PR #7054 完整报告

PaddlePaddle/FastDeploy

[APIServer] Update /config-info endpoint to include version, chat template, and other metadata

合并时间：2026-03-31 21:26

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7054

执行摘要

本次 PR 扩展了 FastDeploy 中 APIServer 的 /config-info 端点，新增版本信息、聊天模板、启动参数和设备元数据字段，以增强监控和调试能力。变更已通过测试，但代码覆盖率有待提升。

功能与动机

根据 PR 描述，当前 /config-info 端点缺乏版本、聊天模板、设备信息和某些服务启动参数的可见性。为此，本次更新旨在丰富端点返回的数据，便于用户了解系统状态和配置。

实现拆解

主要改动在 fastdeploy/entrypoints/openai/api_server.py 的 config_info() 函数中：

- 添加 version_info 字段，调用 get_version_info() 获取 FD/Paddle 版本等。
- 引入 chat_template 字段，直接引用现有变量。
- 构建 server_config 字典，从 args 提取 16 个启动参数。
- 通过 try-except 块获取 device_info，使用 paddle.device.cuda 查询 GPU 设备属性。

测试文件 tests/entrypoints/openai/test_metrics_routes.py 新增了测试用例，验证新字段的正确性和与 args 的匹配。

评论区精华

Review 中无具体讨论，仅由 Jiang-Jia-Jun 批准合并。这表明变更被快速接受，但缺乏深度技术交锋。

风险与影响

风险：

- 测试覆盖率不足（84.61538%），有 4 行代码未覆盖，可能隐藏边界情况。
- device_info 的异常处理使用 Exception，可能掩盖具体错误，影响调试。
- 新增字段可能轻微增加端点响应时间，但影响有限。

影响：

- 用户能获取更丰富的配置信息，提升可观测性。
- 系统端点的功能增强，不影响核心推理流程。

- 团队需确保后续 args 变更时同步更新 server_config。

关联脉络

从历史 PR 看，PR 6992 同样修改了 api_server.py，新增了 /v1/abort_requests 端点，表明 APIServer 模块正持续扩展功能以支持更多运维需求。PR 7082 涉及 APIServer 的测试修复，与本 PR 的测试更新相呼应。整体趋势显示 FastDeploy 在强化 API 层面的可管理性和监控能力。