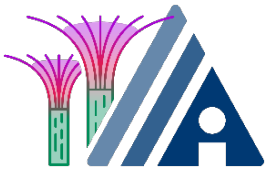


AAAI 2026
January 20 – 27, 2026
Singapore






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

Dr. Shiyu Hu

- Research Fellow in Nanyang Technological University (NTU)
- https://huuuuusy.github.io/
- shiyu.hu@ntu.edu.sg

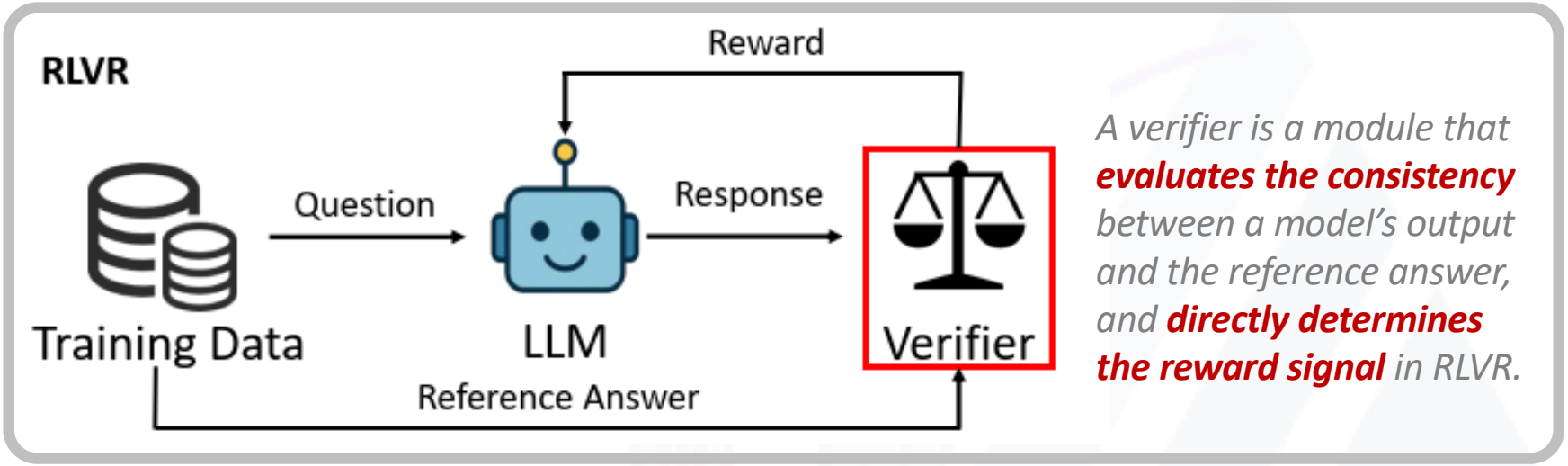
Scan to download this slides

WeChat for the first author

Motivation

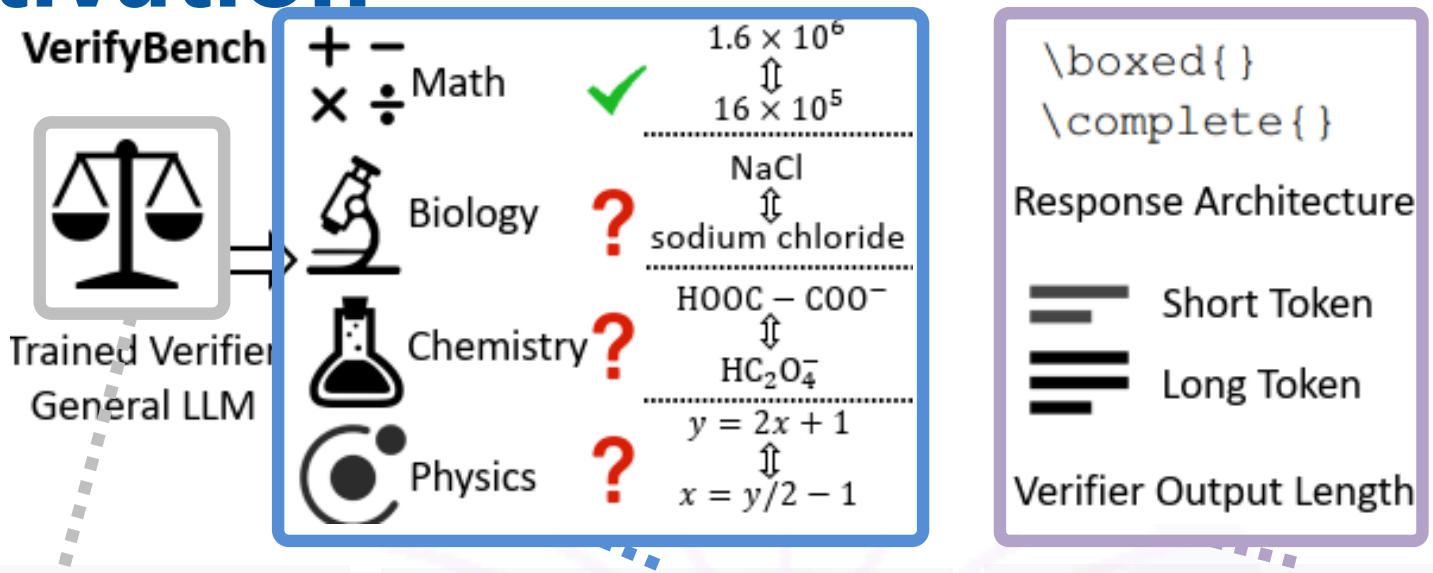
Reinforcement Learning with Verifier (RLVR)

- In RLVR, an LLM generates responses to given questions based on training data.
- The learning process is **guided by rewards assigned by a verifier**, rather than direct supervision from ground-truth answers.
- The verifier compares the model's response with the reference answer and defines the reward signal.



Since rewards are defined by the verifier, **any bias or unreliability in the verifier will be amplified through reinforcement learning**, leading to systematic training errors.

Motivation



Existing Verifier Paradigms

- **Rule-based verifiers:** Rely on strict matching rules or symbolic equivalence checks.
- **Model-based verifiers:** Use trained models or general-purpose LLMs to judge answer correctness.

Challenges Across Disciplines

- The notion of **answer equivalence** varies substantially across domains (e.g., mathematics, chemistry, biology, physics).
- Equivalent answers may **differ in notation, structure, or linguistic form**.

Sensitivity to Evaluation Settings

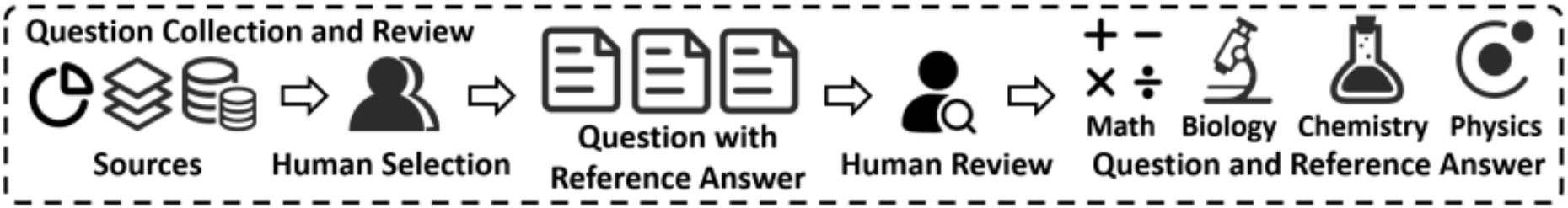
- Verifier performance depends on **response input formats** (final answer vs. full reasoning).
- Verifier behavior is also affected by **output length constraints** (short judgment vs. long explanation).

Despite their widespread use, verifiers **lack a unified benchmark** for systematic comparison across verifier types, disciplines, and evaluation settings.

→ This motivates a unified benchmark (VerifyBench) for comprehensive verifier comparison.

Our Method: VerifyBench

- **Dataset Construction**



Pipeline: Question Collection & Review

- **Sources → Human Selection → Questions w/ Reference Answers → Expert Review → Final QA Pairs**
- **Coverage:** Math / Biology / Chemistry / Physics (multi-disciplinary)

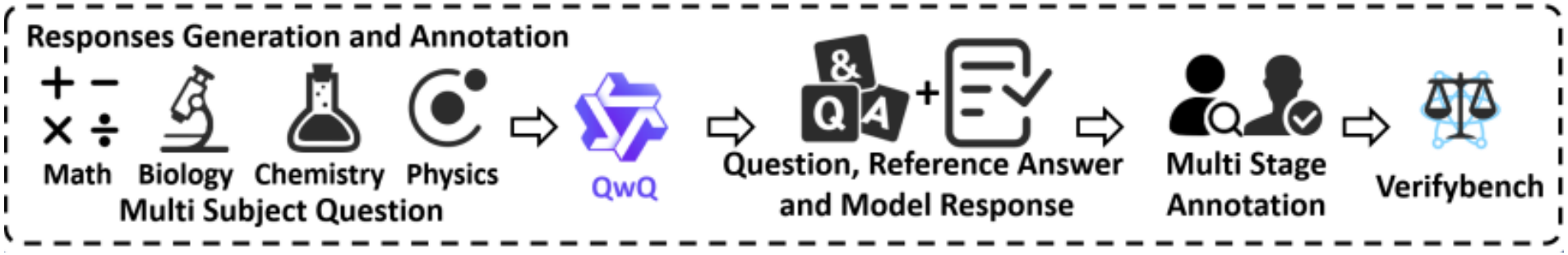
Data Characteristics

- **Multi-disciplinary experts**: reviewed and validated by domain experts
- **University-level**: aligned with university curricula and advanced training
- **Academic competition**: includes competition-style problems for higher difficulty and discriminability

VerifyBench is constructed through a rigorous, expert-driven pipeline to ensure high-quality, multi-disciplinary evaluation data.

Our Method: VerifyBench

- **Dataset Construction**



Standardized Response Generation & Extraction

- Model responses are generated under a **unified protocol**, with a **standardized answer extraction process** applied to all samples.
- Extraction explicitly accounts for **synonymous expressions and symbolic equivalence**, rather than surface-form matching.

Dual Annotation with Cross-domain Consistency Checking

- Correctness labels are obtained through **dual annotation and multi-stage human review**.
- **Cross-domain sampling** is conducted to verify annotation consistency across different disciplines.
- Disagreements are resolved by expert adjudication to ensure reliable final labels.

VerifyBench ensures reliable labels through standardized extraction and dual annotation with cross-domain consistency checks.

Our Method: VerifyBench

- **Data Statistics**

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

Reasoning Depth and Verifier Stress

- **Average question length:** 186 tokens
- **Average model response length:** 4,553 tokens
- These long-form responses reflect **intricate, multi-step reasoning**, rather than short or surface-level answers.

Scale and Annotation Reliability

- **3,989 carefully curated questions**, each paired with a verified reference answer.
- **Balanced label distribution** (45% Correct / 55% Incorrect), avoiding trivial majority-class bias.
- **High inter-annotator agreement (IAA: 0.88–0.92)**, indicating reliable and consistent human judgments.

VerifyBench provides a compact yet high-stress evaluation setting, combining reliable annotations, long-form reasoning, and multi-disciplinary coverage to systematically probe verifier robustness.

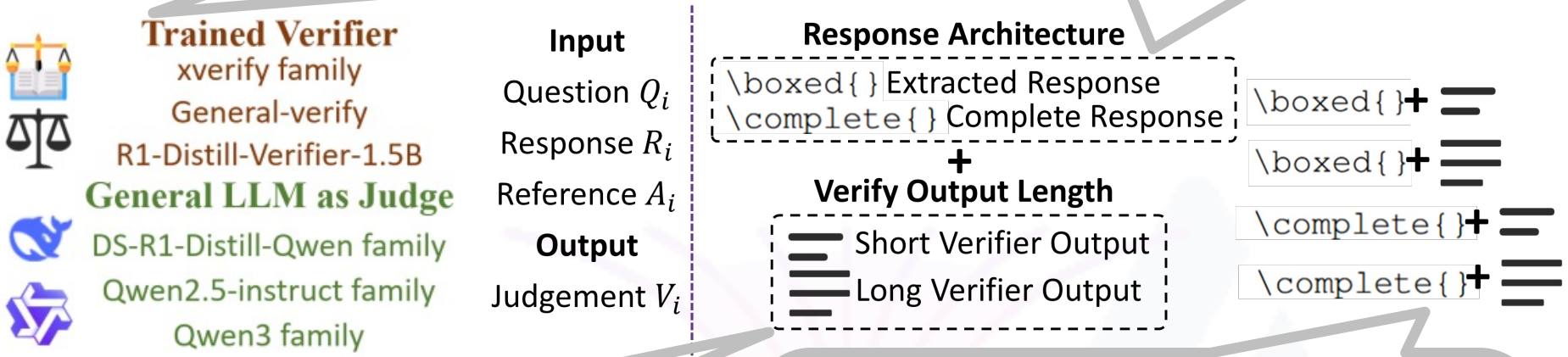
Experiment Setting

Two Categories of Verifiers

To compare **specialized verifiers** against **general-purpose LLM judges** under the same evaluation conditions.

Response Architectures

To study whether access to **intermediate reasoning** affects verifier reliability.



Verifier Output

To analyze how **output verbosity** influences verifier consistency and decision behavior.

Evaluation Protocol

Verifiers are evaluated under **factorized and controlled combinations** of:

- verifier type
- input format (\boxed{} vs. \complete{})
- verifier output length (short vs. long)

We evaluate trained verifiers and general LLM judges under controlled input and output configurations to analyze how verifier type, input format, and output length influence verification behavior.

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

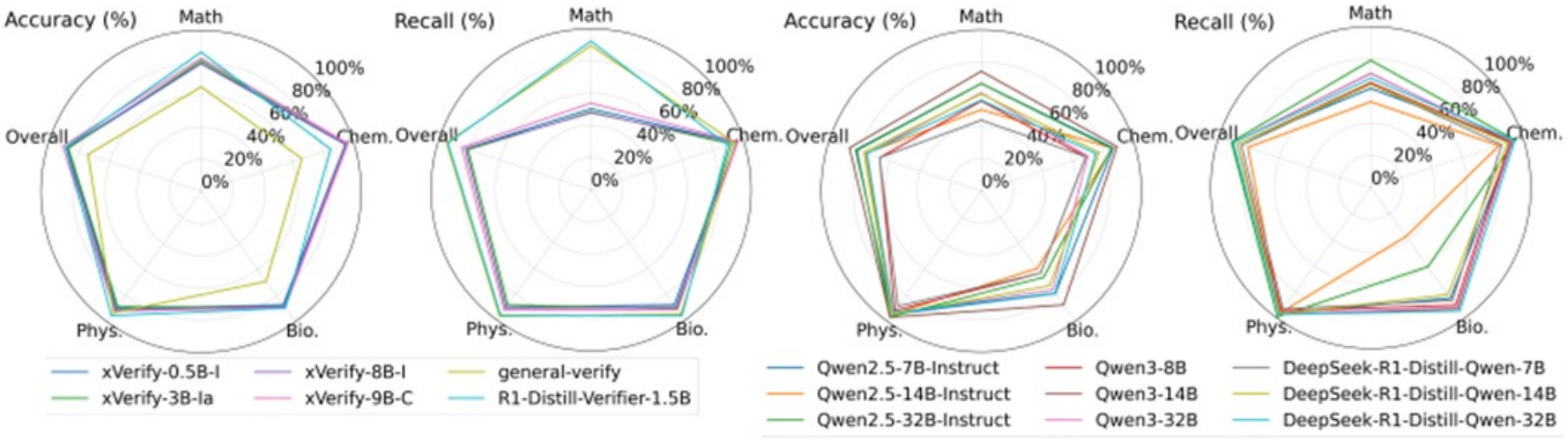


Figure 4: Performance comparison on VerifyBench including mathematics, chemistry, biology and physics. The figure is organized by the trained verifier (left) and the general LLM as the judge (right) with the response in the format `\boxed{}` from QwQ-32B, and the maximum output token size is set to 4k.

- **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**.

Bad Case: Mathematics

- **Correct Answer Rejected due to Expression Equivalence**

Question: Alfred owns a six-sided die where the probability of a number \(x \) appearing on top is directly proportional to \(x^2 \). If he rolls the die twice, what is the probability that the sum of the two numbers he rolls is equal to 7?
Reference Answer: \dfrac{80}{1183}
Model Response: To solve this problem, we need to determine the probability distribution for the outcomes of the die, and then calculate the probability that the sum of two rolls equals 7.\n\n
Step 1: Determine the probability distribution\n\n
The probability of a number \\(x \\) appearing on top is directly proportional to \\(x^2 \\). Let \\(P(x) \\) be the probability of rolling \\(x \\). We can express this as:\n\\[P(x) = kx^2 \\]\nfor \\(x = 1, 2, 3, 4, 5, 6 \\), where \\(k \\) is a constant of proportionality. Since the sum of the probabilities must equal 1, we have:\n\\[\\sum_{x=1}^{6} P(x) = k \\sum_{x=1}^{6} x^2 = k(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = k(1 + 4 + 9 + 16 + 25 + 36) = k \\cdot 91 = 1 \\] \n
Thus, \\(k = \\frac{1}{91} \\).\n\nSo the probability of rolling each number is:\n\\[P(x) = \\frac{x^2}{91} \\]\nfor \\(x = 1, 2, 3, 4, 5, 6 \\).\n\n
Step 2: Calculate the probability of each pair of outcomes that sum to 7\n\nThe pairs that sum to 7 are: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1). We need to find the probability of each pair and sum them up.\n\nThe probability of rolling (1, 6) is:\n\\[P(1) \\cdot P(6) = \\frac{1^2}{91} \\cdot \\frac{6^2}{91} = \\frac{1}{91} \\cdot \\frac{36}{91} = \\frac{36}{8281} \\]\n\n
Similarly, the probability of rolling (2, 5) is:\n\\[P(2) \\cdot P(5) = \\frac{2^2}{91} \\cdot \\frac{5^2}{91} = \\frac{4}{91} \\cdot \\frac{25}{91} = \\frac{100}{8281} \\]\n\n
The probability of rolling (3, 4) is:\n\\[P(3) \\cdot P(4) = \\frac{3^2}{91} \\cdot \\frac{4^2}{91} = \\frac{9}{91} \\cdot \\frac{16}{91} = \\frac{144}{8281} \\]\n\n
The probability of rolling (4, 3) is the same as (3, 4):\n\\[P(4) \\cdot P(3) = \\frac{144}{8281} \\]\n\n
The probability of rolling (5, 2) is the same as (2, 5):\n\\[P(5) \\cdot P(2) = \\frac{100}{8281} \\]\n\n
The probability of rolling (6, 1) is the same as (1, 6):\n\\[P(6) \\cdot P(1) = \\frac{36}{8281} \\]\n\n
Adding these probabilities together, we get:\n\\[\\frac{36}{8281} + \\frac{100}{8281} + \\frac{144}{8281} + \\frac{144}{8281} + \\frac{100}{8281} + \\frac{36}{8281} = \\frac{560}{8281} \\]\n\n
Final Answer\n\\[\\boxed{\\frac{560}{8281}} \\]

Case Setup

- **Task Type:** Probability computation with multi-step mathematical reasoning
- **Key Characteristics:**
  - The model output is **mathematically equivalent** to the reference answer
  - Differences arise from **fraction simplification and equivalent transformations**
- **Manual Verification:** Model answer is correct
- **Verification Outcome: Incorrectly rejected by the verifier**

Why the Verifier Fails

- Fails to recognize **equivalent mathematical expressions**
- Overly relies on **surface-form matching or strict answer extraction rules**
- Lacks robust comparison after **multi-step simplification**

Bad Case: Physics

- **Long-context Answer Extraction and Symbolic Robustness**

Question: A particle has a constant acceleration \(a = 6\mathbf{i} + 4\mathbf{j} \), where \(a \) is in units of \(\mathrm{m} \cdot \mathrm{s}^{-2} \). At \(t = 0 \), its velocity is zero, and its position vector is \(\mathbf{r}_0 = (10\,\mathrm{m})\mathbf{i} \). Find the equation of the trajectory of the particle in the \(Oxy \) plane.
Reference Answer: \(y = \frac{2}{3}x - \frac{20}{3} \)
Model Response: To solve this problem, we need to first find expressions for the velocity and position according to the given conditions, and then use these expressions to derive the trajectory equation.
The given conditions are: the particle's acceleration is \(\mathbf{a} = 6\mathbf{i} + 4\mathbf{j}\) (\(\mathrm{m} \cdot \mathrm{s}^{-2}\)), the initial velocity is \(\mathbf{v}(0) = 0\), and the initial position is \(\mathbf{r}(0) = 10\mathbf{i}\) (\(\mathrm{m}\)).
1. **Find the velocity:**
Velocity is the integral of acceleration. Considering the initial velocity is zero, we can write the expression for velocity as: \mathbf{v}(t) = \int \mathbf{a}\, dt = \int (6\mathbf{i} + 4\mathbf{j})\, dt = 6t\mathbf{i} + 4t\mathbf{j} + \mathbf{C}
Here, \(\mathbf{C}\) is the constant of integration. Because the velocity at \(t=0\) is zero, \(\mathbf{C}=0\). Thus, the velocity is:\mathbf{v}(t) = 6t\mathbf{i} + 4t\mathbf{j}
2. **Find the position:**
Position is the integral of velocity. Considering the initial position \(\mathbf{r}(0) = 10\mathbf{i}\), we obtain:
\mathbf{r}(t) = \int \mathbf{v}(t)\, dt = \int (6t\mathbf{i} + 4t\mathbf{j})\, dt = 3t^2\mathbf{i} + 2t^2\mathbf{j} + \mathbf{D}
Here, \(\mathbf{D}\) is the constant of integration. Since at \(t=0\), the position is \(10\mathbf{i}\), we have \(\mathbf{D} = 10\mathbf{i}\). Therefore, the position is: \mathbf{r}(t) = (3t^2 + 10)\mathbf{i} + 2t^2\mathbf{j}
3. **Find the trajectory equation:**
Now we have the position as a function of time. To find the trajectory equation, we need to eliminate \(t\). Let \(x = 3t^2 + 10\) and \(y = 2t^2\). Solve these two equations to eliminate \(t\):
- From \(y = 2t^2\), we obtain \(t^2 = \frac{y}{2}\).
- Substitute the expression for \(t^2\) into \(x = 3t^2 + 10\), then \(x = 3\left(\frac{y}{2}\right) + 10\).
Simplifying, the trajectory equation is: x = \frac{3}{2}y + 10
Therefore, the equation for the particle's trajectory in the \(Oxy\) plane is \(\boxed{x = \frac{3}{2}y + 10}\).

Case Setup

- **Task Type:** Classical physics problem (kinematics with multi-variable derivations)
- **Key Characteristics:**
  - The correct answer is embedded in a **long reasoning trace**
  - Involves **symbolic equivalence, variable substitution, and LaTeX expressions**
- **Manual Verification:** Model reasoning and final conclusion are correct
- **Verification Outcome: Incorrectly rejected by the verifier**

Why the Verifier Fails

- Difficulty in **locating the final answer within long contexts**
- Lack of robustness to **symbolic equivalence and minor LaTeX variations**
- Treats **parsing or extraction failures** as answer incorrectness

Analysis and Discussion

- **Verifier behavior is not neutral, but systematic:** Verifiers exhibit **distinct and consistent behavioral biases**, such as strict correctness checking or inclusive semantic matching, rather than acting as neutral correctness oracles.

- **Different verifier designs induce different error trade-offs:** Specialized verifiers tend to **reduce false positives at the cost of higher false negatives**, while general LLM judges adopt a more inclusive stance but are prone to **accept loosely matched answers**.

- **Verifier reliability is strongly domain- and format-dependent:** Verifier performance varies significantly across disciplines and answer representations, especially under **long-context reasoning, symbolic equivalence, and expression variation**.

- **VerifyBench enables systematic and controlled verifier evaluation:** VerifyBench provides a **unified benchmark** that disentangles verifier type, input format, and output configuration, making verifier evaluation a **first-class research problem** rather than a by-product of model training.

AAAI 2026
January 20 – 27, 2026
Singapore




Thanks for listening!

2026.01.24 in Singapore

Dr. Shiyu Hu

- Research Fellow in Nanyang Technological University (NTU)
- https://huuuuusy.github.io/
- shiyu.hu@ntu.edu.sg

Scan to download this slides

WeChat for the first author