

PR #34478 完整报告

sgl-project/sglang

[Spec] Support output logprobs with DSpark

合并时间: 2026-08-18 05:23

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34478>

执行摘要

本 PR 为 DeepSeek-V4 DSpark 投机解码路径补齐了 output-logprobs 支持: DSpark 不再拒绝 `return_logprob` 请求, accepted-token 的 logprobs 复用共享的投机 v2 处理器计算, 并重新启用了既有的 DSpark grammar/logprob 集成测试。影响面收敛在 DSpark 投机链路内, 风险有限。注意: 本次提取的上下文不含文件级 diff (changed_files_count 为 0 且与 3 次提交矛盾), 以下实现细节基于提交消息与仓库结构推断。

功能与动机

PR 正文给出三点目标:

- 允许 output-logprob 请求与 DSpark 一起工作;
- 用共享的 speculative v2 处理器计算 accepted-token logprobs;
- 启用既有的 DSpark grammar/logprob 集成覆盖。

投机解码路径一般会生成并被丢弃多余的 draft token, 如果 accepted-token 的 logprob 不能正确投影回最终返回的序列, 下游依赖 token 级 logprobs 的评估、搜索和可视化工具将无法在 DSpark 上工作; 此外 DSpark 的集成覆盖此前被禁用, 回归无保护。该 PR 没有关联 Issue。

实现拆解

当前上下文没有提供文件级的 diff, 以下步骤基于提交消息与仓库结构推断:

1. 放开 output-logprob 限制: DSpark 的投机 worker (推断为 `python/sglang/srt/speculative/dspark_worker.py`) 此前应存在拒绝或跳过 `return_logprob` 请求的守卫, 本次将其移除, 使请求进入投机流水线。
2. 接入共享处理器: accepted-token 的 logprob 计算接到共享的投机 v2 处理器上, 而不是在 DSpark 内部复制一份实现, 避免两套 logprob 语义漂移。
3. 修复 import 顺序: merrymercy 提交 Fix DSpark worker import ordering, 调整了 DSpark worker 模块的导入顺序以满足 Ruff 检查。
4. 重新启用集成覆盖: 恢复既有的 DSpark grammar/logprob 集成测试, 使 DSpark 的语法约束与 logprob 行为回到 CI 回归视野。
5. 测试配套: PR 正文列出 Python 语法编译、Ruff 导入 / 错误检查, 以及 model-backed DSpark 集成覆盖; 具体测试文件路径未被提取, CI 记录显示 Extra test 有一次失败 (Run #32060676187)。

本次提取的上下文未包含该 PR 的任何文件级 diff 与源码片段，无法按仓库要求给出经过整理的相对完整实现片段。若需精确定位，建议重新抓取 PR 34478 的完整 patch，重点关注 DSpark worker 与投机 v2 处理器接缝处的 accepted-token logprob 语义。

评论区精华

- Review 评论为 0，没有技术代码评审。
- 两条 Issue 评论都是 /tag-and-rerun-ci，属于 CI 重跑命令。
- 提交历史显示 merrymercy 在合入前补了一次 Fix DSpark worker import ordering，说明协作过程中有 lint/ 导入顺序的修整回合。

风险与影响

- logprobs 对齐正确性：投机路径中 accepted-token 的 logprob 到底取自 draft 分布还是 target 分布，必须与 specv2 处理器的既有语义一致；不一致会导致 token 与 logprob 静默错位，这是风险最高的点，且在无 diff 的情况下无法核验。
- CI 稳定性：重新启用的 model-backed 集成覆盖在 Extra test 上出现过一次失败（Run #32060676187）；需关注是环境性失败还是 DSpark 行为回归。
- 兼容性与性能：影响面限于 DSpark/DSV4 路径；logprobs 计算带来少量显存与带宽开销，但复用处理器避免了额外的数据搬运。

关联脉络

- 33676（NPU 支持 DeepSeek-V4 DSpark 与 DSV4 缓存管理重构）：DSpark 功能线的奠基 PR，本 PR 在其上补齐 logprobs；
- 35059（[Spec] 仅依 backend 声明解析共享读终点）：specv2 体系内的处理器梳理，本 PR 复用的投机 v2 处理器同属这条线；
- 34801（[PD] HiCache 撤回时保留 decode KV）：同样触及 `base_spec_worker` 与 DSpark 解码路径的调度语义，与本 PR 同属投机解码健壮性方向。