PR #26522 完整报告

sgl-project/sglang

[NemotronH] Fix weight-loading unit test broken by Puzzle support

合并时间：2026-05-28 08:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26522

执行摘要

- 一句话：修复 NemotronH 权重加载测试因配置键缺失而失败
- 推荐动作：合并即可，无需精读。

功能与动机

修复 CI 中 test_nemotron_h_weight_loading.py 的失败，原因是 load_weights 需要 config.max_n_routed_experts 但伪造配置未提供该属性。

实现拆解

将测试文件 test/registered/unit/models/test_nemotron_h_weight_loading.py 中 _make_minimal_model 方法的第 38 行从 SimpleNamespace(n_routed_experts=2) 改为 SimpleNamespace(n_routed_experts=2, max_n_routed_experts=2)，使伪造配置与真实 NemotronHConfig 的行为对齐。

关键文件：

- test/registered/unit/models/test_nemotron_h_weight_loading.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，修复测试夹具缺失的配置属性。

关键符号：_make_minimal_model

评论区精华

无讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：仅修改测试夹具，生产代码不变。
- 影响：影响范围仅限于修复该单元测试在 CI 中的失败，无其他影响。
- 风险标记：暂无

关联脉络

- PR #24429 Support NemotronHPuzzleForCausalLM: 引入 max_n_routed_experts 配置属性，导致本测试失败。
- PR #24434 [NemotronH] Add weight loading unit test: 添加了本 PR 修复的测试用例，但未包含 max_n_routed_experts。