PR #7187 完整报告

PaddlePaddle/FastDeploy

[XPU][Docs] Update Release2.5 Note

合并时间：2026-04-07 18:45

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7187

执行摘要

- 一句话：更新昆仑芯 XPU 文档至 Release 2.5.0 版本，修正版本号和部署命令。
- 推荐动作：此 PR 适合文档维护人员和测试人员精读，以了解版本更新细节和文档优化点；工程师可快速浏览部署命令部分，确保参数正确性，但无需深入技术分析。

功能与动机

根据 PR 标题和 review 评论，变更动机是为配合 FastDeploy Release 2.5.0 版本发布，更新昆仑芯 XPU 相关文档，确保用户指南与软件版本同步，避免因版本过时导致部署错误。

实现拆解

修改了四个文档文件：英文安装指南（docs/get_started/installation/kunlunxin_xpu.md）、英文部署指南（docs/usage/kunlunxin_xpu_deployment.md）及其中文对应版本。关键改动包括：1) 将所有版本号从 2.4.0 更新为 2.5.0，PaddlePaddle-XPU 从 3.3.0 更新为 3.3.1；2) 重构部署表格，新增“快速部署”和“最优部署”命令列，并移除已废弃的 --load-choices 参数；3) 调整表格格式和命令示例，以提升文档清晰度。

关键文件：

- docs/get_started/installation/kunlunxin_xpu.md（模块 文档）：更新安装指南中的版本号，确保用户使用正确的 Docker 镜像和 pip 包，是 XPU 部署的基础文档。
- docs/usage/kunlunxin_xpu_deployment.md（模块 文档）：核心部署文档，更新支持模型表格和命令参数，直接影响用户部署体验和配置准确性。
- docs/zh/get_started/installation/kunlunxin_xpu.md（模块 文档）：中文版本安装指南，同步更新版本号，服务中文用户群体。
- docs/zh/usage/kunlunxin_xpu_deployment.md（模块 文档）：中文版本部署文档，优化表格格式和命令，确保中英文文档一致性。

关键符号：未识别

评论区精华

Copilot 在 review 中指出了多个文档问题：1) 表头括号格式不一致（半角 / 全角混用），建议统一为全角括号；2) 参数错误，如 ERNIE-4.5-300B-A47B（128K）的 --max-model-len 设置为 32768，与上下文长度不匹配，建议修正为 131072；3) 表格中出现重复行，可能造成混淆

；4) 变量命名不一致，如 ${mtp_model_path} 与仓库其他文档不统一；5) 参数冲突，MTP 与 Prefix Caching 不能同时使用，但文档中同时开启。讨论集中在文档准确性和一致性上，作者可能基于评论进行了修正。

- 文档参数错误 (correctness): 建议将 --max-model-len 修正为 131072 以确保准确性，作者可能已采纳。
- 表格格式问题 (style): 建议统一使用中文全角括号“（最优）”，以提升文档美观性。
- 参数冲突风险 (design): 建议明确说明兼容性或调整参数，以避免部署错误；状态可能未完全解决。

风险与影响

- 风险：主要风险是文档准确性风险：如果参数设置错误（如 --max-model-len 值不匹配），用户可能无法正确部署模型，导致性能或功能问题；此外，格式不一致可能影响阅读体验。但由于是纯文档变更，对系统运行无直接影响，风险较低。
- 影响：影响范围限于使用昆仑芯 XPU 硬件的用户，他们依赖这些文档进行安装和模型部署。正确更新的文档能提升用户体验，减少配置错误，促进新版本 adoption；影响程度中等，因文档是用户入口，但无代码逻辑变更。
- 风险标记：文档准确性风险，参数不一致风险

关联脉络

- PR #7101 [Others]Upgrade PaddleFormers to version 1.1.1: 同为版本更新相关的 PR，涉及依赖升级和文档维护，显示团队对软件版本同步的重视。