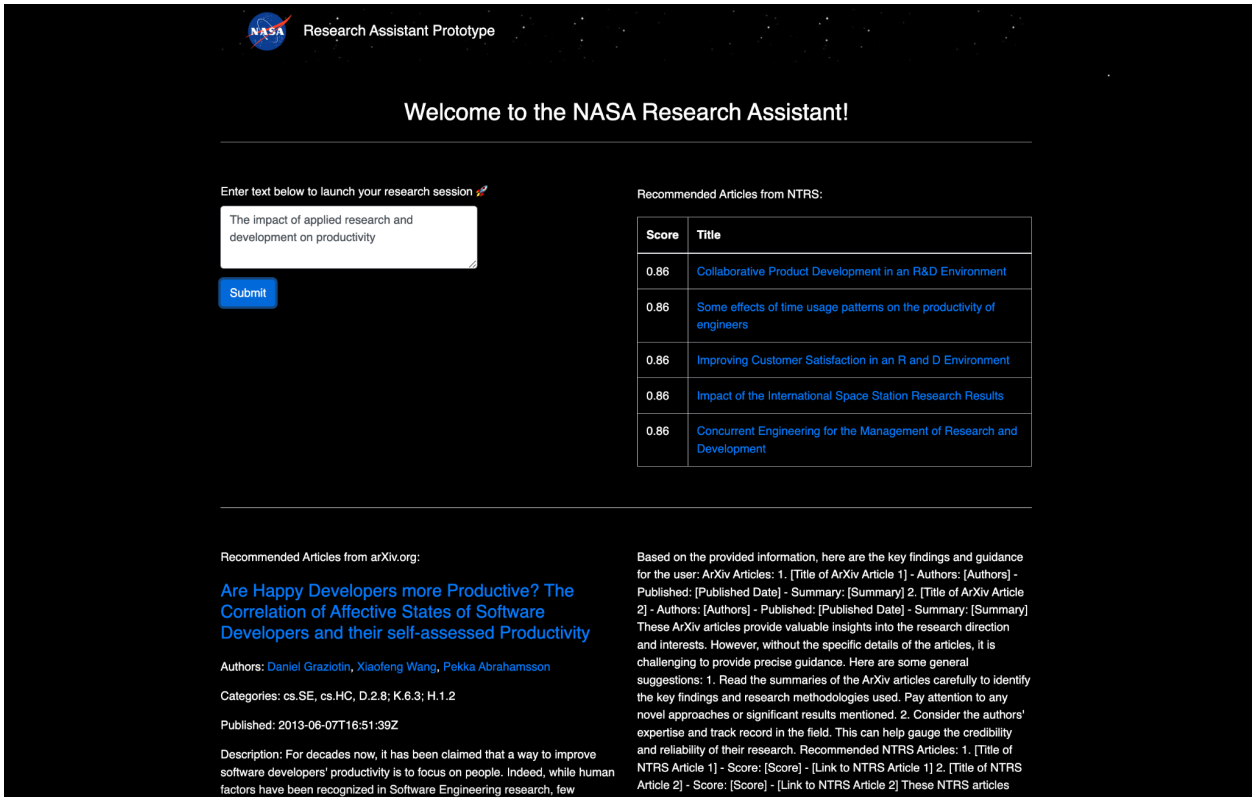
NASA researchers face the challenge of sifting through numerous databases and papers to extract relevant information. Our solution simplifies this process by leveraging the power of vector (“semantic”) search using OpenAI embeddings and PineconeDB.

Users input text, receiving relevant NTRS articles and papers from arXiv.org. The OpenAlex API is then used to retrieve a selected author's body of work, all accessible through our prototype's dynamic interface. This information is fed as context to a ChatGPT prompt, generating a user-friendly research summary (a.k.a., a “completion”).



Recent advancements emphasize deep neural networks and bi-encoder architecture for search. We explore the need for a dedicated "vector store" and demonstrate vector search with OpenAI embeddings and PineconeDB. Our methodology involves utilizing the OpenAI ada2 model for embeddings and PineconeDB for indexing.

Our methodology is centered on:

- **Embedding Generation**: Using OpenAI ada2, we create embeddings with 1536 dimensions for article titles and queries, efficiently managing a substantial dataset.

- **Indexing with PineconeDB**: PineconeDB supports cosine similarity scoring and top-k retrieval, aligning with vector search.

- **Data**: We work with 100,000 NTRS article titles and around 1.1 million arXiv article titles. The OpenAlex API also returns an author's body of work, as mentioned.

Encoding 100,000 NTRS titles took less than 15 minutes at a cost of $0.0001 per 1,000 token embeddings. However, a bottleneck was PineconeDB's limit for "bottom-tier" subscribers, capping databases at 100,000 records, unless they're willing to upgrade and incur a $70 per month cost. Open source alternatives like Postgres' pgvector were considered, but pose technical limitations (the use of Euclidean, or L2, distance instead of cosine similarity for scoring).

In conclusion, our demonstration showcases the efficiency of vector search with OpenAI embeddings and PineconeDB in contemporary AI stacks. Our pitch deck can be found here.