Step 1: Dataset collection

Input:
- Toxic datasets: ParaDetox, APPDIA, Parallel Detoxification
- Non-toxic dataset: Jigsaw Unintended Bias

Process: Combine toxic and selected non-toxic samples
Outcome: Balanced toxic / non-toxic dataset

Step 2: Shapley Value Extraction

Input:
- Models: HateBERT, Toxigen, Detoxtify, ToxDetect
- Balanced dataset

Process: Extract and aggregate Shapely values
Outcome: Enriched dataset with Shapley values

Step 3: Generation of Guided Reasoning

Input:
- Models: Marco-o1, QwQ, OpenO1, Skywork, OpenAI o1
- Enriched dataset

Process: Generate reasoning using instruction + Shapley + target paraphrase
Outcome: Guided reasonings

Step 4: Post-processing and Filtering

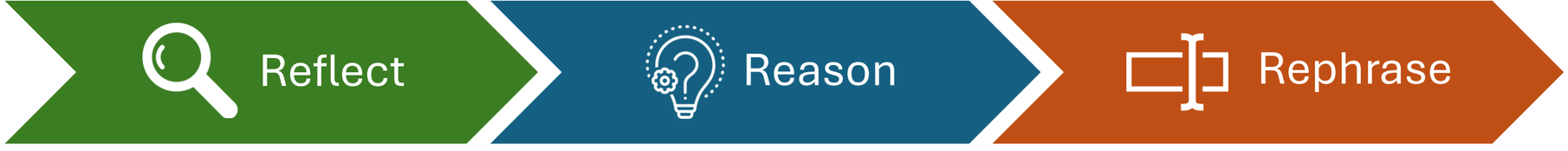
Input:
Guided reasonings
Process: Remove code-switching (Google Translate), remove leakage and restating (Qwen 2.5)
Outcome: Cleaned reasonings

Step 5: Best Model Selection

Input:
Cleaned reasonings
Process: Human A/B Evaluation, JudgeLLM A/B evaluation
Outcome: Best performing models

Step 6: Best Reasoning Selection

Input:
Best-performing models' outputs
Process: Evaluate using JudgeLLM and A/B tournament
Outcome: Final few-shot dataset



Step 7: Comparative Evaluation (ICL)

Input:
Balanced dataset, best-performing models, Qwen 2.5 7B and LlaMA 3.1 8B, baselines (BART, PseudoParaDetox, CAPP, DetoxLLM)
Process: Instruction, n demonstrations, query generation, evaluation metrics (STA, TOX, J, FL, BERT-F1, BLEU, SIM)
Outcome: Comparative results of detoxtified outputs