

# Which Programming Language and Model Work Best With LLM-as-a-Judge For Code Retrieval?

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D.  
Roberts

L. Roberts and D. Roberts

Independent Researcher and New York University

12.10.2025

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature



Figure: Scan QR to connect to the project page.

# Contents

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

## Motivation

L. Roberts and D.  
Roberts

Motivation

## Experiments

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

## Literature

# Code Search and Code Generation-R is for Retrieval

Which Programming Language and Model Work Best With LLM-as-a-Judge For Code Retrieval?

L. Roberts and D. Roberts

## Why this matters?

Developer onboarding, productivity, and maintenance.

## Using a Retrieval Augmented code Generation (RAG)?

Even if you aren't searching code directly, your RAG system is. R (in RAG) is for retrieval.

### Motivation

#### Experiments

Research Questions

Data Construction & Experiments

Findings

#### Literature

## Motivation

## Experiments

Research Questions  
Data Construction &  
Experiments  
Findings

## Literature

# Relevance and Benchmarks

- ▶ Tedium, time consuming, costly, difficult to find annotators, difficult to build data, and difficult to setup good dataset.

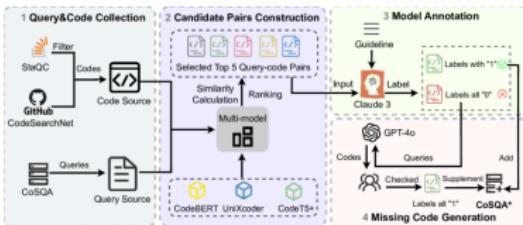


Figure: Construction of CosQA benchmark.

## Why this matters?

A search system is a complex system of interlocking technologies.

Benchmarks allow us to track changes in performance across system changes.

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

# LLM-as-a-judge

- ▶ Given query text and contents of each search result, concatenate with prompt template.
- ▶ Compare these relevance labels with those from human judgment.

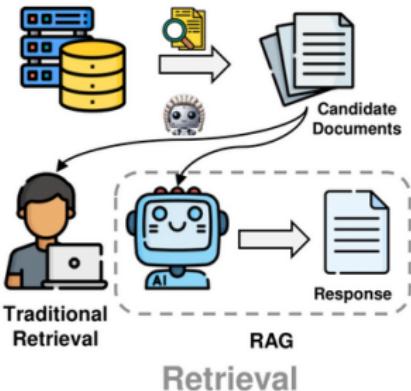


Figure: A standard LLM-as-a-judge workflow.

Motivation

Experiments

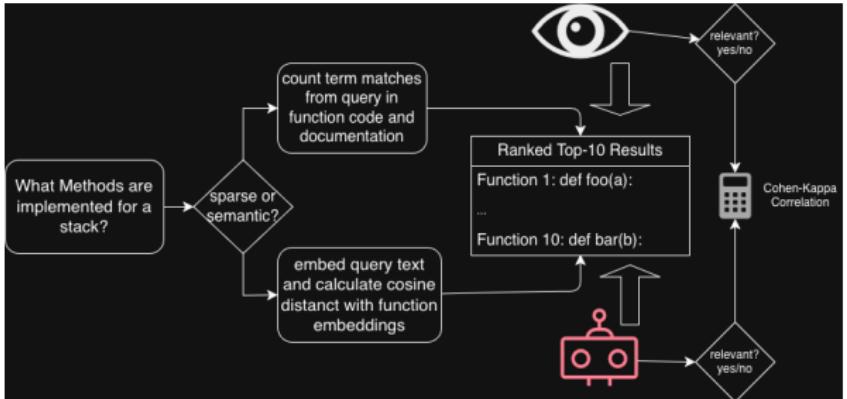
Research Questions

Data Construction &  
Experiments

Findings

Literature

# Experiment Workflow



LLM-as-a-judge annotation experiment process workflow.

- ▶ Sparse search uses BM25 for scoring/ranking.
- ▶ Semantic retrievers are transformers models trained on text & code. CodeBERT is encoder, CodeT5+ is encoder-decoder.
- ▶ Human (as-a-judge) label each query-result pair as relevant or not.
- ▶ LLM-as-a-judge: Nova-lite, gemini-2.0, gpt-4o-mini, Llama-4.

# Research Questions

- ▶ Are there affinities between Programming Languages (PL) and LLMs when evaluating relevance?
- ▶ Does the representation (sparse vs semantic) matter when evaluating relevance?
- ▶ If there are affinities with PLs how to scale benchmarks to other PLs? (most existing benchmarks are in a single programming language)

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D.  
Roberts

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

# Corpus

- ▶ We select git repositories with open source licenses across 5 popular programming languages.
- ▶ Java, Javascript, Go, Python, C.
- ▶ Each repository contains common data structures, Stacks, Priority Queues, Maps, etc.
- ▶ Parse the functions for each repo using a (programming) language parser, tree-sitter.



Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D.  
Roberts

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

# Corpus Statistics

Which Programming Language and Model Work Best With LLM-as-a-Judge For Code Retrieval?

	C	Js	Python	Go	Java	
Stack	✓	✓	✓	✓	✓	L. Roberts and D. Roberts
List	✓	✓	✓	✓	✓	Motivation
Set	✓	✓	✗	✓	✓	Experiments
Map	✓	✓	✗	✓	✓	Research Questions
Ordered Set	✓	✓	✗	✓	✓	Data Construction & Experiments
Tree	✗	✗	✓	✓	✓	Findings
Queue	✓	✓	✓	✓	✓	Literature
Heap	✓	✓	✓	✓	✓	
Trie	✗	✗	✓	✗	✗	
% Docs Absent	25.17	93.87	56.94	54.01	19.91	
# Functions	576	163	144	1,409	844	
Lines of Code	7,285	1803	978	16,567	6,515	
# Doc Tokens	32,762	677	502	17,344	27,057	
# Code Tokens	43,257	12,326	6,724	132,402	42,482	

# Search Queries

- ▶ We construct a list of queries, to cover the 3 main types of queries listed by Broder's taxonomy of web search.
- ▶ Some queries are constructed specifically for a certain data structure whilst others are parametrized by data structure name.
- ▶ Compared 3 retrievers, sparse: BM25, semantic: CodeBERT, and CodeT5+.
- ▶ Relevance labels from humans:

$$\underbrace{2}_{\text{humans}} \times \underbrace{5}_{\text{PLs}} \times \underbrace{3}_{\text{Retriever}} \times \underbrace{33}_{\text{query}} \times \underbrace{10}_{\text{Results size}} = 9,900.$$

Which Programming Language and Model Work Best With LLM-as-a-Judge For Code Retrieval?

L. Roberts and D. Roberts

Motivation

Experiments

Research Questions

Data Construction & Experiments

Findings

Literature

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D.  
Roberts

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

# Transpiler Experiment

- ▶ Translate from Python to C.
- ▶ We use the Cos-QA dataset which contains 19,000 query, code, and relevance markings.
- ▶ Send the translated code to the LLM-as-a-judge.



Transpilers-Noteworthy Tools, Comm. ACM Oct 2025 Cover.

Motivation

Experiments

Research Questions  
Data Construction &  
Experiments

Findings

Literature

# Programming Language Affinities

Cohen- $\kappa$  correlation between human and LLM relevance determinations

	Code T5+		BM25	
	Nova-lite	GPT 4o-mini	Nova-lite	GPT 4o-mini
Python	0.23112	-0.05479	0.25286	0.02932
C	0.00075	0.22783	0.07680	0.36064
Go	0.13167	0.00389	0.28438	0.10107
Js	0.03704	-0.00464	0.14663	0.06904
Java	0.08710	0.19906	0.26513	0.01243

Many clear preferences-regardless of representation.

Many more metrics and model comparisons in the paper.

# Representations

Which Programming Language and Model Work Best With LLM-as-a-Judge For Code Retrieval?

Improvement or Reduction in Human-AI Relevance Annotation Alignment (Cohen- $\kappa$  correlation), Best Sparse and Best Semantic.

L. Roberts and D. Roberts

Motivation

Experiments

Research Questions

Data Construction & Experiments

Findings

Literature

Language	BM25	CodeT5+	% change
Python	0.255286	0.23112	- 9.5%
C	0.36064	0.22783	-36.8 %
Go	0.10107	0.13167	30.3 %
Js	0.14663	0.03704	-74.7 %
Java	0.26513	0.19906	-24.9 %

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D.  
Roberts

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

# Semantic Search Finds Data Structure Synonyms

Search the repo:

ible for a heap data structure?

Codes

Two alternately implemented and named data structures, heap and priority queue are both surfaced at the top with the heap query.

Size

by /Users/rucas/gods/trees/binaryheap/binaryheap.go on 8323d02ee3ca1499478f9cccd7a299fb1c5005780

distance 437

Size

by /Users/rucas/gods/queues/priorityqueue/priorityqueue.go on  
8323d02ee3ca1499478f9cccd7a299fb1c5005780

distance  
4581

Semantic Search finds heap/priority queue synonym.  
Whereas sparse methods like BM25 require thesaurus/query expansion/wordnet etc.

# CosQA Transpiler (Py-2c) Results

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

LLM-as-a-judge on transpiled CosQA data, Python transpiled to C.

<i>nova-lite</i>	not-relevant	relevant	
not-relevant	3845	987	4832
relevant	3171	1018	4189
53.91%	7016	2005	9021
<i>gpt-4o-mini</i>	not-relevant	relevant	
not-relevant	2657	2175	4832
relevant	2060	2129	4189
53.05%	4717	4304	9021

Cohen- $\kappa$ 's are 11.19 % and 5.79% for the nova-lite and gpt-4o-mini models, respectively.

L. Roberts and D. Roberts

Motivation

Experiments

Research Questions

Data Construction & Experiments

Findings

Literature

# Findings Summary

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D.  
Roberts

Motivation

Experiments

Research Questions  
Data Construction &  
Experiments

Findings

Literature

- ▶ There are affinities with LLM-as-a-judge and PLs.
- ▶ Representation matters, perhaps (un)surprisingly, BM25 is a strong candidate for retrieval mechanism for code. Many design consequences from this choice.
- ▶ To scale benchmarks to other PLs consider using a transpiler.

# REFERENCES I

-  Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., and Zhou, M. (2020).  
Codebert: A pre-trained model for programming and natural languages.
-  Hong, J. and Ryu, S. (2025).  
Automatically translating c to rust.  
*Commun. ACM*, 68(11):58–65.
-  Huang, J., Tang, D., Shou, L., Gong, M., Xu, K., Jiang, D., Zhou, M., and Duan, N. (2021).  
CoSQA: 20,000+ web queries for code search and question answering.  
In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5690–5700, Online.  
Association for Computational Linguistics.
-  Lù, X. H. (2024).  
Bm25s: Orders of magnitude faster lexical search via eager sparse scoring.

Which  
Programming  
Language and  
Model Work Best  
With  
LLM-as-a-Judge  
For Code  
Retrieval?

L. Roberts and D. Roberts

Motivation

Experiments

Research Questions

Data Construction &  
Experiments

Findings

Literature

## REFERENCES II



Wang, Y., Le, H., Gotmare, A., Bui, N., Li, J., and Hoi, S. (2023). CodeT5+: Open code large language models for code understanding and generation. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1069–1088, Singapore. Association for Computational Linguistics.



Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Which Programming Language and Model Work Best With LLM-as-a-Judge For Code Retrieval?

L. Roberts and D. Roberts

Motivation

Experiments

Research Questions

Data Construction & Experiments

Findings

Literature