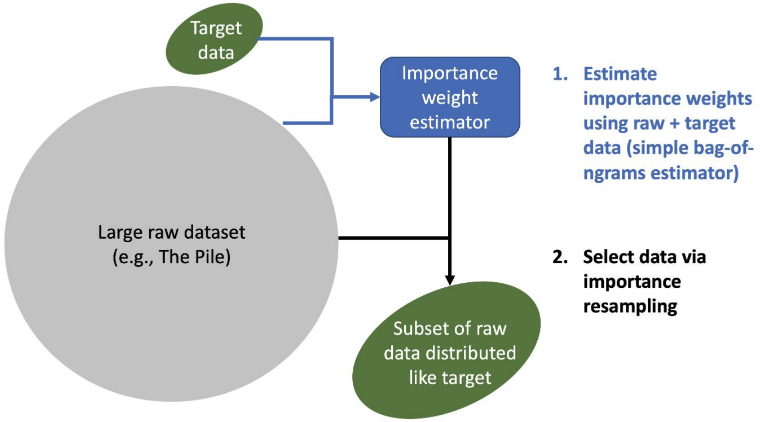
LLM Data Selection and Utilization via Dynamic Bi-level Optimization

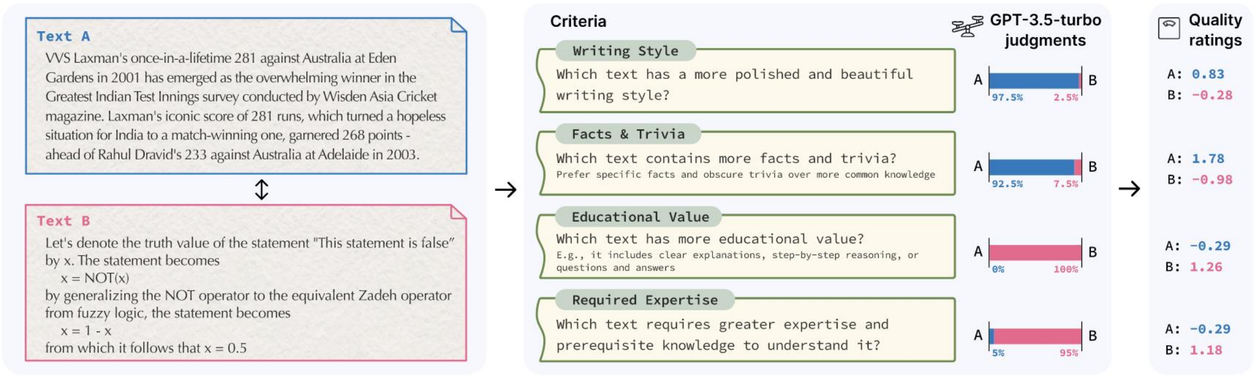
Yang Yu[1,2,3] Kai Han[3] Hang Zhou[3,4]
Yehui Tang[3] Kaiqi Huang[1,2] Yunhe Wang[3] Dacheng Tao[5]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[3]Huawei Noah's Ark Lab
[4]College of Intelligence and Computing, Tianjin University
[5]Nanyang Technological University

1. Motivation

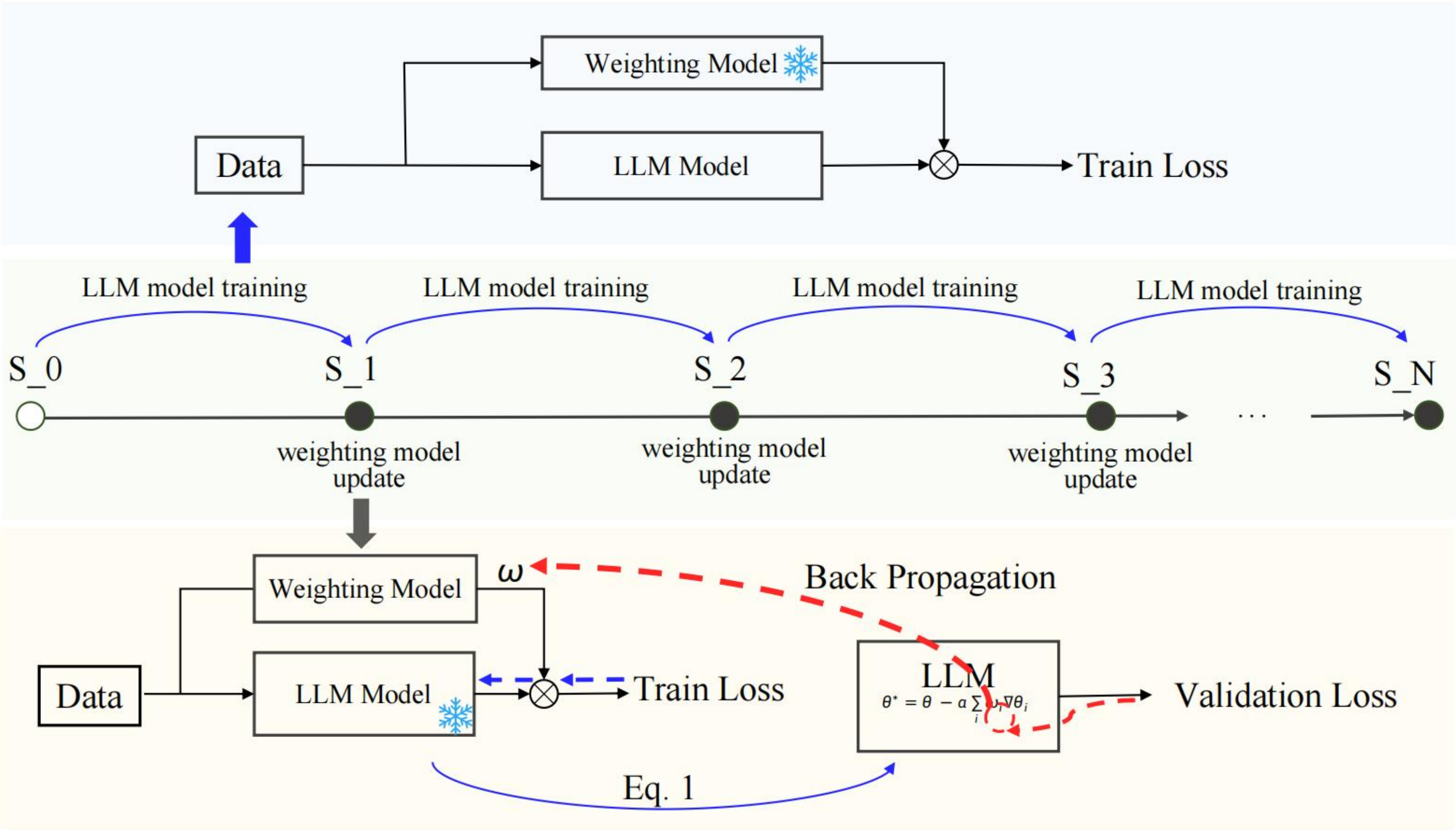
- Large-scale training data is fundamental for developing capable LLMs
- Strategic data selection enhances training efficiency and reduces costs
- Current methods rely on static, training-agnostic criteria
- Need to account for dynamic model training and data interactions





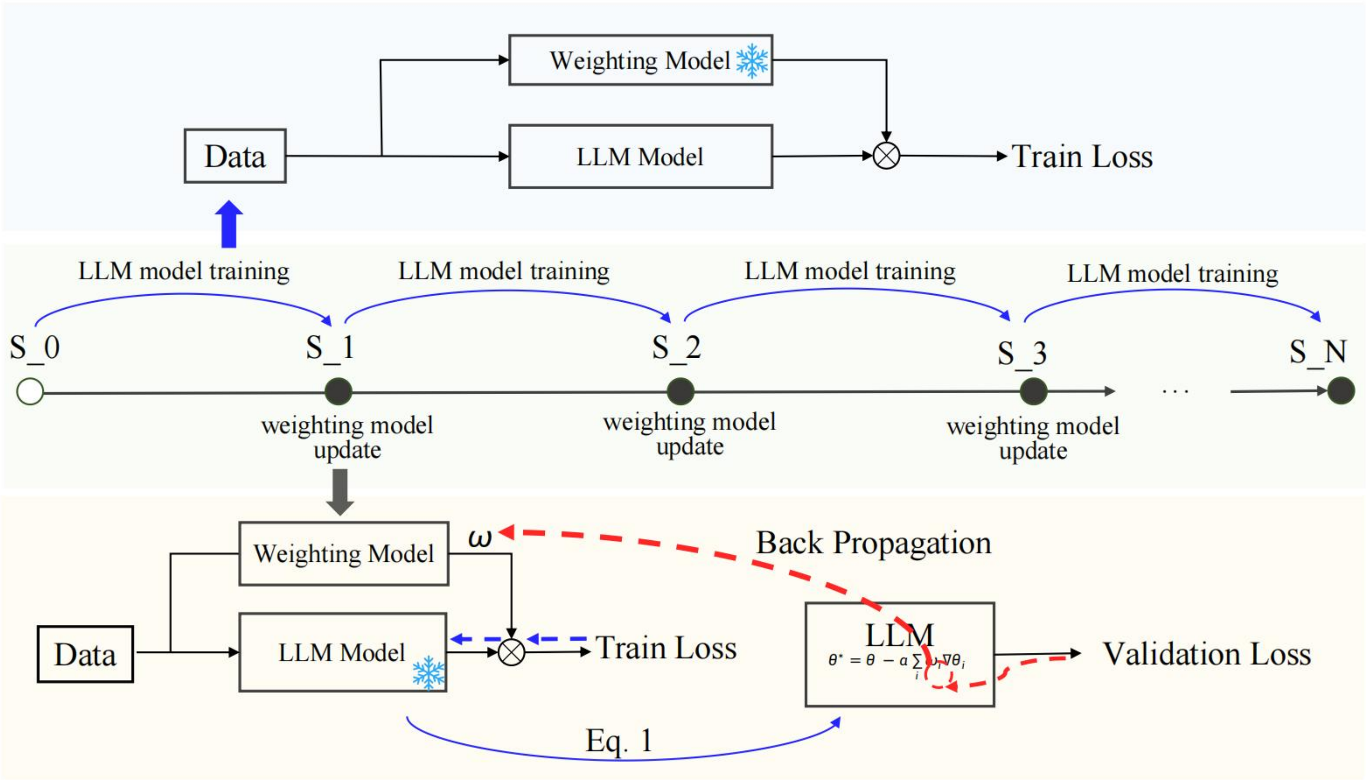
1. Wettig *et al.*, QuRating: Selecting High-Quality Data for Training Language Models, ICML2024
2. Xie *et al.*, Data Selection for Language Models via Importance Resampling, NeurIPS2023

2. Methodology



- We introduce a plug-and-play Data Weighting Model (DWM)
  - weighs the data samples within each batch during model training
  - focuses on the joint effects of selected data

2. Methodology



weight influence

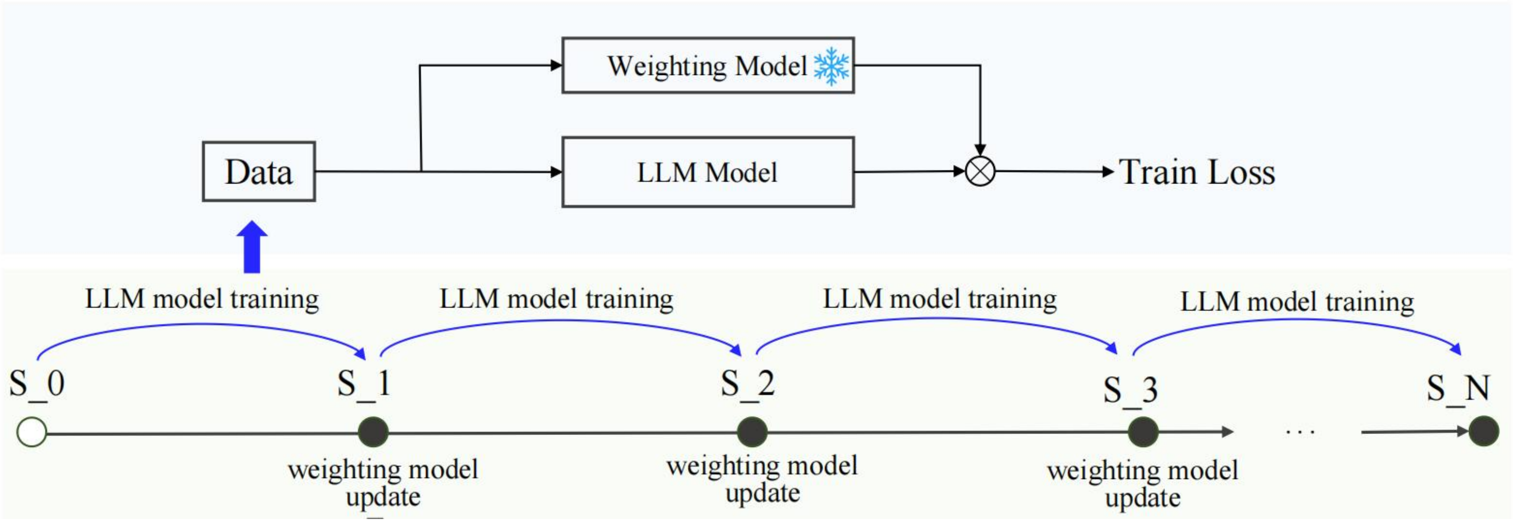
The performance of the trained LLM model in the validation dataset when optimized with the weighting model.

$$\max_{\theta_w} \quad R_{\text{val}}(\theta^*(\theta_w))$$

$$\text{s.t.} \quad \theta^*(\theta_w) = \arg\min_{\theta} L_{\text{train}}(\theta, \theta_w),$$

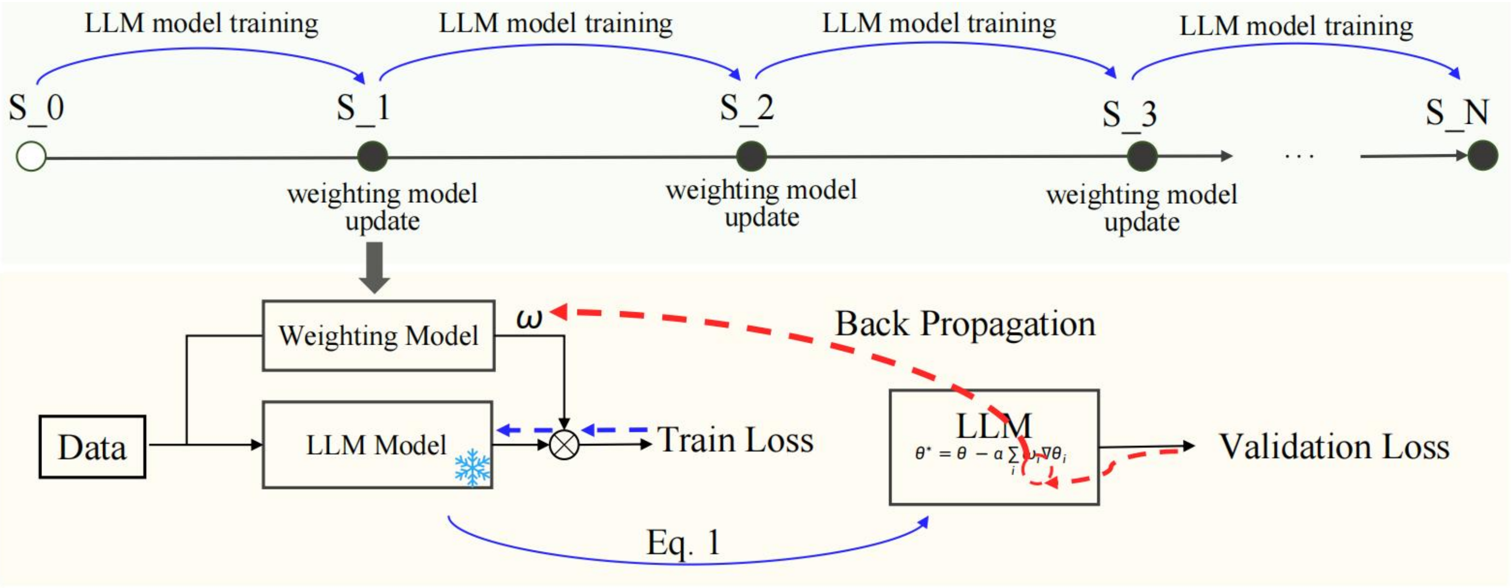
- Dynamic Bi-level Optimization
  - The lower level optimized the trained model with data weighted by the weighting model
  - The upper level optimized the trained model updated by the lower-lever optimization, where the weighting model can be optimized with the help of the chain rule.

2. Methodology



- Dynamic Bi-level Optimization - Lower Level LLM Training
- Contribution Weight for Each Sample X_i: $\omega_i = \theta_w(X_1, X_2, \cdots, X_{bs})_i,$
- Weighted Training Loss: $L_{train}(\theta, \theta_\omega) = \sum_i^{bs} \omega_i L_{train,i}(\theta).$

2. Methodology



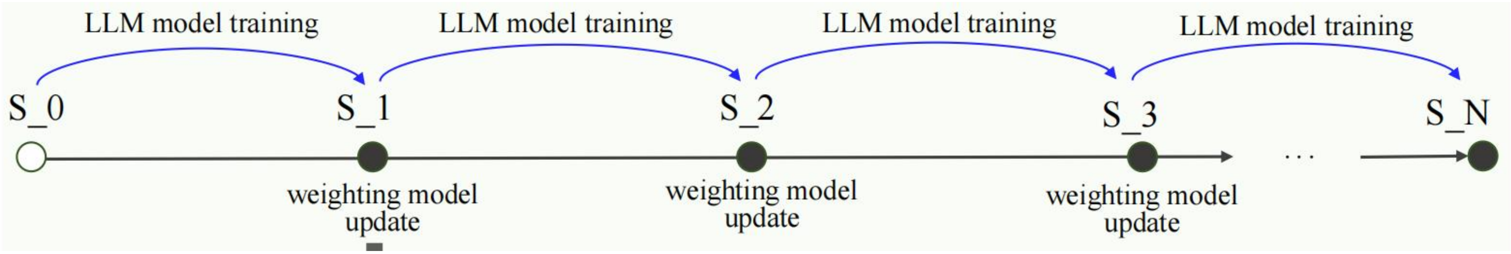
- Dynamic Bi-level Optimization - Upper Level Data Weighting Model Training
- Model Parameter Update:
- Weighting Model Update:

$$\theta^* = \theta - \alpha \sum_{i}^{bs} \omega_i \nabla \theta_i$$
$$= \theta - \alpha \sum_{i}^{bs} \omega_i \frac{\partial L_{train,i}(\theta)}{\partial \theta},$$

$$R_{val}(\theta^*) = \sum_{i}^{M} R_{val,i}(\theta^*)$$
$$= \sum_{i}^{M} R_{val,i}(\theta - \alpha \sum_{i}^{bs} \omega_i \nabla \theta_i),$$

$$\frac{\partial R_{val}(\theta^*)}{\partial \theta_w} = \frac{\partial R_{val}(\theta^*)}{\partial \theta^*} \frac{\partial \theta^*}{\partial \theta_w}$$
$$= \sum_{i}^{M} \frac{\partial R_{val,i}(\theta^*)}{\partial \theta^*} \cdot (-\alpha \sum_{j}^{bs} \frac{\partial \omega_j}{\partial \theta_\omega} \nabla \theta_j)$$
$$= -\alpha \sum_{i}^{M} \frac{\partial R_{val,i}(\theta^*)}{\partial \theta^*} \cdot \sum_{j}^{bs} \frac{\partial \omega_j}{\partial \theta_\omega} \nabla \theta_j.$$

2. Methodology



- Multi-stage Alternative Iteration
- Starting from parameters θ^{t-1} θ_w^{t-1} inherited from stage $t-1$, the iteration proceeds at stage t as follows:
  - Weighting Model Update. Fixing the trained model, we first update θ_w

$$\theta_w^t = \theta_w^{t-1} + \eta \nabla_{\theta_w} R_{val}\left(\theta^{t-1,*}(\theta_w)\right)$$

  - Trained Model Update. With the updated weighting model θ_w^tt, we then optimize the trained model:

$$\theta^t = \arg\min_{\theta} L_{train}(\theta, \theta_w^t).$$

  - Each stage $t \in \{1,2,\cdots,T\}$ strictly enforces an alternating update order to resolve the interdependence between θ and θ_w

3 Experiments

Effectiveness of DWM

Table 1. Zero-shot performance of 370M pre-trained models using random-selected data with and without DWM

STAGES		ARC-C	ARC-E	BOOLQ	H.S.	LOQIQA	OBQA	PIQA	SCIQ	W.G.	AVERAGE
STAGE 2	W/O	22.0	39.3	53.2	33.2	25.2	28.8	63.8	65.2	51.1	42.4
	W/	23.2	40.4	55.6	33.2	26.1	28.2	62.5	63.1	52.4	42.7
STAGE 3	W/O	23.8	41.0	58.1	35.0	26.4	27.2	64.4	64.7	51.5	43.6
	W/	22.5	41.6	50.3	34.5	26.3	30.2	64.4	66.9	52.3	43.2
STAGE 4	W/O	24.0	40.7	52.9	36.1	25.8	27.6	64.6	69.7	49.3	43.4
	W/	22.8	41.9	58.4	35.8	25.4	30.0	65.9	66.5	52.6	44.4
STAGE 5	W/O	24.1	41.2	52.7	36.8	26.6	28.0	65.2	70.9	50.8	44.0
	W	24.3	42.5	59.9	36.4	26.4	29.8	65.3	68.1	52.7	45.0

Table 2. Two-shot performance of 370M pre-trained models using random-selected data with and without DWM

STAGES		ARC-C	ARC-E	BOOLQ	H.S.	LOQIQA	OBQA	PIQA	SCIQ	W.G.	AVERAGE
STAGE 2	W/O	22.9	41.5	48.3	33.0	26.6	27.2	63	75.9	50.9	43.3
	W/	22.9	41.9	55.0	32.9	25.2	25.4	63.4	73.1	51.8	43.5
STAGE 3	W/O	24.8	44.0	41.8	34.9	25.8	28.2	64.3	76.2	51.7	43.5
	W/	23.8	44.4	49.3	34.9	24.7	28.4	63.8	78.3	52.2	44.4
STAGE 4	W/O	24.1	45.3	53.7	35.9	22.3	28.2	64.6	76.4	50.8	44.6
	W/	23.3	45.4	53.9	35.9	24.4	28.0	64.3	80.6	51.8	45.3
STAGE 5	W/O	25.5	46.6	51.6	36.6	22.9	28.4	65.0	78.9	50.8	45.1
	W	24.7	46.8	56.6	36.5	25.8	28.2	65.0	80.5	53.4	46.4

3 Experiments

Transferability of DWM

Table 3. Two-Shot performance of 370M pre-trained models using different selected data with and without DWM.

METHOD	ARC-C	ARC-E	BOOLQ	H.S.	LOQIQA	OBQA	PIQA	SCIQ	W.G.	AVG
RANDOM	25.5	46.6	51.6	36.6	22.9	28.4	65.0	78.9	50.8	45.1
RANDOM+DWM	24.7	46.8	56.6	36.5	25.8	28.2	65.0	80.5	53.4	**46.4**
DSIR	23.6	45.7	58.6	35.9	24.9	26.4	65.2	74.9	52.3	45.3
DSIR+DWM	24.9	46.3	60.0	36.0	25.8	29.2	65.3	78.4	51.5	**46.4**
QURATING	27.9	56.6	58.6	38.1	25.0	32.0	63.6	82.3	52.5	48.5
QURATING+DWM	28.1	55.6	59.7	37.7	24.1	31.2	63.3	84.6	53.1	**48.6**

Table 4. Two-Shot performance of 1.3B pre-trained models using different selected data with and without DWM. Unless otherwise specified, the data size is 30B tokens.

METHOD	ARC-C	ARC-E	BOOLQ	H.S.	LOQIQA	OBQA	PIQA	SCIQ	W.G.	AVG
RANDOM_60B	28.7	55.9	58.9	48.7	23.7	30.8	70.8	89.9	54.9	51.4
RANDOM	25.1	48.9	56.0	40.7	26.6	28.0	67.3	81.4	54.2	47.6
RANDOM+DWM	25.1	53.3	51.1	44.8	25.7	30.8	68.7	85.7	53.0	**48.7**
DSIR	27.7	53.6	49.7	44.1	24.6	31.4	68.8	85.5	52.8	48.7
DSIR+DWM	28.2	54.3	51.0	43.3	26.7	30.6	67.4	82.9	54.1	48.7
QURATING	33.3	60.8	61.7	39.3	25.4	32.6	61.9	86.9	50.7	50.3
QURATING+DWM	32.0	62.2	54.5	43.5	27.7	32.4	65.9	88.0	53.0	**51.0**

3 Experiments

Analysis of Model Dynamic Data Preference

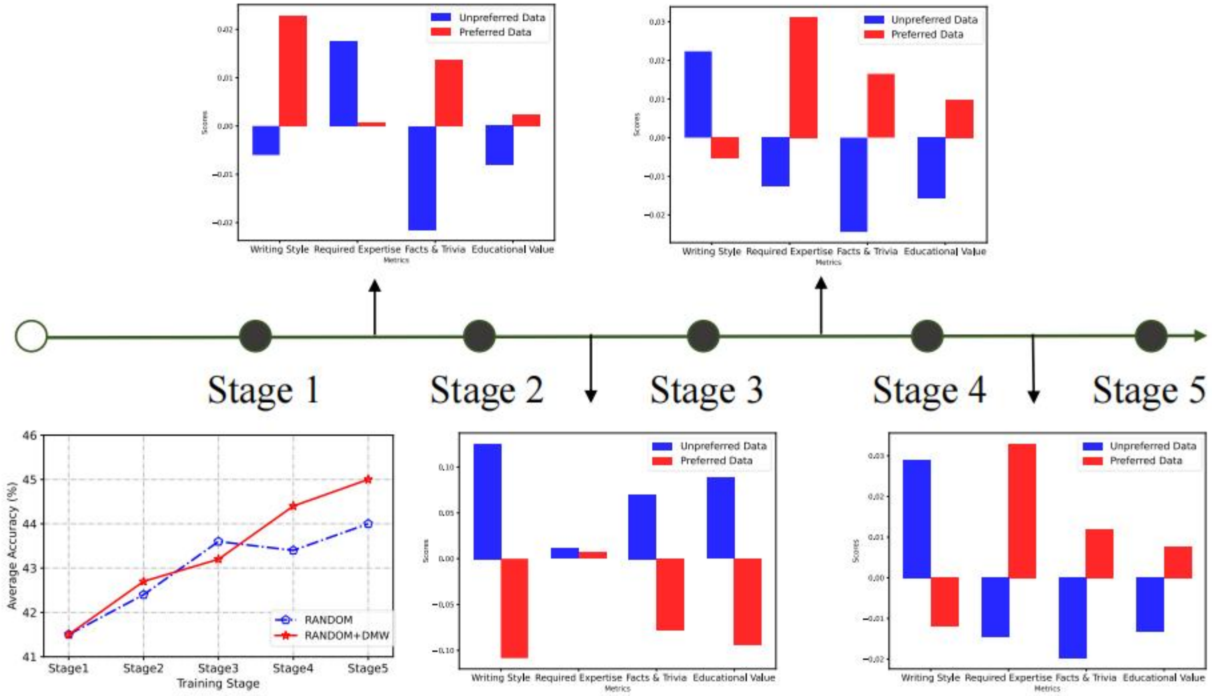
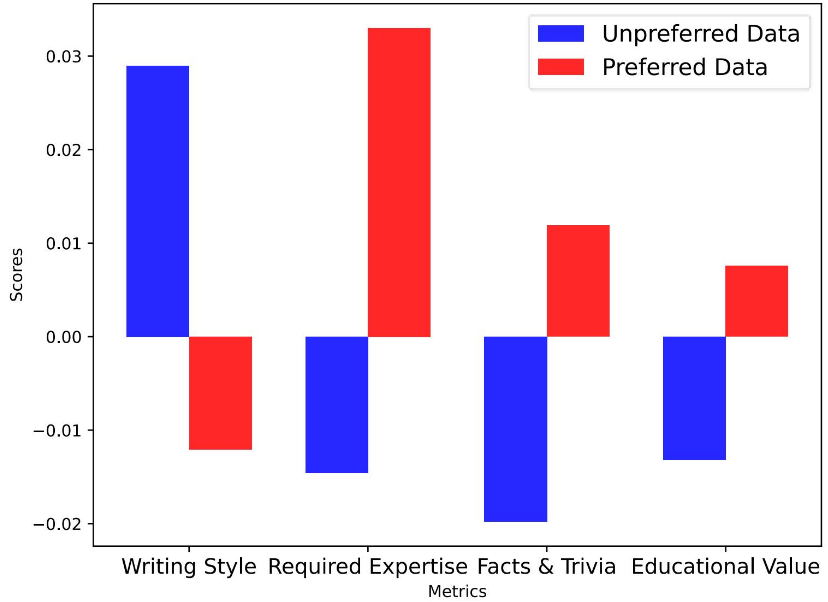


Figure 3. Preferred (red) and unpreferred (blue) data of the weighting model in different training stages, considering properties of writing, expertise, facts and educational values.



4 Conclusion

To this end, this paper

- ✓ proposes a novel bi-level optimization framework with a data weighting model
- ✓ improves the performance of models trained with carefully selected data but also enables models trained with randomly selected data to achieve competitive results
- ✓ demonstrates transferring DWM to larger models yields consistent performance improvements
- ✓ provides insights into how a model's data preferences evolve throughout training

Limitation

- ➢ The additional training cost of introducing the weighting model during training
- ➢ The incompatibility between the model and the high-quality reasoning data when transferring DWM

中国科学院大学
University of Chinese Academy of Sciences

HUAWEI CASIA

Thank you.