PR #24167 完整报告

sgl-project/sglang

[gateway] Align /v1/loads and /model_info with sglang server; drop dead /rerank

合并时间：2026-05-03 02:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24167

执行摘要

本次 PR 将网关的端点路径与 SGLang 服务器规范对齐：新增 /v1/loads 和 /model_info 作为主路径，保留旧名称为废弃别名；移除无法到达后端的 /rerank 路由及其所有相关代码；更新 HTTP 和 PD 上游代理以调用新路径。变更以较小的改动（+10/-275）实现了 API 契约的统一，并清理了死代码。

功能与动机

SGLang 服务器已规范使用 /v1/loads 和 /model_info 路径，而网关之前使用了非规范的 /get_loads 和 /get_model_info，导致内部不一致。此外网关暴露的 /rerank 端点没有对应的后端服务——SGLang 服务器只有 /v1/rerank，因此 /rerank 一直是死代码，无法工作。本次变更加以对齐并清理，降低维护负担。

实现拆解

1. 路由注册调整（server.rs）：在 build_app() 中添加 /v1/loads 和 /model_info 路由，并将 /get_loads 和 /get_model_info 标记为废弃别名（带 TODO 注释）；删除 /rerank 路由及其处理函数 rerank，移除不再需要的 RerankRequest 导入。
2. 上游代理路径更新（router.rs、pd_router.rs）：将 RouterTrait::get_model_info 和 PDRouter::get_model_info 中代理请求的路径从 "get_model_info" 改为 "model_info"。
3. 测试配套更新（mock_worker.rs、test_pd_routing.rs）：将 mock worker 的路由名更新为 /model_info；将 PD 路由覆盖测试矩阵中的 /get_model_info 改为 /model_info，/get_loads 改为 /v1/loads。
4. 测试用例清理（api_endpoints_test.rs）：移除 5 个覆盖 /rerank 的测试用例（test_rerank_success、test_rerank_with_top_k、test_rerank_without_documents、test_rerank_worker_failure、test_rerank_invalid_request），仅保留 test_v1_rerank_compatibility 以验证 /v1/rerank 的正确性。

sgl-model-gateway/src/server.rs

核心路由注册，新增 /v1/loads 和 /model_info 规范路径，删除 /rerank 死代码

```
// 在 build_app() 中注册网关公开的路由
// 规范路径 /v1/loads 和 /model_info 被添加为主路径
// 遗留的 /get_loads 和 /get_model_info 作为废弃别名保留
let app = Router::new()
    .route("/health", get(health))
```

```
        .route("/health_generate", get(health_generate))
        .route("/v1/models", get(v1_models))
        .route("/model_info", get(get_model_info)) // 规范路径
        // TODO: 一个发布周期后移除 /get_model_info 别名
        .route("/get_model_info", get(get_model_info)) // 废弃别名
        .route("/server_info", get(get_server_info))
        .route("/get_server_info", get(get_server_info)) // 废弃别名
        .route("/flush_cache", post(flush_cache))
        .route("/v1/loads", get(get_loads)) // 规范路径
        // TODO: 一个发布周期后移除 /get_loads 别名
        .route("/get_loads", get(get_loads)) // 废弃别名
        // ... 其他路由
        .route("/v1/rerank", post(v1_rerank)); // 已删除旧的 /rerank 路由
```

sgl-model-gateway/tests/api/api_endpoints_test.rs

删除 5 个 /rerank 测试，保留 /v1/rerank 兼容性测试

```
#[tokio::test]
async fn test_v1_rerank_compatibility() {
    let ctx = AppTestContext::new(vec![MockWorkerConfig {
        port: 18110,
        worker_type: WorkerType::Regular,
        health_status: HealthStatus::Healthy,
        response_delay_ms: 0,
        fail_rate: 0.0,
    }])
    .await;

    let app = ctx.create_app().await;

    // 使用规范路径 /v1/rerank 发送请求
    let payload = json!({
        "query": "machine learning algorithms",
        "documents": [
            "Introduction to machine learning concepts",
            "Deep learning neural networks tutorial",
            "Statistical learning theory basics"
        ]
    });

    let req = Request::builder()
        .method("POST")
        .uri("/v1/rerank")
        .header(CONTENT_TYPE, "application/json")
        .body(Body::from(serde_json::to_string(&payload).unwrap()))
        .unwrap();

    let resp = app.oneshot(req).await.unwrap();
    assert_eq!(resp.status(), StatusCode::OK);
```

```
    let body = axum::body::to_bytes(resp.into_body(), usize::MAX).await.unwrap();
    let body_json: serde_json::Value = serde_json::from_slice(&body).unwrap();

    assert!(body_json.get("results").is_some());
    assert!(body_json.get("model").is_some());
    // V1 API 应使用默认模型名
    assert_eq!(body_json["model"], "unknown");

    let results = body_json["results"].as_array().unwrap();
    assert_eq!(results.len(), 3); // 所有文档应被返回

    assert!(results[0]["score"].as_f64().unwrap() >= results[1]["score"].as_f64().unwrap());
    assert!(results[1]["score"].as_f64().unwrap() >= results[2]["score"].as_f64().unwrap());

    // V1 API 默认返回文档
    for result in results {
        assert!(result.get("document").is_some());
    }

    ctx.shutdown().await;
}
```

评论区精华

Gemini Code Assist bot(medium priority):

> 删除的 5 个测试覆盖了成功、top_k、return_documents 等场景，建议将这些测试迁移到 /v1/rerank 端点以保持覆盖。

PR 作者未回复，且 PR 已合并，目前仅保留一个基本的兼容性测试。这可能是一个有待后续补充的缺口。

风险与影响

- 废弃端点移除：虽然保留了 /get_loads 和 /get_model_info 作为别名，但若用户直接依赖未公开的 /rerank 将收到 404，不过该端点从未正常工作，影响有限。
- 测试覆盖减少：删除了 5 个 rerank 测试，仅保留 1 个兼容性测试，若后续对 /v1/rerank 逻辑有改动，可能遗漏边界情况。
- 代理路径一致性：router.rs 和 pd_router.rs 中 get_model_info 调用的上游路径从 "get_model_info" 改为 "model_info"，需确保后端 sglang server 已支持此路径（PR 基于的版本已支持）。

关联脉络

- 关联 PR #21463 做了类似的 server-info 迁移，此 PR 延续了相同的废弃模式。
- 该 PR 是网关端点对齐系列的第一小步，后续计划补齐更多端点（如 LoRA、权重更新、内存占用等）的差距。