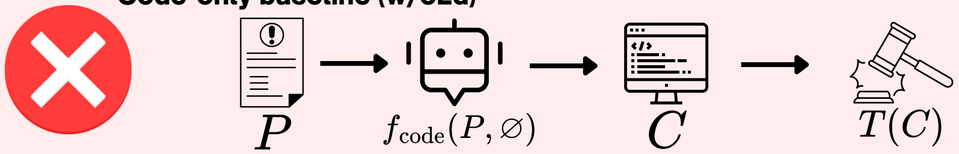
Current Evaluations: *"Conflates algorithmic reasoning/plan with implementation"*

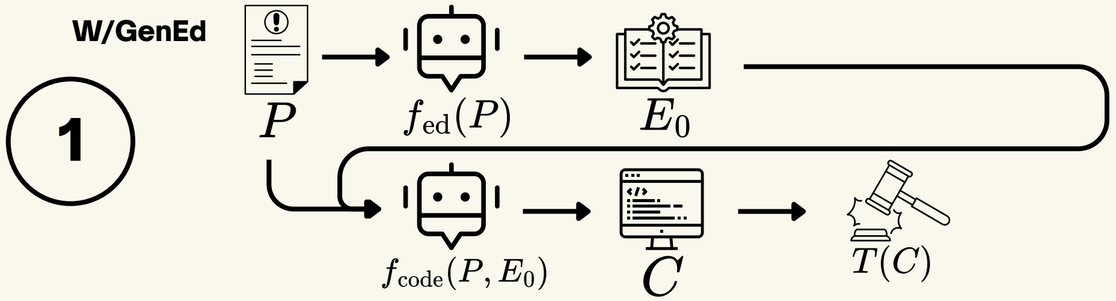
Code-only baseline (w/oEd)



Editorial(algorithmic reasoning/plan) Centric Evaluation

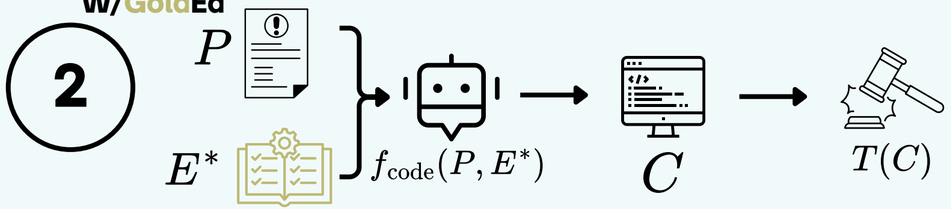
"Isolates model-generated algorithmic reasoning/plan form implementation"

W/GenEd



"Correct algorithmic reasoning/plan provided → isolates implementation limits"

W/GoldEd



LLM-Generated Editorial Annotation

Problem Understanding (PU)

PU-W	PU-M	PU-X	PU-D
Wrong crucial detail Yes/NO	Missing crucial detail Yes/NO	Irrelevant / misleading detail None/Minor/Major	Problem Understanding Difficulty (0 to 5)

Algorithm Description (ALG)

ALG-TAG	Golden-ALG-TAG
The high-level idea/algorithm described in the LLM-generated editorial	The high-level idea/algorithm described in the gold editorial

Algorithm Correctness (ALG-COR)

ALG-COR

The editorial solves/fails the problem under the stated constraints.
Correct / Incorrect

Correct Type

Matches the official solution exactly or uses a different but equally valid approach
Labels: Same as golden / Different from golden

Why incorrect

- Wrong Algorithm
- Correct Algorithm but Incorrect Approach
- Suboptimal but Correct Algorithm
- Suboptimal and Wrong Algorithm

Severity of incorrectness

- Completely Wrong
- Major Edits Needed
- Minor Edits Needed