

融合静态分析与大模型推理的多语言代码异味诊断方法

陈欢 马文亭 姜楠楠 龚丹 田鑫宇 张海南*
哈尔滨华德学院(数据科学与信息工程学院)

DOI:10.32629/pe.v4i4.21750

[摘要] 针对静态规则跨语言适应不足以及大模型置信不稳的问题,构建出证据标准化、约束重复推理、冲突检测以及概率校准协同方法。选用四种语言六百条代码单元,按照项目来划分训练、验证以及测试集,同时比较静态、零样本、工具增强、融合以及消融配置。结果显示,完整模型宏平均指标为0.9315,校准误差为0.0738,四种语言均保持较高水平;去除推理一致性之后校准误差明显增加。该方法兼顾诊断准确性、概率可靠性以及证据可追溯性,能够用于在线评测以及代码审查,并且为争议样本提供复核路径以及参数更新依据。

[关键词] 静态分析; 大语言模型; 代码异味; 多语言诊断; 置信校准

中图分类号: TU311.1 **文献标识码:** A

Multilingual Code Smell Diagnosis via Static Analysis and LLM Reasoning

Huan Chen Wenting Ma Nannan Jiang Dan Gong Xinyu Tian Hainan Zhang*

Harbin Huade College (School of Data Science and Information Engineering)

[Abstract] Aiming at the problems of insufficient cross-language adaptation of static rules and unstable confidence of large models, a collaborative method of evidence standardization, constrained repetitive reasoning, conflict detection and probability calibration is constructed. Select six hundred code units in four languages, divide the training, validation and test sets by project, and compare the static, zero-shot, tool-enhanced, fusion and ablation configurations at the same time. The results showed that the macro average index of the complete model was 0.9315, the calibration error was 0.0738, and all four languages maintained a high level; after removing the consistency of reasoning, the calibration error significantly increased. This method takes into account diagnostic accuracy, probability reliability, and evidence traceability, and can be used for online evaluation and code review, and provides a review path and parameter update basis for disputed samples.

[Key words] static analysis; LLM; code smell; confidence calibration

引言

传统检测依靠阈值规则、语法树模式以及代码度量,结果可复核并且执行迅速,但对命名语义、跨函数依赖以及语言惯例适应不足。大语言模型能够结合上下文来解释代码,已用于修复、定位以及测试生成,不过其判断受提示结构、随机采样以及语义泛化影响,难以独立保证稳定。相关综述指出,模型能力提升并未自动解决证据不可追溯以及评价口径不统一的问题^[1]。诊断链因而需要保留规则、语义以及校准依据,从而形成闭环。

现有工作多沿静态分析或大模型推理单线优化,规则命中、复杂度和耦合信息尚未充分转化为模型可核验的约束证据,语义概率也常缺少校准,高置信误报因而难以进入自动评测闭环。缺陷检测与修复研究表明,大模型具备理解复杂代码的潜力,但工程应用仍受可靠性和成本限制^[2]。配置研究进一步说明,多源证据缺少统一表达和冲突处理时,结论难以保持一致^[3]。据此,

本文组合证据标准化、约束推理、一致性量化与概率校准,并以项目级独立测试验证性能、可靠性、开销及工程适用边界。

1 诊断问题与总体框架

1.1 多语言异味表征

多语言诊断首先面临语法结构以及工程约定不一致的问题。两类静态型语言侧重类、方法以及访问修饰关系,两类动态型语言允许更灵活的函数组织、属性以及闭包;要是直接共享固定阈值,五类异味的判定边界就会随语言漂移。本文把样本统一映射为代码单元,保留语言、单元类型、代码行数、圈复杂度、耦合度以及内聚度等字段,同时将异味转换为二元目标。语言适配解析器只提取结构,不直接决定标签,让融合层能够在统一证据空间来比较,同时避免单一语言的规则偏好扩散。该表示兼顾差异与统一建模,结构见图1。

1.2 静态证据与推理协同

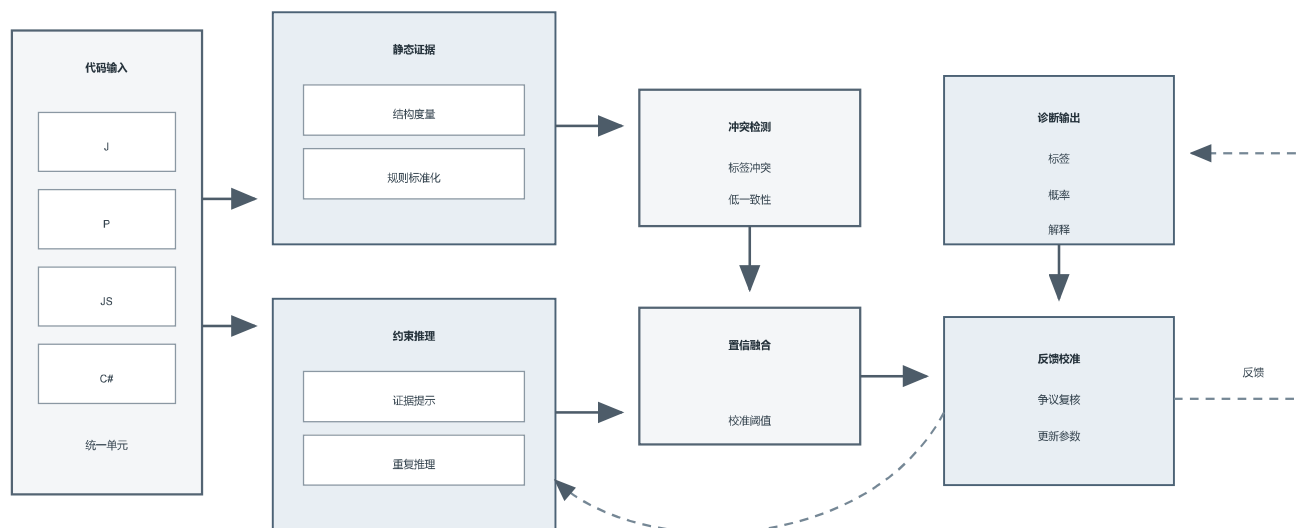


图1 融合诊断框架

图1展示诊断主链以及反馈链。输入代码经过语言适配解析之后形成统一单元，静态链提取度量以及规则投票并且生成标准化证据；推理链把证据、代码摘要以及候选异味约束写入固定模板，借助低温重复推理得到语义概率与一致性。两链在冲突检测节点汇合，对标签冲突、低一致性以及边界概率作标记，再由校准融合模块输出标签、置信度以及解释。相关研究表明，语义归纳可补足传统静态摘要的不足^[4]。反馈链仅依据验证集以及复核结果更新参数，不接触测试集，从而保持评价隔离、结论可追溯以及参数可复现。

2 融合诊断方法设计

2.1 规则证据标准化

静态证据由代码行数、圈复杂度、耦合度、内聚度以及规则投票构成。为消除量纲以及语言分布差异，各特征的均值与标准差只在训练集估计，再按照验证集确定的权重来聚合，计算形式见式(1)。聚合结果经过逻辑函数映射到零至一区间，既形成静态基线的连续得分，也作为约束推理以及最终融合的输入。该过程把离散命中转换为可比较强度，缓解固定阈值在不同语言间造成的突变，同时禁止测试集统计量进入标准化过程，防止评价信息泄露，保证口径一致。

$$z_s = \sum_{i=1}^n w_i (x_i - \mu_i) / (\sigma_i + \epsilon) \quad (1)$$

2.2 大模型约束推理

大模型推理使用固定角色、证据字段和输出结构，要求先核对静态证据，再判断候选异味是否满足结构与语义条件，最后返回标签、概率和简短解释。对同一代码单元执行五次低温推理，并按式(2)计算众数标签的一致率。推理次数、单次标签、示性函数和众数均按式中定义处理；一致性取零点六、零点八或一，分别对应三次、四次或五次结论一致。该值不替代类别概率，而作为独立不确定性证据进入融合。当语义概率较高但一致性偏低时，系统触发冲突标记，避免将随机波动误判为稳定知识，并提供复核入口。

$$q_i = (1/R) \sum_{r=1}^R [y_r = \text{mode}(y_1, \dots, y^R)] \quad (2)$$

2.3 置信度融合与校准

融合层同时接收静态得分、语义对数几率和推理一致性，按式(3)计算异味概率并依据阈值输出标签。三个权重分别控制各类证据贡献，偏置用于校正整体倾向，阈值决定最终类别。实验先在训练集拟合基础参数，再在验证集搜索校准尺度、偏置和阈值；最终尺度取一点八、偏置取零点八、阈值取零点五，测试集只用于一次性评价。静态证据可在语义判断犹豫时提供稳定支撑，大模型能在规则边界附近纠正机械误报，一致性则抑制不稳定推理的权重。

$$p_f = \sigma(\alpha z_s + \beta z_l + \gamma q_l + b), \quad \hat{y} = I(p_f \geq \tau) \quad (3)$$

校准目标旨在让预测概率与实际正确率相匹配。本文以宏平均指标以及期望校准误差构成联合准则：前者平衡异味与正常样本表现，后者衡量置信度与真实正确率的偏差。要是候选参数的宏平均指标接近，那就优先选取校准误差更低者；要是校准导致召回明显下降，那就保留未校准模型作对照。冲突样本依据静态判定、语义判定、一致性以及最终概率归入语义泛化误报、包装调用漏检或者访问器模式漏检。该设计既服务自动评测，还指出补充规则、调整提示以及人工复核的环节，避免指标偏差掩盖可靠性下降。

3 实验设置与评价方案

3.1 数据集与基线配置

实验数据包含六百条受控代码单元记录，四种语言等量分布，共对应八十个项目。每种语言均保持异味样本多于正常样本，五类异味数量相同，避免类别比例掩盖语言效应。数据按照项目编号以七成、一成五以及一成五来划分，形成训练、验证以及测试三个互斥集合，同一项目不跨集合。项目级隔离减少同源代码泄漏，让测试结果更接近未见项目的泛化表现，同时可追踪错误来源以及分布。

经统计表明,四种语言的代码单元数、异味与正常比例完全一致,项目数也均为20,跨语言比较不受规模失衡干扰。双人一致率均超过九成,动态语言之间仍存在细微差异;对应仲裁率均低于一成。JavaScript的动态对象、访问器以及闭包结构让数据类以及特征依恋边界更难统一,不过标签噪声仍处于可控范围。总体一致率超过九成,少量争议经过仲裁消解。

3.2 指标与消融设计

基线囊括静态分析、零样本推理、工具增强、未校准融合、完整融合以及去一致性消融六种配置。分类性能采用精确率、召回率以及宏平均指标,概率可靠性采用十区间期望校准误差,效率采用单代码单元平均时延。所有配置共享测试集、静态规则、提示模板、模型版本以及温度;融合以及消融模型使用相同的静态与推理开销。相关研究强调,缺陷触发条件、测试语言以及定位方式需要多维评价,不能只依据单一准确率^[5]。因此,本文同步检查语言分组、错误类型以及校准变化,同时并列呈现准确性以及推理代价,判断适用边界。

4 结果验证与分析

4.1 多语言诊断性能

独立测试集上,完整模型的宏平均指标达到零点九三一五,四种语言均超过零点九。动态语言中静态证据与语义判断的互补更明显,静态型语言则出现少量正常代码误报。模型识别出绝大多数异味样本,仅有两个漏检和四个误报,召回表现较稳定。错误没有集中于单一异味,而主要位于包装调用、访问器结构及语义边界附近。该结果说明,项目级隔离后模型仍保持跨语言泛化,性能并非来自同项目代码泄漏,也不是对固定规则模式的简单记忆。

表1 多语言诊断性能、校准与效率对比

方法配置			分类性能				可靠性与效率	
模型	静态	推理	精确率	召回率	F1	宏平均 F1	ECE	时延/ms
静态模型	是	否	0.8431	0.7818	0.8113	0.7775	0.1171	8.2
零样本	否	直接	0.8654	0.8182	0.8411	0.8102	0.1070	682.3
工具增强	是	约束	0.9259	0.9091	0.9174	0.8987	0.1681	806.0
未校准	是	约束	0.9615	0.9091	0.9346	0.9218	0.0872	806.0
完整模型	是	约束	0.9298	0.9636	0.9464	0.9315	0.0738	806.0
去一致性	是	约束	0.9615	0.9091	0.9346	0.9218	0.1276	806.0

表1显示,静态模型表现最低,加入工具证据后语义判断明显改善,未校准融合有所提升,但概率可靠性仍有不足。完整模型的宏平均指标达到0.9315,校准误差降至0.0738;去除一致性后,分类指标回落且校准误差明显增大,说明一致性主要改善概

率稳定,而非简单翻转标签。静态链耗时很低,完整链约需八百毫秒。因此,该方法适合离线评测、代码审查和高风险提交,可通过分层触发控制调用频率,不宜直接替代毫秒级编辑器提示。

4.2 组件贡献与错误类型

完整模型在92条测试记录中产生6个错误,其中4个为语义泛化误报,集中于静态型语言的正常样本;两个漏检分别涉及动态语言的包装调用以及访问器模式。去除一致性之后校准误差明显增大,证明该项能够抑制边界样本的过度置信。

5 结论

本文构建出静态证据标准化、约束重复推理、冲突检测以及概率校准一体化的多语言代码异味诊断方法。基于六百条、四种语言以及八十一个项目的隔离数据,完整模型在独立测试集取得零点九三一五的宏平均指标以及零点零七三八的校准误差,较静态模型同时改善分类能力以及置信可靠性。消融结果说明,推理一致性是控制概率可靠性的关键证据。当前限制是时延较高并且范围有限;后续需要借助真实语料来验证,同时研究增量学习以及模型蒸馏。

本文系2025年度黑龙江省教育科学规划重点课题项目,项目名称:基于数智赋能的高校项目式编程教学改革与实践,项目编号:GJB1425150。

本文系2026黑龙江省教育教学科学规划课题-D青年课题项目,项目名称:新工科背景下计算机专业跨学科资源库建设与应用,项目编号:GJD1526019。

本文系2025年度黑龙江省教育科学规划重点课题项目,项目名称:产教融合共同体驱动的“六维协同+四链融合”创新型人才培养研究与实践,项目编号:GJB1425149。

[参考文献]

- [1]许鹏宇,况博裕,苏锐,等.基于大语言模型的自动代码修复综述[J].计算机研究与发展,2025,62(8):2040-2057.
- [2]香佳宏,徐霄阳,孔繁初,等.大模型在软件缺陷检测与修复的应用发展综述[J].软件学报,2025,36(4):1489-1529.
- [3]周书林,李姗姗,董威,等.软件运行时配置研究综述[J].软件学报,2024,35(1):63-86.
- [4]汪海沛,孙福特,张磊.一种利用大语言模型增强静态分析代码摘要能力的方法[J].小型微型计算机系统,2025,46(5):1177-1183.
- [5]何浩辰,李姗姗,贾周阳,等.基础软件性能缺陷检测研究综述[J].计算机学报,2025,48(1):210-233.

*通讯作者:

张海南(1992--),女,汉族,黑龙江绥化人,硕士,助教,研究方向:信息与通信工程。