# Enhancing Tool Selection in Large Language Models: A Hybrid RAG-MCP Approach for Scalable External Tool Integration

**Yash Malode, Mihir Tale, Gleice Chaves, Pooja Salve, Vedant Rajeshirke, Laksh Jethani**
University of Southern California
{malode, tale, gchaves, psalve, rajeshir, ljethani}@usc.edu

## Abstract

As Large Language Models (LLMs) integrate with increasingly large toolsets through protocols like Model Context Protocol (MCP), prompt bloat degrades tool selection accuracy. We propose implementing and extending RAG-MCP with hybrid search techniques to address this scalability challenge. We propose to first implement and validate the original RAG-MCP methodology, and upon successful replication, extend it with hybrid search techniques combining BM25-semantic retrieval and reranking algorithms to potentially improve tool selection accuracy while maintaining computational efficiency.

## 1 Introduction

**Problem Definition**: Given a user query $q$ and a set of $N$ external tools $T = \{t_1, t_2, \ldots, t_N\}$ where each tool $t_i$ has description $d_i$ and schema $s_i$, select the most relevant tool(s) without overwhelming the LLM's context window.

**Example Input:** Query: "Find recent papers about climate change"

**Expected Output:** Selected tool: academic_search_api with parameters {query: "climate change", sort: "recent"}

When $N$ becomes large (50–1000+ tools), traditional approaches including all tool descriptions cause:

- Prompt bloat
- Degraded selection accuracy
- Increased costs

Gan and Sun (2025) showed accuracy drops to 13.62% as tool count increases, motivating scalable retrieval approaches.

## 2 Related Work

**Tool-Augmented LLMs**: Schick et al. (2023) and Yao et al. (2023) developed frameworks for LLM tool use but assume small toolsets.

**RAG for Tool Selection**: Gan and Sun (2025) introduced RAG-MCP using semantic retrieval, achieving 43.13% accuracy versus 13.62% baseline but relying solely on dense embeddings.

**Hybrid Retrieval**: Lewis et al. (2020) established RAG foundations. Hybrid approaches combining sparse (BM25) and dense retrieval show significant improvements (Gao et al., 2023). Khattab and Zaharia (2020) developed efficient late interaction reranking.

## 3 Hypothesis

Upon successful RAG-MCP implementation, we hypothesize that hybrid search (semantic + BM25) will either achieve superior tool selection accuracy or reduce overall task completion time or be an optimal balance versus original RAG-MCP and all-tools-in-prompt approaches.

## 4 Methodology

### 4.1 Baseline Implementation

Reproduce Gan and Sun (2025) methodology: semantic embeddings, FAISS indexing, top-$k$ retrieval.

### 4.2 Hybrid Extensions

**Retrieval**: Combine dense retrieval using `sentence-transformers` for semantic similarity with sparse retrieval using BM25 for exact keyword matching. Dense embeddings capture contextual meaning and handle synonyms, while BM25 excels at precise term matching and rare identifiers. Results from both methods are fused using Reciprocal Rank Fusion (RRF), which combines rankings by summing reciprocal ranks, or weighted linear combination based on query characteristics.

### 4.3 Evaluation

We will evaluate on an MCP benchmark of 50–100 tools spanning diverse domains, comparing the ap-

proaches: all-tools-in-prompt, original RAG-MCP (semantic retrieval), BM25-only, and hybrid variants (BM25 + semantic, optionally with reranking) We will report Accuracy, No. of tokens, and Latency for all the above-mentioned approaches.

## 5 Plan

**Milestone 1:** Implement and reproduce the RAG-MCP baseline to establish a reference system.

**Milestone 2:** Extend the system with a hybrid retrieval module integrating BM25 and semantic embeddings.

**Milestone 3:** Run experiments on tool-selection benchmarks, measure accuracy, efficiency, and latency, and carry out targeted error analysis.

**Milestone 4:** Consolidate results, compare baseline and hybrid methods, and finalize a fully tested implementation with key insights.

## 6 Expected Contributions

First, validate RAG-MCP results on realistic toolset. If successful, demonstrate hybrid retrieval effectiveness and provide systematic evaluation framework. Success criteria: achieve baseline performance ($\sim$43% accuracy), then potentially improve to $>$50% with hybrid approaches and $>$50% token reduction.

## References


Tiantian Gan and Qiyao Sun. 2025. RAG-MCP: Mitigating prompt bloat in LLM tool selection via retrieval-augmented generation. *https://arxiv.org/abs/2505.03275*.

Christopher J.C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Meng Wang. 2023. Retrieval-augmented generation for large language models: A survey. *https://arxiv.org/abs/2312.10997*.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.