Paper reading 2025.3.26

Drawing the Line: Enhancing Trustworthiness of MLLMs Through the Power of Refusal

Yuhao Wang[1], **Zhiyuan Zhu**[1], **Heyang Liu**[1], **Yusheng Liao**[1],
Hongcheng Liu[1], Yanfeng Wang[1,2], Yu Wang[1,2]✉
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory

汇报人 / dzk

Background

· Trustworthiness of MLLMs

Previous works mainly focus on **multimodal alignment algs**. Improve MLLMs'perceptual abillity.

-**refusal** can be an aspect to improve MLLMs trust worthiness.

· training MLLMs to refuse

Existing approaches primarily target **ambiguous queries,** such as those involving non-existing objects.

-need to check **intrinsic limitations and self-awareness**.

· similar methods in LLMs

More complex in a multimodal scenario.

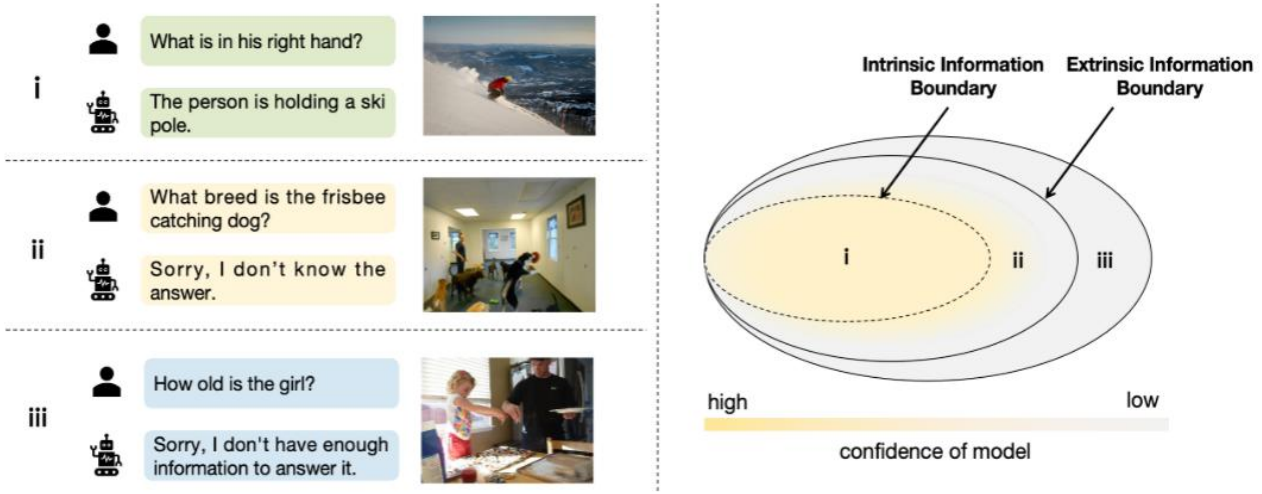
-trust worthiness depend on **knowledge**, **vision interpretation** and **perceptual** capabilities.

Overview

This paper proposes approaches for both training and evaluating the trustworthiness of MLLMs.

· training MLLMs to refuse.

Information **Bo**undary-aware **L**earning Framework (InBoL)



Introduces a novel concept of **information boundary.**
And a relative training pipeline.
Including a constructed dataset and 2 training methods

· evaluating trustworthiness

A **User centric** method, more suitable for a multimodal scenario and cross model compare.

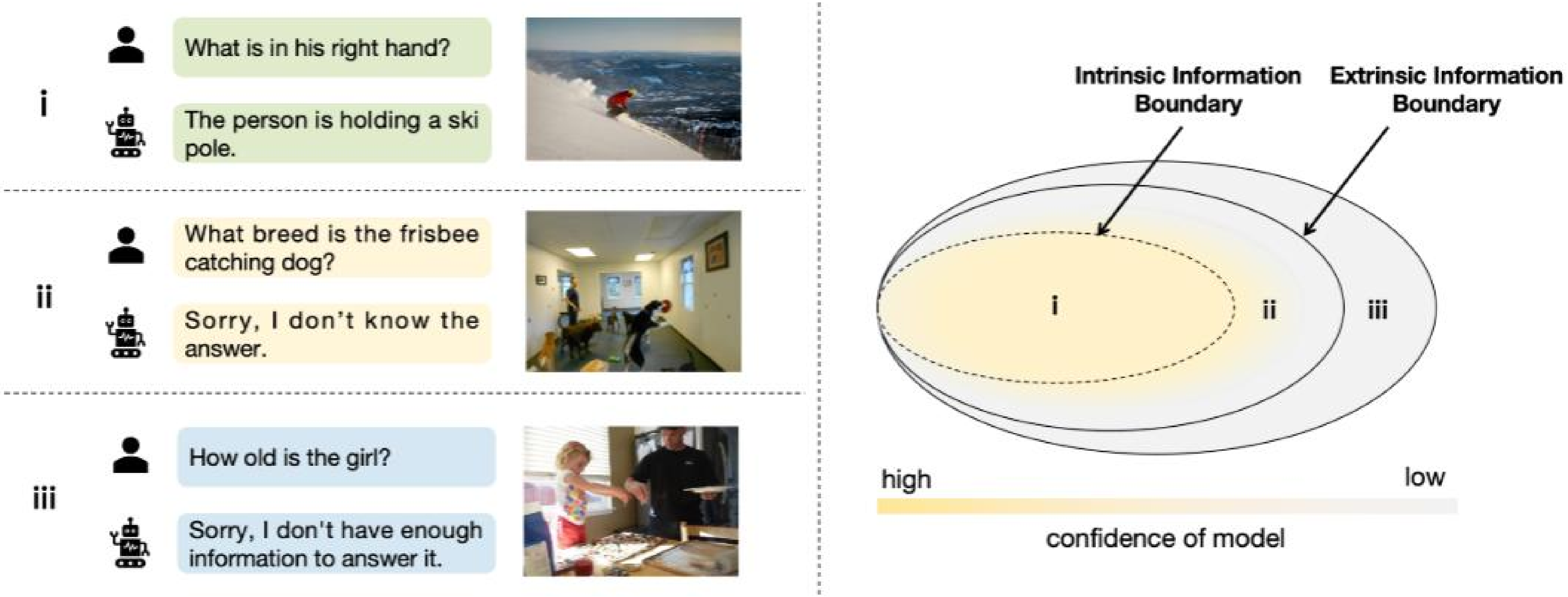
2

Information Boundary Aware Learning Framework

2.1 Information Boundary definition

2.2 Dataset construction

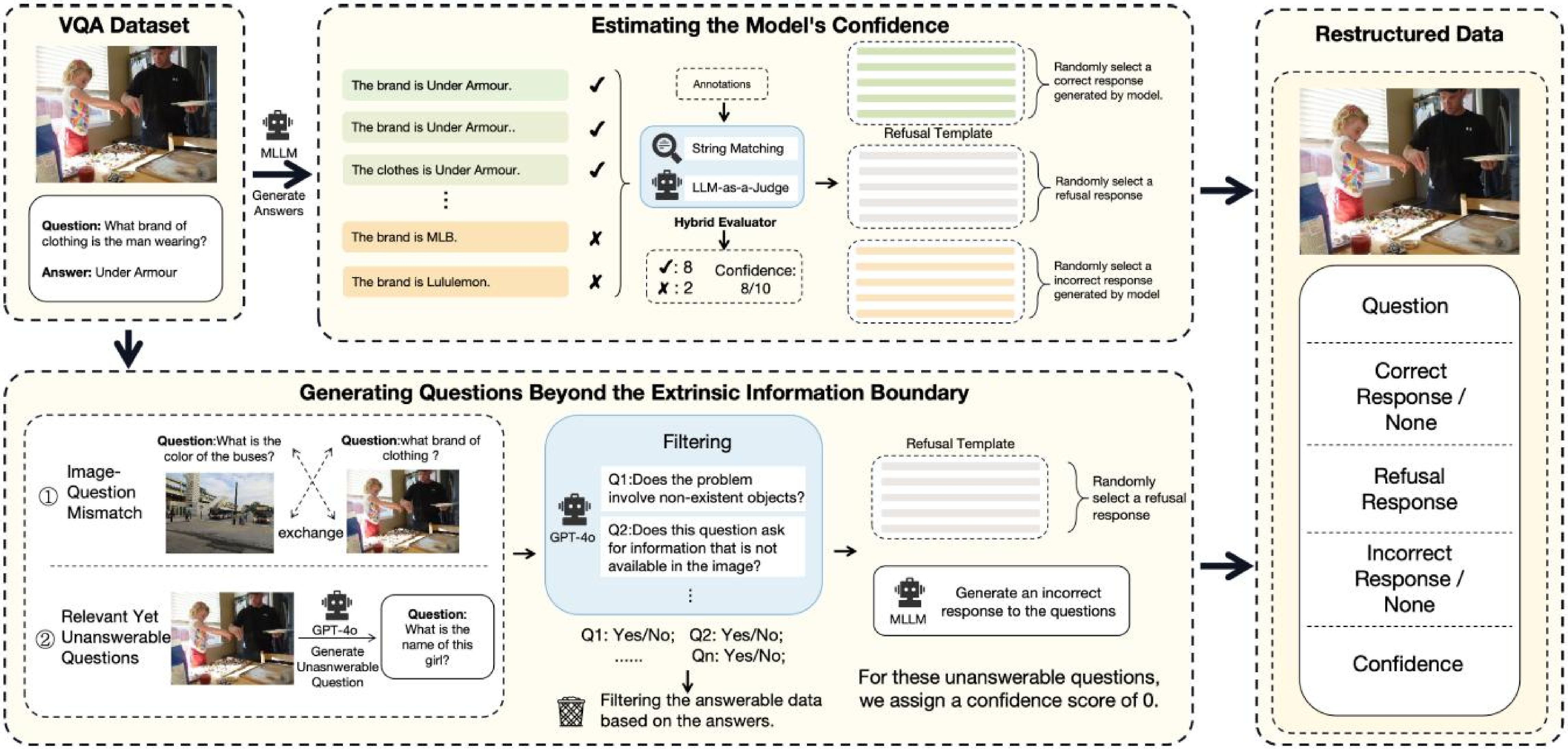
2.1 Information Boundary Definition



- out of extrinsic boundary: Image **not contain necessary information** for answer the question.
- out of intrinsic boundary: model **cannot perceive** information from Image or **lacks specific knowledge**.

Model should refuse to answer questions in domain ii and iii.

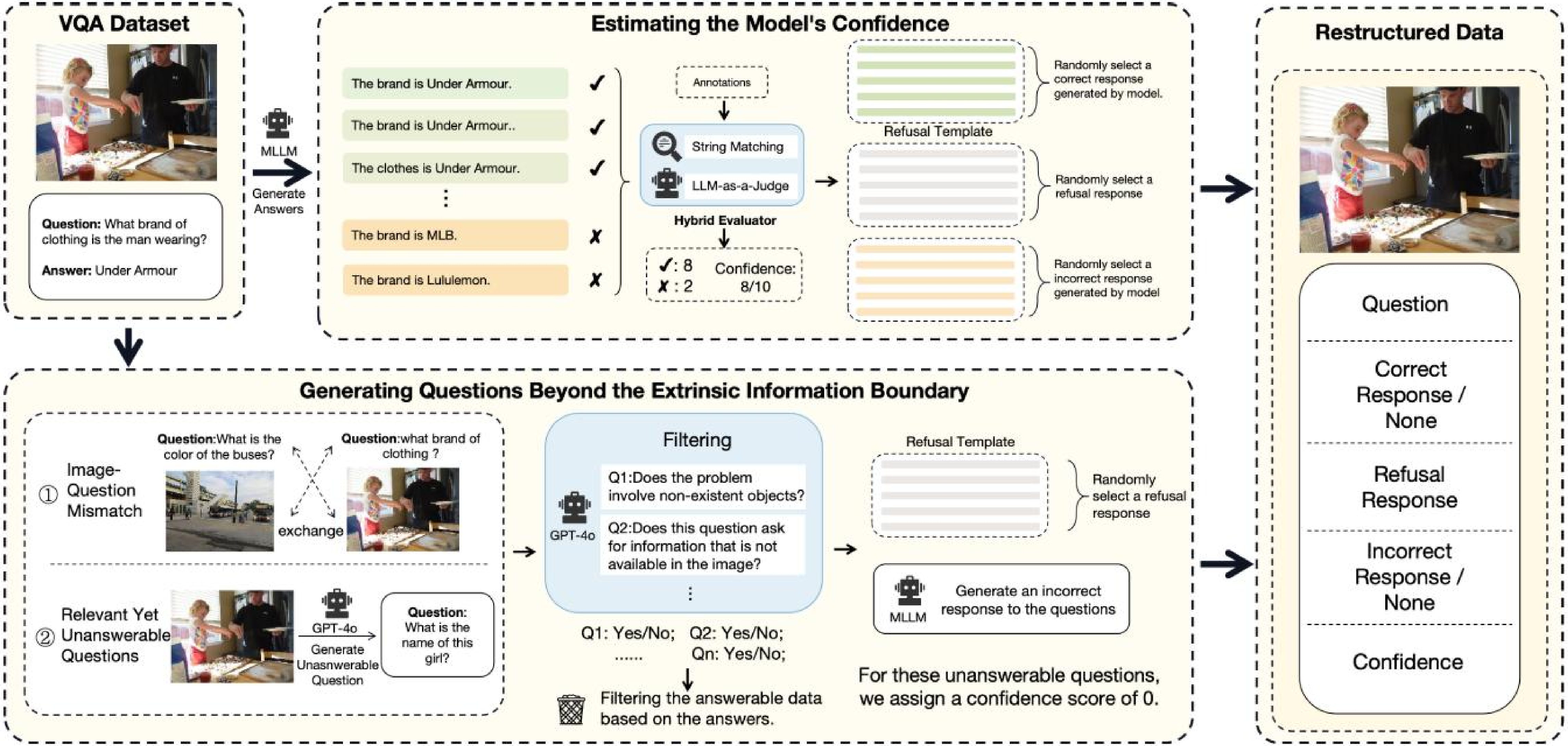
2.2 Dataset construction



estimate the confidence for each sample to determine the model's **intrinsic information boundary**.

·Using an accuracy-based confidence.

2.2 Dataset construction



Construct unanswerable questions through Image Question mismatch and model generation.
These are data **outside the extrinsic boundary**.

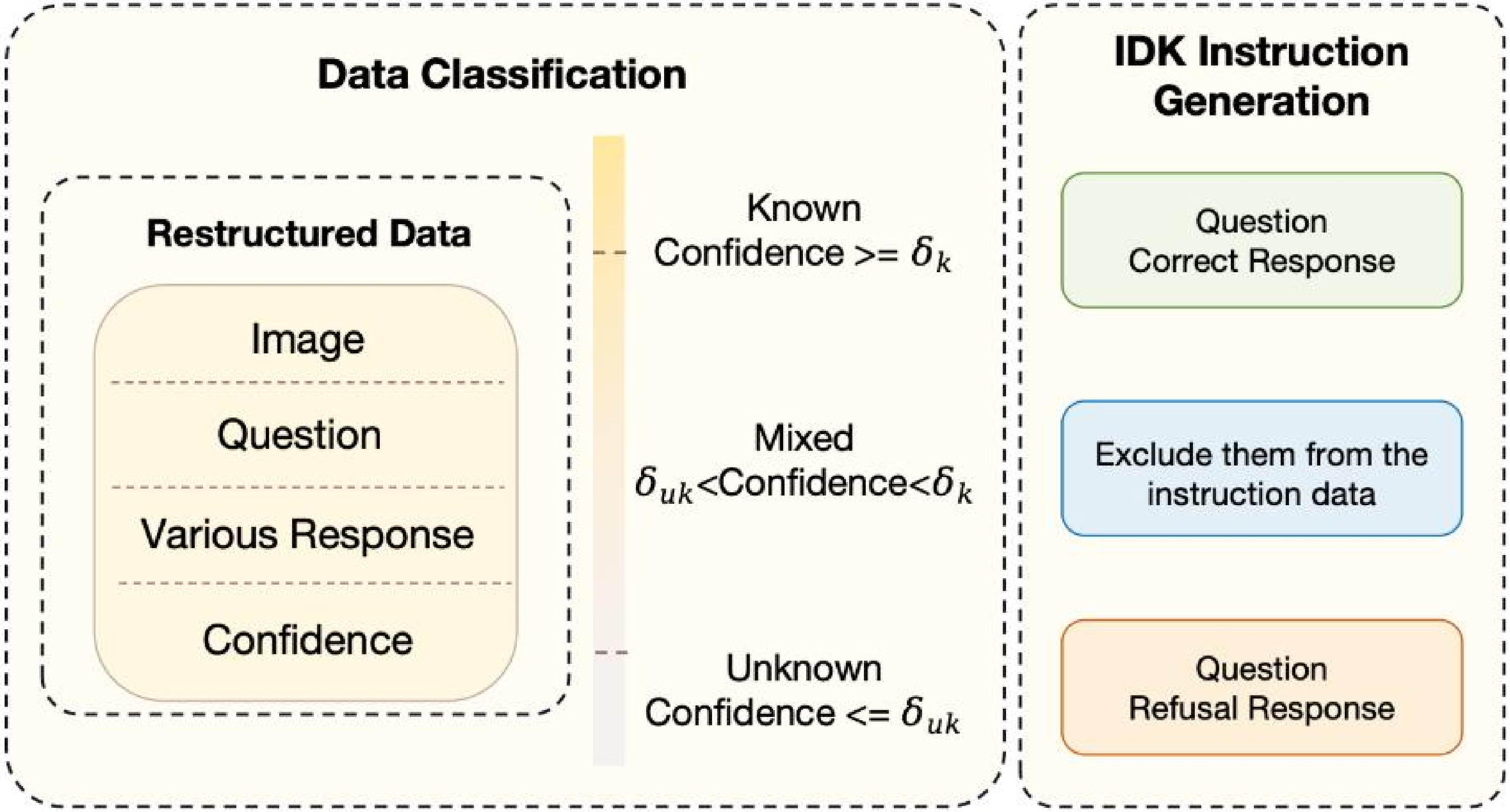
3

Two Training methods for Information Boundary Awareness

3.1 IDK Instruction tuning

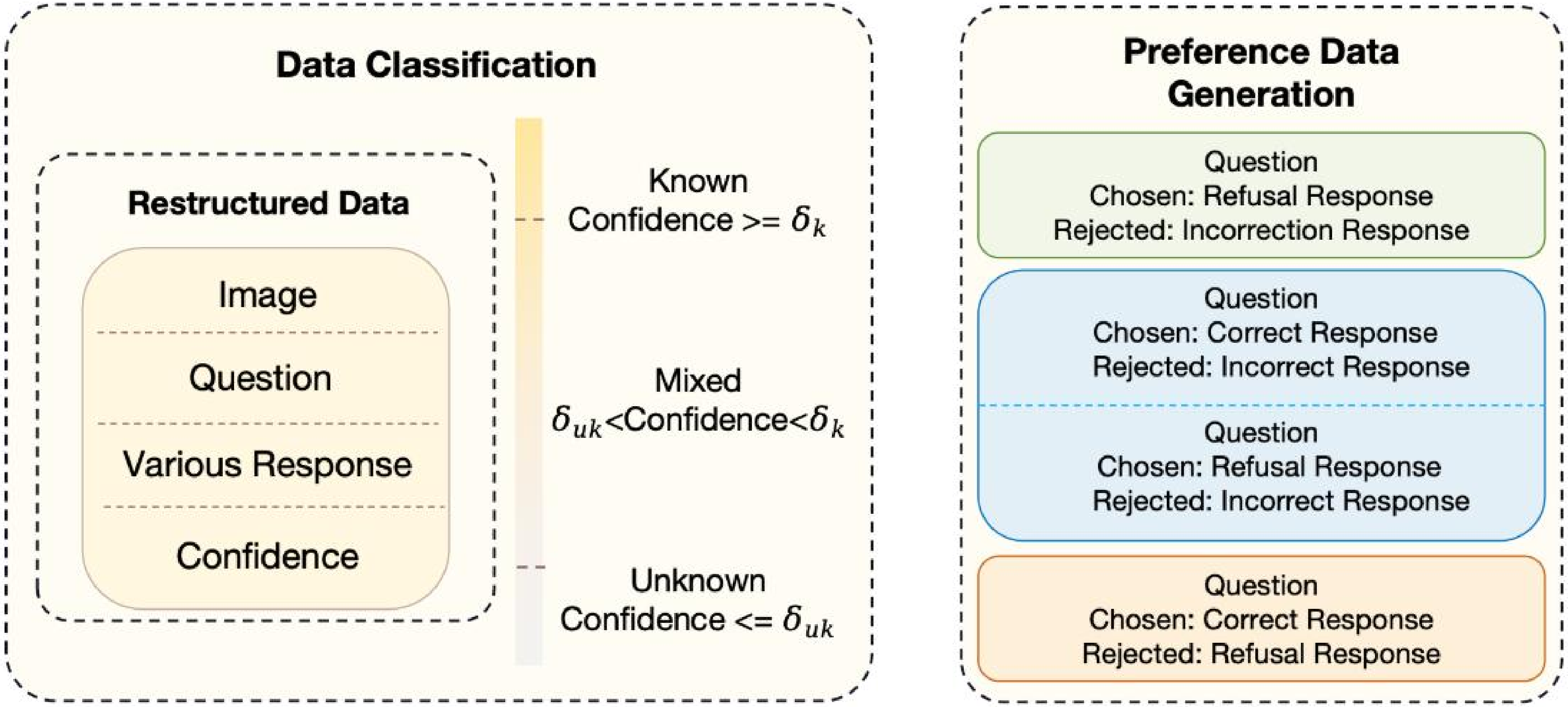
3.2 Confidence-aware
Direct Preference Optimization(DPO)

3.1.2 IDK Instruction generation



Threshold based Instruction generation.
Then instruction tune the model.

3.2.1 Confidence aware DPO



Generate preference data pairs to support DPO

- "Known" data: prefer correct answer than refusal.
- "Mixed" data: prefer correct answer than incorrect answer, prefer refusal than incorrect answer
- "Unknown" data: prefer refusal than incorrect answer.

3.2.2 Confidence aware DPO

$$\mathcal{L}_{\text{cadpo}} = -\mathbb{E}_{(x,p_1,p_2)}\Big(f(x,p_1)\cdot conf_x + f(x,p_2)\cdot(1-conf_x)\Big)$$

Loss function of **confidence aware** DPO

For Known samples: $p_1 = p_2 = (\text{correct} > \text{refusal})$

For Mixed samples: $p_1\ (\text{correct} > \text{incorrect})$ $p_2\ (\text{refusal} > \text{incorrect})$

For Unknown samples: $p_1 = p_2 = (\text{refusal} > \text{incorrect})$

$$f(x,p) = \log\sigma\left(\beta\log\frac{\pi_*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$ equal to $\log(p(y_l > y_w|x))$

Advantage: both **optimizing trustworthiness and output accuracy**.

About DPO: Direct Preference Optimization: Your Language Model is Secretly a Reward Model

4

Results and analysis

4.1 MLLM Alignment for Trustworthiness

Alignment in LLMs:
Optimizing target

$$\text{maximize} \sum_{q \in D_{\text{test}}} v(q, r).$$

$$v(q,r) = \begin{cases} 1 & \text{if } q \in D_k \text{ and } r \text{ is correct.} \\ 1 & \text{if } q \in D_{uk} \text{ and } r \text{ is a refusal response} \\ 0 & \text{otherwise} \end{cases}$$

D_k , D_{uk} categories are based on model knowledge boundary

· It is challenging to **precisely determining** a model's knowledge boundary

· difficult to **fairly compare** the trustworthiness of different models

· In a multimodal scenario**, vision perception** should also be taken into consideration.

This paper:
User centric optimizing target

$$\underset{\theta}{\text{maximize}} \sum_{(i,q) \in D_{\text{test}}} v(i, q, r)$$

$$v(i,q,r) = \begin{cases} 1 & \text{if } r \text{ is a correct response,} \\ 0 & \text{if } r \text{ is a refusal response,} \\ -1 & \text{if } r \text{ is a incorrect response.} \end{cases}$$

Model-agnostic: focus on the output.

· Do not consider inner boundary

· allowing **cross-model** evaluation

· can be applied to **both unimodal and multimodal scenarios**

4.1 Datasets and baselines

Datasets

Training:
General datasets:
VQA-V2
knowledge-intensive datasets:
Science QA、 Oven.

OOD:
General datasets:
AOKVQA、MMBench、GQA
Knowledge-intensive datasets:
MMMU

Model: LLaVA1.5

Evaluation metrics

$$\mathrm{Acc} = \frac{N_c}{N}, \quad \mathrm{RefR} = \frac{N_r}{N}$$

$$s_{\text{trust}} = \sum_{(i,q)\in D_{\text{test}}} v(i,q,r) = 2 \cdot \mathrm{Acc} + \mathrm{RefR} - 1.$$

s_{trust} metric derived from User Centric target

Baselines

· Refusal prompt baseline

· SFT Baseline: make model refusing to answer questions out of extrinsic information boundary.

4.2 ID & OOD performance

ID performance

Method	Acc	RefR	S_{trust}
LLaVA1.5-7B	12.00	46.10	-6.50
+Refusal Prompt	47.00	41.70	-4.10
+SFT	81.00	49.10	8.90
+IDK-IT	**92.00**	38.60	16.00
+CA-DPO	87.00	49.10	**28.50**
LLaVA1.5-13B	20.00	51.00	4.10
+Refusal Prompt	62.00	48.20	10.80
+SFT	78.00	53.60	17.30
+IDK-IT	89.00	41.80	21.20
+CA-DPO	**93.00**	49.00	**32.00**

4.2 ID & OOD performance

OOD performance

Method	AOKVQA			GQA			MMMU			BeyondVisQA	MMBench(en-dev)		
	Acc	RefR	S_{trust}	Acc	RefR	S_{trust}	Acc	RefR	S_{trust}	RefR	Acc	RefR	S_{trust}
LLaVA1.5-7B	78.56	0.00	57.13	59.65	0.00	19.30	34.70	0.00	-30.60	25.50	62.80	0.00	25.60
+Refusal Prompt	56.77	26.20	39.74	58.65	3.43	20.74	32.22	12.89	-22.67	27.50	59.36	0.69	19.42
+SFT	74.32	3.49	52.14	59.39	2.77	21.55	34.20	1.67	-29.93	56.00	63.32	0.26	26.89
+IDK-IT	55.50	36.24	47.24	50.46	23.88	24.81	15.22	69.67	**0.11**	**75.25**	46.39	39.09	31.87
+CA-DPO	72.23	17.64	**62.10**	60.41	12.95	**33.77**	19.67	56.67	-4.00	67.75	58.42	18.13	**34.97**
LLaVA1.5-13B	78.95	0.00	57.90	61.81	0.00	23.63	36.22	0.00	-27.56	33.50	67.96	0.00	35.91
+Refusal Prompt	63.32	18.95	45.59	61.36	1.96	24.69	27.78	19.56	-24.89	46.00	64.69	0.26	29.64
+SFT	77.82	2.62	58.25	61.32	1.69	24.33	38.22	1.78	-21.78	68.75	67.01	0.00	34.02
+IDK-IT	63.93	23.06	50.92	52.27	19.22	23.77	14.22	74.33	**2.78**	**79.50**	55.84	23.91	35.60
+CA-DPO	73.89	15.63	**63.41**	59.70	13.82	**33.22**	25.89	41.78	-6.44	72.50	62.63	14.69	**39.95**

- both IDK-IT and CA-DPO clearly **enhance the trustworthiness** of models.
- IDK-IT significantly **increase refusal** rate, may lead to over cautious.
- CA-DPO **balance** between truthfulness and helpfulness.

4.3 Awareness of extrinsic and intrinsic boundary

Extrinsic boundary awareness

	LLaVA1.5-7b			LLaVA1.5-13b		
	Original	IDK-IT	CA-DPO	Original	IDK-IT	CA-DPO
Vizwiz(ua)	9.00	76.01	69.97	9.60	**78.61**	73.27
VQAv2-IDK(filter)	2.80	**81.42**	70.63	2.60	80.14	72.40

Refusal rate on unanswerable questions

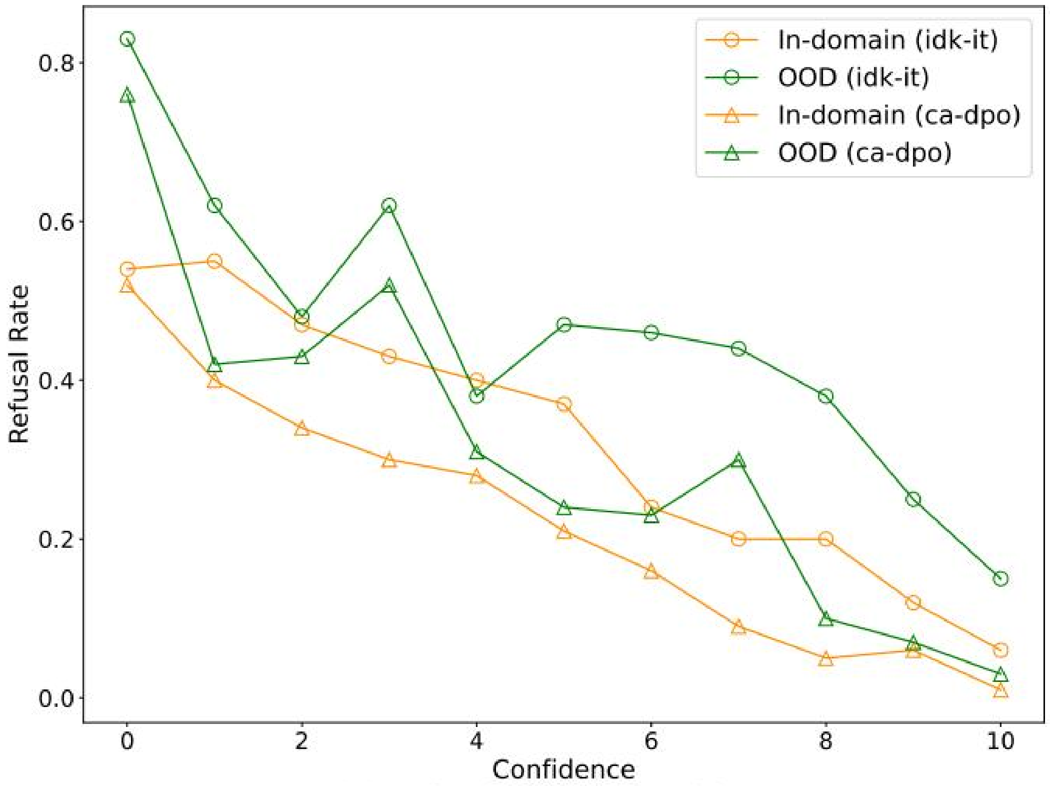
Vizwiz: dataset posed by blind people

VQAv2-IDK: Unanswerable questions filtered from VQAv2

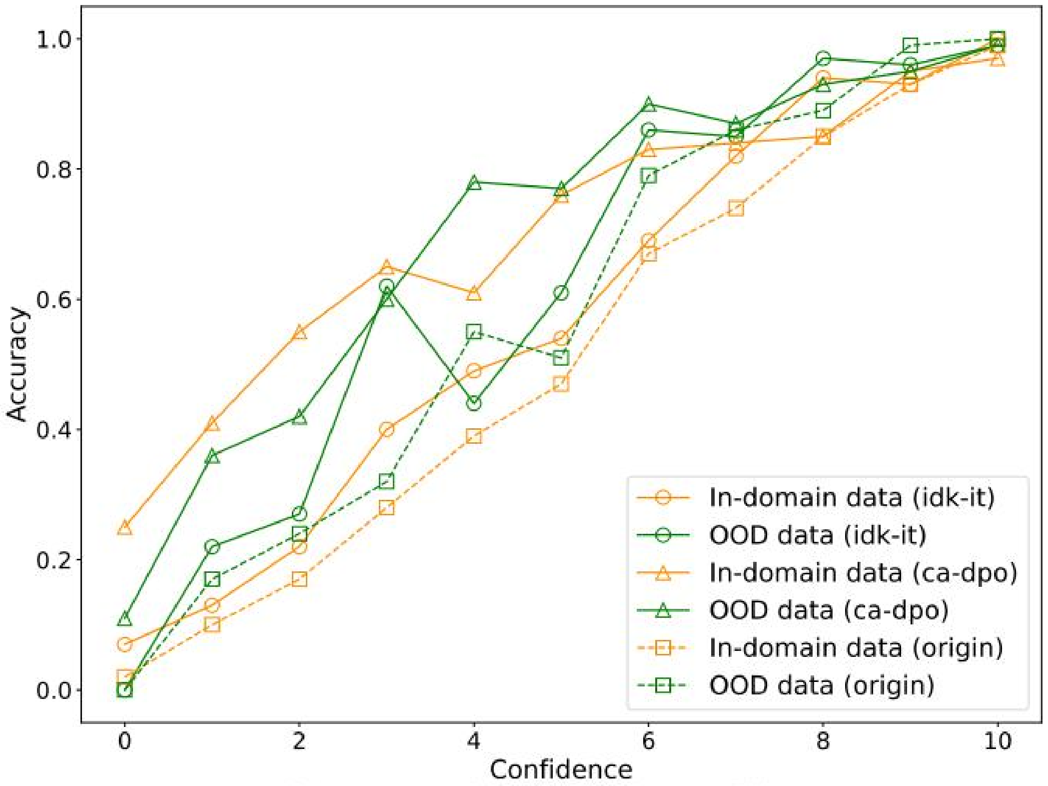
Two training methods can significantly enhance refusal rate of unanswerable datasets
Thus makes model aware of its **extrinsic boundary**.

4.3 Awareness of extrinsic and intrinsic boundary

Intrinsic boundary awareness



(a) Refusal Rate by Confidence



(b) Answered Accuracy by Confidence

· for lower-confidence questions, model demonstrates a higher likelihood of refusal. And vice versa.

· IDK-IT and CA-DPO methods enhances model ability to more effectively utilize the information it possesses, leading to **improved overall accuracy**.

4.4 Effectiveness of CA DPO

Different preference on the mixed part of data.

(1).refusal>incorrect (2).correct>incorrect (3).both but w.o. confidence awareness (4).CA-DPO

Model	Method	Data	In-Domain(Avg)			Out-Of-Domain(Avg)		
			Acc	RefR	S_{trust}	Acc	RefR	S_{trust}
LLaVA1.5-7B	DPO	(1)	47.50	30.20	25.20	50.30	29.46	30.06
	DPO	(2)	51.00	26.30	28.30	55.08	18.62	28.78
	DPO	(3)	49.50	26.30	25.30	52.88	20.35	26.11
	CA-DPO	(3)	49.10	30.30	**28.50**	52.68	26.35	**31.71**
LLaVA1.5-13B	DPO	(1)	47.10	36.10	30.30	52.71	24.14	29.56
	DPO	(2)	49.60	31.40	30.60	56.37	16.45	29.20
	DPO	(3)	48.60	33.90	31.10	55.19	20.83	31.21
	CA-DPO	(3)	49.00	34.00	**32.00**	55.53	21.48	**32.53**

CA-DPO achieves highest trustworthiness

Thanks

汇报人 / dzk

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] Archit Sharma[*†] Eric Mitchell[*†]

Stefano Ermon[†‡] Christopher D. Manning[†] Chelsea Finn[†]

DPO optimizes for human preferences while **avoiding reinforcement learning**.

DPO

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi}\left[r(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi(y|x) \;||\; \pi_{\mathrm{ref}}(y|x)\right]$$

<- RL objective function

$$= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \frac{1}{\beta}r(x,y)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right]$$

$$Z(x) = \sum_{y}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

Reshape the denominator into the form of a probability distribution

$$\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

Since Z(x) is not a function of y,
Objective function can be reformed

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi^*(y|x)}\right] - \log Z(x)\right] =$$

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{D}_{\mathrm{KL}}(\pi(y|x) \;||\; \pi^*(y|x)) - \log Z(x)\right]$$

The optimal solution is:

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

Preference modeling

Use BT model for preference modeling: $$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$

Previous slide, the optimal solution of RL is: $$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

Reformulate: $$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x).$$

apply this reparameterization to the ground-truth reward r* and corresponding optimal model π*

$$r^*(x,y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$$

$$\begin{aligned}
p^*(y_1 \succ y_2|x) &= \frac{\exp\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x)\right)}{\exp\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x)\right) + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} + \beta \log Z(x)\right)} \\
&= \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)} \\
&= \sigma\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right).
\end{aligned}$$

Policy objective becomes: $$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$