

PR #32076 完整报告

sgl-project/sglang

Fix Inkling kernel imports after migration

合并时间: 2026-07-22 22:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32076>

执行摘要

- 一句话: 修复 Inkling JIT 内核迁移后的导入路径
- 推荐动作: 推荐合并。该 PR 是此前 JIT 内核迁移工作 (#32045) 的补丁, 修复了一个阻止 Inkling 模型启动的回归缺陷。变更简洁、验证充分, 值得快速跟进。

功能与动机

PR #32045 将 Inkling 模型的 JIT 内核从 `sglang.jit_kernel` 迁移到 `sglang.kernels.ops.model.inkling`, 但遗漏了 `python/sglang/srt/models/inkling_common/kernels/comm.py` 和 `python/sglang/srt/models/inkling_common/attn.py` 中的三处导入路径更新, 导致 Inkling 模型无法启动。PR body 明确指出: "Inkling cannot launch after #32045 missed three imports while moving its JIT kernels to `sglang.kernels.ops.model.inkling`".

实现拆解

该 PR 通过两个文件的修改, 将三个惰性导入路径从旧 JIT 内核位置迁移到新位置, 并同步更新了文档字符串中的引用路径。

1. `python/sglang/srt/models/inkling_common/kernels/comm.py` (核心修复):
 - 在 `_ar_jit()` 函数中, 将 `from sglang.jit_kernel import inkling_all_reduce` 改为 `from sglang.kernels.ops.model.inkling import inkling_all_reduce`。
 - 在 `_ar_fused_jit()` 函数中, 将 `from sglang.jit_kernel import inkling_ar_fused` 改为 `from sglang.kernels.ops.model.inkling import inkling_ar_fused`。
 - 在 `_ar_ssconv_jit()` 函数中, 将 `from sglang.jit_kernel import inkling_ar_scattered_sconv` 改为 `from sglang.kernels.ops.model.inkling import inkling_ar_scattered_sconv`。
 - 同步更新了 `ar_sconv_norm_fusable()` 和 `scattered_ar_sconv_fusable()` 的文档字符串中内核路径引用。
2. `python/sglang/srt/models/inkling_common/attn.py` (文档更新):
 - 在 `_fused_attn_prologue_verify()` 方法的文档字符串中, 将 `(jit_kernel/inkling_attn_prologue.py)` 更新为 `(kernels/ops/model/inkling/inkling_attn_prologue.py)`。此时的导入语句本身已在 #32045 中正确更新, 这里仅修正注释。

所有变更均为路径字符串替换, 无逻辑改动。

关键文件：

- python/sglang/srt/models/inkling_common/kernels/comm.py（模块 Inkling 内核；类别 source；类型 core-logic；符号 _ar_jit, _ar_fused_jit, _ar_ssconv_jit, ar_sconv_norm_fusable）：修复核心，包含三处惰性导入路径更新和相关文档字符串修正。
- python/sglang/srt/models/inkling_common/attn.py（模块 Inkling 注意力；类别 source；类型 documentation；符号 _fused_attn_prologue_verify）：更新文档字符串中内核路径引用，与 comm.py 中的导入保持文档一致。

关键符号：_ar_jit, _ar_fused_jit, _ar_ssconv_jit, _fused_attn_prologue_verify

关键源码片段

python/sglang/srt/models/inkling_common/kernels/comm.py

修复核心，包含三处惰性导入路径更新和相关文档字符串修正。

```
# python/sglang/srt/models/inkling_common/kernels/comm.py
```

```
@functools.cache
```

```
def _ar_jit():
```

```
    # 旧路径：from sglang.jit_kernel import inkling_all_reduce
```

```
    # 新路径：迁移到 kernels.ops.model.inkling 下
```

```
    from sglang.kernels.ops.model.inkling import inkling_all_reduce
```

```
    return inkling_all_reduce
```

```
@functools.cache
```

```
def _ar_fused_jit():
```

```
    # 旧路径：from sglang.jit_kernel import inkling_ar_fused
```

```
    from sglang.kernels.ops.model.inkling import inkling_ar_fused
```

```
    return inkling_ar_fused
```

```
def _ar_ssconv_jit():
```

```
    # 旧路径：from sglang.jit_kernel import inkling_ar_scattered_sconv
```

```
    from sglang.kernels.ops.model.inkling import inkling_ar_scattered_sconv
```

```
    return inkling_ar_scattered_sconv
```

python/sglang/srt/models/inkling_common/attn.py

更新文档字符串中内核路径引用，与 comm.py 中的导入保持文档一致。

```
# python/sglang/srt/models/inkling_common/attn.py
```

```
def _fused_attn_prologue_verify(self, q, k, v, forward_batch, log_scaling_tau=None):
```

```
    # 旧文档：(jit_kernel/inkling_attn_prologue.py)
```

```
    # 新文档：(kernels/ops/model/inkling/inkling_attn_prologue.py)
```

```
    """Fused target-verify {k/v sconv + save_windows + qk-norm (+ KV store)}
```

```
    (kernels/ops/model/inkling/inkling_attn_prologue.py); returns
```

```
    ``(q, k, v, did_store)``.
```

```
    ...
```

```
    """
```

```
from sglang.kernels.ops.model.inkling.inkling_attn_prologue import (  
    inkling_attn_prologue_verify,  
)  
# ...
```

评论区精华

无相关 review 讨论。维护者 BBuf 快速批准 ("LGTM, thanks.")。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅为导入路径字符串修改和文档字符串更新，不涉及任何逻辑或语义变更。核心风险在于是否遗漏了其他文件中的类似导入（但 PR body 确认已覆盖所有遗漏处）。
- 影响：直接影响 Inkling 模型的可启动性：修复后，Inkling 模型能正常加载其自定义全规约、注意力前导等 JIT 内核。对其他模型无影响。无用户可见的行为变化。
- 风险标记：暂无

关联脉络

- PR #32045 [Kernel] Phase 4 batch-3: migrate tangled JIT subsystems + new groups into kernels.ops (RFC #29630): 该 PR 是 #32045 的补丁，修复了迁移过程中遗漏的三处导入路径。