fix: wrap remaining blocking calls with asyncio.to_thread to prevent DoS #3835

Open **kaiisfree** wants to merge 1 commit into `lm-sys:main` from `kaiisfree:fix/async-blocking-work…`

kaiisfree commented 2 weeks ago

Summary

- Fixes [Security] Denial of Service via Blocking Event Loop in Model Workers (Incomplete Fix for ff66426) #3833
- Wraps 3 remaining synchronous blocking calls with `asyncio.to_thread()` to prevent event loop blocking
- Matches the pattern already applied in `base_model_worker.py:api_generate()` (commit `ff66426`)

Changes

1. `base_model_worker.py:api_get_embeddings()` — wrap `worker.get_embeddings(params)`
2. `multi_model_worker.py:api_generate()` — wrap `worker.generate_gate(params)`
3. `huggingface_api_worker.py:api_generate()` — wrap `worker.generate_gate(params)`

Test plan

- [ ] Verify each wrapped call still functions correctly
- [ ] Confirm event loop is no longer blocked during inference
- [ ] Run existing tests

🤖 Generated with Claude Code


`fix: wrap remaining blocking calls with asyncio.to_thread to prevent DoS` ··· `c9e84b8`

Reviewers

No reviews

Assignees

No one assigned

Labels

None yet

Projects

None yet

Milestone

No milestone

Development

Successfully merging this pull request may close these issues.

[Security] Denial of Service via Blocking Event Loop in Model Workers (Incomplete Fix for ff66426)

1 participant