

PR #3087 完整报告

radixark/miles

feat: expand GLM 53 TITO model support

合并时间: 2026-09-12 05:26

原文链接: <http://prhub.com.cn/radixark/miles/pull/3087>

执行摘要

- 一句话: 新增 GLM-5.3 文本 TITO 支持, 并加固 session verifier 默认参数
- 推荐动作:

功能与动机

已验证的 GLM-5.3 HF 原生 chat template 即使传入 `enable_thinking=False` 也会以 `thinking` 开头生成内容, PR body 明确说明: “The verified GLM-5.3 template starts generation with `thinking` even when `enable_thinking=False`, so it needs a model-specific rendering contract.” 因此需要为 GLM-5.3 建立独立的 TITO 渲染契约, 否则增量 token 与全量重 token 化会失配。GLM-5.1 被有意排除: 其原生模板与 `glm47` 模板契约不兼容, 需要单独修复和验证。

实现拆解

1. TITO 家族登记: 在 `miles/utils/chat_template_utils/tito_tokenizer.py` 中新增 `GLM53TITOTokenizer`, 继承 `GLM47TITOTokenizer`, 复用其 `merge_tokens` 的歧义边界裁剪 (`<luserl>` / `<lobservationl>`) 与 `trailing_token_ids`; `FIXED_TEMPLATE` 设 `template=None` (依赖 checkpoint 原生模板, 不引入捆绑 `jinja`) 并固定 `extra_kwargs={`

关键文件:

- `scripts/tools/verify_session_tito_tokenizer.py` (模块 `scripts/tools`; 类别 `source`; 类型 `core-logic`; 符号 `_print_action_table`, `_with_session_verify_defaults`): 源码主路径; 涉及符号 `_print_action_table`, `_with_session_verify_defaults`; 包含 导入关系调整、控制流调整、配置键调整; +47/-7
- `tests/fast/utils/test_utils/test_session_verify_runner.py` (模块 `session/verify/runner`; 类别 `test`; 类型 `test-coverage`; 符号 `test_session_verify_defaults_select_fsdp_during_backend_preparse`, `test_namespace_to_train_args_preserves_resolved_rollout_gpu_count`, `test_namespace_to_train_args_emits_cuda_graph_max_batch_size`, `test_action_table_uses_requested_cycle_count`): 测试配套; 涉及符号 `test_session_verify_defaults_select_fsdp_during_backend_preparse`, `test_namespace_to_train_args_preserves_resolved_rollout_gpu_count`, `test_namespace_to_train_args_emits_cuda_graph_max_batch_size`, `test_action_table_uses_requested_cycle_count`; 包含 测试覆盖调整、导入关系调整、配

置键调整; +30/-0

- miles/utils/chat_template_utils/tito_tokenizer.py (模块 miles/utils; 类别 source; 类型 core-logic; 符号 GLM53TITOTokenizer) : 源码主路径; 涉及符号 GLM53TITOTokenizer ; 包含 控制流调整、配置键调整、符号定义调整; +16/-1
- tests/fast/utils/test_arguments.py (模块 arguments; 类别 test; 类型 test-coverage; 符号 test_glm53_uses_native_template, test_glm53_rejects_disabling_thinking) : 测试配套; 涉及符号 test_glm53_uses_native_template, test_glm53_rejects_disabling_thinking; 包含 测试覆盖调整、配置键调整、符号定义调整 ; +22/-0
- miles/utils/test_utils/session_verify_runner.py (模块 miles/utils; 类别 source; 类型 core-logic) : 源码主路径; 包含 控制流调整; +4/-0

关键符号: _print_action_table, _with_session_verify_defaults, test_session_verify_defaults_select_fsdp_during_backend_preparse, test_namespace_to_train_args_preserves_resolved_rollout_gpu_count, test_namespace_to_train_args_emits_cuda_graph_max_batch_size, test_action_table_uses_requested_cycle_count, GLM53TITOTokenizer, test_glm53_uses_native_template, test_glm53_rejects_disabling_thinking

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR