RPRA: Predicting an LLM-Judge for Efficient but Performant Inference

Dylan R. Ashley,[1,2,3,4,5] **Gaël Le Lan,**[1] **Changsheng Zhao,**[1] **Naina Dhingra,**[1]
Zhipeng Cai,[1] **Ernie Chang,**[1] **Mingchen Zhuge,**[1,5]
Yangyang Shi[1] **Vikas Chandra,**[1] **and Jürgen Schmidhuber**[2,3,4,5]

[1]Meta Platforms, Inc. [2]IDSIA [3]USI [4]SUPSI [5]KAUST

King Abdullah University of Science and Technology

Dalle Molle Institute for Artificial Intelligence

Abstract. Large language models (LLMs) face a fundamental trade-off between computational efficiency (e.g., number of parameters) and output quality, especially when deployed on computationally limited devices such as phones or laptops. One way to address this challenge is by following the example of humans and have models ask for help when they believe they are incapable of solving a problem on their own; we can overcome this trade-off by allowing smaller models to respond to queries when they believe they can provide good responses, and deferring to larger models when they do not believe they can. To this end, in this paper, we investigate the viability of Predict-Answer/Act (PA) and Reason-Predict-Reason-Answer/Act (RPRA) paradigms where models predict—prior to responding—how an LLM judge would score their output. We evaluate three approaches: zero-shot prediction, prediction using an in-context report card, and supervised fine-tuning. Our results show that larger models (particularly reasoning models) perform well when predicting generic LLM judges zero-shot, while smaller models can reliably predict such judges well after being fine-tuned or provided with an in-context report card. Altogether, both approaches can substantially improve the prediction accuracy of smaller models, with report cards and fine-tuning achieving mean improvements of up to 55% and 52% across datasets, respectively. These findings suggest that models can learn to predict their own performance limitations, paving the way for more efficient and self-aware AI systems.

Predict-Answer/Act & Reason-Predict-Reason-Answer/Act

We formalize **pre-hoc prediction** of an LLM/agentic judge as two paradigms: *Predict-Answer/Act (PA)*, where a model predicts how a judge would score its response before generating it; and *Reason-Predict-Reason-Answer/Act (RPRA)*, which extends PA by having the model reason before predicting and again before responding. In practice, a PA/RPRA model directly responds to queries it is confident it can answer well, and **routes to a larger model** otherwise.

LLM/agentic judges are highly favorable here as they correlate strongly with **human evaluations**, support **flexible evaluation criteria**, and can be easily adapted to **new alignment problems** by modifying their prompts.

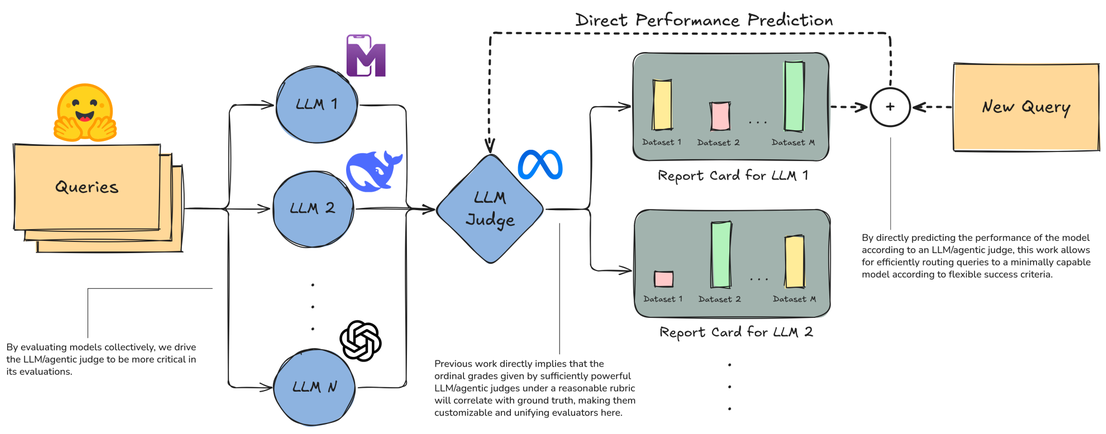


Figure 1: Queries are sent to multiple LLMs whose responses are jointly evaluated by an LLM/agentic judge. The judge's responses are converted into report cards that capture a model's performance across query types. At inference, the model predicts its own capability on a new query—enabling cost-efficient routing.

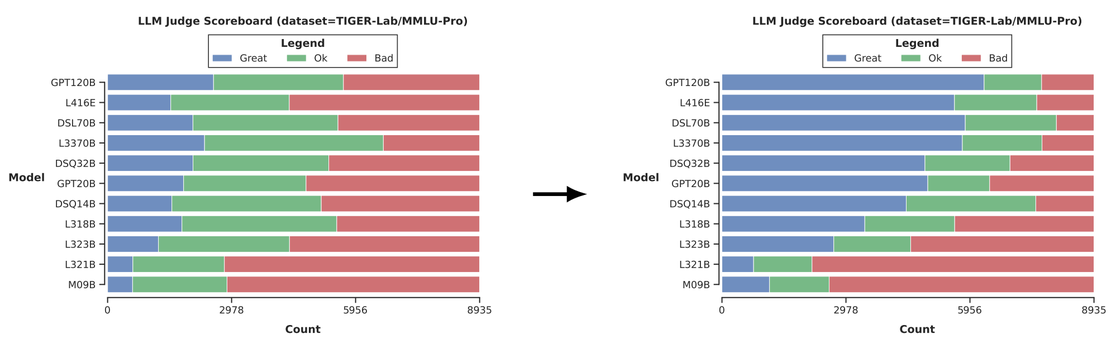
Experimental Setup

We experiment with **five datasets** commonly used in the literature: **MedQA**, **LongFact**, **AIME 2024**, **SciCode**, and **MMLU-Pro**. **Llama 3.3 70B** serves as our judge, grading all model responses for a query *simultaneously* against a predefined rubric (here we use a naive one). We experiment with the following models:

Shorthand	Full Name	# Parameters	Reasoning	Reference
M09B	MobileLLM 0.9B	0.9 Billion	No	Liu et al., 2024
L318B	Llama 3.1 8B Instruct	8.03 Billion	No	Grattafiori et al., 2024
L321B	Llama 3.2 1B Instruct	1.24 Billion	No	Meta AI, 2024
L323B	Llama 3.2 3B Instruct	3.21 Billion	No	Meta AI, 2024
L3370B	Llama 3.3 70B Instruct	70.6 Billion	No	Meta AI, 2024
L416E	Llama 4 Scout 17B 16E Instruct	109 Billion	No	Meta AI, 2025
DSQ14B	DeepSeek R1 Distilled Qwen 14B	14.8 Billion	Yes	Guo et al., 2025
DSQ32B	DeepSeek R1 Distilled Qwen 32B	32.8 Billion	Yes	Guo et al., 2025
DSL70B	DeepSeek R1 Distilled Llama 70B	70.6 Billion	Yes	Guo et al., 2025
GPT20B	GPT OSS 20B	21.5 Billion	Yes	OpenAI, 2025
GPT120B	GPT OSS 120B	120 Billion	Yes	OpenAI, 2025

Table 1: The eleven (11) models considered for our experiments. Parameter counts are taken from the HuggingFace Safetensor values to ensure equal treatment between models.

Joint evaluation yields diversity. Grading multiple responses at once forces the judge to discriminate; grading independently leads to less diversity between the models:



Independent (ablation) *Joint (main experiments)*

Figure 5: Judge score distributions on MMLU-Pro. Independent grading (left) leads to less diversity between the models; joint grading (right) elicits wide, discriminative distributions.

Solution Methods

Our first approach provides models with a *report card*: a summary of the model's historical performance across datasets. This requires **no additional training** and applies to any model, including closed-weight systems.

Our second approach fine-tunes models specifically for performance prediction—enabling PA *without* a report card. Training is needed, but inference no longer pays for report-card tokens. We build training data via the **hindsight trick**, relabeling examples with the judge's scores, then apply supervised fine-tuning.

```
You were tested on the AIME 2024 dataset, which features high-level competition mathematics problems from the American
Invitational Mathematics Examination that require sophisticated mathematical problem-solving skills and multi-step reasoning to
arrive at precise numerical answers. On this dataset, most of your responses were judged to be "{HuggingFaceH4/aime_2024}". This
means that your ability to solve competition-level mathematics problems that require sophisticated problem-solving skills and
multi-step reasoning is "{HuggingFaceH4/aime_2024}".

You were tested on the LongFact dataset, which presents concept-based queries that demand comprehensive, factual responses on
topics like 20th-century events, US foreign policy, accounting principles, and architecture, testing your ability to provide
detailed, accurate information across broad knowledge domains. On this dataset, most of your responses were judged to be
"{claserken/longfact}". This means that your ability to regurgitate factual trivia is "{claserken/longfact}".

You were tested on the MedQA dataset, which focuses specifically on medical question-answering with free-form multiple-choice
questions derived from professional medical board exams, testing clinical knowledge and diagnostic reasoning. On this dataset,
most of your responses were judged to be "{bigbio/med_qa}". This means that your ability to diagnose medical conditions is
"{bigbio/med_qa}".

You were tested on the MMLU-Pro dataset, which contains challenging undergraduate-level multiple-choice questions across diverse
academic disciplines including mathematics, science, history, and humanities, designed to test advanced reasoning and knowledge
beyond standard benchmarks. On this dataset, most of your responses were judged to be "{TIGER-Lab/MMLU-Pro}". This means that
your ability to answer general undergraduate-level examination questions is "{TIGER-Lab/MMLU-Pro}". You were further scored on
different academic subjects within the MMLU-Pro dataset, with the following results:
- Philosophy: {TIGER-Lab/MMLU-Pro/philosophy}
- Mathematics: {TIGER-Lab/MMLU-Pro/math}
- Economics: {TIGER-Lab/MMLU-Pro/economics}
- Engineering: {TIGER-Lab/MMLU-Pro/engineering}
- Physics: {TIGER-Lab/MMLU-Pro/physics}
- Biology: {TIGER-Lab/MMLU-Pro/biology}
- Business: {TIGER-Lab/MMLU-Pro/business}
- History: {TIGER-Lab/MMLU-Pro/history}
- Law: {TIGER-Lab/MMLU-Pro/law}
- Health: {TIGER-Lab/MMLU-Pro/health}
- Chemistry: {TIGER-Lab/MMLU-Pro/chemistry}
- Computer Science: {TIGER-Lab/MMLU-Pro/computer_science}
- Other subjects: {TIGER-Lab/MMLU-Pro/other}
These reflect how your ability to answer undergraduate-level examination questions is affected by the subject.

You were tested on the SciCode dataset, which evaluates scientific computing and programming capabilities through complex
problem-solving tasks that require understanding of scientific concepts, mathematical reasoning, and code implementation across
various scientific domains. On this dataset, most of your responses were judged to be "{SciCode1/SciCode}". This means that your
ability to write source code for scientific work is "{SciCode1/SciCode}".
```

Prompt 5.The full report card template used in our main experiments.

Results

The key observations from the zero-shot results are that **(1)** prediction performance varied wildly based on the dataset used, **(2)** reasoning models exhibited an often better ability to estimate their own performance, and **(3)** smaller models typically performed at or below the random guess accuracy. The **report card** is particularly useful for the smaller and non-reasoning models, enabling them to work in a PA paradigm. The prediction quality of the **fine-tuned** models is generally at or above the performance of even the report card-based predictions.

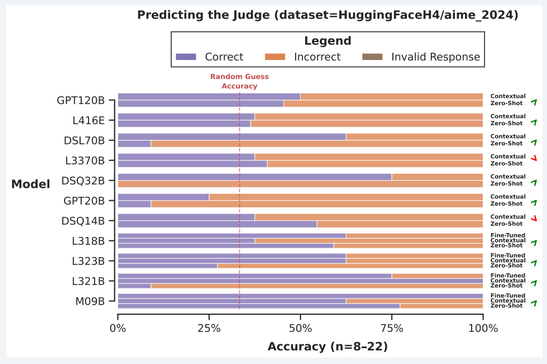


Figure 7: Prediction accuracy on AIME 2024—green arrows indicate improvement over zero-shot. Note the dramatic improvement for the smallest models.

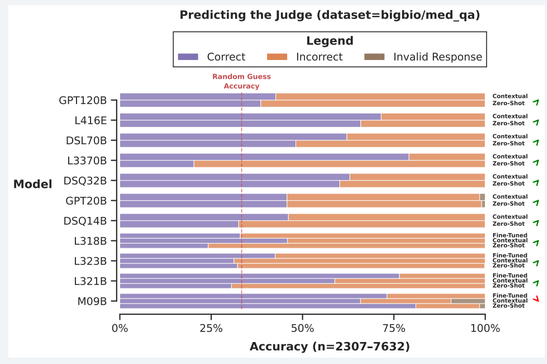


Figure 8: Prediction accuracy on MedQA. Note how dramatic the improvement is for even big non-reasoning models (i.e., L3370B).

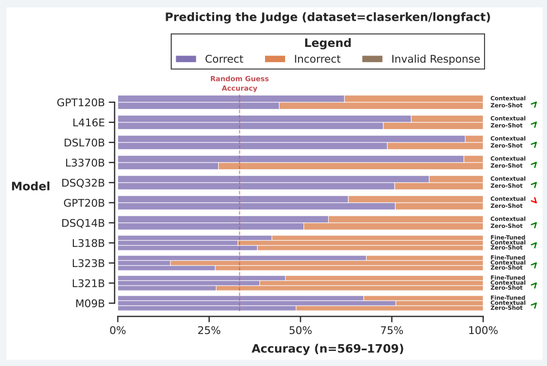


Figure 9: Prediction accuracy on LongFact. Even on a dataset where all models do well, predictions improve in the contextual/fine-tuning setting.

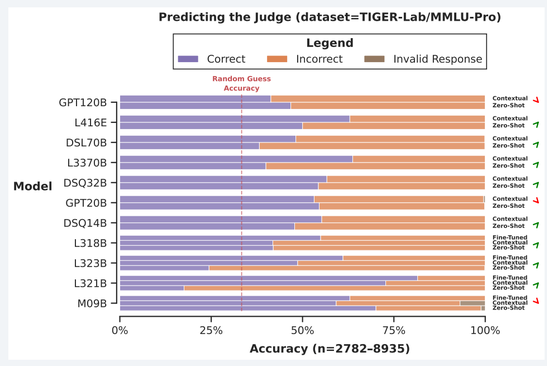


Figure 10: Prediction accuracy on MMLU-Pro. Per-category breakdown of the zero-shot results is in the Appendix.

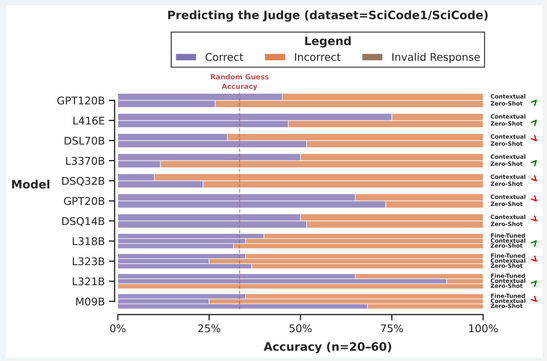


Figure 11: Prediction accuracy on SciCode—the only dataset where the contextual/fine-tuning approach frequently didn't help. We hypothesize the judge was not sophisticated enough to evaluate models properly here.

Dataset	M09B	L321B	L323B	L318B
MedQA	-.09	+.46	+.10	+.09
LongFact	+.19	+.19	+.41	+.04
AIME 2024	+.23	+.66	+.35	+.03
MMLU-Pro	-.07	+.64	+.37	+.13
SciCode	-.33	+.65	-.02	+.08
Mean	-.02	**+.52**	+.24	+.07

Table 2: Summary of the improvement in the prediction accuracy for the fine-tuned models as compared with the zero-shot setting.

Future Work

- **End-to-end routing.** Complete routing systems that realize cost-quality trade-offs in practice.
- **Multi-turn interactions.** Later queries may demand more capable models, requiring considerate routing.
- **Alignment-aware rubrics.** Predicting judges that encode safety, style, or other alignment requirements.
- **Unified prompts.** Integrate predictions into the models so one prompt predicts and generates.
- **Richer training signals.** Reinforcement learning from judge feedback, or human-in-the-loop corrections, as alternatives to the hindsight trick.