

PR #2202 完整报告

radixark/miles

fix(tito): prevent DeepSeek V4 system-tail mismatch

合并时间: 2026-08-05 11:37

原文链接: <http://prhub.com.cn/radixark/miles/pull/2202>

执行摘要

- 一句话: DeepSeek V4 TITO 禁止 system 尾追加, 修复会话不匹配
- 推荐动作: 值得快速浏览: 这是一个典型的 "一行配置修复系统性随机故障" 的案例, 展示了如何通过收窄能力边界而非改写校验语义来根治问题。重点关注 `allowed_append_roles` 的设计 (与 Qwen3.5/Minimax 家族复用同一模式) 以及通过 E2E 阈值保留漂移可见性的思路。不需要精读大量代码。

功能与动机

PR body 报告 DeepSeek V4 Stage 3 出现 0.109375 到 0.1875 的 assistant-text mismatch 率而不是 0。根因有三点: `DeepSeekV4TITOTokenizer.FIXED_TEMPLATE` 继承了 mid-session `system` 支持; `encoding_dsv4.encode_messages` 在 `system` 尾后省略 assistant opener; `compute_session_mismatch` 用 canonical `</think>` 重渲染与采样 token 比较, 于是产生随机性不匹配。

实现拆解

1. 变更入口: `miles/utils/chat_template_utils/tito_tokenizer.py` 中 `DeepSeekV4TITOTokenizer.FIXED_TEMPLATE` 新增 `allowed_append_roles=frozenset({"tool", "user", "assistant"})`, 把可追加角色从默认的全角色集合收窄为编码器 generation-ready 的边界。
2. 核心逻辑: 该约束使 TITO 增量 tokenization 不再在 mid-session `system` 之后追加消息, 从而避免 `encoding_dsv4` 省略 assistant opener 导致的 roundtrip 不一致; 由于没有修改 parser 或 comparator 语义, 其他模型家族不受影响。
3. 测试配套: `tests/fast/utils/chat_template_utils/test_fixed_templates.py` 将 `DeepSeekV4TITOTokenizer` 加入 `test_restricted_fixed_template_excludes_mid_session_system` 参数化列表, 与 Qwen35、Minimax M25/M27 一同断言角色集合。
4. E2E 配套: `tests/e2e/sglang/test_session_server_multi_role/test_deepseekv4.py` 为 CONFIG 增加 `assistant_text_threshold=0.05`, 让未来编码器行为漂移仍能被阈值暴露, 而不是静默失败。
5. 无其他配置、schema 或部署配套改动。

关键文件:

- miles/utils/chat_template_utils/tito_tokenizer.py (模块 会话分词; 类别 source; 类型 core-logic; 符号 DeepSeekV4TITOTokenizer, FIXED_TEMPLATE, FixedTemplate, allowed_append_roles) : 核心修复点: 为 DeepSeekV4TITOTokenizer 的 FIXED_TEMPLATE 增加 allowed_append_roles 限制, 排除 mid-session system 追加, 从根源解决 encoding_dsv4 不产生 assistant opener 导致的 TITO mismatch。
- tests/fast/utils/chat_template_utils/test_fixed_templates.py (模块 模板测试; 类别 test ; 类型 test-coverage; 符号 test_restricted_fixed_template_excludes_mid_session_system) : 将 DeepSeekV4TITOTokenizer 纳入参数化测试, 验证其 allowed_append_roles 与 Qwen3.5、Minimax M25/M27 一致, 防止未来回归。
- tests/e2e/sglang/test_session_server_multi_role/test_deepseekv4.py (模块 V4 E2E; 类别 test; 类型 test-coverage; 符号 CONFIG, assistant_text_threshold) : 在 DeepSeek V4 E2E 配置中增加 assistant_text_threshold=0.05, 保留未来编码器漂移可见性, 同时避免对现有合理波动过度敏感。

关键符号: test_restricted_fixed_template_excludes_mid_session_system, test_deepseekv4

关键源码片段

miles/utils/chat_template_utils/tito_tokenizer.py

核心修复点: 为 DeepSeekV4TITOTokenizer 的 FIXED_TEMPLATE 增加 allowed_append_roles 限制, 排除 mid-session system 追加, 从根源解决 encoding_dsv4 不产生 assistant opener 导致的 TITO mismatch。

```
class DeepSeekV4TITOTokenizer(TITOTokenizer):
    """DeepSeek V4 — official encoder via sglang's ``encoding_dsv4``.

    Like V3.2, V4 ships no jinja chat_template; miles' ``apply_chat_template``
    routes any V4 tokenizer to the ``chat_template_utils.deepseek`` bridge, and
    TITO incremental tokenization rides that same bridge to stay byte-aligned
    with what the runtime serves.
    """

    reasoning_parser = "deepseek-v4"
    tool_call_parser = "deepseekv4"
    chat_template_kwarg_aliases = _DEEPSEEK_MODE_KWARG_ALIASES

    FIXED_TEMPLATE = FixedTemplate(
        template=None,
        extra_kwargs={"drop_thinking": False},
        # 关键修复: 限制中间会话可追加的角色。
        # encoding_dsv4 在 system 消息之后不会追加 assistant 开启标记,
        # 若允许追加 system 会导致 TITO 会话无法 roundtrip,
        # 出现 0.1 到 0.19 的 assistant-text mismatch。
        allowed_append_roles=frozenset({"tool", "user", "assistant"}),
    )
```

```

_DEFAULT_ASSISTANT_START = "< | Assistant | >"

def __init__(
    self,
    tokenizer: Any,
    chat_template_kwargs: dict[str, Any] | None = None,
    assistant_start_str: str | None = None,
):
    super().__init__(
        tokenizer,
        chat_template_kwargs=chat_template_kwargs,
        assistant_start_str=assistant_start_str or self._DEFAULT_ASSISTANT_START,
        special_token_ids={
            tokenizer.convert_tokens_to_ids("< | User | >"),
            tokenizer.convert_tokens_to_ids("< | Assistant | >"),
        },
    )
    self._assistant_id: int = tokenizer.convert_tokens_to_ids("< | Assistant | >")
    self._think_bracket_ids: set[int] = {
        tokenizer.convert_tokens_to_ids("<think>"),
        tokenizer.convert_tokens_to_ids("</think>"),
    }
    self.trailing_token_ids = frozenset({self._assistant_id} | self._think_bracket_ids)
    # sglang 的 dsv4 parser 仅在请求携带 thinking 时分离 reasoning,
    # 这里显式暴露有效渲染模式，供 session server 转发。
    self.chat_template_kwargs = {
        **self.chat_template_kwargs,
        "thinking": deepseek.V4.render_thinking_enabled(self.chat_template_kwargs),
    }

```

tests/fast/utils/chat_template_utils/test_fixed_templates.py

将 DeepSeekV4TITOTokenizer 纳入参数化测试，验证其 allowed_append_roles 与 Qwen3.5、Minimax M25/M27 一致，防止未来回归。

```

@pytest.mark.parametrize(
    "tokenizer_cls",
    [
        DeepSeekV4TITOTokenizer,
        Qwen35TITOTokenizer,
        MinimaxM25TITOTokenizer,
        MinimaxM27TITOTokenizer,
    ],
)
def test_restricted_fixed_template_excludes_mid_session_system(tokenizer_cls):
    # 这些模型家族与中间会话追加 system 不兼容，必须与默认的全角色集合区分；
    # 若未来编码器支持 system tail，应先更新此断言再放开角色集合。
    assert tokenizer_cls.FIXED_TEMPLATE.allowed_append_roles == frozenset(
        {"tool", "user", "assistant"}
    )

```

tests/e2e/sglang/test_session_server_multi_role/test_deepseekv4.py

在 DeepSeek V4 E2E 配置中增加 `assistant_text_threshold=0.05`，保留未来编码器漂移可见性，同时避免对现有合理波动过度敏感。

```
CONFIG = ModelConfig(  
    # 官方 deepseek-ai/DeepSeek-V4-Flash 是 mxfp4 打包的 routed experts,  
    # Hopper MoE 内核无法服务；此 FP8 重打包与 scripts/run_deepseek_v4.py 一致,  
    # tokenizer 字节级相同。  
    model_name="sgl-project/DeepSeek-V4-Flash-FP8",  
    reasoning_parser="deepseek-v4",  
    tool_call_parser="deepseekv4",  
    tito_model="deepseekv4",  
    tp_size=4,  
    ep_size=4,  
    enable_spec=True,  
    cycles=2,  
    # 修复 system 尾不匹配后保留 0.05 漂移阈值，避免未来编码器行为变化被静默吞掉。  
    assistant_text_threshold=0.05,  
    # V4 按前序 assistant 的 tool_calls 顺序排序 tool_result 块，  
    # sentinel tool_call_id 无法 roundtrip，因此使用通用回滚恢复。  
    tool_call_failure_mode="rollback",  
)
```

评论区精华

本 PR 只有一条审评论（Shi-Dong 回复 "LGTM"），没有实质技术讨论。PR body 中作者自述的 review 焦点有两个：一是核对 `DeepSeekV4TITOTokenizer.FIXED_TEMPLATE.allowed_append_roles` 与编码器 generation-ready tail 一致；二是确认 `assistant_text_threshold=0.05` 只存在于 DeepSeek V4 E2E 配置中，不会影响其他模型。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 行为依赖编码器：若未来 `encoding_dsv4` 开始支持 mid-session system 追加，当前限制会导致 V4 会话在 system 后无法追加角色，需同步放宽 `allowed_append_roles`。
 2. E2E 阈值放宽：`assistant_text_threshold=0.05` 允许最多 5% 的 assistant-text mismatch，可能掩盖小规模真实回归；但该配置仅存在于 DeepSeek V4 E2E 文件，不会波及其他模型。
 3. 改动集中在 `tito_tokenizer.py` 的类级常量，Qwen3.5、Minimax M25/M27 已使用相同限制模式，回归概率低。- 影响：影响范围限定在 DeepSeek V4 家族的 session-server TITO 校验：修复后 E2E 全部 64 个样本通过，总 TITO mismatch 与 assistant-text mismatch 均为 0.0。对用户而言，消除了 Stage 3 训练中随机的 assistant-text 验证噪音；对系统而言，没有改变模板渲染、parser 或 mismatch 比较逻辑，其他模型家族零影响。团队后续维护 V4 时需注意 `allowed_append_roles` 与编码

器行为保持一致。 - 风险标记：依赖编码器行为，E2E 阈值放宽，仅修复 DeepSeek V4

关联脉络

- PR #2035 tests: mirror sglang_speculative_algorithm into session-server test namespaces: 同属 session-server 测试领域，共享测试夹具与命名空间，共同完善会话校验相关覆盖。
- PR #2075 session: apply the trained LoRA adapter to session-server rollouts: 同为 session 服务器行为修正，说明 session 路径近期持续演进，TITO 校验与其配合保障会话 roundtrip 正确性。