FinanceBench Evaluator Guide

MinionS

1 Overview

The FinanceBench evaluator runs the MinionS protocol on financial document QA tasks. It requires a local LLM (via SGLang) and a remote LLM (OpenAI/Anthropic).

2 Installation

2.1 Core Dependencies

```
cd /workspace/minions

pip install -e . --break-system-packages
pip install -r requirements.txt --break-system-packages

# Additional packages
pip install tiktoken pydantic ollama openai anthropic together mistralai
pip install mcp sentence-transformers faiss-cpu pdfplumber graphviz
pip install kconfiglib transformers accelerate torch einops pandas
pip install pymupdf rank-bm25

# System packages
apt-get update && apt-get install -y zstd graphviz
```

2.2 Environment Variables

```
export OPENAI_API_KEY="sk-your-key-here"
export HF_TOKEN="hf_your-token-here"
```

3 SGLang Server Setup (Terminal 1)

The local LLM is served via SGLang. Run this in a **separate terminal**:

```
cd /workspace/minions
export HF_TOKEN="hf_your-token-here"

# Install SGLang
pip install -e . --break-system-packages
pip install -r requirements.txt --break-system-packages
pip install "sglang[all]" --break-system-packages

# System dependencies
apt-get update && apt-get install -y libnuma-dev

# Fix PyTorch (if needed)
rm -rf /usr/local/lib/python3.11/dist-packages/torch*
pip install torch==2.9.1 torchvision==0.24.1 torchaudio==2.9.1 \
    --index-url https://download.pytorch.org/whl/cu128
pip install torch_memory_saver==0.0.9 torchao==0.9.0 torchcodec==0.8.0
pip install 'anyio>=3.0,<4.0'
```

```
# Launch server
python -m sglang.launch_server \
    --model-path meta-llama/Llama-3.1-8B-Instruct \
    --port 8000 \
    --host 0.0.0.0 \
    --mem-fraction-static 0.7 \
    --max-running-requests 16 \
    --enable-custom-logit-processor
```

Wait for `Server is ready` before proceeding.

4 Running the Evaluator (Terminal 2)

4.1 Configuration

```
cd /workspace/minions

# Interactive configuration
make menuconfig
```

Key settings:

- **Local Model**: `meta-llama/Llama-3.1-8B-Instruct`
- **Backend**: SGLang
- **SGLang URL**: `http://localhost:8000/v1`
- **Remote Model**: `gpt-4o`
- **Dataset Path**: `/workspace/financebench`

4.2 Run Evaluation

```
# Run with config
python3 evaluate/financebench_evaluator.py config/.config

# Or use make
make run
```

4.3 Options

```
# Run specific samples
python3 evaluate/financebench_evaluator.py config/.config \
    --sample-indices 1,2,3,4,5

# Custom output directory
python3 evaluate/financebench_evaluator.py config/.config \
    --output-dir /tmp/my_results

# Custom prompt overrides
python3 evaluate/financebench_evaluator.py config/.config \
    --prompt-set prompts.json
```

5 Results

Output is saved to `evaluate/results/<run_id>/`:

- `summary.txt` — accuracy and cost summary
- `financebench_results.json` — per-sample results
- `sample_logs/` — detailed logs for each sample
- `report.tex` — LaTeX report (if pdflatex available)

6 Correctness Evaluation

After a run, evaluate correctness with:

```
python3 evaluate/correctness.py evaluate/results/<run_id> \
    --verbose --update-summary
```

7 Troubleshooting

Issue	Solution
Connection refused	Start SGLang server first
API key error	`export OPENAI_API_KEY="..."`
Model not found	Check `CONFIG_LOCAL_MODEL_NAME`
No GPU activity	Disable cache: `CONFIG_USE_CACHE=n`