PR #24323 完整报告

sgl-project/sglang

fix(http): apply SGLANG_TIMEOUT_KEEP_ALIVE in common.py

合并时间：2026-05-08 07:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24323

执行摘要

- 一句话：修复 dummy health check server 硬编码 keep-alive 超时
- 推荐动作：值得合并。此 PR 修复了配置不一致问题，且改动极小，无风险。

功能与动机

PR #19847 使 SGLANG_TIMEOUT_KEEP_ALIVE 在 srt/entrypoints/http_server.py 中可配置，但 dummy health-check server 辅助函数仍使用硬编码的 5s keep-alive 超时，忽略了该环境变量。此 PR 旨在统一配置方式。

实现拆解

1. 在 python/sglang/srt/utils/common.py 的 _launch_uvicorn_dummy_server 函数中，将 uvicorn.Config 的 timeout_keep_alive 参数从硬编码值 5 替换为 envs.SGLANG_TIMEOUT_KEEP_ALIVE.get()。
2. 该环境变量已在 sglang/srt/environ.py 中定义，默认值为 5，因此向后兼容。
3. 无其他文件修改，无测试变更。

关键文件：

- python/sglang/srt/utils/common.py（模块 HTTP；类别 source；类型 core-logic）：唯一修改文件，将 dummy health check server 的 keep-alive 超时从硬编码改为环境变量控制。

关键符号：未识别

关键源码片段

python/sglang/srt/utils/common.py

唯一修改文件，将 dummy health check server 的 keep-alive 超时从硬编码改为环境变量控制。

```
# python/sglang/srt/utils/common.py 中 _launch_uvicorn_dummy_server 函数片段

config = uvicorn.Config(
    app,
    host=host,
    port=port,
```

```
        # 之前是 timeout_keep_alive=5，现在使用环境变量，默认仍为 5
        timeout_keep_alive=envs.SGLANG_TIMEOUT_KEEP_ALIVE.get(),
        loop="auto",
        log_config=None,
        log_level="warning",
    )
```

评论区精华

无 review 评论。机器人 gemini-code-assist 仅表示无反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改一行常量，使用与主服务器相同的环境变量访问模式，默认值一致。回归风险几乎为零。
- 影响：影响范围小：仅影响通过 _launch_uvicorn_dummy_server 启动的健康检查服务器。用户可通过设置 SGLANG_TIMEOUT_KEEP_ALIVE 环境变量统一控制所有 uvicorn 服务器的 keep-alive 超时。
- 风险标记：暂无

关联脉络

- PR #19847 [misc] add env for http keep alive timeout: 原始 PR，引入了 SGLANG_TIMEOUT_KEEP_ALIVE 环境变量但未覆盖 dummy server。