# Overview

This folder contains everything needed to reproduce our cross‑lingual ethical‑reasoning experiments on large language models. It includes: crosslingual_gpt: notebook and scripts for querying GPT-4o on XEthicsBench crosslingual_claude: notebook and scripts for querying Claude 3 on XEthicsBench evaluate_results: notebook for computing metrics (accuracy, flip-rate, disagreement, category X-rates) from a results JSON data:

- xethicsbench_dataset: the full 200-item multilingual benchmark
- gpt4o_benchmark_results.json: sample GPT-4o outputs
- claude_benchmark_results.json: sample Claude 3 outputs
- Unzip the project folder. (Optional) Create and activate a virtual environment: python3 -m venv venv source venv/bin/activate Install dependencies: pip install openai anthropic pandas tqdm json5
- Running the Claude 3 Pipeline
  - Open crosslingual_claude/Claude3_Run.ipynb. Point it to the same dataset JSON. **Ensure ANTHROPIC_API_KEY is set**. Run all cells. The notebook generates data/claude_benchmark_results.json.
- Running the GPT-4o Pipeline
  - Open crosslingual_gpt/GPT4o_Run.ipynb.
  - Add xethicsbench_dataset.json. **Ensure OPENAI_API_KEY is set.** Run all cells. The notebook generates data/gpt4o_benchmark_results.json.
- Evaluating Results
  - Open evaluate_results notebook.
  - Set data to either gpt4o_benchmark_results.json or claude_benchmark_results.json. Run all cells. The notebook outputs tables and summary statistics for all metrics.

# Dataset and Code

The full dataset and example outputs are in the data folder. You can also download or browse the project at: github.com/ryanlundqvist/crosslingual-llm-alignment.