PR #27300 完整报告

sgl-project/sglang

fix(spec): complete CustomSpecAlgo duck-typing interface and guard against drift

合并时间：2026-06-05 06:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27300

执行摘要

- 一句话：修复 CustomSpecAlgo 接口缺失并添加一致性守卫
- 推荐动作：值得精读，展示了如何通过运行时守卫维护鸭子类型接口一致性，可在类似插件系统中借鉴。同时体现了尽早失败（fail-fast）的设计原则。

功能与动机

PR body 指出：CustomSpecAlgo 文档声明其鸭子类型 SpeculativeAlgorithm 枚举，但实际缺失了 is_some()（被 overlap_utils.py 调用）和 is_frozen_kv_mtp() 方法，且硬编码的保留名列表已经漂移（遗漏了 FROZEN_KV_MTP）。根本原因是契约只存在于 docstring，没有强制执行的机制。

实现拆解

1. 补齐接口方法：在 CustomSpecAlgo 中添加 is_some()（返回 True）和 is_frozen_kv_mtp()（返回 False），解决 overlap_utils.py 中调用 spec_algo.is_some() 时可能引发的 AttributeError。
2. 动态推导保留名：将模块级常量 _RESERVED_NAMES 替换为函数 _reserved_names()，该函数通过导入 SpeculativeAlgorithm 枚举并遍历所有成员名称，自动包含新增加的枚举值（如 FROZEN_KV_MTP 之前被遗漏），同时保留手动维护的别名集合 _RESERVED_ALIASES（包含 NEXTN），解决保留名列表漂移问题。
3. 添加接口一致性守卫：实现 _assert_custom_spec_algo_conforms(spec_class)，在 register_algorithm 中调用。该函数使用 vars(SpeculativeAlgorithm) 收集所有以 is_ 或 supports_ 开头的方法名，与 spec_class 的 dir() 对比，若缺少任何方法则抛出 TypeError。由于循环导入问题，守卫在注册时（而非模块加载时）执行。
4. 更新注册逻辑：register_algorithm 中使用 _reserved_names() 替代 _RESERVED_NAMES 做保留名检查，并在构造 spec_class 实例前调用 _assert_custom_spec_algo_conforms，确保插件类符合接口契约。
5. 补充测试：修改 test_spec_registry.py：修复因重命名导致的导入错误；新增测试验证保留名覆盖所有枚举成员、is_some 语义与枚举一致、守卫拒绝缺失方法的子类；在现有接口测试中增加对 is_frozen_kv_mtp() 和 is_some() 的断言。

关键文件：

- python/sglang/srt/speculative/spec_registry.py（模块 推测解码；类别 source；类型 core-logic；符号 CustomSpecAlgo.is_some, CustomSpecAlgo.is_frozen_kv_mtp, _reserved_names, _assert_custom_spec_algo_conforms）：核心变更文件：添加 is_some/is_frozen_kv_mtp 方法，引入动态保留名推导和接口一致性守卫。
- test/registered/unit/spec/test_spec_registry.py（模块 单元测试；类别 test；类型 test-coverage；符号 TestConformanceGuard, test_base_custom_spec_algo_conforms, test_conforming_subclass_passes, test_reserved_names_cover_all_enum_members）：测试配套：新增对保留名派生、is_some 语义、守卫拒绝的测试，覆盖变更的主要路径。

关键符号：CustomSpecAlgo.is_some, CustomSpecAlgo.is_frozen_kv_mtp, _reserved_names, _assert_custom_spec_algo_conforms, register_algorithm

评论区精华

无实质性 review 讨论，变更由作者在 PR body 中详细说明后直接合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：守卫可能在注册时拒绝缺少必要方法的合法插件，但这是设计意图，且错误信息清晰。动态保留名推导确保新增枚举成员自动被保留，减少手动遗漏风险。测试覆盖了正常和异常路径，未发现回归迹象。
- 影响：对用户无直接影响，因为内置算法不受影响。对自定义推测算法开发者，若插件子类未实现全部 is_* / supports_* 方法，注册时会收到 TypeError，开发者需按错误提示补齐方法。此变更有助于提前暴露兼容性问题，提升系统稳定性。
- 风险标记：插件接口兼容性风险，守卫可能过严但符合预期

关联脉络

- PR #26859 FrozenKVMTPVerifyInput: add _draft_preprocess_idle call for when all requests in the verify batch finish in the same iteration: 直接关联：该 PR 引入了 FrozenKVMTP 算法，而本 PR 修复了 CustomSpecAlgo 缺少 is_frozen_kv_mtp 方法的问题，防止该算法注册后出现 AttributeError。
- PR #27193 Replace skip_attn_backend_init with a batch-carried attention plan marker (+ staleness re-plan): 关联近期的 speculative decoding 重构，same spec module 内的接口调整，本 PR 进一步加固了接口一致性。