PR #26551 完整报告

sgl-project/sglang

Remove dead fields and always-False plumbing across SB / FB / LogitsMetadata

合并时间：2026-05-28 18:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26551

执行摘要

- 一句话：清理 ScheduleBatch/ForwardBatch/LogitsMetadata 死字段与始终 -False 逻辑
- 推荐动作：建议其他模块的维护者参考本 PR 的方法：当发现字段仅被写入而不被读取，或标志始终为默认值且无生产者时，应积极清理。本 PR 的清理过程规范（先确认使用历史，再分批提交），值得借鉴。

功能与动机

PR body 指出 'has_stream' 自 2024 年 (#1652) 最后一个读取者移除后已变成只写字段，但仍在 init/filter/merge 中持续维护；'temp_scaled_logprobs' 和 'top_p_normalized_logprobs' 自 2025 年 3 月 (#3988) 引入后从未被任何生产者设置为 True，相关条件分支始终执行默认的 log_softmax 路径。清理这些死字段和管道可以减少维护负担，避免误导未来开发者。

实现拆解

1. Commit 1：移除 ScheduleBatch 死字段- 删除 has_stream 字段声明及其在 init_new、filter_batch、merge_batch 中的赋值代码。 - 删除 temp_scaled_logprobs 和 top_p_normalized_logprobs 字段声明（这些字段之前从未在 SB 中被读取或写入，仅作为副本存在）。-涉及文件：python/sglang/srt/managers/schedule_batch.py，删除10行。
2. Commit 2：移除始终为 False 的 temp_scaled_logprobs / top_p_normalized_logprobs 管道- 移除 LogitsMetadata 和 ForwardBatch 中的同名 bool 字段。 - 删除 LogitsProcessor._expand_metadata_for_logprobs 方法（该方法仅根据标志展开 temperature/top_p 张量，从未实际生效）。 - 删除 logprob.py 中的 compute_temp_top_p_normalized_logprobs 函数，该函数包含温度缩放和 top-p 归一化的条件分支，现在其在 LogitsProcessor.process_input_logprobs 和 process_input_logprobs_by_chunk 中的调用点直接替换为 torch.nn.functional.log_softmax。 - 清理 piecewise_cuda_graph_runner.py 和 two_batch_overlap.py 中为 LogitsMetadata 构造传递这些字段的 kwarg 代码。 - 删除 test_eagle_infer_b.py 中从未被服务端处理的 "temp_scaled_logprobs": True 配置项。 -涉及文件：logits_processor.py、logprob.py、forward_batch_info.py、piecewise_cuda_graph_runner.py、two_batch_overlap.py、test_eagle_infer_b.py。

关键文件：

- python/sglang/srt/layers/logits_processor.py（模块 日志处理器；类别 source；类型 core-logic；符号 _expand_metadata_for_logprobs）：核心变更点：移除 _expand_metadata_for_logprobs 方法调用，将 process_input_logprobs 和 process_input_logprobs_by_chunk 中的 compute_temp_top_p_normalized_logprobs 替换为直接 log_softmax，清理 LogitsMetadata 中的 bool 标志字段。
- python/sglang/srt/layers/utils/logprob.py（模块 logprob 工具；类别 source；类型 core-logic；符号 compute_temp_top_p_normalized_logprobs）：移除了 compute_temp_top_p_normalized_logprobs 函数，该函数包含温度缩放和 top-p 归一化的死分支，以及其对 LogitsMetadata 的依赖。
- python/sglang/srt/managers/schedule_batch.py（模块 调度批处理；类别 source；类型 core-logic）：移除了 has_stream、temp_scaled_logprobs、top_p_normalized_logprobs 三个字段及其在 init_new/filter_batch/merge_batch 中的赋值。
- python/sglang/srt/model_executor/forward_batch_info.py（模块 前向批信息；类别 source；类型 data-contract）：删除了 ForwardBatch 中的 temp_scaled_logprobs 和 top_p_normalized_logprobs 字段声明。
- python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py（模块 分段 CUDA 图；类别 source；类型 data-contract）：移除构造 LogitsMetadata 时传递 temp_scaled_logprobs 和 top_p_normalized_logprobs 的 kwarg。
- python/sglang/srt/batch_overlap/two_batch_overlap.py（模块 批重叠；类别 source；类型 core-logic）：移除 filter_batch 中构造 LogitsMetadata 时传递的 temp_scaled_logprobs 和 top_p_normalized_logprobs 假值。
- test/registered/spec/eagle/test_eagle_infer_b.py（模块 测试；类别 test；类型 test-coverage）：删除了测试配置中硬编码的 "temp_scaled_logprobs": True，该配置从未被服务端支持。

关键符号：_expand_metadata_for_logprobs, compute_temp_top_p_normalized_logprobs

关键源码片段

python/sglang/srt/layers/logits_processor.py

核心变更点：移除 _expand_metadata_for_logprobs 方法调用，将 process_input_logprobs 和 process_input_logprobs_by_chunk 中的 compute_temp_top_p_normalized_logprobs 替换为直接 log_softmax，清理 LogitsMetadata 中的 bool 标志字段。

```
# 文件：python/sglang/srt/layers/logits_processor.py
# process_input_logprobs_by_chunk 中的关键变更：
# 原代码根据 temp_scaled_logprobs/top_p_normalized_logprobs 标志
# 从 logits_metadata 提取 per-token temperature/top_p，
# 并调用 compute_temp_top_p_normalized_logprobs。
# 由于这两个标志始终为 False，直接使用 log_softmax。

# 之前（删除的代码）：
# chunk_temperature = (
# logits_metadata.temperature[global_indices]
```

```
# if logits_metadata.temp_scaled_logprobs
# and logits_metadata.temperature is not None
# else None
# )
# chunk_top_p = (
# logits_metadata.top_p[global_indices]
# if logits_metadata.top_p_normalized_logprobs
# and logits_metadata.top_p is not None
# else None
# )
# chunk_input_logprobs = compute_temp_top_p_normalized_logprobs(
# chunk_input_logprobs, logits_metadata, chunk_top_p, chunk_temperature
# )

# 现在直接 log_softmax:
chunk_input_logprobs = torch.nn.functional.log_softmax(
    chunk_input_logprobs, dim=-1
)

# process_input_logprobs 中也做了同样的替换。
```

python/sglang/srt/layers/utils/logprob.py

移除了 compute_temp_top_p_normalized_logprobs 函数，该函数包含温度缩放和 top-p 归一化的死分支，以及其对 LogitsMetadata 的依赖。

```
# 文件：python/sglang/srt/layers/utils/logprob.py
# 整个函数被移除，因为从未被真正启用：
#
# def compute_temp_top_p_normalized_logprobs(
# last_logits, logits_metadata, top_p=None, temperature=None
# ) -> torch.Tensor:
# """
# 原本会根据 temp_scaled_logprobs / top_p_normalized_logprobs
# 决定是否应用温度缩放或 top-p 归一化。
# 由于两个标志始终为 False，始终走 else 分支 log_softmax。
# """
# ... # 全部删除
#
# 现在导入列表中不再包含该函数，调用方直接调用 torch.nn.functional.log_softmax。
```

评论区精华

该 PR 未引发实质性 Review 讨论。gemini-code-assist 的自动审查评论表示“我没有需要提供的反馈”，未提出任何问题。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更确认了所有移除的字段和分支均无生产者写入或无读取者。潜在风险包括：（1）如果某些外部模块通过 Python 反射（如 hasattr、getattr、pickle）间接依赖这些字段，但根据代码库结构可能性极低；（2）has_stream 字段虽然当前无读，但未来可能被重新用于流式功能，但本 PR 已清理所有写入，未来需要时可基于 Req.stream 重新实现。整体而言，这是安全的死代码清理。
- 影响：对用户：无任何功能变化，API 行为和输出不变。 对系统：移除了无意义的条件判断和字段赋值，带来极微小的性能提升（减少条件分支和内存占用）。 对团队：代码库更清晰，减少了误导性标志，降低了未来维护的认知负担。
- 风险标记：无行为变更，低风险，纯死代码删除

关联脉络

- PR #1652 Remove last reader of has_stream (unknown): 该 PR 移除了 has_stream 字段的最后一个读取者，使字段变成只写无读，本 PR 跟进移除该字段。
- PR #25433 Drop Req half of temp_scaled_logprobs / top_p_normalized_logprobs (unknown): 该 PR 移除了 Req 中的对应标志，本 PR 作为 companion 移除 ScheduleBatch 中的对应字段。
- PR #3988 Introduce temp_scaled_logprobs / top_p_normalized_logprobs plumbing (unknown): 该 PR 引入了这些字段及其管道，但从未设置标志为 True，导致代码死到本 PR 移除。