PR #37292 完整报告

vllm-project/vllm

Fix Mistral yarn warning in Transformers v5

合并时间：2026-04-07 21:23

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37292

执行摘要

本 PR 修复了 Transformers v5 中 Mistral YaRN 模型的警告问题，通过添加版本检查设置 ignore_keys_at_rope_validation 来抑制警告，确保向后兼容性，属于常规维护性 bugfix。

功能与动机

由于 Transformers 库在 v5 版本中移除了 ignore_keys 参数，改为使用 ClassVar（参考 PR #41250），本 PR 的目标是为使用 YaRN 的 Mistral 模型配置设置新的 ClassVar，以消除 Transformers RoPE 验证中的警告。引用 PR body："As of https://github.com/huggingface/transformers/pull/41250 the ignore_keys argument to validate_rope was removed in favour of ClassVars attached to the config classes themselves."

实现拆解

实现集中在 vllm/transformers_utils/configs/mistral.py 文件的 _remap_mistral_yarn_args 函数中：

- 导入变更：添加 from packaging.version import Version 和 from transformers import __version__ as TRANSFORMERS_VERSION。
- 版本检查：添加条件 if Version(TRANSFORMERS_VERSION) >= Version("5.3.0.dev0"): 确保仅在新版本生效。
- 设置忽略键：在条件内设置 config["ignore_keys_at_rope_validation"] = {"apply_yarn_scaling"} 为一个 set。

关键代码逻辑： if Version(TRANSFORMERS_VERSION) >= Version("5.3.0.dev0"): config["ignore_keys_at_rope_validation"] = {"apply_yarn_scaling"}

评论区精华

Review 讨论中，关键交锋如下：

- 版本检查错误：gemini-code-assist[bot] 错误地建议 Transformers v4.42.0 版本检查，hmellor 回应为 "wild hallucination"。
- 精确版本调整：juliendenize 指出："I had to change to make it >5.2.0 for it to work, hence my suggestion. This is due to 5.3.0.dev being < 5.3.0." 最终采用 5.3.0.dev0 确

保兼容性。

- 类型设置：hmellor 建议将 ignore_keys_at_rope_validation 从 list 改为 set 以避免 JSON 序列化错误，juliendenize 确认。

风险与影响

- 风险：版本检查不准确可能导致在旧 Transformers 版本中引发错误或新版本警告未消除；set 类型需确保 Transformers 库正确处理。变更范围小，风险较低。
- 影响：对用户消除警告，无性能影响；对团队提醒关注外部依赖 API 变更。

关联脉络

与近期 PR 的关联：

- PR 39086 修复 Mistral 版本依赖，涉及类似模型兼容性问题。
- PR 38763 处理 PyTorch 版本兼容性，展示版本 guard 设计模式。 这些 PR 共同反映了团队在适配外部库版本变更时的系统化策略。