**Things to add to Praxis**

**4.2 Experimentation Workflow**

Candidates for removal for the reflection workflow cycle:

* Mistral 7B (lowest results)
* Mistral codestral\_mamba - discontinued by Mistral AI because API call to list all the available models doesn't contain this one, only the 22B codstral model and the codestral embeddings model.
* Solar-10.7B (10.7B model, but low results).
* Phixtral-4x2\_8 (duplicates the results of Phixtral-2x2\_8).
* Mistral Nemo (12B model - largest among all, but it performs at the level of 7~8B models). SINCE CODESTRAL\_MAMBA IS DISCONTINUED, I MIGHT AS WELL KEEP NEMO

**This one is probably not really needed**

**4.3.2 Dataset-Specific Prompt Effectiveness**

The dataset-specific wrappers designed for different benchmark formats showed varying effectiveness:

**Function Completion Tasks (HumanEval & BigCodeBench):** [Analysis of how models responded to function header and docstring-based prompts, including discussion of common failure modes and successful patterns.]

**Natural Language Description Tasks (MBPP & LBPP):** [Evaluation of how models handled natural language task descriptions with explicit test cases, including assessment of instruction-following capabilities.]

**4.5.3 Model-Specific Post-Processing Requirements**

See comments for more details.