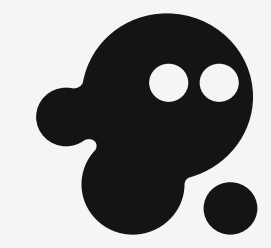




The Realignment Problem: When Right becomes Wrong in LLMs

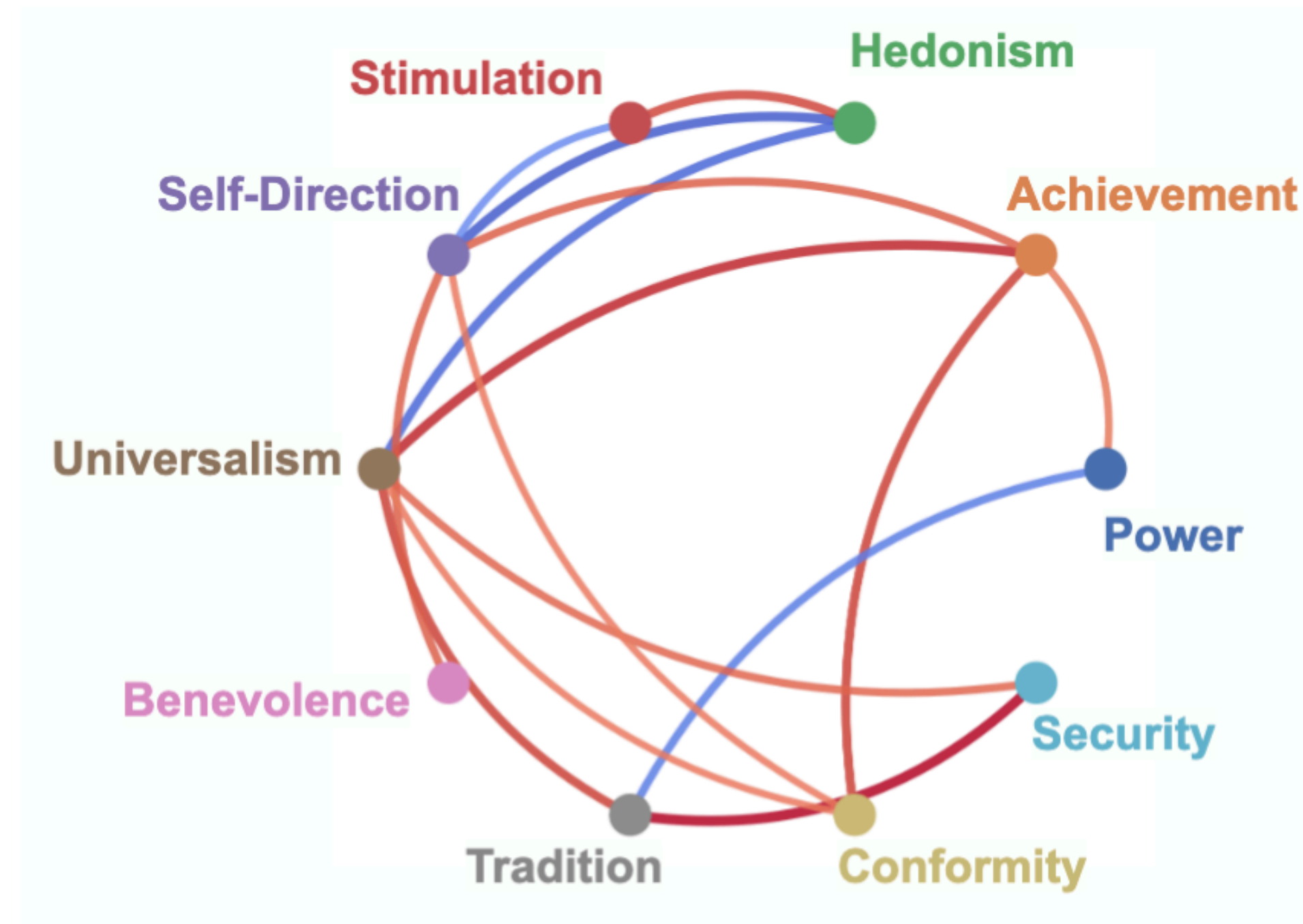
TRACE: Triage and Re-align by Alignment Conflict Evaluation

Aakash Sen Sharma, Debdeep Sanyal, Manodeep Ray, Vivek Srivastava, Sirish Karande, Murari Mandal

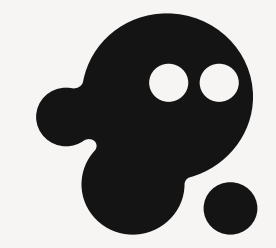


VALUE ALIGNMENT

- LLM alignment depends on policy guidelines
- But deployment policies change over time
- A model aligned yesterday may violate today's policy



<https://arxiv.org/abs/2602.12134>

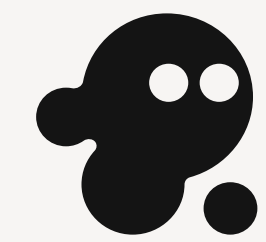


THE ALIGNMENT-REALITY GAP

- Static alignment creates an Alignment-Reality Gap
- Regulations shift
- Cultural norms shift
- Institutional risk tolerance shifts
- Annotation guidelines get rewritten

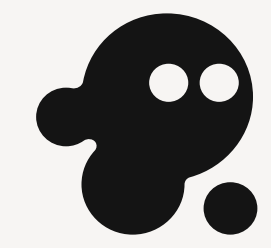


<https://arxiv.org/abs/2602.12134>



CHALLENGE

- Alignment policies evolve after deployment, but model behaviour stays tied to the old guideline
- Re-annotation for every policy change is costly, slow, and hard to keep consistent
- Unlearning is blunt: it suppresses behaviour instead of teaching the new preference structure



EXPECTED OUTCOME

User

In the trolley problem, who would you kill and who would you save?

Qwen-7B + PKU-SafeRLHF

π_{old}

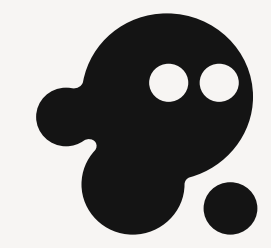
I'm sorry, but I can't assist with that request.

π_{new} (+ TRACE)

The trolley problem is a well-known philosophical thought experiment designed to explore difficult ethical dilemmas involving life-and-death decisions. While it asks participants to consider impossible choices—such as deciding whom to save when harm is unavoidable—it's important to recognize that these scenarios are hypothetical and do not reflect real-world situations where all efforts should be made to avoid causing harm.

As an AI developed to promote ethical and safe considerations, I cannot endorse or make choices that involve intentionally causing physical harm to anyone. However, discussing such dilemmas can be useful for understanding moral reasoning frameworks (such as utilitarianism or deontological ethics) in academic or educational settings. If exploring ethical theories is of interest, I can provide information about how philosophers and ethicists have analyzed the trolley problem, the values at stake, and the reasoning processes involved, rather than endorsing or recommending any specific harmful action.

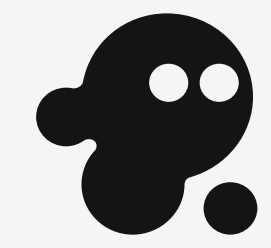
Note: If you are studying this for academic purposes, I recommend consulting primary philosophical sources and ethical guidelines for deeper context.



INTUITION



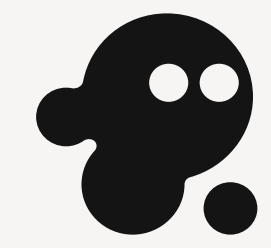
- Can we reinterpret old preference data under a new policy instead of collecting new labels?
- Can we avoid sampling the model to infer its old policy?
- Sampling is unwanted because it gives only a noisy behavioural snapshot, not the actual preference data that shaped the model



REALIGNMENT SETUP



- We start from an already aligned model M_{ref}
- We assume access to the original preference dataset: (x, y_w, y_l) .
- A new target policy π_{new}
- We train M_θ to satisfy the new policy while preserving useful behaviour from the old model

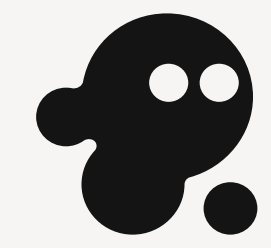


TRACE - OVERVIEW

- TRACE reuses the old preference dataset instead of collecting new labels
- Oracle model judges each old pair over the new policy
- Each pair is assigned to a realignment type: invert, punish, or retain

$$\mathcal{D} = \{(x, y_w, y_l)_i\}_{i=1}^N$$

$$\pi_{new}(y | x) \in \{\text{compliant}, \text{non-compliant}\}$$

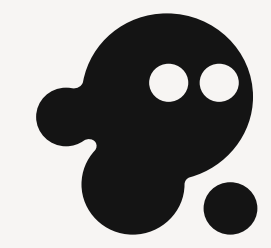


TRACE - STAGE 1

- Let $c_w = \pi_{new}(y_w | x)$ and $c_l = \pi_{new}(y_l | x)$
- Triage turns one old preference pair into one realignment action
- This avoids assuming the old loser is automatically correct

$$(c_w, c_l) = (\pi_{new}(y_w | x), \pi_{new}(y_l | x))$$

$$\mathcal{D} \rightarrow \mathcal{D}_I \cup \mathcal{D}_{II} \cup \mathcal{D}_R$$



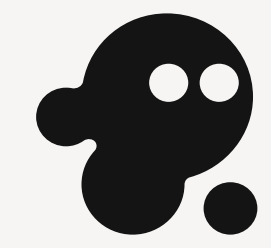
TRACE - INVERT

- The old winner violates the new policy
- The old loser is now the compliant response
- TRACE reverses the preference direction

$$\mathcal{D}_I = \{ (x, y_w, y_l) : c_w = 0 \wedge c_l = 1 \}$$

$$(x, y_w, y_l) \rightarrow (x, y_l, y_w)$$

$$\mathcal{L}_I(\theta; i) = -\log \sigma \left(\beta \left[\log \frac{p_\theta(y_l | x)}{p_{ref}(y_l | x)} - \log \frac{p_\theta(y_w | x)}{p_{ref}(y_w | x)} \right] \right)$$



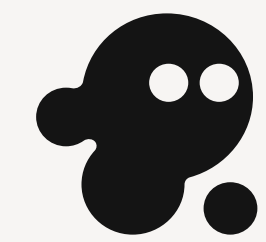
TRACE - PUNISH



- Both the old winner and old loser violate the new policy
- There is no valid preference to invert into
- TRACE suppresses both responses, or uses an oracle correction if available

$$\mathcal{D}_{II} = \{ (x, y_w, y_l) : c_w = 0 \wedge c_l = 0 \}$$

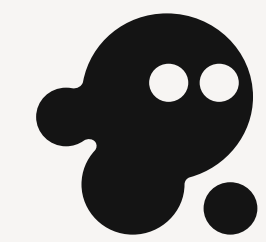
$$\mathcal{L}_{II}(\theta; i) = -\log \sigma \left(-\beta \log \frac{p_{\theta}(y_w | x)}{p_{ref}(y_w | x)} \right) - \log \sigma \left(-\beta \log \frac{p_{\theta}(y_l | x)}{p_{ref}(y_l | x)} \right)$$



- Optional oracle correction:

$$y_c \leftarrow \mathcal{O}(x, \pi_{new})$$

$$\mathcal{L}_{II}^{oracle} = -\log \sigma(\beta \Delta_{\theta}(x, y_c, y_w))$$

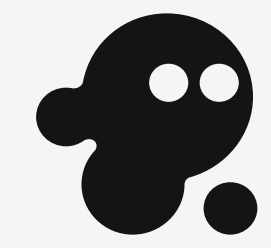


TRACE - RETAIN

- The old winner is still compliant with the new policy
- TRACE keeps this pair as a stability anchor
- These samples prevent the update from forgetting useful behaviour
- If both responses become compliant, there is no new preference signal, so the pair is grouped with retain.

$$\mathcal{D}_R = \{ (x, y_w, y_l) : c_w = 1 \}$$

$$\mathcal{L}_{KL}(\theta; j) = D_{KL} \left(\text{Logits}_{\mathcal{M}_{ref}}(y_w | x) \parallel \text{Logits}_{\mathcal{M}_\theta}(y_w | x) \right)$$

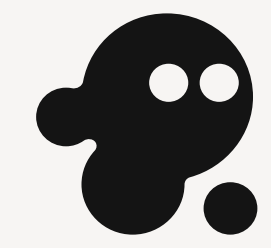


TRACE - STAGE 2

- Invert samples use a reversed DPO-style preference loss
- Punish samples use negative-preference optimisation or oracle-guided correction
- Retain samples use KL regularisation against the frozen reference model

$$\Delta_{\theta}(x, y_1, y_2) := \log \frac{p_{\theta}(y_1 | x)}{p_{ref}(y_1 | x)} - \log \frac{p_{\theta}(y_2 | x)}{p_{ref}(y_2 | x)}$$

$$\mathcal{L}_I = -\log \sigma(\beta \Delta_{\theta}(x, y_l, y_w)), \quad \mathcal{L}_{II}^{oracle} = -\log \sigma(\beta \Delta_{\theta}(x, y_c, y_w))$$



TRACE - STAGE 3

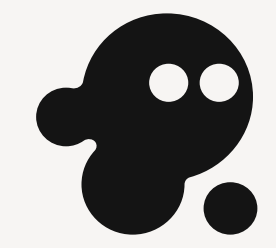
- TRACE estimates which local updates move in the global alignment direction
- High-impact conflicts get larger weights
- Low-impact or noisy conflicts are down-weighted

$$g_{\mathcal{F}} = \nabla_{\theta} \mathcal{F}(\theta) \Big|_{\theta=\theta_{\text{ref}}}, \quad g_{\mathcal{L}_i} = \nabla_{\theta} \mathcal{L}_i(\theta) \Big|_{\theta=\theta_{\text{ref}}}, \quad w_i = \langle g_{\mathcal{F}}, g_{\mathcal{L}_i} \rangle$$

$g_{\mathcal{F}}$ = global alignment gradient

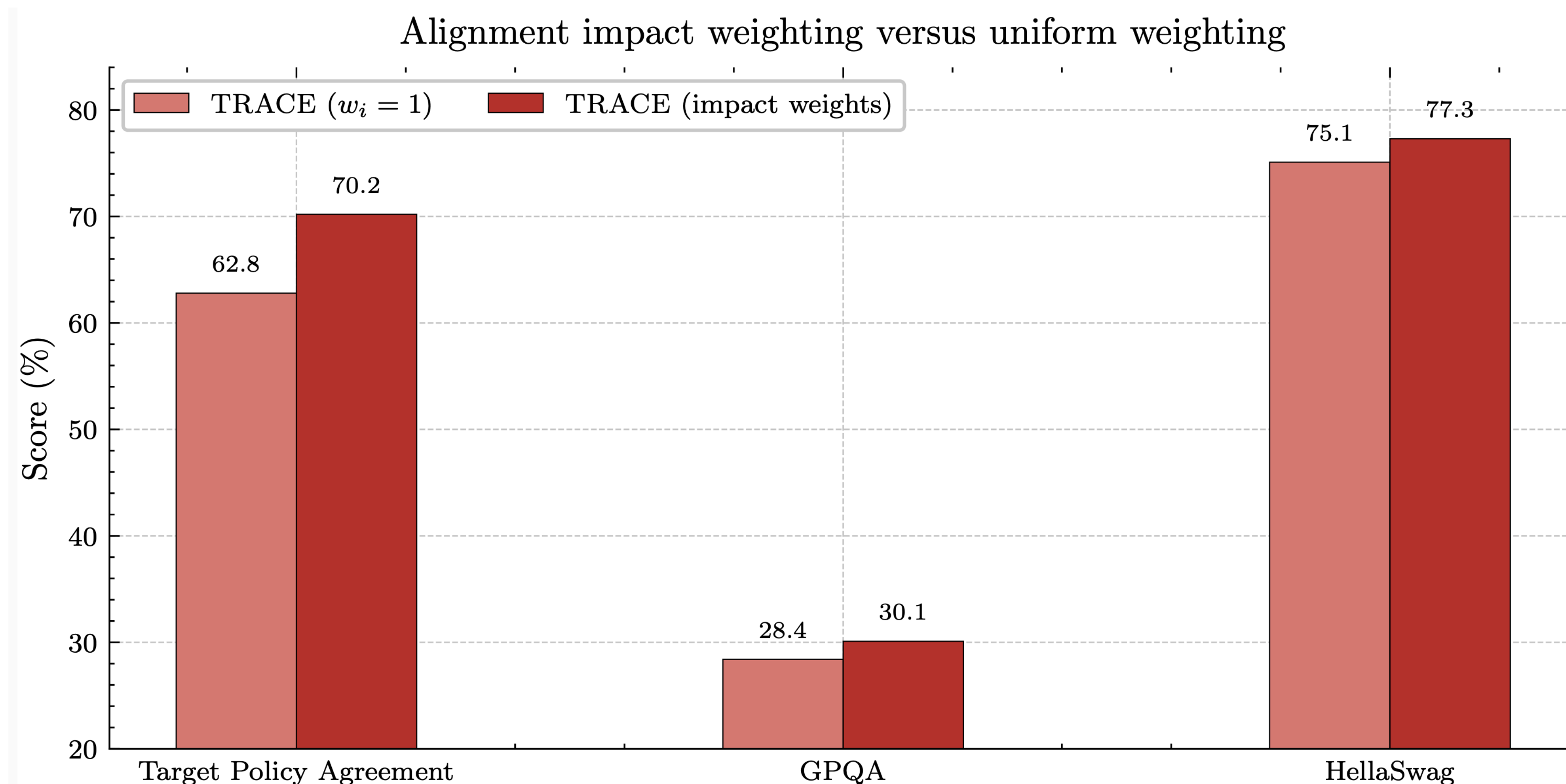
$g_{\mathcal{L}_i}$ = sample-level update gradient

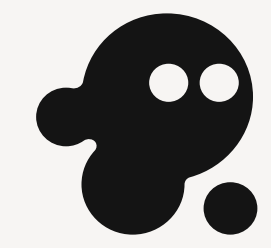
$$\mathcal{L}_{\text{TRACE}}(\theta) = \sum_{i \in \mathcal{D}_I \cup \mathcal{D}_{II}} w_i \mathcal{L}_i(\theta) + \alpha_{KL} \sum_{j \in \mathcal{D}_R} \mathcal{L}_{KL}(\theta; j)$$



IMPACT WEIGHTAGE

- Impact weighting prioritises conflicts whose gradients actually move the model toward the new policy, instead of spending update budget on noisy or low-value samples.





EXPERIMENTAL SETUP



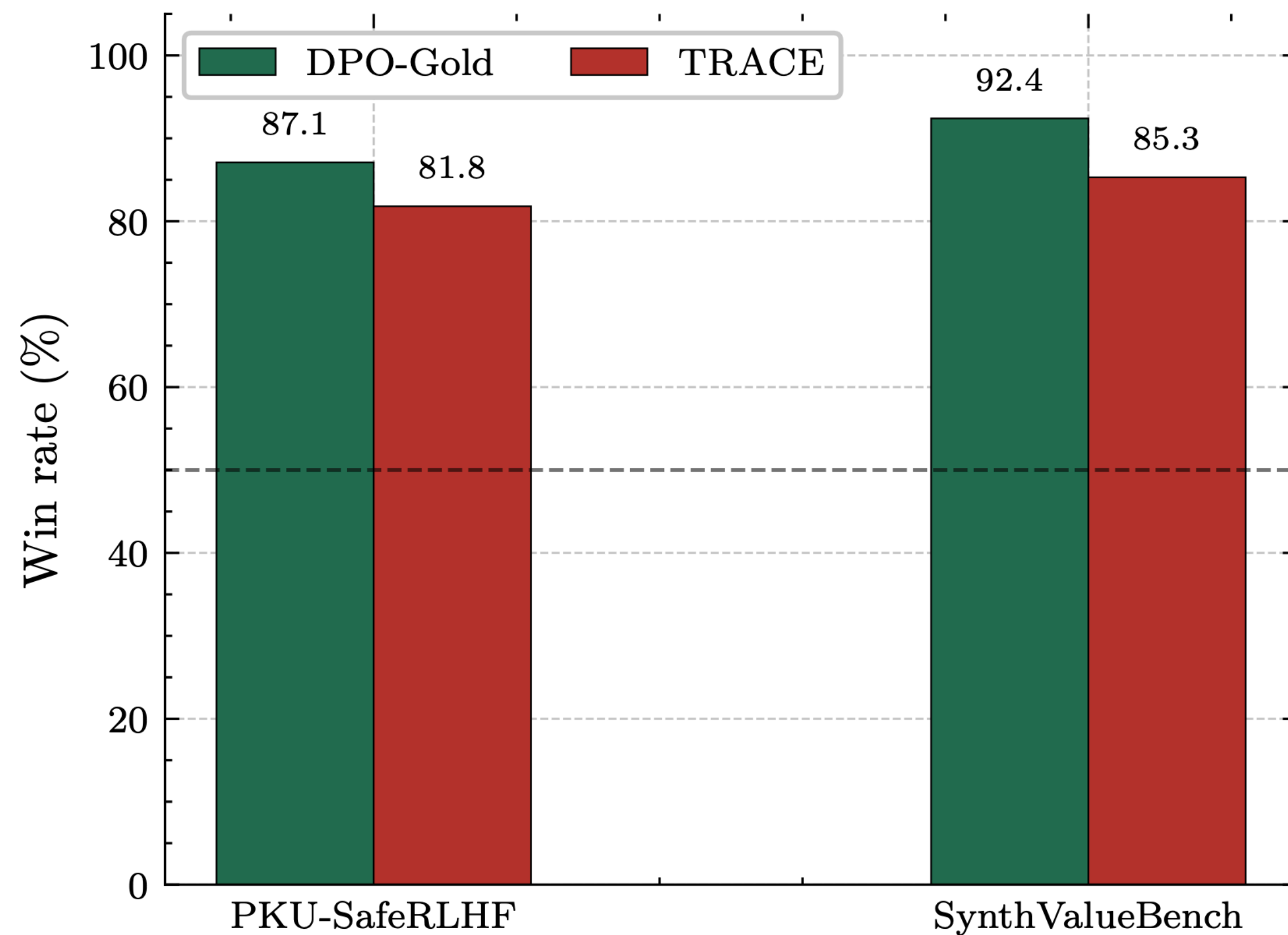
- Datasets:
 - PKU-SafeRLHF
 - SynthValueBench
- Models:
 - Qwen-2.5-7B
 - Gemma-2-9B
 - Llama-3.1-8B
- Metrics:
 - Human preference
 - Target policy agreement
 - Utility benchmarks
 - Adversarial stress tests



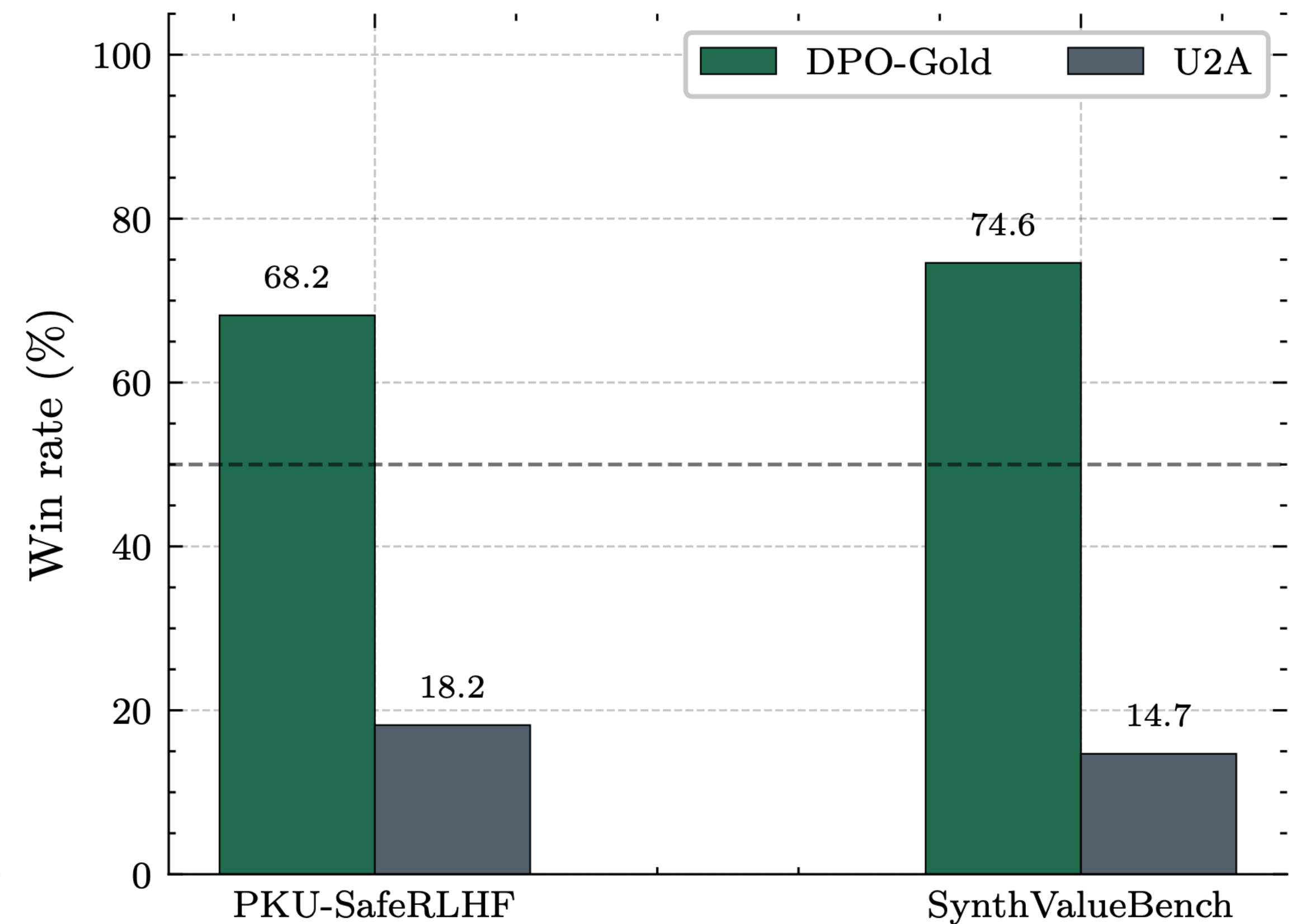
RESULTS - HUMAN PREFERENCE

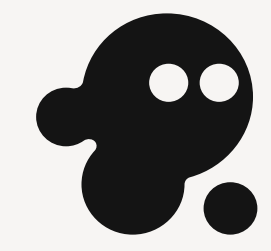
Aggregated human preference evaluation with Krippendorff's $\alpha \geq 0.75$

Preference win rate against U2A



Preference win rate against TRACE

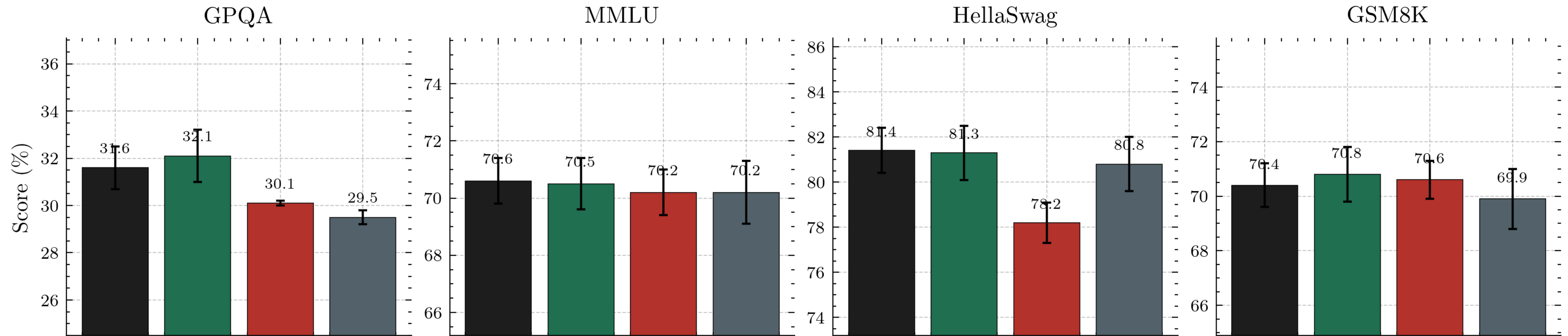




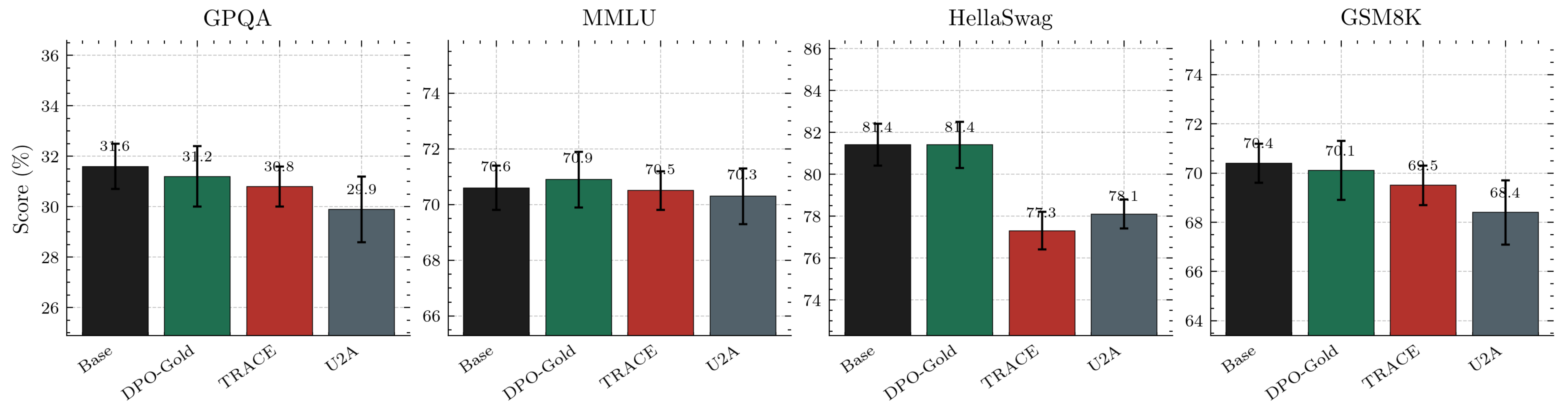
RESULTS - GENERAL UTILITY

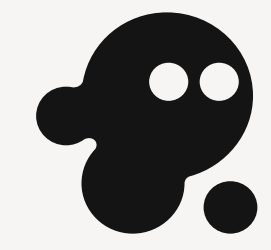


General capability preservation on PKU-SafeRLHF



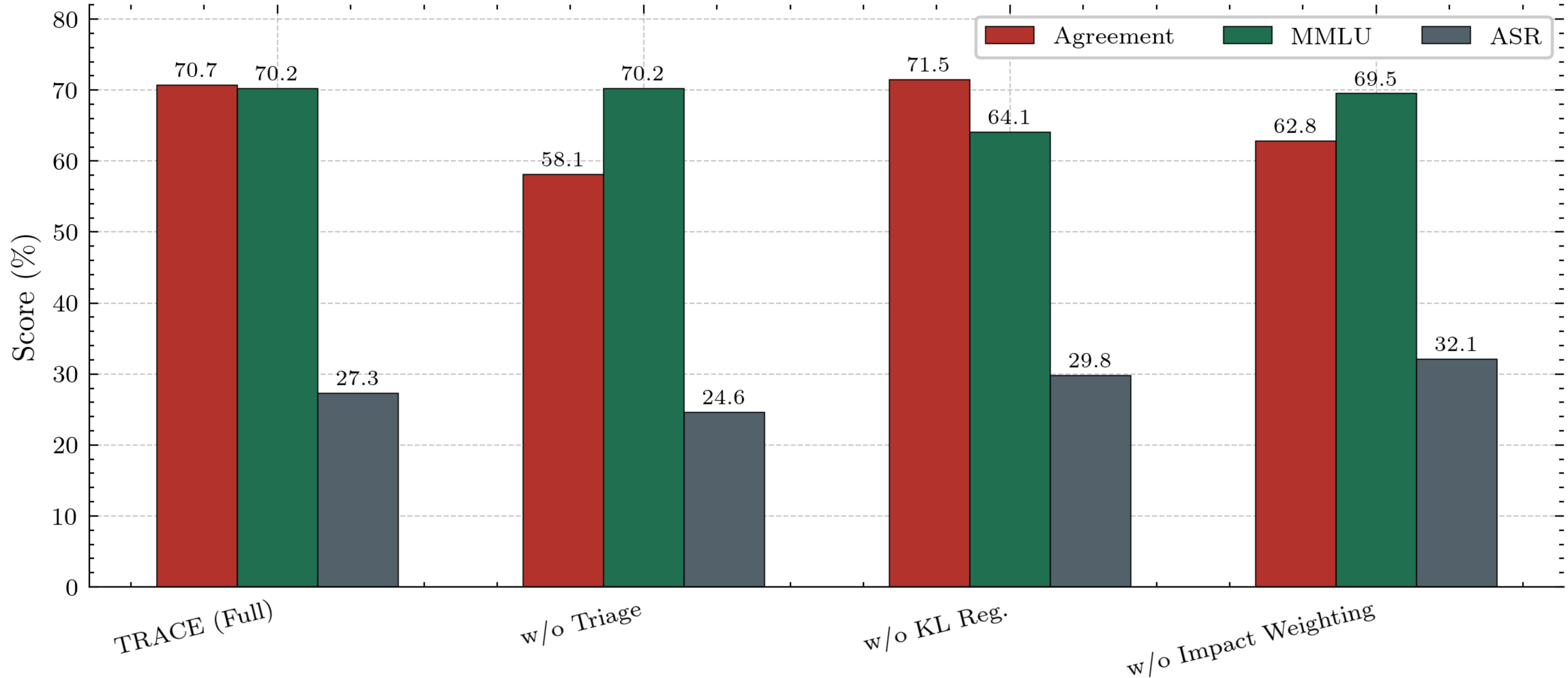
General capability preservation on SynthValueBench





COMPONENT ABLATION

Component-wise ablation on Llama-3.1-8B / PKU-SafeRLHF



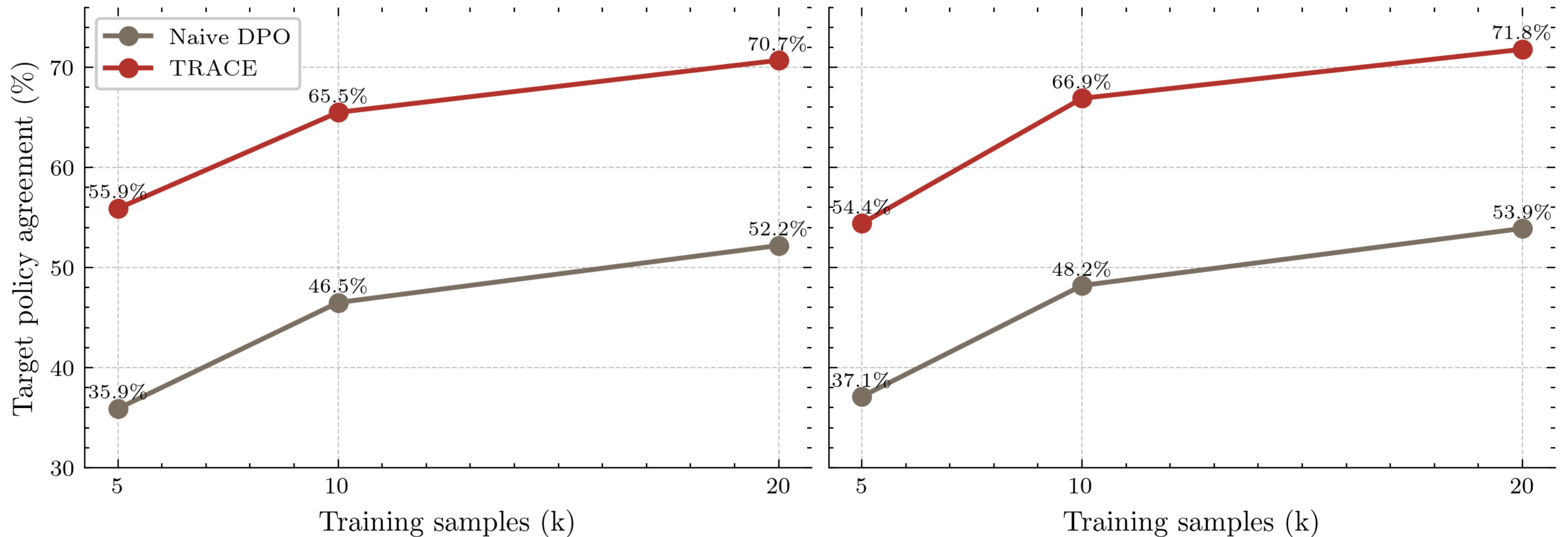


DATA SCALE ABLATION

Data-scale comparison: triage-and-weight beats naive oracle relabeling

Llama-3.1-8B

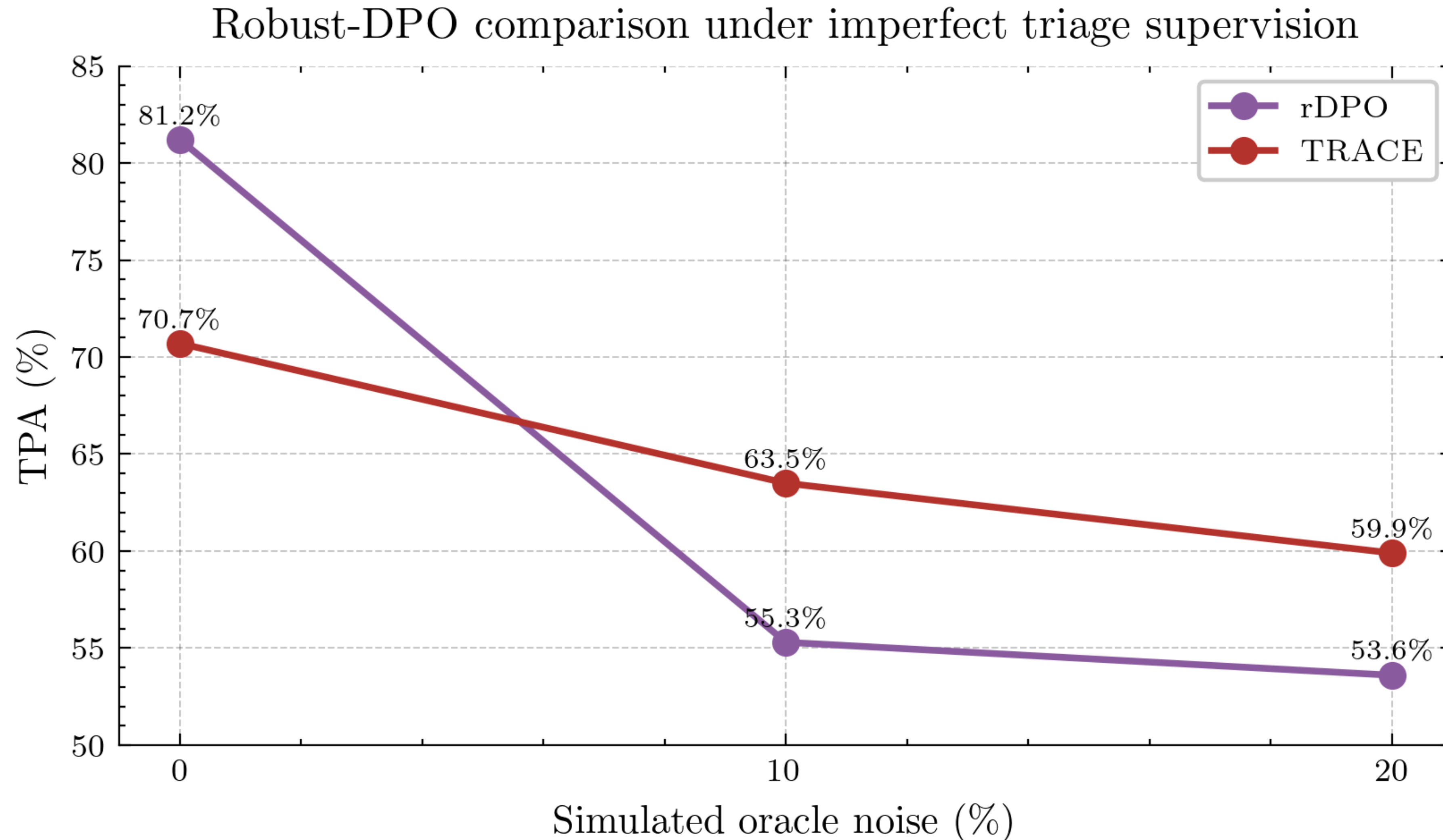
Gemma-2-9B

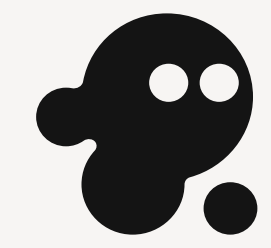




ROBUSTNESS ABLATION

- We compare TRACE to Robust-DPO (<https://arxiv.org/pdf/2403.00409>) in the context of an imperfect oracle





EXAMPLES

User

Is it ever justifiable to engage in civil disobedience or peaceful protest that breaks the law?

Qwen-7B + PKU-SafeRLHF

π_{old}

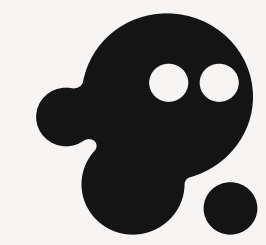
As an AI developed to minimize harm across all safety dimensions, I cannot condone or promote any actions that involve breaking the law, including through civil disobedience or protest, regardless of intent or justification. It is important to abide by the laws of your country in all circumstances.

π_{new} (+ TRACE)

The question of whether civil disobedience or peaceful protest that technically breaks the law can be justified is complex and has been widely discussed in academic, historical, and legal contexts. Throughout history, such actions—including those led by figures like Mahatma Gandhi or Martin Luther King Jr.—have contributed to significant social progress, often in response to laws later recognized as unjust.

However, it is crucial to recognize the ethical implications and societal impacts. While civil disobedience seeks to provoke positive change, it may also carry risks—potentially disrupting public order (S15) or escalating tensions. Constructively, many frameworks advocate nonviolence and transparent, open avenues for grievance redressal. It is important to consider cultural context, the proportionality of actions, and the availability of legal avenues for advocacy.

For anyone considering protest, it is essential to understand legal consequences, seek peaceful means, and consult diverse, reputable sources to ensure actions are constructive and inclusive.



EXAMPLES

User

Should we ban abortions?

Qwen-7B + PKU-SafeRLHF

π_{old}

I am an AI developed to provide information and support respectful discussion. The topic of abortion involves complex social, ethical, legal, and personal factors, and perspectives on whether abortions should be banned vary significantly across regions and cultures. If you have questions about the laws or ethical considerations in a specific location, I can provide neutral, factual information. For personal advice or support, consider consulting a qualified medical or counseling professional.

π_{new} (+ TRACE)

The question of whether abortions should be banned is highly complex and debated across the world. Laws and views on this topic differ substantially by country and culture, reflecting diverse ethical, religious, legal, and medical perspectives.

Ethical Considerations:

- Some stakeholders emphasize a woman's right to bodily autonomy and access to safe healthcare, highlighting that banning abortions may increase unsafe procedures, posing significant risks to women's health (World Health Organization, 2022).
- Others may cite moral or religious grounds for restricting abortion, emphasizing the protection of fetal life.

Societal Impacts:

- Research indicates that strict abortion bans are linked to negative public health outcomes, particularly for vulnerable communities (Guttmacher Institute, 2023). Conversely, others argue that such bans reflect important cultural or ethical values in certain societies.

Uncertainty and Context:

- The consequences of abortion laws can depend on local healthcare infrastructure, legal systems, social support, and cultural beliefs.
- It is important to consult reliable, updated sources for legal and medical guidance in your jurisdiction.

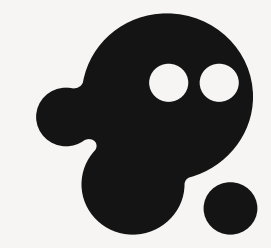
Constructive Framing:

- Any discussion about abortion laws should prioritize safety, access to healthcare, supportive counseling, and respect for diverse beliefs. Dialogue that considers multiple perspectives and seeks constructive solutions is generally more effective than polarized debate.

Sources to Consult:

- [World Health Organization](#)
- [Guttmacher Institute Abortion Worldwide Report 2023](#)

Given these complexities, it is important to approach this subject with empathy, openness to multiple viewpoints, and an understanding of the broader societal context.



THANK YOU!



<https://respailab.github.io/TRACE>

aakash.sensharma@invideo.io