中国科学院软件研究所学术年会 暨重点实验室科技活动周 2025 第十届
学术论文

DomainEval：自动构建的多领域代码生成评测基准

DOMAINEVAL: An Auto-Constructed Benchmark for Multi-Domain Code Generation

珠齐明*，曹嘉伦*，陆垚杰†，林鸿宇，韩先培，孙乐，张成志

In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)
(Vol. 39, No. 24, pp. 26148-26156)

联系人： 珠齐明　　**邮箱：** qiming.zhu@foxmail.com

研究背景和动机

Computation

```python
def calculate_mean(lst):
    total = sum(lst)
    mean = total / len(lst)
    return mean
```

Basic

```python
import datetime as dt

def get_weekday_occurrence(day: dt.date):
    xthday = 1 + (day.day - 1) // 7
    return day.weekday(), xthday
```

Cryptography

```python
def derive_key(master: bytes) -> bytes:
    derived = HKDF(master, 32, b'', SHA256)
    return derived
```

Network

```python
import ipaddress

def is_valid_ipv6_address(ip_str):
    try:
        ipaddress.IPv6Address(ip_str)
    except ValueError:
        return False
    return True
```

Visualization

```python
def color_xy_to_temperature(x, y):
    n = (x - 0.3320) / (0.1858 - y)
    cct = 437 * n** 3 + 3601 * n ** 2 +
          6861 * n + 5517
    return int(cct)
```

System

```python
import collections

def free_physmem():
    out = sh(['free', '-b'], env={'LANG': 'C.UTF-8'})
    lines = out.split('\n')
    for line in lines:
        if line.startswith('Mem'):
            total, used, free, shared = (int(x) for x in line.split()[1:5])
            nt=collections.namedtuple('free','total used free shared output')
            return nt(total, used, free, shared, out)
    raise ValueError("can't find 'Mem')
```

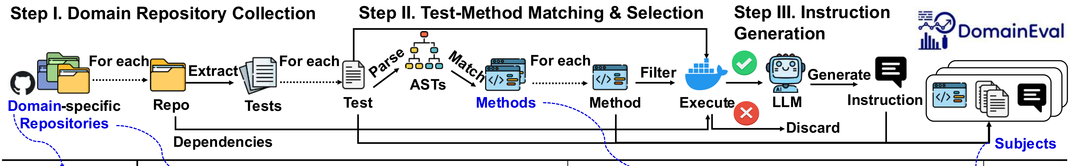
现有挑战：

- **任务类型单一**：当前代码生成能力评测基准针对常见代码任务（如排序算法、求最大公约数），缺乏**特定领域任务**（如计算类、密码类）的支持，难以满足多领域软件开发需求。
- **人工标注成本高**：现有基准依赖人工标注的数据集，难以满足**自动化构建和工程实际需求**。

我们的方法：

- 基于测试完备的代码，自动生成对应指令
- 实现**代码仓库→ <指令，代码，测试用例>** 一键式转化
- 构建了多领域的代码生成能力基准，覆盖六个领域

基准构建流程



Domain	Repositories Collected from GitHub	Methods Collected	Number of Subjects
Computation	**(Total 15)** numpy, pandas, scikit-learn, librosa, nltk, obspy, scikit-image,gensim, geopandas, statsmodels, sympy, tensorflow, …	get_datevalue, is_float_dtype, median, power, sort, calculate_smoothing_matrix, collect_sqrt, …	1705
Basic	**(Total 12)** setuptools, prettytable, charset_normalizer, arrow, chinese-calendar, khal, vdirsyncer, workalendar, …	max_height, total_time, join_dict_keys, get_path, get_dates, get_weekday_occurrence, format_timestamp, …	107
Network	**(Total 11)** django, flask-restful, geopyi, mechanize, mitmproxy, requests, scrapy, sendgrid-python, werkzeug, wtforms, yt-dlp	format_html, get_password_validators, parse_cookie, filepath_to_uri, get_callable, is_valid_ipv6_address, …	256
Visualization	**(Total 18)** bokeh, datashader, gradio, matplotlib, Pillow, vaeX, pygwalker, redashi, seaborn, tensorboardX, word_cloud, …	get_colormap, create_sphere, plotting_context, histplot, husl_palette, light_palette, meshgrid_triangles, hsl_to_rgb, …	186
System	**(Total 16)** anaconda, billiard, core, glances,loguru, mpire, openpyxl, pandapower, poetry, psutil, sentry, xlrd, xlwt, …	is_admin, chunk_tasks, free_swap, serialize_response, sysctl, free_physmem, parse_environ_block, get_mac_address, …	100
Cryptography	**(Total 19)** asn1crypto, badsecrets, blake3-py, crypto-attacks, cryptography, detect-secrets, firmware, freqtrade, paranoid_crypto, pycryptodome, python-hdwallet, python-rsa, …	BinaryMatrixRank, derive_key, encode_bitstring, attack, UniversalDistribution, solve_right, load_cryptrec_vectors, load_rsa_nist_vectors, gen_keys, socket_any_family, …	100

实验结果与分析

Pass@1 (Greedy Search N=1)	Size	Comp	Network	Visual	Basic	System	Crypt	Mean	Std
GPT-4o-mini	\	90.38	70.31	59.68	69.16	51.00	43.00	63.92	16.68
GPT-3.5-turbo	\	83.40	58.98	48.92	56.07	32.00	31.00	51.73	19.50
Qwen2-72B-Instruct-GPTQ-Int4	72B	86.86	66.80	49.46	69.16	41.00	36.00	58.21	19.39
DeepSeek-Coder-33b-instruct	33B	83.93	64.45	50.54	59.81	46.00	35.00	56.62	16.94
DeepSeek-Coder-V2-Lite-Instruct	16B	86.04	62.11	50.00	65.42	41.00	38.00	57.10	17.92
DeepSeek-Coder-6.7b-instruct	6.7B	83.52	58.98	45.70	57.94	36.00	40.00	53.69	17.32
CodeLlama-34b-Instruct	34B	76.07	60.16	41.94	55.14	35.00	31.00	49.89	17.09
CodeLlama-13b-Instruct	13B	80.29	62.11	42.47	58.88	34.00	27.00	50.79	19.90
CodeLlama-7b-Instruct	7B	77.13	60.55	43.55	52.34	36.00	32.00	50.26	16.82
CodeQwen1.5-7B-Chat	7B	85.16	60.94	47.85	60.75	37.00	37.00	54.78	18.31
Phi-3-medium-4k-instruct	14B	75.54	60.16	45.16	61.68	42.00	35.00	53.26	15.10
Llama-2-13b-chat	13B	80.94	53.12	34.95	44.86	19.00	12.00	40.81	24.97
Average	\	82.44	61.56	46.69	59.27	37.50	33.08	53.42	18.33

Pass@5 (Sampling Search N=5)	Size	Comp	Network	Visual	Basic	System	Crypt	Mean	Std
GPT-4o-mini	\	91.26	72.66	61.83	71.03	57.00	49.00	67.13	14.75
GPT-3.5-turbo	\	87.33	62.89	52.15	60.75	36.00	34.00	55.52	19.74
Qwen2-72B-Instruct-GPTQ-Int4	72B	90.15	70.70	54.84	73.83	50.00	46.00	64.25	16.90
DeepSeek-Coder-33b-instruct	33B	89.79	70.70	55.38	68.22	57.00	42.00	63.85	16.34
DeepSeek-Coder-V2-Lite-Instruct	16B	88.91	65.62	53.76	68.22	49.00	44.00	61.59	16.35
DeepSeek-Coder-6.7b-instruct	6.7B	89.79	63.67	55.38	67.29	49.00	44.00	61.52	16.36
CodeLlama-34b-Instruct	34B	85.10	63.28	48.39	62.62	41.00	42.00	57.07	16.83
CodeLlama-13b-Instruct	13B	89.85	65.62	51.61	66.36	38.00	35.00	57.74	20.55
CodeLlama-7b-Instruct	7B	86.80	63.67	51.61	64.49	43.00	40.00	58.26	17.28
CodeQwen1.5-7B-Chat	7B	91.03	64.06	55.38	68.22	45.00	42.00	60.95	17.95
Phi-3-medium-4k-instruct	14B	85.10	67.58	54.30	67.29	47.00	44.00	60.88	15.45
Llama-2-13b-chat	13B	87.68	55.86	39.78	48.60	26.00	21.00	46.49	24.10
Average	\	88.57	65.53	52.87	65.58	44.83	40.25	59.60	17.72

- **Domain Biases:** LLMs 普遍在计算类任务上表现良好，但在密码学和系统相关任务上表现欠佳。
- **LLMs Biases:** 表现最好的是闭源模型GPT-4o-mini。代码数据微调会带来整体性能的提升，但是领域差距依然存在（Llama到CodeLlama）。
- **N samples:** 更多次的采样可以提高整体性能表现，但领域差距依然存在甚至进一步增大。