

PR #29681 完整报告

sgl-project/sglang

fix(mlx): default prefill_aware_swa=False on MlxModelRunnerStub

合并时间: 2026-07-01 13:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29681>

执行摘要

- 一句话: 修复 MLX 后端 prefill_aware_swa 属性缺失崩溃
- 推荐动作: 建议合并。这是一个关键但简单的 bugfix, 修复了 MLX 后端的启动崩溃。值得注意的设计是采用类属性默认值来弥补轻量初始化跳过的状态设置, 与已有的 canary_manager 处理一致。

功能与动机

修复 Issue #29679: 用户在 Apple Silicon (MLX) 上运行 `sglang.launch_server` 时, 无论使用何种模型, 都会立即崩溃, 错误为 `'MlxModelRunnerStub' object has no attribute 'prefill_aware_swa'`。

实现拆解

1. 添加类属性: 在 `python/sglang/srt/hardware_backend/mlx/model_runner_stub.py` 的 `MlxModelRunnerStub` 类中新增 `prefill_aware_swa = False` 类属性。
2. 保持一致模式: 参考已有的 `canary_manager = None` 处理, MLX 路径上的轻量初始化跳过了基类 `ModelRunner.initialize()` 中从模型 `is_prefill_aware_swa()` 方法推导属性的逻辑, 因此直接默认设为 `False`。
3. 无行为变化: 非 MLX 路径不受影响, 因为基类 `ModelRunner` 在完整初始化时仍会正常设置该属性。

关键文件:

- `python/sglang/srt/hardware_backend/mlx/model_runner_stub.py` (模块 MLX 后端; 类别 source; 类型 data-contract): 唯一改动文件, 在 `MlxModelRunnerStub` 类中添加 `prefill_aware_swa` 类属性默认值。

关键符号: 未识别

关键源码片段

`python/sglang/srt/hardware_backend/mlx/model_runner_stub.py`

唯一改动文件, 在 `MlxModelRunnerStub` 类中添加 `prefill_aware_swa` 类属性默认值。

```
class MlxModelRunnerStub(ModelRunner):  
    # ... 其他代码 ...
```

```
# No KV canary on the MLX path. The base ModelRunner installs it via
# install_canary() in its full initialize(), which this lightweight override
# skips. Downstream consumers (scheduler, cuda graph runner, speculative
# workers) all guard with `canary_manager is not None`, so default to None
# as a class attribute to keep those checks working instead of raising
# AttributeError.
canary_manager = None

# No prefill-aware SWA on the MLX path. The base ModelRunner derives this in
# its full initialize() from `model.is_prefill_aware_swa()`, which this
# lightweight override skips (and `_DummyModel` does not implement). The
# scheduler reads `model_runner.prefill_aware_swa` unconditionally when
# admitting a prefill batch, so default to False as a class attribute to keep
# that path working instead of raising AttributeError.
prefill_aware_swa = False

def __init__(self, *args, mlx_pool_size: int | None = None, **kwargs):
    self._mlx_pool_size = mlx_pool_size
    super().__init__(*args, **kwargs)
```

评论区精华

无审核讨论；PR 由 yeahdongcn 直接批准，变更意图清晰，改动简单。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅新增一个布尔类属性，默认值 `False` 与 MLX 路径的实际行为一致（MLX 不使用 prefill-aware SWA），不会影响其他后端。
- 影响：影响范围限于 MLX（Apple Silicon）后端用户：之前 `sglang.launch_server` 启动后立即崩溃，修复后可正常启动并使用。其他用户无影响。
- 风险标记：暂无

关联脉络

- PR #29271 fix: make `write_token` dynamic: 同样涉及 MLX 后端的 bugfix，属于同一维护线。