This project looks at how artificial intelligence and big language models can help suggest where to submit research papers. It's often tough for researchers to pick the right journals or conferences. You have to know your field and what kind of stuff different publications usually want.

We built and tested two ways to tackle this problem. The data was used in different ways , for example a dataset using info from PubMed, grabbing titles, summaries, keywords, and journal details.

We turned this info into a format computers can understand using a the Ollama framework model called bge-small-en, and then saved it in a special database using a csv. When someone searches, we compare their search to the saved info and rank the best publication places.

The second way skipped building a dataset. Instead, we asked a pre-trained language model to suggest places to submit papers based on the paper's summary. Testing the data-driven way showed that it was good at grouping similar topics. The right suggestion was made in the top 3 percent of the data, which means it's pretty good at finding the right places. We also showed evidence articles with the suggestions, which helps people trust the system more.

The other way, just asking the language model, gave reasonable suggestions, even suggesting places that weren't in our PubMed data. But it wasn't always consistent and sometimes made stuff up. So, the two ways show that there's a trade-off: the data-driven way is dependable and clear but limited by the data it has, while the other way is flexible but not as reliable.

What we found out is that these computer-friendly formats are good for grouping similar topics, showing evidence helps build trust, and language models can suggest places beyond what's in existing datasets. We think that the best way to go forward is to use both: a system that uses data to find possible places and then uses language models to give even more suggestions.

Finding the right place to share your research is super important for scientists, but it can be tricky. Every year, tons of papers get sent to different journals and conferences in computer science, AI, and fields like cybersecurity. Getting your paper accepted and seen by others really depends on picking the right one. It can be a real pain, especially if you're just starting out. Usually, people just go with what they already know or ask their colleagues for advice. Common tools like publisher search engines and citation databases such as CrossRef, PubMed, and DBLP are helpful, but they don't always cut it. They usually just look for keywords and don't really understand what your paper is about, so you might miss some good options. They usually use the B25 algorithm.

But things are changing! New and cool ways to find and arrange scientific papers are popping up using things like info finding, how we talk, and suggestion systems. Google Scholar and Semantic Scholar are showing how AI can make things easier, but not many focus on helping researchers find the PERFECT spot for their new papers. And for students in college, to find what they’re looking for faster.. If there are systems that do this, they only use old methods like keyword searches or narrow ways to group papers. The rise in popularity of big language models like BERT and ChatGPT hasn't really touched venue suggestions yet. That's where we can really make a change by looking into the outcome.

Existing systems have three main problems. First, keyword-based models like BM25 are good starting points, but they depend too much on the same words being used. They might not realize that conference and venue basically mean the same thing. Second, embedding-based methods such as Sentence-BERT, SPECTER, and SimCSE try to get a better sense of what the paper means, but they're limited by what they were trained on. For example, SPECTER needs big citation networks, which aren't always around. Lastly, mixing embeddings with LLMs reasoning stuff like retrieval-augmented generation (RAG) hasn't been tried much for venue suggestions. That's why we want to create a system that uses all three types of tech: keywords, embeddings, and LLMs.

So, this project aims to make an AI-powered system that suggests good academic venues using both paper info and those big language models. Here’s what we plan to do: grab and clean up academic info (titles, summaries, keywords) from places like CrossRef, PubMed, DBLP, and Kaggle datasets including ArXiv and Citation Network using API keys. Then, make BM25 as a normal way to compare things, add embedding-based models like Sentence-BERT, SPECTER, and SimCSE for a better understanding, check out hybrid ways to improve results using LLMs and retrieval-augmented generation, build a working model with a website interface (Streamlit with a FastAPI backend), and test everything using normal ways to measure info finding (Precision k, MAP, NDCG) and real-life examples to see if the suggestions are good.

We're trying to answer four questions: First, how do old-school methods like BM25 stack up against embedding-based ones for suggesting venues? Second, can LLMs really make recommendations better and easier to understand? Third, what's the balance between being fast, handling lots of data, and being accurate with BM25, embeddings, and mixed models? Finally, how can we make sure this system is fair and ethical, balancing being right with keeping costs down and watching out for bias and sketchy journals?

This research matters because it can make it easier for researchers, especially those who are new, to pick the right journal or conference and not waste time on bad submissions. It also gives the academic world a way to mix info finding, embeddings, and LLM ranking for venue suggestions. And for info finding and how we talk research, it compares three types of finding models to see what works best. Besides the tech stuff, the project also thinks about being fair and legal, like following GDPR, using author info responsibly, and stopping the system from suggesting bad or fake venues by using journal ranking datasets.

What we're doing is focused on and limited by a few things. We're only using titles, summaries, and keywords from public places like CrossRef, PubMed, DBLP, and Kaggle. We're not looking at the full text of papers because of copyright and money. Also, we're looking at citation-based models, but using them is hard because we don't always have citation networks. LLM-based methods also have problems with being able to handle lots of data, cost, and being repeatable, which we know about. We're testing things with info finding measures and examples, but we're not doing big user tests with real researchers right now.

This paper is split into six parts. Chapter 1 talks about the problem, goals, and why it matters. Chapter 2 looks at past research, from old-school finding models to embeddings and LLM-powered methods, and what's missing. Chapter 3 explains how we're doing things, like getting data, cleaning it up, how the system is set up, picking models, and measuring results. Chapter 4 goes over how we built everything, like the prototype, adding models, and setting up tests. Chapter 5 shows the results, compares them, talks about what's good and bad, and what we learned about ethics. Chapter 6 wraps things up, sums up what we did, answers the questions, and suggests what to do next, like using it for more things and adding better ranking signals.

To sum it up, this project suggests an AI venue finder that fixes the problems of current systems by mixing old-school finding, embeddings, and LLM-based hybrid methods. It's helpful for researchers and valuable for the finding/how we talk community, while also being careful about being able to handle data, being repeatable, and being ethical.