PR #20799 完整报告

sgl-project/sglang

Fix spec v2 + logprob when max_num_token is set

合并时间：2026-04-02 16:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20799

执行摘要

此 PR 修复了 speculative decoding v2 场景下，当设置 max_num_token 时，logprob 数组长度可能大于输出 token 数量的数据对齐问题，通过调整调度输出处理器的切片逻辑确保正确性，属于重要 bugfix，影响 speculative decoding 用户。

功能与动机

动机源于 PR body 中描述的问题：在 max_num_token 设置后，len(logprobs) 可能超过 len(output_ids)，导致数据不一致。这在使用 speculative decoding v2 时可能引发 logprob 发送错误，需要限制对齐以匹配实际输出。

实现拆解

仅修改 python/sglang/srt/managers/scheduler_output_processor_mixin.py 文件中的 stream_output_generation 函数，关键变更点：

- 条件增强：添加 and req.input_token_logprobs_val is not None 条件，确保输入 logprobs 只在 prefill 后计算完毕时发送。
- 切片限制：定义 logprob_end = max(len(output_ids_), 1)，并用于所有输出 logprob 数组的切片（如 output_token_logprobs_val），避免直接使用数组全长。

代码示例： logprob_end = max(len(output_ids_), 1) output_token_logprobs_val.append(req.output_token_logprobs_val[send_output_token_logprobs_offset:logprob_end])

评论区精华

无 review 评论，讨论为空；仅有的评论为作者触发 CI 测试的命令。

风险与影响

- 技术风险：切片逻辑变更可能引入回归，如 logprob_end 计算错误导致 logprob 数据截断或越界；缺少测试覆盖增加风险。
- 影响范围：直接影响使用 speculative decoding v2 并设置 max_num_token 的用户，确保 logprob 数据正确性；对系统调度路径性能影响微小，但需监控。

关联脉络

从历史 PR 看，此 PR 与 speculative decoding 功能演进相关：

- PR #21225 移除了 Ngram 窗口参数，简化配置，可能影响输出处理逻辑。
- PR #21920 迁移 ngram corpus 到 TVM FFI，涉及底层计算，可能间接关联 logprob 生成。这表明项目正持续优化 speculative decoding 的稳定性和性能，本 PR 是其中数据对齐的重要修复。