

# GENCODESEARCHNET

A BENCHMARK TEST SUITE FOR EVALUATING GENERALIZATION IN PROGRAMMING LANGUAGE UNDERSTANDING

Andor Diera, Abdelhalim Dahou, Lukas Galke, Fabian Karl, Florian Sihler, Ansgar Scherp

## Abstract

Code datasets used for training language models consists mainly of the most popular languages and evaluation is predominantly carried out on similar distributions, often overlooking low-resource programming languages. Motivated by the NLP generalization taxonomy proposed by Hupkes et.al., we propose a new benchmark dataset called GenCodeSearchNet (GeCS) which builds upon existing natural language code search datasets to systemically evaluate the programming language understanding generalization capabilities of language models. As part of the full dataset, we introduce a new, manually curated subset StatCodeSearch that focuses on R, a popular but so far underrepresented programming language that is often used by researchers outside the field of computer science. For evaluation and comparison we provide two measures with text-code matching and ranking. We collect several baseline results using fine-tuned BERT-style models and GPT-style large language models in a zero-shot setting.

## Contributions

- A generalization benchmark dataset named **GenCodeSearchNet** that tests text-code matching and ranking
- A novel, manually-curated test dataset named **StatCodeSearch**, consisting of 1,070 text-code pairs from statistical research code written in R.
- Initial baseline results for **RoBERTa**, **CodeBERT**, **CodeT5+**, and **GPT**-based LLMs.

## GenCodeSearchNet

The GeCS dataset includes one fine-tuning set and eight test sets. It contains three previously proposed and a newly introduced dataset:

- **CodeSearchNet** (Husain et al., 2019)
- **CodeSearchNet AdvTest** (Lu et al., 2021)
- **CoSQA** (Huang et al., 2021)
- **StatCodeSearch** (own)

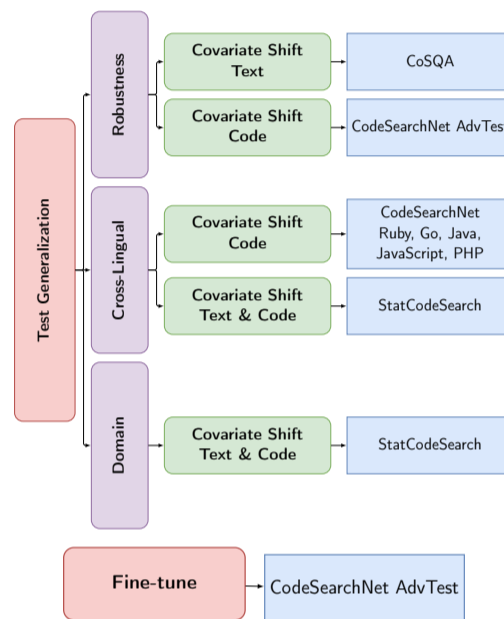


Figure 1: Overview of the benchmark composition

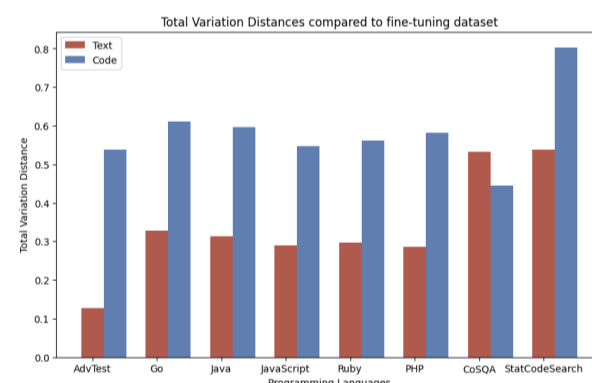


Figure 2: Covariate shifts measured by total variation distance

## StatCodeSearch



2,832 public R projects scraped from the Open Science Framework



Rule-based extraction of 40,041 code comment pairs



Automated filtering with GPT 3.5 resulting in 10,137 code comment pairs



Manual filtering resulting in 1,070 code comment pairs

## Evaluation

We apply two measures for assessing the performance of the models.

**Matching** We test whether a given text-code pair is a matching pair (positive) or not (negative). We evaluate the **Accuracy** on balanced test sets with an equal number of positive and negative examples. The non-matching examples are sampled uniformly within the respective dataset.

**Ranking** To evaluate the ranking in the code search task, we employ **Mean Reciprocal Rank (MRR)**. For each text query, we consider 99 distractors sampled uniformly at random. For each query, the reciprocal of the best ranked correct answer is considered

## Baseline Results

Table 1: Aggregated Performance for Each Generalization Type

Model	Robustness		Cross-Lingual		Domain	
	Acc	MRR	Acc	MRR	Acc	MRR
<b>Fine-tuned</b>						
RoBERTa	0.901	0.214	0.911	0.032	0.895	0.055
CodeBERT	0.931	0.263	0.917	0.057	0.960	0.0251
CodeT5+	0.938	0.115	0.889	0.059	0.905	0.058
<b>Zero-shot</b>						
CodeT5+	-	0.825	-	0.725	-	0.631
GPT 3.5 Turbo	0.448	-	0.607	-	0.627	-
Ada 2	-	0.885	-	0.839	-	0.794

## Take-home Message

- Fine-tuned encoder-only models are strong at matching even in out-of-distribution test sets
- Large-scale embedding models are strong at zero-shot ranking
- Lowest numbers were achieved on the newly introduced StatCodeSearch subset, indicating a challenging distribution shift

## Bibliography

- [1] Junjie Huang et al. "CoSQA: 20,000+ Web Queries for Code Search and Question Answering". 2021
- [2] Dieuwke Hupkes et al. "A taxonomy and review of generalization research in NLP". 2023
- [3] Hamel Husain et al. "CodeSearchNet challenge: Evaluating the state of semantic code search". 2019
- [4] Shuai Lu et al. "CodeXGLUE: A Machine Learning Benchmark Dataset for Code Understanding and Generation". 2021

## Full Paper & Contact



andor.diera@uni-ulm.de