PR #34894 完整报告

vllm-project/vllm

[DOC] Add INT8 W4A8 docs and Arm's supported quantization schemes

合并时间：2026-06-05 00:27

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/34894

执行摘要

- 一句话：新增 INT8 W4A8 量化文档及 Arm CPU 支持表格
- 推荐动作：推荐阅读此 PR 以了解如何正确组织 vLLM 量化文档及其支持硬件表格。对于维护文档的开发者，其中关于 mkdocs 重定向和内容复用的讨论（如 pymdownx.snippets）具有参考价值。

功能与动机

关联 Issue #25169 报告 Arm CPU 量化文档缺失。本 PR 旨在（1）提供 INT8 W4A8 这一新量化方案的文档，（2）在硬件支持表格中新增 Arm CPU 列并标注相应方案的状态，（3）按审查要求将 llm-compressor 配方文档统一组织到专属部分。

实现拆解

1. 创建 INT8 W4A8 文档：新增 docs/features/quantization/llm_compressor/int8_w4a8.md，详细描述了使用 llm-compressor 进行 W4A8 量化的完整流程，包括模型加载、校准数据准备、两种量化配置（Groupwise 和 Channelwise）的应用，以及在 vLLM 中评估的步骤。
2. 搬迁现有 llm-compressor 文档：将 fp8.md、int4.md、int8.md 从 docs/features/quantization/ 移至 docs/features/quantization/llm_compressor/ 目录，并统一调整了安装命令（强调使用独立环境避免冲突）、将折叠的代码示例展开为直接可用的 Python 块。
3. 更新硬件支持表格：在 docs/features/quantization/README.md 的兼容性表格中新增 "Arm CPU" 列，并添加了 "llm-compressor INT8 (W4A8)" 和 "llm-compressor INT8 (W8A8)" 的行，明确标注 Arm CPU 上的支持状态；同时将量化格式列表重组，增加了 LLM Compressor 子列表。
4. 添加 mkdocs 重定向：在 mkdocs.yaml 的 redirect_maps 中增加了三条规则，将旧路径 features/quantization/fp8.md、features/quantization/int4.md、features/quantization/int8.md 分别映射到新位置，确保现有书签和外部链接不失效。注意未包含 int8_w4a8.md 的重定向，因为该页面为新增，不存在旧路径。
5. 其他文档调整：同步修改了 docs/features/quantization/llm_compressor/README.md 的链接引用，以适应目录结构变化。

关键文件：

- docs/features/quantization/llm_compressor/int8_w4a8.md（模块 量化文档；类别 docs；类型 documentation；符号 preprocess, tokenize, oneshot, GPTQModifier）：新增的 INT8 W4A8 量化核心文档，包含完整的使用流程和代码示例，是本 PR 的主要新增内容。
- docs/features/quantization/llm_compressor/int4.md（模块 量化文档；类别 docs；类型 rename-or-move；符号 preprocess, tokenize）：从根目录搬入 llm-compressor 子目录，并同步更新了安装说明和环境隔离提示。
- docs/features/quantization/llm_compressor/int8_w8a8.md（模块 量化文档；类别 docs；类型 rename-or-move；符号 preprocess, tokenize）：INT8 W8A8 文档搬迁并调整内容，保持与 INT4 文档一致的格式。
- docs/features/quantization/llm_compressor/fp8.md（模块 量化文档；类别 docs；类型 rename-or-move）：FP8 文档搬迁，并添加了评估说明和环境隔离建议。
- mkdocs.yaml（模块 网站配置；类别 config；类型 configuration）：增加了文档重定向配置，确保搬迁后的旧路径仍可访问。
- docs/features/quantization/README.md（模块 量化文档；类别 docs；类型 documentation）：更新了硬件支持表格，新增 Arm CPU 列和对应行；重组了量化格式列表。
- docs/features/quantization/llm_compressor/README.md（模块 量化文档；类别 docs；类型 rename-or-move）：文件搬迁（无内容变更），维持目录入口。

关键符号：preprocess, tokenize, oneshot, GPTQModifier

关键源码片段

docs/features/quantization/llm_compressor/int8_w4a8.md

新增的 INT8 W4A8 量化核心文档，包含完整的使用流程和代码示例，是本 PR 的主要新增内容。

```
# 使用 llmcompressor 进行 W4A8 量化

from llmcompressor import oneshot
from llmcompressor.modifiers.quantization import GPTQModifier

# 配置量化配方：权重 INT4，激活 INT8，忽略 lm_head
recipe = [
    GPTQModifier(
        targets="Linear",
        scheme="W4A8",
        ignore=["lm_head"],
        dampening_frac=0.01,
    ),
]

# 应用量化，使用校准数据集 ds
oneshot(
    model=model,
    dataset=ds,
```

```
    recipe=recipe,
    max_seq_length=2048,
)

# 保存量化后的模型及 tokenizer
save_dir = MODEL_ID.split("/")[1] + "-W4A8-Dynamic-Per-Token-Groupwise"
model.save_pretrained(save_dir)
tokenizer.save_pretrained(save_dir)
```

mkdocs.yaml

增加了文档重定向配置，确保搬迁后的旧路径仍可访问。

```
# docs/features/quantization/ 旧路径重定向到新位置
plugins:
  - redirects:
      redirect_maps:
        features/quantization/fp8.md: features/quantization/llm_compressor/fp8.md
        features/quantization/int4.md: features/quantization/llm_compressor/int4.md
        features/quantization/int8.md: features/quantization/llm_compressor/int8_w8a8.md
        # 原有的重定向保留
        serving/openai_compatible_server.md: serving/online_serving/README.md
```

评论区精华

1. 文档搬迁决策：dsikka 要求将 INT8 W4A8 示例放在 llm-compressor 专属 section。作者不仅移动了该文档，还将 FP8、INT4、INT8 文档一并搬迁至 llm_compressor 目录。
2. 文档内容修正：gemini-code-assist 自动审查发现 README 表格重复行、残留 HTML 标签、W4A8 描述错误和评估命令不一致。作者修复了大部分问题，但评估命令参数保留与其他方案一致的设计。
3. 重定向准确性：hmellor 指出 mkdocs.yaml 中不应包含新增页面 int8_w4a8.md 的重定向，因为该页面从未存在过。作者移除了该条目。
4. 文档格式优化：hmellor 建议使用 mkdocs-material 的 content tabs 展示两种量化方案，并使用 pymdownx.snippets 复用公共内容。作者应用了 content tabs，将 snippets 复用延迟到后续清理 PR。

- llm-compressor 文档应放入专属 section (design): 作者将所有 llm-compressor 相关文档（FP8、INT4、INT8）一并移动到 llm_compressor 目录，并更新了 README 索引。
- 评估命令不一致及描述错误 (correctness): 作者修正了绝大部分问题，但评估命令坚持与其他方案文档保持一致，未完全采纳审查建议。
- mkdocs.yaml 中不应包含不存在的页面重定向 (correctness): 已移除不存在的页面的重定向。
- 使用 mkdocs-material content tabs 和 pymdownx.snippets (documentation): 作者使用了 content tabs 来展示 groupwise/channelwise，但 snippets 的复用推迟到后续清理 PR。

风险与影响

- 风险：

- 文档渲染兼容性：新文档使用了 mkdocs-material 的 content tabs 和步骤编号，不同版本渲染效果可能略有差异，需通过文档构建预览验证。
- 链接失效：搬迁后旧文件被移除，若外部网站或内部其他文档直接引用旧路径将导致 404；本 PR 已通过 mkdocs 重定向缓解该风险。
- 评估命令差异：文档中的评估命令与 PR 描述中使用的命令在参数上不一致（如 num_fewshot 和额外参数），可能使用户复现的准确率结果略有出入。作者主张与其他方案文档一致，但用户若参考 PR 描述则可能混淆。
- 环境隔离提示：新增了使用独立虚拟环境的建议，但用户若跳过该提示可能导致 llm-compressor 与 vLLM 依赖冲突，这是一个操作性风险。

- 影响：
  - 用户：量化文档更完整，Arm CPU 用户可清晰了解哪些方案可用及如何使用；llm-compressor 配方集中放置降低了导航复杂度。
  - 系统：无运行时影响，仅涉及文档和配置。
  - 团队：文档结构更加清晰，便于未来新增其他 llm-compressor 方案时保持组织一致性。
  - 风险标记：文档链接失效（重定向缓解），评估命令参数差异，环境隔离提示可能被忽略，渲染兼容性需验证

关联脉络

- 暂无明显关联 PR