PR #1769 完整报告

THUDM/slime

Support FP8 conversion for Qwen3.5

合并时间：2026-03-29 13:45

原文链接：http://prhub.com.cn/THUDM/slime/pull/1769

执行摘要

本 PR 为 Qwen3.5 模型添加了 FP8 转换支持，通过修改转换脚本过滤特定权重键名，以适配该模型的特殊结构。这是一项有意义的改进，影响范围限定于使用该转换脚本的用户，风险较低但缺少测试覆盖。

功能与动机

PR 旨在支持 Qwen3.5 模型的 FP8 量化转换。动机源于 Qwen3.5 模型可能包含如 'conv1d'、'A_log' 等特殊权重键名，这些在原有转换逻辑中未被排除，可能导致转换错误或兼容性问题。因此，需要扩展过滤条件以确保正确性。

实现拆解

实现集中在 tools/convert_hf_to_fp8.py 文件中的 process_file 函数。关键改动是在权重过滤条件中新增了以下检查：

```
and "conv1d" not in key
and "A_log" not in key
and "dt_bias" not in key
and "in_proj_a" not in key
and "in_proj_b" not in key
```

这扩展了原有逻辑，排除 Qwen3.5 模型中可能不适合 FP8 量化的权重，避免错误处理。

评论区精华

本 PR 未收到任何 review 评论，因此无讨论内容可供分析。

风险与影响

- 风险：过滤条件可能不完整，导致其他 Qwen3.5 权重键名被遗漏；缺少测试覆盖，增加回归风险；未经 review，逻辑正确性未验证。
- 影响：影响使用 tools/convert_hf_to_fp8.py 进行 Qwen3.5 模型 FP8 转换的用户，提升兼容性，但不会波及系统其他部分。

关联脉络

从近期历史 PR 看，如 PR 1721（添加 Qwen3.5-4B 支持）和 PR 1719（修复 Qwen3.5 启动脚本），表明仓库在持续增强 Qwen3.5 模型的支持生态。本 PR 是该趋势的一部分，专注于转换流程的适配，共同推动模型可用性。