# CS839 Project Proposal

Skylar Hou, Abigale Kim, Minh Phan, Mark Tervo

November 2025

## 1 Motivation

Dense retrieval uses neural networks to encode queries and documents into vectors, retrieving documents based on vector similarity. These methods achieve strong performance but require massive labeled datasets - often hundreds of thousands of query-document pairs [1, 2]. In many real-world scenarios, creating such labels is expensive and time-consuming for new domains or specialized applications.

**How can we improve dense retrieval accuracy without domain-specific training?** HyDE [3] proposes generating hypothetical documents as a solution: an instruction-following LLM generates a hypothetical document that answers the query, capturing relevance patterns of what a relevant document should look like. An encoder then retrieves real corpus documents similar to this generated document, achieving strong retrieval performance without domain-specific labels.

## 2 Hypothesis

We extend HyDE to table search, which has practical applications in wikitable and dataset retrieval. Our hypothesis is that generating hypothetical tables will improve zero-shot table retrieval performance compared to directly encoding queries.

## 3 Methodology

We will evaluate on the WikiTables dataset [4]. An instruction-following LLM will generate hypothetical tables from queries using chain-of-thought and few-shot prompting. Generated tables will be encoded with pre-trained contrastive encoders and used to retrieve real tables via embedding similarity. We will compare against direct query encoding using standard retrieval metrics. Optional extensions include incorporating pseudo-relevance feedback to iteratively refine hypothetical tables and evaluating whether hypothetical table generation improves sparse retrieval methods such as BM25.

# References


[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

[2] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

[3] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*, 2022.

[4] Shuo Zhang and Krisztian Balog. Ad hoc table retrieval using semantic similarity. In *WWW*, 2018.