R2L Lab Conclusion

Who cares? What difference does the author's results make? Discuss the broader impact of these technologies.
Information retrieval plays a critical role in how we navigate today's vast sea of information. With so much content available, it's simply impossible to read everything in order to decide what's relevant. Whenever we use a search engine or a recommendation system, we're engaging in information retrieval. Early approaches, like Boolean retrieval, relied on exact matches between queries and documents. But in most cases, a perfect match isn't available, and we also need to prioritize results based on relevance—this is where ranked retrieval comes in. These papers explore how to better do ranked retrieval, either though sparse retrieval or dense retrieval. In both cases, a ranking is determined by scoring the relevance of each document relative to the query. Finally, RAG extends these ideas by integrating information retrieval directly into large language models, enabling them to ground their generation in relevant external knowledge.

What are the risks? What are the potential failure modes or downsides of these approaches?
Sparse retrieval works, but it is unable to capture semantic meaning of terms (ex: "car" and "vehicle" mean similar things, but they are treated as separate entities). On the other hand, with dense retrieval, we work in a latent space, hence it is much harder to verify the correctness of what these representations actually mean. For RAG, it relies on a good retrieval system to work. If provided with poor documents that contain incorrect /limited information, the generator's performance will still be suboptimal.

Synthesis: Briefly explain how these three technologies fit together. How do sparse and dense retrieval support the RAG framework? What are the pros and cons of using one retrieval method over the other in a RAG system?
Sparse retrieval was the initial technology used to do information retrieval, i.e. using sparse vectors that would count the frequency of terms, and use those sparse vectors to match a query to a given document. However, sparse retrieval is unable to capture semantic meaning of words (since it sees "villain" and "bag guy" as completely different things even though they mean the same thing), and thus may result in suboptimal information retrieval.

Other the other hand, Dense retrieval (DPR) is able to capture semantic meaning, because it first encodes the text in a latent space via embeddings, and uses the dot product of these embeddings to quantify how similar two different pieces of texts are. However, prior to transformers the popularization of GPUs, and large pre-training, learning a solid latent representation was very difficult. However, thanks to those recent innovations listed, dense retrieval has consistently been shown to perform better in information retrieval that sparse retrieval.

Both sparse and dense retrieval can support the RAG framework, as the implementation of the retrieval system can either be implemented via sparse or dense retrieval. These retrieval systems can be used to ensure that the LLM doesn't "hallucinate", by retrieving relevant information from the relevant document that the LLM can use to generate a better answer, as opposed to purely relying on its parametric memory.