PR #21537 完整报告

sgl-project/sglang

[NPU] recover accuracy for gemma3-4b-it from 54% to 72% (reduced by transformer5.3)

合并时间：2026-05-13 16:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21537

执行摘要

- 一句话：修复 Gemma3 NPU 因 transformers 升级导致的精度下降
- 推荐动作：值得合并：这是一个精准的 bugfix，以极小改动（+2/-1）修复了因上游依赖变更导致的精度回归，并提供了明确的测试数据证明改善。建议在类似模型变更中同步检查 transformers 5.x 的兼容性。

功能与动机

PR body 明确指出："Previously, accuracy of gemma 3-4b-it was degraded by the update of transformers from 4.57.1 to 5.3.0." 说明动机是恢复因依赖升级导致的精度回退。

实现拆解

该 PR 仅修改一个文件 python/sglang/srt/models/gemma3_causal.py，在 Gemma3Model.__init__ 中调整了全局 RoPE 参数的构建逻辑：

1. 提取 factor 参数：从 config.rope_parameters["full_attention"]["factor"] 中读取线性缩放因子。
2. 修改 rope_type：将原来的 "rope_type": "default" 改为 "rope_type": "linear"，以匹配 transformers 5.3 中 full attention 使用的线性缩放 RoPE。
3. 保留原有结构：仅变更 global_config 的 rope_parameters 字典，local_config 保持不变。

变更仅涉及 3 行（+2/-1），无测试或配置配套改动。

关键文件：

- python/sglang/srt/models/gemma3_causal.py（模块 模型；类别 source；类型 data-contract）：唯一修改的文件，核心变更：调整全局 RoPE 配置以适配 transformers 5.3，恢复 Gemma3 NPU 精度。

关键符号：Gemma3Model.init

关键源码片段

python/sglang/srt/models/gemma3_causal.py

唯一修改的文件，核心变更：调整全局 RoPE 配置以适配 transformers 5.3，恢复 Gemma3 NPU 精度。

```
# In Gemma3Model.__init__, building global RoPE config for full attention
# Previously (transformers v4 path) used "rope_type": "default" without factor.
# Transformers v5 provides nested rope_parameters: {"full_attention": {"rope_type": ..., "rope_
theta": 1000000, "factor": 1.0}}
# The fix reads the factor and sets rope_type to "linear" to enable linear scaling.
global_config = copy.deepcopy(config)
global_config.rope_parameters = {
    "rope_theta": global_theta,
    "factor": config.rope_parameters["full_attention"]["factor"], # NEW: read factor from v5
    config
    "rope_type": "linear", # CHANGED: from "default" to "linear"
}
self.rotary_emb = Gemma3RotaryEmbedding(config=global_config)
```

评论区精华

核心讨论围绕变更的必要性展开：

- reviewer iforgetmyname 质疑："bad changes, why we need to change rope parameters?"（为什么需要改变 RoPE 参数？），但 PR 作者未在讨论中回应。
- gemini-code-assist[bot] 建议使用模块级变量 _is_npu 代替函数调用 is_npu()，但该建议针对的是最终未包含在提交中的 NPU 条件判断代码（当前变更无 NPU 硬件判断）。
- sglang-npu-bot 两次批准该 PR。
- 变更必要性质疑 (question)：未直接回应，但 PR 最终被批准合并，且精度测试数据支持变更的必要性。
- NPU 条件判断代码风格 (style)：该建议针对的 NPU 条件判断未出现在最终提交中，已过时。

风险与影响

- 风险：风险较低。变更仅影响 Gemma3 模型初始化时的 RoPE 配置，且仅针对 transformers 5.3+ 版本（通过 rope_parameters 嵌套结构触发新路径）。潜在风险：
  - 若 transformers 回退至 4.x，rope_parameters["full_attention"]["factor"] 可能不存在，但已有 else 分支处理 v4 flat 格式，不产生错误。
  - 缺少单元测试覆盖该分支。
- 影响：影响范围有限：
  - 仅影响 Gemma3-4b-it 模型在 NPU 设备上的推理精度（从 54% 恢复至 71.5%）。
  - 不影响其他模型或 GPU 设备。
  - 无性能影响，仅初始化时多读取一个配置字段。
  - 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR