PR #25115 完整报告

sgl-project/sglang

[Doc]: add interns2preview in cookbook

合并时间：2026-05-13 12:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25115

执行摘要

- 一句话：新增 Intern-S2-Preview 部署文档
- 推荐动作：对于需要部署 Intern-S2-Preview 模型的用户，建议精读此文档。文档中提供的 MTP 多 token 预测配置值得关注，展示了 SGLang 对新模型的高级特性支持。

功能与动机

提供 Intern-S2-Preview 模型在 SGLang 上的部署指南，包括标准模式和 MTP 模式，以及视觉输入调用示例，帮助用户快速上手。

实现拆解

实现步骤如下：

1. 在 docs_new/cookbook/autoregressive/InternLM/ 下新增 Intern-S2-Preview.mdx，内容包括模型介绍、SGLang 安装、标准部署命令（tp=8）、MTP 多 token 预测配置（NEXTN 算法）、以及基于 OpenAI 客户端的视觉输入调用示例。
2. 修改 docs_new/docs.json，在 InternLM 分组下的 pages 数组中增加 cookbook/autoregressive/InternLM/Intern-S2-Preview 路径，使文档导航侧边栏能正确索引该页面。
3. 修改 docs_new/cookbook/autoregressive/intro.mdx，在 InternLM 卡片列表中添加一个指向 Intern-S2-Preview 的新卡片链接，方便用户从介绍页直接访问。 此 PR 不涉及任何源代码或测试变更，纯文档补充。

关键文件：

- docs_new/cookbook/autoregressive/InternLM/Intern-S2-Preview.mdx（模块 部署文档；类别 docs；类型 documentation）：新增的模型部署文档主体，包含部署命令和调用示例。
- docs_new/docs.json（模块 导航配置；类别 config；类型 configuration）：文档导航配置文件，新增 Intern-S2-Preview 页面路径。
- docs_new/cookbook/autoregressive/intro.mdx（模块 文档目录；类别 docs；类型 documentation）：文档介绍页新增 Intern-S2-Preview 卡片链接。

关键符号：未识别

关键源码片段

docs_new/cookbook/autoregressive/InternLM/Intern-S2-Preview.mdx

新增的模型部署文档主体，包含部署命令和调用示例。

```
# 标准部署方案：使用 8 张 GPU 张量并行，加载 Intern-S2-Preview
sglang serve \
    --model-path internLM/Intern-S2-Preview \
    --tp 8 \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --mem-fraction-static 0.8 \
    --host 0.0.0.0 \
    --port 30000

# MTP 多 token 预测方案：启用 NEXTN 投机解码以加速推理
SGLANG_ENABLE_SPEC_V2=1 \
sglang serve \
    --model-path internLM/Intern-S2-Preview \
    --tp 8 \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --mamba-scheduler-strategy extra_buffer \
    --speculative-algo 'NEXTN' \
    --speculative-eagle-topk 1 \
    --speculative-num-steps 3 \
    --speculative-num-draft-tokens 4 \
    --mem-fraction-static 0.8 \
    --host 0.0.0.0 \
    --port 30000

# 配置建议：
# - 使用 --reasoning-parser qwen3 来分离推理流与最终输出
# - 使用 --tool-call-parser qwen3_coder 以支持工具调用
# - MTP 模式需要同时设置 --mamba-scheduler-strategy 和 --speculative-algo 'NEXTN'
# - 若权重加载缓慢，可添加 --model-loader-extra-config='{"enable_multithread_load": "true","num_
threads": 64}'
```

评论区精华

无 review 讨论，由维护者直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：纯文档变更，风险极低。唯一风险是部署命令或配置建议可能存在错误，但经过维护者审查批准，且模型官方配置稳定，风险可控。
- 影响：对用户：提供了清晰的 Intern-S2-Preview 部署指南，降低上手成本。对系统：无运行时影响。对团队：增加了文档维护内容，但属于常规文档补充。

- 风险标记：文档变更

关联脉络

- PR #24730 [Cookbook]: add Laguna-XS.2 (Poolside): 同为 cookbook 文档新增 PR，均涉及部署指南和导航配置更新。