ISCAS 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

COFT：提升大语言模型的零样本生成能力
COFT: Making Large Language Models Better Zero-Shot Learners for Code Generation

李维佳，钱永杰，高科，陈海鑫，王歆妤，童煜晨，
李玲，武延军，赵琛

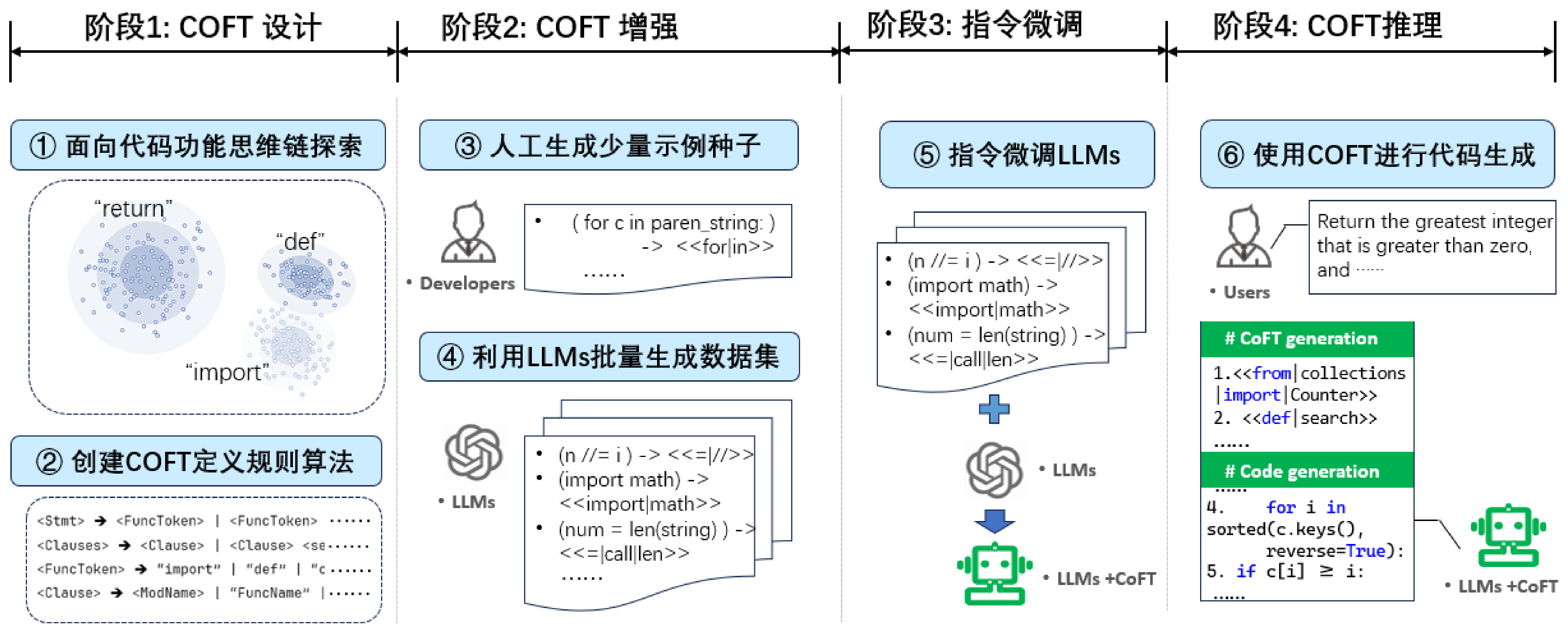
发表在ICPC 2025

主要联系人：李维佳，liweijia221@mails.ucas.ac.cn

工作介绍

思维链提示机制已有效提升了大语言模型在多种自然语言处理任务中的表现，包括复杂的零样本学习场景。然而，传统的 CoT 方法在代码生成领域收效甚微。这一局限性的根源在于代码的逻辑结构和表征形式与自然语言之间存在巨大差异。鉴于训练和部署成本，通过先进的提示技术和指令微调来提升轻量级大语言模型的性能至关重要。我们提出 COFT（功能触发链, Chain of Functional Triggers），这是一种专为代码生成任务设计的新型思维链策略。COFT 的设计基于以下重要观察：一个专用于代码生成的思维链应清晰地指明每个关键步骤的核心功能，同时采用编程领域中普遍使用的标准标识符来聚焦任务的核心概念。实验结果证明适当的COFT设计与指令微调相结合，能够充分激发出大语言模型的潜能。

方法设计



① **COFT设计：**探索最常见的代码功能触发词，并制定COFT生成规则

② **COFT增强：**数据构建首先用人工标注少量种子数据；其次使用大语言模型根据上一步的输出批量生成COFT-Instruct指令微调数据集

③ **指令微调：**使用COFT-Instruct数据集微调大语言模型

④ **COFT推理：**对微调后的大语言模型使用COFT进行代码生成任务

实验评估

Method	Year	InsT	HumanEval	MBPP	BCB(full)	BCB(hard)
StarCoder [10]	**2023**	✗	**33.6**	**43.3**	**19.9**	**2.7**
(15B, foundation model for the following methods)						
OurMethod		✓	68.9	45.4	22.7	6.1
Absolute Improvement			**(35.3↑)**	**(2.1↑)**	**(2.8↑)**	**(5.4↑)**
CodeLlama [14]	**2023**	✗	**36.0**	**47.0**	**31.3**	**8.1**
(13B, foundation model for the following methods)						
OurMethod		✓	64.6	51.0	34.4	9.5
Absolute Improvement			**(28.6↑)**	**(4.0↑)**	**(3.1↑)**	**(1.4↑)**
CodeLlama [14]	**2023**	✗	**33.5**	**41.4**	**27.1**	**4.1**
(7B, foundation model for the following methods)						
OurMethod		✓	63.4	45.6	30.0	12.2
Absolute Improvement			**(29.9↑)**	**(4.2↑)**	**(2.9↑)**	**(8.1↑)**
DeepseekCoder [16]	**2023**	✗	**48.2**	**59.4**	**41.8**	**12.8**
(6.7B, foundation model for the following methods)						
OurMethod		✓	73.7	64.0	45.2	16.9
Absolute Improvement			**(25.5↑)**	**(4.6↑)**	**(3.4↑)**	**(4.1↑)**

表1：COFT在基座大模型的提升效果

Method	Year	InsT	HumanEval	MBPP	BCB(full)	BCB(hard)
OctoCoder [12]	2023	✓	46.2	43.5	5.1	2.7
OurMethod		✓	67.9	41.8	21.6	6.8
CodeLLama-13B-instruct [14]	2023	✓	42.7	49.4	27.5	5.4
OurMethod		✓	66.5	50.2	32.5	10.1
CodeLLama-7B-instruct [14]	2023	✓	34.8	44.4	23.4	4.1
OurMethod		✓	62.8	46.4	28.4	9.5
WaveCoder-DS-6.7B [13]	2024	✓	64.0	62.8	43.6	15.5
OurMethod		✓	76.2	61.8	43.4	15.5

表2：COFT对其他微调方法的提升效果

Format	HumanEval	MBPP	BCB(full)	BCB(hard)
DeepseekCoder	48.2	59.4	41.8	12.8
COFT	73.7	64.0	45.2	16.9
non-COFT	68.2	59.8	42.2	14.2
first-FT only	70.7	61.6	43.0	15.5
"|"-only	79.2	63.0	44.9	16.2
comma style	76.8	62.4	45.5	17.6

表3：消融实验

1. 相比于所有基线模型，COFT表现出更优的正确率，在零样本场景带来了最多35.3%的正确率提升
2. COFT在复杂任务上带来的提升优于其他参数微调方法
3. COFT应用于其他微调方法后效果进一步提升
4. 探索合适的COFT格式能够进一步提升大语言模型的代码生成能力