

PR #34271 完整报告

sgl-project/sglang

Muse Glimmer Cookbook

合并时间: 2026-08-10 18:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34271>

执行摘要

- 一句话: 新增 Muse Glimmer cookbook 文档与部署配置
- 推荐动作: 值得快速浏览而非精读: 若你关心 SGLang cookbook 文档体系如何用 jsx 配置驱动 Deployment 组件、如何组织多硬件x多量化验证矩阵, 该 PR 是很好的模板样例; 若你负责文档导航或链接维护, 建议关注 Llama→Meta 目录迁移是否还有遗漏的外部引用。技术价值主要在于注释中记录的实测参数 (如 RTX 5090 上 fp8 KV cache 与 draft 量化对 max_total_num_tokens 的影响), 对在该硬件上部署 Muse Glimmer 的工程师有直接参考价值。

功能与动机

PR body 仅说明 "Adds a cookbook page for Muse Glimmer". 从文档内容看, 动机是为新发布的 Meta Muse Glimmer 30B 模型提供标准化的 cookbook 入口: 模型支持多种量化格式 (BF16、GGUF、NVFP4、MLX) 与 DFlash 投机解码, 需要一份文档阐明各格式的能力边界 (如 GGUF/NVFP4/MLX 仅文本、BF16 支持图像输入) 以及实测验证过的启动参数组合。

实现拆解

1. 新增模型部署配置数据文件: 新增 docs/src/snippets/configs/meta-models/muse-glimmer.jsx (522 行) 与 muse-glimmer-benchmarks.jsx (23 行)。前者以声明式 config 描述硬件矩阵 (b200/h200/rtx5090/rtx6000/dgx-spark/mac)、量化选项、DFlash 投机解码 flags、docker 镜像与 benchmark 命令模板, 并内置多组带 verified: true 的启动参数 cell; 后者枚举了可复现 benchmark 的硬件x量化x策略组合。两文件共同构成 Deployment/Playground 组件的 data-contract。
2. 新增 cookbook 页面: 新增 docs/cookbook/autoregressive/Meta/MuseGlimmer.mdx (205 行), 导入上述配置并渲染 Deployment/Playground 组件, 正文分 Deployment、Playground、1. Model Introduction、2. Configuration Tips、3. Advanced Usage (Reasoning/Tool Calling/Multimodal/Apple Silicon MLX) 等节, 并链接到 HF 资源。该页同时是 docs/docs.json 导航的新入口。
3. 导航与重定向配套: 在 docs/docs.json 中将原 Llama group 改为 Meta group, 加入 MuseGlimmer 页面并把 Llama4/Llama3.3-70B/Llama3.1 三页路径从 cookbook/autoregressive/Llama/ 迁移到 cookbook/autoregressive/Meta/; 同时更新 docs/scripts/gen_redirects.py 中 /basic_usage/llama4 的重定向目标, 防止旧链接失效。

4. 其他引用同步：更新 docs/cookbook/autoregressive/intro.mdx 的卡片（标题 Llama → Meta, href 指向 MuseGlimmer, logo 换为新增的 docs/cards/logos/meta.png），并修正 docs/docs/supported-models/generative_models.mdx 中 Llama-4 文档链接。
5. 文件迁移：Llama1/Llama3.3-70B/Llama4 三个 mdx 文件执行 rename-or-move 至 Meta 目录，内容不变。

关键文件：

- docs/src/snippets/configs/meta-models/muse-glimmer.jsx（模块 文档组件；类别 source；类型 data-contract）：核心数据契约文件：以声明式 config 定义 Muse Glimmer 的硬件矩阵、量化选项、DFlash 投机解码参数与全部 verified 启动参数组合，是 Deployment/Playground 组件的渲染依据，也是本 PR 的主体新增内容。
- docs/cookbook/autoregressive/Meta/MuseGlimmer.mdx（模块 菜谱文档；类别 other；类型 dependency-wiring）：新增的 cookbook 正文页面，是用户实际阅读的入口文档，串联起配置导入、部署步骤、量化格式能力边界与高级用法说明。
- docs/docs.json（模块 站点配置；类别 config；类型 configuration）：导航配置：把 Llama 分组整体改为 Meta 分组并接入 MuseGlimmer 页面，同时更新旧链接重定向目标，属于站内导航与链接完整性配套。

关键符号：未识别

关键源码片段

docs/cookbook/autoregressive/Meta/MuseGlimmer.mdx

新增的 cookbook 正文页面，是用户实际阅读的入口文档，串联起配置导入、部署步骤、量化格式能力边界与高级用法说明。

```
---
title: Muse Glimmer
---

<!-- 页面通过导入 jsx 数据契约，把命令面板、benchmark 复现、Playground 一次性接入 -->
import { Deployment } from "/src/snippets/_deployment.jsx";
import { config } from "/src/snippets/configs/meta-models/muse-glimmer.jsx";
import { benchmarks } from "/src/snippets/configs/meta-models/muse-glimmer-benchmarks.jsx";

<Deployment config={config} benchmarks={benchmarks} />

<!-- Playground 做验证矩阵之外的自由探索 -->
import { Playground } from "/src/snippets/_playground.jsx";

<Playground config={config} />

## 2. Configuration Tips

<!-- 文档核心价值：把每种量化格式的能力边界与坑位写清楚 -->

**The GGUF format is text only.** SGLang has no `mmproj` path.
```

You cannot use the vision GGUF files. Use the BF16 checkpoint for multimodal input.

****DFlash with a GGUF target model**** needs `--speculative-draft-load-format auto`. Without this flag, the draft model uses the `gguf` load format from the target model. The loader then rejects the draft directory.

****Apple Silicon uses an MLX checkpoint, not the GGUF files.**** The MLX backend has no GGUF path. Serve one of the three `RadixArk/Muse-Glimmer-*-MLX` artifacts with `SGLANG_USE_MLX=1`. Keep `--disable-radix-cache` — the windowed KV storage for the sliding-window layers requires it — and set `SGLANG_MLX_CACHE_LIMIT_GB=8`. Speculative decoding is not available on the MLX backend.

docs/docs.json

导航配置：把 Llama 分组整体改为 Meta 分组并接入 MuseGlimmer 页面，同时更新旧链接重定向目标，属于站内导航与链接完整性配套。

```
// docs/docs.json (节选)：导航分组与旧链接重定向同步修改。
// 1) 旧 /basic_usage/llama4.html 的重定向目标从 Llama/ 目录改为 Meta/ 目录
{
  "source": "/basic_usage/llama4.html",
  "destination": "/cookbook/autoregressive/Meta/Llama4"
}

// 2) 导航分组：原 "Llama" group 整体改名为 "Meta"，并新增 MuseGlimmer 页面
{
  "group": "Meta",
  "pages": [
    "cookbook/autoregressive/Meta/MuseGlimmer",
    "cookbook/autoregressive/Meta/Llama4",
    "cookbook/autoregressive/Meta/Llama3.3-70B",
    "cookbook/autoregressive/Meta/Llama3.1"
  ]
}
```

评论区精华

review 评论为空，唯一 comment 是 mintlify bot 的 docs 预览部署通知（状态 Ready，预览地址 `lmsysorg-muse-glimmer-cookbook.mintlify.site`），zijiexia 直接 APPROVED，没有实质技术讨论。因此讨论高亮仅能反映：该 PR 无人工 review 交锋，通过 Mintlify 预览验证文档渲染。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 导航 / 重定向一致性风险：Llama 系列 cookbook 从 Llama/ 目录整体迁到 Meta/，若仓库内其他文档（如 docs.json、intro.mdx、generative_models.mdx）之外还有引用旧

链接的页面未同步，会造成 404；gen_redirects.py 仅补了 /basic_usage/llama4 一条，旧 /cookbook/autoregressive/Llama/* 的站内链接是否被覆盖未完全可见。

2. 文档数据契约维护风险：muse-glimmer.jsx 中通过大段注释记录了 RTX 5090 上实测的 mem-fraction-static 0.9、kv-cache-dtype fp8_e4m3、draft fp8 量化等结论，这些数值与硬件强相关，后续模型或内核行为变化时文档可能失真，且目前无自动化校验。
3. 无测试配套：本次改动含 522 行配置 JS 与 205 行 mdx，但无对应测试；Deployment/Playground 组件对 config 的 schema 依赖没有回归保障。
4. CI 状态：PR Test (Extra) 显示失败 (x)，可能与文档构建或 lint 相关，需确认是否已解决 (PR 最终已合并)。- 影响：对用户：Meta Muse Glimmer 30B 的部署者获得一份带可交互命令面板的完整指南，覆盖多种硬件（含 RTX 5090/RTX 6000 消费级 Blackwell、DGX Spark、Apple Silicon）与量化格式，显著降低上手成本。对系统：不涉及运行时代码，仅影响文档站点导航与重定向；Llama 系列 cookbook 路径迁移属于结构性整理，需要关注外部书签 / 旧链接。对团队：为后续模型 cookbook 建立了 "配置数据文件 + benchmarks 枚举 + mdx 页面" 的标准化模板（与近期 Inkling-Small、Muse Glimmer 相关 PR 的模式一致）。影响程度：中低，纯文档变更但涉及导航结构调整。- 风险标记：导航重定向未全面核对，配置数据无自动化校验，缺少测试配套，CI Extra 失败未留痕

关联脉络

- PR #34281 Muse Glimmer Cookbook: install from the PR branch: 同一功能线：同为 Muse Glimmer cookbook 的配套调整（安装方式改为从 PR 分支构建），说明该 cookbook 后续仍在迭代。
- PR #34278 Muse Glimmer Cookbook: install from the PR branch: 与 34281 内容重复的姊妹 PR，同样针对 MuseGlimmer.mdx 的安装指引，属于同一文档演进序列。
- PR #34250 Update dspark draft path in Inkling small cookbook: 同为 cookbook 文档维护 PR，参考其 jsx 配置 + mdx 页面模式可理解本仓库 cookbook 的标准化写法。