PR #41059 完整报告

vllm-project/vllm

[CI] Add temperature to bfcl eval, default greedy

合并时间：2026-04-30 05:01

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41059

执行摘要

- 一句话：BFCL 评估脚本默认使用贪婪解码
- 推荐动作：该 PR 简单直接，无需深入审查。但可注意文件参数索引的调整（原第 6 个参数 output_dir 变为第 7 个），确保兼容性。

功能与动机

之前 BFCL 评估使用默认温度 0.001，导致 vLLM 日志警告 temperature 0.001 is less than 0.01, which may cause numerical errors nan or inf in tensors. We have maxed it out to 0.01。同时，非零温度带来随机性，影响结果可复现性。

实现拆解

1. 在脚本头部添加注释和变量声明：BFCL_TEMPERATURE - Temperature (default: 0.0) 并读取环境变量 TEMPERATURE="${BFCL_TEMPERATURE:-0.0}"。
2. 在调用评估 Python 脚本时，将温度作为第 6 个命令行参数传入：python3 - "$MODEL" "$TEST_CATEGORY" "$NUM_THREADS" "$PORT" "$API_TYPE" "$TEMPERATURE" "$OUTPUT_DIR" 并调整后续参数索引。
3. 在 Python 内联脚本中将该参数转为浮点数，并传递给 generate() 函数的 gen_kwargs["temperature"]。

关键文件：

- .buildkite/scripts/tool_call/run-bfcl-eval.sh（模块 CI 脚本；类别 other；类型 core-logic）：核心变更文件：新增环境变量 BFCL_TEMPERATURE，默认贪婪解码，传递到生成函数。

关键符号：未识别

关键源码片段

.buildkite/scripts/tool_call/run-bfcl-eval.sh

核心变更文件：新增环境变量 BFCL_TEMPERATURE，默认贪婪解码，传递到生成函数。

```
# 新增环境变量定义
# BFCL_TEMPERATURE - Temperature (default: 0.0)

# 读取环境变量，默认 0.0
```

```
TEMPERATURE="${BFCL_TEMPERATURE:-0.0}"

# 将温度作为第 6 个参数传递给 Python 脚本
python3 - "$MODEL" "$TEST_CATEGORY" "$NUM_THREADS" "$PORT" "$API_TYPE"
"$TEMPERATURE" "$OUTPUT_DIR" << 'PYEOF'
# ... 内嵌 Python 脚本中:
temperature = float(sys.argv[6])
output_dir = sys.argv[7] if len(sys.argv) > 7 and sys.argv[7] else os.getcwd()
# ... 后续生成时使用
gen_kwargs["temperature"] = temperature
generate(**gen_kwargs)
PYEOF
```

评论区精华

gemini-code-assist[bot] 指出评估步骤（evaluate）可能也使用默认温度导致结果目录不匹配，但作者 yzong-rh 回复 evaluate 并不使用 temperature 参数，该顾虑已解除。

- 评估步骤是否也需要设置温度 (correctness): 作者回复 evaluate 不使用 temperature 参数，无需修改。

风险与影响

- 风险：评估脚本仅修改 CI 脚本，不影响 vLLM 核心运行时。风险极低：默认温度改为 0.0 可能使部分原本依赖随机性的模型在 BFCL 评估上表现略有不同，但这是作者希望达到的修复效果。
- 影响：影响范围限定在 BFCL CI 评估流程：
  - 消除此前 0.001 温度引发的数值警告。
  - 默认贪婪解码减少 run-to-run 方差，使评估结果更稳定。
  - 通过环境变量 BFCL_TEMPERATURE 保留灵活性。
  - 风险标记：暂无

关联脉络

- 暂无明显关联 PR