PECC: Problem Extraction and Coding Challenges

Patrick Haller, Jonas Golde, Alan Akbik
Humboldt Universität zu Berlin

Can Language Models Understand and Code as You Narrate?

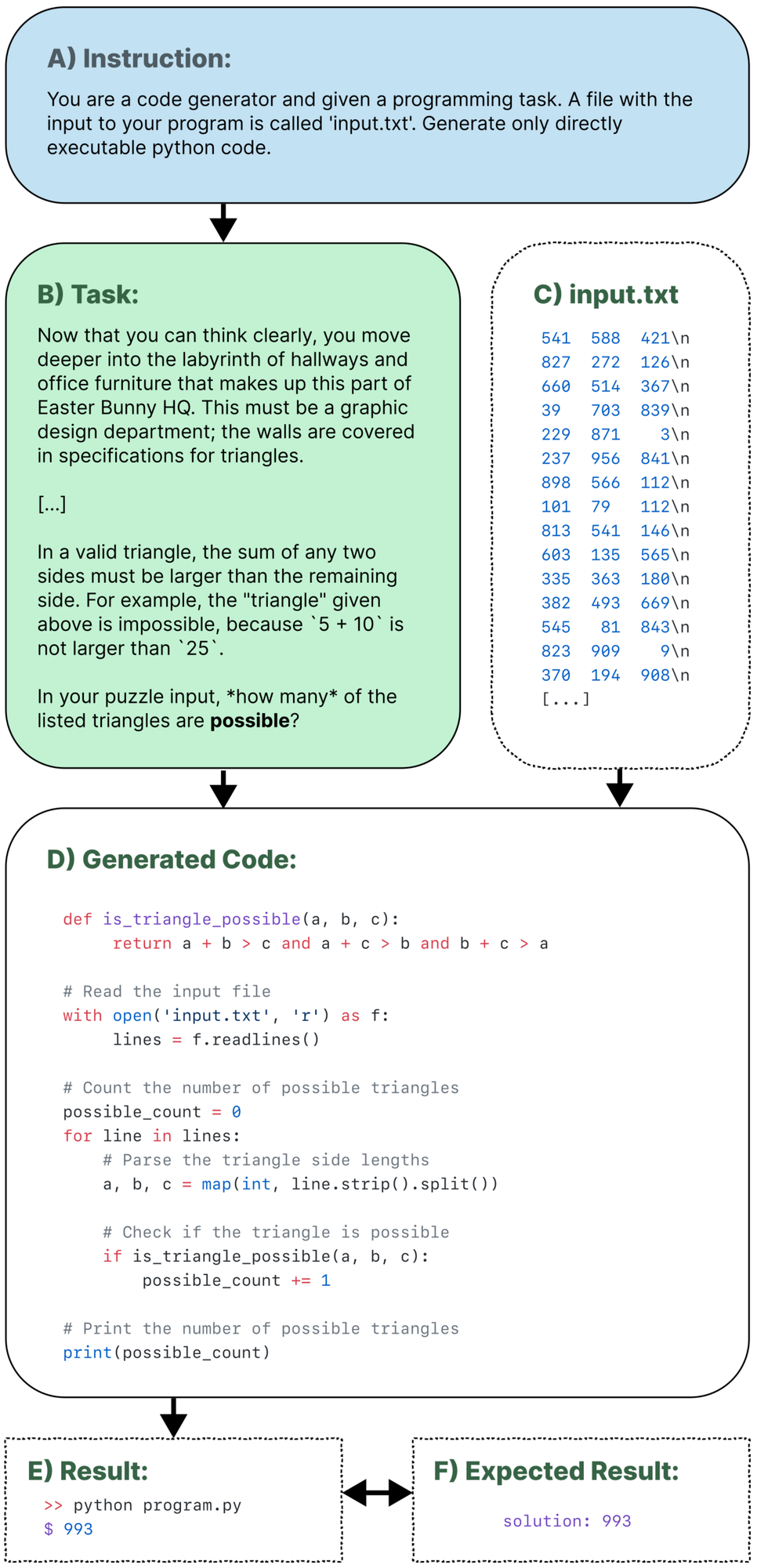
Novel Benchmark to evaluate LLMs' abilities to

- Understand prose-style problem statements.
- Extract underlying problems.
- Generate appropriate code solutions.

	APPS	HumanEval	DS-1000	PECC
Programming Language	Python	Python	Python	Universal
Avg. Program Length	18	6	3.6	26 (AoC) /19 (Euler)
Number of Problems	10,000	164	1000	2352
Domain	Programming	Programming	Programming	Math & Programming
Evaluation	Test Cases	Python Code	Test Cases + Surface Form Constraints	Explicit Result

PECC Evaluation Pipeline

- **Instruction Stage**: Define the task and context to the LLM, setting foundational instructions and problem statements.
- **Code Generation:** LLM generates executable Python code from narrative or neutral problem descriptions.
- **Execution and Result Assertion:** Code is executed, and outputs are compared against expected results to validate correctness.



Results and Findings

Evaluation with Pass@3 Accuracy on all 4 subsets and average.

Model	AoC	AoC-Concise	Euler	Euler-Stories	Average
claude-haiku	51.28	46.26	7.07	6.08	27.67
gpt-3.5-turbo	50.00	29.85	8.19	6.95	23.75
codechat-bison	21.17	17.60	4.59	2.61	11.49
chat-bison	17.09	13.78	2.36	0.62	8.48
mixtral-instruct	15.31	13.01	2.86	2.23	8.35
phi3-instruct	10.13	13.00	3.35	2.23	7.18
wizardlm-2-7b	5.87	6.89	1.24	0.87	3.72
llama3-8b-instruct	1.53	6.38	4.47	0.0	3.1

- Strong difference in AoC and Euler subset scores
- Proprietary LMs perform significantly better in chat-based evaluation

Conclusion

Challenging chat-based interaction not only requires a capable pre-trained LM, but also robust instruction fine-tuning.

Acknowledgments Alan Akbik and Patrick Haller are supported by the Deutsche Forschungsgemein- schaft (DFG, German Research Foundation) un- der Emmy Noether grant "Eidetic Represen- tations of Natural Language" (project number 448414230). Alan Akbik is furthermore supported under Germany's Excellence Strategy "Sci- ence of Intelligence" (EXC 2002/1, project number 390523135). We thank all reviewers for their valu- able comments. Jonas Golde is supported by the German Federal Ministry of Economic Affairs and Climate Action (BMWK) as part of the project ENA (KK5148001LB0).