PR #21746 完整报告

sgl-project/sglang

[diffusion] Fix typo

合并时间：2026-03-31 17:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21746

执行摘要

此 PR 修复了扩散模型模块中一个注释行的拼写错误，将 return_s•oftmax_lse 更正为 return_softmax_lse。变更仅影响代码可读性，不改变运行时行为，但 review 中指出了潜在的兼容性问题未被解决。

功能与动机

动机是修复注释中的拼写错误（"Fix typo."），以提高代码准确性和可读性。这源于开发过程中的小错误修正。

实现拆解

修改了文件 python/sglang/multimodal_gen/runtime/layers/usp.py 的第 213 行： - # logger.warning(f"Warning: return_s•oftmax_lse is only supported for FlashAttentionImpl") + # logger.warning(f"Warning: return_softmax_lse is only supported for FlashAttentionImpl") 该变更仅修正注释字符串，不涉及任何代码逻辑。

评论区精华

review 中，gemini-code-assist[bot] 评论道：

> "This commented-out warning identifies a significant limitation: the attn_callable_adapter is tightly coupled with FlashAttentionImpl... consider adding an explicit compatibility check..." 此评论揭示了注释背后的设计问题，但 PR 作者未回应，变更只限于拼写错误修复。

风险与影响

风险：直接风险极低，因为是注释变更。但 review 中提到的兼容性问题（如 return_softmax_lse=True 和 attn_metadata=None 在其他后端可能导致崩溃）仍存在，未被本 PR 解决。 影响：对用户和系统无影响；对团队，提高了代码可读性，但遗留了潜在的设计隐患。

关联脉络

与本 PR 相关的历史 PR 包括：

- PR #21664 "[diffusion] Fix Flux.2"：同属扩散模型模块的 bugfix。
- PR #21621 "[AMD] Fix CI multimodal-gen-test-1-gpu-amd for gen model"：涉及扩散模型和 JIT 内核的修复。 这些 PR 表明扩散模型模块是活跃的开发区域，常有小修复和优化。