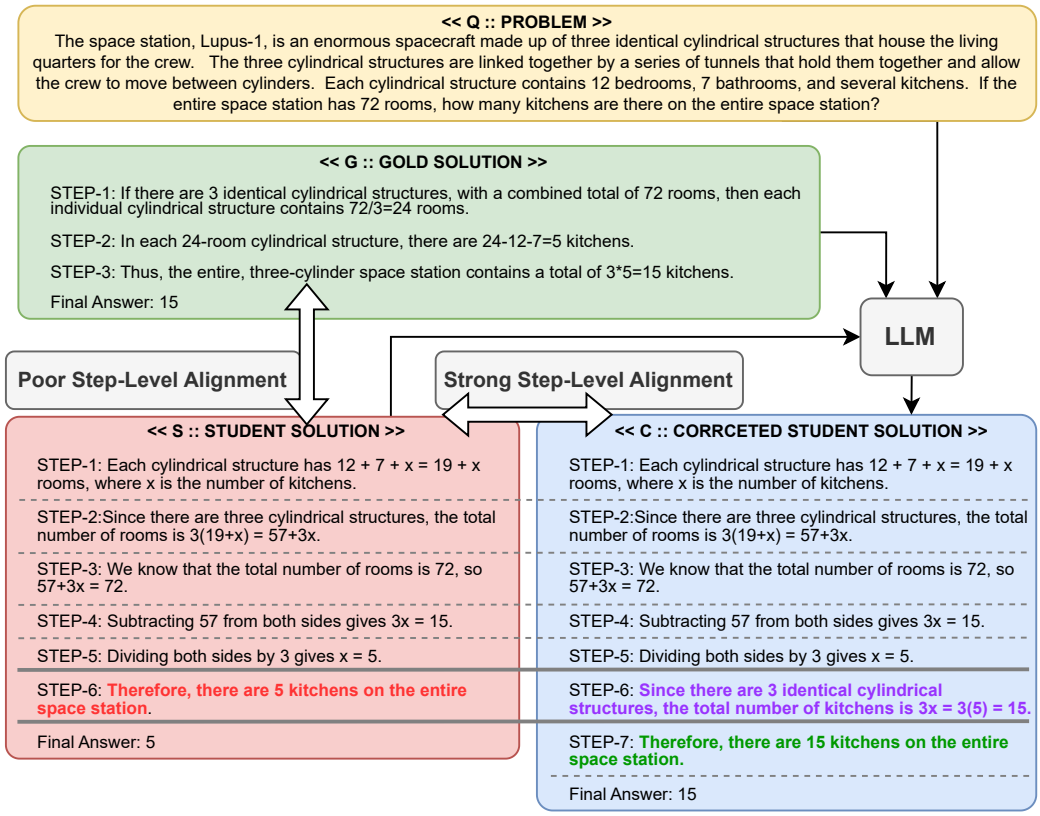
LLMs cannot spot math errors, even when allowed to peek into the solution

KV Aditya Srivatsa, Kaushal Kumar Maurya, Ekaterina Kochmar
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

EMNLP 2025 Suzhou, China 中国苏州

Mohamed bin Zayed University of Artificial Intelligence



Problem Statement

Error localization in reasoning chains is a crucial precursor to effective evaluation and automated tutoring.

RQ-1: Can LLMs accurately predict the first error step in an incorrect math solution using the reference solution (**w-GS**)?

RQ-2: Can localization improve using an intermediate corrected solution (**w-Cor**)?

Method

Data: LLMs' error localization performance tested on annotated stepwise student solutions from the VtG [1] and PRM800K [2] datasets.

Three experimental settings to predict the first error step ID in the student solution:

(1) **w/o-S <Q,S>:** Input contains only problem text (Q) and incorrect student solution (S).

(2) **w-GS <Q,S,G>:** Input contains Q, S, and a gold annotated reference solution (G).

(3) **w-Cor <Q,S,C>:** Input contains Q, S, and a corrected version of S (C) (See below).

Corrected Student Solution (C):

1. Generated using respective LLMs used for localization later. Prompted to correct the student's solution while sticking to the student's approach.

2. Manual evaluation shows most models produce **accurate corrections >93% times.**

Results

 # LLMs cannot localize well: μ = 49% & 29%

 # Even with ref soln: w-GS > w/o-S for most cases, but w-GS remains low (54% & 36%).

 # Intermediate corrections help: (57% & 40%) w-Cor scores highest for most data-LLM pairs.

Model	VtG			PRM800K		
Random	18.32			9.52		
	w/o-S	*w-GS*	*w-Cor*	*w/o-S*	*w-GS*	*w-Cor*
Llama3-70B	42.51	49.50	**61.28**	19.64	24.12	**33.03**
Llama3.1-70B	49.10	57.98	**64.17**	24.46	34.23	**38.39**
Llama3.1-405B	49.90	62.38	**64.77**	24.12	39.54	**47.86**
GPT-4o	54.49	63.57	**64.57**	39.29	43.72	**49.40**
Qwen2.5-72B-Math	**45.01**	30.44	19.10	21.86	**28.50**	21.47
LearnLM-1.5-Pro	54.89	**64.07**	63.67	42.51	49.69	**51.13**

Table 1: First error step localization accuracy (in %) on VtG and PRM800K datasets. For each task, within each dataset, the **bold** value represents the highest accuracy per LLM, whereas the underlined value represents the overall highest accuracy.

References

[1] Daheim, N., Macina, J., Kapur, M., Gurevych, I., & Sachan, M. (2024). Stepwise verification and remediation of student reasoning errors with large language model tutors. EMNLP 2024. ACL. https://doi.org/10.18653/v1/2024.emnlp-main.478

[2] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let's verify step by step. arXiv preprint arXiv:2305.20050. https://arxiv.org/abs/2305.20050

Data & Code | Presentation | arXiv Paper

Analysis

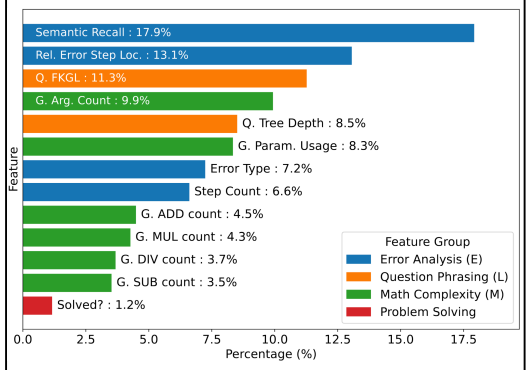
Q.What impacts error localization the most?

Black-box **feature importance study** reveals:

Semantic recall b/w the student solution (S) and the reference solution (G or C) has the greatest impact.

Questions' phrasing and **math complexity** rank among the most impactful.

LLMs' ability to solve the underlying math problem does not correlate well with their ability to spot errors in an incorrect solution (Chi-squared Test: $p > 0.01$; $\phi < 0.2$).



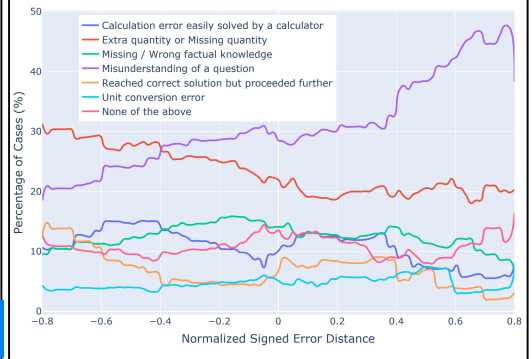
Q. Does the type of error play a role?

Tracking different error types relative to their actual location reveals:

Question-independent errors are **uniformly distributed**, e.g., calculation or unit-conversion mistakes.

Errors due to question misunderstanding are **predicted much later than they occur**.

Errors involving missing or extra variables are **predicted much before that actually occur**.



Q.Why does Qwen2.5-72B-Math perform so poorly across settings?

Despite being tuned on math problem data, Qwen2.5-Math scores consistently lower than other models in most cases.

Manual evaluation reveals that the model struggles to (1) follow instructions to spot first error location and (2) to rectify first error in C.

Likely because the model is **tuned for problem solving rather than critique or other metareasoning.**