

PR #28613 完整报告

sgl-project/sglang

docs: add DeepSeek-V4 compressed state dtype tip

合并时间: 2026-06-18 14:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28613>

执行摘要

本 PR 在 DeepSeek-V4 cookbook 文档中新增了关于 `SGLANG_DSV4_COMPRESS_STATE_DTYPE` 环境变量的配置说明，指导用户通过设置为 `bf16` 来降低压缩注意力状态池的 GPU 内存占用。纯文档变更，无代码或测试修改。

功能与动机

DeepSeek-V4 引入了混合压缩注意力机制 (hybrid compressed attention)，新增了 `SGLANG_DSV4_COMPRESS_STATE_DTYPE` 环境变量用于控制 C4/C128 压缩注意力状态池的 dtype。PR body 明确说明目的是“Document the new DeepSeek-V4 compressed attention state dtype setting in the cookbook.”，解决用户缺乏相关配置文档的问题。

实现拆解

- 新增文档小节：在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 的“EPLB + DeepEP Waterfill”小节之前，新增了一个独立的“Compressed attention state dtype”小节。
- 参数说明：详细介绍了 `SGLANG_DSV4_COMPRESS_STATE_DTYPE` 支持的值：`float32/fp32`（默认）和 `bfloat16/bf16`。
- 使用示例：提供了一个使用 `bf16` 启动服务的命令行示例，并强调该设置仅影响压缩注意力状态池，不改变模型权重精度或主 KV 缓存 dtype。
- 效果说明：补充了使用 `bf16` 后的效果：每个压缩状态槽的 GPU 内存占用降低，自动池大小分配下可容纳更多槽位，启动日志中将显示更大的 `c4_state` 和 `c128_state` 池大小。并建议默认使用 `float32` 以保持最保守的行为。

无（纯文档变更，无代码片段可展示）。

评论区精华

无讨论，审核人 Fridge003 已批准。

风险与影响

- 风险：无。纯文档变更，不影响任何代码运行。
- 影响：对用户而言，获得了明确的内存优化指导，有助于在长上下文场景中更好地使用 DeepSeek-V4。影响范围仅限于阅读该文档的用户。

关联脉络

本 PR 与近期 DeepSeek 系列文档更新相关（如 PR#28201 更新 fp8 KV cache 文档），属于持续完善大模型部署文档的一部分。