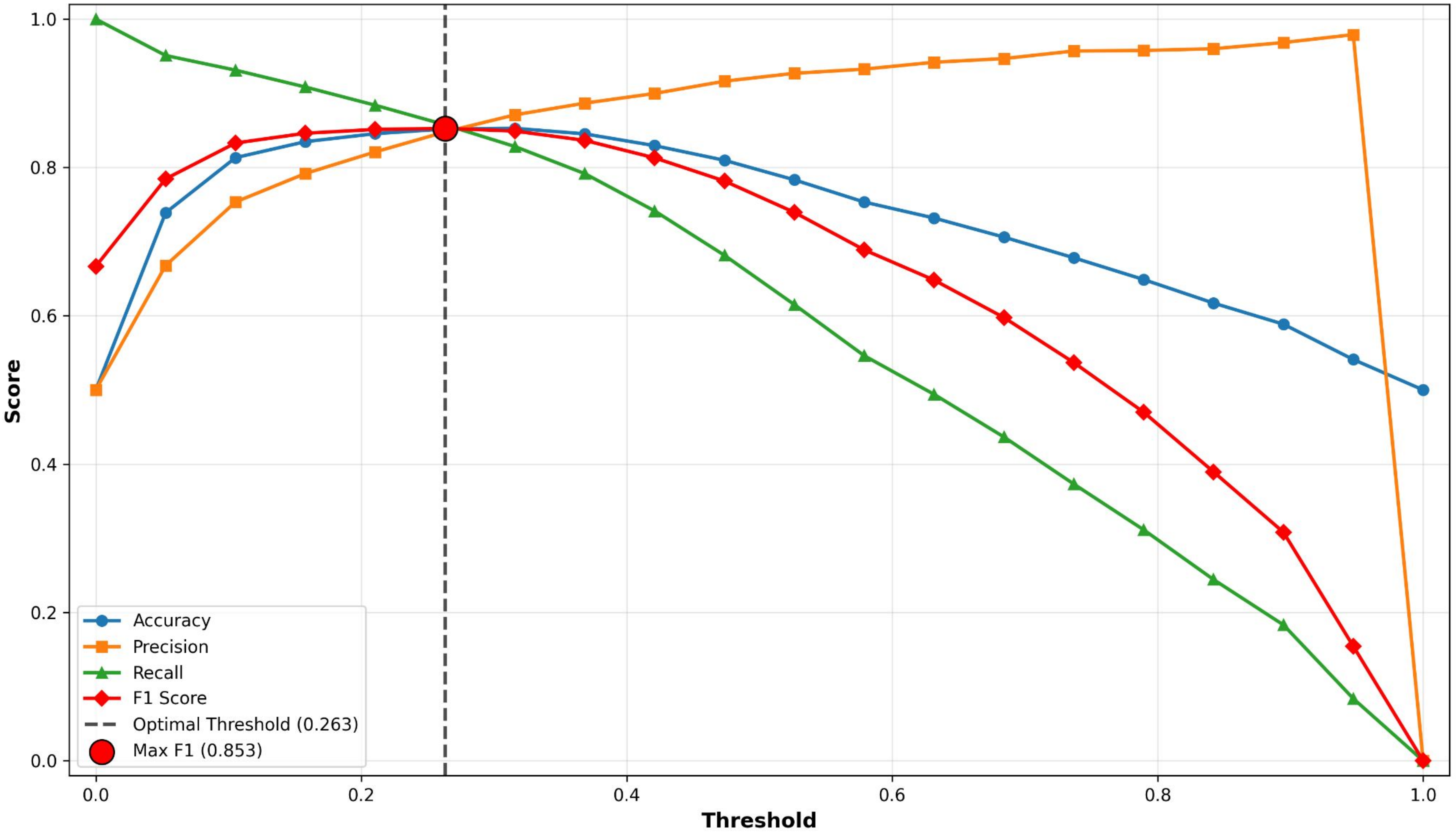


Figure 8: Confidence threshold gridsearch across accuracy, precision, recall, and f1 score on RoBERTa model

Three-Tiered Hallucination Reporting

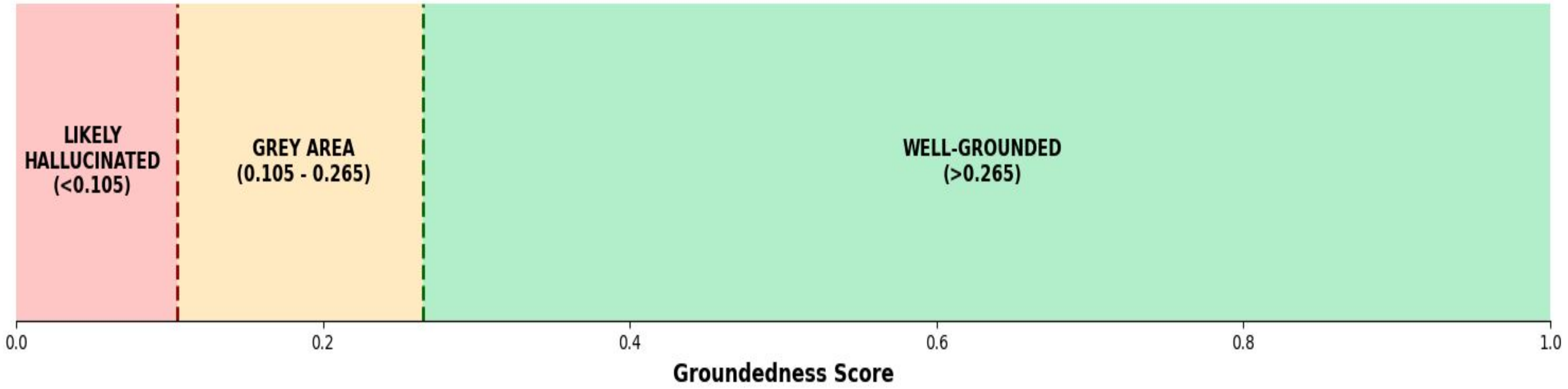


Figure 9: Hallucination detection and reporting decision framework

False Negatives - RoBERTa Struggles With Numbers

Question	What is the correlation between any two adjacent time points in the first case of the repeated measures ANOVA model?
Excerpt	In the first case of the repeated measures ANOVA model, we assumed $\rho = 0$…
Claim	0.7
Answer	0.7

False Positives - RoBERTa Struggles Similar Words

Question	What could be removed by aligning each spectrum to the mean spectrum and re-calibrating the m/z positions via the internal calibrants?
Excerpt	These m/z shifts could be removed by aligning each spectrum to the mean spectrum…
Claim	baseline shifts
Answer	m/z shifts

Hallucination Detection Limitations & Solutions

Model struggles

- Numbers
- Similar text
- Large text
- Formatting issues

Solutions

- Numerical features layer
- Fine-tune
- Small chunk size, Simpler queries
- Stringent processing

Technical Implementation & Extensibility

Document Loader

Imports text, PDF, and Markdown files whilst extracting metadata for processing.

Recursive Chunker

Splits text into 512-character chunks using intelligent separators.

SciNCL Embeddings

Creates domain-specific embeddings with sentence-transformers.

ChromaDB Storage

Stores and manages vectors with metadata filtering capabilities.

Key Use Cases

- Semantic search for research papers
- Building knowledge bases
- Finding similar documents
- Research assistant tools for large datasets

Future Roadmap

- Fine-tune embeddings
- Hybrid search
- Advanced metadata filtering
- Scalable distributed processing

Two-Stage Retrieval Architecture

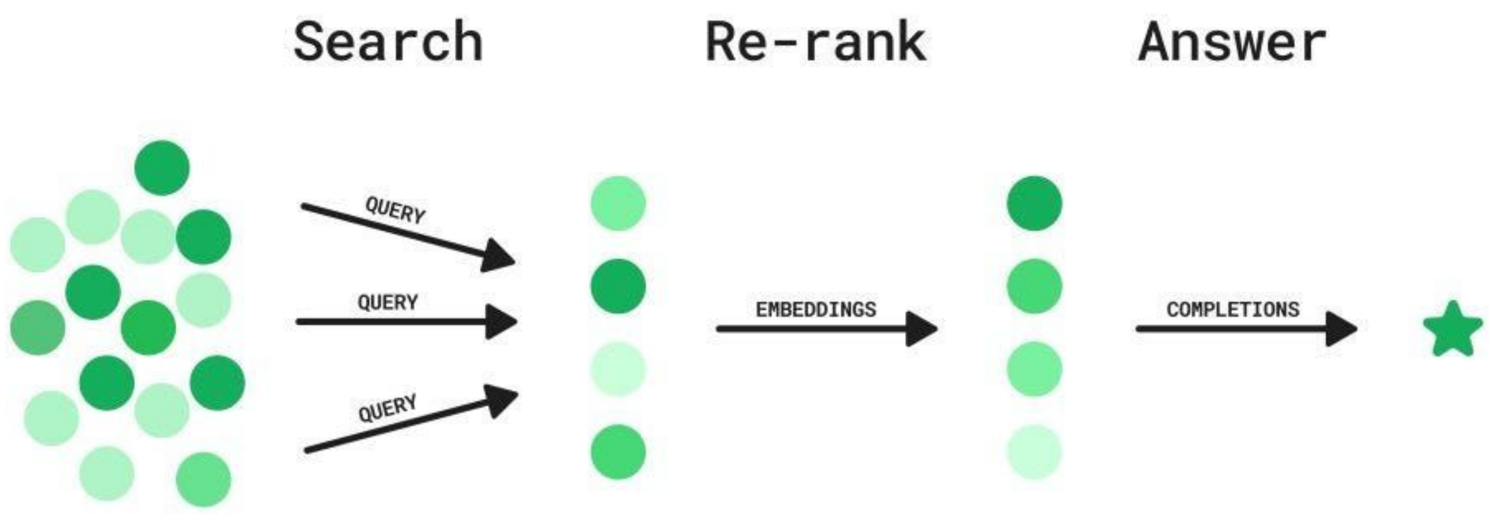
Stage 1: Embedding Retrieval

Converts queries and documents into vector format, finding the most relevant results based on similarity in meaning.

Stage 2: Reranker Refinement

Uses a transformer model to re-check top results for better accuracy, focusing on context and deeper meaning.

Domain-Specific Embedding Fine-Tuning



Custom Dataset Training

Trains on lab data to better understand technical terms and domain-specific language.

Pipeline Evaluation & Optimisation

Evaluation Metrics: Measure precision and recall to assess how well the system finds and ranks results whilst tracking speed and quality.

Next Steps: Test the full system with real lab data, compare against baseline models, and use results to fine-tune performance.

Linear Adapter Fine-Tuning Overview

Goal: Adapt a pre-trained embedding model to a specific domain without retraining the entire network.

Base Model Setup

Start with a pre-trained sentence embedding model (like all-MiniLM or SciNCL).

Freeze the original model weights to retain general language understanding.

Add Linear Adapter Layer

Insert a lightweight linear layer between the encoder and output. This layer learns domain-specific patterns whilst keeping the model efficient.

Evaluation & Validation

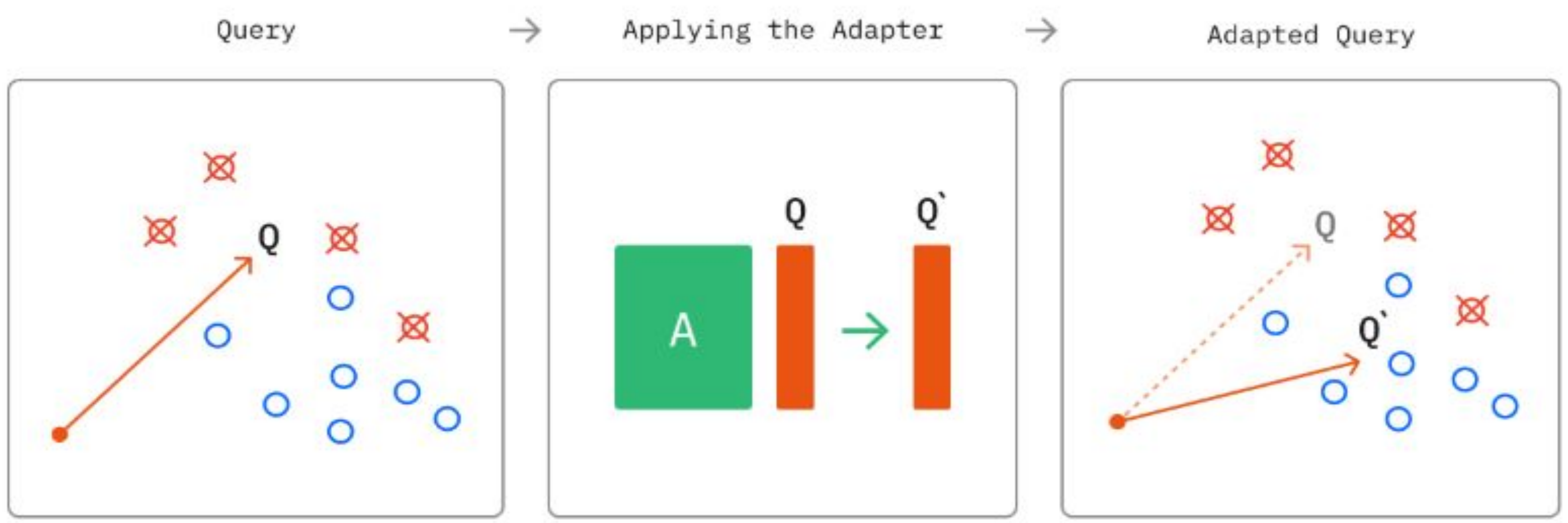
Measure similarity accuracy using metrics like cosine similarity, Precision@k, and Recall. Adjust learning rate and adapter size for optimal balance.

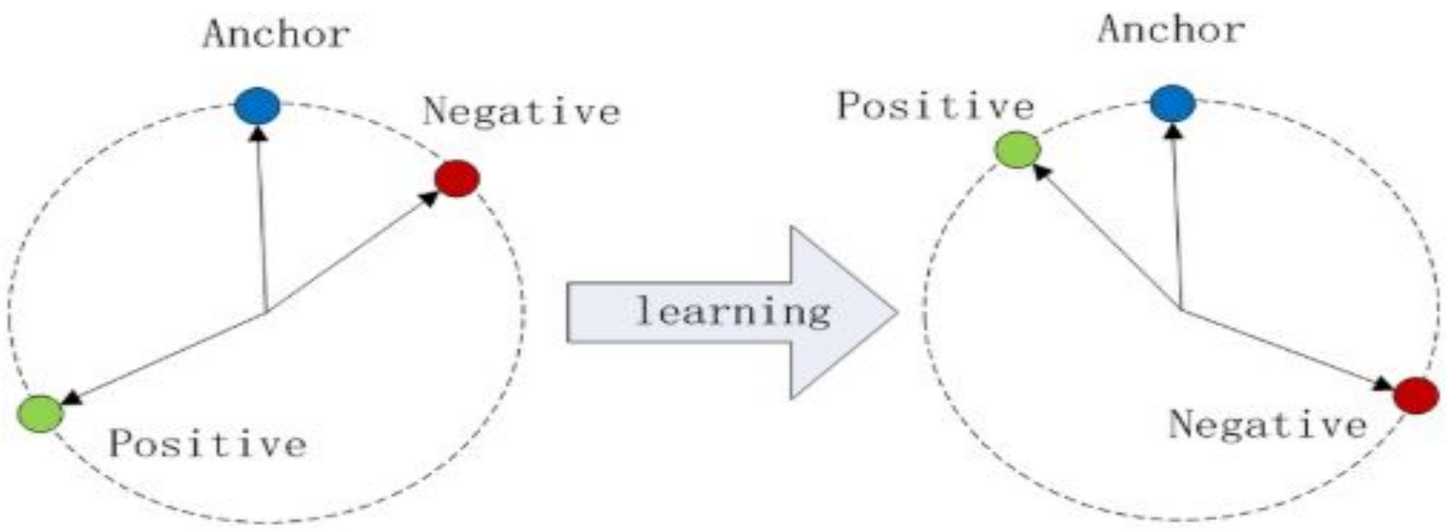
Training Phase

Fine-tune only the adapter using your custom dataset. Use contrastive or triplet loss to align similar sentences closer in vector space.

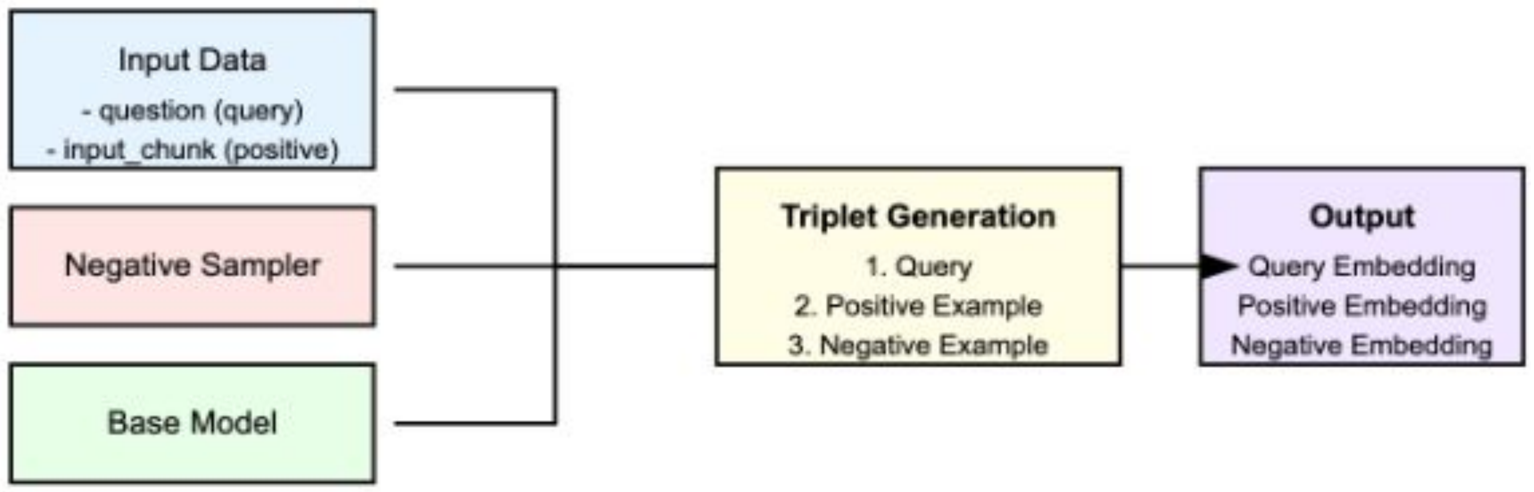
Deployment

Merge adapter outputs with the base model to generate enhanced embeddings. Integrate into the retrieval or RAG pipeline for improved semantic search performance.





$$L = max(d(A,P) - d(A,N) + margin, 0)$$



Milestones Met

~~Phase 1: Midphase~~

~~Week 1-2~~

~~Data & Processing~~

- ~~Extraction & Cleaning~~
- ~~Strategic Chunking~~
- ~~Metadata tagging~~
- ~~Embedding generation and vector store~~

Phase 1: Endphase

Week 3-5

Integration

- Generation Model
- Retrieval & Reranker
- Agent Architecture
- ~~Hallucination Detection~~

Phase 2: Midphase

Week 6-7

Evaluation

- Test Set Generation
- Evaluate System
- Hallucination Mitigation
- Optimizations

Phase 2: Endphase

Week 8-10

App & Deployment

- CI/CD
- Slack App
- Documentation

Next Steps

- ➢ Generation Model
- ➢ Retrieval & Reranker
- ➢ Agent Architecture

Phase 1: Endphase

Week
3-5

Integration

Thank You