# layton-eval:<br>Sideways: a Benchmark for LLM/VLM Reasoning

Raphael Vienne

February 5, 2026

**Abstract**

Traditional large language model (LLM) benchmarks face a fundamental trade-off between scalability and ground truth. Human evaluation is accurate but expensive; automated evaluation ("LLM-as-a-judge") scales, but exhibits systematic errors such as self-preference bias, provider/family bias, position (order) bias, and limited sensitivity to subtle logical flaws [8, 10, 11, 13].

We introduce **layton-eval**, a high-difficulty reasoning benchmark built from Professor Layton–style riddles that require both correct answers and sound justifications. To reconcile speed with validity, we combine automated judging with *Prediction-Powered Inference* (PPI) [14], treating a multi-model "jury" as a noisy proxy calibrated against a human-labeled gold set. We further propose a *stratified bootstrap* rectification procedure that matches a model's jury-score distribution, mitigating regression-to-the-mean effects and reporting uncertainty via confidence intervals and rank spread.

## Contents

# 1 Introduction

## 1.1 Motivation and problem statement

A main motivation for **layton-eval** is *capability coverage*: many widely used benchmarks do not probe the breadth of human-level problem-solving, and instead over-index on linear, instruction-following reasoning.

Concretely, popular benchmarks often emphasize academic and professional knowledge (e.g., MMLU [1]), school-style math (GSM8K [2]), science QA (ARC [3]), or code synthesis (HumanEval [4]). Broader evaluation suites exist (e.g., HELM [5] and BIG-bench [6]), but coverage of lateral, puzzle-like reasoning and structured wordplay remains limited. On the multimodal side, benchmarks such as MMMU [7] primarily target expert-level perception and domain knowledge rather than divergent "thinking sideways."

In contrast, **layton-eval** targets reasoning modes that are common in real puzzles and games, including **non-linear (lateral) thinking**, **out-of-the-box logic**, **divergent reasoning**, and **wordplay**. It also naturally extends to **visual reasoning** for VLMs.

A second motivation is the classical trade-off between:

- **Scalability:** automated scoring is fast and cheap.
- **Ground truth:** human scoring is accurate but costly and slow.

This can be viewed as an analogue of the *bias–variance trade-off*. Purely human evaluation tends to have *lower bias* (closer to the target notion of correctness) but *higher variance* because only a small number of items can be labeled (and annotators disagree). Purely automated judging tends to have *lower variance* (large $n$ is inexpensive) but can introduce *systematic bias* from judge failure modes (e.g., self-preference, style bias) [8, 13]. Prediction-Powered Inference (PPI) aims to strike a balance: use the proxy to reduce variance by scoring many items, and use a gold set to estimate and correct the proxy's bias [14, 16].

Automated judging is known to be fallible, especially on multi-step reasoning tasks where plausible explanations can conceal subtle hallucinations [11, 13]. Relatedly, rigorous evaluation protocols for RAG systems in due diligence settings combine human annotation with LLM-judge labels and PPI-style calibration [17].

# 2 Benchmark

## 2.1 layton-eval dataset and task

**layton-eval** uses Professor Layton-style riddles—multi-step logic puzzles—as benchmark items.

The **layton-eval** dataset was obtained by scraping the Layton Wiki at Fandom (`https://layton.fandom.com/wiki`)[1]. The dataset is available on Hugging Face at `rvienne/layton-eval`.

The raw scraped material then underwent substantial manual and automated processing, including:

1. curation of riddles that do not require running the game's verification engine,
2. classification of riddles into text-based versus vision-based items, and
3. for each riddle, generation of a gold-standard justification from the riddle context and answer.

In total, we curated 503 riddles, of which 186 form the `LLM` split and 317 form the `VLM` split.

The resulting benchmark is organized into two splits: an `LLM` split (text-only riddles) and a `VLM` split (vision-based riddles), targeting language-only and multimodal models respectively.

Each item requires:

- a final **answer**, and
- a **justification** that can be checked for logical validity.

This setting stress-tests model reasoning beyond short-form multiple-choice accuracy. Moreover, since some riddles can be formulated as multiple-choice questions (MCQs), a model may occasionally arrive at the correct final answer by chance; in contrast, there is no way to produce a logically valid justification "out of luck." This makes the benchmark more robust overall.

# 3 Background and method

## 3.1 Judge fallibility

Even frontier models used as judges are imperfect proxies for human reasoning [8, 9, 12, 11, 13]. Common failure modes include:

- **Being "fooled" by plausible justifications** that contain subtle errors or hallucinated steps [11, 13].
- **Self-preference bias:** a judge may rate outputs from its own provider more favorably [8, 13].
- **Family/provider bias:** stylistic familiarity can be mistaken for correctness [8, 12].
- **Position/order bias:** verdicts can shift when the same outputs are presented in a different order [10].

[1] This work uses material from the Professor Layton Wiki at Fandom and all subsequent datasets are licensed under the Creative Commons Attribution–ShareAlike License.

### 3.2 Prediction-Powered Inference (PPI)

*Prediction-Powered Inference* (PPI) is a framework for estimating a quantity defined in terms of *expensive ground-truth labels* (here: human correctness judgments), while leveraging a much larger set of *cheap but imperfect proxy labels* (here: automated jury scores) [14, 16].

At a high level, PPI works by:

1. **Scoring at scale with a proxy:** use the proxy to label many examples, which greatly reduces variance because we can afford a large $n$.
2. **Calibrating with a gold set:** additionally collect a smaller set of human labels, and estimate the proxy's *systematic error* (bias) by comparing proxy vs. human labels on that overlap.
3. **Rectifying the proxy estimate:** adjust the proxy-based estimate using the estimated error, producing an estimator that remains anchored to human judgment while retaining much of the proxy's sample-efficiency.

This is a good fit for **layton-eval** because (i) fully human evaluation of hundreds of riddles across many models is prohibitively expensive, but (ii) purely automated judging is known to be systematically biased on subtle multi-step reasoning. PPI explicitly embraces the jury as a *noisy measurement device* and uses the gold set to correct it, yielding scores that are both scalable (proxy-labeled large set) and statistically grounded in human correctness.

All artifacts needed to recompute the PPI estimates (human annotations and per-judge outputs for each prediction) are released as a companion dataset on Hugging Face at `rvienne/layton-eval-ppi`.

### 3.3 Gold set calibration

We collect a smaller human-annotated **gold set**, with an annotation budget set so that the total volume of annotations is approximately $\sim 3\times$ the number of riddles in each split. Predictions were selected for human annotation via stratified sampling over **riddle ID**, **provider**, and **model within provider**, to ensure broad coverage and avoid over-representing any single model family. We compute PPI *independently* on each split (`LLM` and `VLM`). Human labels are pure booleans (cast to floats 0.0 or 1.0), while jury scores are discrete-valued averages in $[0, 1]$. For gold items, we compute residuals between these values (the calibration "delta").

### 3.4 Rectified benchmark scoring

We use the estimated residual to rectify scores on the larger unlabeled set, yielding a point estimate that is mathematically anchored to human judgment while still leveraging the scale of automated scoring.

### 3.5 Jury setup (dynamic, provider-aware)

We use a 3-judge panel selected from a fixed pool of 4 candidate frontier judges:

- `gpt-5.1` (reasoning high),
- `gemini-3-pro` (reasoning high),
- `claude-opus-4.5` (with "thinking" enabled; 32k tokens),
- `mistral-large-2512` (with "thinking" disabled).

By default, the jury is {`gpt-5.1`, `gemini-3-pro`, `claude-opus-4.5`}. To mitigate self-preference and same-provider effects, if the evaluated model is from OpenAI, Google (Gemini), or Anthropic, we exclude the corresponding judge and include `mistral-large-2512` instead.

### 3.6 Jury score (discrete average of boolean correctness)

Each judge produces a structured verdict indicating whether (i) the final *answer* is correct and (ii) the *justification* is correct (both booleans). We define the judge-level correctness indicator as their boolean product, i.e., *answer* $\wedge$ *justification.*

The final jury score for an item is the average of this indicator across the 3 judges, yielding a floating-point value with discrete support (e.g., 0, 1/3, 2/3, 1).

### 3.7 Stratified bootstrapping

A key risk in calibration is **regression to the mean**: if the gold set does not match the evaluated model's difficulty profile, rectification can unfairly pull scores toward a global average [15].

We address this with a stratified bootstrap procedure [15, 18]:

1. **Binning:** compute the evaluated model's jury-score distribution.
2. **Matching:** resample (with replacement) from the human-labeled pool to mirror that distribution, creating a "virtual model" calibration set.
3. **Iteration:** repeat (e.g., 10,000 times) to form a distribution of rectified scores.

From this distribution we report:

- a **95% confidence interval** (CI) by taking the 2.5% and 97.5% percentiles of the rectified-score distribution,

- a midpoint **point estimate** $P$ defined as the center of that interval, and
- the CI **half-width** $W$ such that, at 95% certainty, the true rectified performance lies in $[P - W,\, P + W]$.

We compute **rank spread** afterwards, once we obtain such a distribution for each candidate model, by comparing worst-case and best-case ranking scenarios induced by these 95% intervals.

### 3.8 Key technical pillars

- **Jury ensembling:** a 3-judge ensemble drawn from a pool of four frontier judges (with provider-aware selection) to reduce individual variance and mitigate provider-specific biases.
- **Structured justifications:** JSON-formatted rationales to reduce "lucky guesses" and enable more consistent judging.
- **Self-preference mitigation:** dynamic jury selection excluding the evaluated model's provider.
- **Rank sensitivity:** confidence-interval-based rank spread to communicate statistical "jitter."

## 4 Results

### 4.1 LLM split

| Rank | Model | Hints | Score | 95% CI (±) | Rank Spread | Provider |
|---:|---|---:|---:|---:|:---:|---|
| 1 | `gemini-3-flash-high` | 0 | 85.2 | 1.4 | [1] ↔ [2] | gemini |
| 2 | `gemini-3-pro-high` | 0 | 83.9 | 1.1 | [1] ↔ [3] | gemini |
| 3 | `gpt-5.1-2025-11-13-high` | 0 | 83.3 | 0.1 | [2] ↔ [3] | openai |
| 4 | `gpt-5.2-2025-12-11-high` | 0 | 80.4 | 0.3 | [4] ↔ [5] | openai |
| 5 | `claude-opus-4-5-20251101-thinking-32k` | 0 | 79.6 | 0.6 | [4] ↔ [5] | anthropic |
| 6 | `moonshotai-kimi-k2.5-thinking` | 0 | 73.4 | 1.4 | [6] ↔ [6] | together |
| 7 | `claude-opus-4-5-20251101-no-thinking` | 0 | 70.2 | 0.9 | [7] ↔ [7] | anthropic |
| 8 | `moonshotai-kimi-k2-thinking` | 0 | 66.7 | 1.7 | [8] ↔ [8] | together |
| 9 | `mistral-large-2512` | 0 | 48.7 | 1.4 | [9] ↔ [9] | mistral |
| 10 | `qwen-qwen3-vl-235b-a22b-instruct-fp8` | 0 | 39.0 | 1.4 | [10] ↔ [10] | doubleword |

Table 1: Leaderboard on the `LLM` split of **layton-eval** (0-hint setting). Scores are reported with 95% CI half-widths from the stratified bootstrap rectification procedure.

### 4.2 VLM split

| Rank | Model | Hints | Score | 95% CI (±) | Rank Spread | Provider |
|---|---|---|---|---|---|---|
| 1 | `gemini-3-flash-high` | 0 | 46.4 | 1.9 | [1] ↔ [2] | gemini |
| 2 | `gemini-3-pro-high` | 0 | 46.1 | 1.9 | [1] ↔ [2] | gemini |
| 3 | `gpt-5.2-2025-12-11-high` | 0 | 33.4 | 1.9 | [3] ↔ [4] | openai |
| 4 | `gpt-5.1-2025-11-13-high` | 0 | 32.3 | 2.1 | [3] ↔ [5] | openai |
| 5 | `moonshotai-kimi-k2.5-thinking` | 0 | 28.9 | 1.8 | [4] ↔ [5] | together |
| 6 | `claude-opus-4-5-20251101-thinking-32k` | 0 | 25.1 | 1.8 | [6] ↔ [7] | anthropic |
| 7 | `claude-opus-4-5-20251101-no-thinking` | 0 | 23.0 | 1.9 | [6] ↔ [7] | anthropic |
| 8 | `mistral-large-2512` | 0 | 12.6 | 1.3 | [8] ↔ [8] | mistral |
| 9 | `qwen-qwen3-235b-a22b-instruct-2507-tput` | 0 | 9.3 | 1.5 | [9] ↔ [9] | together |

Table 2: Leaderboard on the `VLM` split of **layton-eval** (0-hint setting). Scores are reported with 95% CI half-widths from the stratified bootstrap rectification procedure.

**Interpreting the LLM–VLM gap.** We observe a large performance gap between the `LLM` and `VLM` splits. Part of this difference is expected: visual riddles are often harder to solve in the original games, and visual reasoning is plausibly less common (and less emphasized) in frontier-model training and evaluation.

However, we also suspect an input-level and architectural component: Nintendo DS-era assets can be low-resolution and visually noisy, and current VLM perception stacks may simply not have the "right eyes" for this regime (e.g., resolution limits, aliasing, compression artifacts), which may also be underrepresented in typical pretraining corpora. This suggests that progress toward human-level puzzle-solving may be constrained not only by reasoning ability, but also by structural perceptual affordances of the model architecture.

### 4.3 Interactive data exploration and in-depth analysis

To facilitate exploration beyond the aggregate leaderboard, we provide a web interface to browse the full evaluation data, including overall results, breakdowns by riddle category, and per-riddle results:

`https://vienneraphael.github.io/layton-eval/`

## 5 Reproducibility

All code, prompts, and evaluation scripts needed to reproduce the experiments and to self-report model performance are available in the **layton-eval** repository:

`https://github.com/vienneraphael/layton-eval`

# 6 Cost Report

We report approximate inference costs incurred during evaluation (including the cost of judging where applicable). The costs in Table 3 are the *batched* costs. We relied on providers' batch APIs; for the *estimated unbatched* costs, we assume a default 50% batch economy (i.e., batched cost is $\approx$ 50% of unbatched cost), and a 66% batch economy for Doubleword.

| **Provider** | **Batched Cost (USD)** | **Est. Unbatched (USD)** | **Notes** |
|---|---:|---:|---|
| Anthropic | 104.00 | 208.00 | 2 models (including judge) |
| OpenAI | 63.00 | 126.00 | 2 models (including judge) |
| Gemini | 3.47 | 6.94 | 2 models (including judge) |
| Together | 21.00 | 42.00 | 3 models |
| Mistral | 1.20 | 2.40 | 1 model (including judge) |
| Doubleword | 0.13 | 0.39 | 1 model (no judge; 66% batch economy) |
| **Total** | **192.80** | **385.73** | |

Table 3: Approximate inference costs by provider for this evaluation (batched), with a rough unbatched estimate assuming a 50% batch economy by default (66% for Doubleword).

# 7 Conclusion

**layton-eval** probes lateral, puzzle-like reasoning—in both text-only (`LLM`) and vision-based (`VLM`) settings—and combines scalable automated judging with Prediction-Powered Inference (PPI) to remain anchored to human correctness.

Our results indicate that there is still substantial room for improvement on both splits: even top frontier models systematically fail some riddle categories, suggesting that key forms of non-linear reasoning and robust justification are not yet reliably acquired.

Despite this gap, the relative ranking produced by **layton-eval** is highly consistent with other widely used benchmarks and with the broader community signal from LMArena (now `arena.ai`) [19], reinforcing the benchmark's external validity.

Finally, we view **layton-eval** as a living evaluation: we will continue to expand coverage and update the leaderboard as new models emerge, and we look forward to tracking how future generations close the remaining reasoning and perception gaps.

# Acknowledgements

This work uses material from the Professor Layton Wiki at Fandom and all subsequent datasets are licensed under the Creative Commons Attribution–ShareAlike License.

# References

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

[3] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[4] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021.

[5] Rishi Bommasani, Drew A. Hudson, and Percy Liang et al. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*, 2022.

[6] Aarohi Srivastava and many others. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[7] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*, 2023.

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

[9] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[10] Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2406.07791*, 2024.

[11] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-a-Judges. *arXiv preprint arXiv:2406.12624*, 2024.

[12] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *arXiv preprint arXiv:2406.18403*, 2024.

[13] Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing Judging Bias in Large Reasoning Models: An Empirical Study. *arXiv preprint arXiv:2504.09946*, 2025.

[14] Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Prediction-Powered Inference. *arXiv preprint arXiv:2301.09633*, 2023.

[15] Anastasios N. Angelopoulos and Stephen Bates. Bootstrapped Prediction-Powered Inference. *arXiv preprint*, 2024.

[16] Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient Prediction-Powered Inference. *arXiv preprint arXiv:2311.01453*, 2023.

[17] Grégoire Martinon, Alexandra Lorenzo de Brionne, Jérôme Bohard, Antoine Lojou, Damien Hervault, and Nicolas J-B. Brunel. Towards a Rigorous Evaluation of RAG Systems: The Challenge of Due Diligence. *arXiv preprint arXiv:2507.21753*, 2025.

[18] Tijana Zrnic and Emmanuel J. Candès. Active Statistical Inference. *arXiv preprint arXiv:2403.03208*, 2024.

[19] LMArena. `arena.ai` (accessed 2026). `https://arena.ai`.