SLM + RAG

CS6101 Project 03 — Jonathan Chen, Shyamal Narang, JF Koh

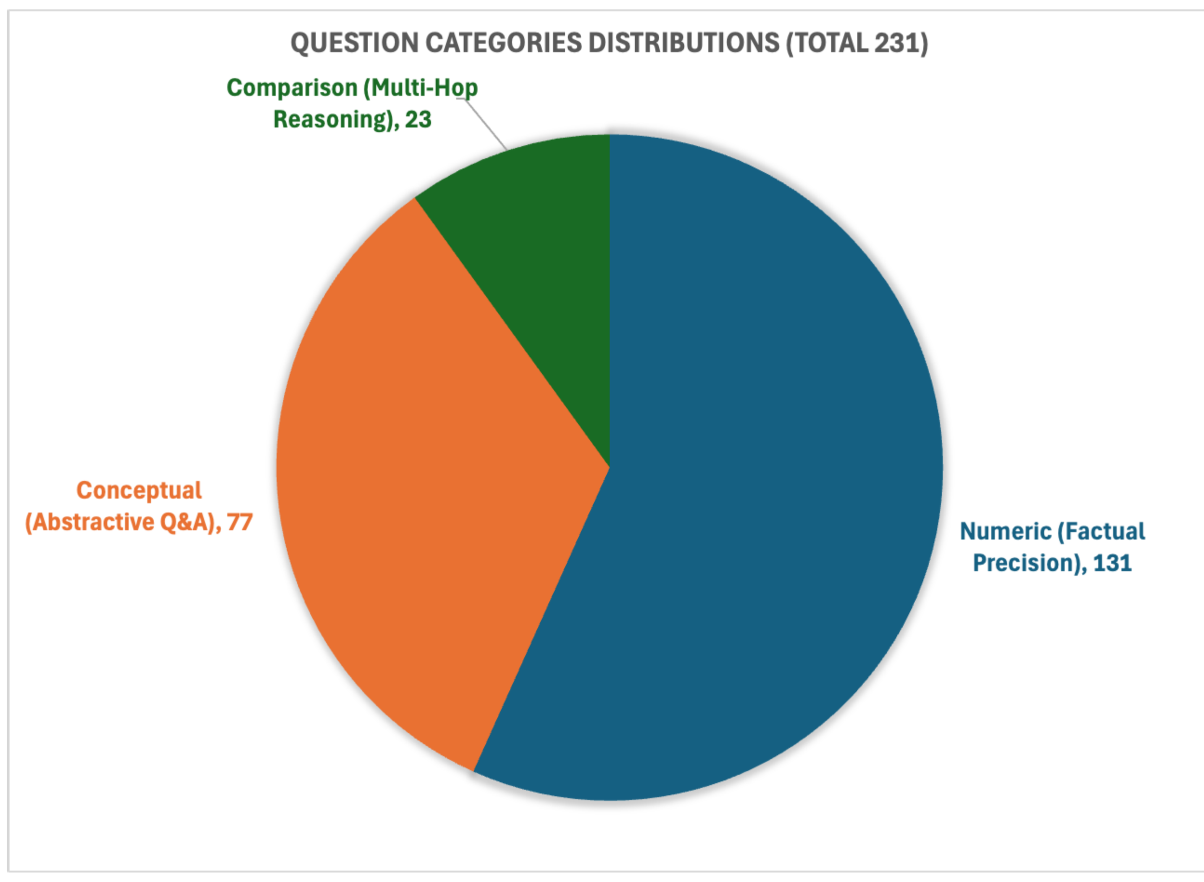
Abstract

With the rise of Generative AI, Small Language Models (SLMs) offer a more practical and affordable solution for industrial deployment compared to Large Language Models (LLMs). For integrating domain-specific knowledge, Retrieval-Augmented Generation (RAG) is often a more effective strategy than model fine-tuning. This project investigates the efficacy of RAG systems built upon SLMs. We will evaluate their performance and challenges across various question categories using product technical documents. The 2nd objective is to analyse whether the underlying language model size (SLM vs. LLM) significantly impacts the overall performance and reliability of the RAG system.

Dataset

Source: 48 camera lens manuals in PDF format downloaded from internet.
Content: image, text, simple table, etc.
Total question number: 229
Question categories:

- **Conceptual**: What is the purpose of Tamron's "full time manual focus" feature?"
- **Numeric**: What is the magnification ratio (Max. Mag. Ratio) of the 35-150mm F/2.8-4 Di VC OSD (Model A043) at the 150mm focal length?
- **Comparison**: How does the 25-200mm F/2.8-5.6 Di III VXD G2 (Model A075) compare to its predecessor (Model A071) regarding the starting focal length?



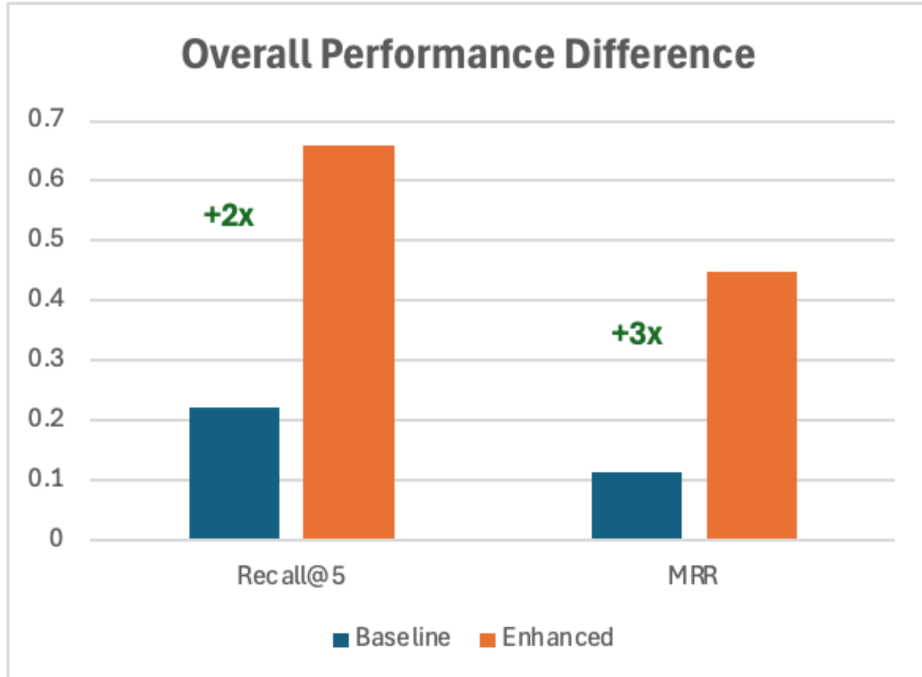
Preprocessing

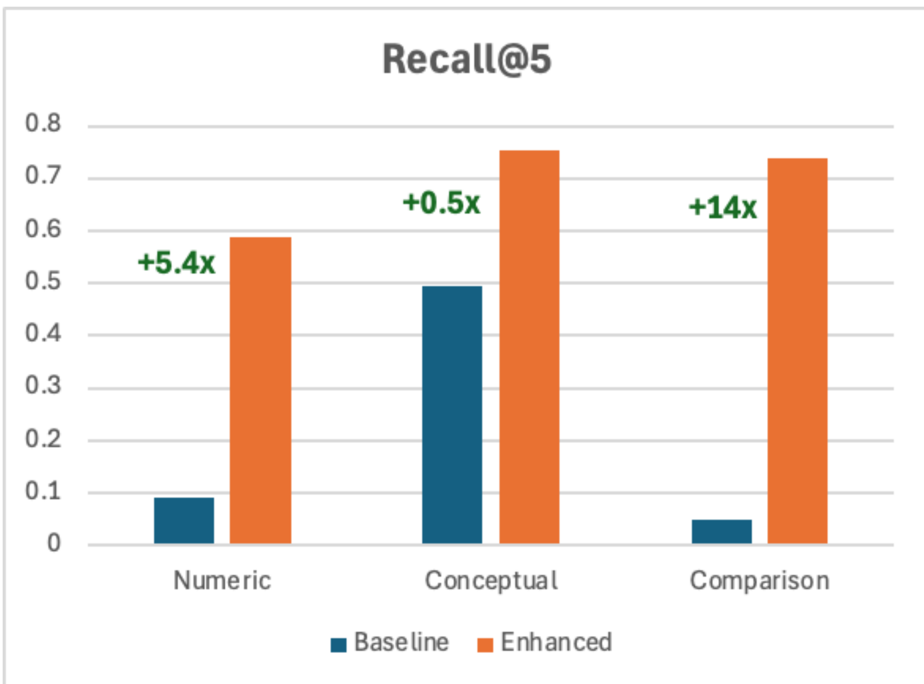
1. Spec-aware chunking
2. BM25 Tokenization to preserve f-numbers, ratios, units, model codes, acronyms
3. Hybrid scoring
   - -Adaptive BM25 weight (0.75 for numeric questions, 0.25 for text)
   - -Multiplicative boosts: doc_id (2.5x), spec_table (1.4x), spec_category(1.2x^n), units (1.15x), numeric proximity (1.3x)
   - -Additive doc_id bump - finding right doc first
   - -Phrase-specific boosts (1.35-1.45x), e.g. Daisangen, XLD+LD, etc.
4. Query preprocessing
   - -Multi-variation generation (acronym expansion, model prefix, unit normalization)
   - -Widened candidate pool
   - -Max BM25 score across variations
   - -Doc-ID augmentation
5. Answer matching
   - -Fuzzy numeric signals (ratios, f-numbers, num+unit)
   - -Acronym tolerance (XLD -> XLD2)
   - -Phrase subset matching

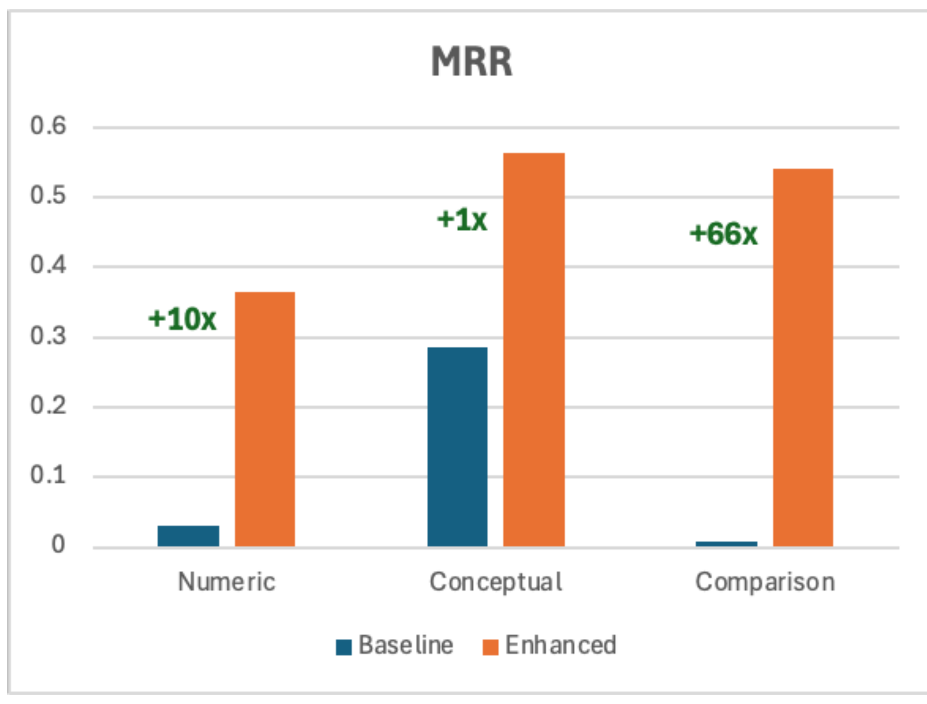
Retrieval

Performance boost from preprocessing tunings compared with vanilla retrieval.

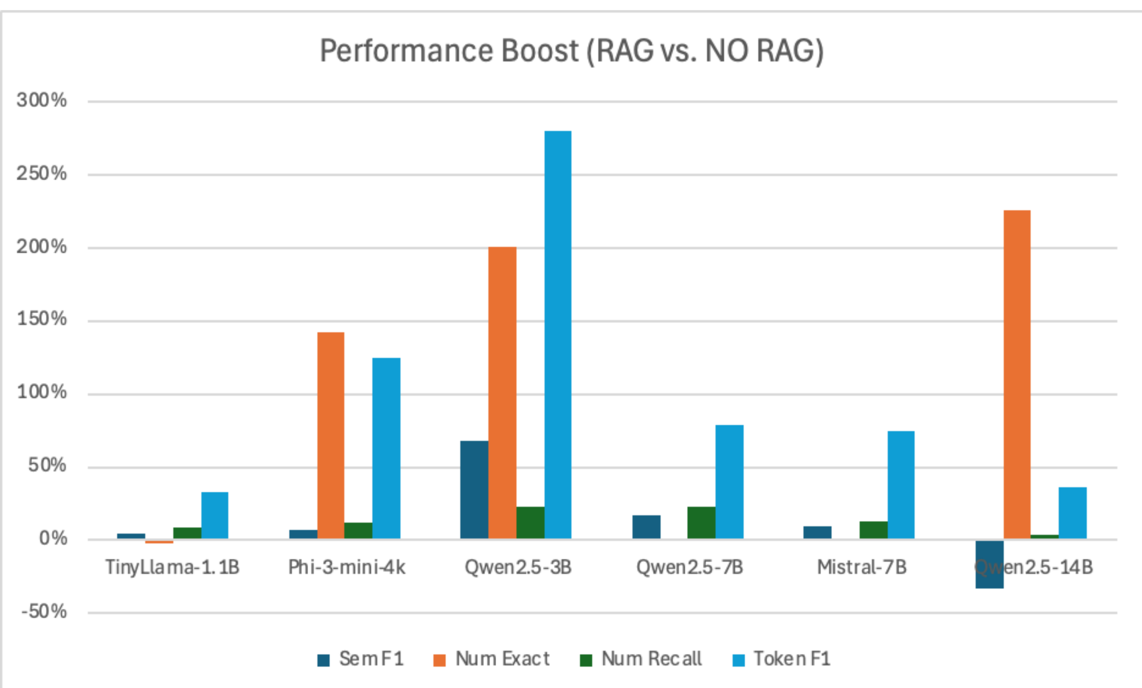
For domain specific document, carefully crafting preprocessing steps is the most rewarding activity in RAG.

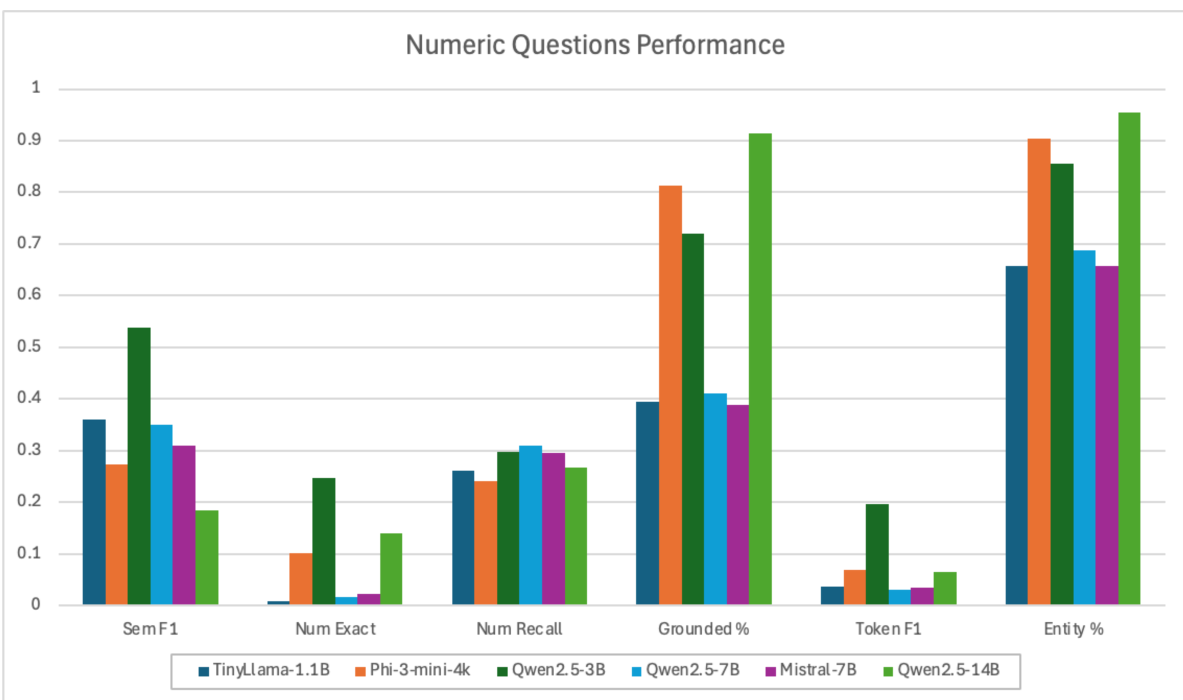


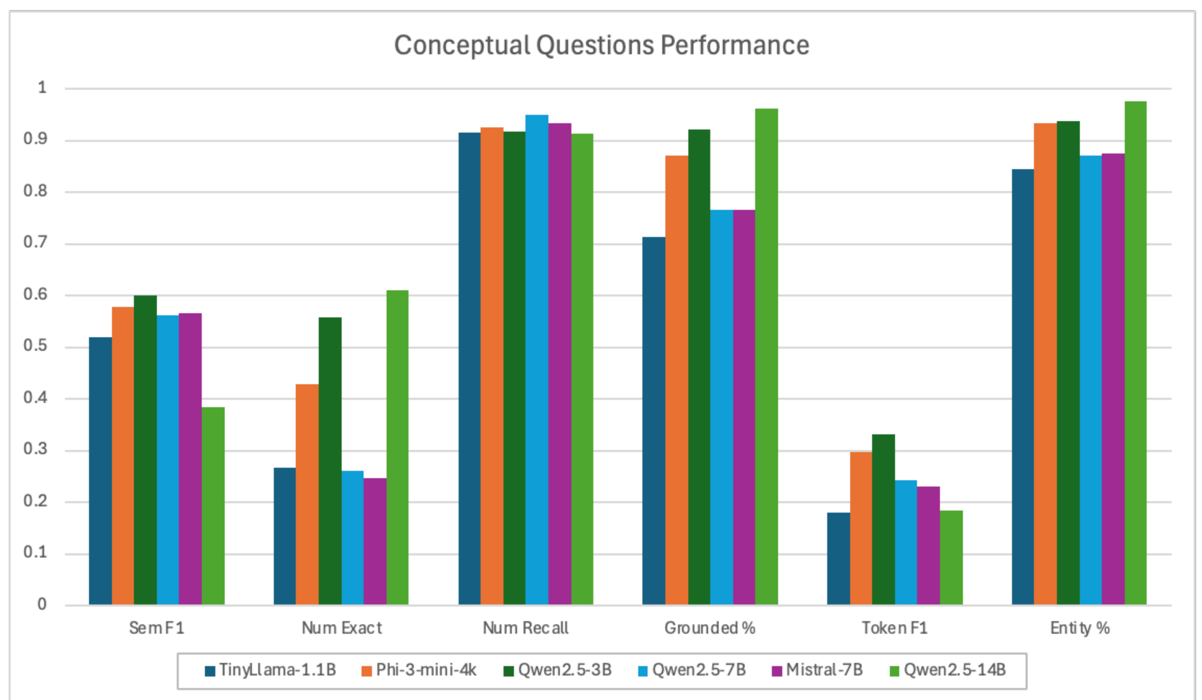




Generation







Model performance comparison

Conclusion

1. Preprocessing is foundation of performance of RAG
2. RAG can significantly boost performance of LM (+200%), especially in handling numeric specs.
3. LM size doesn't matter in overall performance of RAG