Maintenance

Cached Configs and Apptainer Images

In the `/model-weights` folder, the following files exists:

- `environment.yaml`: Cached environment config
- `models.yaml`: OLD cached model config. This is from pre-v0.8.0, it is no longer being updated, cached model config path was determined directly from the package instead of read from cached environment config. This is kept for backward compatibility.
- `models_v0.8.0.yaml`: NEW cached model config. This is the real "models.yaml" being used by latest versions after v0.8.0. Consider removing the old cached model config and rename this file as `models.yaml` in the near future.
- `vector-inference_x.y.z.sif`: OLD Vector Inference image, this is from pre-v0.8.0, where only vLLM is baked in and the tag is based on vLLM version, discard in the future.
- `vector-inference_latest.sif`: OLD Vector Inference image symlink, this is from pre-v0.8.0 pointing to `vector-inference_x.y.z.sif`, discard in the future.
- `vector-inference-vllm_x.y.z.sif`: Vector Inference image with vLLM baked in, tagged by vLLM version number. Remove old ones as see fit.
- `vector-inference-sglang_x.y.z.sif`: Vector Inference image with SGLang baked in, tagged by SGLang version number. Remove old ones as see fit.
- `vector-inference-vllm_latest.sif`: Vector Inference vLLM image symlink, always point to the latest working version.
- `vector-inference-sglang_latest.sif`: Vector Inference SGLang image symlink, always point to the latest working version.

Model weights tracking

Use `MODEL_TRACKING.md` in the `vec-inf-maintenance` repo to keep track of cached model weights.

- `MODEL_TRACKING.md`: Cached model weights list (i.e. downloaded weights in `/model-weights`), and whether or not it's currently supported by `vec-inf` (i.e. cached model config)

Revision #1
Created 2026-04-10 20:12:37 UTC
Updated 2026-04-10 20:23:06 UTC