UFO-RL: Uncertainty-Focused Optimization for Efficient Reinforcement Learning Data Selection

Yang Zhao[♠], Kai Xiong[♠],Xiao Ding[♠†],Li Du[♡],YangouOuyang[♠],Zhouhao Sun[♠],
Jiannan Guan[♠],Wenbin Zhang[♣], Bin Liu[♣],Dong Hu[♣], Ting Liu[♠]and Bing Qin[♠]
[♠]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
[♡]Beijing Academy of Artificial Intelligence, Beijing, China
[♣]Du Xiaoman Technology (Beijing) Co., Ltd.
{yangzhao, kxiong, xding, oyyo, hzsun, jnguan, tliu, qinb}@ir.hit.edu.cn
duli@baai.ac.cn
zhangwenbin,liubin,hudong@duxiaoman.com

Yang Zhao
Harbin Institute of Technology

2025.12

The Challenge: The High Cost of RL for LLMs

- **The Goal:** Reinforcement Learning (RL) is a powerful paradigm for enhancing the complex reasoning abilities of Large Language Models.

- **The Bottleneck:** RL is extremely costly because it requires multiple interactions with the environment (i.e., multi-sampling per instance) to evaluate and optimize its policy, creating a massive computational overhead.

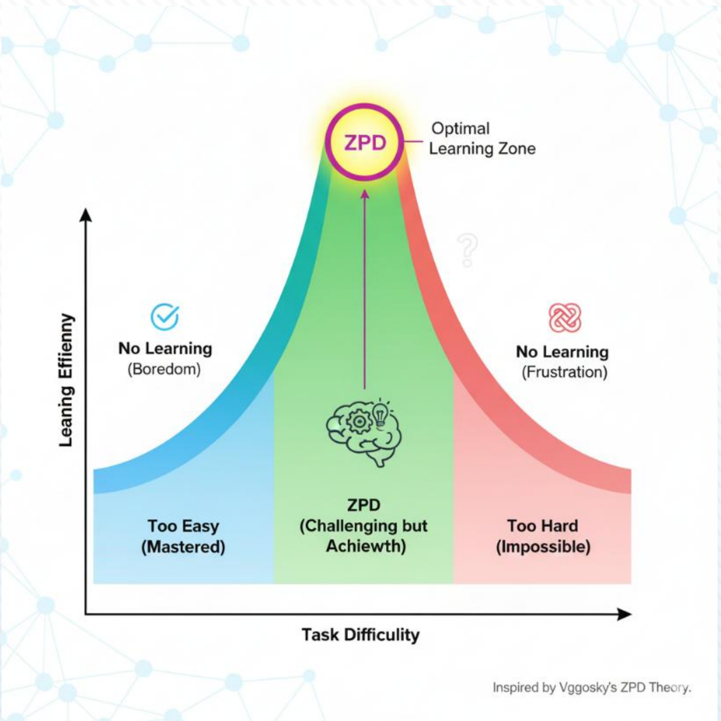
- **The Need**: This calls for a new data selection strategy guided by the "Less is More" principle—to dramatically improve training efficiency by focusing on the most valuable data.

The Core Idea: The Zone of Proximal Development

- **Inspiration from Cognitive Science:** The "Zone of Proximal Development" (ZPD) theory suggests that optimal learning occurs on tasks that are challenging but not impossible.

- **Hypothesis for LLMs:** We hypothesize that LLMs learn best from data they have not yet mastered but show the potential to comprehend.
  - We call this "**Fuzzy Data**"—where the model's understanding is incomplete or uncertain.

- **The Goal:** Identify and focus training on this "fuzzy" middle ground, avoiding data that is either too easy or too hard.



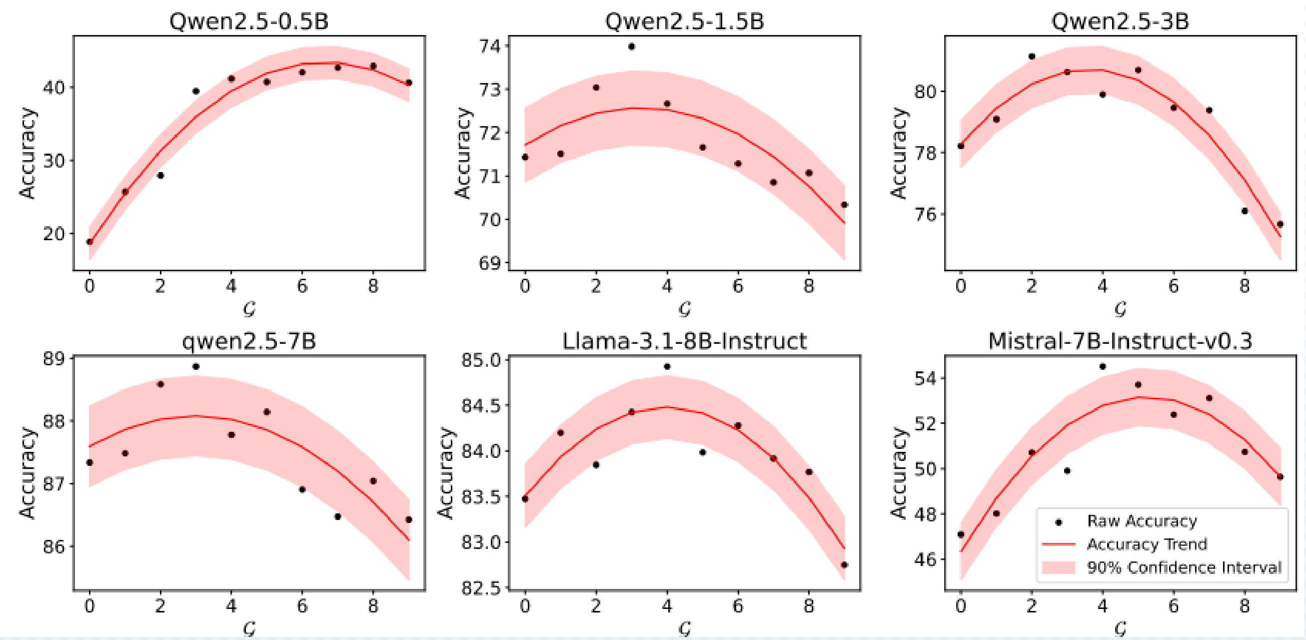
Preliminary Study: Validating the ZPD Hypothesis

Method:

- Used multi-sampling accuracy as a proxy for data difficulty.
- Partitioned the GSM8K dataset into 10 bins based on accuracy (from hardest to easiest).
- Trained models exclusively on data from each bin.

Finding:

- A clear non-monotonic relationship was observed.
- Performance peaks when training on data of intermediate difficulty, strongly supporting the ZPD hypothesis.



The UFO-RL Framework: A New Solution

- Problem with the Preliminary Approach: Using multi-sampling to find the ZPD is self-defeating—it relies on the very bottleneck we want to eliminate.

- Introducing UFO-RL: Uncertainty-Focused Optimization for Reinforcement Learning.

- Core Innovation: A lightweight and scalable framework that uses a computationally efficient, single-pass uncertainty estimation technique.
  - It completely avoids multi-sampling for data selection.
  - It efficiently identifies "fuzzy data" within the model's ZPD for training

Core Technology: Single-Pass Confidence Estimation

How It Works:

- Generate a single complete answer sequence $\{y_1, \dots, y_T\}$for an input x_i .
- Define the Confidence Score as the average log-probability of the output tokens:

$$Conf(x_i) = \frac{1}{T}\sum_{t=1}^{T} \log P(y_t | x_i, y < t)$$

- Select the top 10% of samples with a "fuzziness score" that prioritizes confidence values near the dataset mean.

Advantages:

- Extremely Fast: Requires only a single forward pass, achieving up to a 185x speedup in data evaluation over multi-sampling.
- Fine-Grained: Provides a continuous uncertainty signal, unlike discrete rewards.
- Consistent: Shows strong correlation with multi-sampling accuracy.

Experimental Setup

- **Models:** Qwen2.5 (0.5B-7B), Llama3.1-8B, Mistral-7B
- **Training Datasets:** GSM8K and the more challenging DAPO-MATH-17K
- **Evaluation Datasets:**
  - In-Domain: GSM8K Near-Domain: Math500 Out-of-Domain: MMLU
- **Baselines:**

Full Data	Random	High Conf	Low Conf	Acc_{Filter}	UFO_{ours}
100% data	10% random	10% easiest	10% hardest	removing 0% and 100% accuracy data	10% mid-uncertainty data

	Train Set	Test Set	Full Data	High Conf	Low Conf	Random	Acc_{Filter}	UFO_{ours}
Qwen 2.5-7B	GSM8K	GSM8K	91.88	71.09	91.20	90.93	<u>91.35</u>	**92.03**
		Math500	75.00	74.40	75.60	75.64	<u>76.20</u>	**76.40**
		MMLU	69.64	69.25	**69.44**	69.14	69.26	<u>69.43</u>
	DAPO-MATH-17K	GSM8K	92.03	84.75	81.34	87.43	<u>88.09</u>	**91.16**
		Math500	75.80	76.40	<u>77.20</u>	75.46	75.60	**77.40**
		MMLU	70.33	68.88	68.88	<u>69.17</u>	68.91	**69.69**

Key Result 1: Performance

Less is More:

- Training on just 10% of the data selected by UFO-RL achieves performance comparable to or even surpassing training on the full dataset.

Enhanced Generalization:

- On near-domain benchmarks like Math500, UFO-RL often outperforms full-data RL, suggesting better generalization.

Superior Selection:

- UFO-RL consistently outperforms other data reduction baselines like random sampling or focusing on extreme-difficulty samples

Increased Stability: .

- On the challenging DAPO-MATH-17K dataset, UFO-RL demonstrates resilience and avoids catastrophic performance drops seen with other methods

Key Result 2: Computational Efficiency

Data Evaluation Speedup:

- The single-pass confidence estimation is up to **185x faster** than calculating multi-sample accuracy.

Method	Model	Time	SpeedUp	Model	Time	SpeedUp
Accuracy Confidence	Qwen2.5-0.5B	3337s 18s	×185	Qwen2.5-1.5B	3712s 45s	×82
Accuracy Confidence	Qwen2.5-3B	4186s 89s	×47	Qwen2.5-7B	6827s 175s	×39
Accuracy Confidence	Llama3.1-8B	11426s 186s	×61	Mistral 7B	8335s 146s	×57

Overall Training Time Reduction:

- By processing only 10% of the data, UFO-RL achieves up to a 16x reduction in total RL fine-tuning time compared to the full-data baseline.

Method	Model	Time (s)	Speedup	Method	Time (s)	Speedup
UFO Full Data	Qwen2.5-0.5B	140 1815	×13	Qwen2.5-1.5B	407 5694	×14
UFO Full Data	Qwen2.5-3B	739 10224	×14	Qwen2.5-7B	1154 12959	×11
UFO Full Data	Llama3.1-8B	1219 14040	×12	Mistral 7B	1454 22955	×16

Thanks