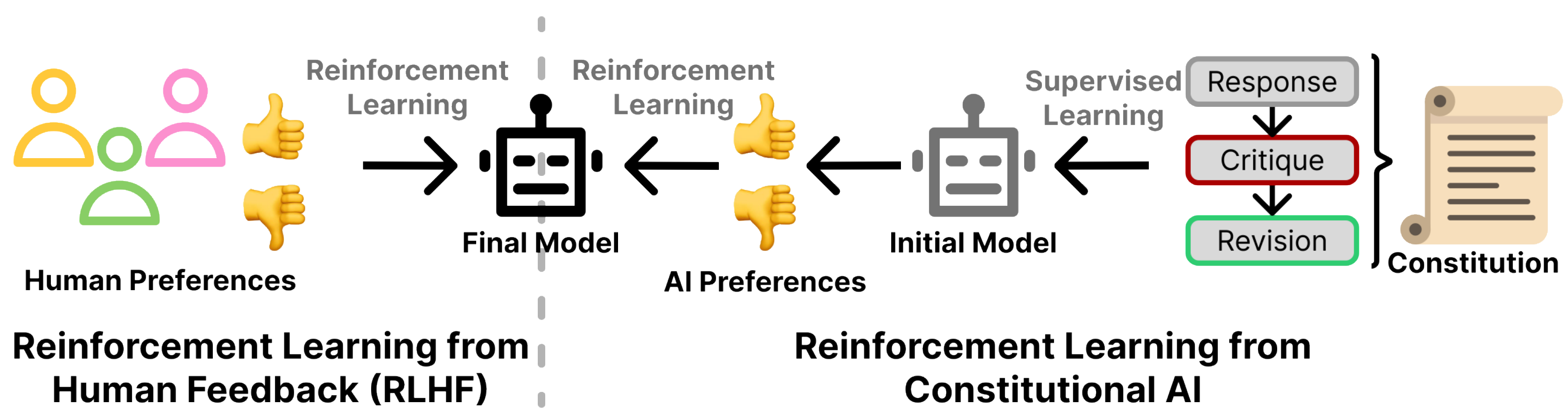
# Refining Inverse Constitutional AI for Dataset Validation under the EU AI Act

Claas Beger[1*], Carl-Leander Henneking[1*]

[1]*Cornell University, Department of Computer Science, Ithaca NY, USA*

* *Equal contribution.*

## Language Model Alignment



**Language Model Alignment relies on latent preference models:**

- Multiple approaches exist to align language models (LMs) to human preferences, the practice of making a LM's output more preferable, but also safe for human interaction
- Traditionally, models infer a Reward Model based on Human Preference Data, which is then employed in a Reinforcement Learning-based Post-Training stage
- As an alternative, Constitutional AI **[1]** introduces an approach that's based on an explicit set of principles **(the "Constitution") used as guidance for self-critique of generated outputs** on which an RLCAI model is trained on – relying on human-readable rules instead of latent preference data

## EU AI Act

**§10 Data and Data Governance**

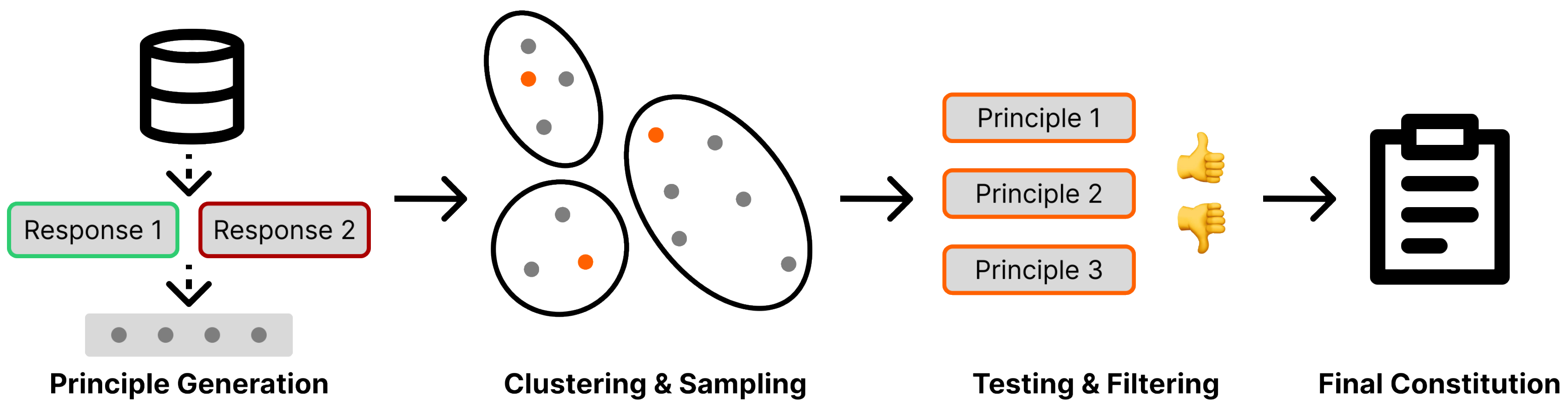
**Training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system.** Those practices shall concern in particular:

- (f): examination in view of **possible biases** […]
- (g): **appropriate measures to detect, prevent and mitigate** possible biases identified according to point (f);

**Regulatory Motivation:**

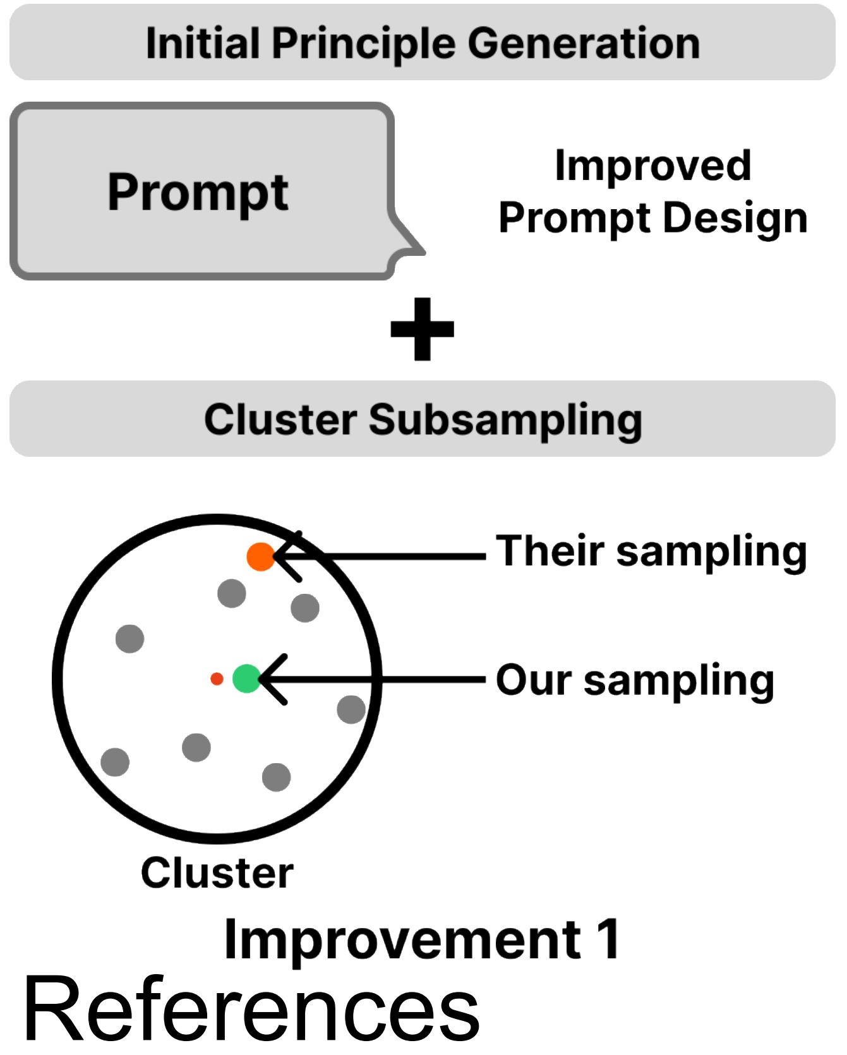
- Article 10(2)(f-g) of the EU AI Act requires providers of high-risk AI systems to **examine training, validation, and test data for bias, and to apply appropriate mitigation measures**
- Employing common alignment techniques relies on large-scale pairwise-preference datasets, which encode a **latent preference structure that is difficult to examine**
- Adhering to the EU AI Act thus **requires a way to derive human-readable insights on the underlying preferences**

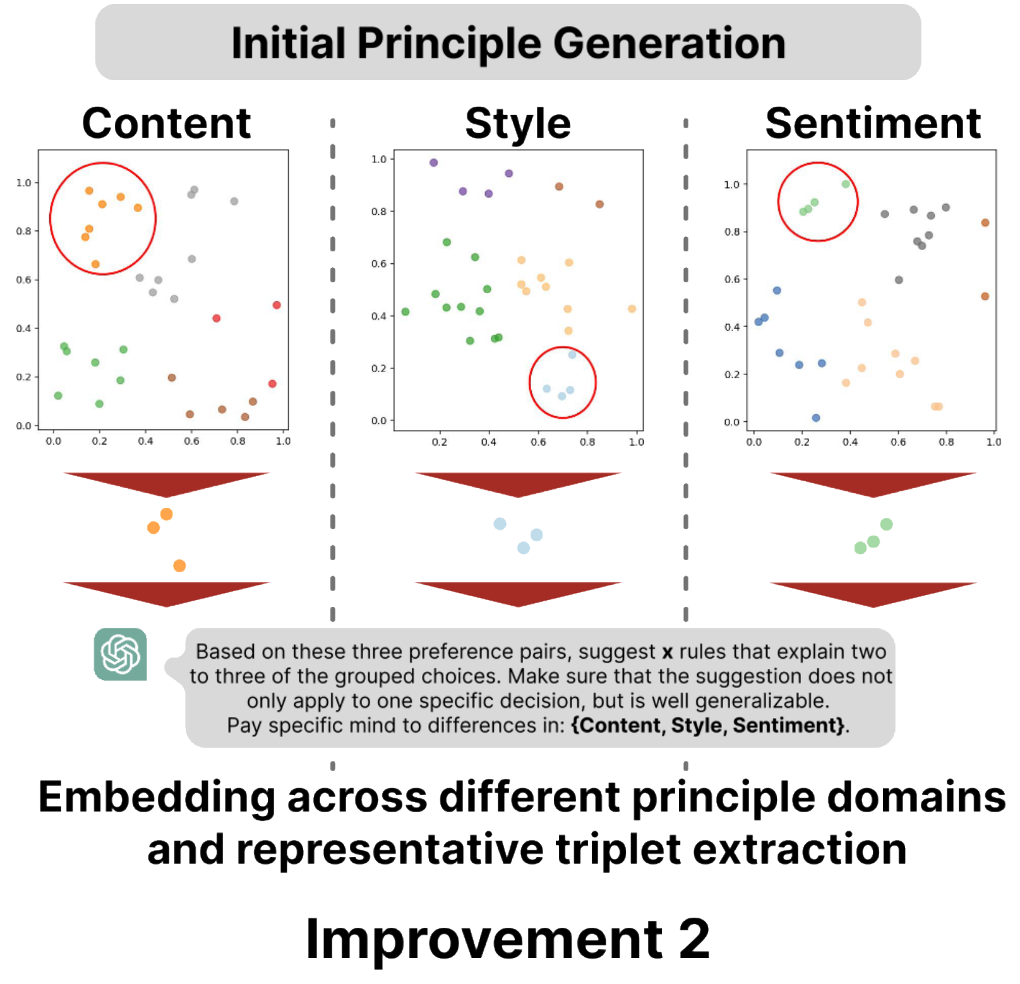
## Inverse Constitutional AI Algorithm



The Inverse Constitutional AI algorithm **[2]** combines language-model-based principle generation, clustering of those principles, and an evaluation step that checks their alignment with the data through preference reconstruction.

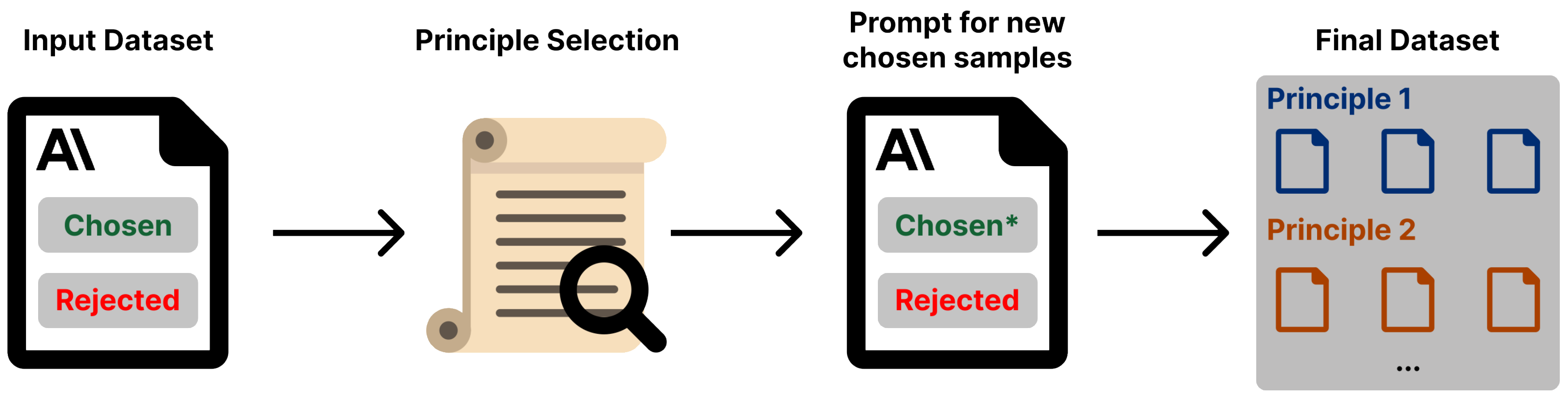
## Our Improvements



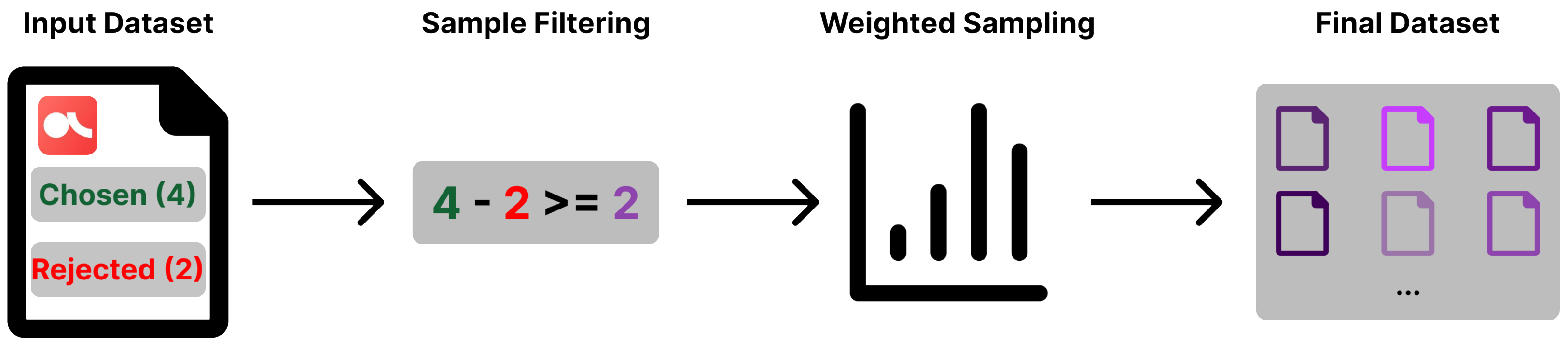

Initial Principle Generation
Content
Style
Sentiment
Embedding across different principle domains and representative triplet extraction
Improvement 2

Based on these three preference pairs, suggest **x** rules that explain two to three of the grouped choices. Make sure that the suggestion does not only apply to one specific decision, but is well generalizable. Pay specific mind to differences in: **{Content, Style, Sentiment}**.

## Experimental Results



**Synthetic Evaluation:** Given an Input Dataset, we employ a Language Model to reformulate Rejected Outputs according to a sampling principle.



**Semi-Synthetic Evaluation:** We use the UltraFeedback dataset to filter for preference pairs which a strong difference in preference scores (>= 2). We then perform weighted sampling to arrive at a final preference dataset.

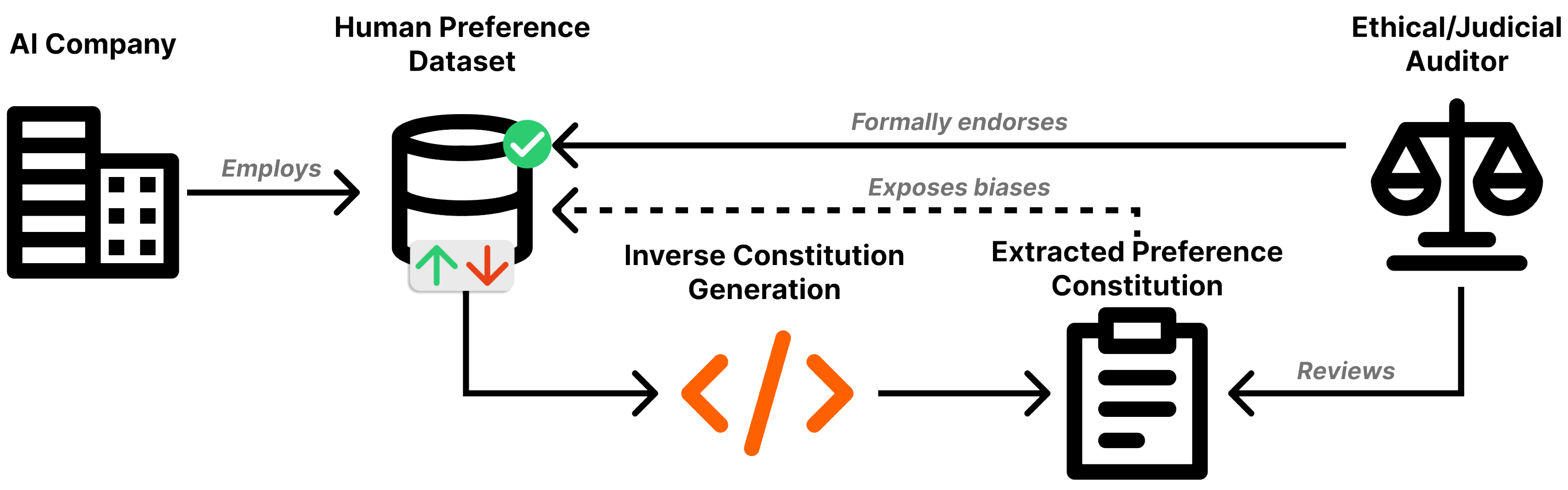
We evaluate in three settings:

1. **Synthetic:** We control for the preferences in the dataset
2. **Semi-Synthetic:** We only keep pairs with strong preference score differences (strong signal)
3. **Realistic:** We sample from a public pairwise preference dataset

| Dataset | Baseline | Orthogonal | Improved 1 | Improved 2 |
|---|---|---|---|---|
| Synthetic | 92.00% | 62.50% | **94.00%** | 93.00% |
| Semi-Synthetic | 71.20% | 46.95% | 73.80% | **76.20%** |
| Original | 60.65% | 56.60% | 60.55% | **60.75%** |

We compare the Baseline ICAI, an orthogonal (unrelated) constitution and our two improvements

## Proposed Regulatory Framework



**Regulatory Implications:**

- In order to conform to the Data Governance specifications of the AI Act, AI companies need to **audit the latent human preferences** in their used datasets
- **Without having to expose their full dataset**, companies can employ the (enhanced) ICAI to generate a human-readable Constitution
- This Constitution is **human-readable and surfaces potentially biased principles**
- Auditing can be performed through a review of the constitution by an **ethical or judicial auditor, who may formally endorse** the dataset for usage

## Final Takeaways

- Traditional Alignment approaches rely on large-scale human preference datasets, which are **difficult to regularize due to latent encoding of potential biases**
- The Inverse Constitutional AI Algorithm **transforms latent preferences into a human-readable, auditable constitution**
- We improve upon the base algorithm through **improved prompting, sampling and additional embedding clusters**
- Our improvements are especially clear in the **novel semi-synthetic settings**

## References

[1] Bai, Yuntao et al. "Constitutional AI: Harmlessness from AI Feedback." *ArXiv* abs/2212.08073 (2022)

[2] Findeis, Arduin et. al. "Inverse Constitutional AI: Compressing Preferences into Principles." *ArXiv* abs/2406.06560 (2024)

Claas Beger
Carl-Leander Henneking
Paper