Not All Proofs Are Equal: Evaluating LLM Proof Quality Beyond Correctness



Ivo Petrov



Dimitar I. Dimitrov



Jasper Dekoninck



Martin Vechev

INSAIT

ETH zürich

SRILAB

NEURAL INFORMATION PROCESSING SYSTEMS

Background

AI for Mathematics

July 21, 2025 Research

Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

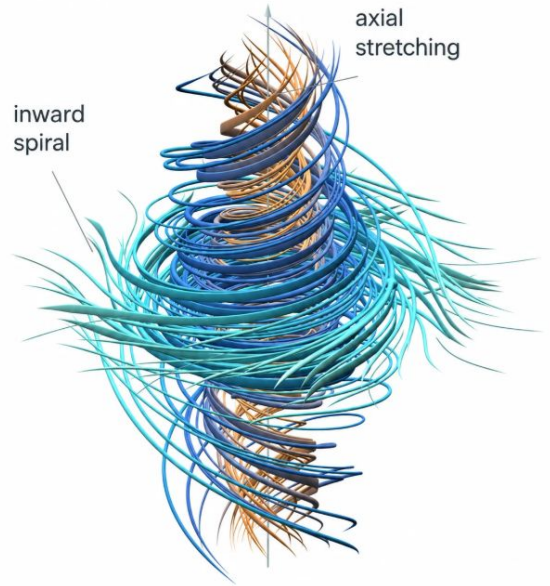
Thang Luong and Edward Lockhart

Share



September 8, 2026 Research Publication

On the Navier–Stokes Millennium Prize Problem



Proof Quality

Takeaways from the Navier-Stokes resolution:



Terence Tao
@tao@mathstodon.xyz

A concrete example of how AI advances in solving key open problems could inihibit the future development of a mathematical field can be found in the global regularity problem for the incompressible Navier-Stokes equations (as well as its counterpart for the Euler equations). Until recently, this problem was on track to be one of the most promising examples of an AI-assisted success story, in which AI tools could be used both to solve the problem and to set the stage for the next round of progress in the field. While the problems remain open for now, there is now an increasingly realistic scenario in which a primarily AI-generated solution to the problem appears, but in a fashion that contaminates the problem as a source of further advances.

AI solved one of math's hardest problems. Humanity learned nothing (so far)

"It's a terribly written paper," Buckmaster said.

"The paper doesn't explain what parts are important. What parts are routine? How does the idea feed into other places?" added Gómez-Serrano.

ProofRank

Towards Evaluating LLM Proof Quality

Main Challenges

“Quality” is inherently subjective

Main Challenges

“Quality” is inherently subjective

“Quality” depends on the reader

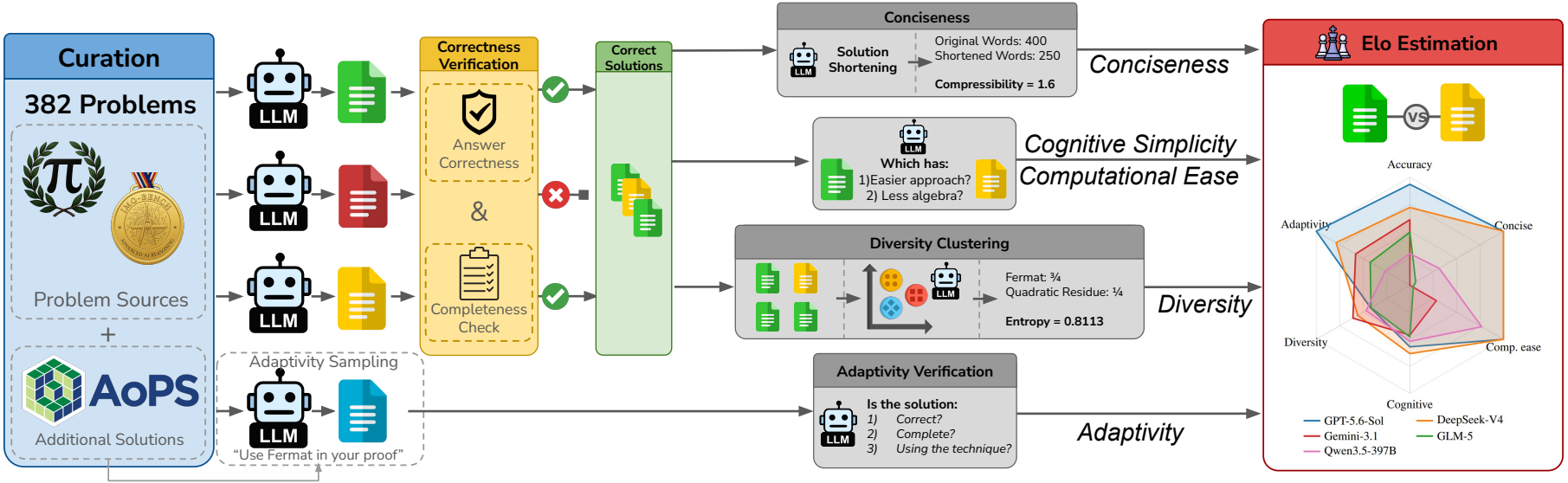
Main Challenges

“Quality” is inherently subjective

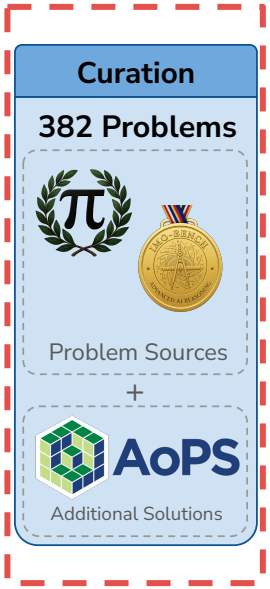
“Quality” depends on the reader

Ranking models should be scalable

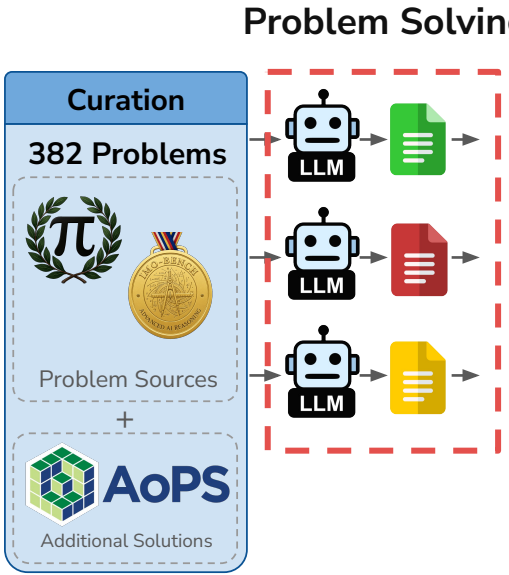
ProofRank Overview



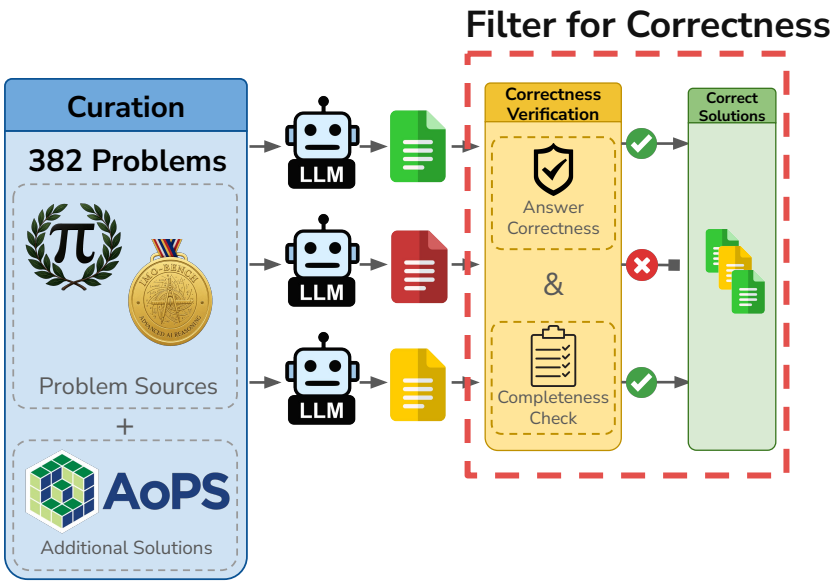
ProofRank Overview



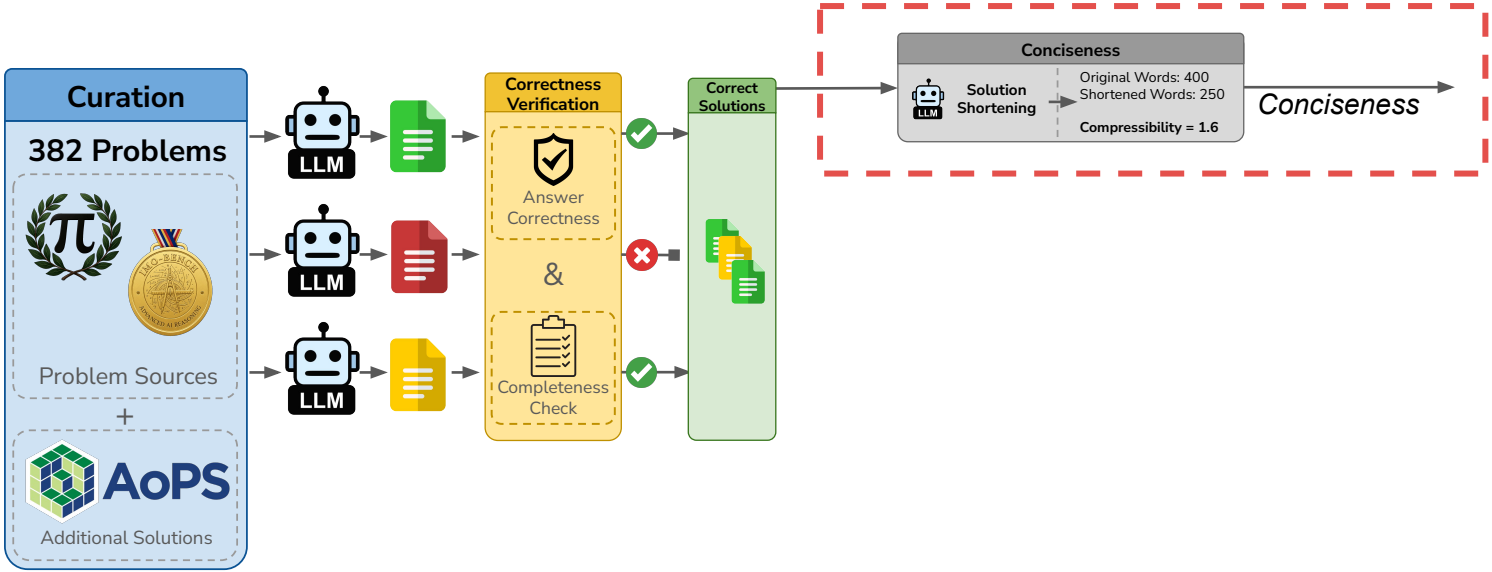
ProofRank Overview



ProofRank Overview

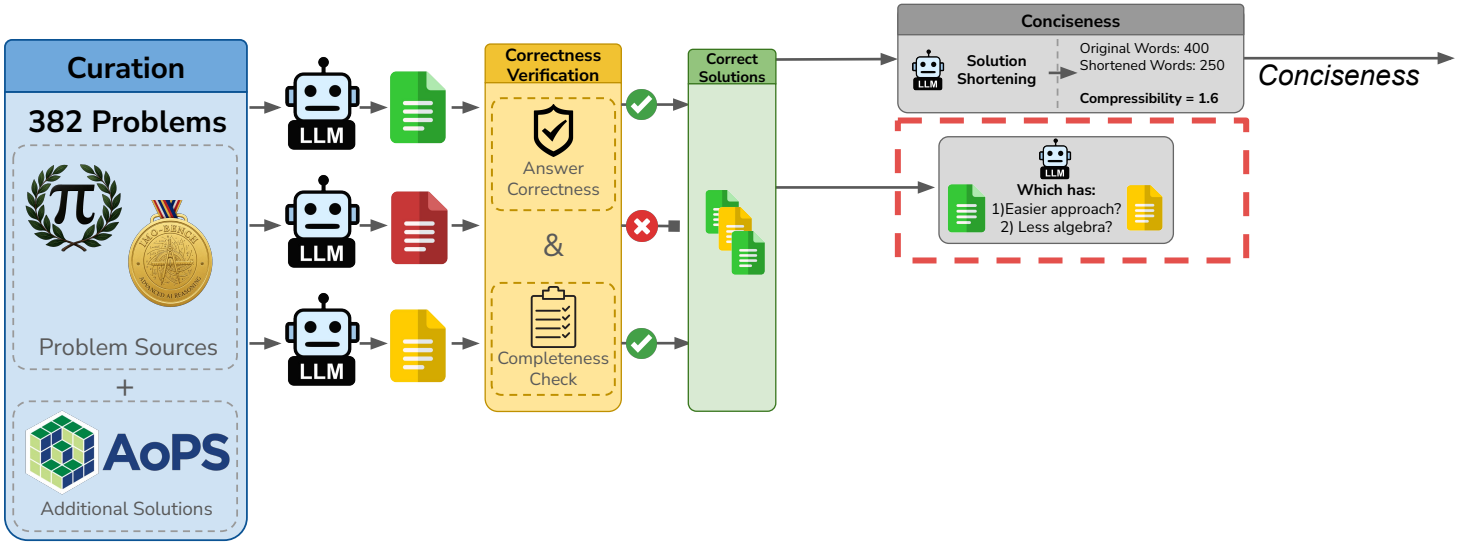


ProofRank Overview

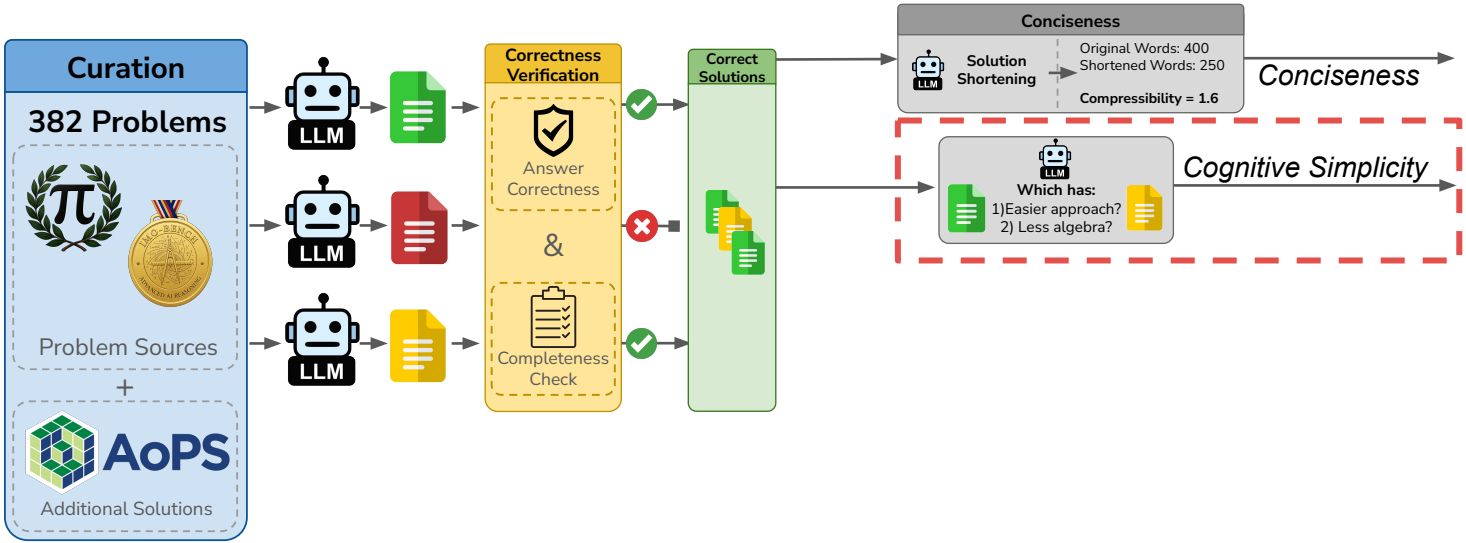


Conciseness: How brief can LLMs present their arguments?

ProofRank Overview

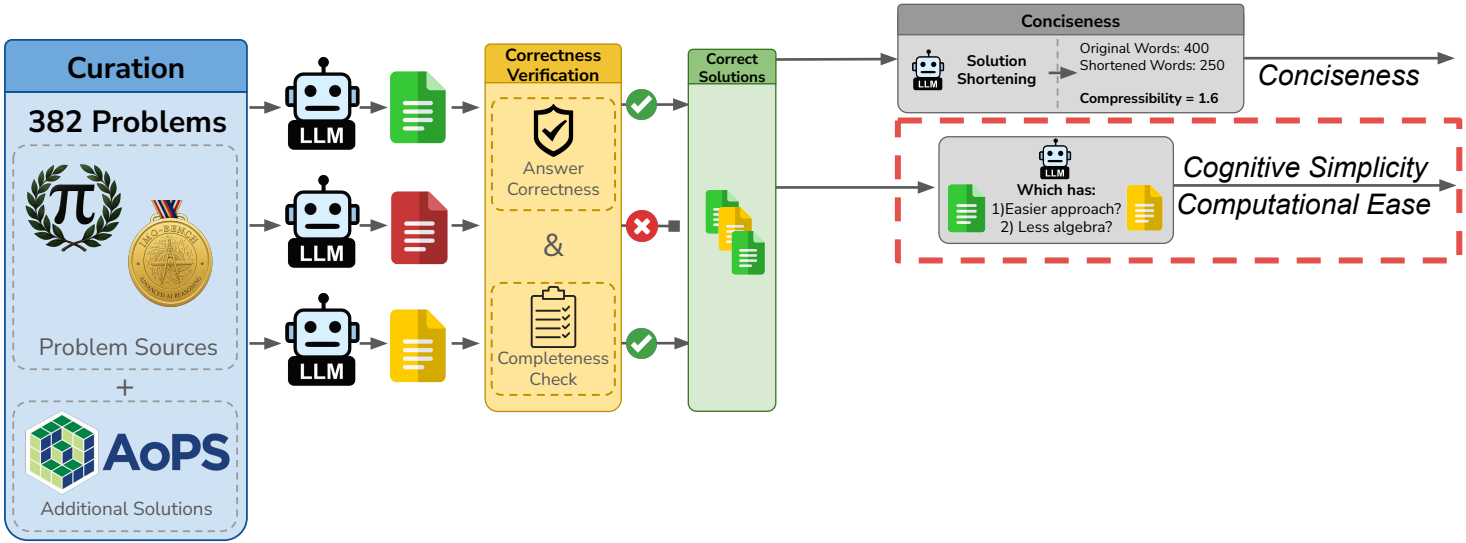


ProofRank Overview



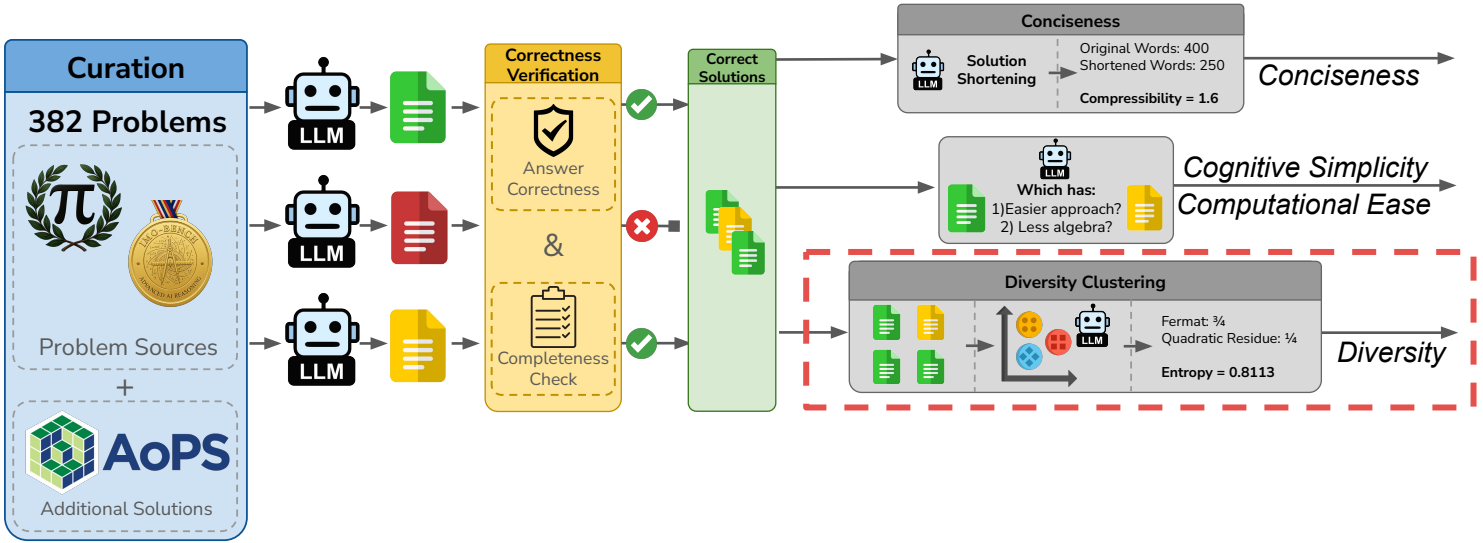
Cognitive Simplicity: Which LLM used more accessible methods?

ProofRank Overview



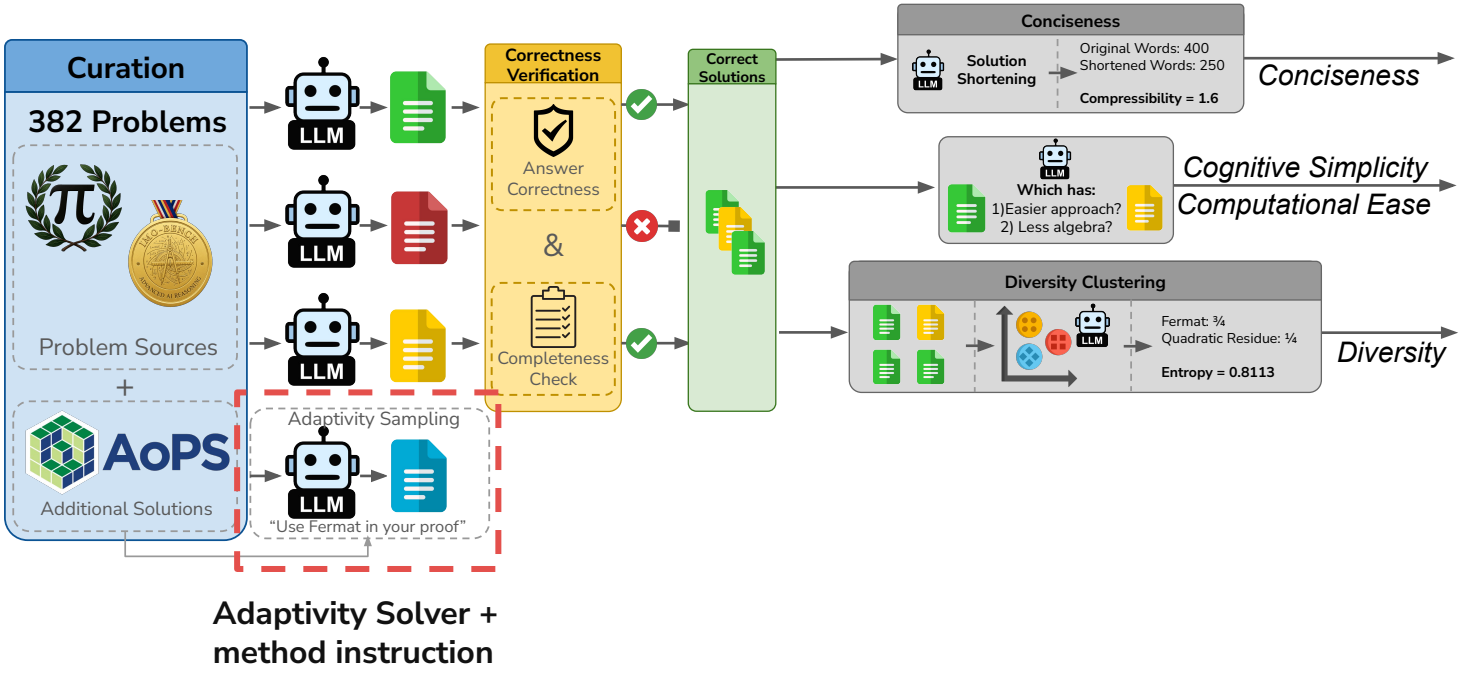
Computational Ease: Which proof used less algebra?

ProofRank Overview

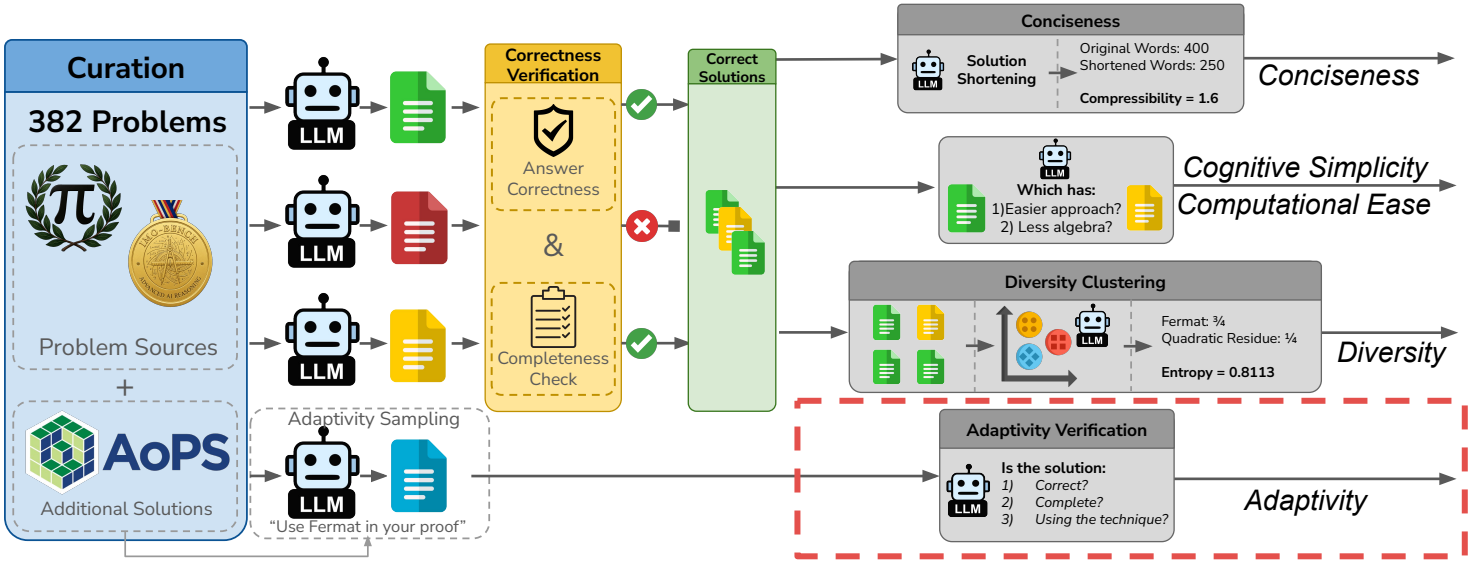


Diversity: Do LLMs naturally produce different proof approaches?

ProofRank Overview

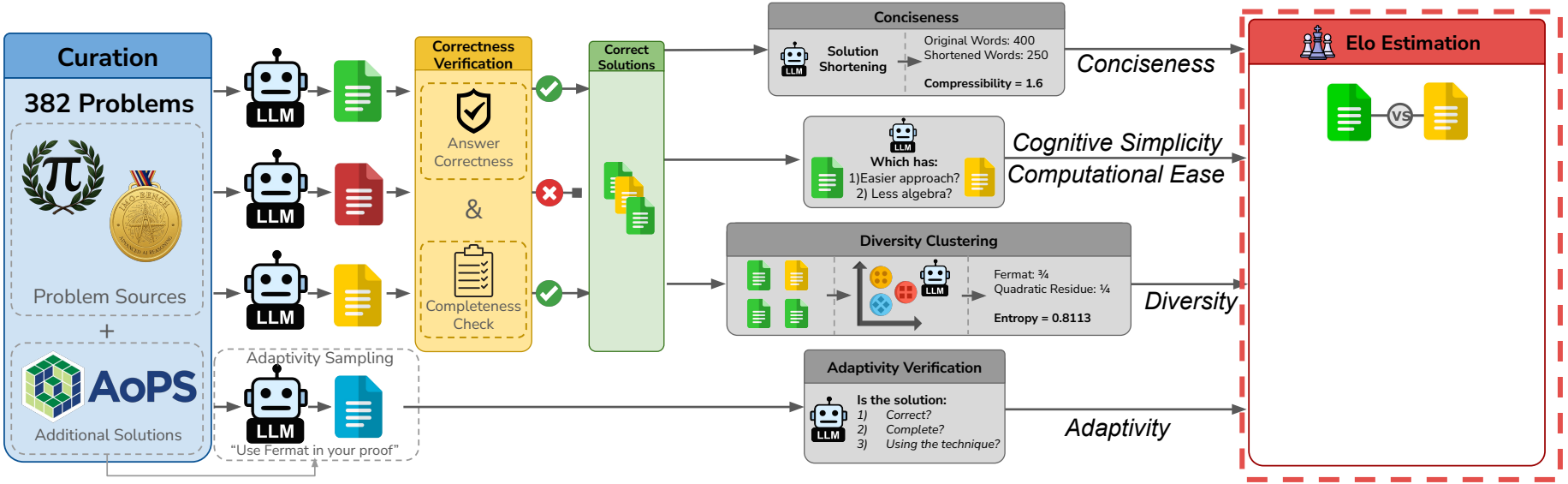


ProofRank Overview

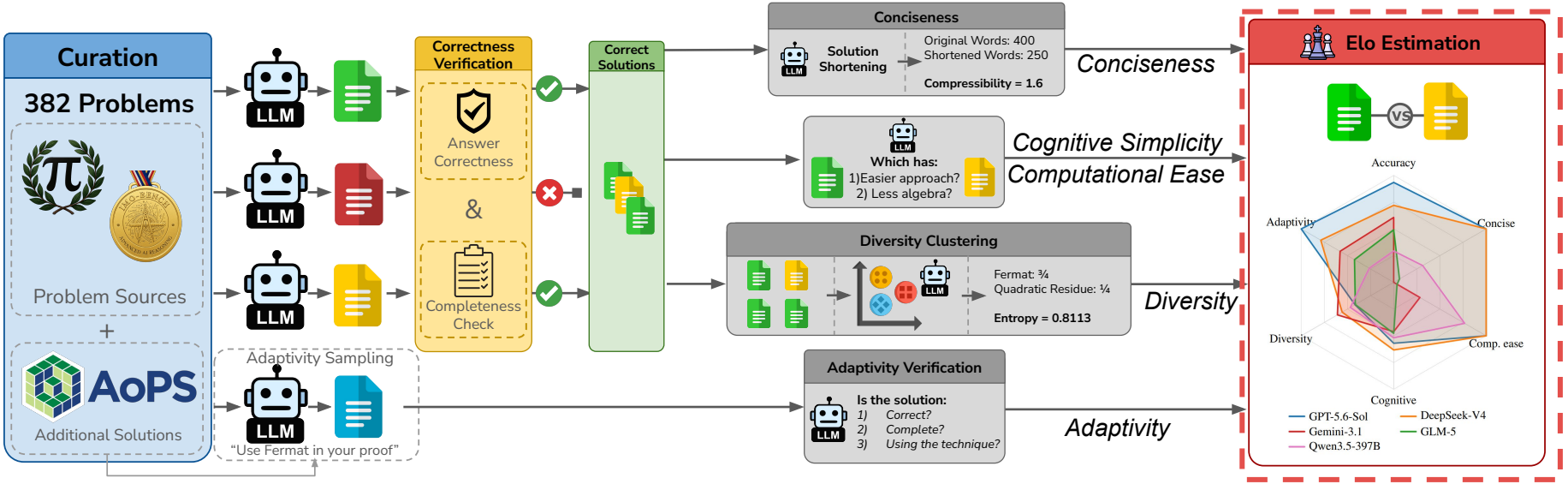


Adaptivity: Can LLMs successfully follow proof directions?

ProofRank Overview



ProofRank Overview



Results

Main Results

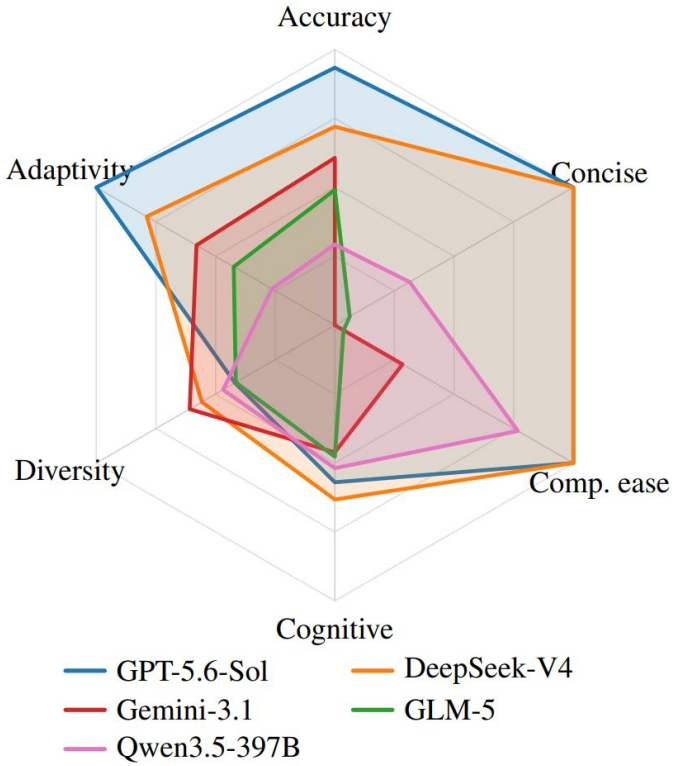
- **11 frontier models**

Main Results

- 11 frontier models
- Each evaluated on >= ~100 proofs

Main Results

- **11 frontier models**
- **Each evaluated on >= ~100 proofs**



LLMs have **different** proof **quality** signatures

Main Results

	Conciseness	Computational Ease	Cognitive Simplicity	Diversity	Adaptivity	Accuracy
GPT-5.6-SOL	1944	1609	1256	1135	**1619**	**93.4%**
GPT-5.4	1295	1566	**1338**	1236	1484	86.3%
DEEPSEEK-V4-PRO	**1973**	**1636**	1306	1246	1431	72.0%
GEMINI-3.7-FLASH	463	852	*1096*	1227	1314	71.2%
GEMINI-3.1-PRO	*165*	1027	1169	1287	1264	60.7%
KIMI-K2.5-THINK	1463	1215	1098	1253	1067	53.5%
GLM-5	850	*829*	1182	1132	1139	49.1%
DEEPSEEK-V3.2	1144	988	1266	**1311**	1049	47.9%
GPT-OSS-120B	1383	1064	1098	1179	*999*	41.1%
STEP-3.5-FLASH	1345	1151	1145	1194	1168	34.3%
GROK-4.1-FAST	1407	990	1184	1129	1031	31.4%
GEMINI-3-FLASH	1117	1260	1247	*1094*	1025	29.8%
QWEN3.5-397B-A17B	1051	1414	1215	1175	1011	*29.3%*
GEMINI-3-FLASH-LOW	1217	937	1196	1081	975	16.5%
QWEN3-8B	750	1209	1324	1067	939	8.1%

Main Results

	Conciseness	Computational Ease	Cognitive Simplicity	Diversity	Adaptivity	Accuracy
GPT-5.6-SOL	1944	1609	1256	1135	**1619**	**93.4%**
GPT-5.4	1295	1566	**1338**	1236	1484	86.3%
DEEPSEEK-V4-PRO	**1973**	**1636**	1306	1246	1431	72.0%
GEMINI-3.7-FLASH	463	852	*1096*	1227	1314	71.2%
GEMINI-3.1-PRO	*165*	1027	1169	1287	1264	60.7%
KIMI-K2.5-THINK	1463	1215	1098	1253	1067	53.5%
GLM-5	850	*829*	1182	1132	1139	49.1%
DEEPSEEK-V3.2	1144	988	1266	**1311**	1049	47.9%
GPT-OSS-120B	1383	1064	1098	1179	*999*	41.1%
STEP-3.5-FLASH	1345	1151	1145	1194	1168	34.3%
GROK-4.1-FAST	1407	990	1184	1129	1031	31.4%
GEMINI-3-FLASH	1117	1260	1247	*1094*	1025	29.8%
QWEN3.5-397B-A17B	1051	1414	1215	1175	1011	*29.3%*
GEMINI-3-FLASH-LOW	1217	937	1196	1081	975	16.5%
QWEN3-8B	750	1209	1324	1067	939	8.1%

The **most accurate** models also rate **highest**

Main Results

	Conciseness	Computational Ease	Cognitive Simplicity	Diversity	Adaptivity	Accuracy
GPT-5.6-SOL	1944	1609	1256	1135	**1619**	**93.4%**
GPT-5.4	1295	1566	**1338**	1236	1484	86.3%
DEEPSEEK-V4-PRO	**1973**	**1636**	1306	1246	1431	72.0%
GEMINI-3.7-FLASH	463	852	*1096*	1227	1314	71.2%
GEMINI-3.1-PRO	*165*	1027	1169	1287	1264	60.7%
KIMI-K2.5-THINK	1463	1215	1098	1253	1067	53.5%
GLM-5	850	*829*	1182	1132	1139	49.1%
DEEPSEEK-V3.2	1144	988	1266	**1311**	1049	47.9%
GPT-OSS-120B	1383	1064	1098	1179	*999*	41.1%
STEP-3.5-FLASH	1345	1151	1145	1194	1168	34.3%
GROK-4.1-FAST	1407	990	1184	1129	1031	31.4%
GEMINI-3-FLASH	1117	1260	1247	*1094*	1025	29.8%
QWEN3.5-397B-A17B	1051	1414	1215	1175	1011	*29.3%*
GEMINI-3-FLASH-LOW	1217	937	1196	1081	975	16.5%
QWEN3-8B	750	1209	1324	1067	939	8.1%

Stronger **Gemini** models are **extremely verbose**

Main Results

	Conciseness	Computational Ease	Cognitive Simplicity	Diversity	Adaptivity	Accuracy
GPT-5.6-SOL	1944	1609	1256	1135	**1619**	**93.4%**
GPT-5.4	1295	1566	**1338**	1236	1484	86.3%
DEEPSEEK-V4-PRO	**1973**	**1636**	1306	1246	1431	72.0%
GEMINI-3.7-FLASH	463	852	*1096*	1227	1314	71.2%
GEMINI-3.1-PRO	*165*	1027	1169	1287	1264	60.7%
KIMI-K2.5-THINK	1463	1215	1098	1253	1067	53.5%
GLM-5	850	*829*	1182	1132	1139	49.1%
DEEPSEEK-V3.2	1144	988	1266	**1311**	1049	47.9%
GPT-OSS-120B	1383	1064	1098	1179	*999*	41.1%
STEP-3.5-FLASH	1345	1151	1145	1194	1168	34.3%
GROK-4.1-FAST	1407	990	1184	1129	1031	31.4%
GEMINI-3-FLASH	1117	1260	1247	*1094*	1025	29.8%
QWEN3.5-397B-A17B	1051	1414	1215	1175	1011	*29.3%*
GEMINI-3-FLASH-LOW	1217	937	1196	1081	975	16.5%
QWEN3-8B	750	1209	1324	1067	939	8.1%

Cognitive simplicity and diversity are **most uniform**

Steerability

Can these metrics be guided towards through prompting?

Steerability

Can these metrics be guided towards through prompting?

- **Conciseness:** Ask for briefer (and much briefer) proofs

Steerability

Can these metrics be guided towards through prompting?

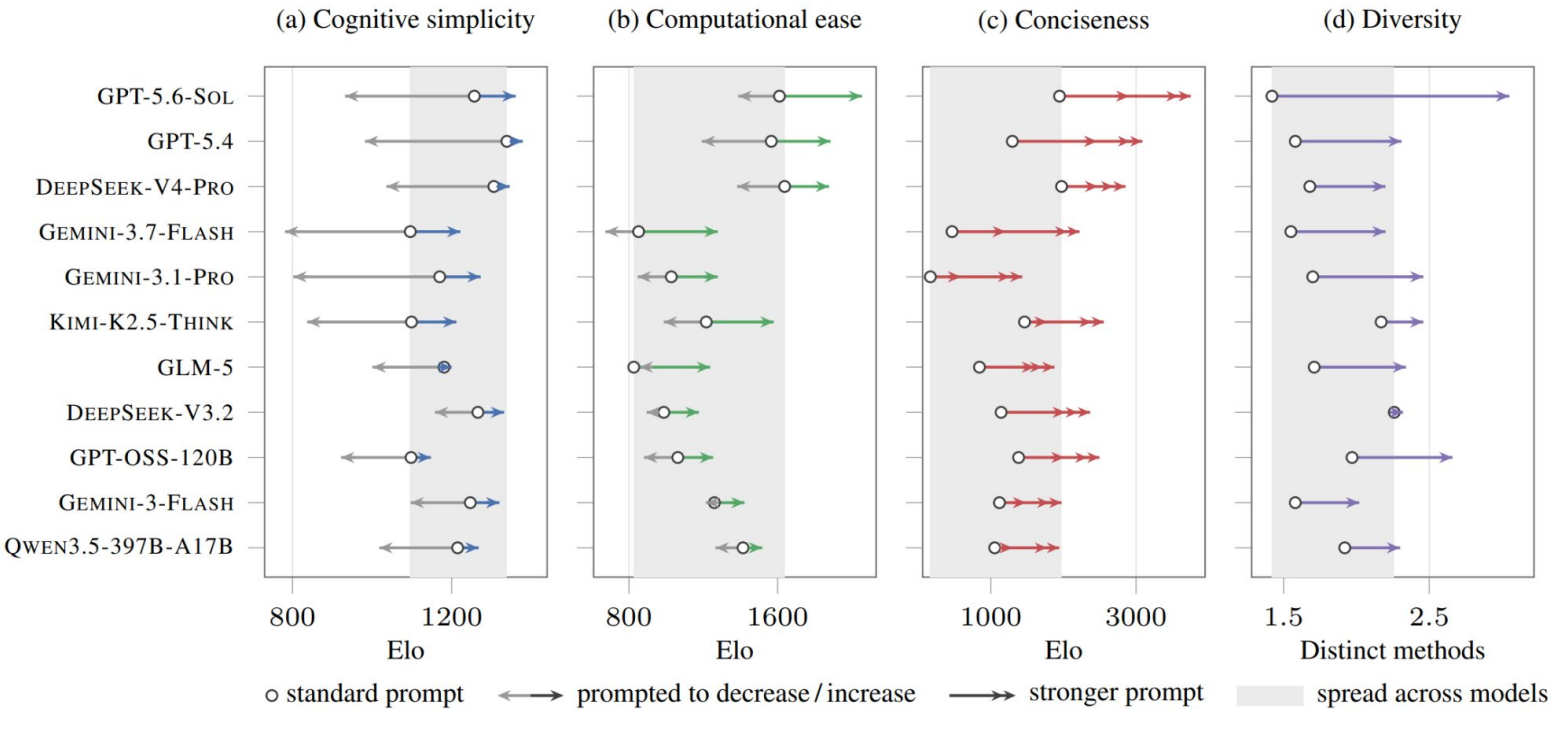
- **Conciseness:** Ask for briefer (and much briefer) proofs
- **Computational ease/Cognitive simplicity:** Ask for simpler/more complex proofs

Steerability

Can these metrics be guided towards through prompting?

- **Conciseness:** Ask for briefer (and much briefer) proofs
- **Computational ease/Cognitive simplicity:** Ask for simpler/more complex proofs
- **Diversity:** Ask for 4 different solutions simultaneously (measures forced diversity)

Steerability



(Almost) **Every** model + metric **can be steered**

Steerability

	Prompt	ΔElo	ΔAcc (%)		Inv.	ρ
		median	median	worst		
Cognitive	easier	+66	−1.8	−14.7	7	0.88
	harder	−262	−1.1	−11.8	16	0.57
Computation	lighter	+250	−2.3	−13.4	12	0.79
	heavier	−181	+0.5	−13.4	5	0.94
Conciseness	mild	+634	−5.5	−15.6	9	0.85
	strong	+1114	−10.9	−31.7	10	0.82

Model ranking under the **same** prompt is **mostly preserved**

Steerability

	Prompt	ΔElo	ΔAcc (%)		Inv.	ρ
		median	median	worst		
Cognitive	easier	+66	−1.8	−14.7	7	0.88
	harder	−262	−1.1	−11.8	16	0.57
Computation	lighter	+250	−2.3	−13.4	12	0.79
	heavier	−181	+0.5	−13.4	5	0.94
Conciseness	mild	+634	−5.5	−15.6	9	0.85
	strong	+1114	−10.9	−31.7	10	0.82

Steering can, however, come at an **utility cost**

Additional Results

Human vs LLM solutions

- ~ 24% of LLM solutions use techniques not found in the human set

- Most unmatched solutions appear as **obscure techniques** or **computation-heavy** approaches

Additional Results

Human vs LLM solutions

- ~ 24% of LLM solutions use techniques not found in the human set

- Most unmatched solutions appear as **obscure techniques** or **computation-heavy** approaches

Judging Incorrect Proofs?

- **Weaker models score higher** all-around as simpler, less clear, high-level arguments are compared against correct proofs

- Filtering for correctness is **crucial**



Thank You!