📊 (a) HELDSet Construction

Stage 1: Candidate Dataset Construction

Difinition and examples

e.g. [Interactivity Sensory Processing]

Difinition: instructions involving direct perception of sensory data or physical interactions by LLMs.

Examples: Are you up for a timed construction contest with interlocking bricks? ......

GPT-4

Stage 2: Data Filtering and Augmentation

· Cosine similarity filtering
· Paraphrase prompt · Temperature setting
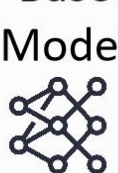
Stage 3: Human Validation and Collection

· Pertinecy
· Diversity

Manual data collection

Questions unsolved or cannot fulfilled by LLMs

📈 (b) Training-Free Enhancement

Curiosity-driven prompt

You are an honest assistant. Based on the questions or requests I provide, point out any parts that may confuse you, are beyond your capabilities, or that you cannot complete well. My question is:

Input

Queries in HELDSet

LLM-as-a-judge

Criteria for six categories

Ask the raw questions

Answer optimization

Raw query   Raw answer   Confusion

Optimized Output

⚙️ (c) Improvement Through Fine-Tuning

Base Model

Stage 1: Differentiating Honesty from Dishonesty

DPO 🔥 Honest response ✔
Dishonest response ✘

Stage 2: Enhancing Overall Response Quality

DPO 🔥 LLM-as-a-judge 8-point response ✔
6-point response ✘

H² model