PR #23921 完整报告

sgl-project/sglang

[SKILL] Sync SGLang skill docs

合并时间：2026-04-28 17:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23921

执行摘要

- 一句话：统一 torch profiler 分析脚本并更新 CI 技能文档
- 推荐动作：建议合并，代码质量良好，review 建议可后续单独修复 runner 命名问题。值得关注的设计决策是 profiler 分析脚本的统一化架构，特别是跨框架的 canonicalize_framework 和 normalize_repo_relative_path 设计。

功能与动机

此 PR 旨在将 SGLang 专用的 sglang-torch-profiler-analysis skill 升级为统一的 llm-torch-profiler-analysis，以支持 SGLang、vLLM、TensorRT-LLM 等框架，并提供更完善的 fuse/overlap 目录。同时刷新过时的 CI skill 文档，补充 B200 JIT 覆盖、B200-small/GB200 套件状态以及 target_stage 和 include_wheel_build 的 kernel-wheel 行为。

实现拆解

1. 迁移与统一化：创建 llm-torch-profiler-analysis 目录，将原 sglang-torch-profiler-analysis 下的脚本迁移过来并大幅增强，新增 profile_common.py 作为共享模块，提供跨框架的工具函数（如 canonicalize_framework、normalize_repo_relative_path），并新增 analyze_llm_torch_profile.py 作为统一入口，替代旧的 analyze_sglang_torch_profile.py。
2. 新增辅助工具：新增 render_triage_markdown_bundle.py 用于将多个 triage 报告合并为 Markdown，新增 probe_llm_server.py 用于探测服务器，新增 make_trtllm_py_executor_override.py 用于 TensorRT-LLM 的 profiler 注入。
3. 内核与重叠分析增强：在 triage_kernel_helpers.py 和 triage_overlap_helpers.py 中新增低信号路径过滤（LOW_SIGNAL_FUNCTION_TOKENS、LOW_SIGNAL_PATH_TOKENS）、Stage 标注支持（StageAnnotation、StageWindow）以及 Qwen-style shared expert 融合模式。
4. CI 文档刷新：更新 .claude/skills/ci-workflow-guide/SKILL.md 和 .claude/skills/write-sglang-test/SKILL.md，补充 B200 测试套件说明、runner 标签修正（如 stage-b-kernel-unit-1-gpu-b200 和 stage-c-test-4-gpu-b200-small 的 runner 名称）。

5. 清理旧文件：删除原 sglang-torch-profiler-analysis 目录下的多个脚本，确保技能目录整洁。

关键文件：

- .claude/skills/llm-torch-profiler-analysis/scripts/analyze_llm_torch_profile.py（模块 分析入口；类别 source；类型 dependency-wiring；符号 build_triage_parser, parse_triage_args, resolve_profile_targets, build_mapping_kernel_map）：统一分析入口，替代旧 SGLang 专有版本，支持多框架命令行参数
- .claude/skills/llm-torch-profiler-analysis/scripts/profile_common.py（模块 共享工具；类别 source；类型 dependency-wiring；符号 _normalize_text_cached, normalize_text, canonicalize_framework, framework_display_name）：共享工具模块，提供跨框架的字符串规范化、框架别名转换等核心功能
- .claude/skills/llm-torch-profiler-analysis/scripts/triage_kernel_helpers.py（模块 内核分析；类别 source；类型 rename-or-move；符号 end_ts, TimedEventIndex, FrameResolution, StageAnnotation）：内核分析核心模块，新增低信号路径过滤和 Stage 标注支持
- .claude/skills/llm-torch-profiler-analysis/scripts/triage_overlap_helpers.py（模块 重叠分析；类别 source；类型 rename-or-move；符号 group_events_by_stage, choose_window_events, render_ascii_timeline, is_low_signal_scope）：重叠分析模块，新增 Stage 感知和低信号 Scope 过滤
- .claude/skills/ci-workflow-guide/SKILL.md（模块 CI 文档；类别 docs；类型 documentation）：CI 工作流文档，补充了 B200 测试套件和 kernel-wheel 构建说明
- .claude/skills/write-sglang-test/SKILL.md（模块 测试文档；类别 docs；类型 documentation）：测试技能文档，补充了 B200 测试套件 runner 说明

关键符号：build_triage_parser, parse_triage_args, resolve_profile_targets, build_mapping_kernel_map, stage_index, sample_kernels_for_mapping, group_events_by_stage, choose_window_events, render_ascii_timeline, is_low_signal_scope, normalize_match_text, normalize_text, canonicalize_framework, normalize_repo_relative_path, coerce_optional_int

关键源码片段

.claude/skills/llm-torch-profiler-analysis/scripts/analyze_llm_torch_profile.py

统一分析入口，替代旧 SGLang 专有版本，支持多框架命令行参数

```
# analyze_llm_torch_profile.py - 统一 LLM torch-profiler triage 入口
# 支持 SGLang / vLLM / TensorRT-LLM 框架

import argparse
import sys
from collections import defaultdict
from pathlib import Path
from typing import Dict, List, Optional, Sequence, Tuple
```

```
import triage_kernel_helpers as kernel_helpers
import triage_overlap_helpers as overlap_helpers
from profile_common import (
    discover_trace_targets,
    framework_display_name,
    load_server_args,
    load_trace_json,
    parse_stage,
    resolve_framework,
    run_profiler,
)

MIN_RENDER_SHARE_PCT = 1.0
MAPPING_KERNEL_SAMPLE_LIMIT_PER_NAME = 16


def build_triage_parser() -> argparse.ArgumentParser:
    parser = argparse.ArgumentParser(
        prog="analyze_llm_torch_profile.py",
        description=(
            "Compact LLM torch-profiler triage entrypoint for SGLang, vLLM, and "
            "TensorRT-LLM. "
            "This prints three tables: kernel mapping, overlap opportunities, "
            "and fuse opportunities. "
            "Use either a single trace / profile input or a mapping + formal two-trace pair."
        ),
    )
    # --framework 支持自动检测或指定框架
    parser.add_argument(
        "--framework",
        type=str,
        default="auto",
        choices=["auto", "sglang", "vllm", "trtllm", "tllm", "tensorrt-llm"],
        help=(
            "Serving framework. Use auto to detect from trace contents, path hints, "
            "or URL features."
        ),
    )
    # --input / --url / --output-dir 等参数省略以节省篇幅
    return parser
```

.claude/skills/llm-torch-profiler-analysis/scripts/profile_common.py

共享工具模块，提供跨框架的字符串规范化、框架别名转换等核心功能

```
# profile_common.py - 统一 LLM torch-profiler 技能脚本共享工具

from functools import lru_cache

# 框架别名映射：将各种输入转换为标准内部键
```

```
FRAMEWORK_LABELS = {
    "auto": "auto",
    "sglang": "SGLang",
    "vllm": "vLLM",
    "trtllm": "TensorRT-LLM",
}

@lru_cache(maxsize=65536)
def _normalize_text_cached(text: str) -> str:
    """带缓存的文本规范化，将空白字符压缩为单空格。"""
    text = text.strip()
    if not text:
        return ""
    for token in (" ", "\t", "\n", "\r", "\v", "\f"):
        if token in text:
            return " ".join(text.split())
    return text

def normalize_text(value: object) -> str:
    """统一文本规范化入口。"""
    return _normalize_text_cached(value if isinstance(value, str) else str(value))

def canonicalize_framework(value: object) -> str:
    """将框架名称别名转换为标准键（sglang / vllm / trtllm）。"""
    lowered = normalize_text(value).lower().replace("_", "-")
    aliases = {
        "": "auto",
        "auto": "auto",
        "sglang": "sglang",
        "sgl": "sglang",
        "vllm": "vllm",
        "trt": "trtllm",
        "tllm": "trtllm",
        "trtllm": "trtllm",
        "tensorrt-llm": "trtllm",
        "tensorrtllm": "trtllm",
    }
    return aliases.get(lowered, "auto")
```

评论区精华

Review 评论主要指出 CI 文档中的 runner 命名不一致问题：

- stage-b-kernel-unit-1-gpu-b200 的 suite 名为单 GPU，但 runner 却配置了 4-gpu-b200，可能属于标签错误或 runner 不可用导致的特例。
- stage-c-test-4-gpu-b200-small 的 runner 名称缺少 -low-disk 后缀，与 CI 工作流文件中的实际 runner 不匹配，建议修改为 4-gpu-b200-low-disk。 上述问题均未在 PR 中修复，但评论已记录。

- B200 单 GPU suite runner 命名不一致 (question): 未在 PR 中修复，可能需要确认 runner 可用性后添加注释或调整命名。
- B200-small suite runner 缺少 low-disk 后缀 (style): reviewer 给出了修改建议，但 PR 未采纳。
- GB200 suite 状态更新 (documentation): 文档更新正确，无争议。

风险与影响

- 风险：主要风险集中在脚本路径和 API 变更上：sglang-torch-profiler-analysis 下的旧脚本被删除，如果用户（如 AI 辅助工具或其他维护者）硬编码了旧路径，可能会找不到脚本。但 .claude/skills 目录主要用于 AI 辅助，影响面小。CI 文档的 runner 标签与真实 CI 配置可能存在不一致，若未同步修正可能导致开发者参考错误的配置。
- 影响：直接影响 .claude/skills 下的 AI 辅助技能脚本和文档，不涉及 SGLang 运行时逻辑。对使用 torch profiler 分析的开发者有利，统一后的工具支持更多框架并提供了更丰富的分析选项。CI 文档的更新帮助团队了解 Blackwell 测试套件，降低维护成本。
- 风险标记：脚本路径变更，CI 文档与实际可能不一致

关联脉络

- 暂无明显关联 PR