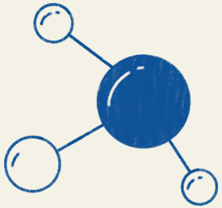
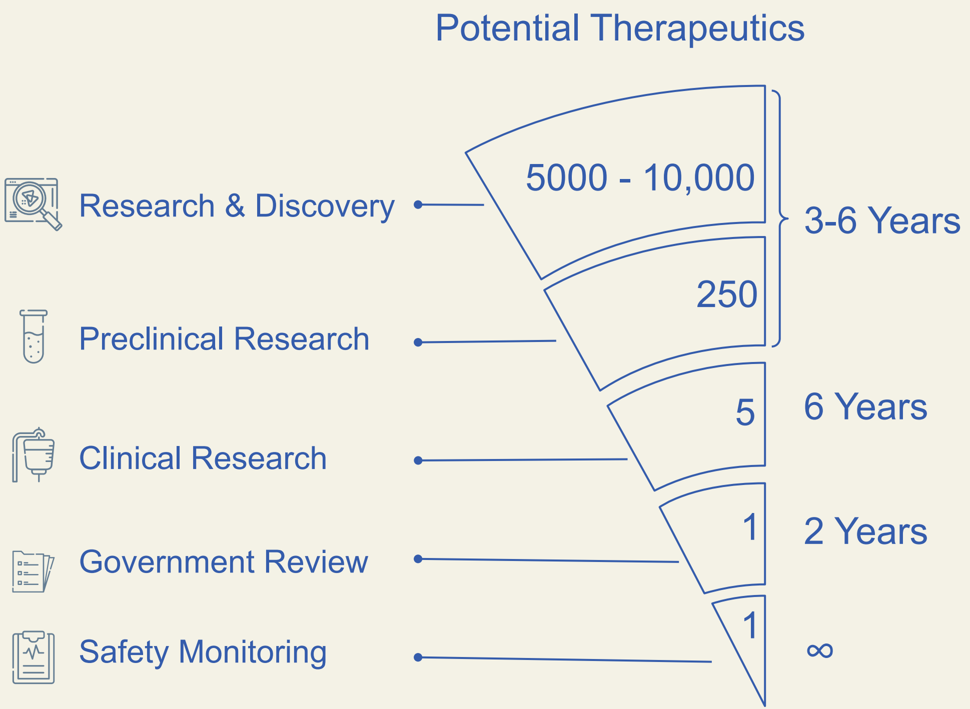
Therapeutic Accelerator GPT (TA.GPT)



Nicholas Lee, Vani Vijayakumar, Nic Brathwaite

Our mission is to use artificial intelligence to accelerate pharmaceutical development by reducing time spent on literature reviews more time in the lab

Finding an Edge



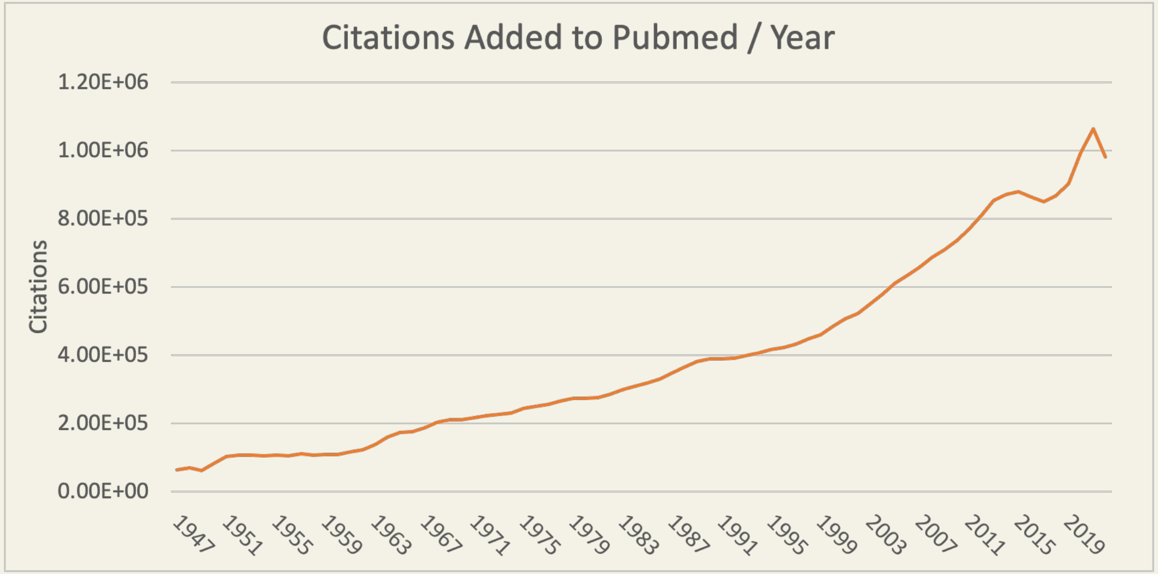
Users of the Product:

- Researchers
- Leadership Teams
- Clinicians

~ 12% Approval rate after clinical testing (Hay et al., 2014)

~ $2.6 B Average Development Cost over 7 - 10 years

Staying On Top of Publications



More than biomedical 1 million papers, added to PubMed database each year *(Landhuis, E. (2016))*

Researchers spend ~ **10 Hours Per month** Reading *(Van Noorden, 2014)*

In Biomedical Research **irreproducibility ranges from 75% - 90%** and **85% of is wasted** at-large (*Six Factors Affecting Reproducibility in Life Science Research and How to Handle Them*, n.d.)

Competition

- OpenAI
- IBM
- Ai2
- Microsoft Biogpt
- Google Health
- PaperGPT — A smarter way to search and discover biomedical papers

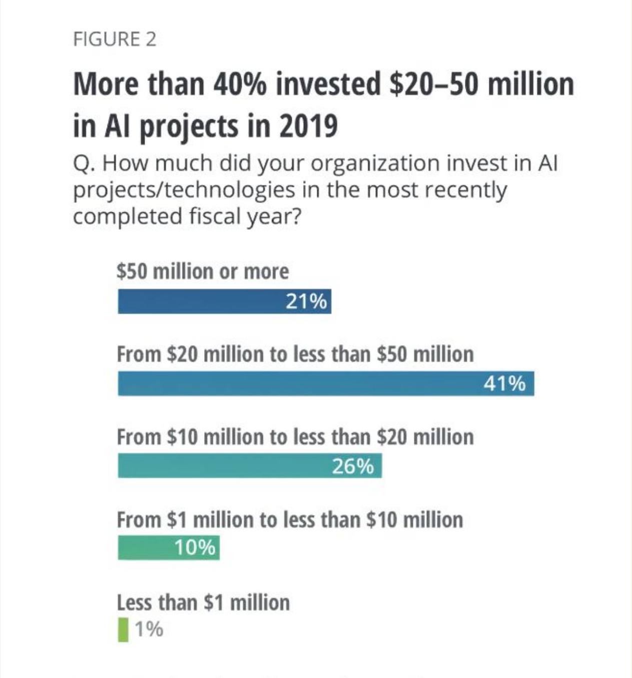
Our Key Differentiators

1. Summarizations and Q&A trained on full medical corpus
2. Citations of relevant articles
3. Time Saver

Biotherapeutics AI Market Space

- **USD 15.4 billion in 2022** was the global artificial intelligence in healthcare market size value
    - *(Artificial Intelligence In Healthcare Market Size Report, 2030. (July, 2023)*

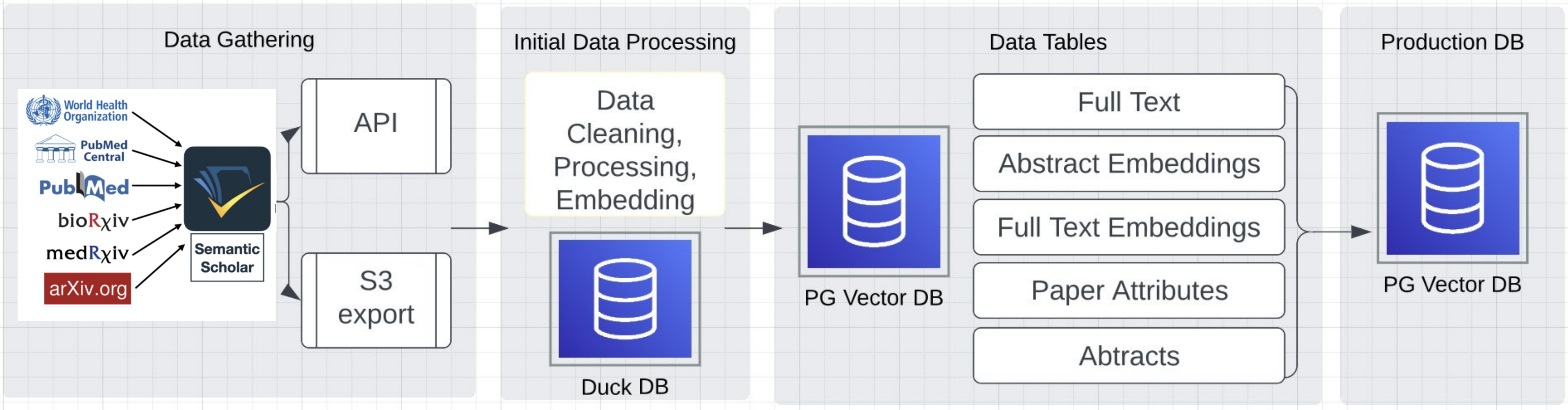
- The McKinsey Global Institute estimates **$100 billion in value** could be generated annually improving the efficiency of research and clinical trials
    - *(How Big Data Can Revolutionize Pharmaceutical R&D | McKinsey, n.d.)*



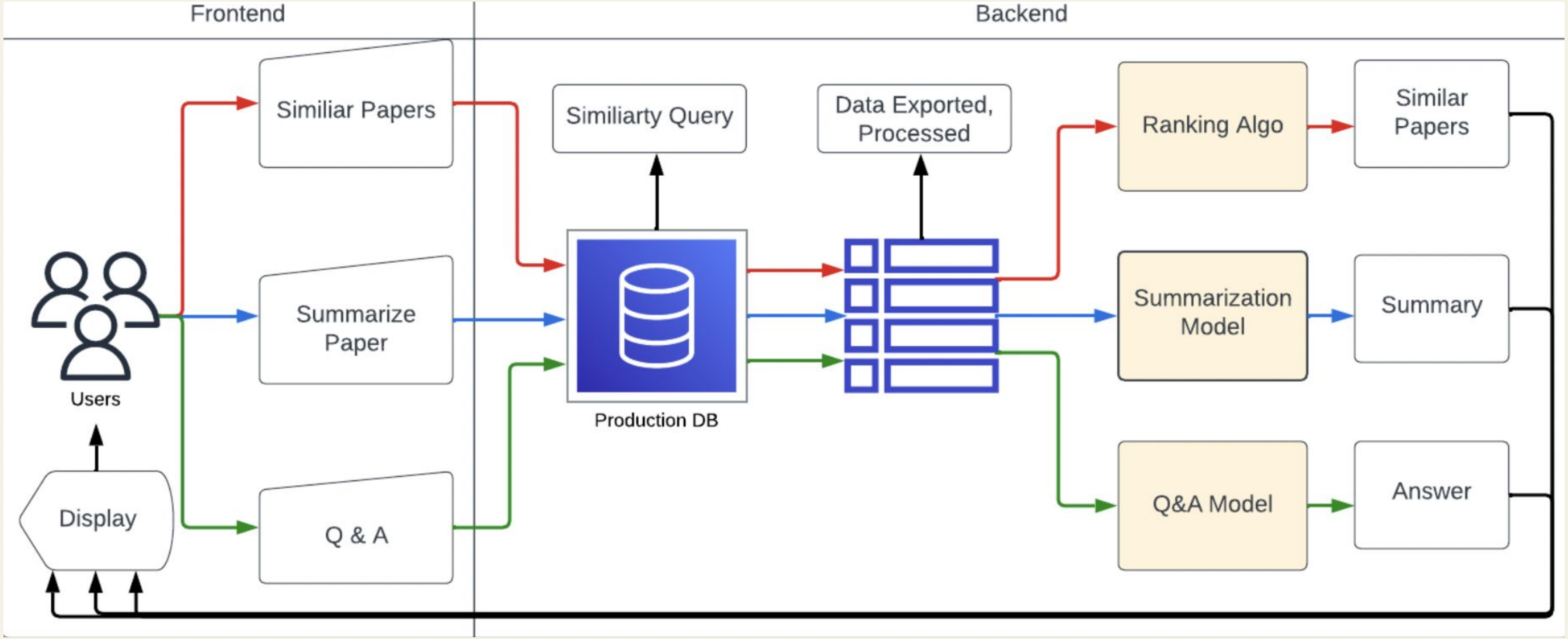
Deloitte Insights (July, 2023)

The Data and App Architecture

Data Set & Pipeline



App Workflow



Demo

The Models

Model Evaluation

Metrics: F1 (precision), ROUGE (coherence), Human verified accuracy + relevance

Limitations: High cost of generating human written summaries and answers. Using abstracts as substitute for human written summary allowed us to benchmark models in a low-cost and standardized way.

Similarity - F1

Model	F1 Score
Random	32.5
Sent-Bert	67.5
SciBert	59.6
ELMO	69.0
Specter	**80.0**

Summarization - ROUGE + Human Verified

ROUGE score	Abstract Summarization (LangChain)	Full Text Summarization (LangChain)
F1	.24	.17
Recall	.17	.13
Precision	.53	.32

- Lower ROUGE scores are expected as information is lost between an abstract & a summary (low recall).

- Using human generated summaries as a benchmark for model evaluation will increase the scores.

Q&A - Human Verified

- High cost of getting human generated responses

- Lack of comparable proxy

Model 1: Similarity

The Model:

- Pull the most similar papers based on paper ID or Text string.
- Allows users to quickly expand knowledge base of papers in seconds, without subject matter experts.
- Narrows down scope of documents needed for LLMs to 10-15 papers out of 400k papers.

How it works:

- Specter embeddings are built on Scibert & trained on the citation graph.
- Minimizes the cosine difference of Specter embedding vectors and reranks on trigram similarity.
- Considers both the citation graph as well as specific n-grams in a sentence to determine similarity.

Challenges:

- Finding embedding model to represent complex scientific text.
- Runtime in large corpus.

Model 2: Summarization (Langchain)

The model:

- Langchain Framework is compatible with OpenAI (Summarization Chain)
- Allows backend to directly load abstracts and full text documents for summarization
- Adjustable settings for the OpenAI LLM for next word predictions

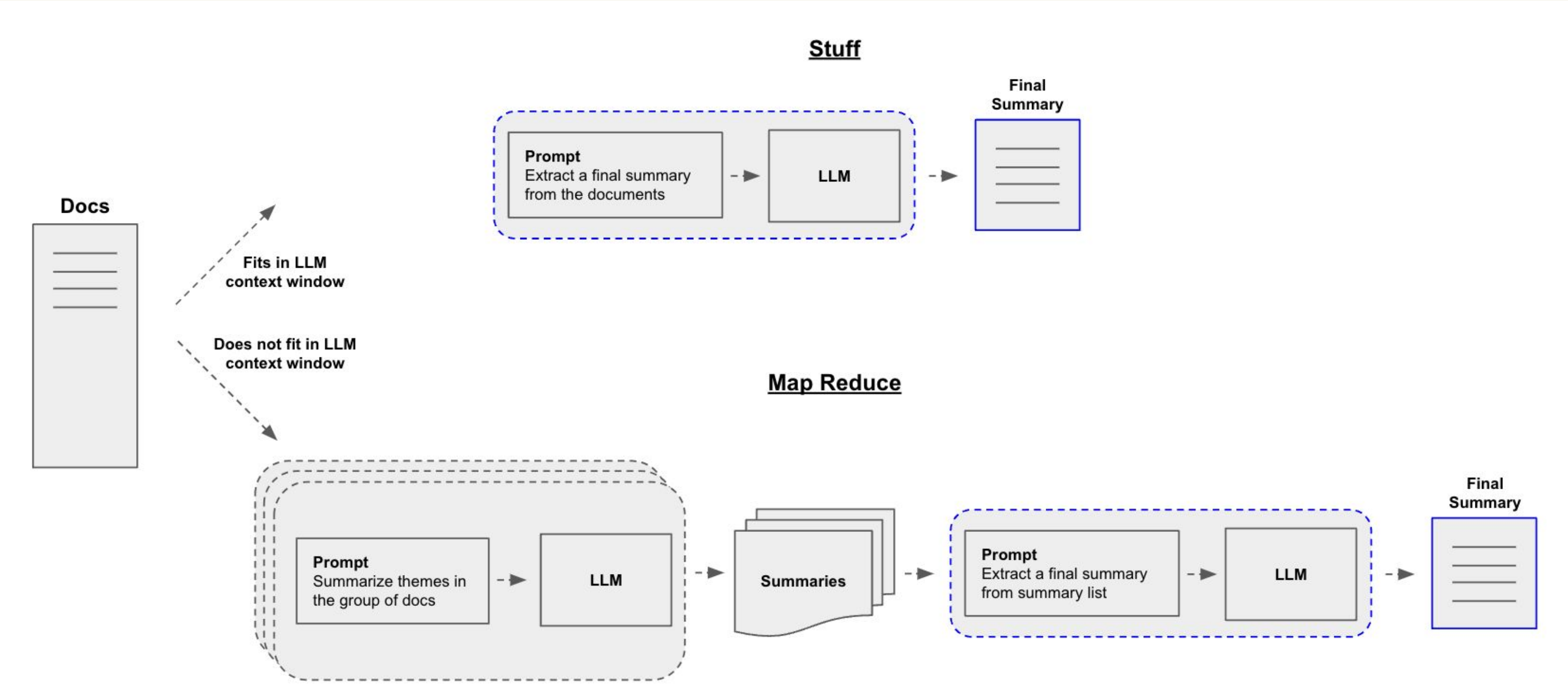
How it works:

- Langchain provides packages and functions that separate large text documents into chunks prior to being loaded into the model
- Chains come with adjustable parameters to utilize document objects with the desired LLM
- Customizable chains alter the method of splitting up documents and applying our LLM model to each piece

Challenges:

- Loading and separating documents in a consistent manner
- Choosing which model
- Integrating the similarity model and Postgress DB

Langchain (MapReduce Method)



Model 3: Summarization (HuggingFace)

The Model(s):

- T5 - transformer trained on multiple tasks using encoder and decoder structures
- LED - Longform Encoder Decoder useful for tokenizing and summarizing large bodies of text
- BioGPT - Transformer trained on medical research documents for classification and research purposes

How They Work:

- Each model has a tokenizer used to convert text into numeric representations. The models themselves have a generate method that when prefixed with a task for conditional generation, use the prior tokens to create a response. Each generate method consists of parameters that adjust how new sequences and words are selected in the response.

Challenges:

- Parameter Tuning
- Tokenization & Formatting

Model 4: Q&A (Langchain)

The Model:

- Langchain question and answer chain
- OpenAI LLM

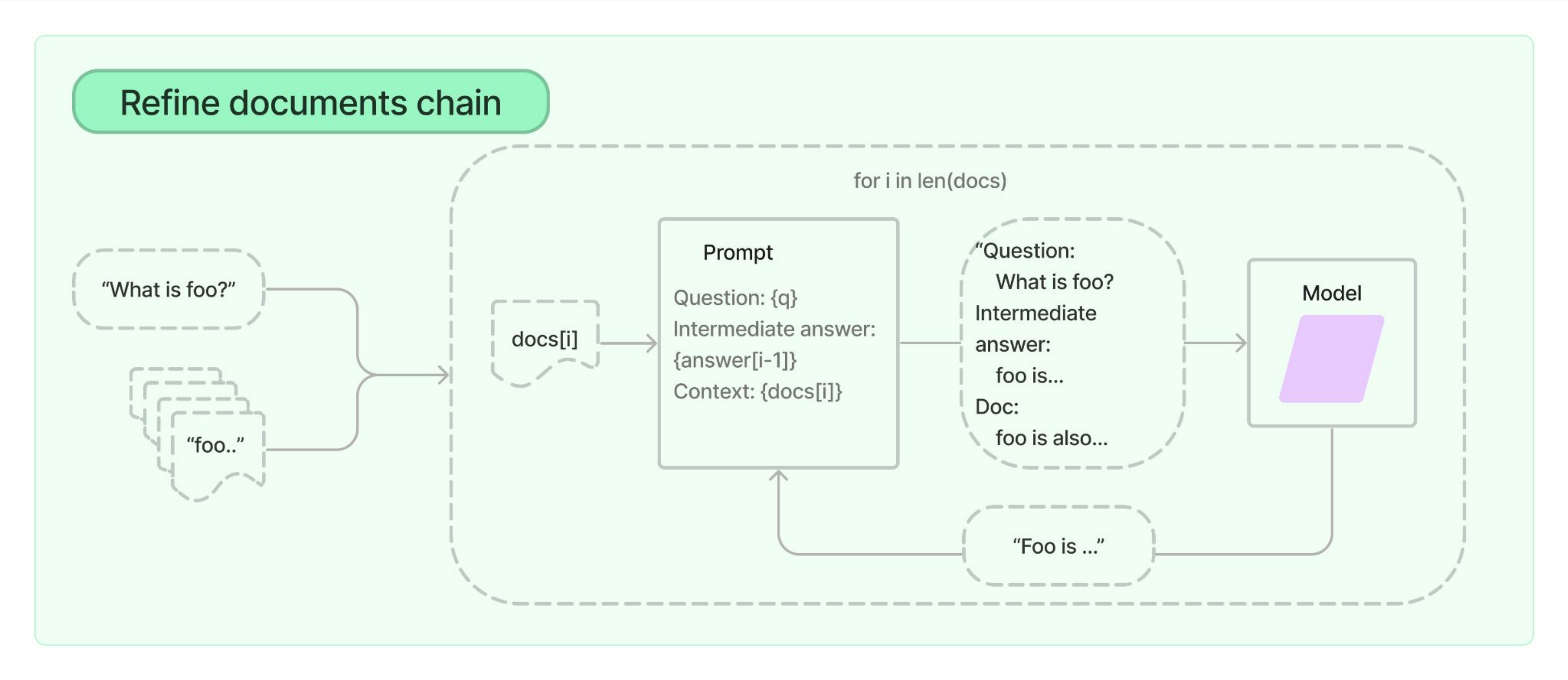
How it Works:

- A VectorStore or Document object with LangChain processes the relevant text
- Input a prompt with a question
- The q&a chain uses the same text splitting techniques to extract segments of a document when generating its response to the given prompt

Challenges:

- Text Preprocessing
- Chain Assemblance

Langchain (Refine Method)



Challenges & Next Steps

Challenges:

- Storing and accessing the research papers
- Quick runtime for output
- Evaluating the fluency and accuracy of responses

Next Steps:

- Finalizing the VectorStore to access papers
- Combining models for optimized outputs
- Uploading company research to our DB
- Customer ranking system for papers

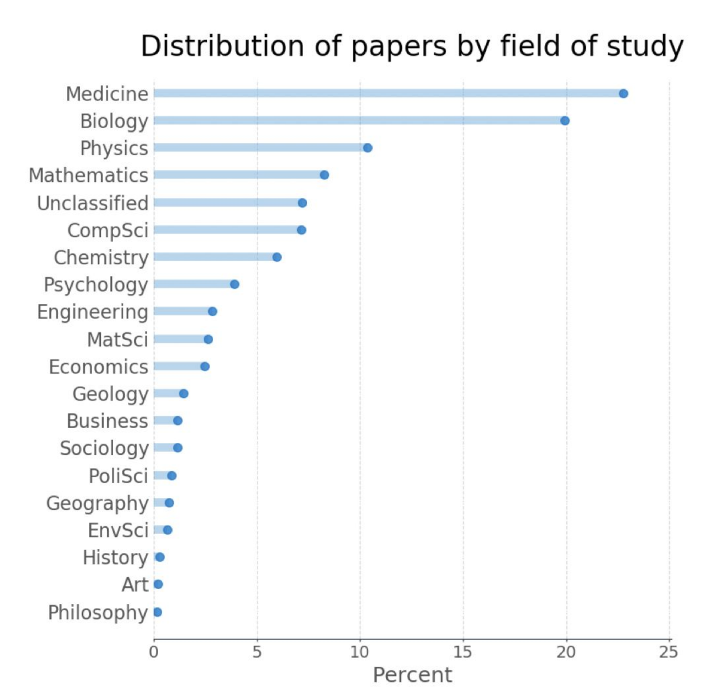
Thank You!

Appendix

References

- Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, *535*(7612), Article 7612. https://doi.org/10.1038/nj7612-457a
- *Artificial Intelligence In Healthcare Market Size Report, 2030*. (n.d.). Retrieved July 11, 2023, from https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market
- *Scaling up AI across the life sciences value chain*. (n.d.). Deloitte Insights. Retrieved July 11, 2023, from https://www2.deloitte.com/us/en/insights/industry/life-sciences/ai-and-pharma.html
- Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., & Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology*, *32*(1), 40–51. https://doi.org/10.1038/nbt.2786

Data Set



Total papers	81.1M
Papers w/ PDF	28.9M (35.6%)
Papers w/ bibliographies	27.6M (34.1%)
Papers w/ GROBID full text	8.1M (10.0%)
Papers w/ LaTeX full text	1.5M (1.8%)
Papers w/ publisher abstract	73.4M (90.4%)
Papers w/ DOIs	52.2M (64.3%)
Papers w/ Pubmed IDs	21.5M (26.5%)
Papers w/ PMC IDs	4.7M (5.8%)
Papers w/ ArXiv IDs	1.7M (2.0%)
Papers w/ ACL IDs	42k (0.1%)

Model Evaluation - T5 v Langchain

- We compared both the T5 model and Langchain ROUGE scores to evaluate which model to use in our final application.
- Due to the high cost and time to generate human written summaries to compare to our model generated summaries, in all model evaluations we compared summaries of a paper (full text/abstract) to a paper's abstract.
- This comparison isn't perfect, as the summaries are <5% the length of a full text article and <25% the length of an abstract, so there will be lower recall as information is lost when summarizing.
- While the Langchain model scored lower on F1 score, we though the summaries generated by Langchain read better than T5 (human verification) and were faster to generate (~15 sec LangChain v 3 mins T5), so we moved forward with Langchain.

ROUGE score	**Abstract Summarization (T5)**	**Full Text Summarization (T5)**	**Abstract Summarization (LangChain)**	**Full Text Summarization (LangChain)**
F1	.41	.44	.24	.17
Recall	.31	.28	.17	.13
Precision	.86	.98	.53	.32

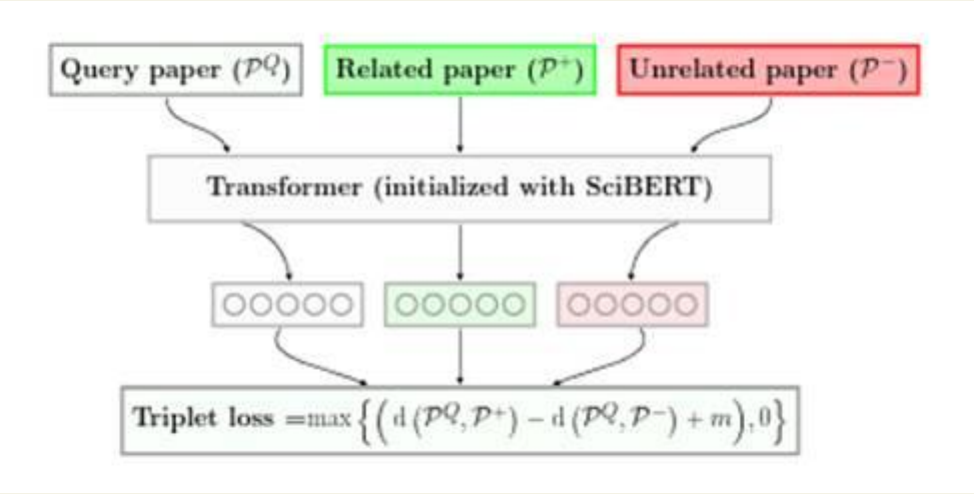
Similarity - Specter Embeddings

- Built off of existing LM SciBERT, trained on corpus of 1.14M papers (3.1B tokens).

- Trained on paper citations with the goal of adapting output representations so they are more similar for papers that share a citation link. This training is done on 146k papers (26.7M tokens).

- Loss function is described to the right, maximizing difference between related paper and unrelated paper.

- Additional training between direct citation and secondary citation.



Similarity - Specter Embeddings

Task →	Classification		User activity prediction				Citation prediction				Recomm.		Avg.
Subtask →	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model ↓ / Metric →	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nD̂CG	P@̂1	
Random	4.8	9.4	25.2	51.6	25.6	51.9	25.1	51.5	24.9	51.4	51.3	16.8	32.5
Doc2vec (2014)	66.2	69.2	67.8	82.9	64.9	81.6	65.3	82.2	67.1	83.4	51.7	16.9	66.6
Fasttext-sum (2017)	78.1	84.1	76.5	87.9	75.3	87.4	74.6	88.1	77.8	89.6	52.5	18.0	74.1
SIF (2017)	78.4	81.4	79.4	89.4	78.2	88.9	79.4	90.5	80.8	90.9	53.4	19.5	75.9
ELMo (2018)	77.0	75.7	70.3	84.3	67.4	82.6	65.8	82.6	68.5	83.8	52.5	18.2	69.0
Citeomatic (2018)	67.1	75.7	81.1	90.2	80.5	90.2	86.3	94.1	84.4	92.8	52.5	17.3	76.0
SGC (2019a)	76.8	82.7	77.2	88.0	75.7	87.5	**91.6**	**96.2**	84.1	92.5	52.7	18.2	76.9
SciBERT (2019)	79.7	80.7	50.7	73.1	47.7	71.1	48.3	71.7	49.7	72.6	52.1	17.9	59.6
Sent-BERT (2019)	80.5	69.1	68.2	83.3	64.8	81.3	63.5	81.6	66.4	82.8	51.6	17.1	67.5
SPECTER (Ours)	**82.0**	**86.4**	**83.6**	**91.5**	**84.5**	**92.4**	88.3	94.9	**88.1**	**94.8**	**53.9**	**20.0**	**80.0**

Similarity - Difference Metric

- Jaccard Similarity - Higher Score for more words in common, similar words divided by number of words in corpus

- Cosine Similarity / Euclidean Distance - Difference of embedded vectors, similar in this case as embeddings are same length.

- Euclidean Distance preferred for general categorization while Cosine preferred for text similarity.

- **Chose Cosine Similarity** - faster implementation in PG Vector, more flexible in case future document embeddings are different lengths.

Similarity - Reranking on Trigrams

Difference of embeddings gets us to ballpark of similar papers, though doesn't factor in enough information about specific words in text.

Re-rank top 1k similar papers based on similarity of specific words in abstracts:

1. Lexize Function to standardize text and remove stop words

2. Trigrams similarity of first 500 characters, capture main point of paper as well as any special terms

Ethical Considerations

- Misinformation
  - Hallucinations
  - Irreproducible results in publications
- Verifying Information
  - Abstractive Summarizations
  - Q&A
- Copyright issues
- Proprietary information from companies