PR #2760 完整报告

radixark/miles

feat(tito): support Qwen3.8 27B and Flash-Next templates

合并时间：2026-09-01 04:12

原文链接：http://prhub.com.cn/radixark/miles/pull/2760

执行摘要

- 一句话：Qwen3.8 系列新增双 TITO 家族名共享固定渲染模板
- 推荐动作：值得精读，重点看三处设计决策：(1) 两个公共枚举映射到同一个 tokenizer 类的工厂写法，如何避免复制 renderer；(2) 固定默认 reasoning 模式并显式拒绝冲突 kwargs 的保护机制，防止缓存前缀静默失配；(3) FIXME 驱动的临时行为显式化，为后续 request-argument 优先级重构留出清晰的演进路径。若团队要新增类似“共享序列化契约、不同公共名”的模型家族，可直接复用本 PR 的模式。

功能与动机

Qwen3.8 将 reasoning effort 序列化进 system prefix，当 Miles 供应 pretokenized input_ids 时需要一个家族专属 renderer。27B 与 Flash-Next 的 tokenizer 共享该序列化契约，但模型架构仍需不同的公共 TITO 家族名。作者明确说明本 PR 只支持注册的默认 reasoning 模式：per-request 定制需要单独的 request-argument 契约重构，否则缓存的 session prefix 可能静默改变渲染模式。

实现拆解

实现按以下 4 步展开：

1. 新增 tokenizer 家族与枚举注册：在 miles/utils/chat_template_utils/tito_tokenizer.py 中定义 Qwen38SmallTITOTokenizer，继承 Qwen3TITOTokenizer，tool_call_parser = "qwen3_coder"，FIXED_TEMPLATE 指向新模板并固定 preserve_thinking=True, reasoning_effort="xhigh"（带 FIXME 注释标明临时性）；TITOTokenizerType 新增 QWEN38_SMALL = "qwen38small" 与 QWEN4_EXP = "qwen4exp"，get_tokenizer_class 中用 case cls.QWEN38_SMALL | cls.QWEN4_EXP 让两个公共名共享同一个实现类，避免复制 Jinja renderer。
2. 新增固定模板miles/utils/chat_template_utils/templates/qwen3.8_small_and_flash_next_fixed.jinja（167 行）：从 HuggingFace 上 Qwen3.8-27B commit 1d4bf0f 与 Qwen3.8-Flash-Next commit de4b8e4 的 tokenizer_config.json 复制而来，仅删除 synthetic-prefix 的 no-user guard；核心包含 reasoning effort 校验与 system 前缀指令注入、vision 占位符渲染、tool_call 格式、multi_step_tool 反向扫描逻辑。
3. 扩展验证器参数透传：miles/utils/test_utils/session_verify_runner.py 的 namespace_to_train_args 新增 --sglang-kv-cache-dtype、

--sglang-mamba-full-memory-ratio 两个 opt-in 参数的序列化，默认 None 时不输出，确保 Qwen3.8 的 FP8 KV cache 与 Mamba 全内存驻留配置能从 verify runner 传到训练命令，同时不影响其他家族默认值。

4. 测试与文档配套：新增 tests/e2e/sglang/test_session_server_multi_role/test_qwen38small.py（2 GPU H200 lane，ModelConfig 使用 Qwen3.8-27B-FP8、fp8_e4m3、mamba_full_memory_ratio=4.59、2 cycles，注册两个 mismatch_rate CI gate）；fast 测试覆盖枚举 dispatch（test_tito_tokenizer.py）、固定模板解析表（test_fixed_templates.py，新增 QWEN38_SMALL / QWEN4_EXP 两行共享同一模板）、kwargs 覆盖保护参数化、pretokenized append-only 契约（test_pretokenized_chat.py，Qwen3.8 因默认保留思维被排除在跨 user turn 负向测试外）、session 预 tokenize 冒烟（test_session_pretokenized_e2e.py）、参数序列化（test_session_verify_runner.py、test_arguments.py）；文档更新 docs/user-guide/agentic-rollout.md。

关键文件：

- miles/utils/chat_template_utils/tito_tokenizer.py（模块 模板渲染；类别 source；类型 core-logic；符号 Qwen38SmallTITOTokenizer）：核心源码：新增 Qwen38SmallTITOTokenizer 类，并在 TITOTokenizerType 枚举与 get_tokenizer_class 工厂中注册 qwen38small / qwen4exp 两个共享实现的家族名。
- miles/utils/chat_template_utils/templates/qwen3.8_small_and_flash_next_fixed.jinja（模块 渲染模板；类别 other；类型 core-logic）：新增 167 行固定 Jinja 模板，是 Qwen3.8 两个家族共享的渲染核心：包含 reasoning effort 校验与 system 前缀指令注入、vision 占位符、tool_call 渲染与 multi_step_tool 逻辑。
- miles/utils/test_utils/session_verify_runner.py（模块 参数透传；类别 source；类型 core-logic；符号 namespace_to_train_args）：扩展 namespace_to_train_args，新增 FP8 KV cache dtype 与 Mamba memory ratio 两个 opt-in 参数的透传，是 E2E 验证 Qwen3.8 配置的关键通道。
- tests/e2e/sglang/test_session_server_multi_role/test_qwen38small.py（模块 E2E 测试；类别 test；类型 test-coverage；符号 test_qwen38small）：新增 E2E 测试：在 2 GPU H200 lane 上用 Qwen3.8-27B-FP8 验证 qwen38small 家族，配置 FP8 KV cache、Mamba ratio 4.59，并注册两个 mismatch_rate CI gate。
- 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_registered_kwargs_cannot_be_overridden）：固定模板注册表的强制覆盖测试：新增 QWEN38_SMALL / QWEN4_EXP 两行共享同一模板与 xhigh kwargs，并把 kwargs 覆盖保护测试参数化到 Qwen38SmallTITOTokenizer。
- tests/fast/utils/test_utils/test_session_verify_runner.py（模块 参数测试；类别 test；类型 test-coverage；符号 test_namespace_to_train_args_omits_model_mamba_cache_config_by_default, test_namespace_to_train_args_emits_model_mamba_cache_config）：验证 namespace_to_train_args 对新增 FP8 KV cache 与 Mamba ratio 参数的默认省略与显式输出行为，并断言 run_one 能透传这两个配置。
- tests/fast/utils/chat_template_utils/test_pretokenized_chat.py（模块 渲染测试；类别 test；类型 test-coverage）：把 Qwen3.8 家族纳入 append-only 契约测试，并因默认保留思维将其排除在跨 user turn 压缩负向测试之外，验证渲染不变量。

- tests/fast/utils/test_arguments.py（模块 参数测试；类别 test；类型 test-coverage；符号 test_qwen38_families_resolve_default_template）：新增 test_qwen38_families_resolve_default_template，验证两个家族名经参数解析后落到同一个模板路径与 xhigh kwargs。
- docs/user-guide/agentic-rollout.md（模块 用户文档；类别 docs；类型 documentation）：用户文档更新：补充 Qwen3.8 两个家族名与限制说明，保持模型选择表与实现一致。

关键符号：Qwen38SmallTITOTokenizer, TITOTokenizerType.get_tokenizer_class, namespace_to_train_args, test_qwen38small, test_qwen38_families_resolve_default_template, test_registered_kwargs_cannot_be_overridden, test_namespace_to_train_args_omits_model_mamba_cache_config_by_default, test_namespace_to_train_args_emits_model_mamba_cache_config

关键源码片段

miles/utils/chat_template_utils/tito_tokenizer.py

核心源码：新增 Qwen38SmallTITOTokenizer 类，并在 TITOTokenizerType 枚举与 get_tokenizer_class 工厂中注册 qwen38small / qwen4exp 两个共享实现的家族名。

```
# Qwen3.8 家族：两个公共名共享同一个序列化实现。
# Qwen3.8 会把 reasoning effort 写进 system prefix，因此 Miles 供应
# pretokenized input_ids 时必须由家族专属 renderer 处理，否则前后缀会失配。
class Qwen38SmallTITOTokenizer(Qwen3TITOTokenizer):
    """Qwen3.8 reasoning-effort 模板 + Qwen3 token 边界。"""

    tool_call_parser = "qwen3_coder"

    FIXED_TEMPLATE = FixedTemplate(
        template="qwen3.8_small_and_flash_next_fixed.jinja",
        # FIXME: 在 request-argument 优先级统一之前，临时固定 reasoning effort 为
        # xhigh，避免缓存的 session prefix 因 per-request 参数而静默改变渲染模式。
        extra_kwargs={"preserve_thinking": True, "reasoning_effort": "xhigh"},
        allowed_append_roles=frozenset({"tool", "user", "assistant"}),
    )


class TITOTokenizerType(StrEnum):
    # ... 省略其他家族枚举 ...
    # qwen38small 对应 Qwen3.8-27B，qwen4exp 对应 Qwen3.8-Flash-Next。
    # 二者 tokenizer 序列化契约相同，但公共名分开，给模型架构独立演进留空间。
    QWEN38_SMALL = "qwen38small"
    QWEN4_EXP = "qwen4exp"

    @classmethod
    def get_tokenizer_class(cls, t: TITOTokenizerType) -> type[TITOTokenizer]:
        """Resolve the concrete ``TITOTokenizer`` subclass for *t*."""
        match t:
            # ... 省略其他 case ...
```

```
        case cls.QWEN38_SMALL | cls.QWEN4_EXP:
            # or-pattern：两个公共名指向同一个实现类，避免复制 Jinja renderer。
            return Qwen38SmallTITOTokenizer
        # ... 省略其他 case ...
        case _:
            raise ValueError(f"Unknown TITOTokenizerType: {t!r}")
```

miles/utils/chat_template_utils/templates/qwen3.8_small_and_flash_next_fixed.jinja

新增 167 行固定 Jinja 模板，是 Qwen3.8 两个家族共享的渲染核心：包含 reasoning effort 校验与 system 前缀指令注入、vision 占位符、tool_call 渲染与 multi_step_tool 逻辑。

```
{# reasoning effort 解析与指令注入：这是 Qwen3.8 与 Qwen3 系列的关键差异，
   effort 被序列化进 system 前缀，本地渲染必须与远端 SGLang 默认一致，
   否则 pretokenized 前缀会与远端生成失配。 #}
{%- set reasoning_instructions = '' %}
{%- if enable_thinking is undefined or enable_thinking is true %}
    {%- set resolved_reasoning_effort = reasoning_effort|default('xhigh') %}
    {%- if resolved_reasoning_effort not in ('xhigh', 'medium', 'low') %}
        {{- raise_exception('Unexpected reasoning effort ' ~ reasoning_effort ~
            '. Supported types are xhigh (default), medium, and low.') }}
    {%- endif %}
    {# xhigh 与 low 注入不同的思考指令；medium 走默认空指令。 #}
    {%- if resolved_reasoning_effort == 'xhigh' %}
        {%- set reasoning_instructions = 'Reasoning effort is set to xhigh. Please think
            carefully through the task, validate key assumptions, consider plausible
            alternatives, and prioritize correctness, consistency, and clarity in the
            final answer.' %}
    {%- elif resolved_reasoning_effort == 'low' %}
        {%- set reasoning_instructions = 'Reasoning effort is set to low. Keep your
            thinking brief and focused, moving directly to the conclusion without
            unnecessary elaboration.' %}
    {%- endif %}
{%- endif %}

{# 有工具时 system 前缀包含指令；否则把指令单独拼进 system 消息。 #}
{%- if tools and tools is iterable and tools is not mapping %}
    {{- '<|im_start|>system\n' }}
    {%- if reasoning_instructions %}
        {{- reasoning_instructions + '

' }}
    {%- endif %}
    {{- "# Tools

You have access to the following functions:

<tools>" }}
    {%- for tool in tools %}
```

```
        {{- "\n" }}
        {{- tool | tojson }}
    {%- endfor %}
    {{- "\n</tools>" }}
{%- else %}
    {%- if messages[0].role == 'system' %}
        {%- set content = render_content(messages[0].content, false, true)|trim %}
        {%- if content %}
            {{- '<|im_start|>system\n' + (reasoning_instructions + '

'
                if reasoning_instructions else '') + content + '<|im_end|>\n' }}
        {%- elif reasoning_instructions %}
            {{- '<|im_start|>system\n' + reasoning_instructions + '<|im_end|>\n' }}
        {%- endif %}
    {%- elif reasoning_instructions %}
        {{- '<|im_start|>system\n' + reasoning_instructions + '<|im_end|>\n' }}
    {%- endif %}
{%- endif %}
```

评论区精华

本 PR 无实质性的 review 评论交锋。claude[bot] 前后触发两次自动 review，最终结论为“Code review found no issues”；人工评审者 Shi-Dong 评论“@claude review always”并直接 APPROVED（LGTM!）。真正的设计讨论体现在 PR body 与 commit 演进中：作者明确解释了 qwen38small / qwen4exp 两个外部名共享一个序列化实现的取舍，以及固定 xhigh 是直到 request-argument 优先级统一前的临时 workaround。

- claude[bot] 自动代码审查 (other): 自动审查未发现问题，PR 通过。
- Shi-Dong 人工批准 (other): 人工评审通过，无未解决疑虑。

风险与影响

- 风险：
  1. reasoning effort 被硬编码为 xhigh：Qwen38SmallTITOTokenizer.FIXED_TEMPLATE 中的 FIXME 已标注此行为临时性；在当前契约下，用户无法在 session 内请求 low / medium 推理档位，且 xhigh 会导致更长的 thinking token 与更高的推理成本。若未来取消固定而 request-argument 优先级未统一，缓存的 session prefix 会静默改变渲染模式，破坏不变量。
  2. Flash-Next 仅 tokenizer 层落地：PR body 明确 Qwen3.8-Flash-Next 的完整执行仍需独立 sglang-miles-qwen38next runtime/image，本 PR 的 qwen4exp 单独使用可能无法端到端跑通。
  3. namespace_to_train_args 直接访问新属性：ns.sglang_kv_cache_dtype / ns.sglang_mamba_full_memory_ratio 若调用方 Namespace 未提供会抛 AttributeError；测试 _build_args 已显式补 None 默认，但其他调用路径需同步跟进。

4. 模板差异：从 HF 复制时删除了 synthetic-prefix no-user guard，若存在空 system 前缀场景，本地渲染与 HF 原生模板可能存在细微行为差异，依赖 fast 测试的 append-only 契约兜底。
5. E2E 时长：test_qwen38small 曾在 4 GPU lane 耗时 1h32m，后续迁到 2 GPU lane 缩短到 10-22 分钟，但仍属较重的 CI 门禁。 - 影响：对用户：新增 --tito-model qwen38small（Qwen3.8-27B）与 --tito-model qwen4exp（Qwen3.8-Flash-Next）两个可用家族名，均默认 preserve_thinking=true、reasoning_effort=xhigh；现有 qwen3 / qwen35 / qwen36 / qwennext 家族行为完全不变，属纯增量。对系统：固定模板渲染链路、枚举工厂 dispatch、验证器参数透传三处扩展，全部 opt-in。对团队：确立了一个"多个公共家族名共享同一实现类"的模板复用模式，并为后续 Qwen3.8 系列其他变体（预留 qwen38 名称空间）与 request-argument 契约重构铺路。 - 风险标记：固定 reasoning effort 为临时行为，Flash-Next 依赖独立 runtime，参数透传新属性依赖调用方提供，模板与 HF 源存在删改差异，E2E 门禁耗时较长

关联脉络

- PR #2759 feat(tito): support Qwen3.5 and Qwen3.6 templates: 同一 tito_tokenizer.py 演进线的直接前作，确立了家族级 FIXED_TEMPLATE + 枚举注册 + 固定模板测试的模式，本 PR 完全沿用该模式扩展到 Qwen3.8 系列。
- PR #2785 docs: add the Qwen3.8-Flash-Next RL recipe page: 同一模型系列（Qwen3.8-Flash-Next）的 RL 配方文档，与本 PR 的 qwen4exp tokenizer 支持构成模型可用的前后端配套。
- PR #2791 test(tito): use FP8 Qwen3.6 checkpoint: 同属 TITO 会话测试线，且同样采用 FP8 checkpoint 验证模式；本 PR 的 E2E 也选用 Qwen3.8-27B-FP8 与 fp8_e4m3 KV cache，延续该模式。
- PR #2818 fix(megatron): keep SFT logits in model precision: 与 2764 同属 training_utils 精度链路的后续修复，2760 虽然不涉及 loss 精度，但同处 megatron 训练与 session 渲染的交叉区域，可作为上下文参考。