PR #1721 完整报告

THUDM/slime

feat: add Qwen3.5-4B model support

合并时间：2026-03-22 16:26

原文链接：http://prhub.com.cn/THUDM/slime/pull/1721

执行摘要

该 PR 在 slime 仓库中添加了 Qwen3.5-4B 模型的配置脚本，补全了模型支持链条，使用户能方便地启动和部署该模型。变更仅限于脚本文件，风险低，影响正面。

功能与动机

根据 PR 描述，仓库已有 Qwen3.5 核心代码支持（如 slime_plugins/models/qwen3_5.py 等文件），但缺少 4B 版本的预设配置脚本。添加此脚本后，下游启动脚本可以直接引用专用配置，简化了密集混合 Qwen3.5-4B 模型的部署流程。作者在 PR body 中明确表示："This PR only adds the missing 4B preset script so that downstream launch scripts can source a dedicated config for the dense hybrid Qwen3.5-4B model。"

实现拆解

实现仅涉及一个文件：slime/scripts/models/qwen3.5-4B.sh。该 shell 脚本定义了 Qwen3.5-4B 的架构参数，关键内容如下：

```
MODEL_ARGS=(
   --spec "slime_plugins.models.qwen3_5" "get_qwen3_5_spec"
   --disable-bias-linear
   --qk-layernorm
   --group-query-attention
   --num-attention-heads 16
   --num-query-groups 4
   --kv-channels 256
   --num-layers 32
   --hidden-size 2560
   --ffn-hidden-size 9216
   --use-gated-attention
   --normalization RMSNorm
   --apply-layernorm-1p
   --position-embedding-type rope
   --norm-epsilon 1e-6
   --rotary-percent 0.25
   --swiglu
   --vocab-size 248320
   --rotary-base 10000000
   --attention-output-gate
)
```

脚本使用 Qwen3.5 特定的 spec 入口点，并设置了如层数、隐藏大小等关键参数。作者已验证这些参数与 HuggingFace Qwen3.5-4B 模型配置匹配，确保正确性。此变更属于 scripts/models 模块，无核心代码改动。

评论区精华

本次 PR 没有 review 评论，因此无讨论内容。变更直接由作者提交并由 zhuzilin 合并，表明变更较为直接，未引发争议。

风险与影响

- 风险：主要风险在于配置参数的正确性，但作者已进行验证；脚本语法错误可能导致启动失败，但由于是简单参数设置，且无核心代码变更，风险可控。无回归风险或安全漏洞。
- 影响：对用户而言，新增了 Qwen3.5-4B 模型支持，扩展了模型选择，提升部署便利性；对系统，仅添加配置文件，不影响现有功能和性能，兼容性良好；对团队，简化了配置工作流，减少手动参数设置，便于后续维护和扩展。

关联脉络

从历史 PR 分析看，类似变更集中在模型配置脚本的修复和优化，体现了仓库对配置脚本的持续维护趋势。相关 PR 包括：

- PR #1719：修复 Qwen3-235B-A22B 启动脚本的 JSON 格式问题，涉及类似配置脚本的修改。
- PR #1689：修复 shell 脚本中变量引用导致的 glob 扩展问题，与本 PR 的脚本风格相关。
- PR #1700：修复 qwen3-4B 脚本的 GPU 检测问题，属于同一模型系列配置的优化。 这些 PR 共同显示了仓库在模型配置支持方面的演进，本 PR 是这一趋势的延续，补全了 Qwen3.5 系列的配置支持，为未来类似模型添加提供参考。