AiDD AI$^+$ 软件研发数字峰会
AI$^+$ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·深圳 11.24-25

AI$^+$
software
Development
Digital
summit

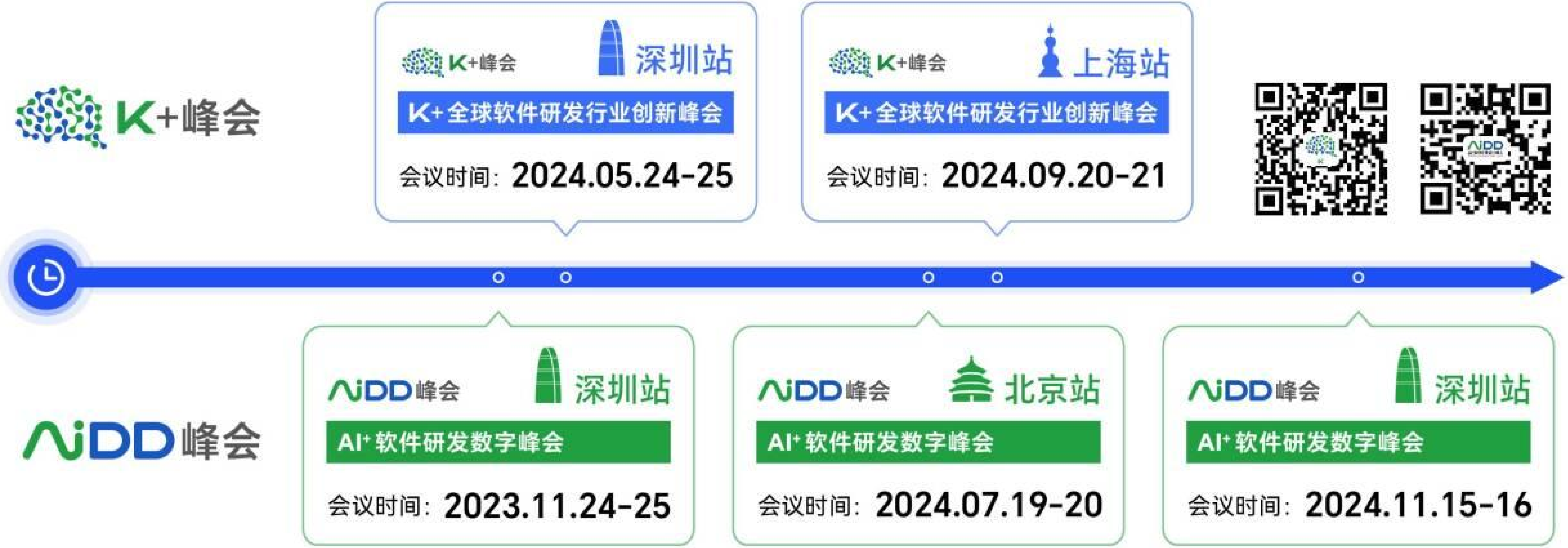
蚂蚁代码大模型的评测实践

申敏 蚂蚁集团

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINKing



K+峰会

K+峰会 深圳站
K+全球软件研发行业创新峰会
会议时间：2024.05.24-25

K+峰会 上海站
K+全球软件研发行业创新峰会
会议时间：2024.09.20-21

AiDD峰会

AiDD峰会 深圳站
AI⁺软件研发数字峰会
会议时间：2023.11.24-25

AiDD峰会 北京站
AI⁺软件研发数字峰会
会议时间：2024.07.19-20

AiDD峰会 深圳站
AI⁺软件研发数字峰会
会议时间：2024.11.15-16

演讲嘉宾

申敏

蚂蚁集团-测试开发专家

蚂蚁集团测试开发专家，研究方向：大模型在代码领域的评测技术。
长期投入蚂蚁支付、账务、计收费等业务领域质量保障工作，熟悉企业级编码风格及要求，当前，负责蚂蚁百灵大模型CodeFuse系列的代码能力评测。

目 录
CONTENTS

1. 前言
2. 代码大模型的评测内容
3. 代码大模型的评测方法
4. 代码大模型的评估基准
5. 代码大模型多任务评估
6. 展望

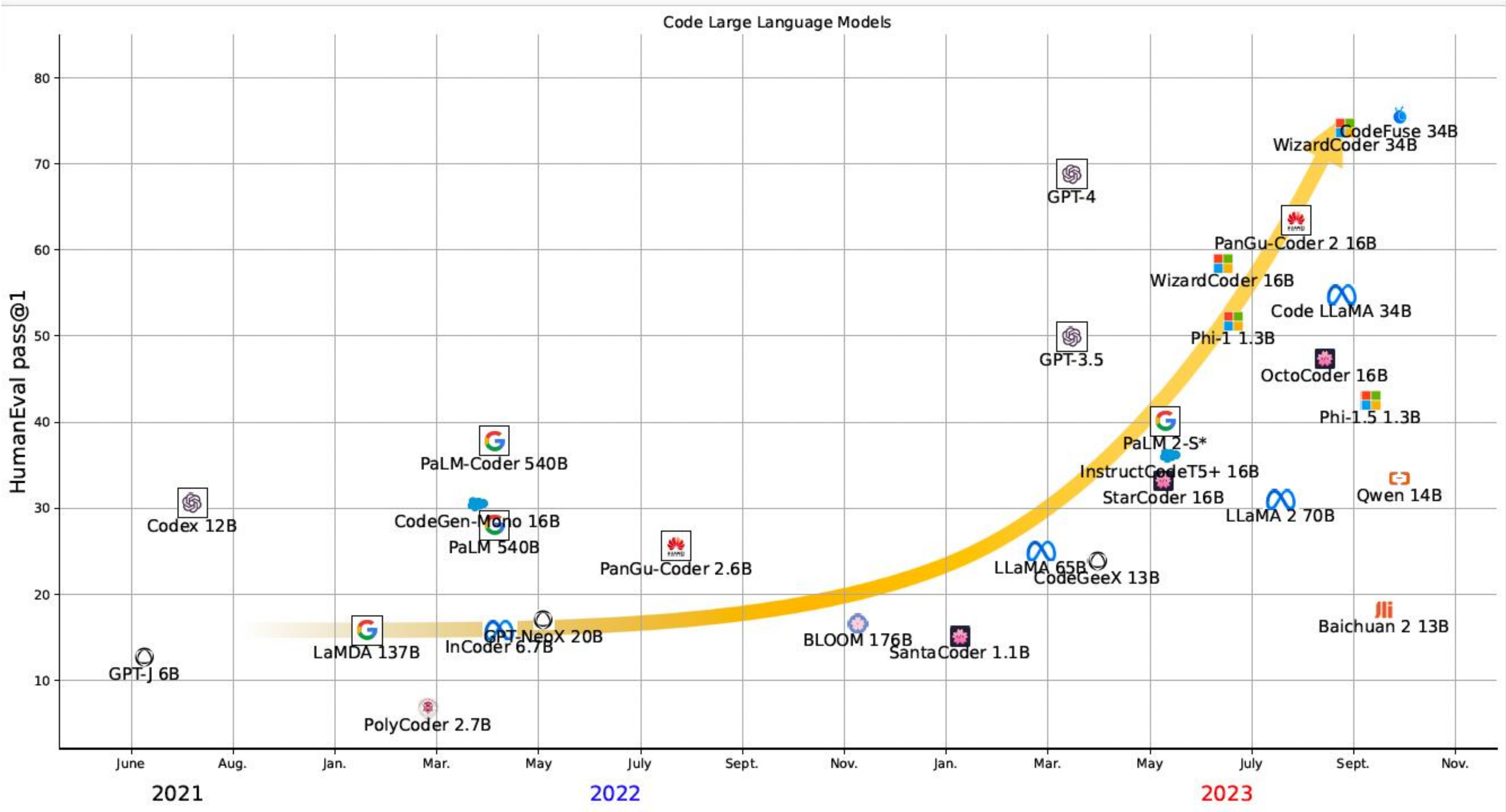
PART 01

前言

前言：模型发展与模型评估

2023年大模型呈爆发式增长，截至2023年7月，中国累计有130个大模型问世，国外大模型138个，其中，美国大模型114个。 [赛迪顾问《IT2023》]

模型发布必然离不开模型评估；AIGC编程是模型落地最为广泛的场景之一，充分衡量方能更好的运用or选用。

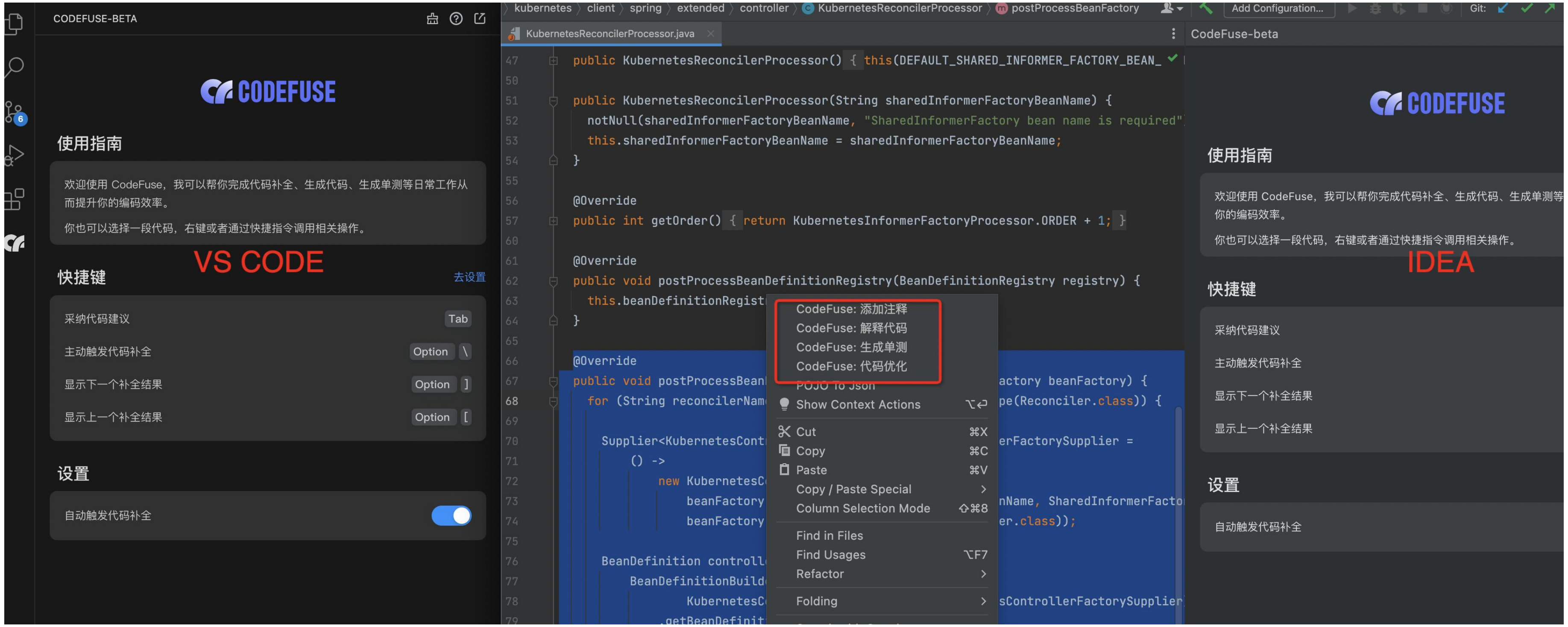


AI驱动**软件研发**全面进入数字化时代

前言：CodeFuse 让研发变的更简单

CodeFuse 是一款为国内开发者提供智能研发服务的产品，该产品是基于蚂蚁集团自研的基础大模型进行微调的**代码大模型**。

CodeFuse 具备**代码补全、添加注释、解释代码、生成单测，以及代码优化等功能**，以帮助开发者更快、更轻松地编写代码。



CODEFUSE-BETA

使用指南

欢迎使用 CodeFuse，我可以帮你完成代码补全、生成代码、生成单测等日常工作从而提升你的编码效率。

你也可以选择一段代码，右键或者通过快捷指令调用相关操作。

VS CODE

快捷键 去设置

采纳代码建议	Tab
主动触发代码补全	Option \
显示下一个补全结果	Option]
显示上一个补全结果	Option [

设置

自动触发代码补全 ●

kubernetes 〉 client 〉 spring 〉 extended 〉 controller 〉 KubernetesReconcilerProcessor 〉 postProcessBeanFactory

KubernetesReconcilerProcessor.java

```
public KubernetesReconcilerProcessor() { this(DEFAULT_SHARED_INFORMER_FACTORY_BEAN_

public KubernetesReconcilerProcessor(String sharedInformerFactoryBeanName) {
  notNull(sharedInformerFactoryBeanName, "SharedInformerFactory bean name is required"
  this.sharedInformerFactoryBeanName = sharedInformerFactoryBeanName;
}

@Override
public int getOrder() { return KubernetesInformerFactoryProcessor.ORDER + 1; }

@Override
public void postProcessBeanDefinitionRegistry(BeanDefinitionRegistry registry) {
  this.beanDefinitionRegist
}

@Override
public void postProcessBean                              actory beanFactory) {
  for (String reconcilerNam                              pe(Reconciler.class)) {

    Supplier<KubernetesCont                              erFactorySupplier =
        () ->
            new KubernetesC
                beanFactory                              nName, SharedInformerFacto
                beanFactory                              er.class));

    BeanDefinition controll
        BeanDefinitionBuild
                KubernetesC                              sControllerFactorySupplier
            .getBeanDefinit
```

CodeFuse: 添加注释
CodeFuse: 解释代码
CodeFuse: 生成单测
CodeFuse: 代码优化
POJO To Json
Show Context Actions ⌥↩
Cut ⌘X
Copy ⌘C
Paste ⌘V
Copy / Paste Special ›
Column Selection Mode ⇧⌘8
Find in Files
Find Usages ⌥F7
Refactor ›
Folding ›

Add Configuration... Git:

CodeFuse-beta

使用指南

欢迎使用 CodeFuse，我可以帮你完成代码补全、生成代码、生成单测等你的编码效率。

你也可以选择一段代码，右键或者通过快捷指令调用相关操作。

IDEA

快捷键

采纳代码建议

主动触发代码补全

显示下一个补全结果

显示上一个补全结果

设置

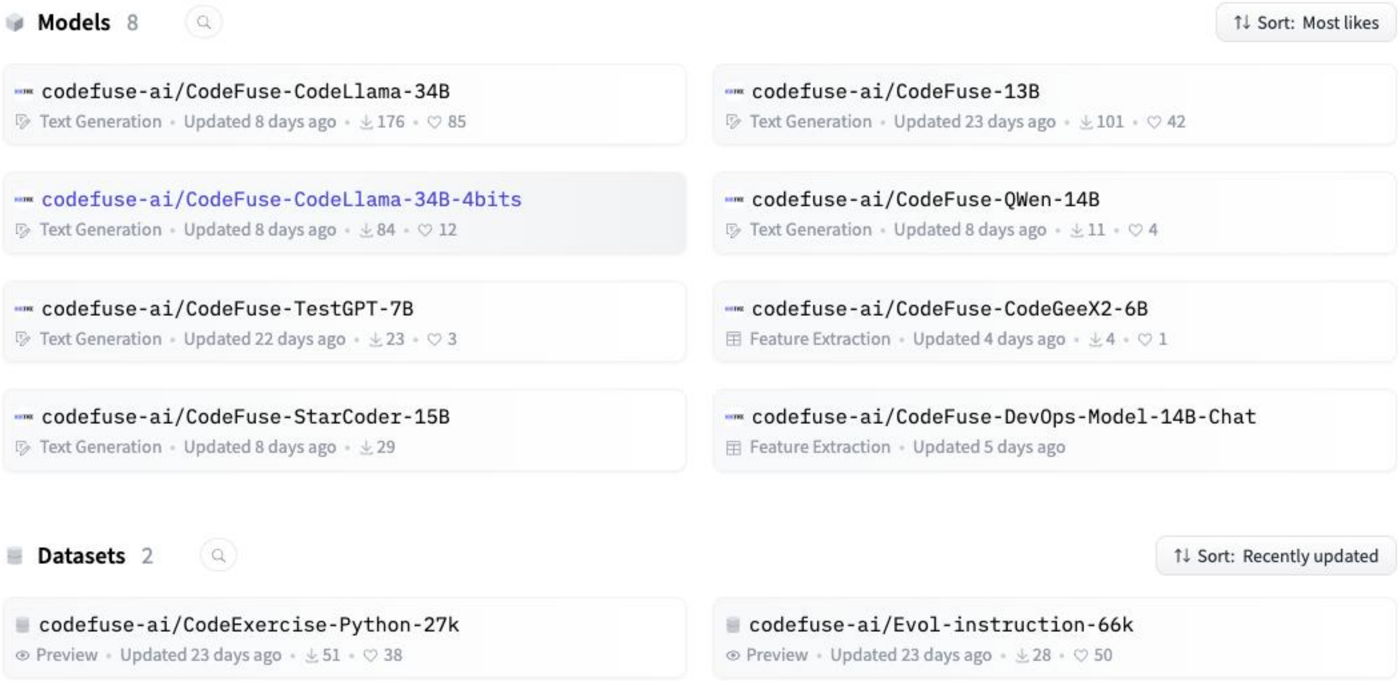
自动触发代码补全

目前支持10款IDE
40+编程语言
包括python/java/js等

官网:https://codefuse.alipay.com/welcome/product

▶ 前言：CodeFuse模型

CodeFuse 模型：旨在支持整个软件开发生命周期的大型代码语言模型（Code LLMs），涵盖设计、需求、编码、测试、部署、运维等关键阶段。



Models 8

⇅ Sort: Most likes

codefuse-ai/CodeFuse-CodeLlama-34B
Text Generation • Updated 8 days ago • ⤓176 • ♡ 85

codefuse-ai/CodeFuse-13B
Text Generation • Updated 23 days ago • ⤓101 • ♡ 42

codefuse-ai/CodeFuse-CodeLlama-34B-4bits
Text Generation • Updated 8 days ago • ⤓84 • ♡ 12

codefuse-ai/CodeFuse-QWen-14B
Text Generation • Updated 8 days ago • ⤓11 • ♡ 4

codefuse-ai/CodeFuse-TestGPT-7B
Text Generation • Updated 22 days ago • ⤓23 • ♡ 3

codefuse-ai/CodeFuse-CodeGeeX2-6B
Feature Extraction • Updated 4 days ago • ⤓4 • ♡ 1

codefuse-ai/CodeFuse-StarCoder-15B
Text Generation • Updated 8 days ago • ⤓29

codefuse-ai/CodeFuse-DevOps-Model-14B-Chat
Feature Extraction • Updated 5 days ago

Datasets 2

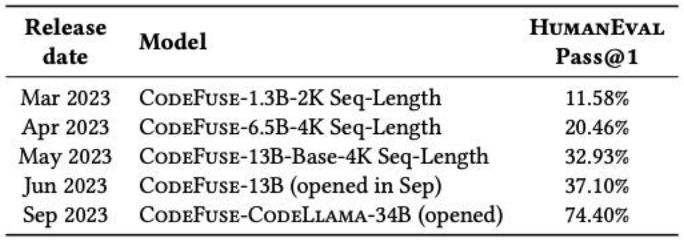
⇅ Sort: Recently updated

codefuse-ai/CodeExercise-Python-27k
Preview • Updated 23 days ago • ⤓51 • ♡ 38

codefuse-ai/Evol-instruction-66k
Preview • Updated 23 days ago • ⤓28 • ♡ 50

业界开源评测集上的roadmap

Table 1: CodeFuse project roadmap.

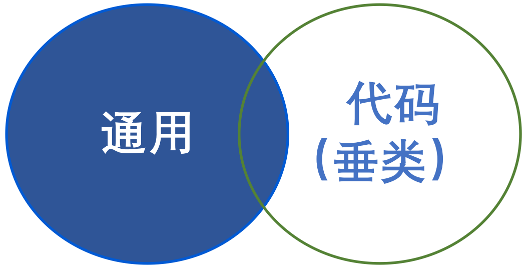


Release date	Model	HumanEval Pass@1
Mar 2023	CodeFuse-1.3B-2K Seq-Length	11.58%
Apr 2023	CodeFuse-6.5B-4K Seq-Length	20.46%
May 2023	CodeFuse-13B-Base-4K Seq-Length	32.93%
Jun 2023	CodeFuse-13B (opened in Sep)	37.10%
Sep 2023	CodeFuse-CodeLlama-34B (opened)	74.40%

CodeFuse系列论文：

https://arxiv.org/abs/2311.02303
https://arxiv.org/abs/2310.06266

前言：大模型时代如何评估代码大模型



伴随蚂蚁代码大模型的投产，我们发现代码领域打榜与实际投产存在一定的差异，基于此，我们探索并构建了适合企业项目的代码大模型的评测范式。

多维：　代码能力、基础能力、安全能力等多维度

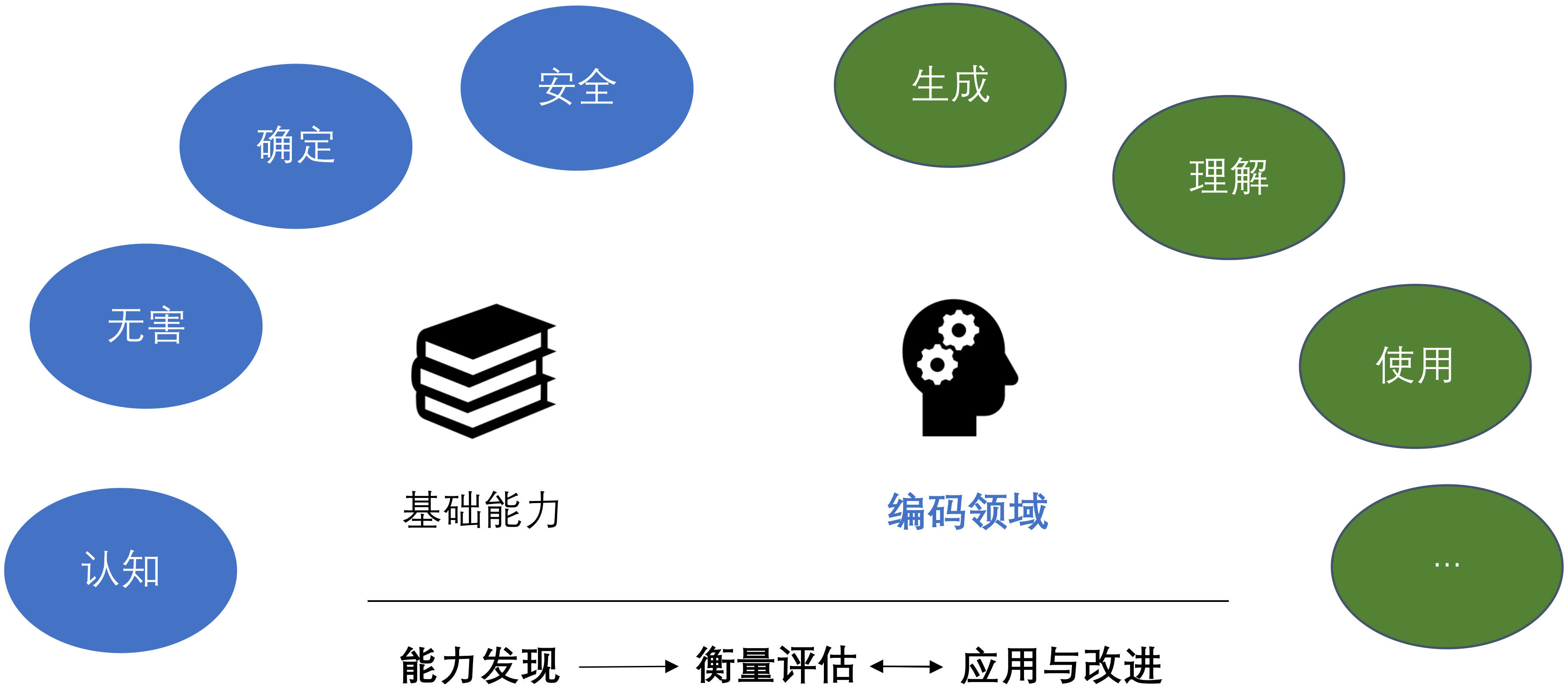
多样：　多编码语言、编码规范、跨项目编码等任务多样

可信：　样本纯净、标准开放、客观中立

PART 02
代码大模型的评估内容

代码大模型的评估内容

通过评估内容客观衡量代码大模型的各项能力，发现其长处和短处，反馈并协助模型做针对性的改进。



确定

安全

无害

认知

基础能力

生成

理解

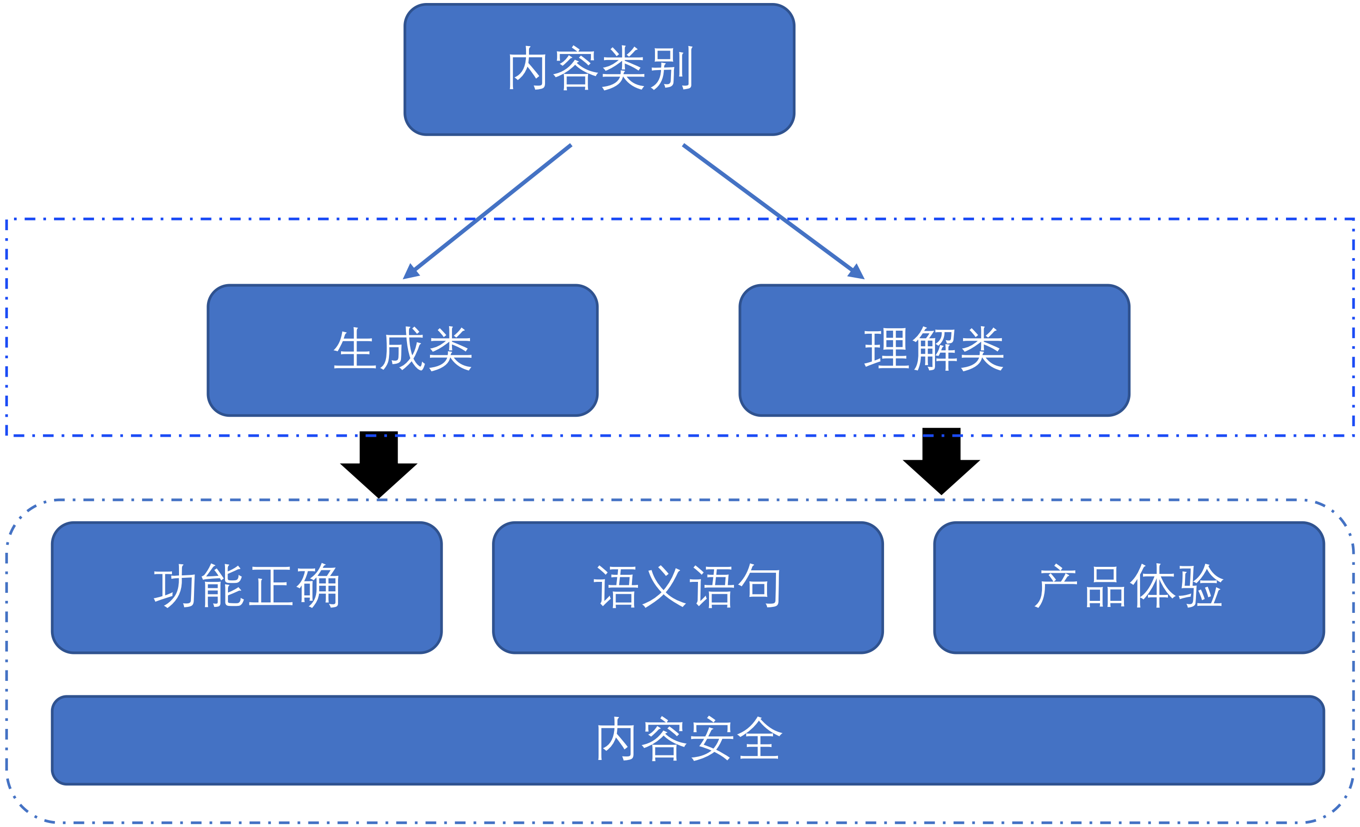
使用

…

编码领域

能力发现 → 衡量评估 ↔ 应用与改进

代码大模型的评估内容



内容类别

生成类

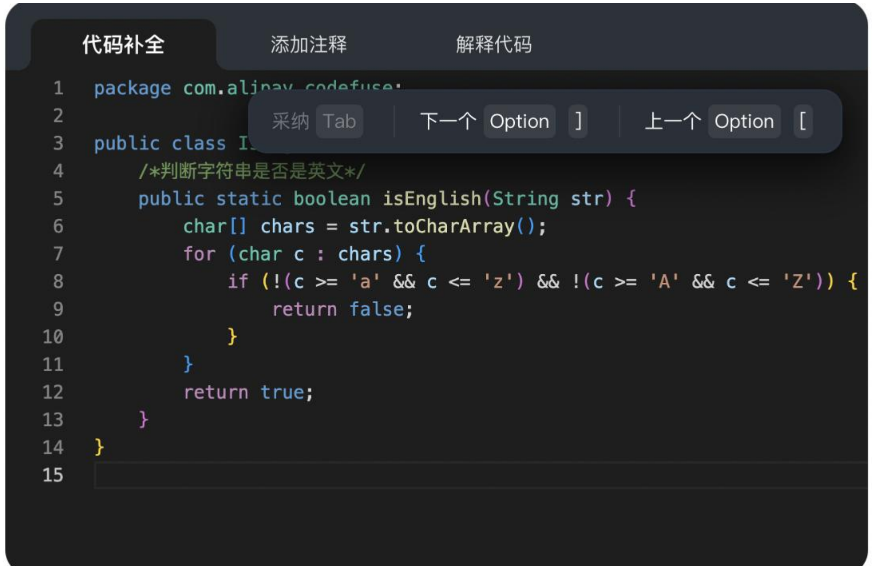
理解类

功能正确

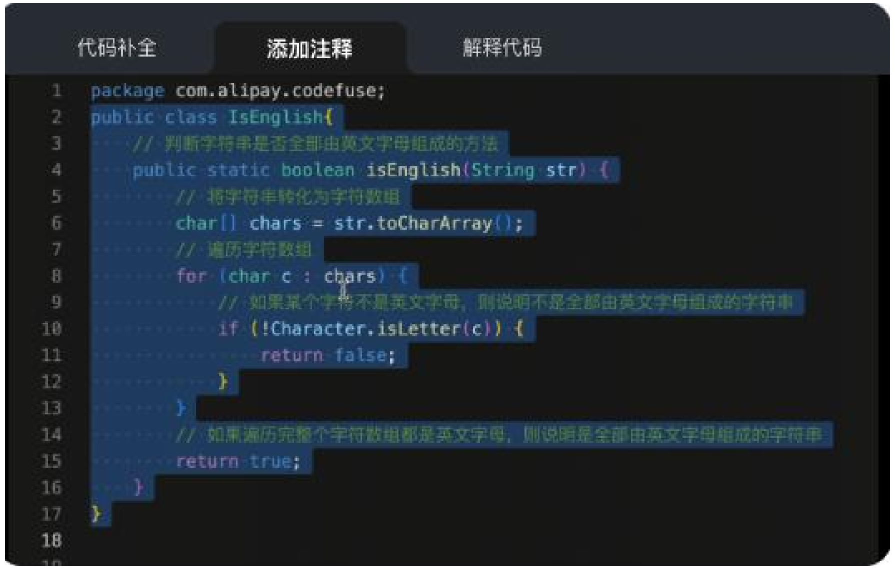
语义语句

产品体验

内容安全



```
package com.alipay.codefuse;

public class I
    /*判断字符串是否是英文*/
    public static boolean isEnglish(String str) {
        char[] chars = str.toCharArray();
        for (char c : chars) {
            if (!(c >= 'a' && c <= 'z') && !(c >= 'A' && c <= 'Z')) {
                return false;
            }
        }
        return true;
    }
}
```



```
package com.alipay.codefuse;
public class IsEnglish{
    // 判断字符串是否全部由英文字母组成的方法
    public static boolean isEnglish(String str) {
        // 将字符串转化为字符数组
        char[] chars = str.toCharArray();
        // 遍历字符数组
        for (char c : chars) {
            // 如果某个字符不是英文字母，则说明不是全部由英文字母组成的字符串
            if (!Character.isLetter(c)) {
                return false;
            }
        }
        // 如果遍历完整个字符数组都是英文字母，则说明是全部由英文字母组成的字符串
        return true;
    }
}
```

PART 03
代码大模型的评估方法

代码大模型的评估方法

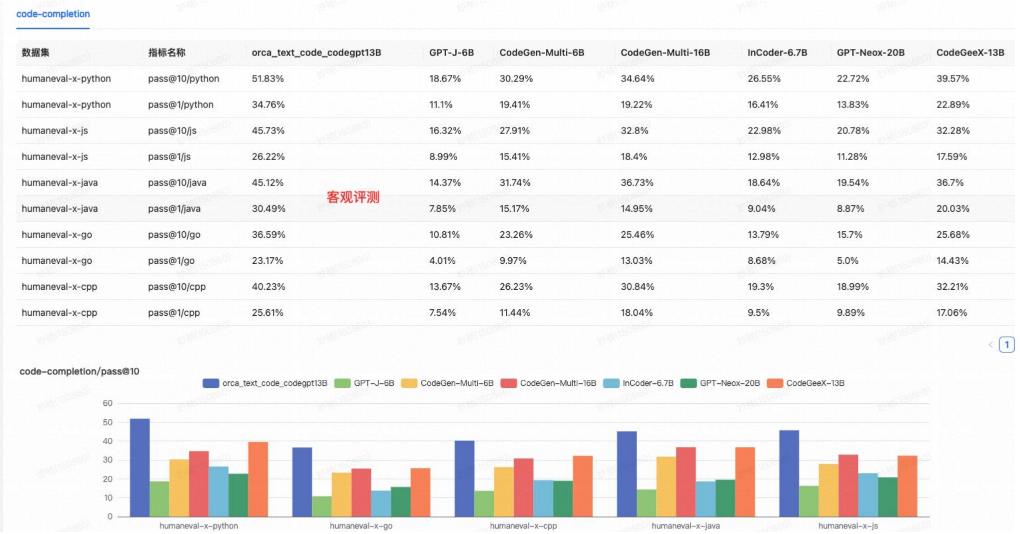
按照是否由**人来评估**模型性能，评测方法分为主观和客观

客观评测

基于评测基准对生成内容进行各维度量化评估

主观评测

组织多位专业人员通过人模交互观察模型表现
并根据基础标准、专家知识和经验综合评估

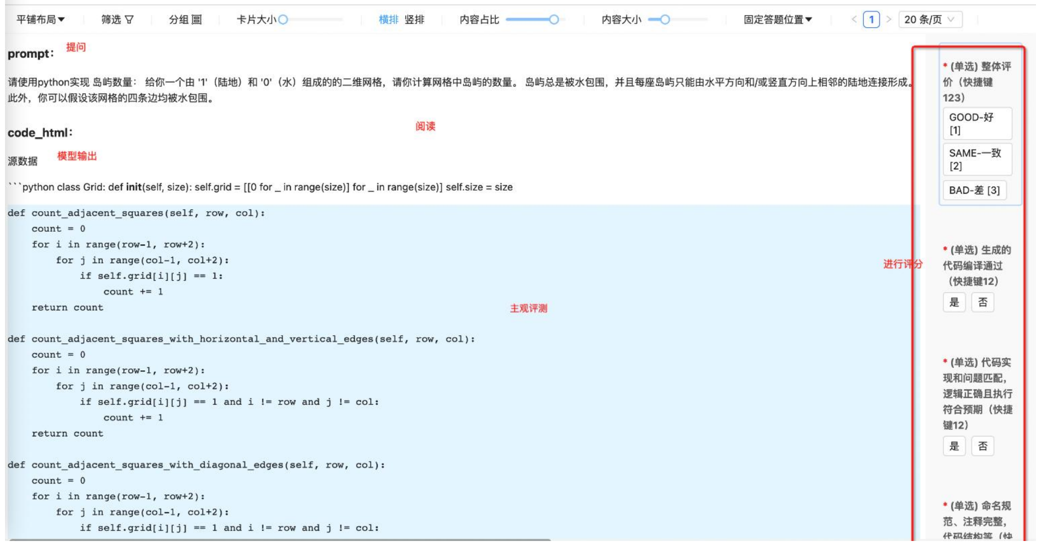


code-completion

客观评测

数据集	指标名称	orca_text_code_codegpt13B	GPT-J-6B	CodeGen-Multi-6B	CodeGen-Multi-16B	InCoder-6.7B	GPT-Neox-20B	CodeGeeX-13B
humaneval-x-python	pass@10/python	51.83%	18.67%	30.29%	34.64%	26.55%	22.72%	39.57%
humaneval-x-python	pass@1/python	34.76%	11.1%	19.41%	19.22%	16.41%	13.83%	22.89%
humaneval-x-js	pass@10/js	45.73%	16.32%	27.91%	32.8%	22.98%	20.78%	32.28%
humaneval-x-js	pass@1/js	26.22%	8.99%	15.41%	18.4%	12.98%	11.28%	17.59%
humaneval-x-java	pass@10/java	45.12%	14.37%	31.74%	36.73%	18.64%	19.54%	36.7%
humaneval-x-java	pass@1/java	30.49%	7.85%	15.17%	14.95%	9.04%	8.87%	20.03%
humaneval-x-go	pass@10/go	36.59%	10.81%	23.26%	25.46%	13.79%	15.7%	25.68%
humaneval-x-go	pass@1/go	23.17%	4.01%	9.97%	13.03%	8.68%	5.0%	14.43%
humaneval-x-cpp	pass@10/cpp	40.23%	13.67%	26.23%	30.84%	19.3%	18.99%	32.21%
humaneval-x-cpp	pass@1/cpp	25.61%	7.54%	11.44%	18.04%	9.5%	9.89%	17.06%

code-completion/pass@10



prompt: 提问

请使用python实现 岛屿数量： 给你一个由 '1'（陆地）和 '0'（水）组成的的二维网格，请你计算网格中岛屿的数量。 岛屿总是被水包围，并且每座岛屿只能由水平方向和/或竖直方向上相邻的陆地连接形成。此外，你可以假设该网格的四条边均被水包围。

code_html: 阅读

源数据 模型输出

```python class Grid: def **init**(self, size): self.grid = [[0 for _ in range(size)] for _ in range(size)] self.size = size

```
def count_adjacent_squares(self, row, col):
    count = 0
    for i in range(row-1, row+2):
        for j in range(col-1, col+2):
            if self.grid[i][j] == 1:
                count += 1
    return count

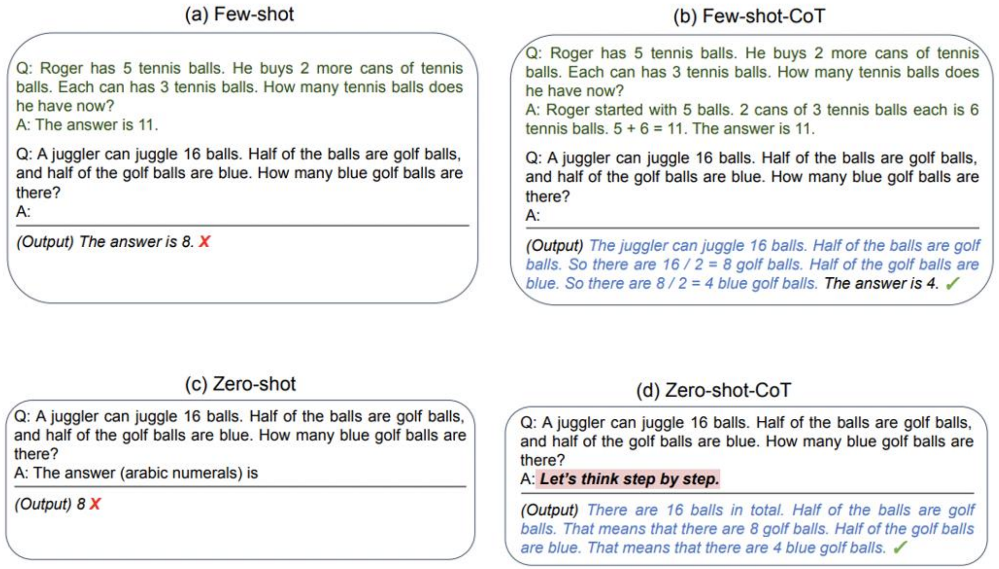
def count_adjacent_squares_with_horizontal_and_vertical_edges(self, row, col):
    count = 0
    for i in range(row-1, row+2):
        for j in range(col-1, col+2):
            if self.grid[i][j] == 1 and i != row and j != col:
                count += 1
    return count

def count_adjacent_squares_with_diagonal_edges(self, row, col):
    count = 0
    for i in range(row-1, row+2):
        for j in range(col-1, col+2):
            if self.grid[i][j] == 1 and i != row and j != col:
```

主观评测

进行评分

* (单选) 整体评价（快捷键123）

GOOD-好 [1]

SAME-一致 [2]

BAD-差 [3]

* (单选) 生成的代码编译通过（快捷键12）

是 否

* (单选) 代码实现和问题匹配，逻辑正确且执行符合预期（快捷键12）

是 否

* (单选) 命名规范、注释完整，代码结构等（快
```

代码大模型的评估方法

按照prompts设置方法评测又可分为：零样本（zero-shot）、小样本（few-shot）、零样本思维链（zero-shot-cot）、小样本思维链（few-shot-cot）

代码生成能力目前大部分采用的策略是：零样本（zero-shot）



(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

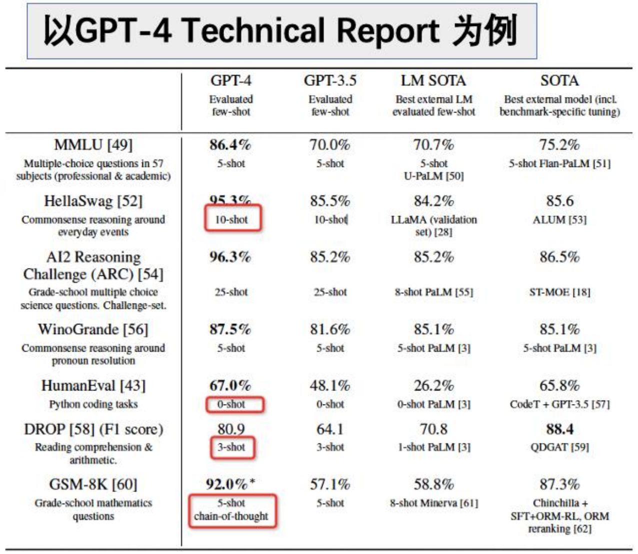
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓



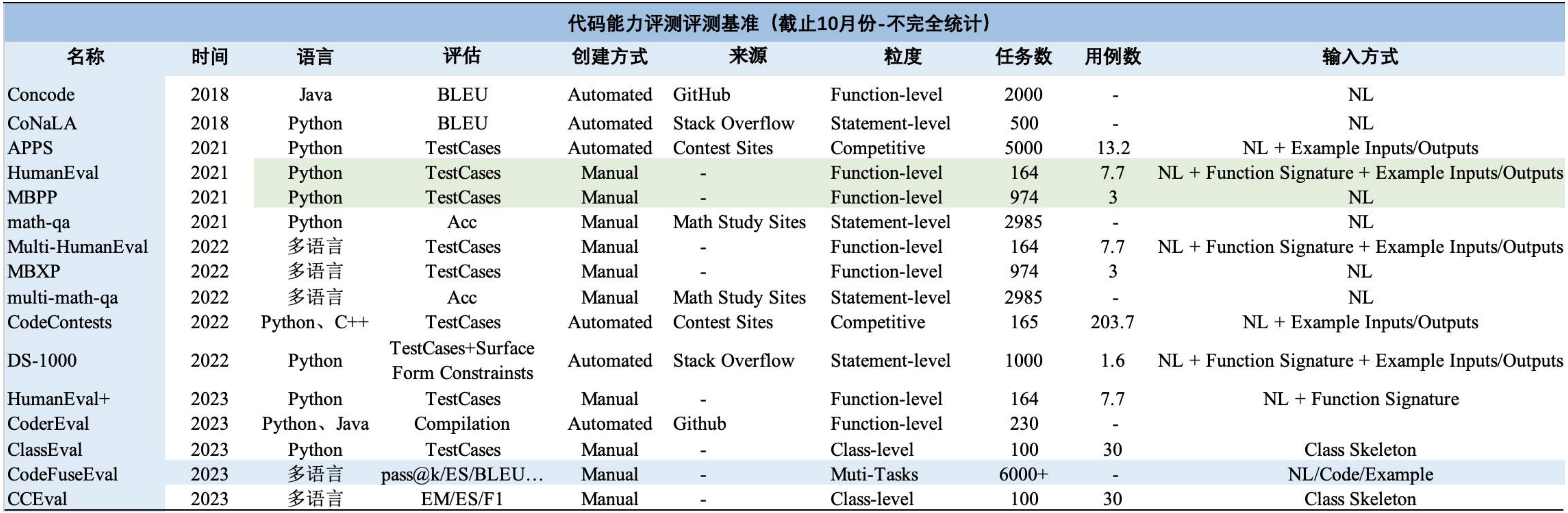
以GPT-4 Technical Report 为例

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	**95.3%** 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	**96.3%** 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	**87.5%** 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	**88.4** QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

PART 04
代码大模型评估基准

评估基准

模型评估基准是优化模型，了解差距，衡量不同架构模型的同类场景性能的最有效的工具。

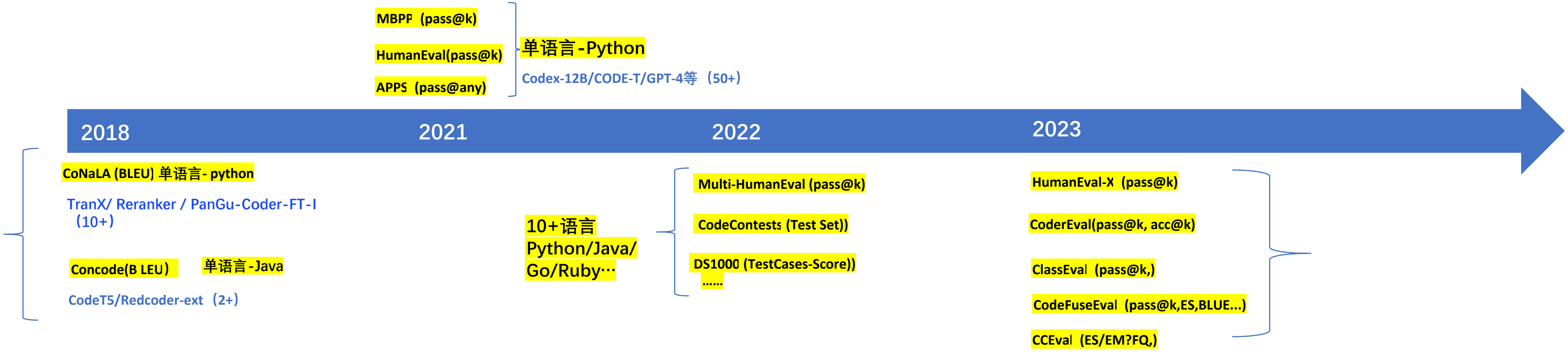


代码能力评测评测基准（截止10月份-不完全统计）

名称	时间	语言	评估	创建方式	来源	粒度	任务数	用例数	输入方式
Concode	2018	Java	BLEU	Automated	GitHub	Function-level	2000	-	NL
CoNaLA	2018	Python	BLEU	Automated	Stack Overflow	Statement-level	500	-	NL
APPS	2021	Python	TestCases	Automated	Contest Sites	Competitive	5000	13.2	NL + Example Inputs/Outputs
HumanEval	2021	Python	TestCases	Manual	-	Function-level	164	7.7	NL + Function Signature + Example Inputs/Outputs
MBPP	2021	Python	TestCases	Manual	-	Function-level	974	3	NL
math-qa	2021	Python	Acc	Manual	Math Study Sites	Statement-level	2985	-	NL
Multi-HumanEval	2022	多语言	TestCases	Manual	-	Function-level	164	7.7	NL + Function Signature + Example Inputs/Outputs
MBXP	2022	多语言	TestCases	Manual	-	Function-level	974	3	NL
multi-math-qa	2022	多语言	Acc	Manual	Math Study Sites	Statement-level	2985	-	NL
CodeContests	2022	Python、C++	TestCases	Automated	Contest Sites	Competitive	165	203.7	NL + Example Inputs/Outputs
DS-1000	2022	Python	TestCases+Surface Form Constrainsts	Automated	Stack Overflow	Statement-level	1000	1.6	NL + Function Signature + Example Inputs/Outputs
HumanEval+	2023	Python	TestCases	Manual	-	Function-level	164	7.7	NL + Function Signature
CoderEval	2023	Python、Java	Compilation	Automated	Github	Function-level	230	-	
ClassEval	2023	Python	TestCases	Manual	-	Class-level	100	30	Class Skeleton
CodeFuseEval	2023	多语言	pass@k/ES/BLEU…	Manual	-	Muti-Tasks	6000+	-	NL/Code/Example
CCEval	2023	多语言	EM/ES/F1	Manual	-	Class-level	100	30	Class Skeleton

采纳原则：　有效未被污染的　　　多样多维：多语言，多任务，多维衡量

评估基准的演进



2018

CoNaLA (BLEU) 单语言- python

TranX/ Reranker / PanGu-Coder-FT-I（10+）

Concode(B LEU） 单语言-Java

CodeT5/Redcoder-ext（2+）

2021

MBPP（pass@k)

HumanEval(pass@k)

APPS（pass@any)

单语言-Python

Codex-12B/CODE-T/GPT-4等（50+）

10+语言 Python/Java/ Go/Ruby…

2022

Multi-HumanEval (pass@k)

CodeContests (Test Set))

DS1000 (TestCases-Score))

……

2023

HumanEval-X（pass@k)

CoderEval(pass@k, acc@k)

ClassEval（pass@k,)

CodeFuseEval（pass@k,ES,BLUE...)

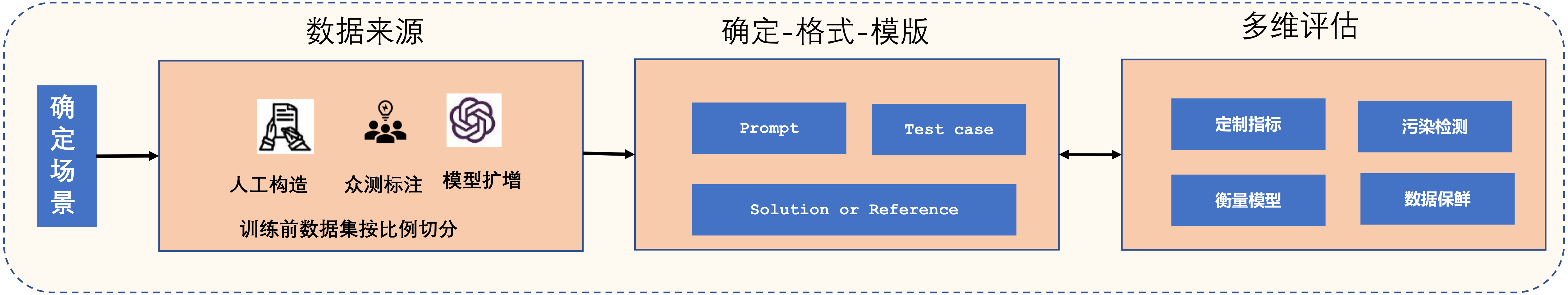
CCEval（ES/EM?FQ,)

单语言 静态指标 单轮 → 多语言 动态指标 单轮 → 多语言 动静指标 单轮

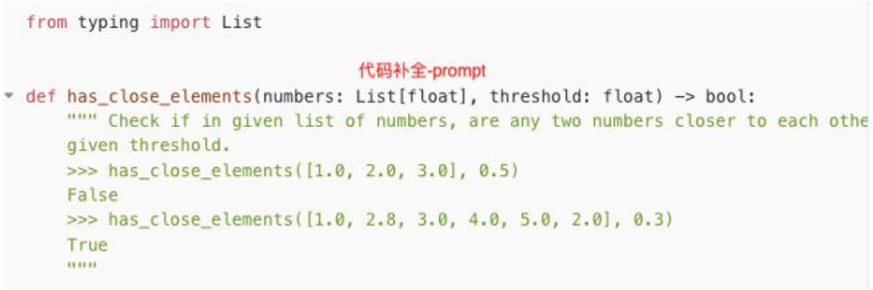
▶ 评估基准（自建）

自建是开源的延伸:

1. 更贴合企业特定场景如衡量模型对企业项目代码的生成和理解能力
2. 补充开源评测集，全面有梯度的衡量模型性能

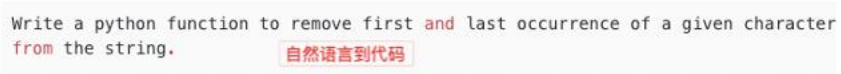


评测样本prompt&格式



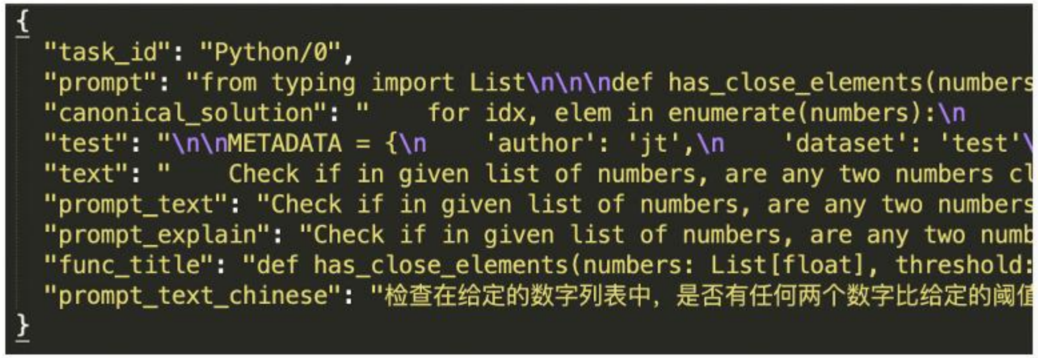
代码补全-prompt

```
from typing import List


def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """ Check if in given list of numbers, are any two numbers closer to each othe
    given threshold.
    >>> has_close_elements([1.0, 2.0, 3.0], 0.5)
    False
    >>> has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0], 0.3)
    True
    """
```

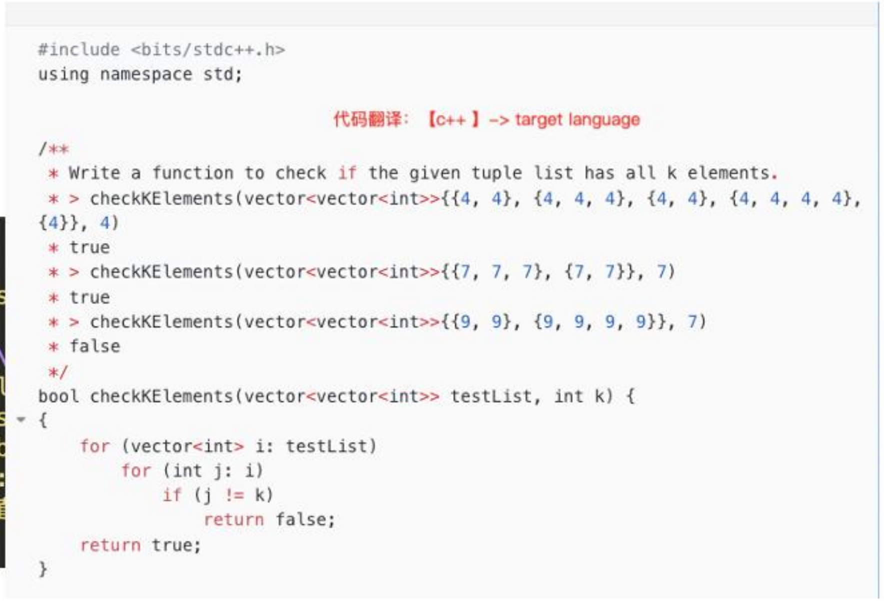


```
Write a python function to remove first and last occurrence of a given character
from the string.
```

自然语言到代码

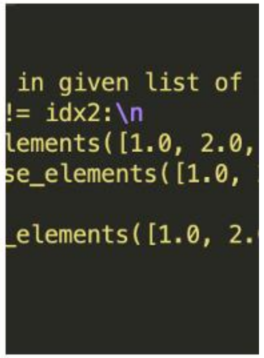


```
{
  "task_id": "Python/0",
  "prompt": "from typing import List\n\n\ndef has_close_elements(numbers
  "canonical_solution": "    for idx, elem in enumerate(numbers):\n
  "test": "\n\nMETADATA = {\n    'author': 'jt',\n    'dataset': 'test'\
  "text": "    Check if in given list of numbers, are any two numbers cl
  "prompt_text": "Check if in given list of numbers, are any two numbers
  "prompt_explain": "Check if in given list of numbers, are any two numb
  "func_title": "def has_close_elements(numbers: List[float], threshold:
  "prompt_text_chinese": "检查在给定的数字列表中，是否有任何两个数字比给定的阈值
}
```



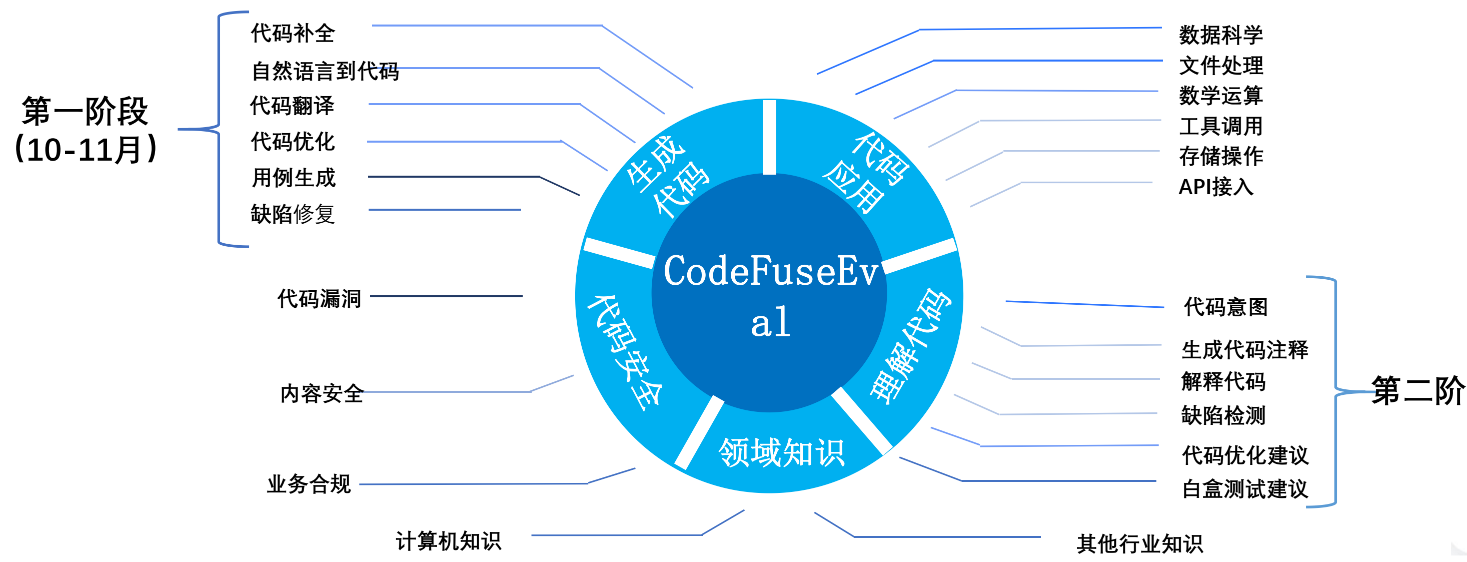
代码翻译：【c++ 】-> target language

```
#include <bits/stdc++.h>
using namespace std;

/**
 * Write a function to check if the given tuple list has all k elements.
 * > checkKElements(vector<vector<int>>{{4, 4}, {4, 4, 4}, {4, 4}, {4, 4, 4, 4},
{4}}, 4)
 * true
 * > checkKElements(vector<vector<int>>{{7, 7, 7}, {7, 7}}, 7)
 * true
 * > checkKElements(vector<vector<int>>{{9, 9}, {9, 9, 9, 9}}, 7)
 * false
 */
bool checkKElements(vector<vector<int>> testList, int k) {
{
    for (vector<int> i: testList)
        for (int j: i)
            if (j != k)
                return false;
    return true;
}
```



```
 in given list of
!= idx2:\n
lements([1.0, 2.0,
se_elements([1.0,

_elements([1.0, 2.
```

多任务评估基准



第一阶段（10-11月）

- 代码补全
- 自然语言到代码
- 代码翻译
- 代码优化
- 用例生成
- 缺陷修复

代码漏洞

内容安全

业务合规

计算机知识

CodeFuseEval

生成代码

代码应用

理解代码

领域知识

代码安全

数据科学

文件处理

数学运算

工具调用

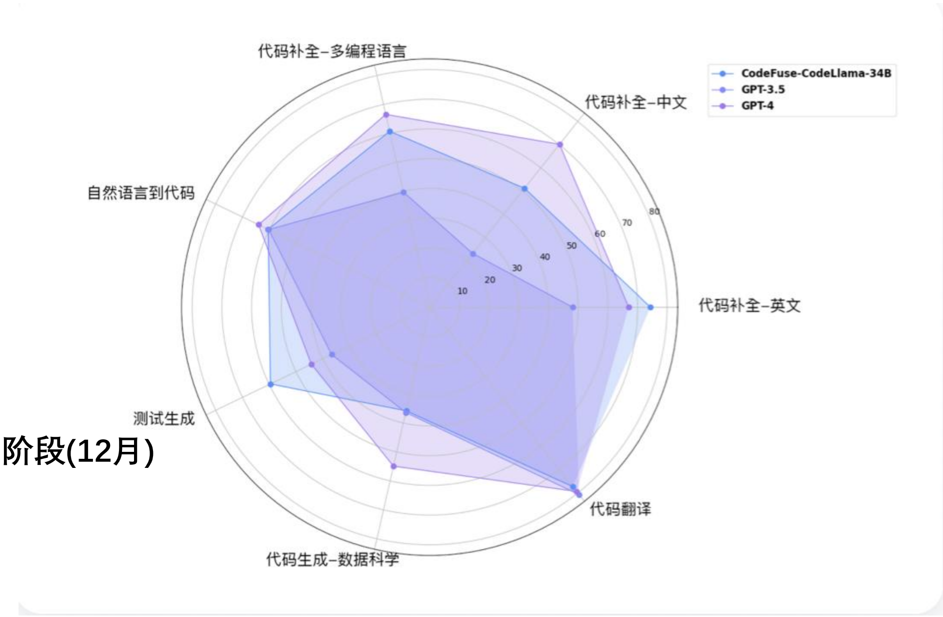
存储操作

API接入

第二阶段(12月)

- 代码意图
- 生成代码注释
- 解释代码
- 缺陷检测
- 代码优化建议
- 白盒测试建议

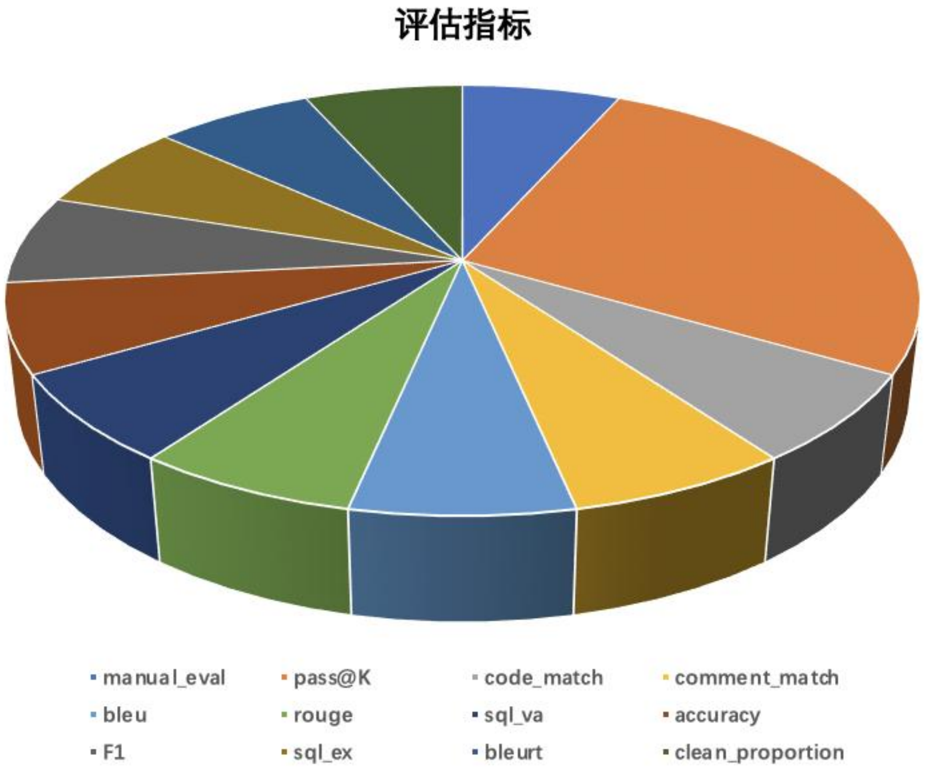
其他行业知识

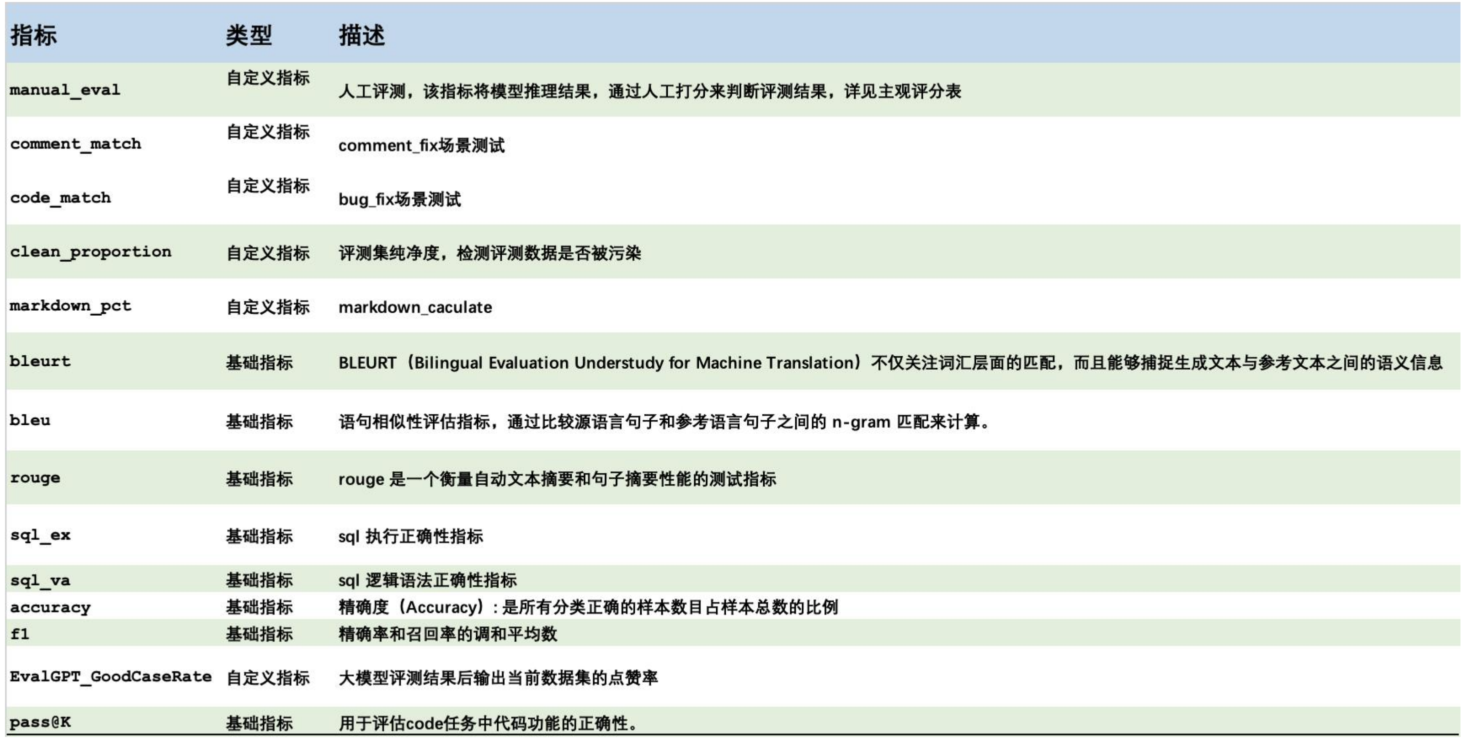


开源地址：https://github.com/codefuse-ai/codefuse-evaluation

多类型评测指标

生成结果要求：有用 真实 无害



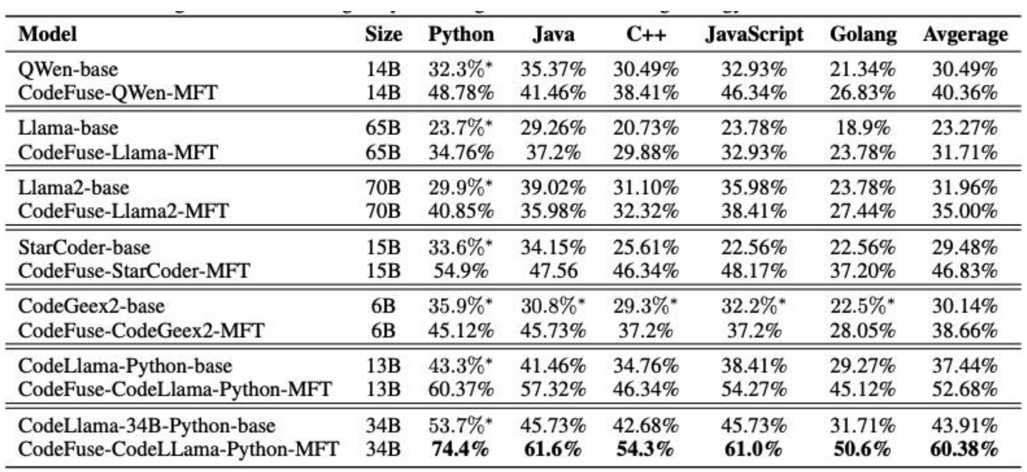


指标	类型	描述
manual_eval	自定义指标	人工评测，该指标将模型推理结果，通过人工打分来判断评测结果，详见主观评分表
comment_match	自定义指标	comment_fix场景测试
code_match	自定义指标	bug_fix场景测试
clean_proportion	自定义指标	评测集纯净度，检测评测数据是否被污染
markdown_pct	自定义指标	markdown_caculate
bleurt	基础指标	BLEURT（Bilingual Evaluation Understudy for Machine Translation）不仅关注词汇层面的匹配，而且能够捕捉生成文本与参考文本之间的语义信息
bleu	基础指标	语句相似性评估指标，通过比较源语言句子和参考语言句子之间的 n-gram 匹配来计算。
rouge	基础指标	rouge 是一个衡量自动文本摘要和句子摘要性能的测试指标
sql_ex	基础指标	sql 执行正确性指标
sql_va	基础指标	sql 逻辑语法正确性指标
accuracy	基础指标	精确度（Accuracy）：是所有分类正确的样本数目占样本总数的比例
f1	基础指标	精确率和召回率的调和平均数
EvalGPT_GoodCaseRate	自定义指标	大模型评测结果后输出当前数据集的点赞率
pass@K	基础指标	用于评估code任务中代码功能的正确性。

PART 05
代码大模型多任务评估

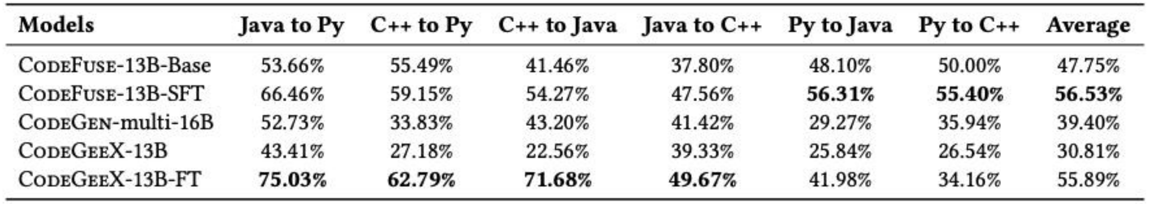
代码大模型多任务评估

多任务评估让我们全面了解模型的专业能力和基础性能、检验模型的泛化能力、并辅助模型优化。



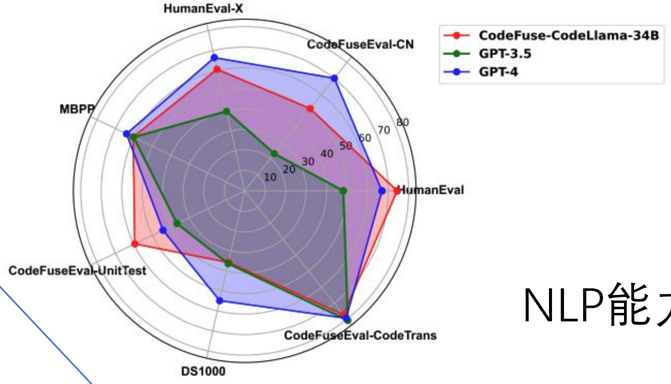
Model	Size	Python	Java	C++	JavaScript	Golang	Avgerage
QWen-base	14B	32.3%*	35.37%	30.49%	32.93%	21.34%	30.49%
CodeFuse-QWen-MFT	14B	48.78%	41.46%	38.41%	46.34%	26.83%	40.36%
Llama-base	65B	23.7%*	29.26%	20.73%	23.78%	18.9%	23.27%
CodeFuse-Llama-MFT	65B	34.76%	37.2%	29.88%	32.93%	23.78%	31.71%
Llama2-base	70B	29.9%*	39.02%	31.10%	35.98%	23.78%	31.96%
CodeFuse-Llama2-MFT	70B	40.85%	35.98%	32.32%	38.41%	27.44%	35.00%
StarCoder-base	15B	33.6%*	34.15%	25.61%	22.56%	22.56%	29.48%
CodeFuse-StarCoder-MFT	15B	54.9%	47.56	46.34%	48.17%	37.20%	46.83%
CodeGeex2-base	6B	35.9%*	30.8%*	29.3%*	32.2%*	22.5%*	30.14%
CodeFuse-CodeGeex2-MFT	6B	45.12%	45.73%	37.2%	37.2%	28.05%	38.66%
CodeLlama-Python-base	13B	43.3%*	41.46%	34.76%	38.41%	29.27%	37.44%
CodeFuse-CodeLlama-Python-MFT	13B	60.37%	57.32%	46.34%	54.27%	45.12%	52.68%
CodeLlama-34B-Python-base	34B	53.7%*	45.73%	42.68%	45.73%	31.71%	43.91%
CodeFuse-CodeLLama-Python-MFT	34B	**74.4%**	**61.6%**	**54.3%**	**61.0%**	**50.6%**	**60.38%**

Table 5: Performance(pass@1) comparison of CodeFuse with previous models on code translation using greedy decoding

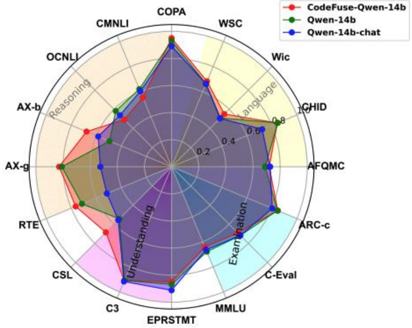


Models	Java to Py	C++ to Py	C++ to Java	Java to C++	Py to Java	Py to C++	Average
CodeFuse-13B-Base	53.66%	55.49%	41.46%	37.80%	48.10%	50.00%	47.75%
CodeFuse-13B-SFT	66.46%	59.15%	54.27%	47.56%	**56.31%**	**55.40%**	**56.53%**
CodeGen-multi-16B	52.73%	33.83%	43.20%	41.42%	29.27%	35.94%	39.40%
CodeGeeX-13B	43.41%	27.18%	22.56%	39.33%	25.84%	26.54%	30.81%
CodeGeeX-13B-FT	**75.03%**	**62.79%**	**71.68%**	**49.67%**	41.98%	34.16%	55.89%

代码能力



NLP能力



PART 06
展望

展望

评测对象（模型）：

1. 评测任务多样化，贴合企业级编码场景
2. 多维评估：技能、效能、服务（稳定/鲁棒）

评测技术（基准）：

1. 开源开放，快速迭代，匹配模型能力增长速度
2. 框架升级，模型评测，提高主观评测效能及可重复性

代码大模型评估
技能
效能
鲁棒
稳定

THANKS