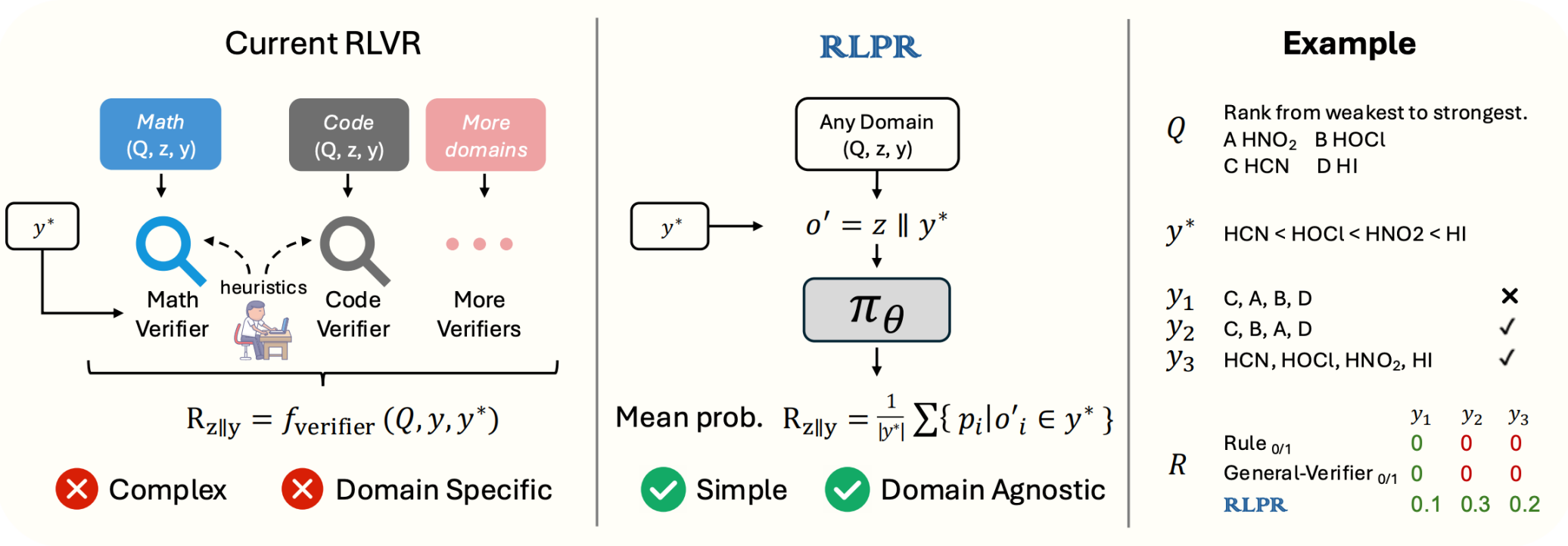
RLPR: Scaling RLVR to General Domain without Verifiers

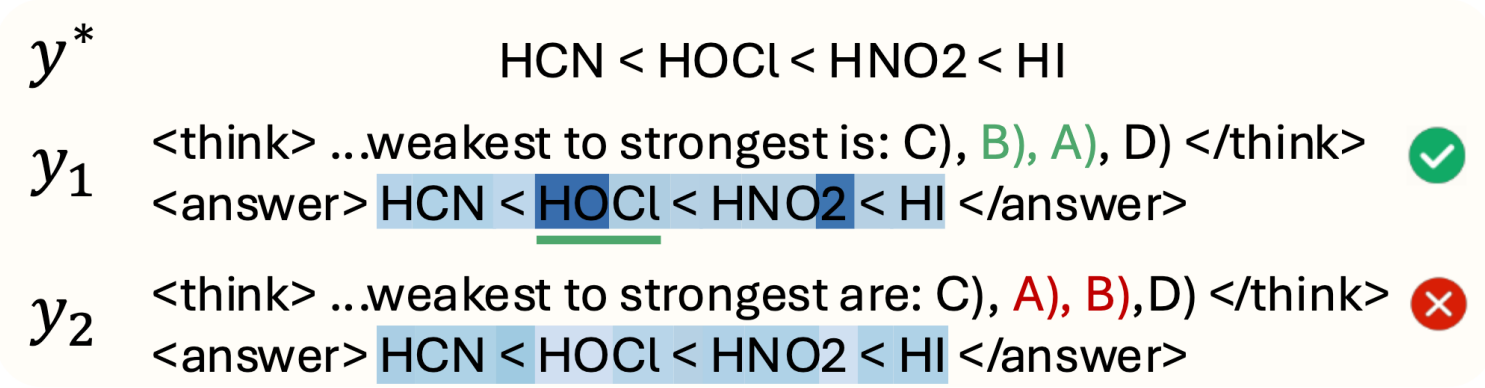
- **Challenge:**
  - Rule-based verifiers construction is laborious and domain specific.
  - Verifier models require extensive data annotation and complicate RL systems.



RLPR: Scaling RLVR to General Domain without Verifiers

- **Challenge:**
  - Rule-based verifiers construction is laborious and domain specific.
  - Verifier models require extensive data annotation and complicate RL systems.

- **Insight:** LLM's intrinsic probability of generating a correct answer directly indicates its own evaluation of the reasoning reward (i.e., how well the reasoning process leads to the correct answer



RLPR: Scaling RLVR to General Domain without Verifiers

- **Probability-based reward**: Using average token probability is much more robust than naive likelihood-based reward signal.
  - For instance, (0.01, 0.7, 0.9) and (0.05, 0.7, 0.9) yield vastly different likelihood score.

$$r = f_{\text{seq}}(\{p_i | o'_i \in y^*\}),$$

- **Reward debiasing:** Reducing influence of unobserved latent factors other than reasoning quality by introduce a baseline without reasoning z.

$$U_r = U_z + U_{\text{others}}, \qquad \hat{r} = \text{clip}(0, 1, r - r'),$$

- **Standard deviation filtering**: Keep only prompts with reward standard deviation larger than a dynamic threshold.

RLPR: Scaling RLVR to General Domain without Verifiers

- RLPR significantly improves the reasoning capability, even surpassing methods using dedicate verifier models by +1.0 on average across seven benchmarks..

Model	Base	Verifier	MMLU-Pro Avg@2	GPQA Avg@4	TheoremQA Avg@2	WebInst. Avg@2	MATH-500 Avg@2	Minerva Avg@2	AIME 24 Avg@16	General -	All -
Gemma Models											
Gemma2-2B-it	Base	–	27.9	19.3	16.4	33.5	26.6	15.9	0.0	24.3	19.9
RLVR	Inst	Rule	31.6	25.8	20.1	**52.3**	**30.7**	16.5	**0.2**	32.4	25.3
RLPR	Inst	✗	**33.5**	**28.5**	**21.2**	52.0	30.4	**17.1**	**0.2**	**33.8**	**26.0**
Llama Models											
Llama3.1-8B-Inst	Base	–	46.4	31.6	31.3	54.7	50.1	32.7	4.2	40.5	35.6
RLVR	Inst	Rule	49.3	36.0	32.0	60.2	51.9	35.2	4.6	44.4	38.5
RLPR	Inst	✗	**53.6**	**36.5**	**35.5**	**68.5**	**54.1**	**39.0**	**8.8**	**48.5**	**42.3**
Qwen Models											
Qwen2.5-7B	–	–	45.3	32.4	41.4	60.4	63.0	37.6	6.5	44.9	40.9
Qwen2.5-7B-Inst	Base	–	54.5	34.2	47.3	72.6	75.4	49.4	9.4	52.2	49.0
Oat-Zero	Math	Rule	45.8	**38.8**	53.3	71.5	80.8	52.1	**29.8**	52.4	53.2
PRIME	Math	Rule	39.5	32.1	47.7	54.5	76.4	45.5	20.4	43.4	45.2
SimpleRL-Zoo	Math	Rule	46.9	38.4	51.1	70.3	77.1	51.0	26.5	51.7	51.6
TTRL	Base	Rule	51.1	34.1	48.8	68.0	**82.1**	52.8	15.8	50.5	50.4
SimpleRL-Zoo	Base	Rule	54.1	36.2	49.5	70.7	76.3	49.2	14.8	52.6	50.1
RLVR	Base	Rule	55.1	36.2	52.2	75.3	76.5	54.9	17.7	54.7	52.6
General Reasoner	Base	Model	55.4	37.4	52.1	74.5	77.0	51.7	16.0	54.8	52.0
VeriFree	Base	✗	53.8	36.7	47.6	72.5	73.5	49.0	12.5	52.6	49.4
RLPR	Base	✗	**56.0**	37.6	**55.4**	**75.5**	78.0	**56.5**	16.3	**56.1**	**53.6**

RLPR: Scaling RLVR to General Domain without Verifiers

- RLPR exceeds concurrent VeriFree by 7.8 on TheoremQA and 7.5 on Minerva

Model	Base	Verifier	MMLU-Pro Avg@2	GPQA Avg@4	TheoremQA Avg@2	WebInst. Avg@2	MATH-500 Avg@2	Minerva Avg@2	AIME 24 Avg@16	General -	All -
Qwen Models											
Qwen2.5-7B	–	–	45.3	32.4	41.4	60.4	63.0	37.6	6.5	44.9	40.9
Qwen2.5-7B-Inst	Base	–	54.5	34.2	47.3	72.6	75.4	49.4	9.4	52.2	49.0
Oat-Zero	Math	Rule	45.8	**38.8**	53.3	71.5	80.8	52.1	**29.8**	52.4	53.2
PRIME	Math	Rule	39.5	32.1	47.7	54.5	76.4	45.5	20.4	43.4	45.2
SimpleRL-Zoo	Math	Rule	46.9	38.4	51.1	70.3	77.1	51.0	26.5	51.7	51.6
TTRL	Base	Rule	51.1	34.1	48.8	68.0	**82.1**	52.8	15.8	50.5	50.4
SimpleRL-Zoo	Base	Rule	54.1	36.2	49.5	70.7	76.3	49.2	14.8	52.6	50.1
RLVR	Base	Rule	55.1	36.2	52.2	75.3	76.5	54.9	17.7	54.7	52.6
General Reasoner	Base	Model	55.4	37.4	52.1	74.5	77.0	51.7	16.0	54.8	52.0
VeriFree	Base	✗	53.8	36.7	47.6	72.5	73.5	49.0	12.5	52.6	49.4
RLPR	Base	✗	**56.0**	37.6	**55.4**	**75.5**	78.0	**56.5**	16.3	**56.1**	**53.6**

RLPR: Scaling RLVR to General Domain without Verifiers

- RLPR shows better robustness and learning efficiency, while VeriFree exhibits great sensitivity to the choice of prompt and achieves lower performance.

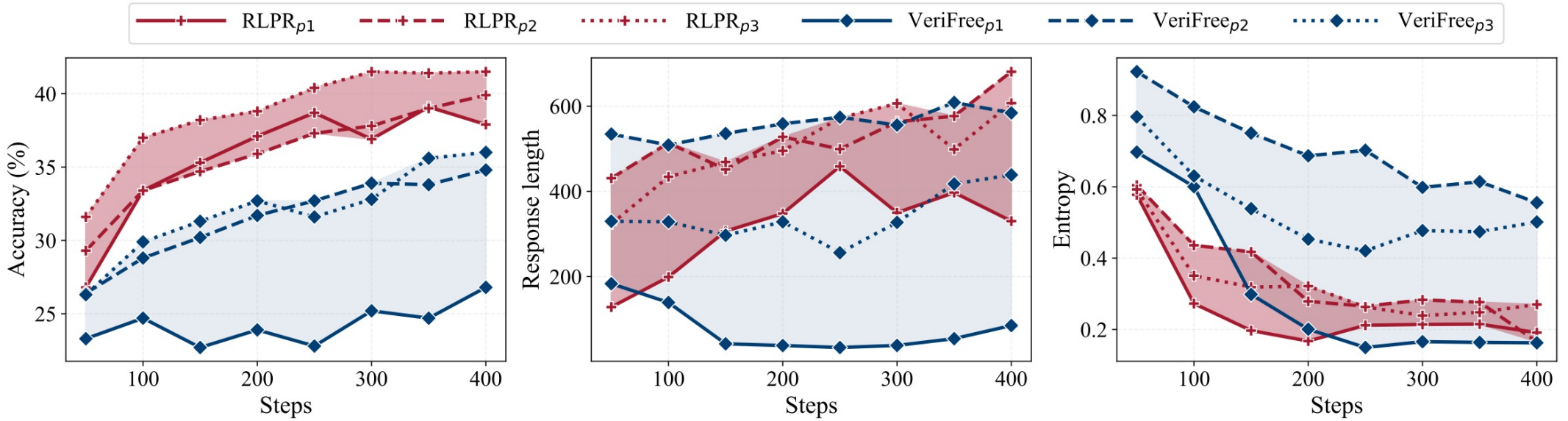


Figure 5: Robustness across different training prompt templates. **RLPR** yields consistently higher performance compared with VeriFree. Left: average performance on seven benchmarks. Middle: response length. Right: response entropy during training.

RLPR: Scaling RLVR to General Domain without Verifiers

- Probability-based reward (PR) quality: we rank responses for each prompt according to the respective rewards and compuate the ROC-AUC metric.
  - PR consistently outperforms rule-based and model-based verifiers.
  - 0.5B small-sized model is enough to generate high-quality rewards.

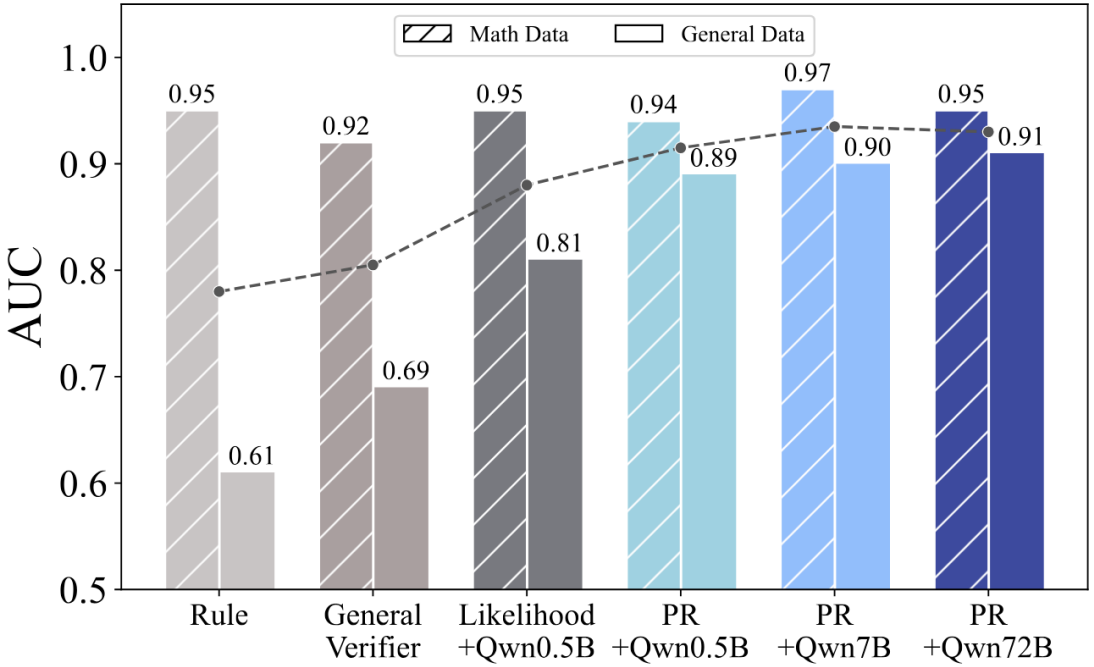


Figure 4: Reward quality comparison. We report the AUC on both math data and general data, and highlight the average score with the dashed line. Qwn: Qwen2.5 models.

RLPR: Scaling RLVR to General Domain without Verifiers

- **Ablation Study**: all components are crucial, and using likelihood instead of token average probabiltiy can significantly hinder the training stability and final performance.

- **PR on verifiable domains**: we study the effectiveness of RLPR on math domains using the dataset from PRIME, and observe that including PR can further improve the utilization of mathematic data.

Method	TheoremQA	Minerva
RLPR	**55.4**	**56.5**
w/o debiasing	52.7-2.7	54.1-2.4
w/o std-filtering	52.5-2.9	55.1-1.4
w/o token prob.	33.5-21.9	34.2-22.3

Table 3: Ablation experimental results. Token prob.: token probability average. Avg@2 results are reported.

Reward	TheoremQA	Minerva
Rule-based	44.8	**50.0**
Rule-based + PR	**48.8**	49.0

Table 4: Experimental results of different rewards on mathematical data. Avg@2 results are reported. We combine rule-based reward and PR by summarizing advantages.

RLPR: Scaling RLVR to General Domain without Verifiers

- **Important questions in LLM+RL**
- Data scaling
- Effective exploration
- Training Efficiency