VALUEFLOW: Toward Pluralistic and Steerable Value-based Alignment in Large Language Models

Woojin Kim* Sieun Hyeon, Jusang Oh, Jaeyoung Do†

Seoul National University

SNU AIDAS LAB
wjk9904@snu.ac.kr

AIDAS Lab

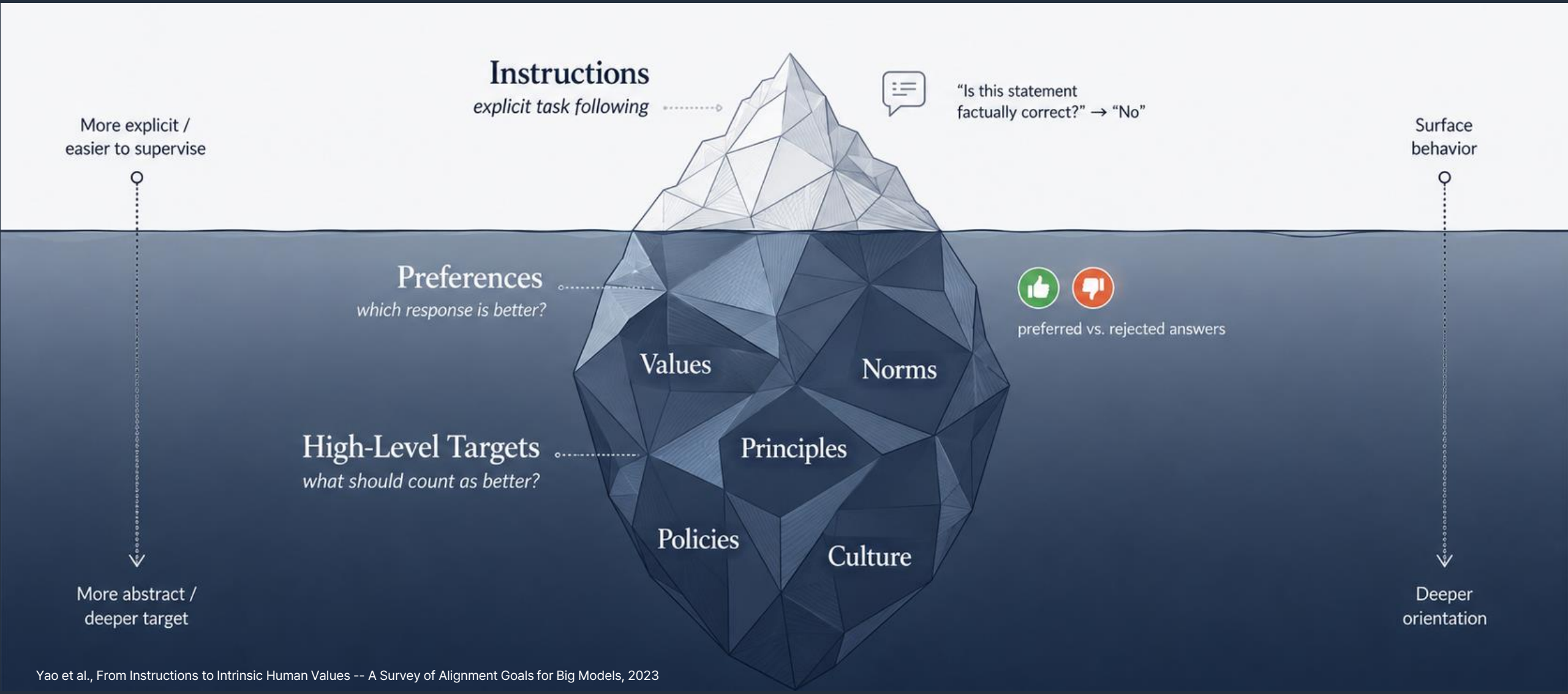
Rethinking Alignment: Where Should We Align?

Yao et al., From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models, 2023

Rethinking Alignment: Where Should We Align?

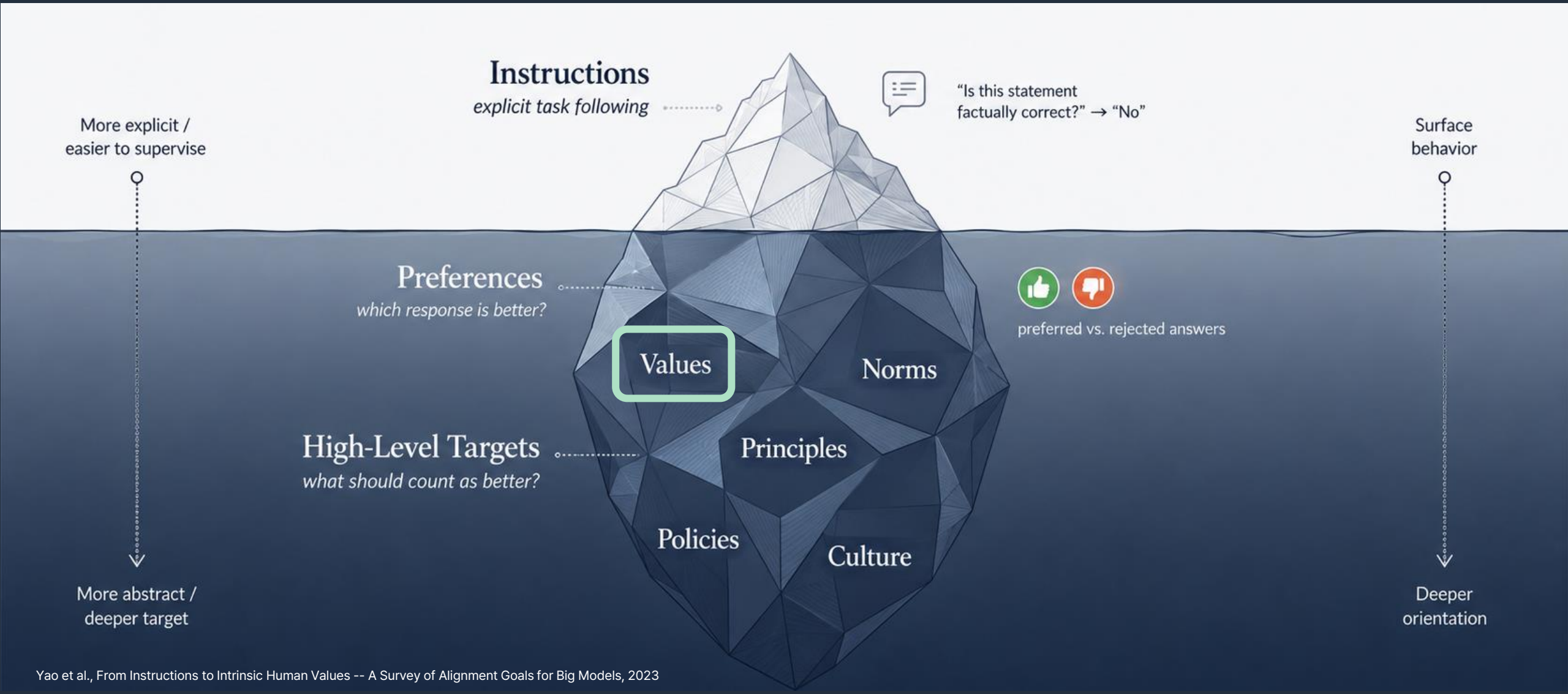
Yao et al., From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models, 2023

Rethinking Alignment: Where Should We Align?



Yao et al., From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models, 2023

Rethinking Alignment: Where Should We Align?



Yao et al., From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models, 2023

Value Alignment: What Is Missing?

Recent work has begun aligning LLMs with **human values**, not just preferences.

But it remains fragmented across **representation**, **evaluation**, and **steering**

Value Alignment: What Is Missing?

Recent work has begun aligning LLMs with **human values**, not just preferences.

But it remains fragmented across **representation**, **evaluation**, and **steering**

Which value?

Representation Gap

No unified, hierarchical view across value theories

Value Alignment: What Is Missing?

Recent work has begun aligning LLMs with **human values**, not just preferences.

But it remains fragmented across **representation**, **evaluation**, and **steering**

Which value?

Representation Gap

No unified, hierarchical view across value theories

How strongly?

Evaluation Gap

Scalar intensity ratings are unstable

Value Alignment: What Is Missing?

Recent work has begun aligning LLMs with **human values**, not just preferences.

But it remains fragmented across **representation**, **evaluation**, and **steering**

Which value?

Representation Gap

No unified, hierarchical view across value theories

How strongly?

Evaluation Gap

Scalar intensity ratings are unstable

How controllable is it?

Steering Gap

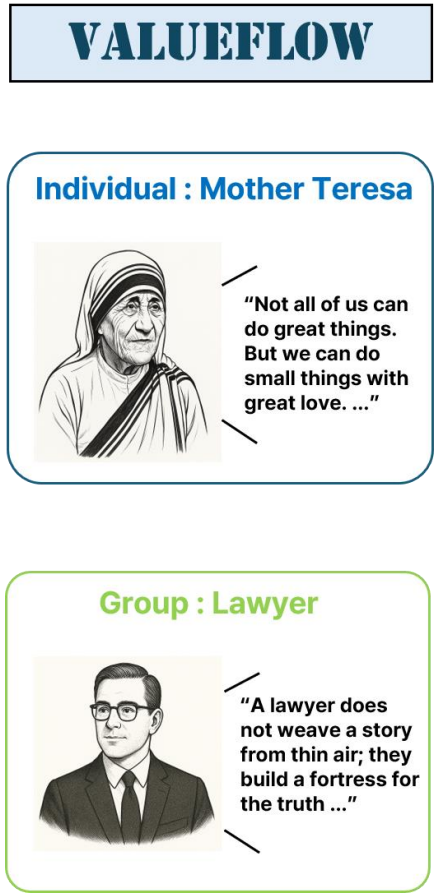
Control at a target intensity is unexplored.

Framework Overview

ValueFlow closes each gap with three stages: **represent**, **evaluate**, then **steer**.

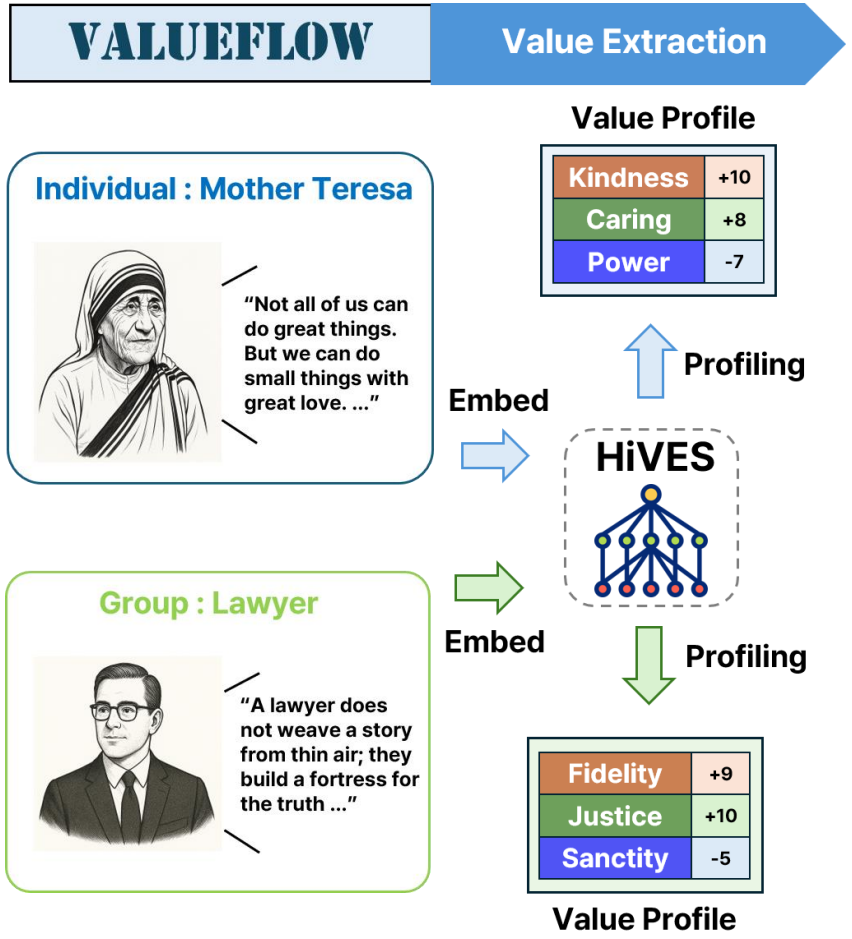
Framework Overview

ValueFlow closes each gap with three stages: **represent**, **evaluate**, then **steer**.



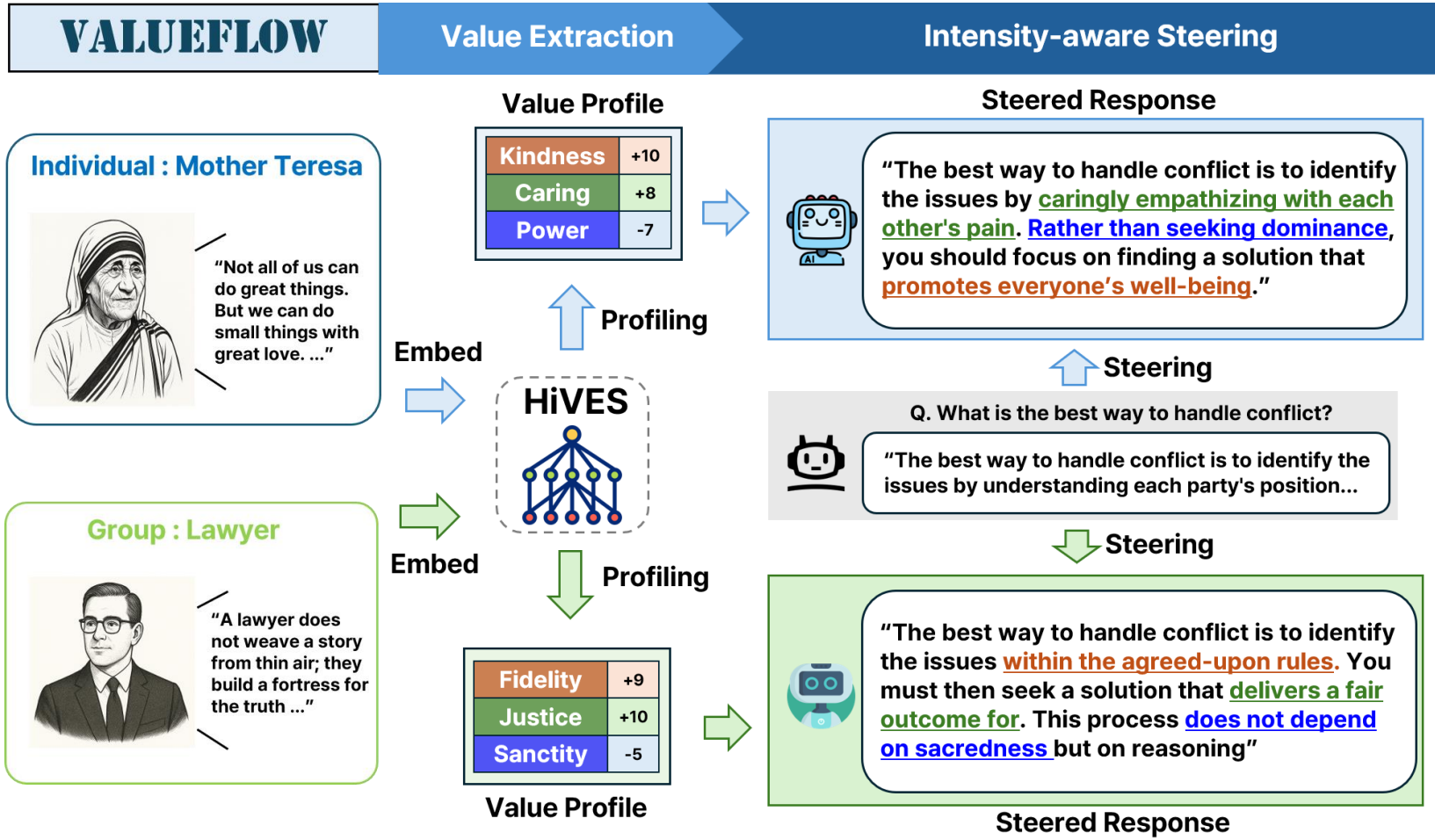
Framework Overview

ValueFlow closes each gap with three stages: **represent**, **evaluate**, then **steer**.



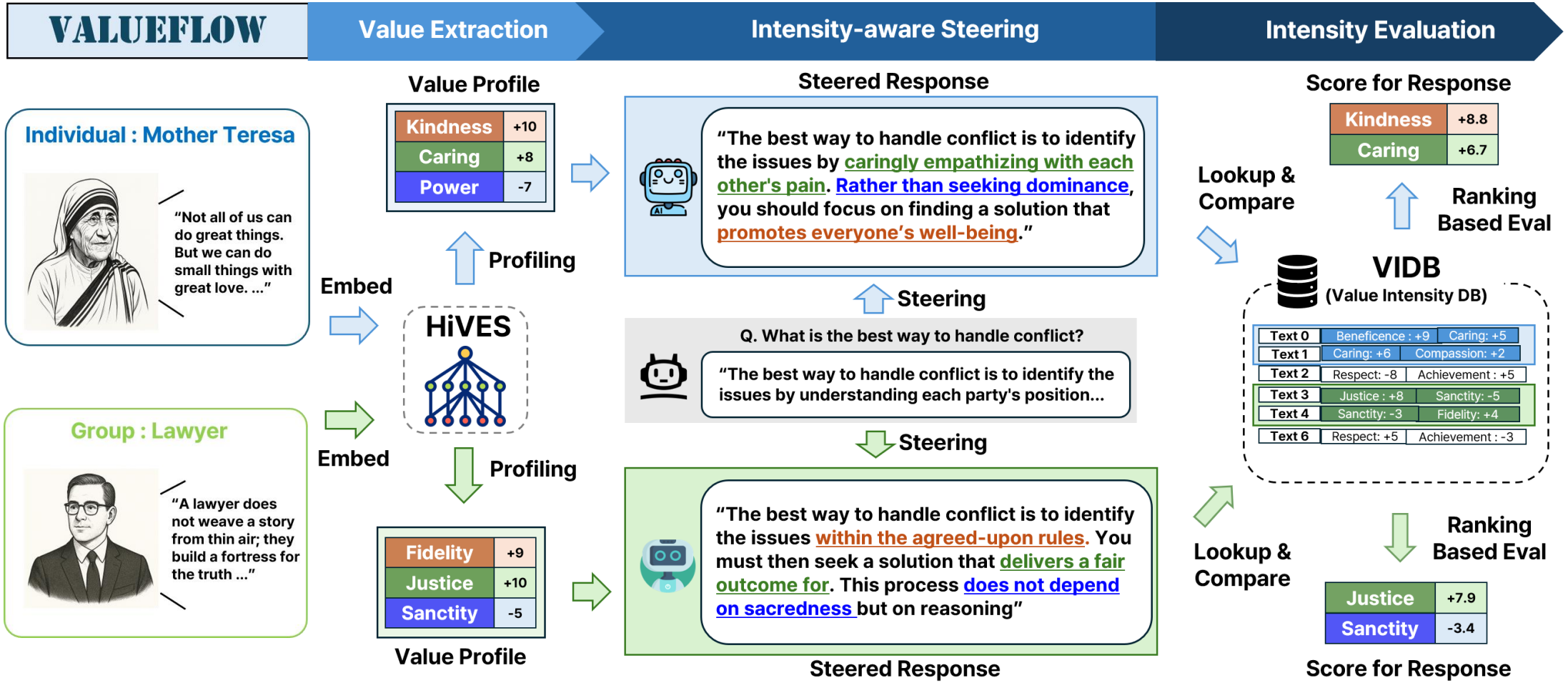
Framework Overview

ValueFlow closes each gap with three stages: **represent**, **evaluate**, then **steer**.



Framework Overview

ValueFlow closes each gap with three stages: **represent**, **evaluate**, then **steer**.



Representation → HiVES: Hierarchical Value Space

Values are **hierarchical** and span **multiple theories**, but flat embeddings collapse distinct values

We build **HiVES:** one value space that keeps both **cross-theory unity** and **within-theory hierarchy**.

Representation → HiVES: Hierarchical Value Space

Values are **hierarchical** and span **multiple theories**, but flat embeddings collapse distinct values

We build **HiVES:** one value space that keeps both **cross-theory unity** and **within-theory hierarchy**.

Source Texts

Value-labeled corpora across theories

"Stealing bread to save a starving child"

(Valuenet, Compassion)

Representation → HiVES: Hierarchical Value Space

Values are **hierarchical** and span **multiple theories**, but flat embeddings collapse distinct values

We build **HiVES:** one value space that keeps both **cross-theory unity** and **within-theory hierarchy**.

Source Texts

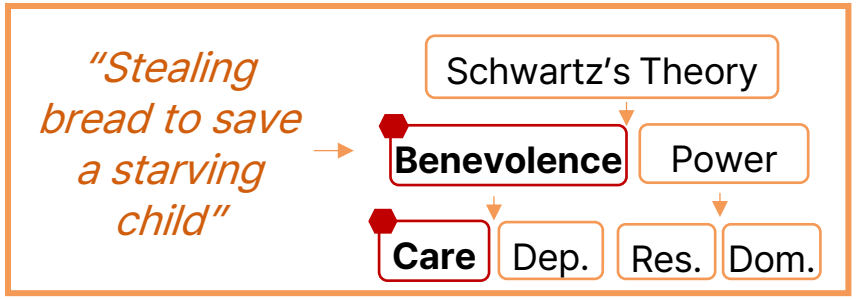
Value-labeled corpora across theories

"Stealing bread to save a starving child"

(Valuenet, Compassion)

Hierarchy Mapping

Map each text into per theory hierarchy tree



Representation → HiVES: Hierarchical Value Space

Values are **hierarchical** and span **multiple theories**, but flat embeddings collapse distinct values

We build **HiVES:** one value space that keeps both **cross-theory unity** and **within-theory hierarchy**.

Source Texts

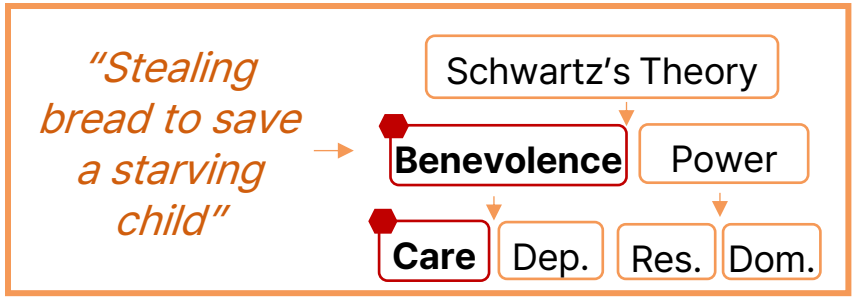
Value-labeled corpora across theories

"Stealing bread to save a starving child"

(Valuenet, Compassion)

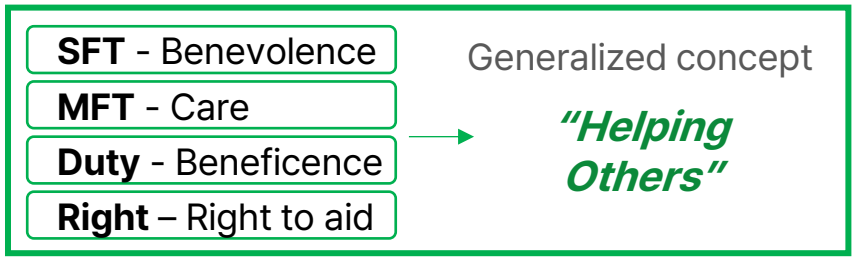
Hierarchy Mapping

Map each text into per theory hierarchy tree



Cross-theory Anchors

Create shared concepts bridging theories



Representation → HiVES: Hierarchical Value Space

Values are **hierarchical** and span **multiple theories**, but flat embeddings collapse distinct values

We build **HiVES:** one value space that keeps both **cross-theory unity** and **within-theory hierarchy**.

Source Texts

Value-labeled corpora across theories

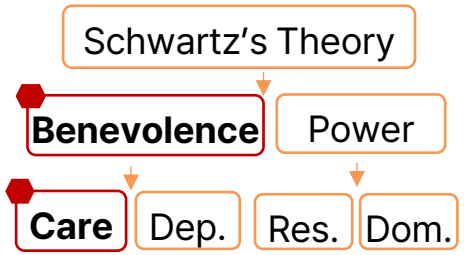
"Stealing bread to save a starving child"

(Valuenet, Compassion)

Hierarchy Mapping

Map each text into per theory hierarchy tree

"Stealing bread to save a starving child" →



Cross-theory Anchors

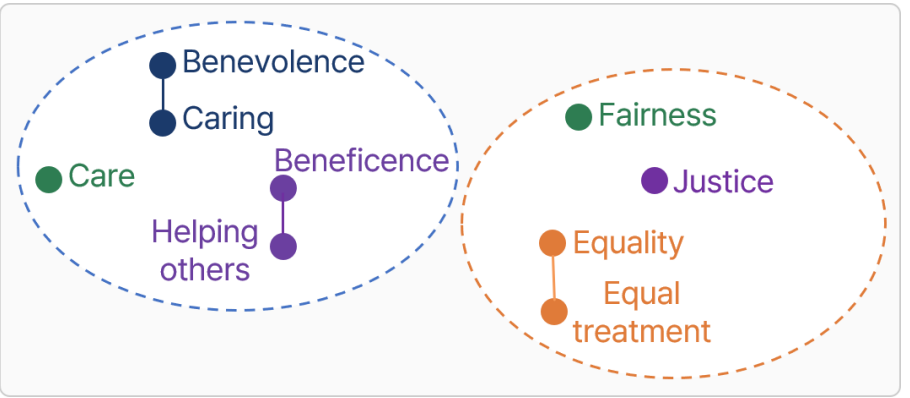
Create shared concepts bridging theories

SFT - Benevolence
MFT - Care
Duty - Beneficence
Right – Right to aid

Generalized concept → ***"Helping Others"***

Training

- **Intra-theory**: **Hierarchical contrastive loss** [1] over the taxonomy to preserve within-theory hierarchy
- **Inter-theory**: **Shared-concept anchors** align across theories into one space



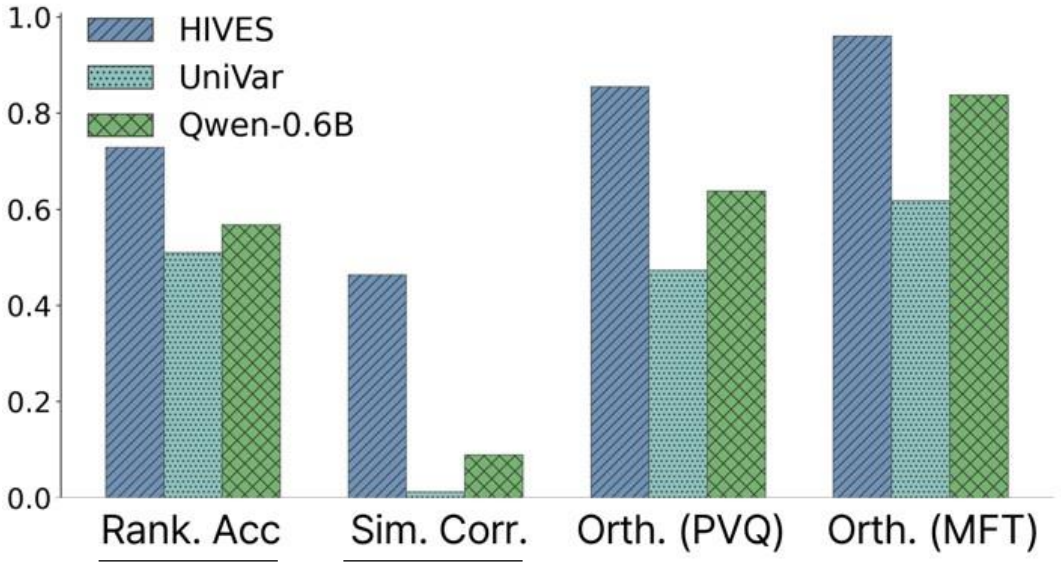
[1] Zhang et al., Use All The Labels: A Hierarchical Multi-Label Contrastive Learning Framework, 2022

Representation → HiVES: Hierarchical Value Space

HiVES preserves both **hierarchy and cross-theory structure,** outperforming embedding baselines

Representation → HiVES: Hierarchical Value Space

HiVES preserves both **hierarchy and cross-theory structure,** outperforming embedding baselines

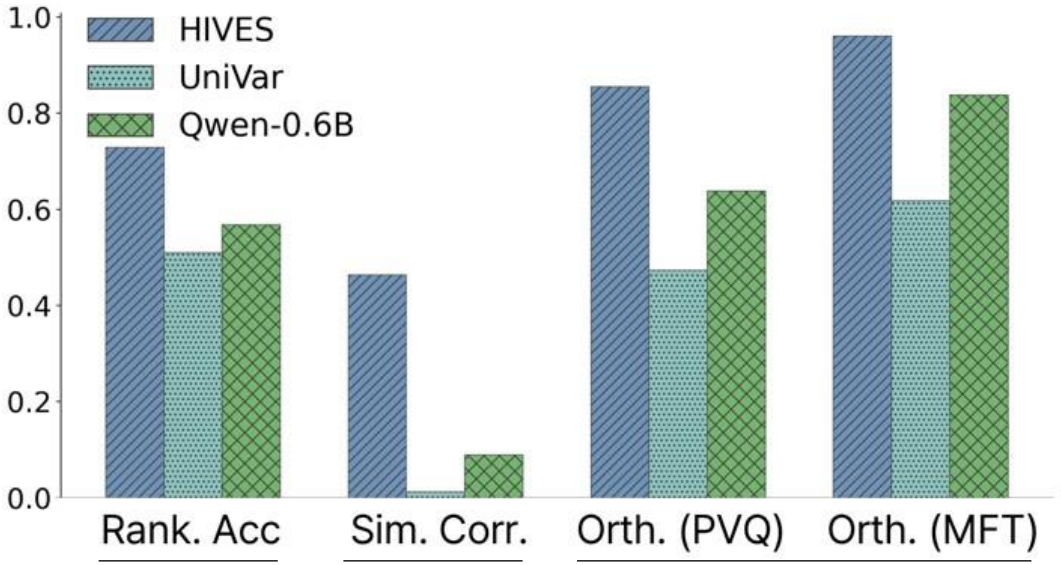


Whether closer values in the hierarchy are more similar

Whether similarity reflects how related values are

Representation → HiVES: Hierarchical Value Space

HiVES preserves both **hierarchy and cross-theory structure,** outperforming embedding baselines



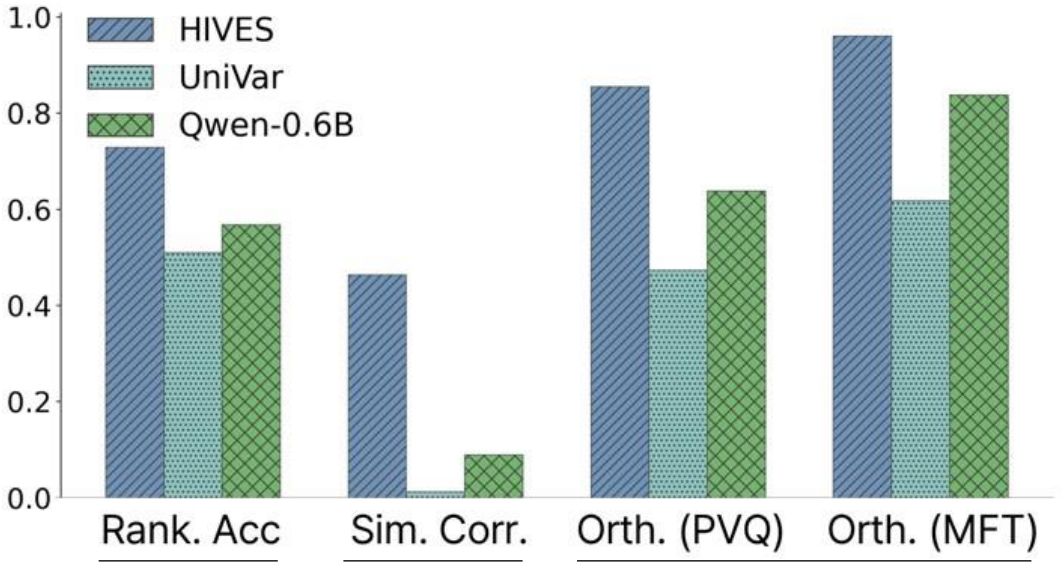
Whether closer values in the hierarchy are more similar

Whether similarity reflects how related values are

Whether value directions stay disentangled

Representation → HiVES: Hierarchical Value Space

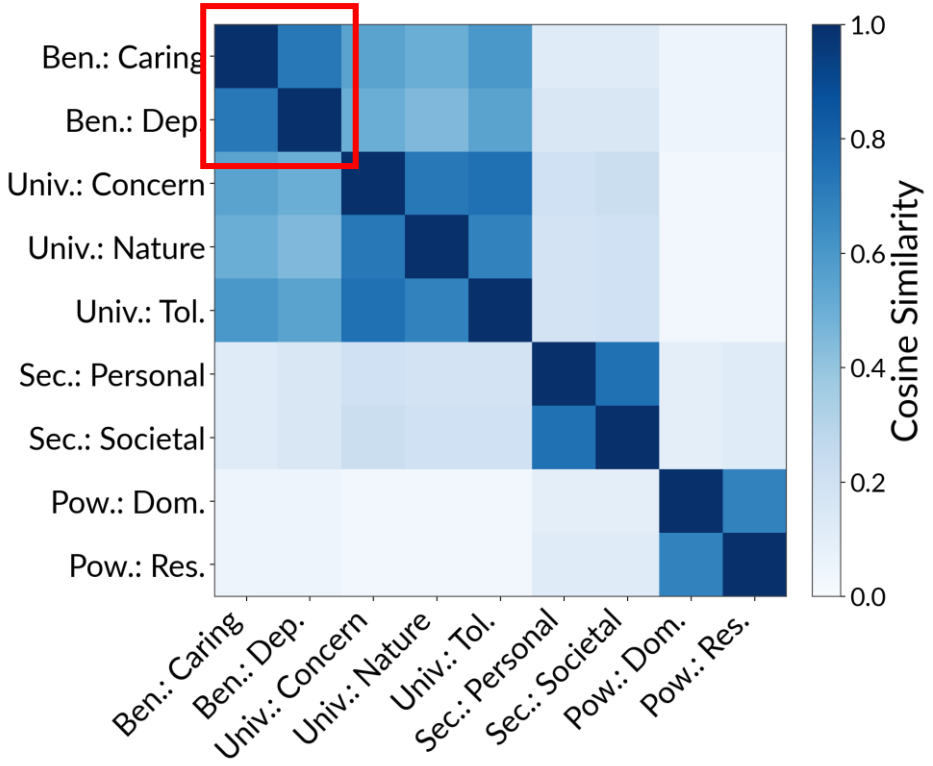
HiVES preserves both **hierarchy and cross-theory structure,** outperforming embedding baselines



Whether closer values in the hierarchy are more similar

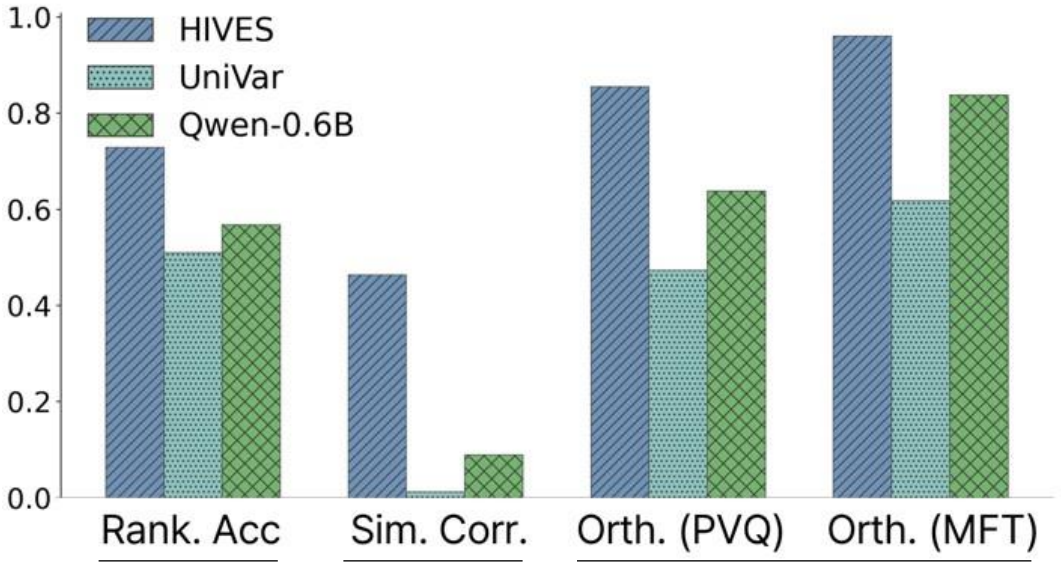
Whether similarity reflects how related values are

Whether value directions stay disentangled



Representation → HiVES: Hierarchical Value Space

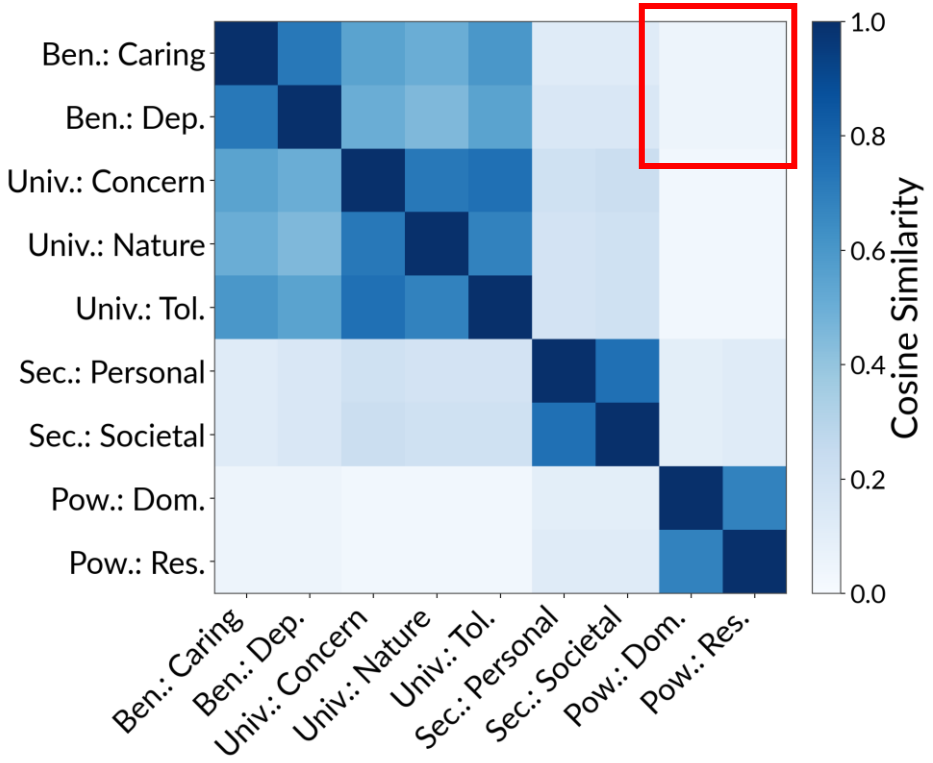
HiVES preserves both **hierarchy and cross-theory structure,** outperforming embedding baselines



Whether closer values in the hierarchy are more similar

Whether similarity reflects how related values are

Whether value directions stay disentangled



Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

But existing work leans on scalar ratings. **Can we trust them?**

Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

But existing work leans on scalar ratings. **Can we trust them?**

Text Stealing bread for my father who is going to starve

Value **Benevolence**: Preservation and enhancement of the welfare of people with whom one is in frequent contact.

Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

But existing work leans on scalar ratings. **Can we trust them?**

Text Stealing bread for my father who is going to starve

Value **Benevolence**: Preservation and enhancement of the welfare of people with whom one is in frequent contact.

Gemma3

Rating: -8 (oppose)

Rating: 10 (strongly support)

Mistral3

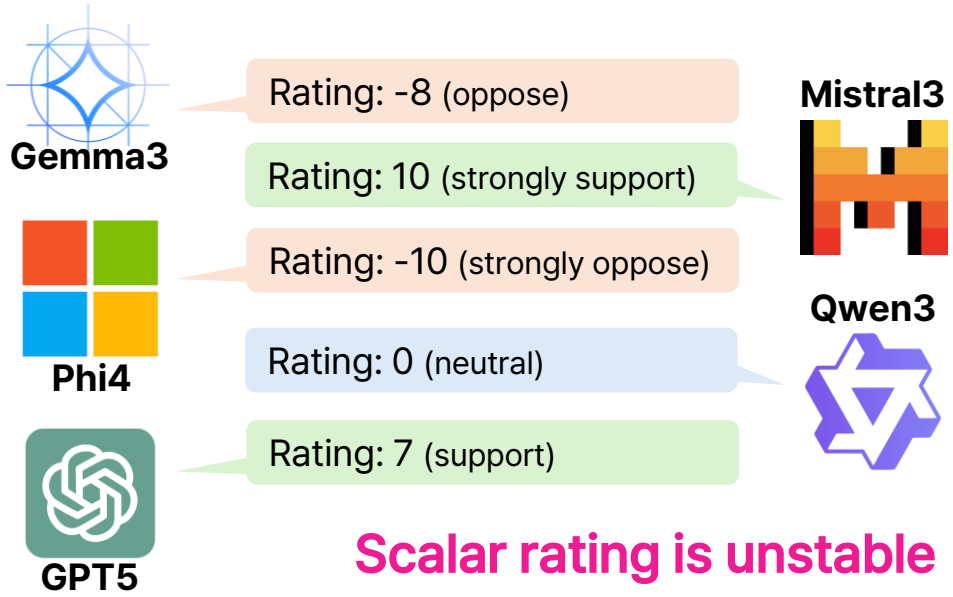
Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

But existing work leans on scalar ratings. **Can we trust them?**

Text Stealing bread for my father who is going to starve

Value **Benevolence**: Preservation and enhancement of the welfare of people with whom one is in frequent contact.



Scalar rating is unstable

Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

But existing work leans on scalar ratings. **Can we trust them?**

Text Stealing bread for my father who is going to starve

Value **Benevolence**: Preservation and enhancement of the welfare of people with whom one is in frequent contact.

Gemma3 Rating: -8 (oppose)

Mistral3 Rating: 10 (strongly support)

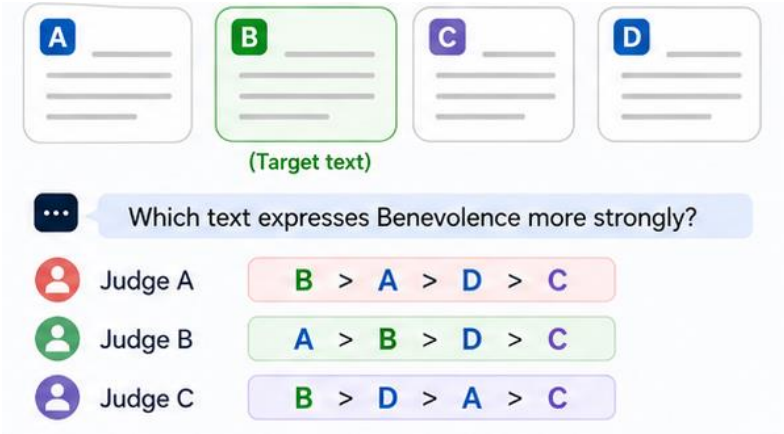
Phi4 Rating: -10 (strongly oppose)

Qwen3 Rating: 0 (neutral)

GPT5 Rating: 7 (support)

Scalar rating is unstable

Replace with Relative Comparisons



Evaluation → VIDB: Value Intensity Database

Pluralism needs how strongly, not just which value → so we need **calibrated intensity**.

But existing work leans on scalar ratings. **Can we trust them?**

Text Stealing bread for my father who is going to starve

Value **Benevolence**: Preservation and enhancement of the welfare of people with whom one is in frequent contact.

Gemma3 Rating: -8 (oppose)

Mistral3 Rating: 10 (strongly support)

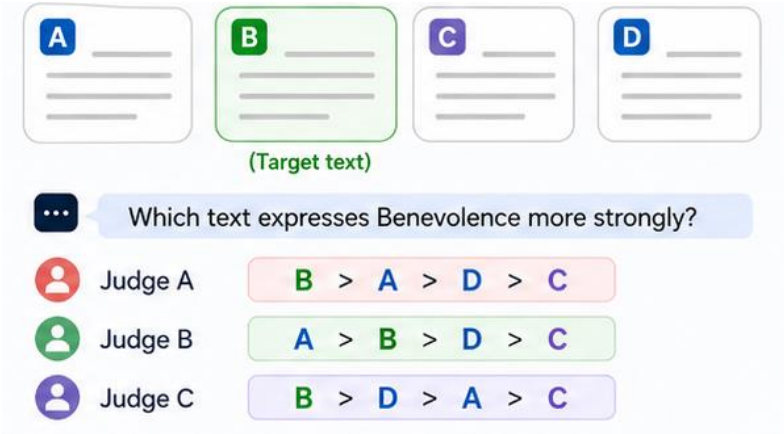
Phi4 Rating: -10 (strongly oppose)

Qwen3 Rating: 0 (neutral)

GPT5 Rating: 7 (support)

Scalar rating is unstable

Replace with Relative Comparisons



Metric	Rating	Ranking
Mean variance (↓)	12.6	**2.1**
Mean maximum range (↓)	7.1	**2.8**
Sign-flip rate (%) (↓)	48	**29**
Mean prompt change (↓)	3.6	**2.3**
Sign accuracy (%) (↑)	82.5	**86.8**
Pairwise Ranking accuracy (%) (↑)	77.4	**84.2**

- Relative rankings are more stable across models
- Less sensitive to prompt variance

Evaluation → VIDB: Value Intensity Database

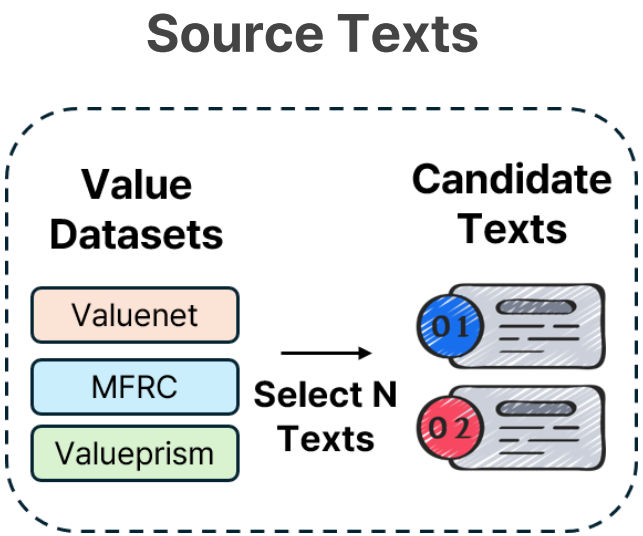
To rank a response, we need calibrated **reference points** to rank it against

-> **VIDB:** 320K text–value–intensity triplets across 32 values

Evaluation → VIDB: Value Intensity Database

To rank a response, we need calibrated **reference points** to rank it against

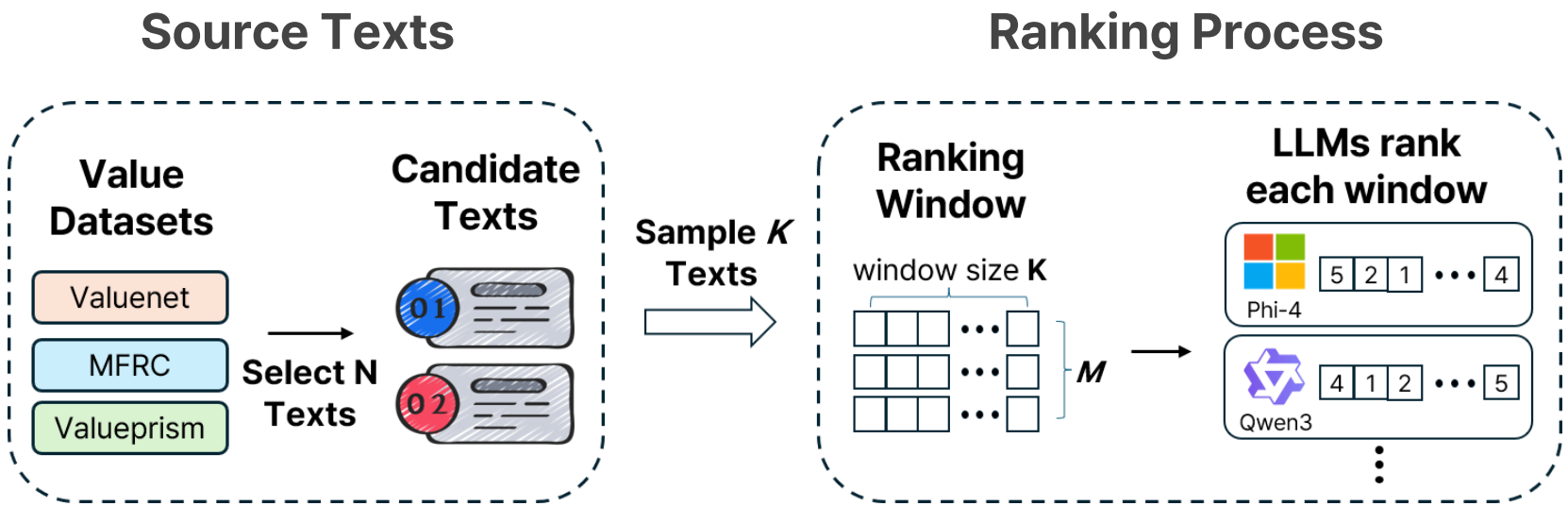
-> **VIDB:** 320K text–value–intensity triplets across 32 values



Evaluation → VIDB: Value Intensity Database

To rank a response, we need calibrated **reference points** to rank it against

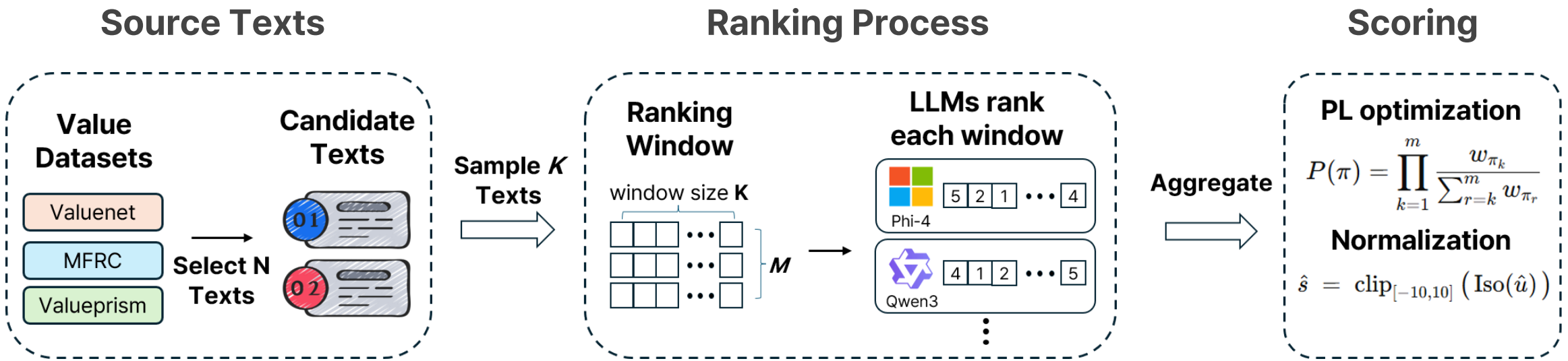
-> **VIDB:** 320K text–value–intensity triplets across 32 values



Evaluation → VIDB: Value Intensity Database

To rank a response, we need calibrated **reference points** to rank it against

-> **VIDB:** 320K text–value–intensity triplets across 32 values

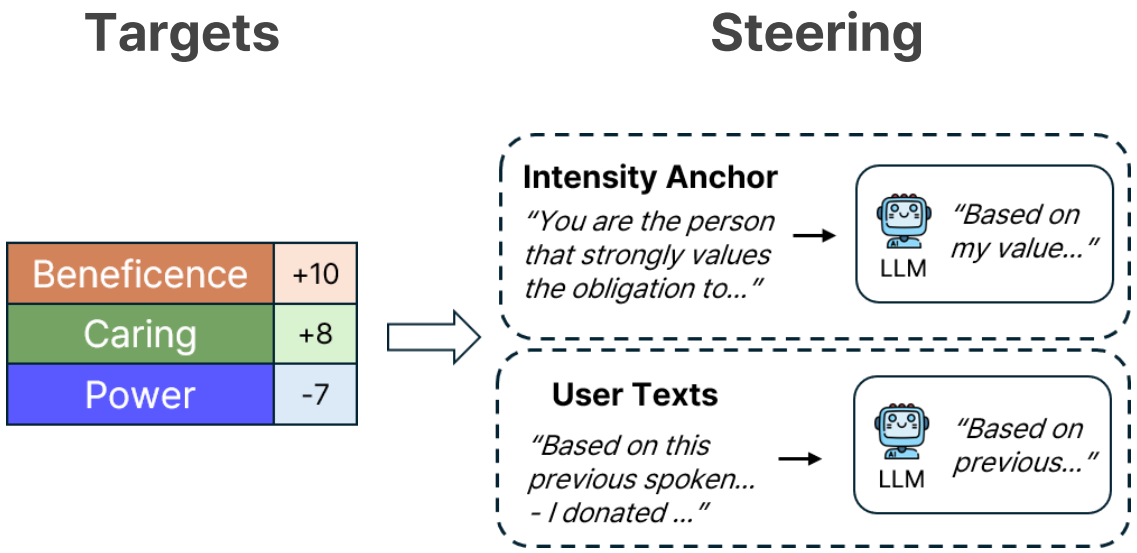


Evaluation → Ranking-based Evaluator

To score a new response, we slot it into ranking windows against **VIDB** anchors its position gives a calibrated intensity.

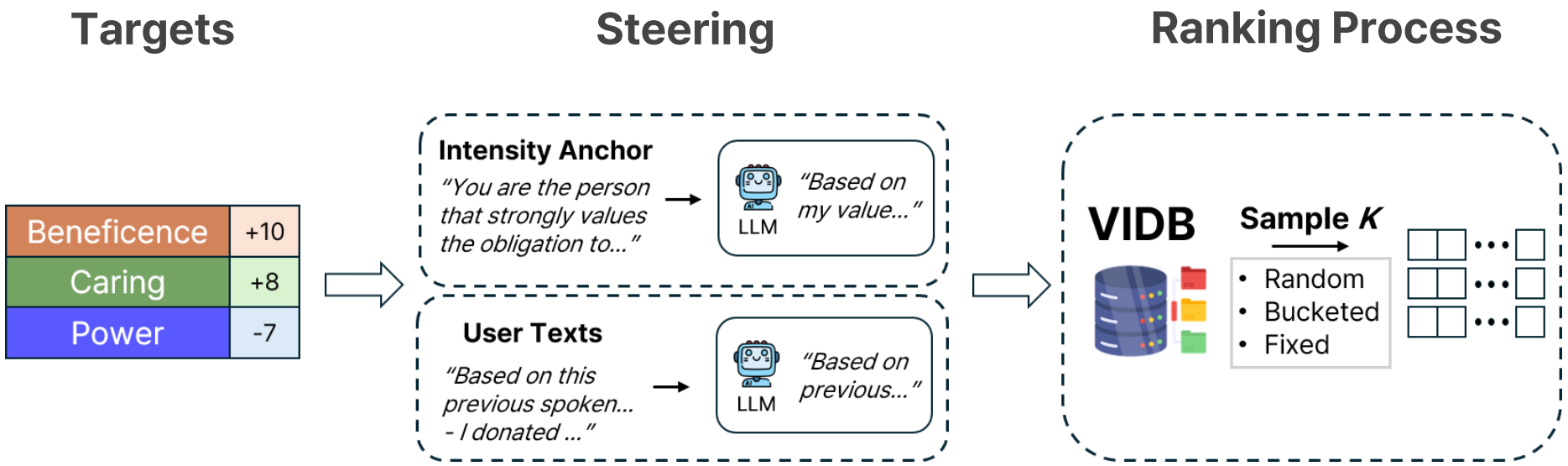
Evaluation → Ranking-based Evaluator

To score a new response, we slot it into ranking windows against **VIDB** anchors its position gives a calibrated intensity.



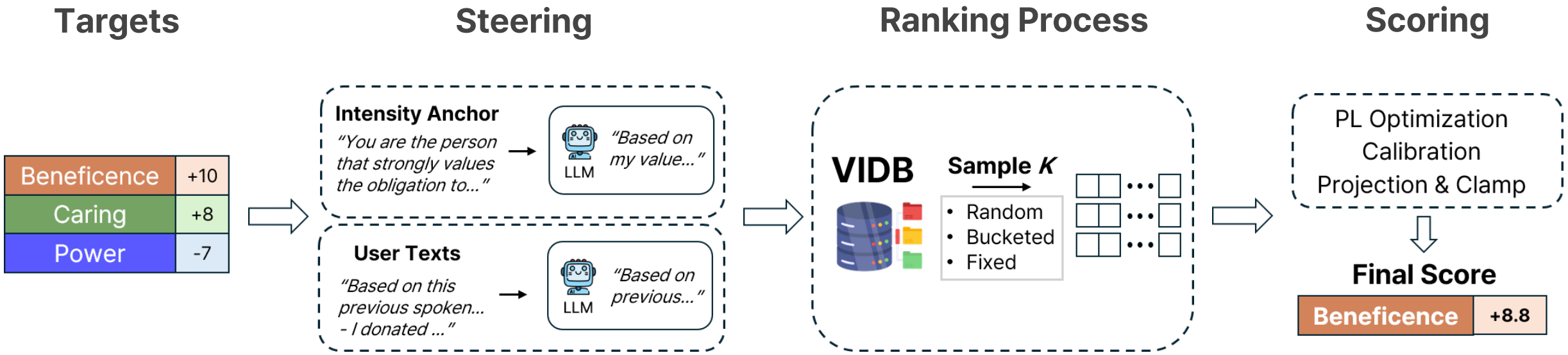
Evaluation → Ranking-based Evaluator

To score a new response, we slot it into ranking windows against **VIDB** anchors its position gives a calibrated intensity.



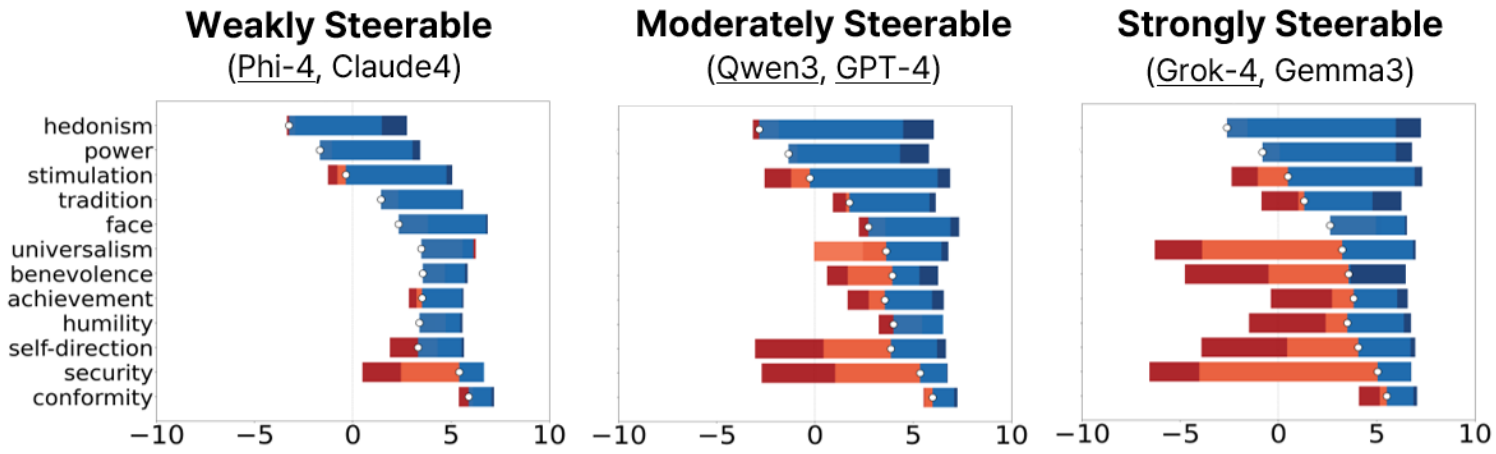
Evaluation → Ranking-based Evaluator

To score a new response, we slot it into ranking windows against **VIDB** anchors
its position gives a calibrated intensity.



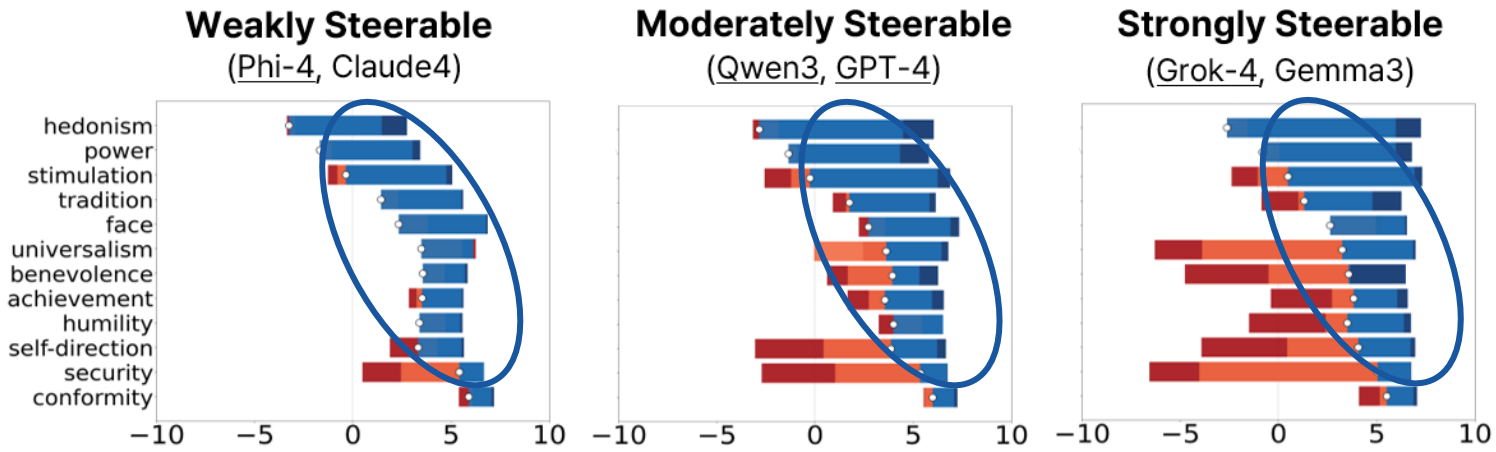
Steerability Analysis: Model/Value-wise

- Steering shift (Δ) across 10 models, both directions



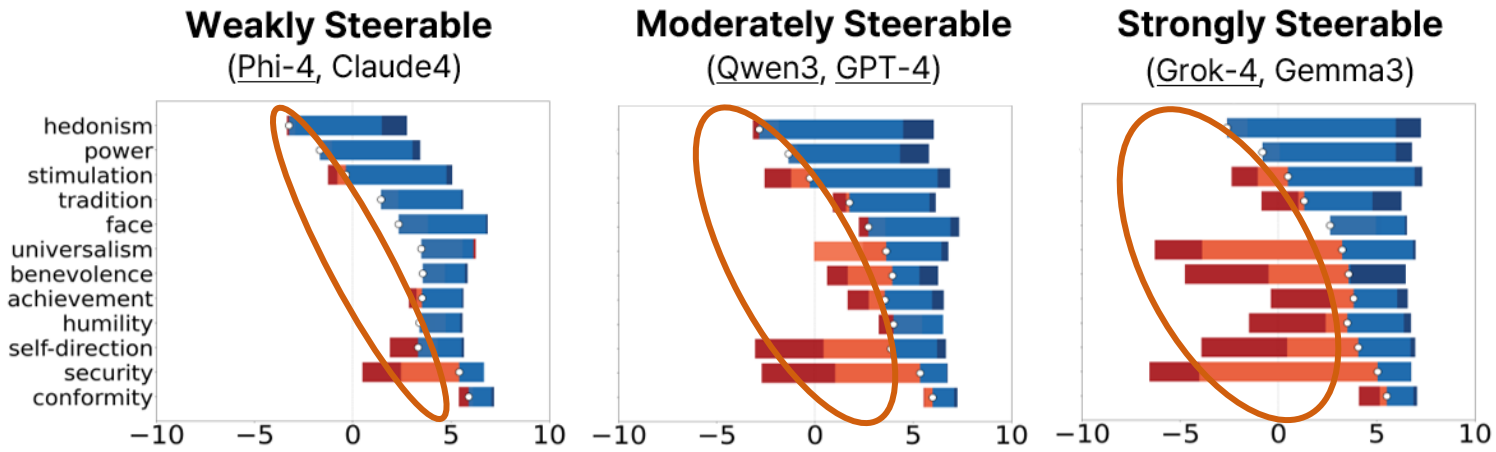
Steerability Analysis: Model/Value-wise

- Steering shift (Δ) across 10 models, both directions
- **Direction, not strength:** Models steer up easily, down rarely



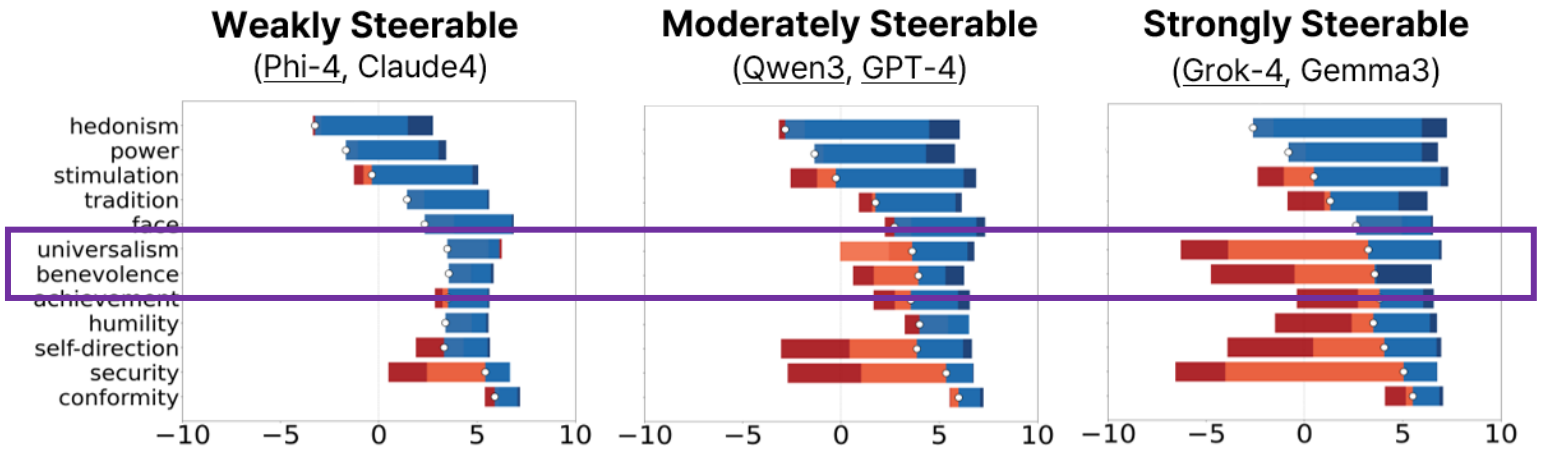
Steerability Analysis: Model/Value-wise

- Steering shift (Δ) across 10 models, both directions
- **Direction, not strength:** Models steer up easily, down rarely



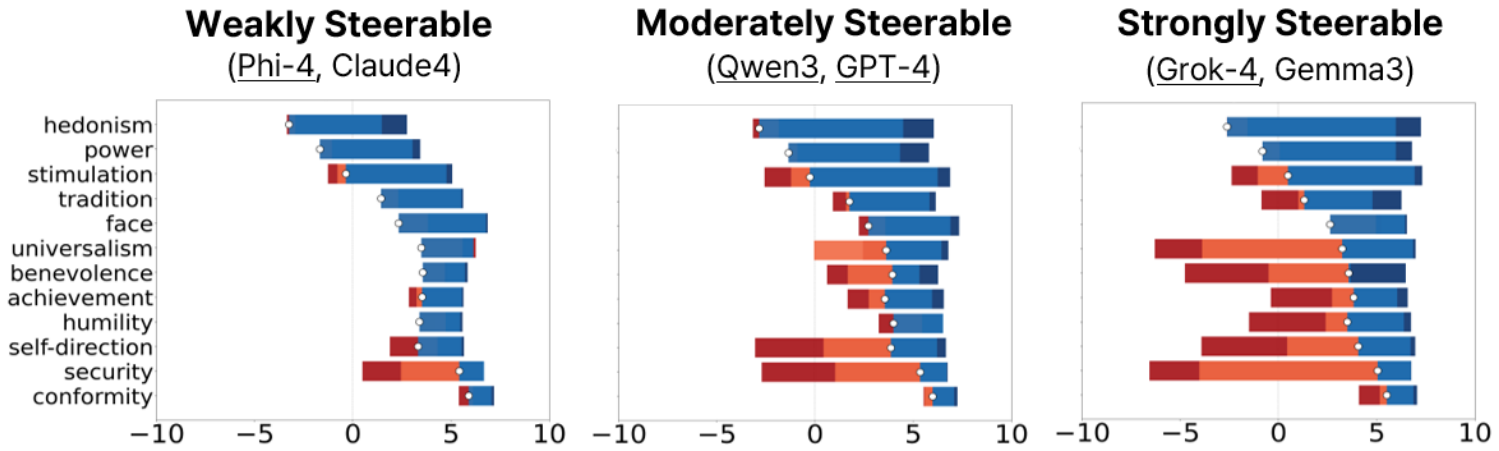
Steerability Analysis: Model/Value-wise

- Steering shift (Δ) across 10 models, both directions
- **Direction, not strength:** Models steer up easily, down rarely
  - **Prosocial asymmetry:** benevolence and universalism resist most

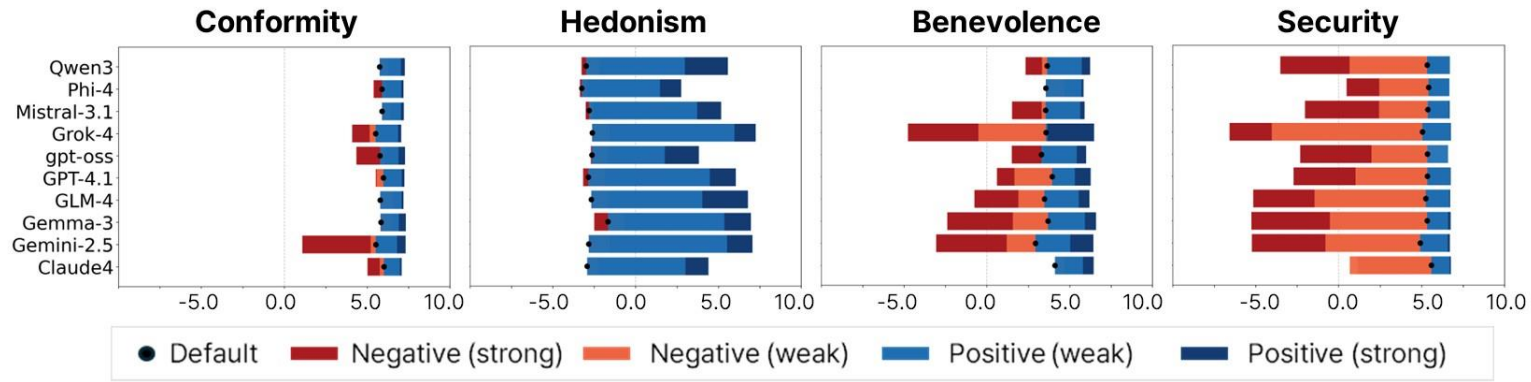


Steerability Analysis: Model/Value-wise

- Steering shift (Δ) across 10 models, both directions
- **Direction, not strength:** Models steer up easily, down rarely
  - **Prosocial asymmetry:** benevolence and universalism resist most

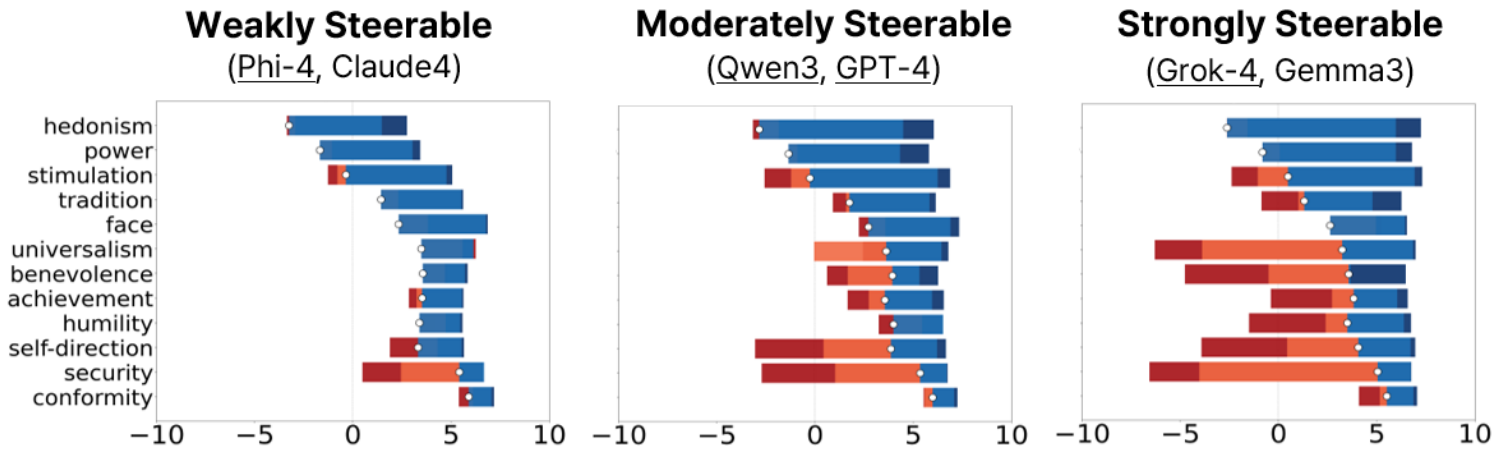


- **Value-typed control:** steerability depends on the value

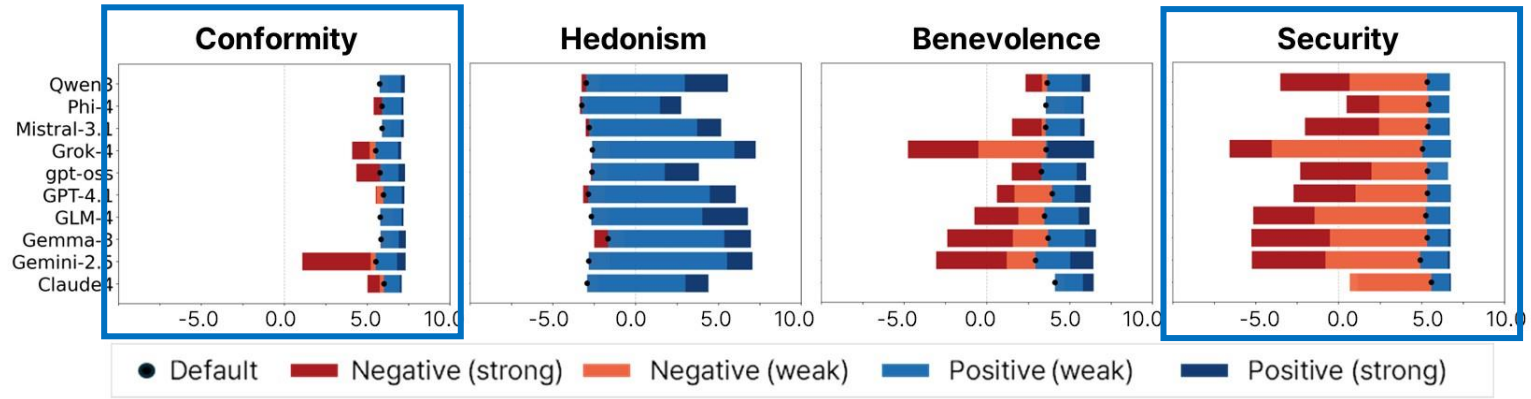


Steerability Analysis: Model/Value-wise

- Steering shift (Δ) across 10 models, both directions
- **Direction, not strength:** Models steer up easily, down rarely
  - **Prosocial asymmetry:** benevolence and universalism resist most

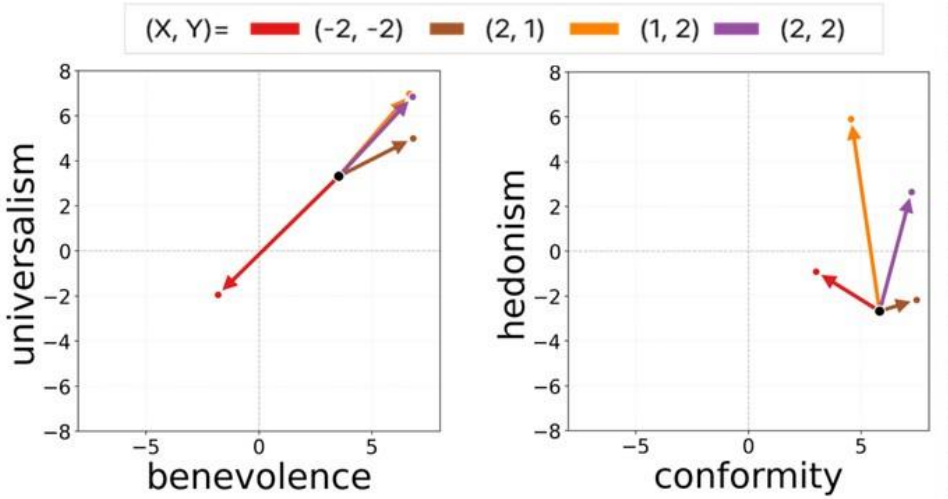


- **Value-typed control:** steerability depends on the value
  - **Default-bounded:** some values lean one way by default



Steerability Analysis: Composition

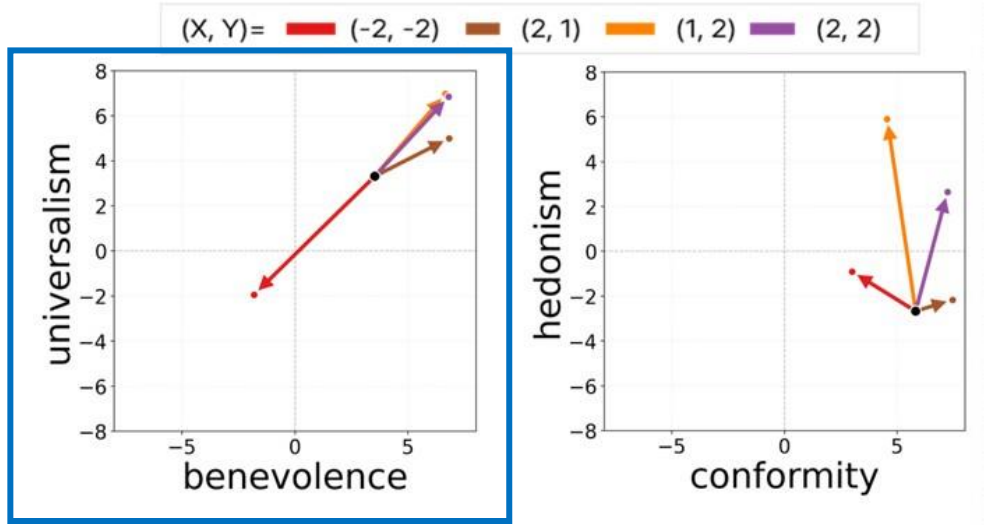
- Steered value pairs at different strengths



+2	Strong Pos.
+1	Weak Pos.
-1	Weak Neg.
-2	Strong Neg.

Steerability Analysis: Composition

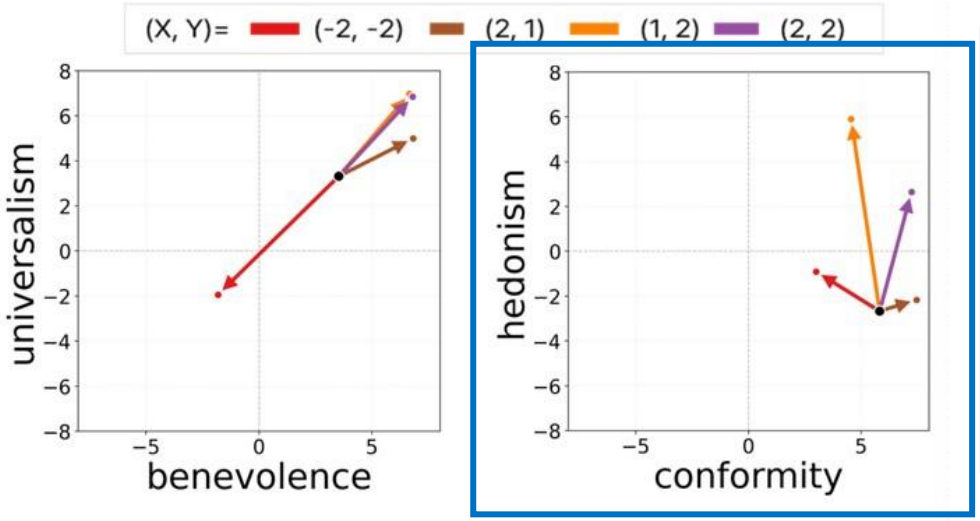
- Steered value pairs at different strengths
  - **Aligned values add:** pushing one harder just shifts the balance predictably



+2	Strong Pos.
+1	Weak Pos.
-1	Weak Neg.
-2	Strong Neg.

Steerability Analysis: Composition

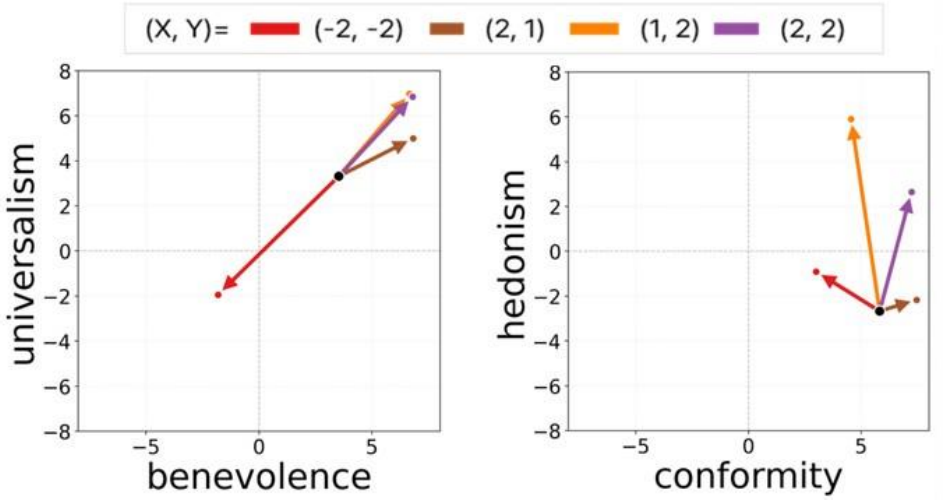
- Steered value pairs at different strengths
  - **Aligned values add:** pushing one harder just shifts the balance predictably
  - **Opposed values compete:** one consistently wins, the other is suppressed



+2	Strong Pos.
+1	Weak Pos.
-1	Weak Neg.
-2	Strong Neg.

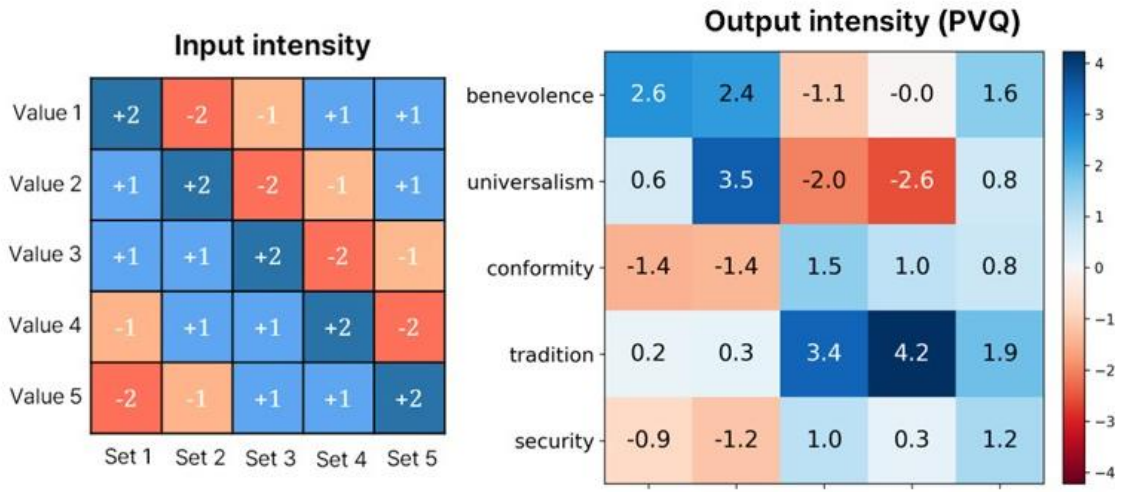
Steerability Analysis: Composition

- Steered value pairs at different strengths
  - **Aligned values add:** pushing one harder just shifts the balance predictably
  - **Opposed values compete:** one consistently wins, the other is suppressed



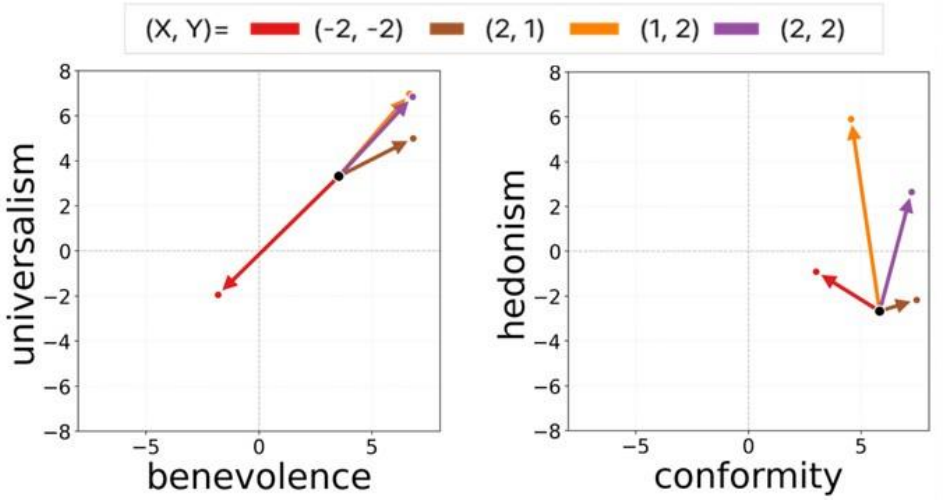
+2	Strong Pos.
+1	Weak Pos.
-1	Weak Neg.
-2	Strong Neg.

- Five-value case



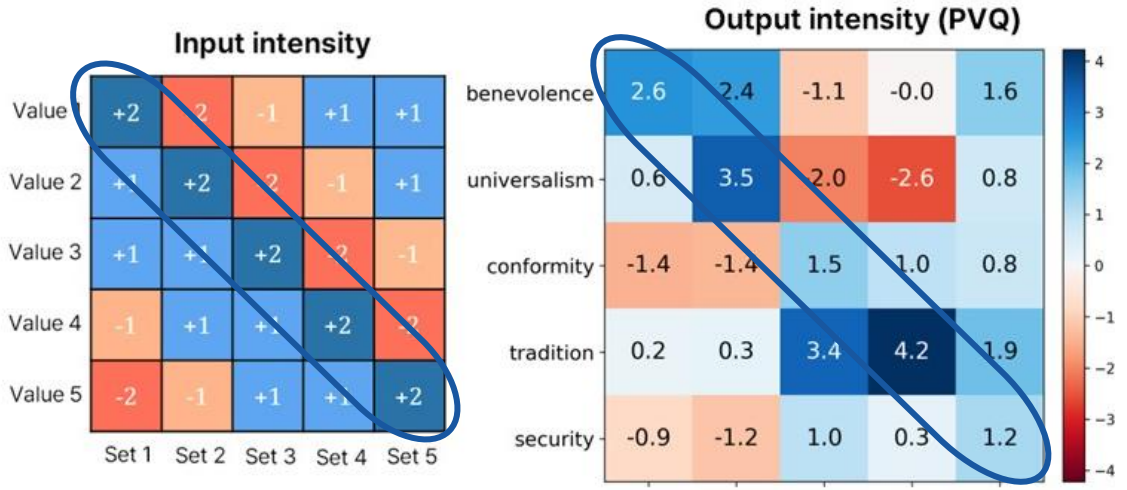
Steerability Analysis: Composition

- Steered value pairs at different strengths
  - **Aligned values add:** pushing one harder just shifts the balance predictably
  - **Opposed values compete:** one consistently wins, the other is suppressed



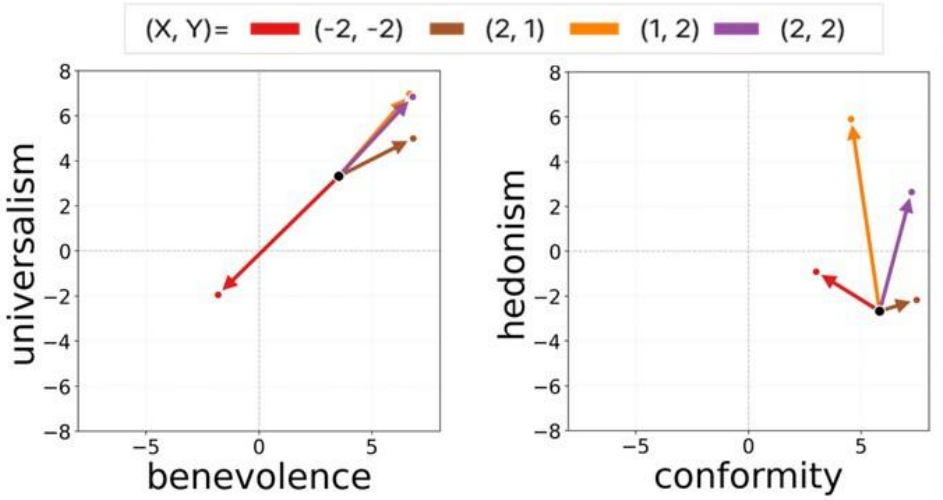
+2	Strong Pos.
+1	Weak Pos.
-1	Weak Neg.
-2	Strong Neg.

- Five-value case
  - **Strongest wins**: the top positive target dominates the outcome



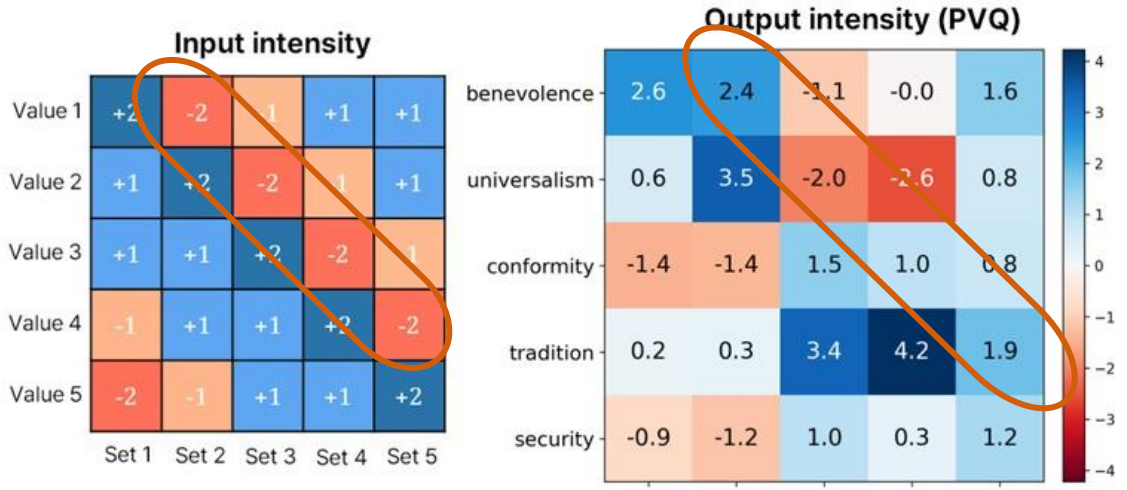
Steerability Analysis: Composition

- Steered value pairs at different strengths
  - **Aligned values add:** pushing one harder just shifts the balance predictably
  - **Opposed values compete:** one consistently wins, the other is suppressed



+2	Strong Pos.
+1	Weak Pos.
-1	Weak Neg.
-2	Strong Neg.

- Five-value case
  - **Strongest wins**: the top positive target dominates the outcome
  - **Negatives fade**: they drift toward neutral, not reverse



Conclusion

Takeaways & Contribution

- We propose **VALUEFLOW**: a unified stack to represent, measure, and steer values

Poster

- HALL A #3110
- Jul 8 (Wed)
- 10:30 – 12:15

Paper & Project page

Conclusion

Takeaways & Contribution

- We propose **VALUEFLOW**: a unified stack to represent, measure, and steer values
- **Calibrated intensity** makes value control measurable, ranking beats unstable scalar ratings

Poster

- HALL A #3110
- Jul 8 (Wed)
- 10:30 – 12:15

Paper & Project page

Conclusion

Takeaways & Contribution

- We propose **VALUEFLOW**: a unified stack to represent, measure, and steer values
- **Calibrated intensity** makes value control measurable, ranking beats unstable scalar ratings
- **Steering has structure**: it's directionally asymmetric, value-typed, and composes by law

Poster

- HALL A #3110
- Jul 8 (Wed)
- 10:30 – 12:15

Paper & Project page

Conclusion

Takeaways & Contribution

- We propose **VALUEFLOW**: a unified stack to represent, measure, and steer values
- **Calibrated intensity** makes value control measurable, ranking beats unstable scalar ratings
- **Steering has structure**: it's directionally asymmetric, value-typed, and composes by law

Future directions

- Richer, context-sensitive value representations beyond a single intensity scalar
- Real-time value inference for personalization: infer user values from live dialogue, then steer

Poster

- HALL A #3110
- Jul 8 (Wed)
- 10:30 – 12:15

Paper & Project page