

# DATASET REQUIREMENTS SPECIFICATION

## ComBack: Requirements Specification

Owner: Ming Zhong; Created: 2024/03/19; Last updated: 2024/05/21

### Vision

ComBack is the first public dataset for compiler backend development, it comprises 178 backends for mainstream compilers and three tasks for common backend development scenarios: Statement-Level Completion, Next-Statement Suggestion, and Code Generation.

### Motivation

Compiler backend development is a laborious and time-consuming task, lacking effective automation methods in the community. To mitigate this challenge, we propose ComBack, which can be utilized to fine-tune models. This approach aims to facilitate the automatic completion and generation of backend code using the fine-tuned model, thereby reducing redundant development efforts and enhancing manual efficiency.

### Intended uses

ComBack is intended is for non-profit research and educational purposes.

### Non-intended uses

ComBack is not intended is for commercial uses, because all source data in ComBack are obtained from open-source repositories.

### Data mocks

```
//Inputs:
unsigned FPReg = getFPReg(STI);
...
adjustReg(MBB.LastFrameDestroy, DL, SPReg, FPReg, ~StackSize+RVFI->getVarArgsSaveSize());
//Ground Truth:
MachineInstr::FrameDestroy;
```

(a) Statement-Level Completion

```
//Inputs:
unsigned maxCallFrameSize = MFI -> getMaxCallFrameSize();
...
maxCallFrameSize = (maxCallFrameSize + AlignMask) & ~AlignMask;
//Ground Truth:
MFI -> setMaxCallFrameSize(maxCallFrameSize);
```

(b) Next-Statement Suggestion

```
//Inputs:
getPointerRegClass: Returns a TargetRegisterClass used for pointer values.
Target-Specific Value: Sparc, SP::i64RegsRegClass, SP::IntRegsRegClass.
//Ground Truth:
TargetRegisterClass *SparcRegisterInfo::getPointerRegClass(MachineFunction &MF, unsigned Kind) {
    return Subtarget.is64Bit() ? &SP::i64RegsRegClass : &SP::IntRegsRegClass;
}
```

(c) Code Generation

### Creation requirements

- GitHub. Crawling from open-sourced repositories related to "LLVM Backend" and "GCC Backend".
- LLVM Official Released Website. Downloading LLVM released version 2.0.1 - 17.0.1.
- GCC Official Released Website. Downloading GCC released version 3.0 - 13.0.

### Sign-off grid

Name	Role	Date
Ming Zhong	Owner	2024/04/29

### Instance requirements

For completion tasks:

- Coverage. The proportion of tokens in the instance relative to the entire function is over 30%.
- Ground Truth. Assuming each sample contains  $n$  statements, for Statement-Level Completion, the first  $n-1$  statements along with the preceding 50%-90% of tokens from the  $n_{th}$  statement as input. The subsequent 10%-50% of tokens from the  $n_{th}$  statement served as ground truth. For Next-Statement Suggestion, the preceding  $n-1$  statements serve as input, while the  $n_{th}$  statement serves as the ground truth.

For Code Generation:

- Function Description. ComBack only contains functions with natural language descriptions, discarding those without such descriptions.

### Distributional requirements

- Train/validation/Test. 80%/10%/10% of the entire dataset.

### Data processing requirements

- Length. For completion tasks, Instance with input lengths exceeding 512 tokens or output lengths exceeding 128 tokens are filtered. For Code Generation, input exceeding 256 tokens or ground truth surpassing 512 tokens were removed.

### Performance requirements

- Accuracy. Improvement of accuracy across three tasks for fine-tuned models with ComBack.

### Maintenance requirements

The data should be regularly updated with new released version of LLVM and GCC, and other open-sourced LLVM and GCC backends repositories in GitHub.

### Sharing requirements

ComBack is available at <https://huggingface.co/datasets/docz-ict/ComBack>, it can be shared under CC-BY-4.0 license.

### Caveats and risks

No risks.

### Data ethics

No ethical implications.

### Changelog

Editor	Comments	Date
Ming Zhong	Fix Errors	2024/05/21

# DATASET DESIGN DOCUMENT

## ComBack: Design Document

Owner: *Ming Zhong*; Created: 2024/03/19; Last updated: 2024/05/21

### Overview

**Dataset Name:** ComBack.

**Primary Data Type(s):** Code, text.

**Data Content:** Code.

**Funding:** None

### Objective

ComBack is the first public dataset for compiler backend development, which aims to enhance programmers' efficiency by fine-tuning language models based on it.

### Version

Initial Version: V1.0.

### Background

Compiler backend development is a laborious and time-consuming task, lacking effective automation methods in the community. To mitigate this challenge, we propose ComBack, which can be utilized to fine-tune models and facilitate the automatic completion and generation of backend code.

### Sources

Source data contains 21 GCC repositories and 296 LLVM repositories in GitHub with "LLVM, GCC, Backend" as keywords. Additionally, source data also comprises source code of GCC versions 3.0 to 13.0 and LLVM versions 2.0.1 to 17.0.1.

### Annotations

Function descriptions and target-specific values are features in ComBack, they will be filtered in code completion tasks and serve as input for Code Generation.

### Data Quality

Quality are measured by reliability and feature representation. Labels in ComBack are function descriptions and target-specific values, the first one is crawled or extracted from websites and source code without manual intervention. The latter one is firstly labeled by a script with rules, then double-checked by manual efforts.

**Related Datasets:** TenSet, Circuit Net 2.0

**Dataset Discovery Process:** Doing search on mainstream websites and databases, including arxiv, GitHub, Huggingface, etc.

**Survey:** TenSet is a tensor program performance dataset. Circuit Net 2.0 is a dataset for chip design environment.

	ComBack	TenSet	Circuit Net 2.0
<b>Documentation and DOI</b>	<a href="https://huggingface.co/datasets/docz-ict/ComBack">https://huggingface.co/datasets/docz-ict/ComBack</a>	<a href="https://github.com/tlc-pack/tenset">https://github.com/tlc-pack/tenset</a>	<a href="https://circuitnet.github.io/">https://circuitnet.github.io/</a>
<b>Motivation and Intended use</b>	Backend Development	Performance Prediction	Chip Design
<b>Size, Sampling and Filtering</b>	422,735 samples for 3 tasks.	13,848 tasks.	10,791 samples.
<b>Annotation and Labels</b>	Target-Specific Values	Network and Subgraph	Congestion, DRV, etc.

### Characteristics

**Expected Characteristics:** Statement-Level Completion contains 161,124 data samples containing 14.2M token. Next-Statement Suggestion contains 216,315 samples with 19.7M tokens. Code Generation contains 45,296 samples with 6.4M tokens. All datasets are divided into train/validation/test sets with an ratio of 80%:10%:10%.

### Privacy Handling

All source data are collected from open-source repositories, thus there is no privacy issue.

### Maintenance

The owner and his team will be responsible for maintain the dataset. They will regularly extend the dataset with backends in newly released GCC, LLVM and open-sourced projects. As the dataset has been open-sourced at Huggingface, it will be easy to recover from former version when issues arise.

### Sharing

ComBack is available at <https://huggingface.co/datasets/docz-ict/ComBack>, it can be shared under CC-BY-4.0 license.

### Caveats

No caveats.

### Data Ethics

No ethical considerations.

### Work estimates

It takes about a week to collect data and a month for data pre-processing and labeling data.

### Author Statement

We bear all responsibility in case of violation of rights.