

PR #34626 完整报告

sgl-project/sglang

[Docs] Use Meta's canonical Muse Glimmer GGUF filename

合并时间: 2026-08-13 07:07

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34626>

执行摘要

本 PR 将 Muse Glimmer cookbook 配置中 GGUF 模型文件名更新为 Meta 官方规范命名 ([muse-glimmer-30B-kquant-17gb.gguf](#) → [Muse-Glimmer-30B-KQuant-17GB-Q4_K_M.gguf](#))。纯文档字符串变更，全仓库唯一旧名引用，无功能影响，已获批准合入。

功能与动机

Meta 将 [meta-models/Muse-Glimmer-30B-GGUF](#) 仓库工件重命名为编码量化类型的规范命名方案 ([hub discussion #5](#))，并请求下游仓库同步。PR body 指出 cookbook 的 GGUF 单元格通过 `modelName["defaultlgguf"]` 解析 `MODEL_NAME`，这是仓库中仅存的旧名称引用（已验证全仓库搜索四个旧名称均无其他匹配）。由于 Hub 仓库目前新旧名称并存 ([43c7ead](#))，本变更属于前瞻性对齐而非修复已损坏的链接。

实现拆解

- 修改配置文件：在 `docs/src/snippets/configs/meta-models/muse-glimmer.jsx` 中，将 `modelName` 对象的 `"defaultlgguf"` 值更新为 `Muse-Glimmer-30B-KQuant-17GB-Q4_K_M.gguf`。这是 cookbook 配置的唯一入口，`MODEL_NAME` 占位符被命令和 `curl` 模板引用，修改会自动传播到所有生成的 GGUF 单元格。
- 明确不更新的三个旧名称：
 - `muse-glimmer-30B-kquant-dynamic.gguf` (→ `Q4_K_XL`)：cookbook 不提供该量化选项；
 - `mmproj-kquant.gguf`：SGLang 没有 `mmproj` 路径，且文档已注明 GGUF 路由仅文本；
 - `dflash-kquant.gguf`：GGUF DFlash 草稿模型未被使用，`rtx5090 + gguf + dflash` 单元格实际使用 HF 草稿模型 `meta-models/Muse-Glimmer-30B-assistant`。
- 无测试配套：纯文档字符串变更，PR 已声明不适用单元测试和精度验证。

`docs/src/snippets/configs/meta-models/muse-glimmer.jsx`

cookbook 配置源文件，其中的 `modelName` 映射定义了不同量化格式的模型标识符，本次修改的 `defaultlgguf` 条目是所有 GGUF 部署单元格的命令生成基础。

```
export const config = {
  // ...
  modelName: {
    // GGUF 文件名跟随 Meta 官方规范命名，编码量化类型：
    // 旧名 muse-glimmer-30B-kquant-17gb.gguf → 新名 Muse-Glimmer-30B-KQuant-17GB-Q4_K_M.gguf
```

```
"defaultlbf16": "meta-models/Muse-Glimmer-30B",
"defaultlgguf": "meta-models/Muse-Glimmer-30B-GGUF/Muse-Glimmer-30B-KQuant-17GB-Q4_K_M.gguf",
"defaultlnvfp4": "RadixArk/Muse-Glimmer-NVFP4",
"defaultlmlx-q4": "RadixArk/Muse-Glimmer-q4-MLX",
"defaultlmlx-q4km": "RadixArk/Muse-Glimmer-q4km-gs128-MLX",
"defaultlmlx-q4k-dyn": "RadixArk/Muse-Glimmer-q4k-dynamic-MLX",
},
// ...
}
```

评论区精华

无 review 评论。审阅者 zijiexia 直接批准 (APPROVED) 。

风险与影响

极低风险：仅一行文档配置字符串，无代码逻辑或运行时影响。潜在轻微风险：新文件名含大写字母，若未来旧名称从 Hub 移除而新名称解析异常，会出现链接失效；但 PR body 已确认新名称当前可用，且新旧并存，实际影响可忽略。该变更不影响用户功能，仅影响 cookbook 部署命令中生成的 URL。

关联脉络

本 PR 属于 cookbook 文档体系的持续演进（近期有 Qwen3.8、GLM-5.2、Nemotron 3.5 Lightning 等 cookbook 新增 / 更新，见 PR#34587、#34379、#34573），是模型工件重命名后下游仓库主动对齐的典型维护动作，不涉及代码功能演进。