PR #21416 完整报告

sgl-project/sglang

Update Nemotron Example docs to include Super v3 and Nano 4B

合并时间：2026-03-26 00:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21416

执行摘要

本次 PR 更新了 SGLang 的生成模型文档，添加了 NVIDIA Nemotron 3 Super 和 Nano 4B 条目，确保文档反映最新的模型支持，对系统无技术影响，用户可获取更新信息。

功能与动机

动机基于 PR 正文表述："Update docs to provide new generation of NVIDIA Nemotron models in the supported model examples"，旨在保持文档同步，以支持平台扩展能力。这解决了用户可能无法获取最新模型信息的问题，便于他们选择和使用适当的模型。

实现拆解

变更集中在单一文件 docs/supported_models/text_generation/generative_models.md，具体改动如下：

- 添加了 NVIDIA Nemotron 3 Super 条目：| **NVIDIA Nemotron 3 Super**(NVIDIA) |nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4| The [NVIDIA Nemotron](https://www.nvidia.com/en-us/ai-data-science/foundation-models/nemotron/) 3 Super is a 120B-parameter MoE model (12B active) delivering high-quality reasoning and generation for enterprise AI agents. |
- 添加了 NVIDIA Nemotron 3 Nano 条目：| **NVIDIA Nemotron 3 Nano**(NVIDIA) |nvidia/NVIDIA-Nemotron-3-Nano-4B-BF16| The [NVIDIA Nemotron](https://www.nvidia.com/en-us/ai-data-science/foundation-models/nemotron/) 3 Nano is a compact model designed for efficient edge and enterprise deployment with strong reasoning capabilities. | 此实现直接扩展了支持模型列表，无其他技术变更。

评论区精华

Review 中无深度讨论，仅有自动化工具和简单批准：

- gemini-code-assist[bot] 评论："This pull request updates the generative_models.md documentation by adding entries for two new NVIDIA Nemotron models... There are no review comments to address."
- b8zhong 直接批准，表明变更无争议，被快速接受。

风险与影响

- 风险分析：风险极低，主要集中于文档准确性。变更仅涉及文本更新，无代码逻辑改动，但需验证添加的模型标识符和描述是否正确，避免误导用户。无回归、性能、安全或兼容性风险。
- 影响分析：影响用户文档，提升了信息的时效性和完整性；对系统核心功能无影响，维护团队需确保文档与其他组件同步。影响范围限于用户参考文档，程度轻微。

关联脉络

与其他文档更新 PR 类似，如 PR #21373（整合扩散模型文档）和 PR #20846（更新 Ascend 文档），体现 SGLang 项目对文档维护的持续投入。这些 PR 共同展示了跨模块文档同步的趋势，确保用户体验一致性。未发现直接的功能演进关联，更多是常规维护流程的一部分。