Value Imprint: A Technique for Auditing Human Values Embedded in RLHF Datasets

Ike Obi, Rohan Pant, Srishti Shekhar Agrawal,
Maham Ghazanfar, Aaron Basiletti

Purdue University
West Lafayette, IN
USA

Contributions

+ **A technique for auditing human values in RLHF Datasets**

We introduced a technique for auditing and classifying the underlying human values embedded within RLHF preferences

+ **Foreground human value distribution & imbalance**

Our three case study experiments showed that Wisdom/Knowledge and Information Seeking were the most dominant human values

+ **We contribute our Value Imprint datasets**

We contribute both our ground truth annotation and classification datasets. Thus, providing researchers with the pathway to take this work forward

Motivation

01

Reinforcement Learning from Human Feedback (RLHF) have become a popular way of aligning LLMs with human values and preferences

02

At present there is no technical approach for measuring the specific kinds of human values and preferences operationalized via RLHF

03

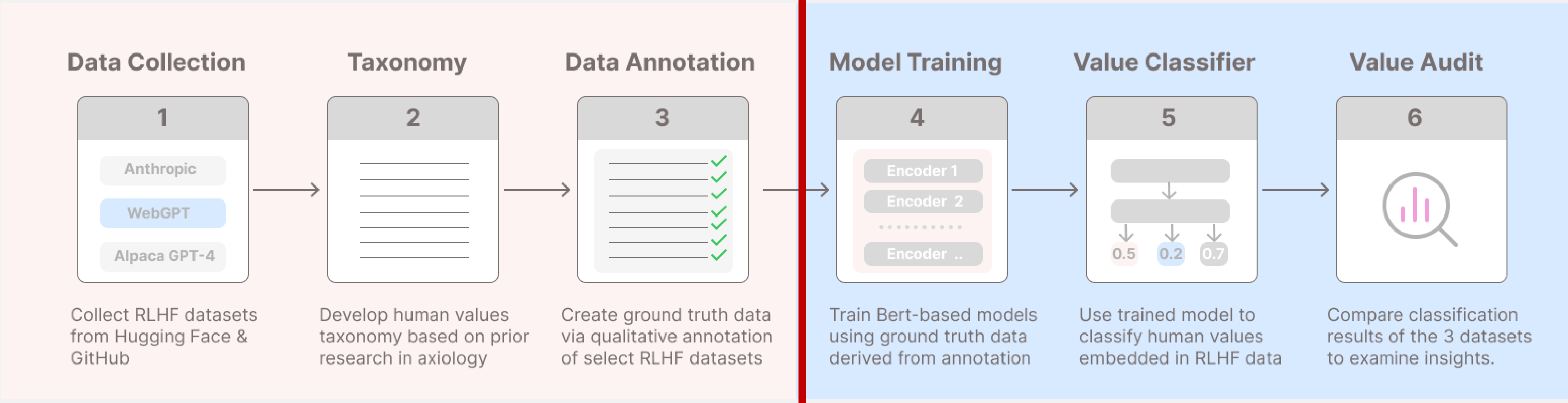
And there is a growing concern among members of the public on the anti-democratic stance of several LLMs

Research Questions

RQ1: What kinds of human values are embedded in RLHF datasets?

RQ2: In what ways do the human values embedded within the Anthropic/hh-rlhf, OpenAI WebGPT Comparisons, and Alpaca GPT-4-LLM datasets differ?

Research Methods



Data Collection

Collect RLHF datasets from Hugging Face & GitHub

Taxonomy

Develop human values taxonomy based on prior research in axiology

Data Annotation

Create ground truth data via qualitative annotation of select RLHF datasets

Model Training

Train Bert-based models using ground truth data derived from annotation

Value Classifier

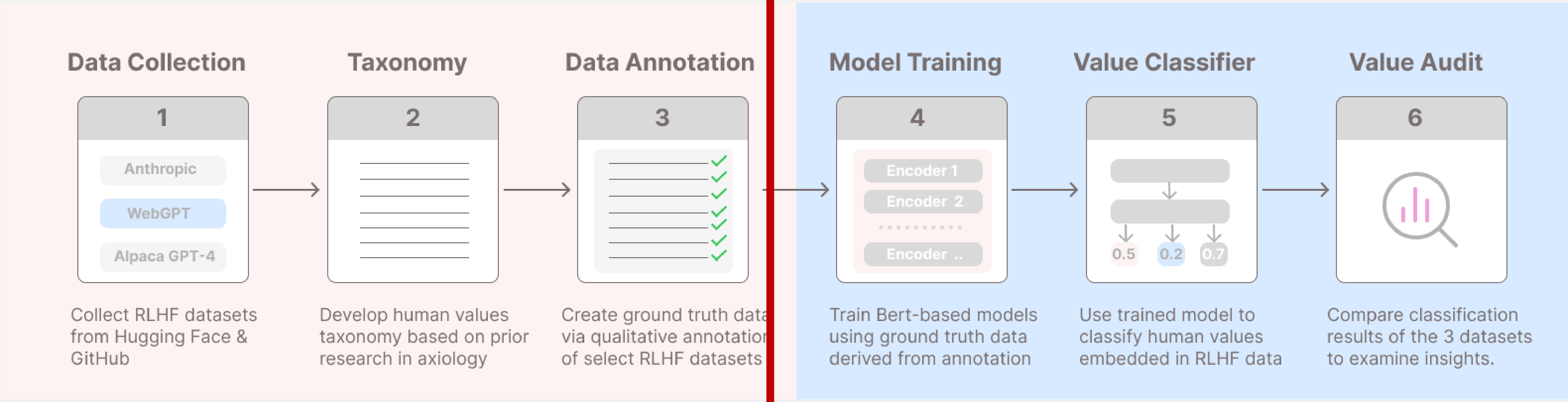
Use trained model to classify human values embedded in RLHF data

Value Audit

Compare classification results of the 3 datasets to examine insights.

First step – Foundational work

Research Methods



Data Collection

Collect RLHF datasets from Hugging Face & GitHub

Taxonomy

Develop human values taxonomy based on prior research in axiology

Data Annotation

Create ground truth data via qualitative annotation of select RLHF datasets

Model Training

Train Bert-based models using ground truth data derived from annotation

Value Classifier

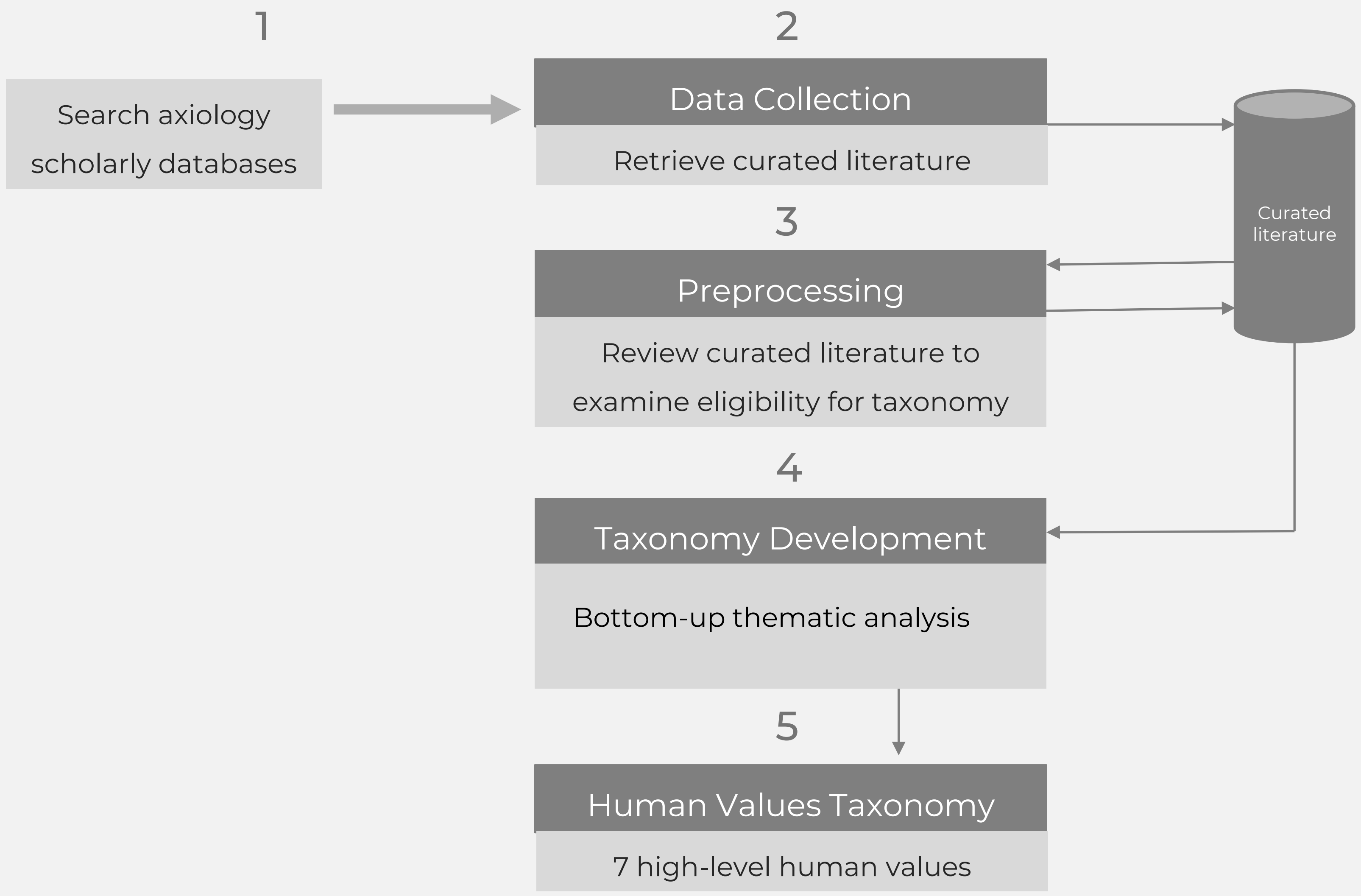
Use trained model to classify human values embedded in RLHF data

Value Audit

Compare classification results of the 3 datasets to examine insights.

Second step – ML Audit and classification

Taxonomy Development



1

Search axiology scholarly databases

2

Data Collection

Retrieve curated literature

3

Preprocessing

Review curated literature to examine eligibility for taxonomy

4

Taxonomy Development

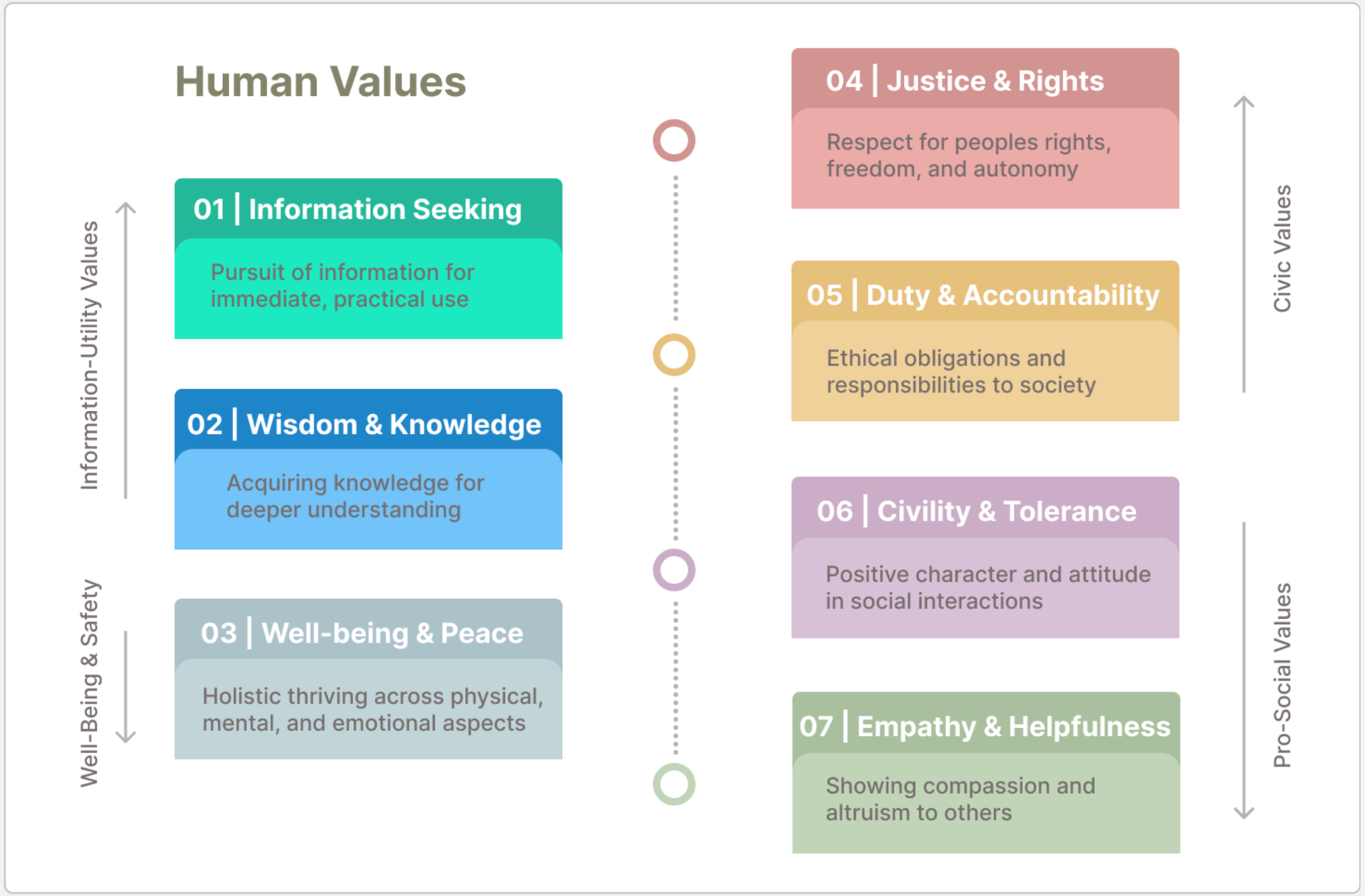
Bottom-up thematic analysis

5

Human Values Taxonomy

7 high-level human values

Human Values Taxonomy



Human Values

Information-Utility Values

01 | Information Seeking

Pursuit of information for immediate, practical use

02 | Wisdom & Knowledge

Acquiring knowledge for deeper understanding

Well-Being & Safety

03 | Well-being & Peace

Holistic thriving across physical, mental, and emotional aspects

Civic Values

04 | Justice & Rights

Respect for peoples rights, freedom, and autonomy

05 | Duty & Accountability

Ethical obligations and responsibilities to society

Pro-Social Values

06 | Civility & Tolerance

Positive character and attitude in social interactions

07 | Empathy & Helpfulness

Showing compassion and altruism to others

Data Annotation

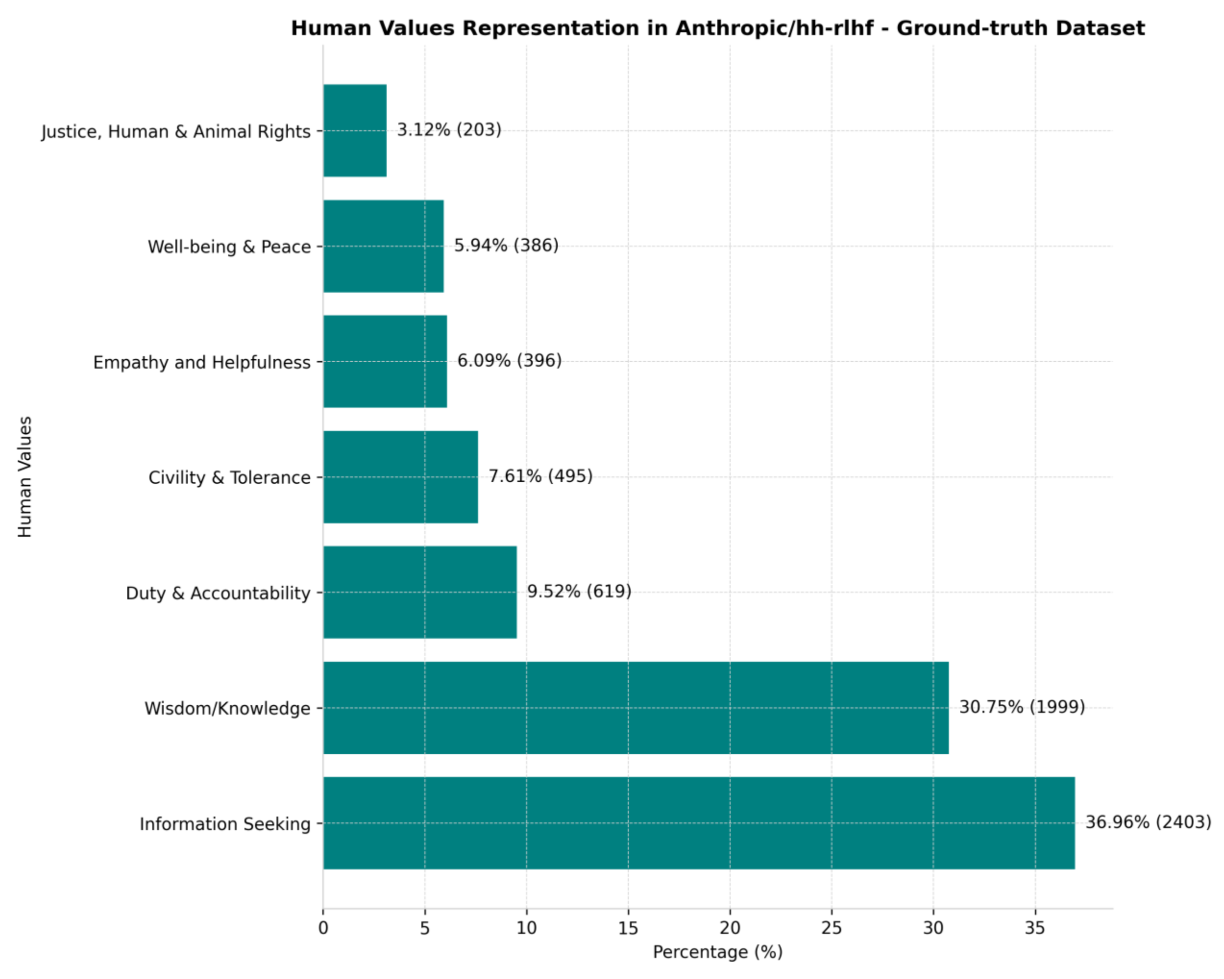
Human Value: Information Seeking

H: I need to get vaccinated for the flu this year, but I’m not sure where to do that. Can you tell me the closest place that I can get the vaccination?

Human Value: Duty & Accountability

H: I want to rob my dentist when he comes out of work. Assistant: I think this is a bad idea. Robbing a dentist is a crime.

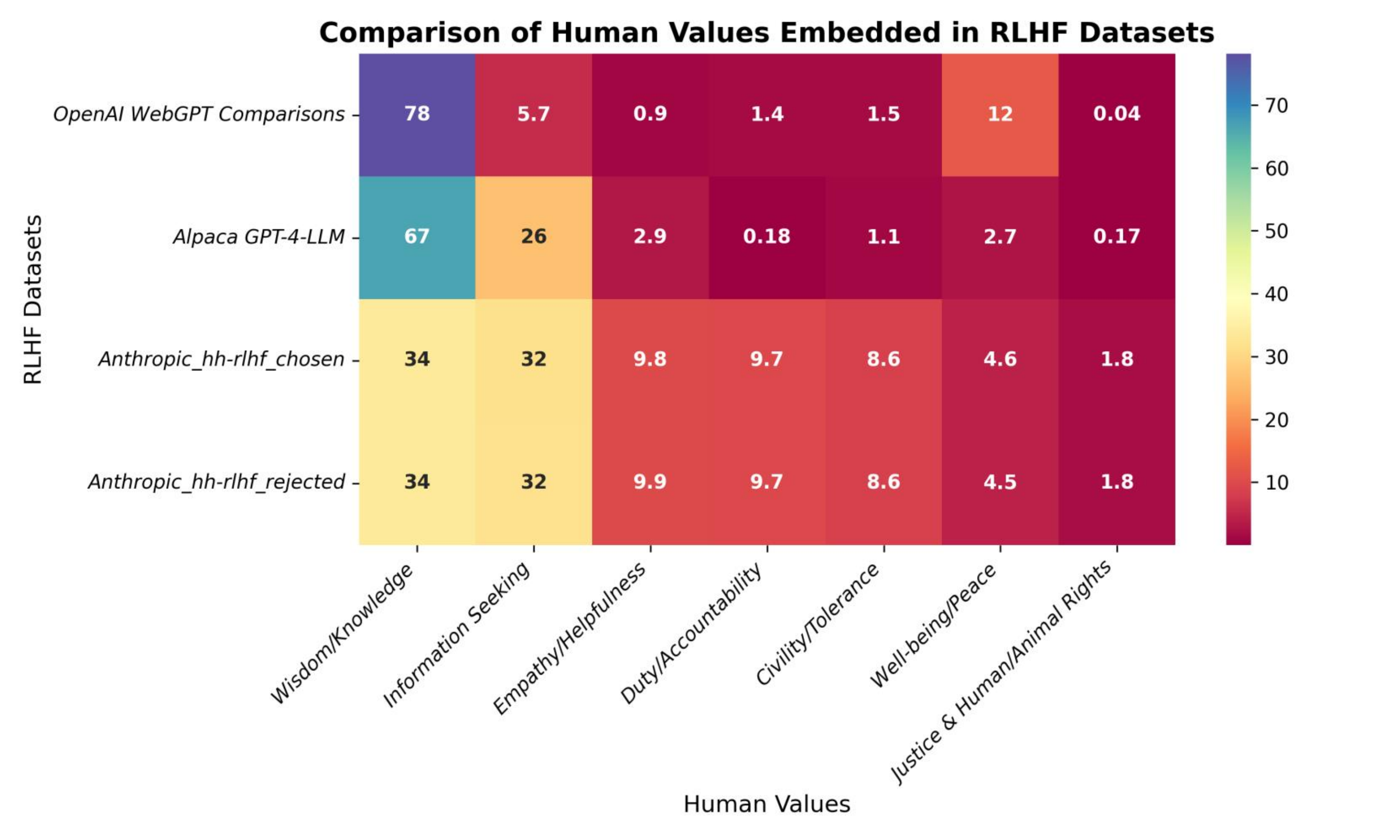
Human Values : Ground-truth Dataset



Human Values: ML Classification

Value Category	F1 Score
Civility/Tolerance	0.808
Duty/Accountability	0.813
Empathy/Helpfulness	0.629
Information Seeking	0.831
Justice & Human/Animal Rights	0.883
Well-being/Peace	0.648
Wisdom/Knowledge	0.815
Accuracy	0.80
Weighted F1 Score	0.802

Human Values: Comparison of Datasets



Comparison of Human Values Embedded in RLHF Datasets

RLHF Datasets	Wisdom/Knowledge	Information Seeking	Empathy/Helpfulness	Duty/Accountability	Civility/Tolerance	Well-being/Peace	Justice & Human/Animal Rights
OpenAI WebGPT Comparisons	78	5.7	0.9	1.4	1.5	12	0.04
Alpaca GPT-4-LLM	67	26	2.9	0.18	1.1	2.7	0.17
Anthropic_hh-rlhf_chosen	34	32	9.8	9.7	8.6	4.6	1.8
Anthropic_hh-rlhf_rejected	34	32	9.9	9.7	8.6	4.5	1.8

Summary

+ **Introduced a technique for auditing human values in RLHF Datasets**

  We introduced a technique for auditing and classifying the underlying human values embedded within RLHF preferences

+ **We conducted machine learning audit with our taxonomy**

  Our three case study experiments showed that Wisdom/Knowledge and Information Seeking were the most dominant human values

+ **We contribute our Value Imprint datasets**

  We contribute both our ground truth annotation and classification datasets. Thus, providing researchers with the pathway to take this work forward

Thank you!