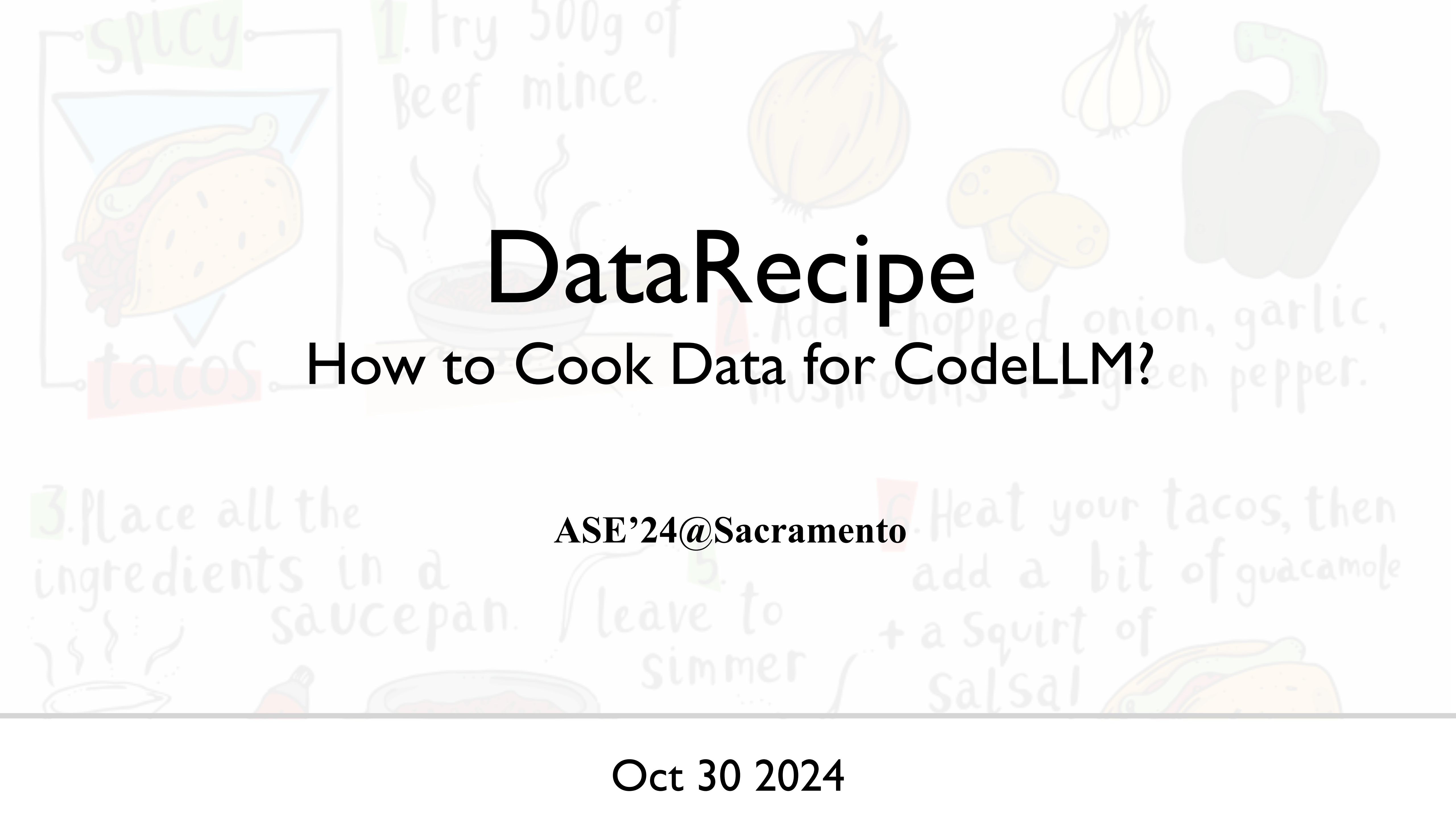




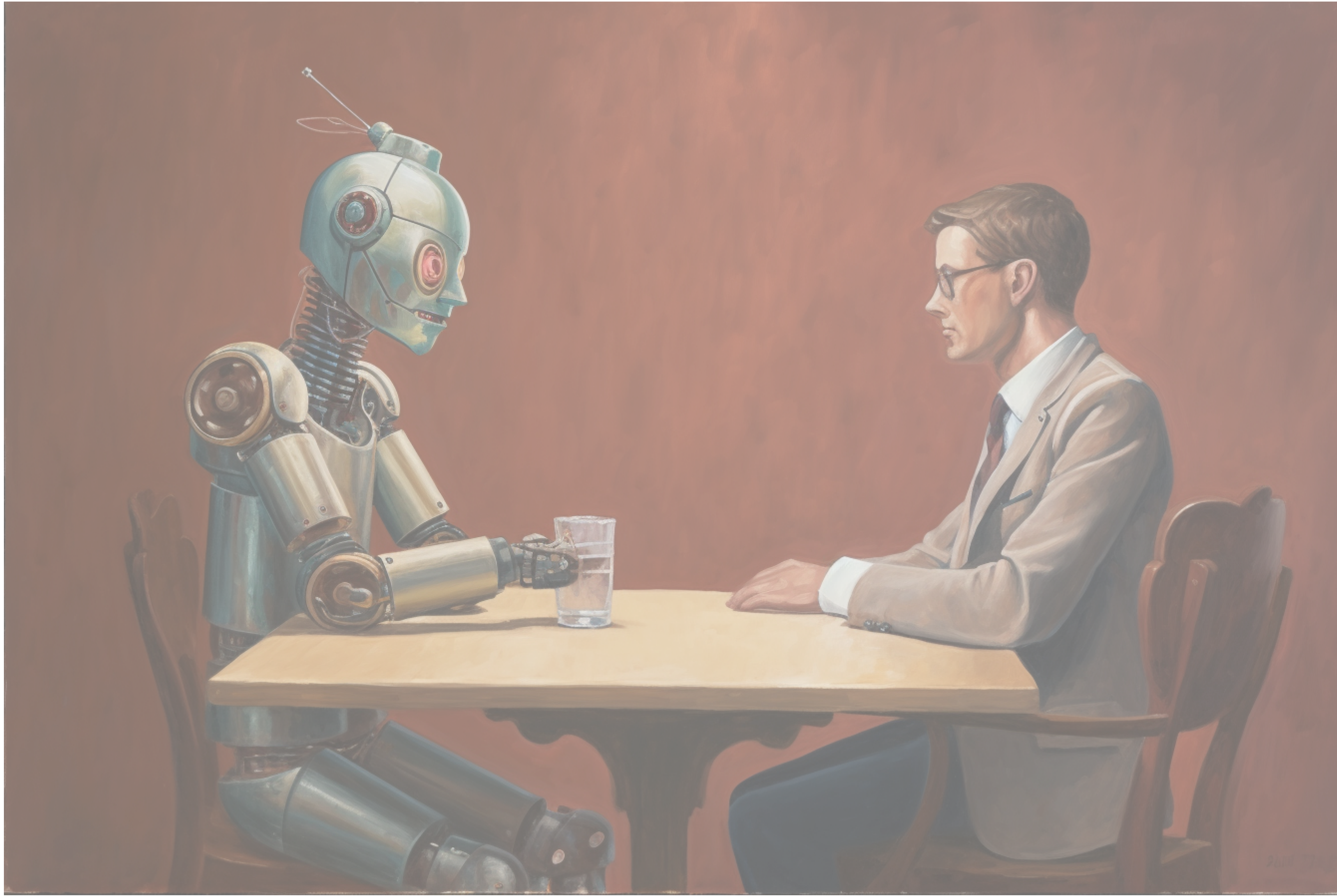
DataRecipe

How to Cook Data for CodeLLM?

ASE'24@Sacramento

Oct 30 2024

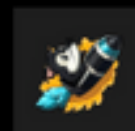
Large Language Model



CodeLlama

-  WizardCoder
-  Phind-CodeLlama
-  OpenAssistant-Codellama
-  NexusRaven

Mistral

-  Mistralic
-  SQLCoder-7B
-  Mistral-OpenOrca

StarCoder

-  SQLCoder-15B
-  WizardCoder-15B

Llama2

-  OpenChat
-  CodeLlama
-  Nous-Hermes
-  Vicuna
-  NSQL
-  Redmond-Puffin
-  Llama2-32K
-  OpenAssistant-Llama2
-  WizardLM

GPT

-  GPT-3.5-Turbo
-  GPT-4

Stable

-  StableLM
-  StableCode
-  OpenAssistant-StableLM

Pythia

-  Pythia-ChatBase
-  OpenAssistant-Pythia

Claude

-  Claude-2

PaLM

-  PaLM-2

Falcon

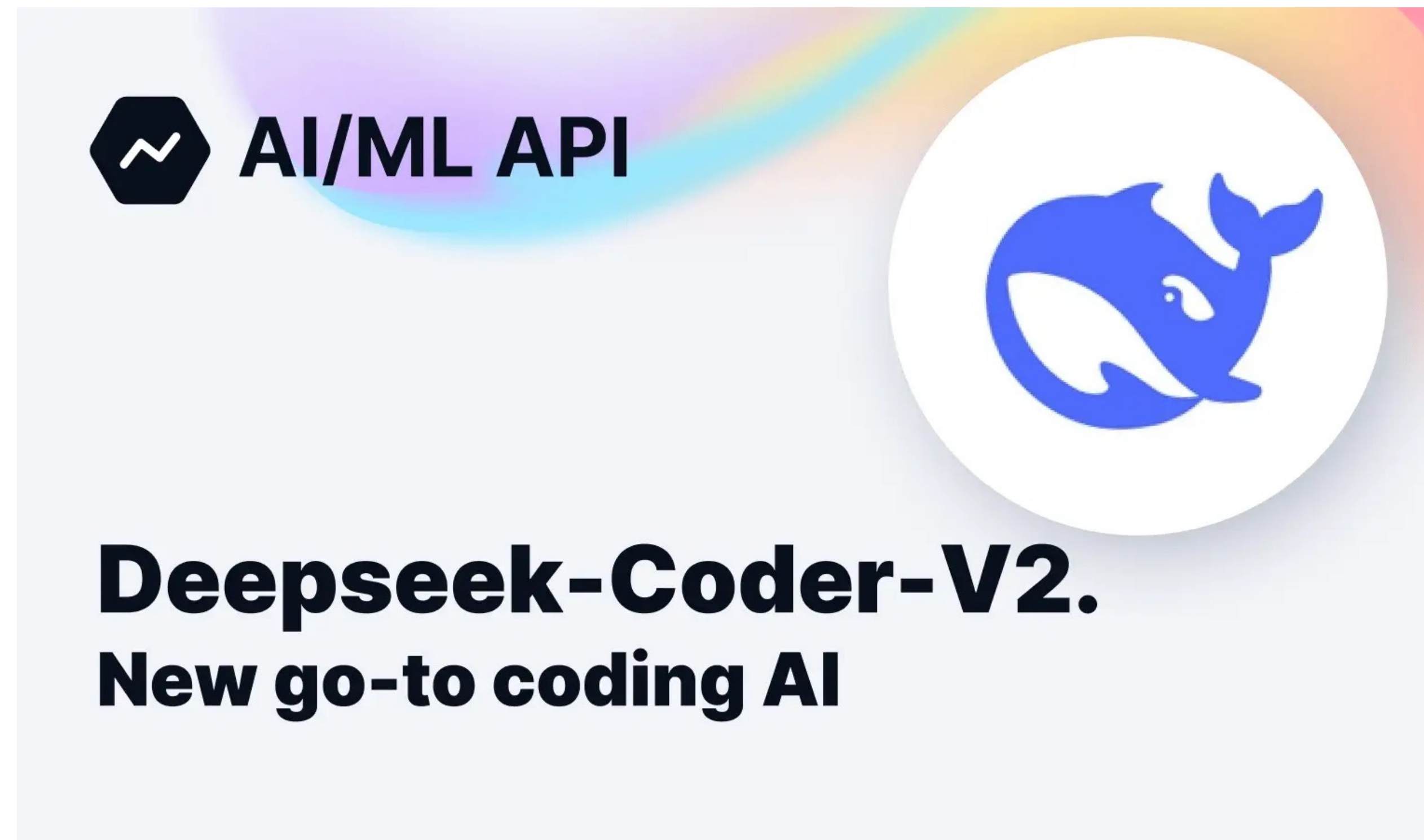
-  Falcon

Qwen

-  Qwen

Map of LLMs

What is a CodeLLM?

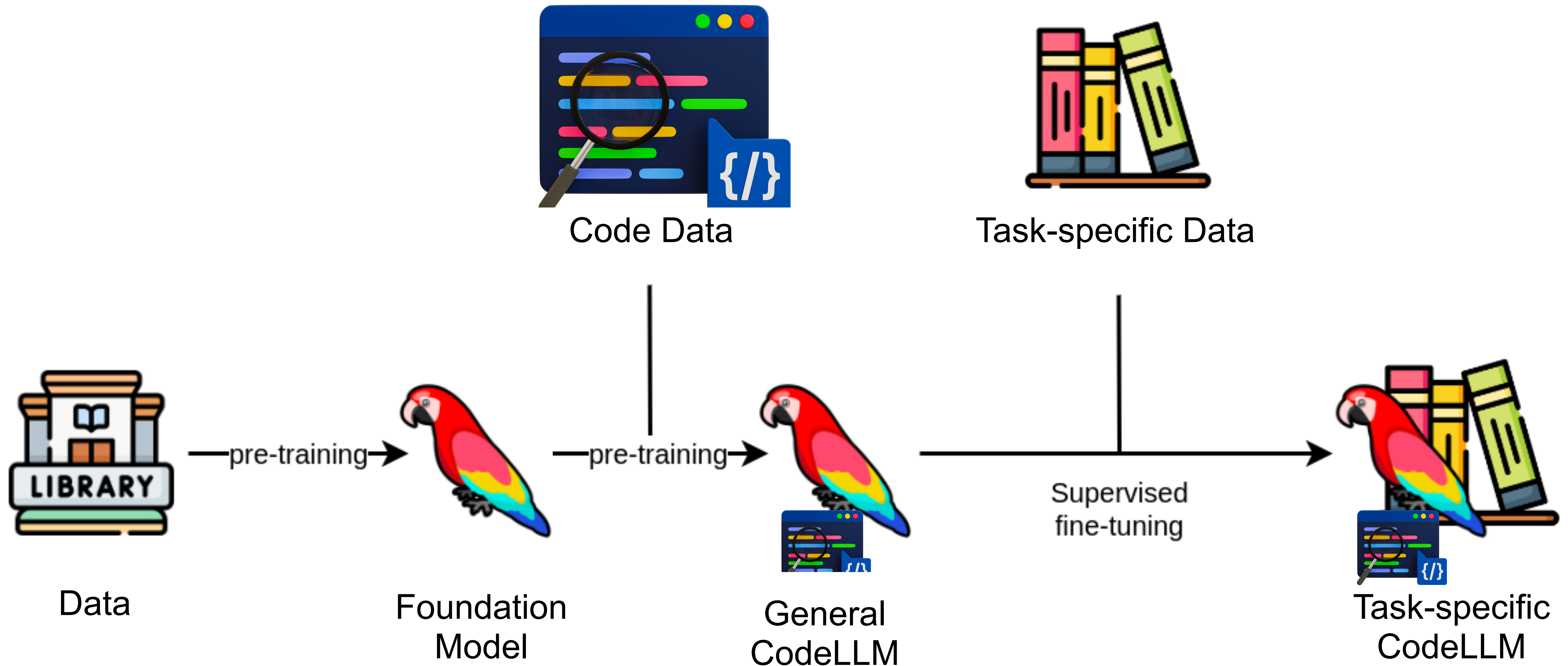


CodeLLM refers to an LLM that can support coding tasks.

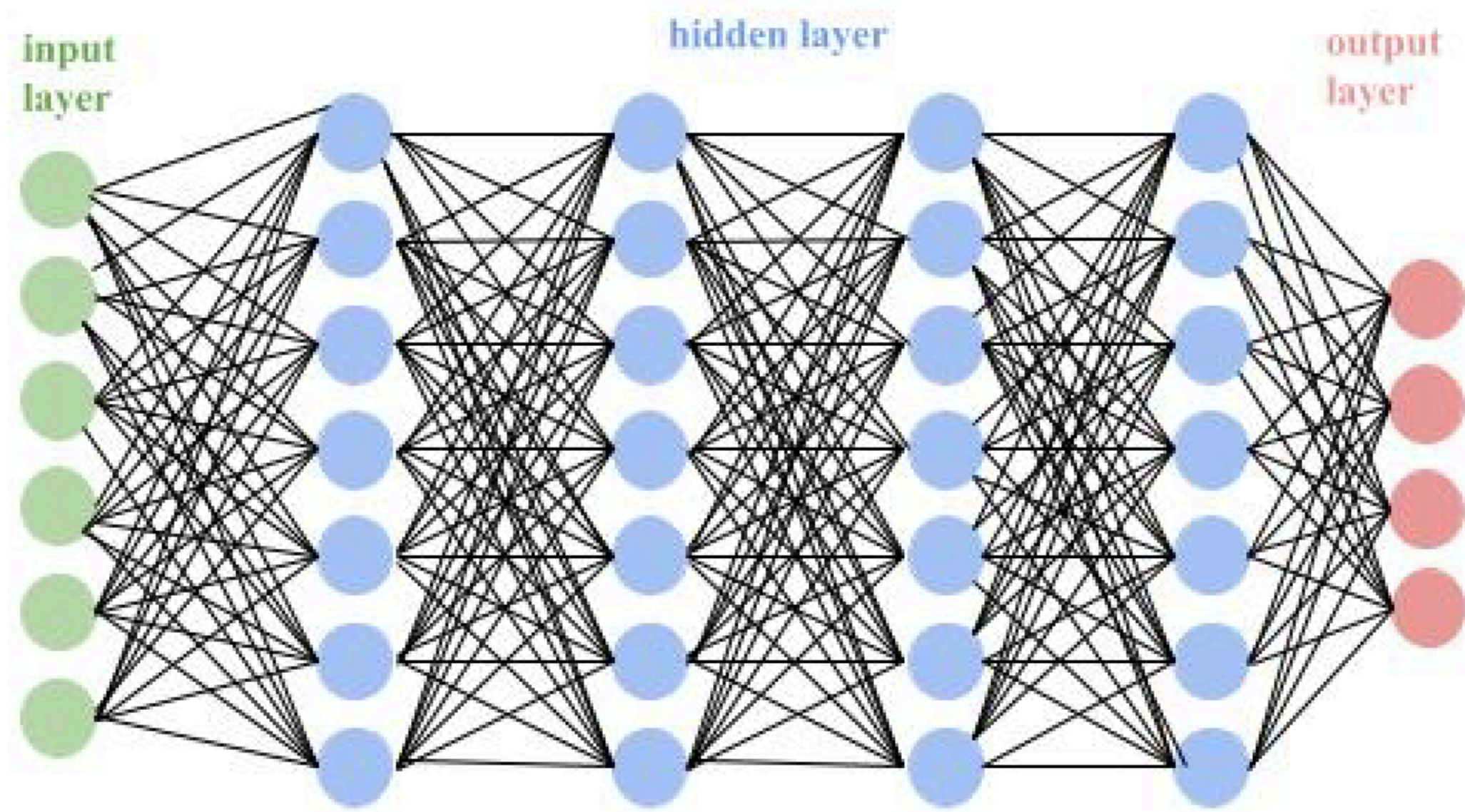
- Code Generation
- Code Review
- Code Vulnerability Detection

...

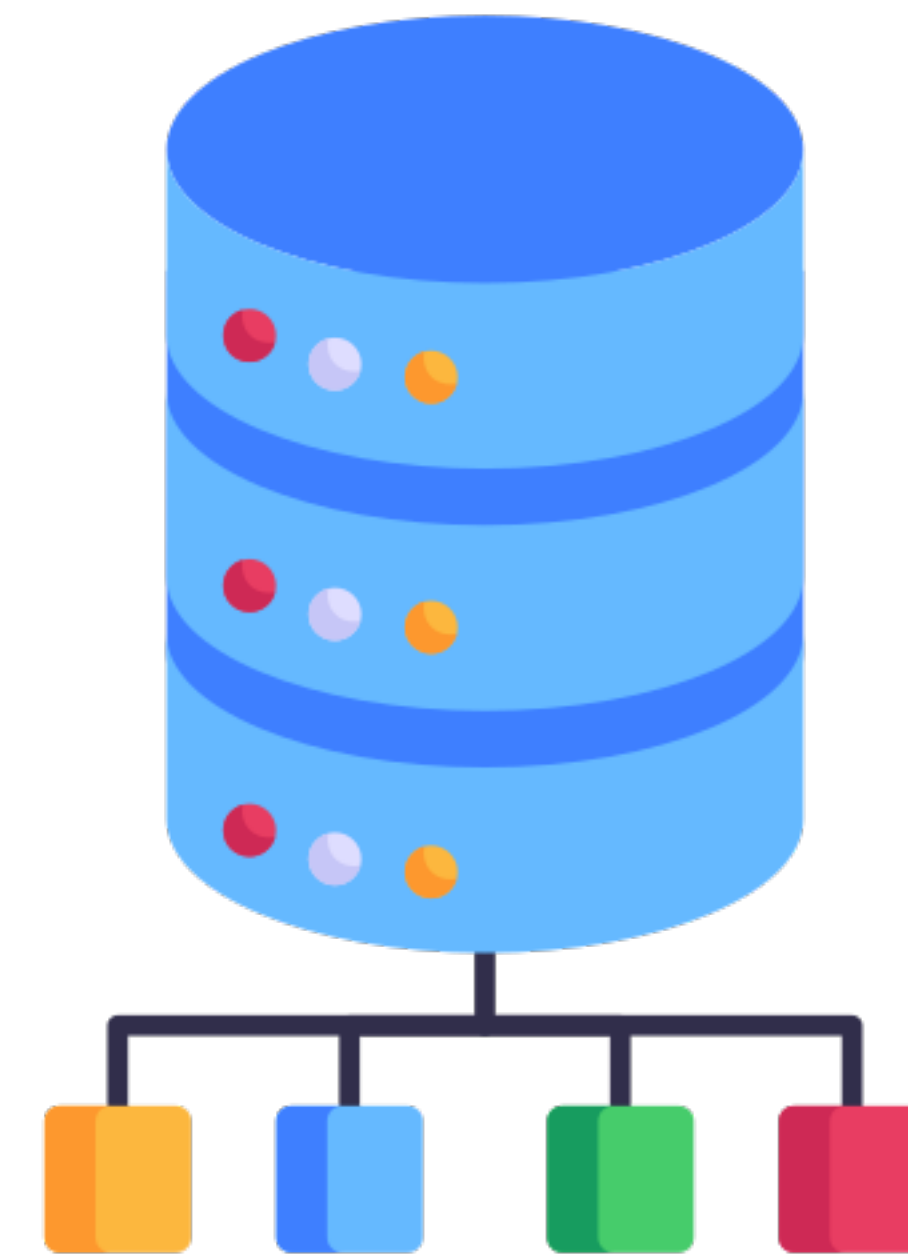
How to build a CodeLLM?



What are the keys in a CodeLLM?

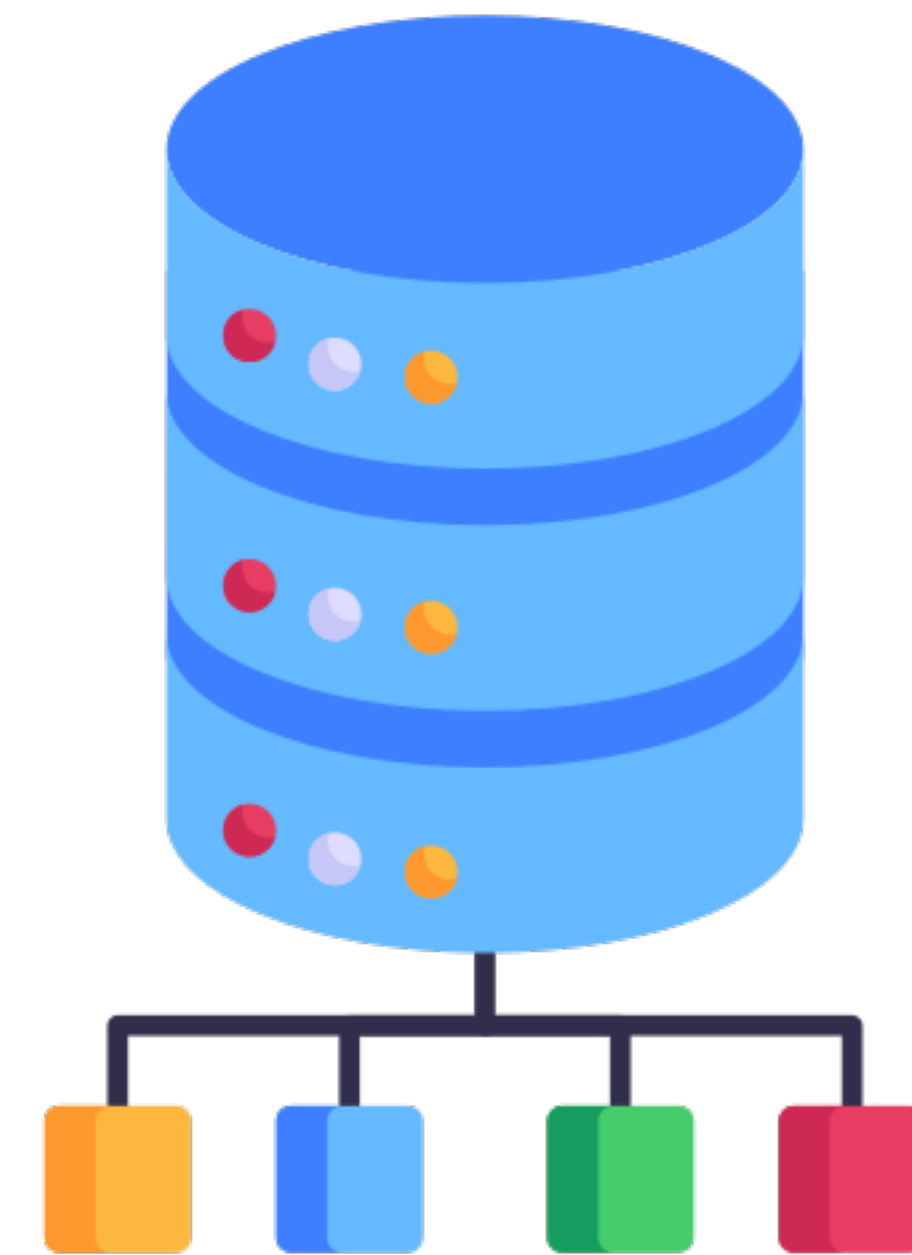


Architecture



Data

What are the keys in a CodeLLM?



Data

What can be the Data for a CodeLLM?



Code Generation

<Query, Code>

Code Review

<Review, Code>

Code Completion

<pre-Code, post-Code>

Code Search

<Query, Code>

Vulnerability Detection

<Code, Y/N>

Code Summarization

<Summary, Code>



What is the
high-quality data?

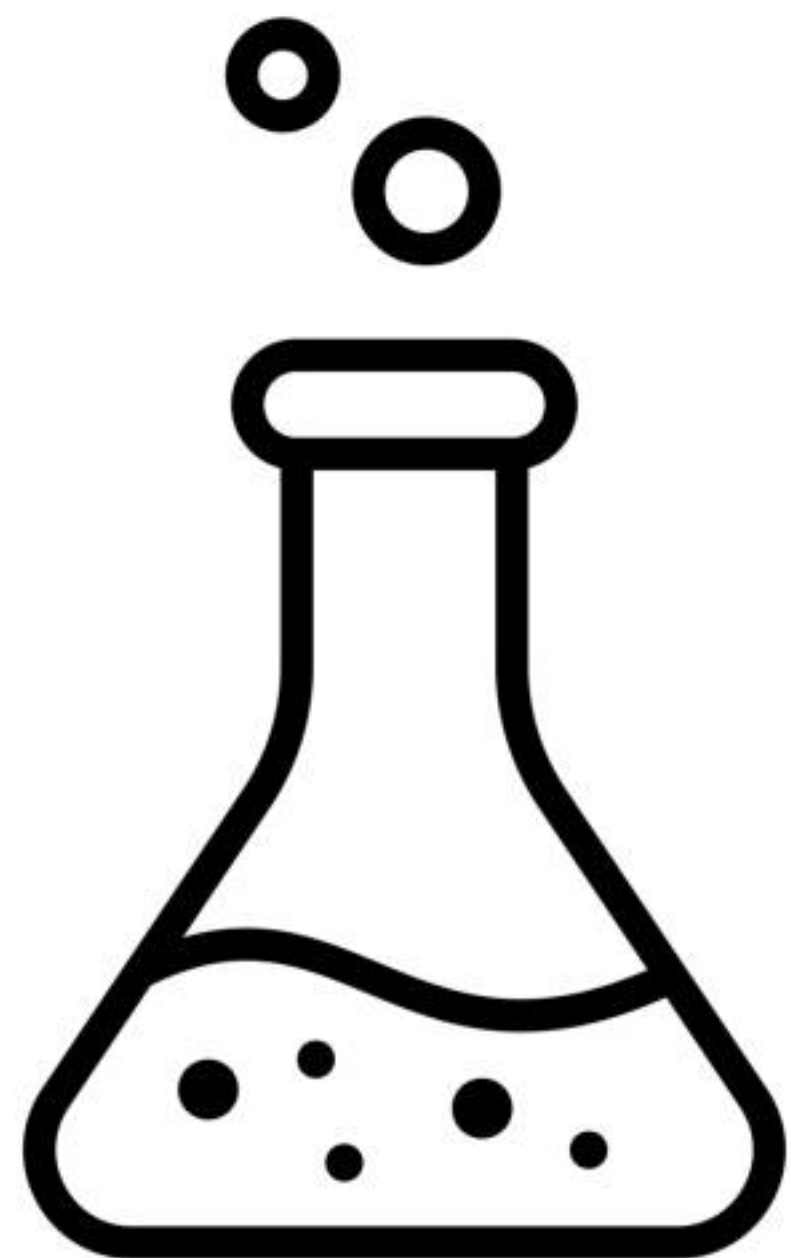
Nobody reveals because...





Let's discover
the recipe ourselves

The ingredients?



Target Task



Datasets



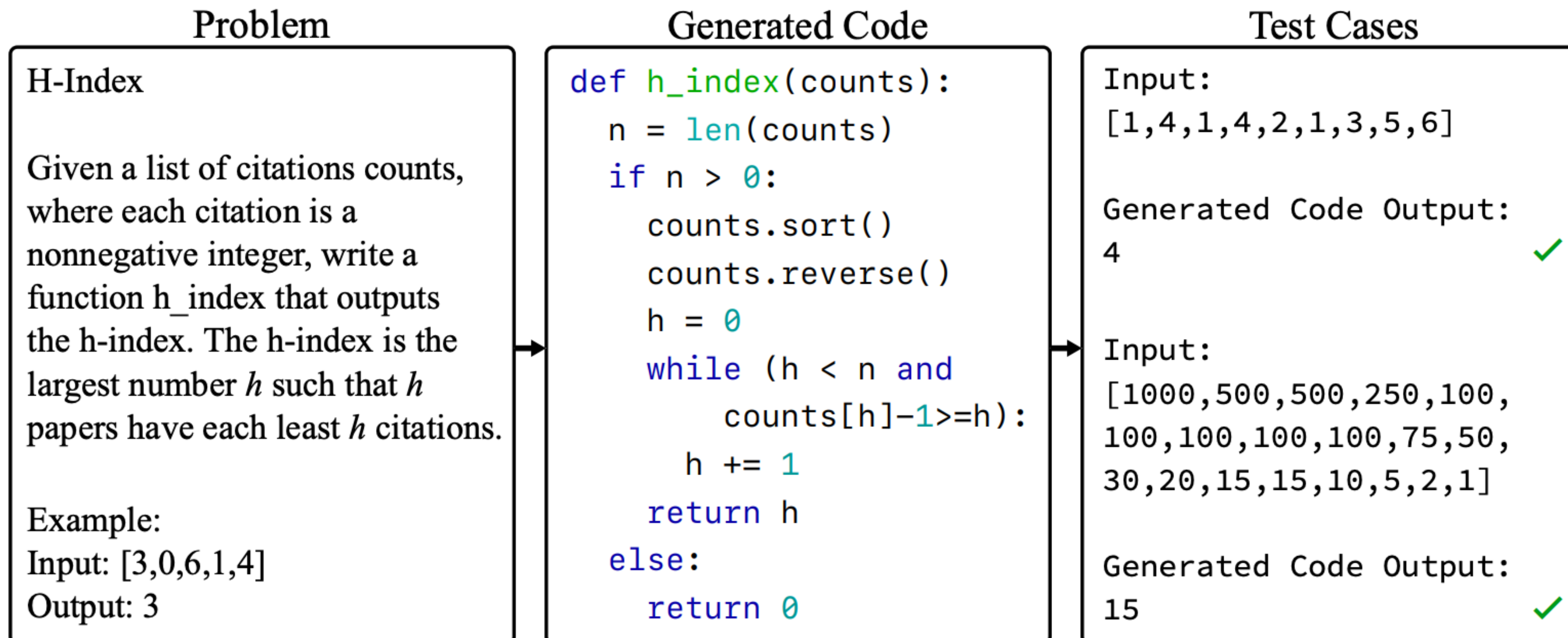
Models

What task do we target?

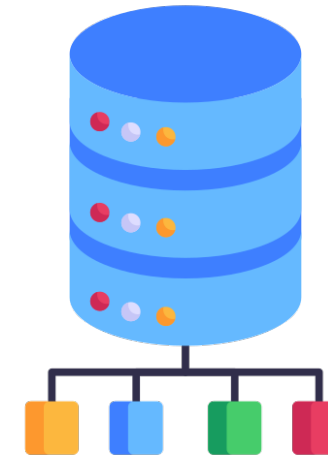


Code Generation

Data format: <Query, Code>



Do we have data?



MBPP

MBPP

Write a function to find the median of two sorted arrays of the same size.

```
def get_ludic(n):
    ludics = []
    for i in range(1, n + 1):
        ludics.append(i)

    ...
    return ludics
```

CoNaLa

CoNaLa

get a list of items from nested list 'li' where third element of each item contains string 'ar'.

```
[x for x in li if 'ar' in x[2]]
```

CodeAlpaca

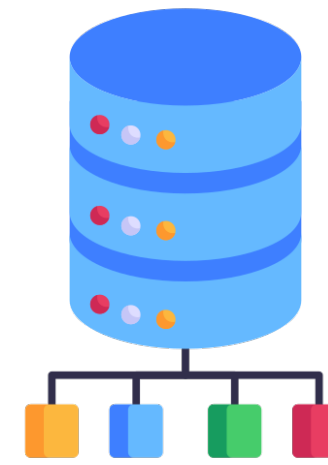
CodeAlpaca

Create an API using Django Rest Framework to retrieve data from a PostgreSQL database table.

```
import viewsets, routers from myapp.models
import MyModel
class MyModelViewSet(viewsets.ModelViewSet):
    ...
    router = routers.DefaultRouter()
    router.register(r'mymodel', MyModelViewSet)
```

ODEX

Do we have data?



MBPP

CoNaLa

CodeAlpaca

DS-1000

ODEX

DS-1000

Problem: How can I know the (row, column) index of the minimum of a numpy array/matrix? For example, if `A = array([[1, 2], [3, 0]])`, I want to get (1, 1) Thanks!

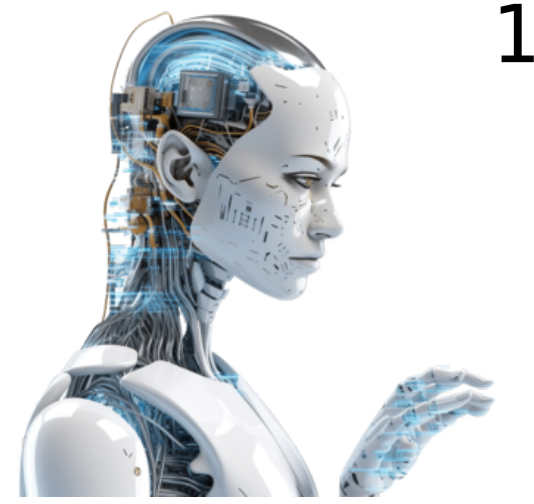
A: `<code> import numpy as np a = np.array([[1, 2], [3, 0]]) </code> result = ... # put solution in this variable BEGIN SOLUTION <code>.`

```
import numpy as np
def generate_test_case(test_case_id):
    ...
def exec_test(result, ans):
    ...
def test_execution(solution: str):
    ...
    assert exec_test(test_env["result"], expected_result)
```

ODEX

convert items in 'T1' to integers.

```
def f_642154(T1):
    return tuple(map(int, T1))
...
def f_642154(T1):
    return (T1 * 1).astype(int)
```



What are the best known models?

General CodeLLM

Normal-sized Language Model



CodeGen - 350M



StarCoder - 1B



DeepSeek-Coder - 1.3B

Large Language Model



DeepSeek-Coder - 6.7B



MagiCoder - 6.7B



CodeLlama - 7B

Foundation Model

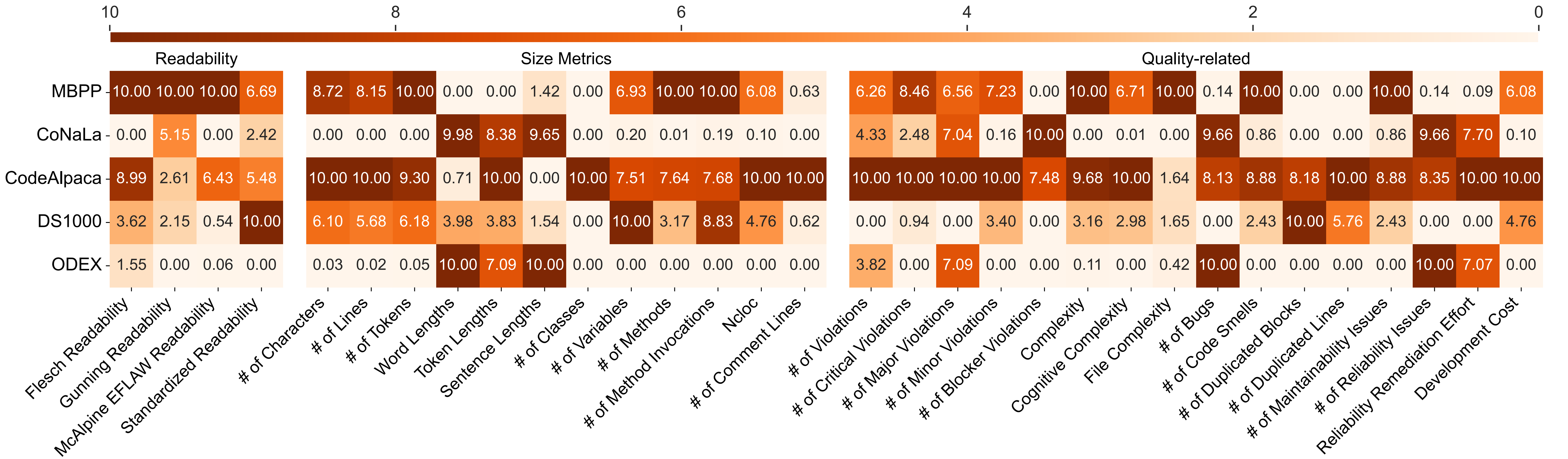
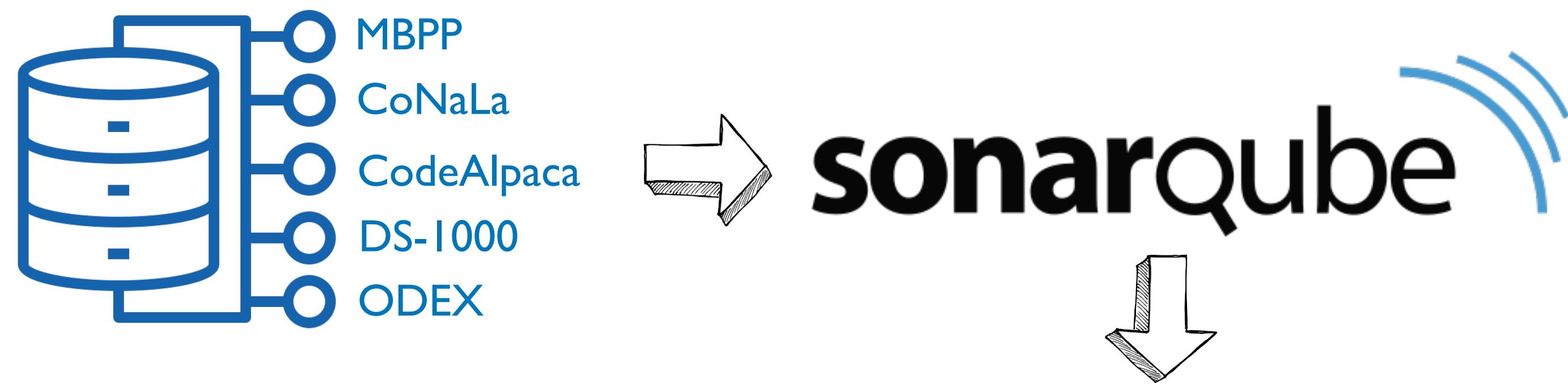


Llama-2 - 7B



Before we cook

Preliminary Study - Basic Characteristics



Preliminary Study - Semantic Correlation



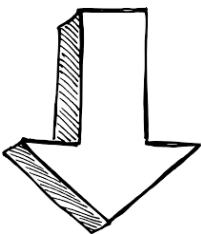
ChatGPT-4

[1] Python Exercise Dataset

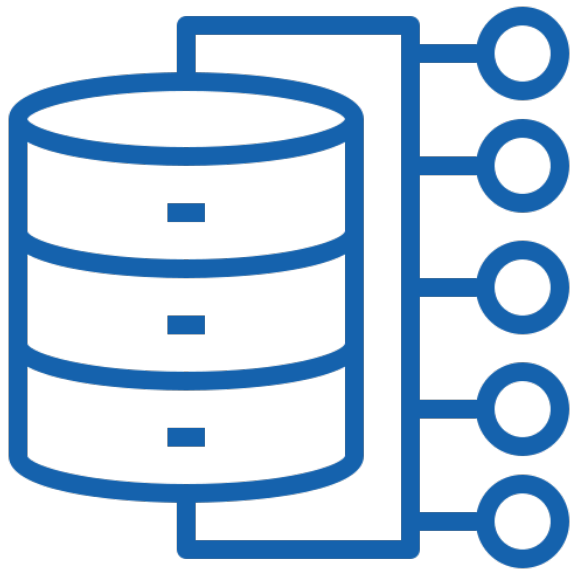
THE CODEEXERCISES DATASET

This is a small *synthetic exercises* dataset consisting of less than 180M tokens of Python exercises and solutions. Each exercise is a docstring of a function that needs to be completed. The goal of this dataset is to align the model to perform function completion tasks based on natural language instructions. This dataset was also generated by GPT-3.5, where the main means of eliciting diversity is by constraining the function names. For this dataset in particular, we conduct explicit decontamination and alternative evaluations in the following sections to ensure that problems similar to those from HumanEval benchmark are not seen during finetuning. Example exercise:

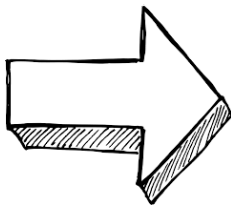
```
def valid_guessing_letters(word: str, guesses: List[str]) -> List[str]:
    """
    Returns a list of valid guessing letters, which are letters that have not been guessed
    yet and are present in the word.
    Parameters:
    word (str): The word to guess.
    guesses (List[str]): A list of letters that have already been guessed.
    Returns:
    List[str]: A list of valid guessing letters.
    """
    valid_letters = []
    for letter in word:
        if letter not in guesses and letter not in valid_letters:
            valid_letters.append(letter)
    return valid_letters
```



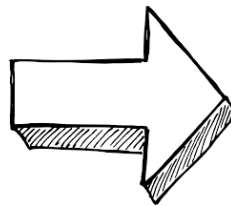
742 <Query, Code> pairs



- MBPP
- CoNaLa
- CodeAlpaca
- DS-1000
- ODEX



deepseek
DeepSeek-Coder-
6.7B-Instruct



Dataset	Semantic Correlation
MBPP _{train}	83.94
CoNaLa _{train}	91.71
CodeAlpaca _{train}	81.74
DS-1000 _{train}	79.92
ODEX _{train}	91.08

[1] Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O. and Salim, A., 2023. **Textbooks are all you need.** *arXiv preprint arXiv:2306.11644*.

Preliminary Study - Design Difference

Write a function to find the median of two sorted arrays of the same size.

```
def get_ludic(n):
    ludics = []
    for i in range(1, n + 1):
        ludics.append(i)
    ...
    return ludics
```

get a list of items form nested list 'li' where third element of each item contains string 'ar'.

```
[x for x in li if 'ar' in x[2]]
```

Create an API using Django Rest Framework to retrieve data from a PostgreSQL database table.

```
import viewsets, routers from myapp.models
import MyModel
class MyModelViewSet(viewsets.ModelViewSet):
    ...
    router = routers.DefaultRouter()
    router.register(r'mymodel', MyModelViewSet)
```

Problem: How can I know the (row, column) index of the minimum of a numpy array/matrix? For example, if A = array([[1, 2], [3, 0]]), I want to get (1, 1) Thanks!

A: `<code> import numpy as np a = np.array([[1, 2], [3, 0]]) </code> result = ... # put solution in this variable BEGIN SOLUTION <code>.`

```
import numpy as np
def generate_test_case(test_case_id):
    ...
def exec_test(result, ans):
    ...
def test_execution(solution: str):
    ...
    assert exec_test(test_env["result"], expected_result)
```

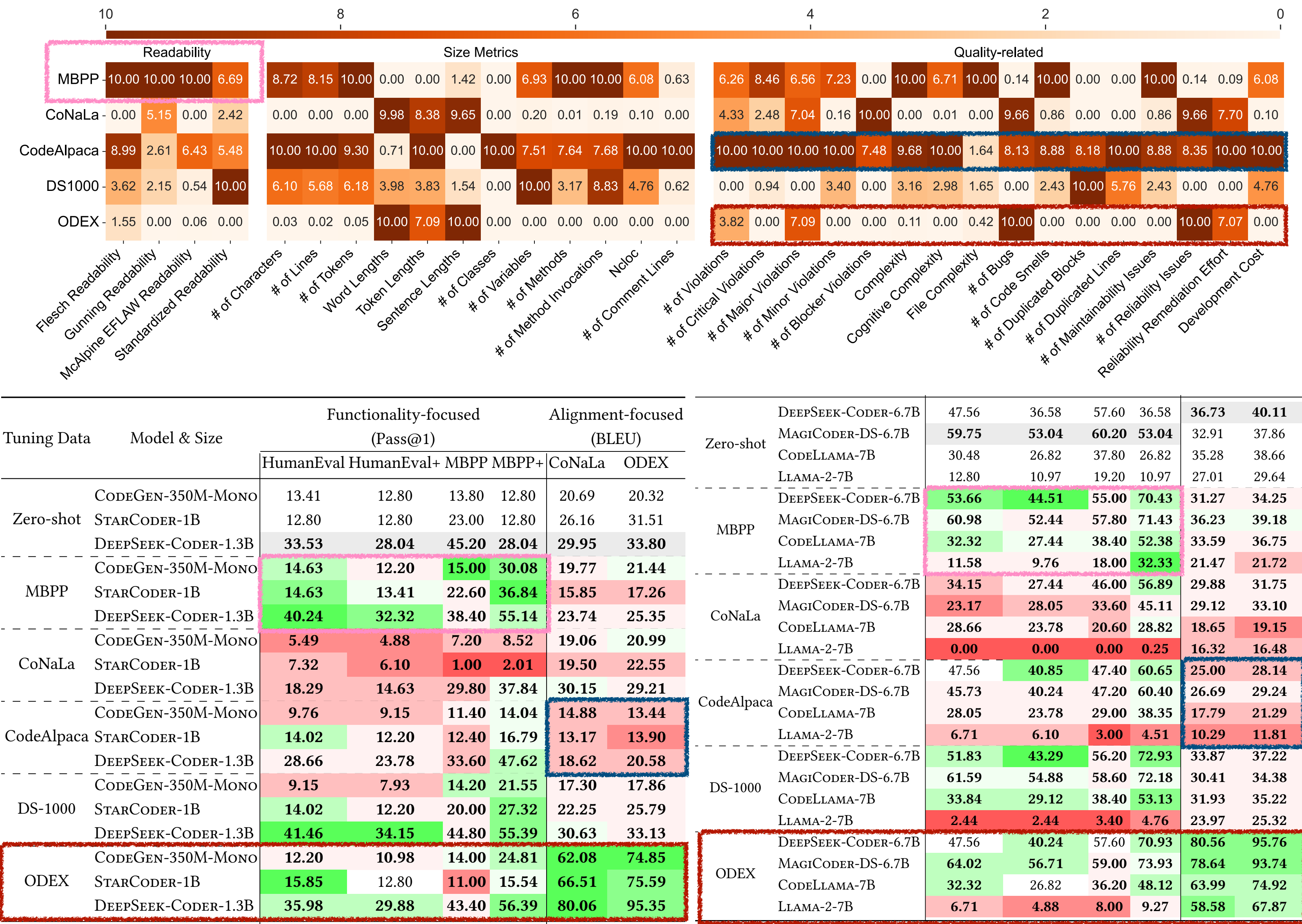
convert items in 'T1' to integers.

```
def f_642154(T1):
    return tuple(map(int, T1))
...
def f_642154(T1):
    return (T1 * 1).astype(int)
```




Let the
cooking begin..

Zero-shot & Initial fine-tuning



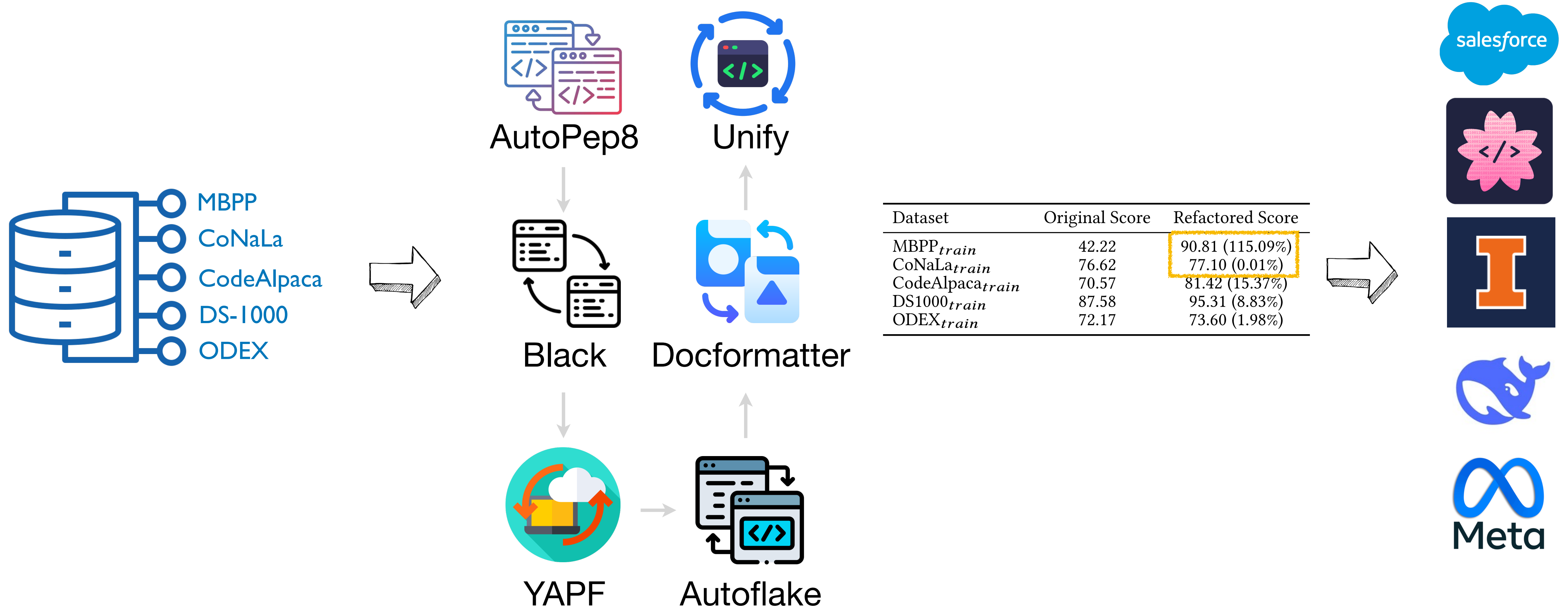
Research Questions

RQ1: Can automating **code**-level **refactoring** on dataset enhance model performance?

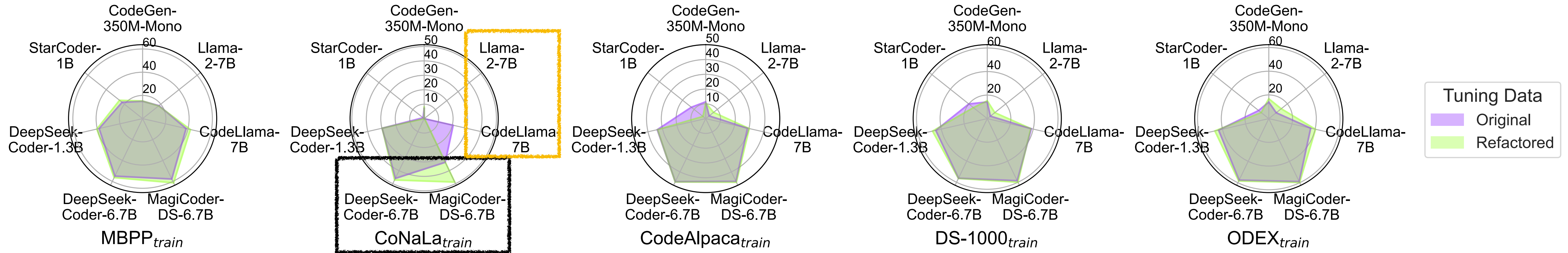
RQ2: How does **refining** pair-level **semantic correlations** on dataset impact model performance?

RQ3: Does **aligning** dataset-level **design** on dataset amplify model performance?

RQ1. Automated Refactoring - Cooking



RQ1. Automated Refactoring - Results



Tuning Data	Model & Size	Functionality-focused (Pass@1)				Alignment-focused (BLEU)	
		HumanEval	HumanEval+	MBPP	MBPP+	CoNaLa	ODEX
MBPP _{refactored}	CODEGEN-350M-MONO	14.63	12.20	15.20	29.57	17.39	18.31
	STARCORDER-1B	16.46	15.24	25.40	42.61	15.12	18.99
	DEEPSEEK-CODER-1.3B	35.98	30.49	40.40	58.90	22.10	26.30
CoNaLa _{refactored}	CODEGEN-350M-MONO	7.32	6.10	8.20	11.53	18.79	19.16
	STARCORDER-1B	8.54	7.32	0.00	0.00	23.69	24.62
	DEEPSEEK-CODER-1.3B	16.46	12.20	29.80	42.61	26.34	28.00
CodeAlpaca _{refactored}	CODEGEN-350M-MONO	11.59	9.76	10.60	13.78	9.41	9.45
	STARCORDER-1B	13.41	10.37	2.40	5.01	8.66	11.28
	DEEPSEEK-CODER-1.3B	28.05	23.17	33.40	46.87	17.52	19.32
DS-1000 _{refactored}	CODEGEN-350M-MONO	12.20	10.37	15.20	20.80	18.79	20.80
	STARCORDER-1B	10.37	9.15	15.00	22.81	21.58	24.96
	DEEPSEEK-CODER-1.3B	36.59	30.49	47.60	57.14	27.53	29.98
ODEX _{refactored}	CODEGEN-350M-MONO	14.02	10.98	16.80	24.31	62.82	71.66
	STARCORDER-1B	14.02	12.80	8.20	11.78	64.09	71.73
	DEEPSEEK-CODER-1.3B	31.71	26.83	46.60	55.39	63.92	71.20

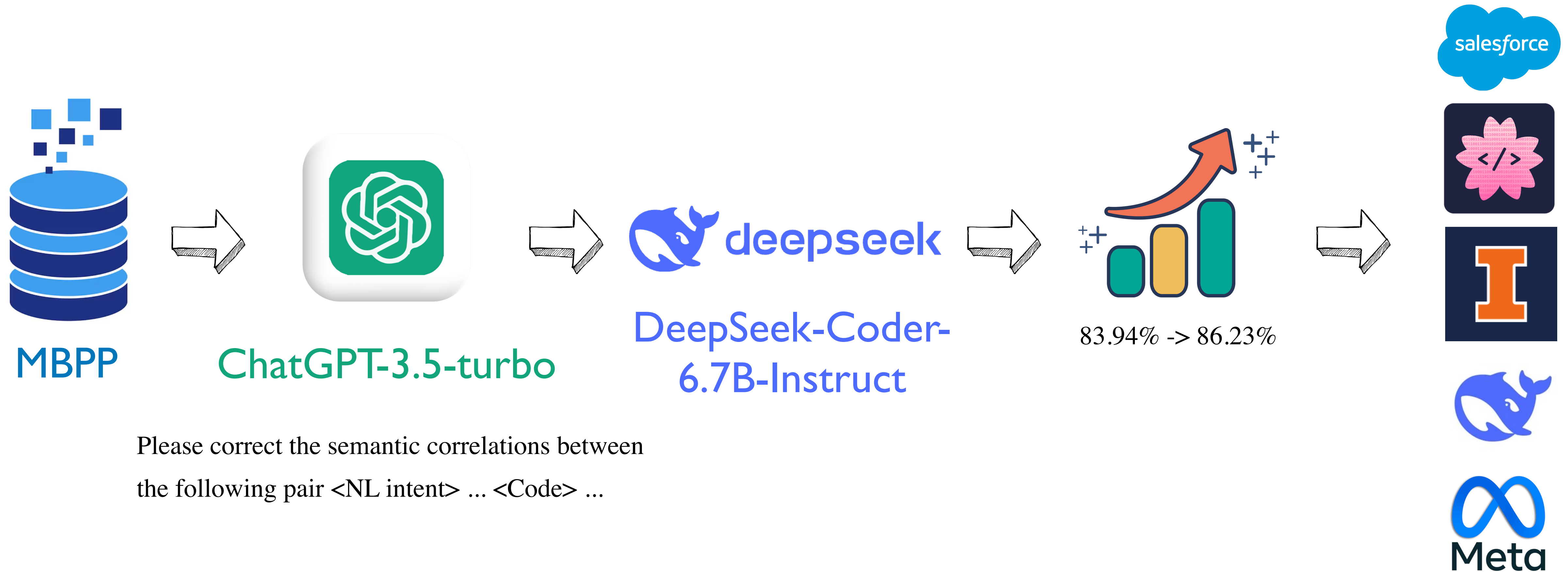
MBPP _{refactored}	DEEPSEEK-CODER-6.7B	51.83	43.29	56.40	69.92	33.04	37.12
	MAGICODER-DS-6.7B	58.54	52.44	61.40	75.19	32.42	34.71
	CODELLAMA-7B	34.15	28.66	42.20	56.14	31.57	36.12
CoNaLa _{refactored}	LLAMA-2-7B	10.98	9.15	17.6	31.08	20.14	22.72
	DEEPSEEK-CODER-6.7B	33.54	30.49	47.20	57.39	31.70	34.68
	MAGICODER-DS-6.7B	42.68	37.80	48.80	61.65	32.71	35.00
CodeAlpaca _{refactored}	CODELLAMA-7B	15.85	14.02	0.40	1.50	23.43	23.45
	LLAMA-2-7B	0.00	0.00	0.00	0.00	15.36	15.60
	DEEPSEEK-CODER-6.7B	42.07	37.80	47.60	58.90	22.61	25.50
DS-1000 _{refactored}	MAGICODER-DS-6.7B	42.07	34.15	47.80	58.65	23.47	26.31
	CODELLAMA-7B	21.95	17.68	30.20	41.35	13.80	15.56
	LLAMA-2-7B	7.32	6.10	6.80	10.28	7.14	8.90
ODEX _{refactored}	DEEPSEEK-CODER-6.7B	50.00	42.68	56.40	72.43	34.20	36.56
	MAGICODER-DS-6.7B	57.93	52.44	60.00	75.19	32.72	35.74
	CODELLAMA-7B	32.32	26.83	37.80	53.38	33.67	37.92
ODEX _{refactored}	LLAMA-2-7B	4.88	3.66	8.20	12.78	23.88	25.60
	DEEPSEEK-CODER-6.7B	47.56	40.85	58.40	68.92	70.35	81.30
	MAGICODER-DS-6.7B	59.76	54.27	59.80	72.93	67.76	78.63
ODEX _{refactored}	CODELLAMA-7B	32.32	26.83	40.40	54.89	55.08	60.78
	LLAMA-2-7B	9.76	6.10	13.40	20.05	51.57	58.17



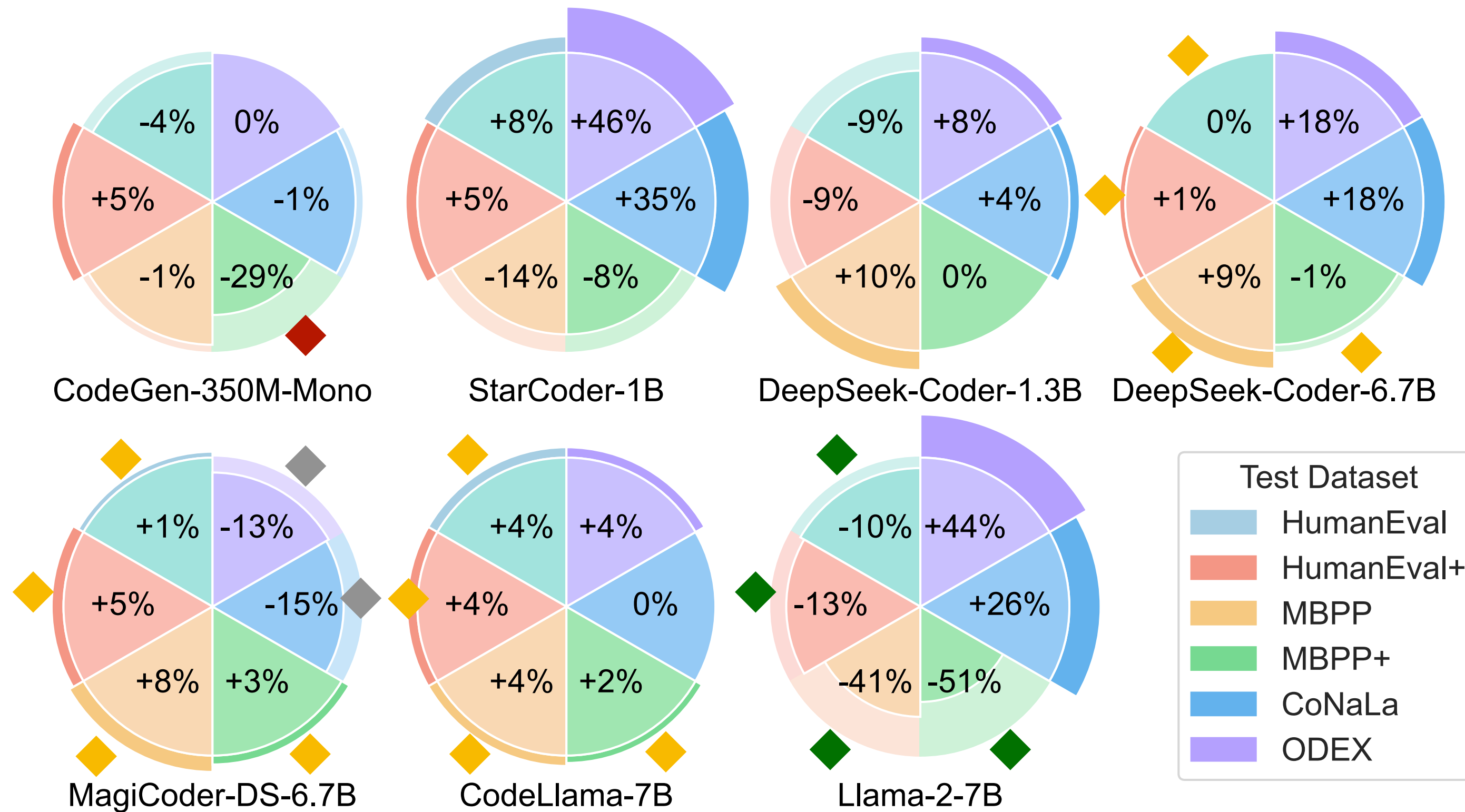
Automated refactoring can support certain combinations.

Data characteristics should be carefully considered for a CodeLLM.

RQ2. Semantic Correlational Correction - Cooking



RQ2. Semantic Correlational Correction - Results



Tuning Data	Model & Size	Functionality-focused (Pass@1)				Alignment-focused (BLEU)	
		HumanEval	HumanEval+	MBPP	MBPP+	CoNaLa	ODEX
MBPP _{semantic}	CODEGEN-350M-MONO	14.02	12.80	14.80	21.30	19.59	21.51
	STARCODER-1B	15.85	14.02	19.40	33.83	21.45	25.19
	DEEPSEEK-CODER-1.3B	36.59	29.27	42.40	55.14	24.78	27.41
	DEEPSEEK-CODER-6.7B	53.66	45.12	60.20	69.67	36.81	40.34
	MAGICODER-DS-6.7B	61.59	54.88	62.20	73.93	30.70	34.17
	CODELLAMA-7B	33.54	28.66	39.80	53.63	33.67	38.14
	LLAMA-2-7B	10.37	8.54	10.60	15.79	26.98	31.22

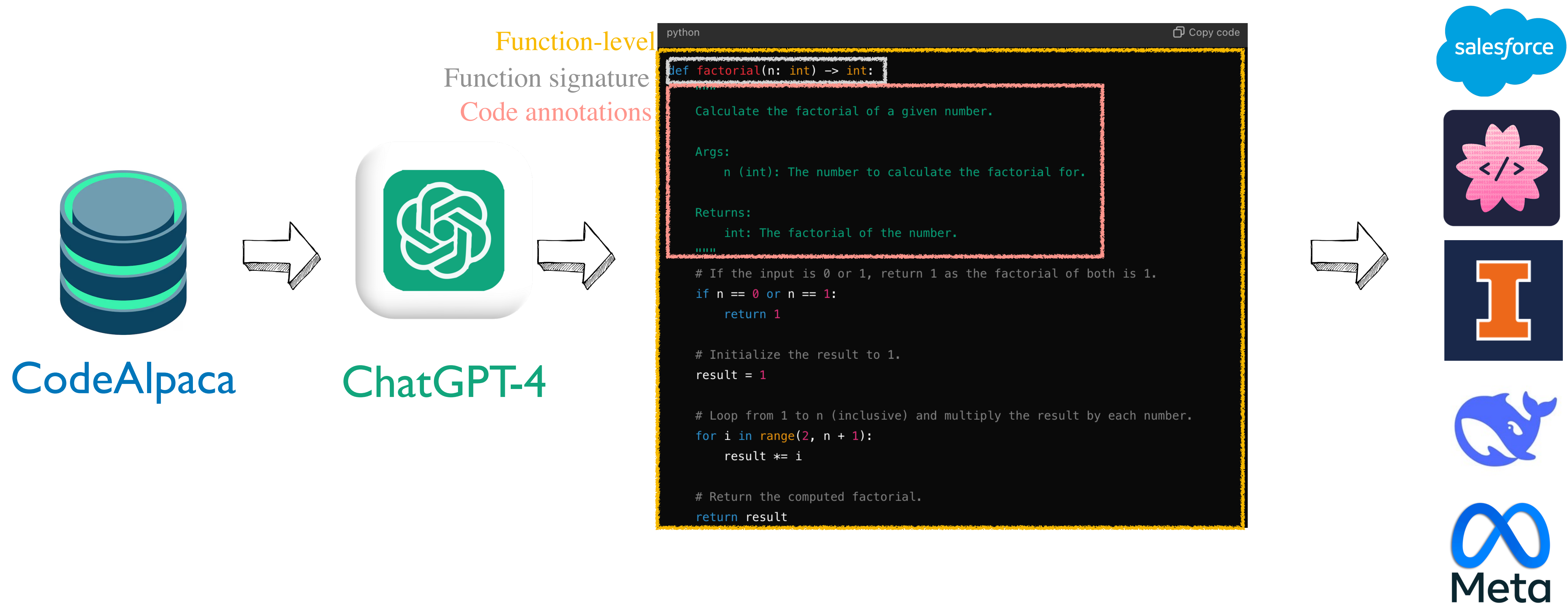
Tuning Data	Model & Size	Functionality-focused (Pass@1)				Alignment-focused (BLEU)	
		HumanEval	HumanEval+	MBPP	MBPP+	CoNaLa	ODEX
MBPP _{semantic}	CODEGEN-350M-MONO	14.02	12.80	14.80	21.30	19.59	21.51
	STARCODER-1B	15.85	14.02	19.40	33.83	21.45	25.19
	DEEPSEEK-CODER-1.3B	36.59	29.27	42.40	55.14	24.78	27.41
	DEEPSEEK-CODER-6.7B	53.66	45.12	60.20	69.67	36.81	40.34
	MAGICODER-DS-6.7B	61.59	54.88	62.20	73.93	30.70	34.17
	CODELLAMA-7B	33.54	28.66	39.80	53.63	33.67	38.14
	LLAMA-2-7B	10.37	8.54	10.60	15.79	26.98	31.22

It was generally better than Automatic Refactoring.

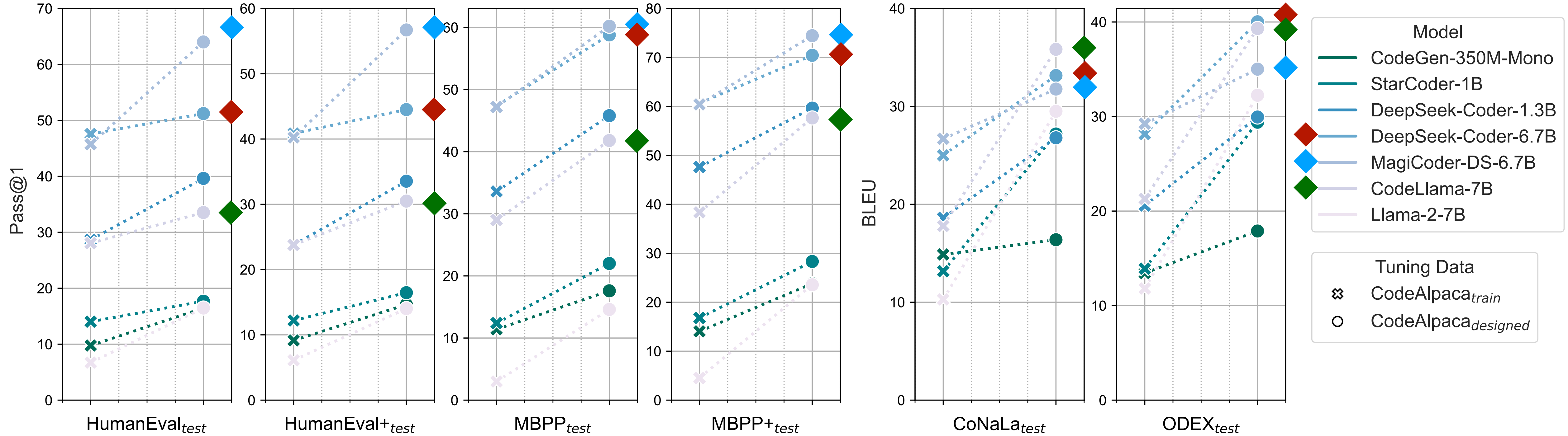
However, we can also confirm that model's characteristics needs to be reviewed for training a CodeLLM.



RQ3. Data Design Alignment - Cooking



RQ3. Data Design Alignment - Results



Tuning Data	Model & Size	Functionality-focused (Pass@1)				Alignment-focused (BLEU)	
		HumanEval	HumanEval+	MBPP	MBPP+	CoNaLa	ODEX
CodeAlpaca _{designed}	CODEGEN-350M-MONO	16.46	14.63	17.60	23.81	16.38	17.89
	STARCORDER-1B	17.68	16.46	22.00	28.32	27.18	29.40
	DEEPSEEK-CODER-1.3B	39.63	33.54	45.80	59.65	26.79	29.95
	DEEPSEEK-CODER-6.7B	51.22	44.51	58.80	70.43	33.15	40.02
	MAGICODER-DS-6.7B	64.02	56.71	60.20	74.44	31.77	35.01
	CODELLAMA-7B	33.54	30.49	41.80	57.64	35.83	39.31
	LLAMA-2-7B	16.46	14.02	14.60	23.56	29.50	32.25

From now on..

We train a CodeLLM after aligning the design of input data.

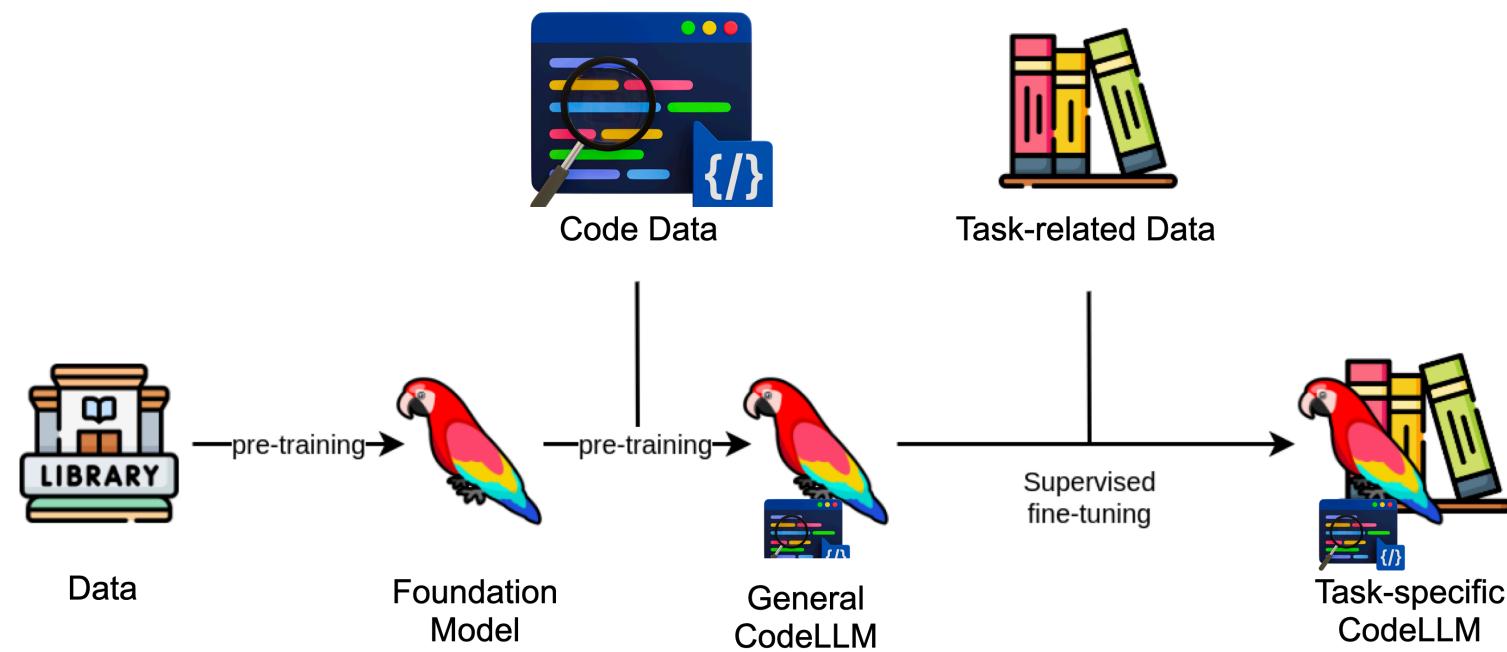
- Function-level
- Function signature
- Code annotation



Summary

5

How to build a CodeLLM?



9

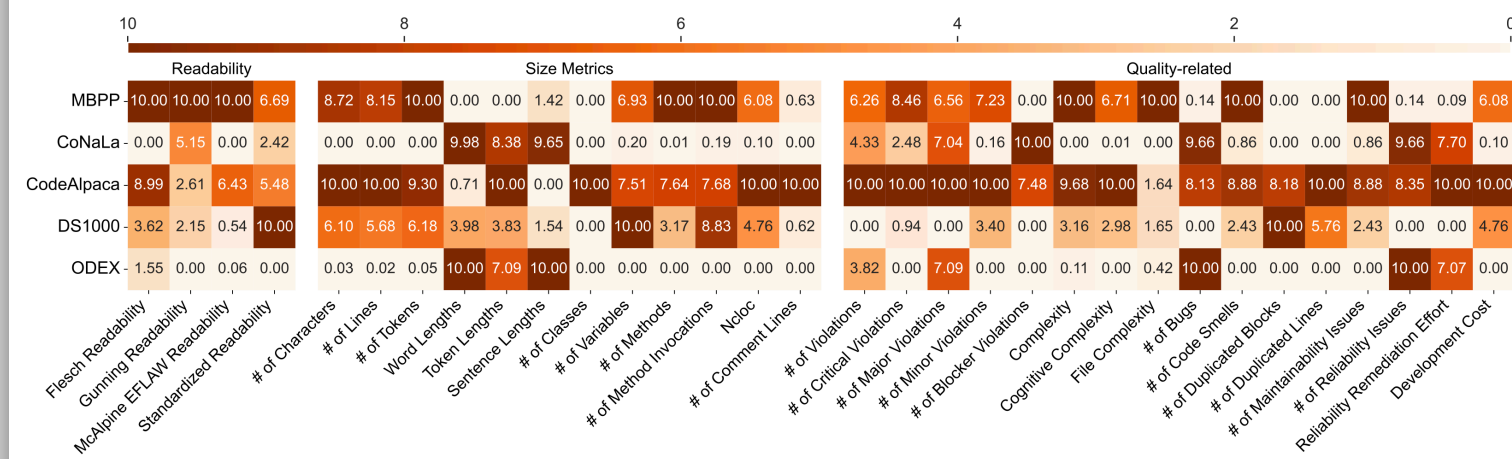
What can be the Data for a CodeLLM?



- Code Generation** <Query, Code>
- Code Review** <Review, Code>
- Code Completion** <pre-Code, post-Code>
- Code Search** <Query, Code>
- Vulnerability Detection** <Code, Y/N>
- Code Summarization** <Summary, Code>

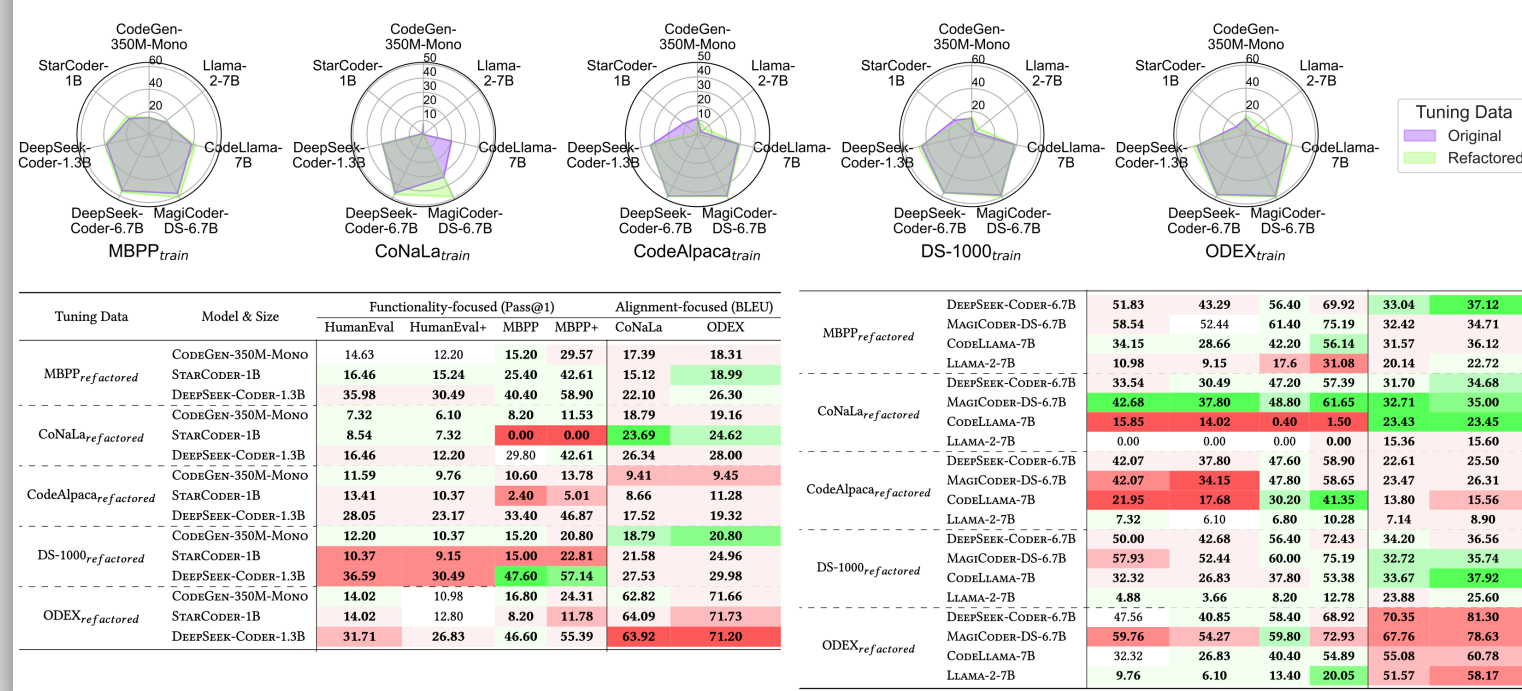
18

Preliminary Study - Basic Characteristics



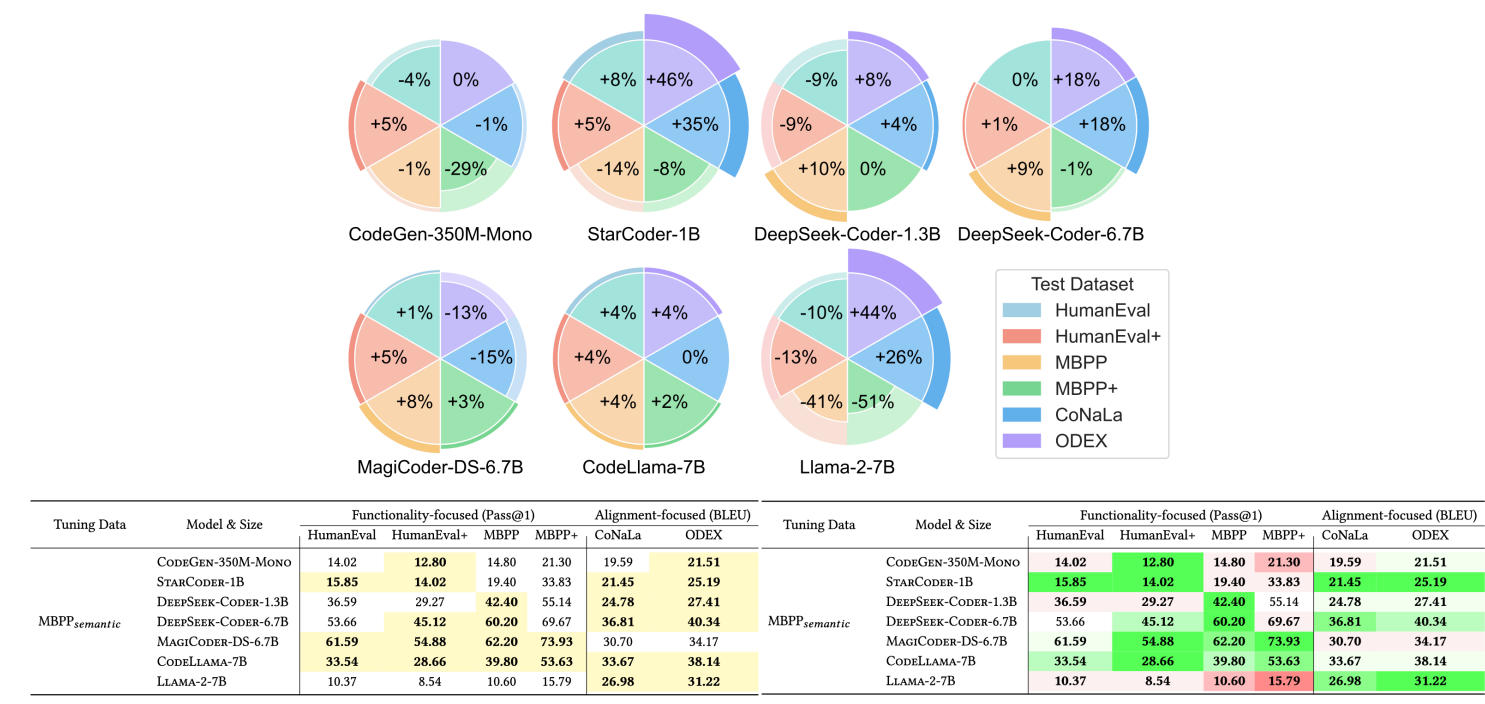
24

RQ1. Automated Refactoring



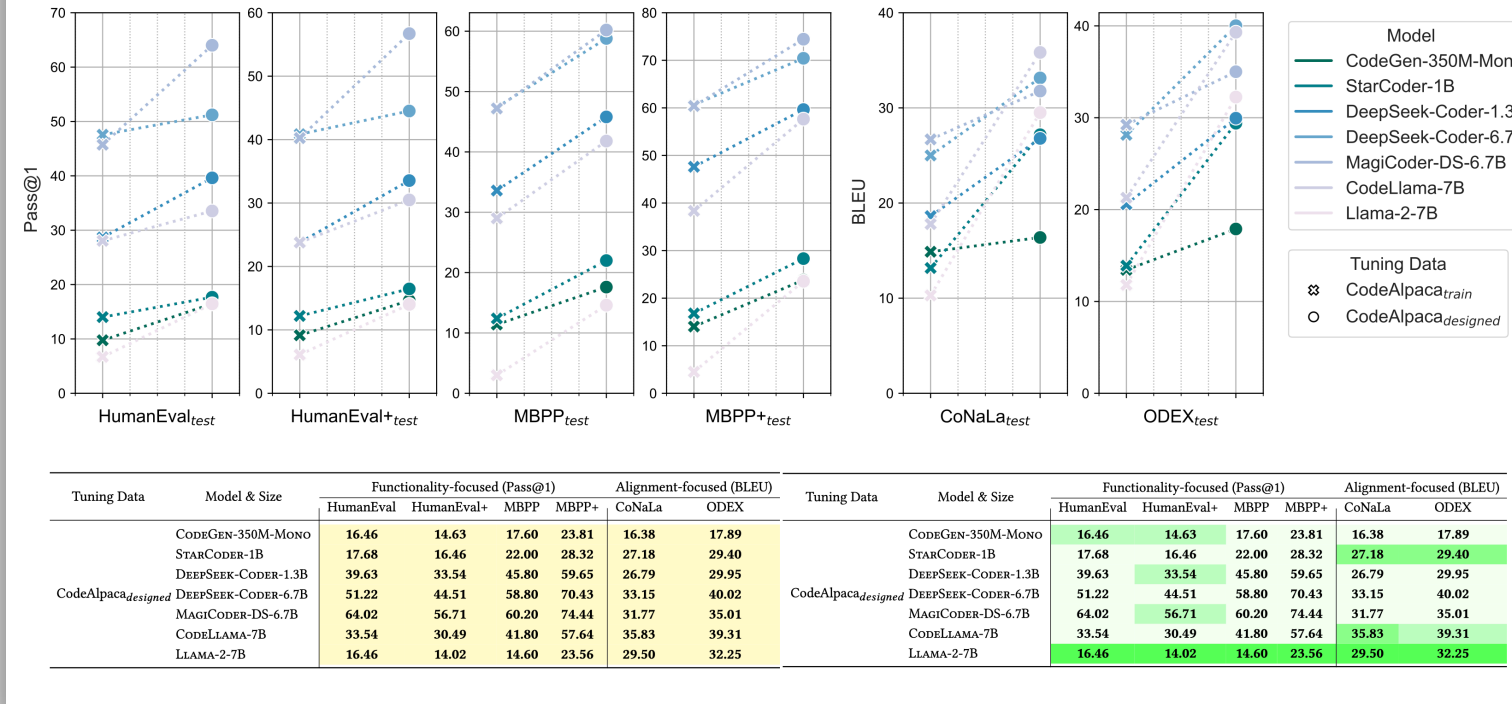
25

RQ2. Semantic Correlational Correction



26

RQ3. Data Design Alignment



Kisub Kim: falconlk00@gmail.com