PR #44244 完整报告

vllm-project/vllm

[Benchmark] Enable reasoning-model (thinking) benchmarking via `--chat-template-kwargs` for client-rendered datasets

合并时间：2026-06-03 13:49

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44244

执行摘要

- 一句话：支持推理模型基准测试的思考模式
- 推荐动作：该 PR 值得精读，特别是对负责基准测试和推理性能分析的工程师。其设计简洁、聚焦，通过最小的 CLI 改动解决了一个实际的基准测试盲区。建议关注后续是否扩展支持更多数据集。

功能与动机

在 custom 和 speed_bench 数据集中，提示词在客户端通过 tokenizer.apply_chat_template() 渲染并发送到 /v1/completions 端点。由于该调用未接收 chat_template_kwargs，无法在这些数据集上启用推理模型的思考模式。PR 说明中提到："there was no way to benchmark reasoning models in their reasoning ("thinking") mode on these datasets"，并指出 --extra-body 和 --default-chat-template-kwargs 均不适用于客户端预渲染路径。

实现拆解

1. 添加 CLI 参数：在 vllm/benchmarks/serve.py 的 add_cli_args 函数中新增 --chat-template-kwargs 参数，类型为 json.loads，默认值 None。
2. 传递参数到数据集采样：在 vllm/benchmarks/datasets/datasets.py 的 get_samples 函数中，对 custom 和 speed_bench 数据集，通过 getattr(args, "chat_template_kwargs", None) 获取参数并传递给 dataset.sample 方法。
3. 在 CustomDataset.sample 中使用：向 apply_chat_template 调用中解包 chat_template_kwargs（或空字典），以将参数传递到模板渲染过程。
4. 新增单元测试：创建 tests/benchmarks/test_custom_dataset_chat_template_kwargs.py，使用 _RecordingTokenizer 桩记录传递给 apply_chat_template 的 kwargs，验证参数正确传递且默认行为不变。

关键文件：

- vllm/benchmarks/serve.py（模块 基准测试；类别 source；类型 core-logic）：添加 --chat-template-kwargs CLI 参数，是用户交互的入口。
- vllm/benchmarks/datasets/datasets.py（模块 基准测试；类别 source；类型 core-logic）：在数据采样逻辑中传递参数到 CustomDataset.sample 和 SpeedBench.sample，是参数

生效的核心路径。

- tests/benchmarks/test_custom_dataset_chat_template_kwargs.py（模块 基准测试；类别 test；类型 test-coverage；符号 _RecordingTokenizer, init, apply_chat_template, call）：新增单元测试，验证参数正确传递和默认行为不变，确保变更质量。

关键符号：add_cli_args, get_samples, CustomDataset.sample, SpeedBench.sample

关键源码片段

vllm/benchmarks/serve.py

添加 --chat-template-kwargs CLI 参数，是用户交互的入口。

```
# vllm/benchmarks/serve.py
# 在 add_cli_args 函数中新增参数组
parser.add_argument(
    "--chat-template-kwargs",
    type=json.loads, # 将 JSON 字符串解析为 dict
    default=None, # 默认不传递额外 kwargs
    help="A JSON string of kwargs forwarded to the tokenizer's "
         "apply_chat_template when a dataset renders prompts client-side "
         "(e.g. custom / speed_bench). "
         "Example: '{\"thinking\": true}' to enable reasoning models.",
)
```

vllm/benchmarks/datasets/datasets.py

在数据采样逻辑中传递参数到 CustomDataset.sample 和 SpeedBench.sample，是参数生效的核心路径。

```
# vllm/benchmarks/datasets/datasets.py 中的 get_samples 函数
# 为 custom 数据集传递 chat_template_kwargs
if args.dataset_name == "custom":
    dataset = CustomDataset(...)
    input_requests = dataset.sample(
        ...,
        skip_chat_template=args.skip_chat_template,
        chat_template_kwargs=getattr(args, "chat_template_kwargs", None), # 新增
        ...
    )
# 为 speed_bench 数据集传递
elif ...:
    "speed_bench": lambda: SpeedBench(...).sample(
        ...,
        chat_template_kwargs=getattr(args, "chat_template_kwargs", None), # 新增
        ...
    ),
```

tests/benchmarks/test_custom_dataset_chat_template_kwargs.py

新增单元测试，验证参数正确传递和默认行为不变，确保变更质量。

```
# tests/benchmarks/test_custom_dataset_chat_template_kwargs.py
# 最小化 tokenizer 桩，记录传递给 apply_chat_template 的 kwargs
class _RecordingTokenizer:
    """记录 kwargs 而不加载真实模型/模板。"""
    def __init__(self) -> None:
        self.captured_kwargs: dict | None = None
        self.chat_template = "dummy-template"

    def apply_chat_template(
        self,
        conversation,
        add_generation_prompt: bool = True,
        tokenize: bool = False,
        **kwargs,
    ) -> str:
        self.captured_kwargs = kwargs
        return conversation[0]["content"]

# 测试参数被正确转发
def test_chat_template_kwargs_forwarded(tmp_path: Path) -> None:
    tok = _RecordingTokenizer()
    get_samples(_args(tmp_path / "data.jsonl", {"thinking": True}), tok)
    assert tok.captured_kwargs == {"thinking": True}

# 测试默认情况下不传递额外参数
def test_chat_template_kwargs_default_is_noop(tmp_path: Path) -> None:
    tok = _RecordingTokenizer()
    get_samples(_args(tmp_path / "data.jsonl", None), tok)
    assert tok.captured_kwargs == {}
```

评论区精华

该 PR 仅有一条 reviewer benchislett 的评论 "LGTM"，无其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅添加可选 CLI 参数并传递到特定数据集，不影响现有默认行为。测试覆盖了两种场景（传递参数和默认无参数）。需注意：如果其他数据集未来也添加 chat_template_kwargs 支持，需保持接口一致性；当前 custom_image、custom_audio 等数据集未支持，可能造成用户困惑。
- 影响：直接影响：使用 custom 或 speed_bench 数据集进行推理模型基准测试的用户现在可以通过 --chat-template-kwargs '{"thinking": true}' 启用思考模式，获得更真实的接受长度测量（如 speculative decoding 中的 MTP）。对其他用户无影响，因为参数默认 None 且未改变已有逻辑。团队需要维护新增的测试文件。
- 风险标记：低风险，新增 CLI 参数

关联脉络

- PR #42191 [Perf] Apply single-pass min_larger finding and binary search in Triton Top-p path.: 同为性能相关，涉及采样逻辑，本 PR 的 --chat-template-kwargs 可以用于更精确地测量类似 Top-p 采样优化的实际效果。