AAAI 2026
January 20 – 27, 2026
Singapore

VerifyBench: A Systematic Benchmark for Evaluating Reasoning Verifiers Across Domains

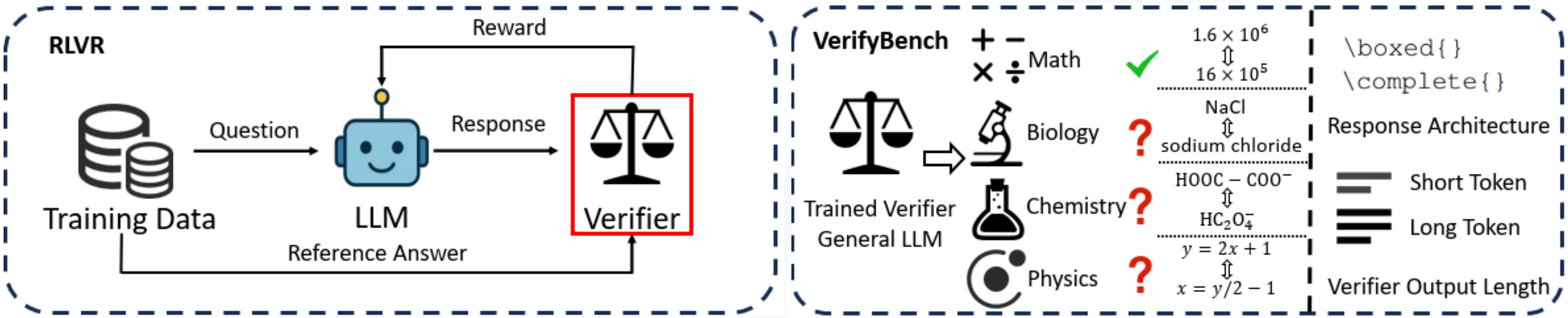
Xuzhao Li[1*†], **Xuchen Li**[2,3,4*], **Shiyu Hu**[5], **Yongzhen Guo**[1‡], **Wentao Zhang**[4,6‡]

[1]Ant Group
[2]Institute of Automation, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]Zhongguancun Academy
[5]Nanyang Technological University
[6]Peking University
xuzhaoli2001@gmail.com, xuchenli1030@gmail.com, yongzhen.gyz@antgroup.com, wentao.zhang@pku.edu.cn

Motivation

- In Reinforcement Learning with Verifier (RLVR), the verifier is tasked with assessing **the consistency between LLM outputs and reference answers**;
- however, existing solutions exhibit significant deficiencies:
  - **Limitation 1**: **Rule-based verifiers** suffer from poor generalization and an inability to handle flexible linguistic expressions.
  - **Limitation 2**: **Model-based verifiers** (both specialized and general-purpose) lack systematic evaluation across diverse domains and scenarios.

The absence of **a unified benchmark for comprehensive verifier comparison** hinders the further advancement of RLVR.

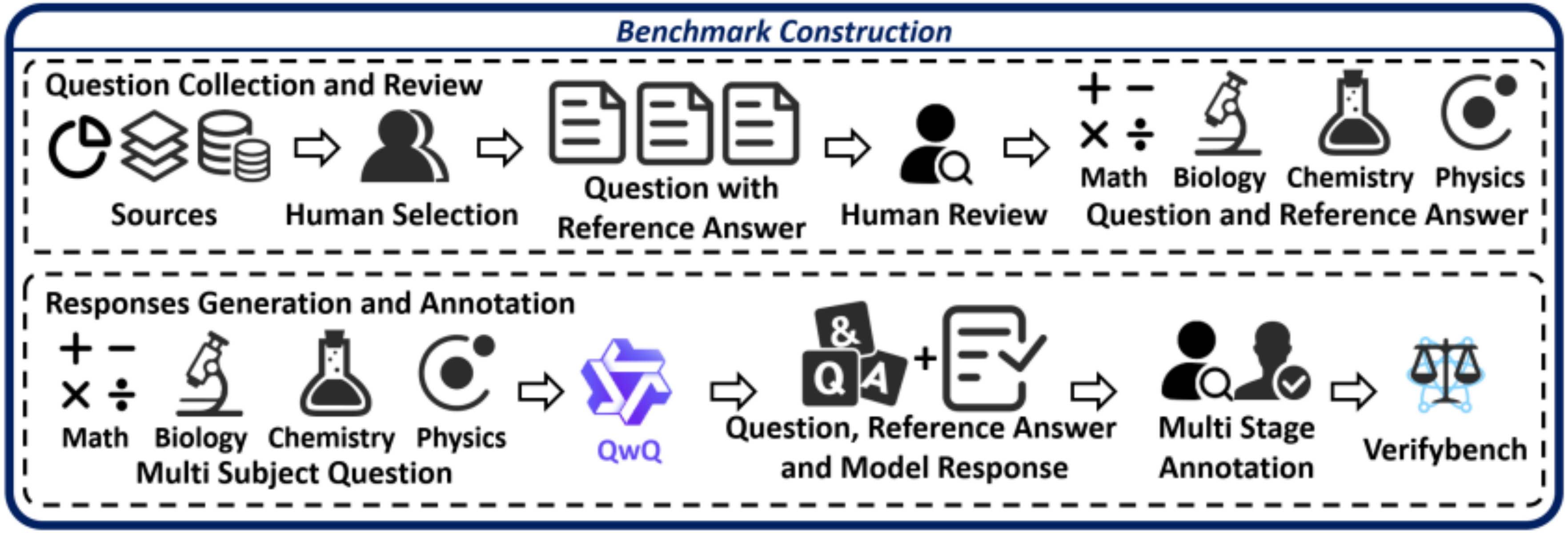


Contributions

- **Data Perspective:** We propose VerifyBench, a benchmark comprising 3,989 **expert-level problems** across **four core domains** (Mathematics, Physics, Chemistry, and Biology). It features meticulous human annotations and **diverse Chain-of-Thought (CoT) response candidates**.
- **Methodological Perspective**: We design a four-dimensional experimental framework (**Verifier Type × Input Format × Output Length × Domain**) to systematically evaluate and compare **specialized verifiers against general-purpose LLMs**.
- **Findings Perspective**: Our analysis reveals critical insights, including the **precision-recall trade-off** in verification, **sensitivity to input structural formatting**, and limitations in **cross-domain generalization**, providing concrete directions for future optimization.

Benchmark Construction

- **Problem Collection:** Problems were curated by **multi-disciplinary experts**, covering **university-level** curricula and **academic competition** challenges. Each entry is paired with a standard reference answer.
- **Response Generation**: Detailed CoT responses were generated using the QwQ-32B model, with the final answers encapsulated within **\boxed{}** tags for **standardized extraction**.
- **Two-stage Annotation and Annotation Guidelines: A dual-annotation process and cross-domain sampling** for cross-validation. The annotation criteria permit **synonymous expressions and symbolic equivalence**.



Data Statistics

- The VerifyBench dataset consists of **3,989 high-quality entries**, distributed nearly equally across **the four domains**. A defining characteristic of the benchmark is its depth and complexity: **the average response length** reaches 4,553 tokens, reflecting **the intricate multi-step reasoning** required for expert-level problems.

Statistic	Value
Total Questions	3,989
Average Question Length	186 tokens
Average Model Response Length	4,553 tokens
Total Annotated Instances	3,989
Label Distribution (Correct / Incorrect)	45% / 55%
Inter-Annotator Agreement (IAA)	0.88 – 0.92

Statistics of VerifyBench

Semantic Diversity

Logical Discernment

Cross-Domain Awareness

Characteristic

Experiment Setting

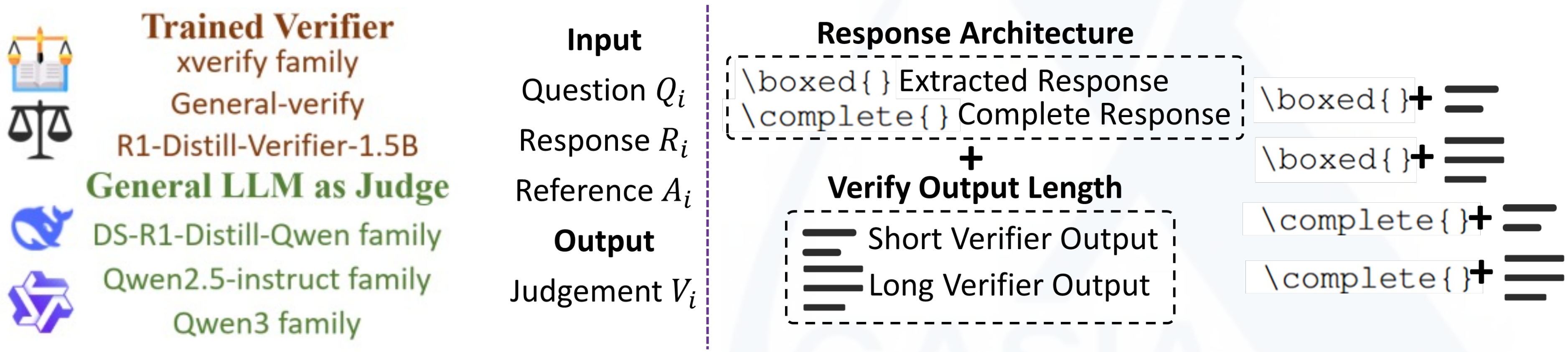
Evaluation Baseline:

Specialized Verifiers: Including the xVerify series, R1-Distill-Verifier-1.5B, and other models specifically fine-tuned for verification tasks.

General-purpose LLMs: Including the Qwen2.5/3 series and DeepSeek-R1-Distill-Qwen series.

Experiment Setting:

A systematic configuration that benchmarks verifiers by integrating diverse **response architectures** and varying **output lengths** across multiple domains.



Results and Analysis

Verifier	Mathematics	Chemistry	Biology	Physics	Overall
Trained Verifier					
xVerify-0.5B-I	79.38%\54.14%	94.67%\94.46%	87.15%\91.07%	92.38%\92.81%	88.64%\84.76%
xVerify-3B-Ia	81.18%\56.91%	95.03%\96.21%	88.76%\90.18%	91.88%\92.24%	89.38%\85.35%
xVerify-8B-I	81.58%\55.80%	**96.28%****96.50%**	89.16%\91.07%	92.58%\92.69%	90.06%\85.47%
xVerify-9B-C	82.78%\64.92%	96.18%**96.50%**	**89.96%****92.86%**	93.98%\94.86%	90.86%\88.66%
general-verify	68.77%**88.12%**	75.13%\88.63%	73.09%\85.71%	94.32%\97.72%	79.01%\93.03%
R1-Distill-Verifier-1.5B	76.18%\81.22%	80.71%\86.30%	77.91%\78.57%	88.77%\89.50%	81.91%\86.36%
General LLM as Judge					
Qwen2.5-7B-Instruct	77.88%\47.51%	89.45%\93.59%	54.62%\29.46%	86.66%\86.64%	82.35%\75.90%
Qwen2.5-14B-Instruct	78.08%\59.12%	90.15%\86.01%	62.25%\34.82%	91.67%\92.47%	84.75%\80.21%
Qwen2.5-32B-Instruct	81.08%\71.27%	85.28%**96.50%**	62.25%\48.21%	95.39%\97.37%	81.20%\88.36%
Qwen3-8B	70.77%\74.31%	88.04%\93.88%	81.53%\83.04%	95.09%\97.37%	84.38%\90.79%
Qwen3-14B	**85.39%**\80.11%	92.61%\95.92%	84.34%**92.86%**	**96.99%**\98.52%	**91.11%****93.68%**
Qwen3-32B	74.67%\70.72%	83.82%\93.00%	83.40%\91.96%	95.89%\97.15%	84.69%\90.31%
DS-R1-Distill-Qwen-7B	64.66%\74.59%	75.38%\85.42%	74.80%\77.68%	91.78%\97.03%	77.07%\88.72%
DS-R1-Distill-Qwen-14B	76.18%\78.73%	81.06%\85.13%	68.67%\64.29%	95.69%\97.26%	83.02%\88.66%
DS-R1-Distill-Qwen-32B	72.37%\78.73%	79.10%**96.50%**	72.29%\85.71%	96.59%**99.43%**	81.85%\93.50%

Table 3: Performance comparison on VerifyBench, with results shown in terms of Accuracy/Recall. The table is organized by the trained verifier and the general LLM as the judge with the `\complete{}` response from QwQ-32B, and the maximum output token size is set to 4k. "DS" denotes DeepSeek. The best results are highlighted in bold.

- **Specialized verifiers prioritize correctness** and **reject ambiguous or loosely matched responses**, aiming to **reduce false positives**.
- **General LLMs** adopt **a more inclusive stance**, recognizing broader expression forms and redundant reasoning.

Results and Analysis

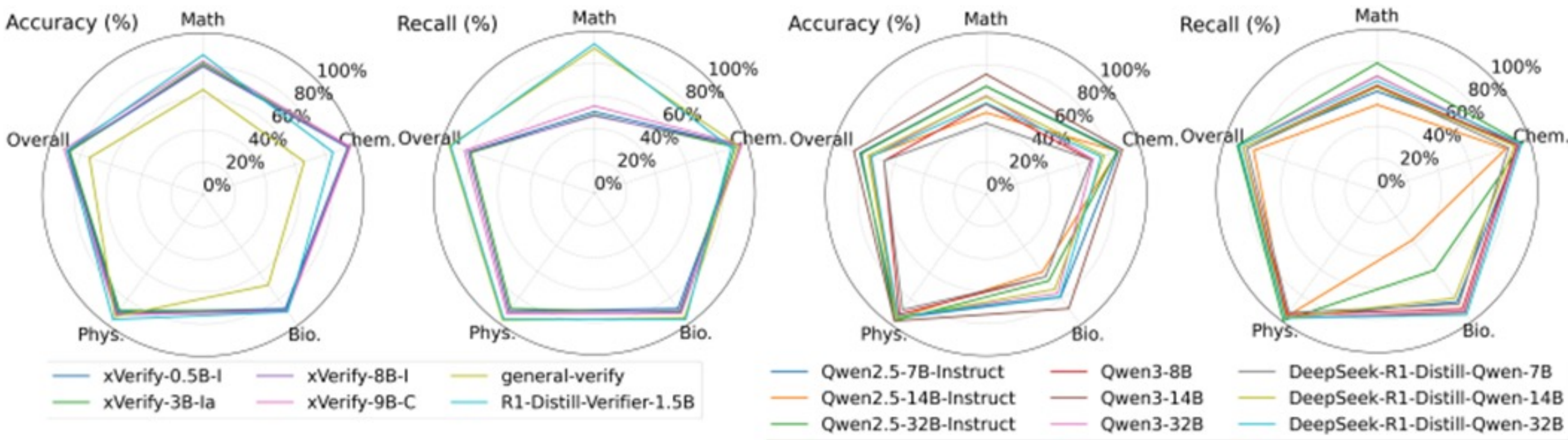


Figure 4: Performance comparison on VerifyBench including mathematics, chemistry, biology and physics. The figure is organized by the trained verifier (left) and the general LLM as the judge (right) with the response in the format `\boxed{}` from QwQ-32B, and the maximum output token size is set to 4k.

- **Specialized verifiers** consistently achieve **higher accuracy**, especially in fields demanding **strict semantic consistency**.
- **General LLMs** exhibit **greater sensitivity** to input/output conditions.
- Verifiers should ideally produce direct, structured outputs **without relying on the extraction of model response or verifier's judgment result**. This reduces **engineering overhead and minimizes error propagation**.

Thanks!

Shiyu Hu
January 24, 2026