WHAT MAKES LARGE LANGUAGE MODELS REASON IN (MULTI-TURN) CODE GENERATION?

Kunhao Zheng*, Juliette Decugis*, Jonas Gehring, Taco Cohen,
Benjamin Negrevergne, Gabriel Synnaeve

Dauphine | PSL UNIVERSITÉ PARIS
LAMSADE UMR 7243
Meta AI

Motivation

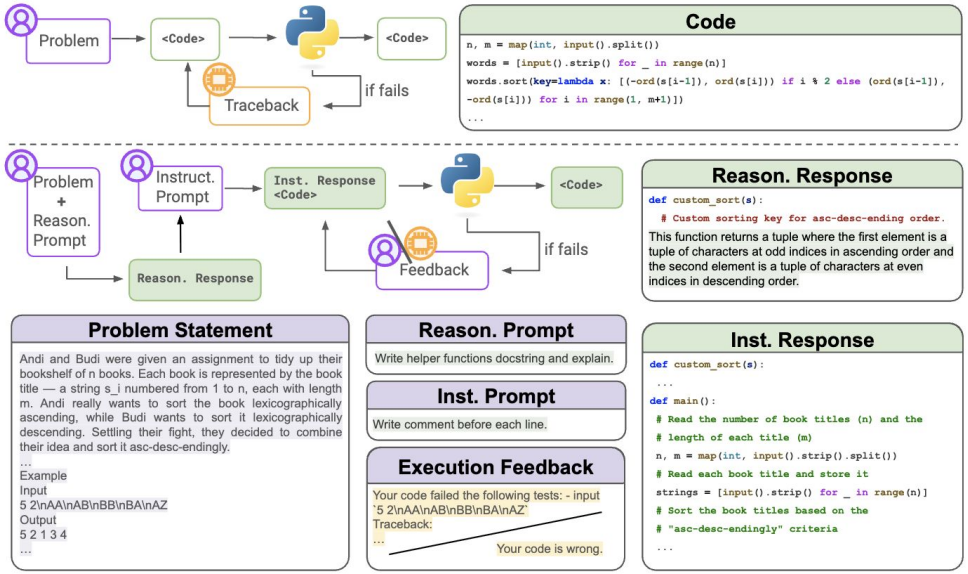

Figure 1: **Our framework for evaluating LLM multi-turn code generation techniques. Top**: In the default multi-turn setting, given a programming problem, the model generates a code solution, interacts with the runtime environment to gather execution feedback and retries in case of failure. **Bottom**: On top of the default setting, we gather *reasoning* (Reason.) prompts, *instruction* (Inst.) prompts, and *execution feedback* prompts. The problem statement is augmented with a *reasoning* prompt. After generating an answer to the *reasoning* prompt, an *instruction* prompt determines how program code should be generated. The *execution feedback* prompts vary in granularity, ranging from a binary pass or fail indicator to detailed tracing information.

Q: Can we find a trade-off between methods using:

- massive single-turn sampling
- complex prompt engineering

Target Benchmark (Competitive programming):

- CodeContests
- TACO
- LiveCodeBench

Contributions

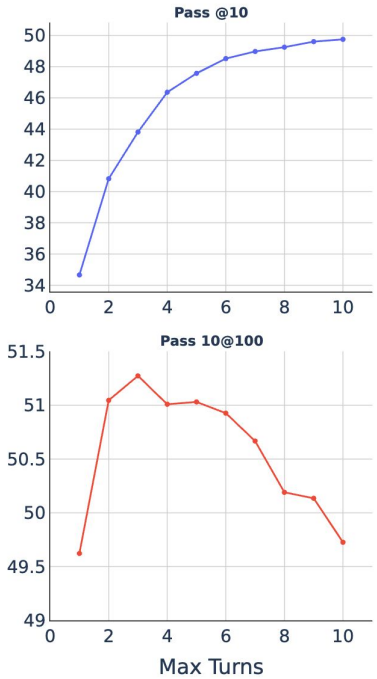

Figure 2: **Scaling number of turns is not compute optimal.** Pass@10 (**Top**) and pass 10@100 (**Bottom**) on CodeContests test set when increasing the number of turns with Llama 3.1 70B.

Comprehensive experimental survey of CoT and execution feedback in both single-turn and multi-turn setup under *equal compute constraints*

3 Key Insights:

- ***Single turn***: reasoning (NL → NL) + instruction (NL → Code) is best, CoT can degrade performance
- ***Multi-turn***: can be worse than single-turn under fixed sampling budget
- ***Rejection Sampling Fine-tuning***: Reasoning can be learned via fine-tuning on multi-turn CoT data

Prompting and Feedback Space

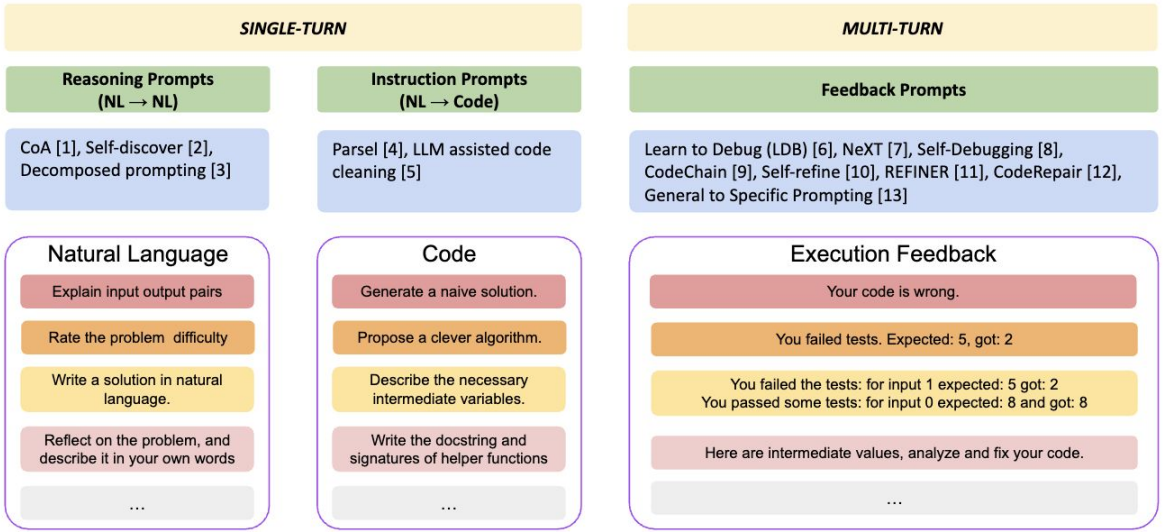

Figure 3: **Prompting space explored in our survey.** We explore chain of thought prompts at three different levels: before the first code generation (*reasoning* prompts), with code generation (*instruction* prompts), and after the first code generation (*execution feedback*). The corresponding works from the single-turn and multi-turn reasoning and code generation literature are: [1] Gao et al. (2024), [2] Zhou et al. (2024), [3] Khot et al. (2023), [4] Zelikman et al. (2023), [5] Jain et al. (2024), [6] Zhong et al. (2024), [7] Ni et al. (2024), [8] Chen et al. (2024), [9] Le et al. (2024), [10] Madaan et al. (2024), [11] Paul et al. (2024), [12] Tang et al. (2024), [13] Li et al. (2023a).

Single turn

Main take-away: CoT works best for hard problems, large models and high sampling

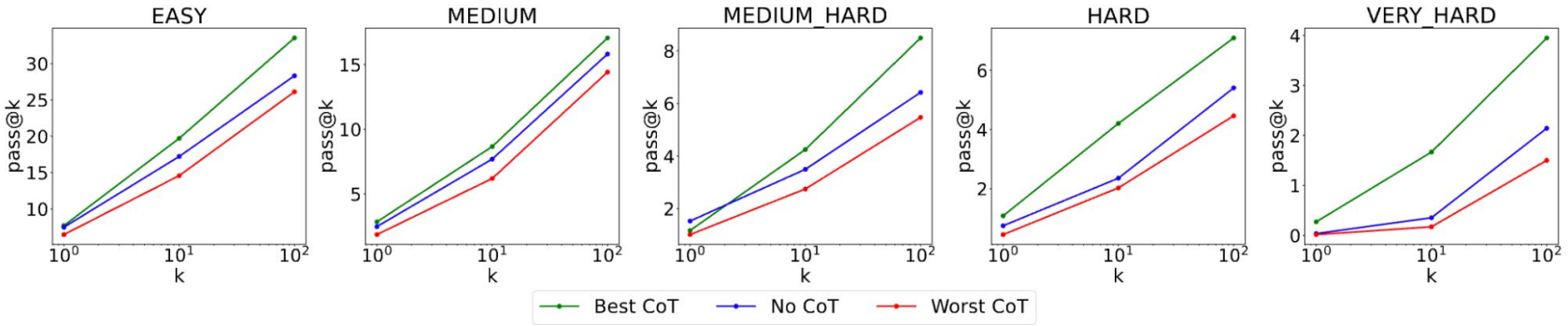

Figure 4: **CoT helps most on hard examples.** From a set of 8 reasoning and 6 instruction prompts commonly used on competitive coding benchmarks, we extract the pass rate of the best and worst prompts amongst all $63 = (8+1) \times (6+1)$ combinations (including no *reasoning* or no *instruction*) for Llama 3.0 8B. We compare on different difficulty split of the TACO dataset. The relative gain from a tailored CoT increases with problem difficulty and sampling size.

Multi-turn

Main take-away: execution feedback granularity impacts model's multi-turn behavior (exploration v.s. exploitation)

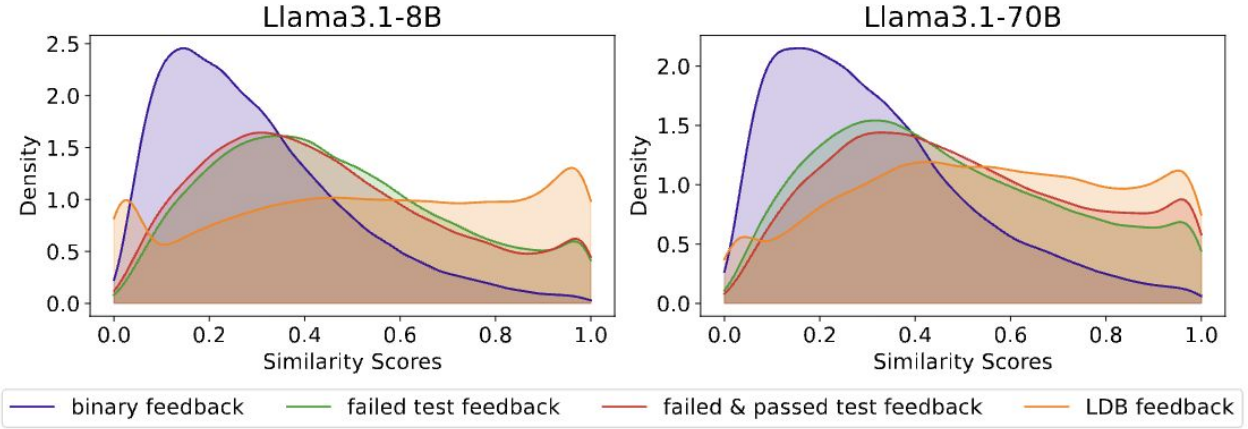

Figure 6: **Fine-grained feedback induces exploitative behavior.** Distribution of consecutive code similarity scores within dialog for different types of feedback, obtained from Llama 3.1 8B and 70B samples (temperature 1.0). The higher the similarity scores between consecutive codes in the same dialog, the more the model exhibits exploitative behavior.

CoT Rejection Sampling

Main take-away: models can internalize reasoning behavior without any specific prompting through rejection sampling finetuning.

Table 1: **Up to +10% with multi-turn CoT.** Pass $n@k$ (%) on CodeContests test set with high temperature (1.0) and large sampling budget. In the multi-turn setting, we use a maximum of 3 code attempts (i.e., 3 turns) with the "failed tests" feedback. The pass $n@k$ is calculated from 200 trajectories for both single-turn and multi-turn settings. We also report the pass rates for Llama 3.1 70B after Rejection Sampling Fine-tuning (RFT) (Section 5.3).

Model	Variants	CodeContests / Test			
		1@3	10@30	33@100	100@300
Llama 3.0 8B		2.9	8.0	12.6	-
	+ CoT	$3.4_{+0.5}$	$11.7_{+3.7}$	$\mathbf{17.3}_{+4.7}$	-
	+ Multi-turn	$2.4_{-0.5}$	$8.0_{+0.0}$	$12.8_{+0.2}$	16.7
	+ Multi-turn CoT	$2.8_{-0.1}$	$9.8_{+1.8}$	$14.9_{+2.3}$	19.4
Llama 3.0 70B		9.6	18.9	23.1	-
	+ CoT	$10.4_{+0.8}$	$26.0_{+7.1}$	$33.0_{+9.9}$	-
	+ Multi-turn	$10.1_{+0.5}$	$21.0_{+2.1}$	$26.7_{+3.6}$	32.7
	+ Multi-turn CoT	$11.1_{+1.5}$	$26.5_{+7.6}$	$\mathbf{34.3}_{+11.2}$	40.4
Llama 3.1 8B		7.7	18.2	23.8	-
	+ CoT	$8.0_{+0.3}$	$19.5_{+1.3}$	$\mathbf{26.1}_{+2.3}$	-
	+ Multi-turn	$7.0_{-0.7}$	$18.8_{+0.6}$	$24.5_{+0.7}$	30.4
	+ Multi-turn CoT	$6.9_{-0.8}$	$19.4_{+1.2}$	$26.0_{+2.2}$	31.5
Llama 3.1 70B		24.1	42.3	49.8	-
	+ CoT	$26.4_{+2.3}$	$47.8_{+5.5}$	$54.8_{+5.0}$	-
	+ Multi-turn	$24.1_{+0.0}$	$43.8_{+1.5}$	$51.6_{+1.8}$	56.2
	+ Multi-turn CoT	$27.7_{+3.6}$	$48.4_{+6.1}$	$\mathbf{55.3}_{+5.5}$	59.6
Llama 3.1 $70B^{RFT}$		26.2	45.1	50.9	-
	+ Multi-turn	$29.7_{+3.5}$	$50.5_{+5.4}$	$\mathbf{57.2}_{+6.3}$	61.1

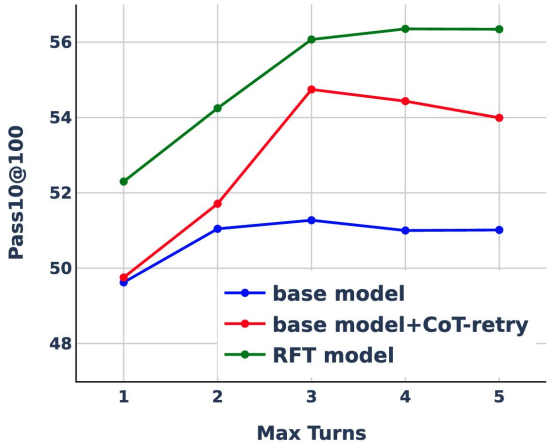

Figure 7: *Reasoning* and *execution feedback* prompts, and RFT, enhance both single- and multi-turn performance for Llama 3.1 70B.

CoT Rejection Sampling

Main take-away: models can internalize reasoning behavior without any specific prompting through rejection sampling finetuning.

Table 1: **Up to +10% with multi-turn CoT.** Pass $n@k$ (%) on CodeContests test set with high temperature (1.0) and large sampling budget. In the multi-turn setting, we use a maximum of 3 code attempts (i.e., 3 turns) with the "failed tests" feedback. The pass $n@k$ is calculated from 200 trajectories for both single-turn and multi-turn settings. We also report the pass rates for Llama 3.1 70B after Rejection Sampling Fine-tuning (RFT) (Section 5.3).

Model	Variants	CodeContests / Test			
		1@3	10@30	33@100	100@300
Llama 3.0 8B		2.9	8.0	12.6	-
	+ CoT	$3.4_{+0.5}$	$11.7_{+3.7}$	$\mathbf{17.3}_{+4.7}$	-
	+ Multi-turn	$2.4_{-0.5}$	$8.0_{+0.0}$	$12.8_{+0.2}$	16.7
	+ Multi-turn CoT	$2.8_{-0.1}$	$9.8_{+1.8}$	$14.9_{+2.3}$	19.4
Llama 3.0 70B		9.6	18.9	23.1	-
	+ CoT	$10.4_{+0.8}$	$26.0_{+7.1}$	$33.0_{+9.9}$	-
	+ Multi-turn	$10.1_{+0.5}$	$21.0_{+2.1}$	$26.7_{+3.6}$	32.7
	+ Multi-turn CoT	$11.1_{+1.5}$	$26.5_{+7.6}$	$\mathbf{34.3}_{+11.2}$	40.4
Llama 3.1 8B		7.7	18.2	23.8	-
	+ CoT	$8.0_{+0.3}$	$19.5_{+1.3}$	$\mathbf{26.1}_{+2.3}$	-
	+ Multi-turn	$7.0_{-0.7}$	$18.8_{+0.6}$	$24.5_{+0.7}$	30.4
	+ Multi-turn CoT	$6.9_{-0.8}$	$19.4_{+1.2}$	$26.0_{+2.2}$	31.5
Llama 3.1 70B		24.1	42.3	49.8	-
	+ CoT	$26.4_{+2.3}$	$47.8_{+5.5}$	$54.8_{+5.0}$	-
	+ Multi-turn	$24.1_{+0.0}$	$43.8_{+1.5}$	$51.6_{+1.8}$	56.2
	+ Multi-turn CoT	$27.7_{+3.6}$	$48.4_{+6.1}$	$\mathbf{55.3}_{+5.5}$	59.6
Llama 3.1 $70B^{RFT}$		26.2	45.1	50.9	-
	+ Multi-turn	$29.7_{+3.5}$	$50.5_{+5.4}$	$\mathbf{57.2}_{+6.3}$	61.1

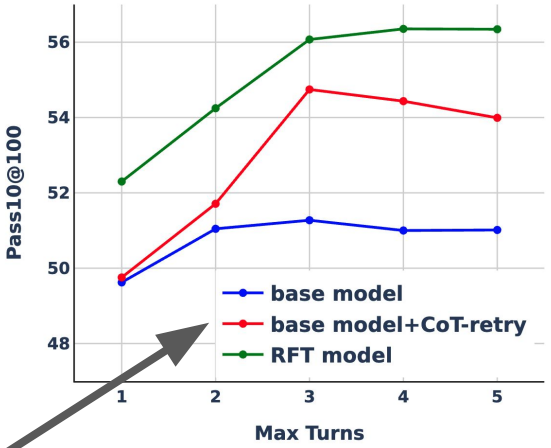

Figure 7: *Reasoning* and *execution feedback* prompts, and RFT, enhance both single- and multi-turn performance for Llama 3.1 70B.

CoT Rejection Sampling

Main take-away: models can internalize reasoning behavior without any specific prompting through rejection sampling finetuning.

Table 1: **Up to +10% with multi-turn CoT.** Pass $n@k$ (%) on CodeContests test set with high temperature (1.0) and large sampling budget. In the multi-turn setting, we use a maximum of 3 code attempts (i.e., 3 turns) with the "failed tests" feedback. The pass $n@k$ is calculated from 200 trajectories for both single-turn and multi-turn settings. We also report the pass rates for Llama 3.1 70B after Rejection Sampling Fine-tuning (RFT) (Section 5.3).

Model	Variants	CodeContests / Test			
		1@3	10@30	33@100	100@300
Llama 3.0 8B		2.9	8.0	12.6	-
	+ CoT	$3.4_{+0.5}$	$11.7_{+3.7}$	$\mathbf{17.3}_{+4.7}$	-
	+ Multi-turn	$2.4_{-0.5}$	$8.0_{+0.0}$	$12.8_{+0.2}$	16.7
	+ Multi-turn CoT	$2.8_{-0.1}$	$9.8_{+1.8}$	$14.9_{+2.3}$	19.4
Llama 3.0 70B		9.6	18.9	23.1	-
	+ CoT	$10.4_{+0.8}$	$26.0_{+7.1}$	$33.0_{+9.9}$	-
	+ Multi-turn	$10.1_{+0.5}$	$21.0_{+2.1}$	$26.7_{+3.6}$	32.7
	+ Multi-turn CoT	$11.1_{+1.5}$	$26.5_{+7.6}$	$\mathbf{34.3}_{+11.2}$	40.4
Llama 3.1 8B		7.7	18.2	23.8	-
	+ CoT	$8.0_{+0.3}$	$19.5_{+1.3}$	$\mathbf{26.1}_{+2.3}$	-
	+ Multi-turn	$7.0_{-0.7}$	$18.8_{+0.6}$	$24.5_{+0.7}$	30.4
	+ Multi-turn CoT	$6.9_{-0.8}$	$19.4_{+1.2}$	$26.0_{+2.2}$	31.5
Llama 3.1 70B		24.1	42.3	49.8	-
	+ CoT	$26.4_{+2.3}$	$47.8_{+5.5}$	$54.8_{+5.0}$	-
	+ Multi-turn	$24.1_{+0.0}$	$43.8_{+1.5}$	$51.6_{+1.8}$	56.2
	+ Multi-turn CoT	$27.7_{+3.6}$	$48.4_{+6.1}$	$\mathbf{55.3}_{+5.5}$	59.6
Llama 3.1 70B^{RFT}		26.2	45.1	50.9	-
	+ Multi-turn	$29.7_{+3.5}$	$50.5_{+5.4}$	$\mathbf{57.2}_{+6.3}$	61.1

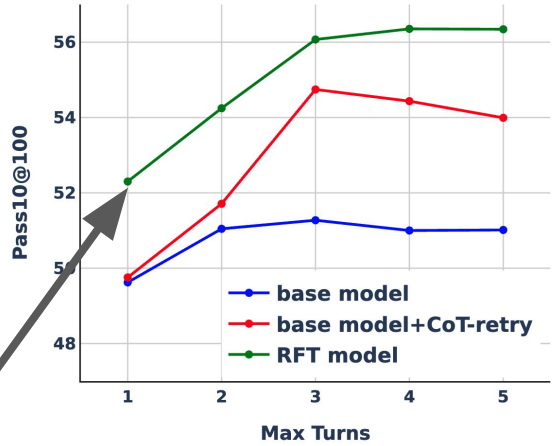

Figure 7: *Reasoning* and *execution feedback* prompts, and RFT, enhance both single- and multi-turn performance for Llama 3.1 70B.

Conclusion

Limitations / future work:

- Only focus on linear trajectories (not tree structures, back-tracking)
- Use the number of LLM forwards / evaluations (pass n@k) not number of tokens
- More advanced training beyond SFT / RFT