The Erdős Institute

Corporate NLP Project from Aware

Pawsitive Retrieval

The performance of modern generative AI systems, such as OpenAI's ChatGPT or Google's Gemini, relies on the quality of the data these systems are fed. For instance, if a user asks for a summary of the differences between puppies and kittens, the generative model must have the most relevant data about puppies and kittens available as `context` in order to produce a useful summary. But where does this relevant data come from? Currently, many systems use some form of Retrieval-Augmented Generation (RAG), which is simply a fancy way to say a clever search algorithm sifts through millions (or billions!) of pieces of content, ranks those pieces of content in terms of relevance, and then sends them on to a Large Language Model (LLM) to be summarized.

Teams that choose this project will replicate an internal RAG-building effort at Aware, a Columbus software startup that uses AI in the digital workplace to help customers reduce operational costs and drive insights from their internal data (such as those derived from HR surveys, Slack conversations, or email). The teams will be provided with more than 5 million public Reddit submissions and comments. The objective will be to build a system that can, given a user's query, identify and rank the most relevant content in these messages. As Aware deals with large, heterogeneous data sets, two key priorities for their teams emerge: (1) the system must be fast (sub-second response is a requirement), and (2) there needs to be some way to measure the qualitative performance of the result sets across retrieval methods.