UNIVERSITY of WASHINGTON

LeetPrompt: Leveraging Human Intelligence to Study Large Language Models

Sebastin Santy Ayana Bharadwaj Sahith Dambekodi Alex Albert Cathy Yuan Ranjay Krishna

Presenting LeetPrompt: Rigorous Evaluation with Real Human Interactions

LeetPrompt is an online platform populated with problems that users can attempt to solve by invoking LLMs with custom instructions. **As a dual objective platform**, LeetPrompt allows users to solve problems with LLMs while the platform automatically gathers evaluation metrics and user-behavior insights for researchers.

A Need for Better Benchmarking

Past attempts are limited due to model evaluation on a fixed set of LLM ”instructions” (colloquially referred to as prompts) or remain largely anecdotal when shared on forums and social media.

Findings

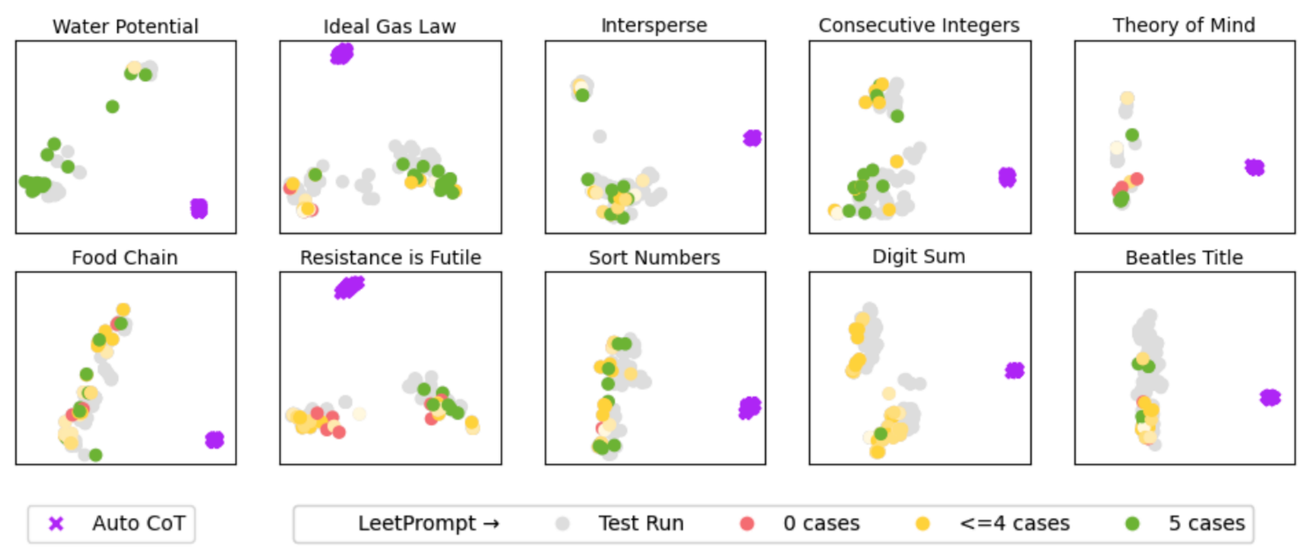
We conduct a within-subjects user study (N = 32) across 10 problems from 5 domains: biology, physics, math, programming, and general knowledge. By analyzing the 1853 instructions used to invoke GPT-4, we present the following findings:

(1) Participants are able to design instructions for all tasks, including those that problem setters deemed unlikely to be solved. Sometimes, incorrect instructions can still result in correct LLM behavior.

(2) All automatic mechanisms fail to generate instructions to solve all tasks, including zero-shot, few-shot, zero-shot CoT, few-shot CoT, and auto-CoT.

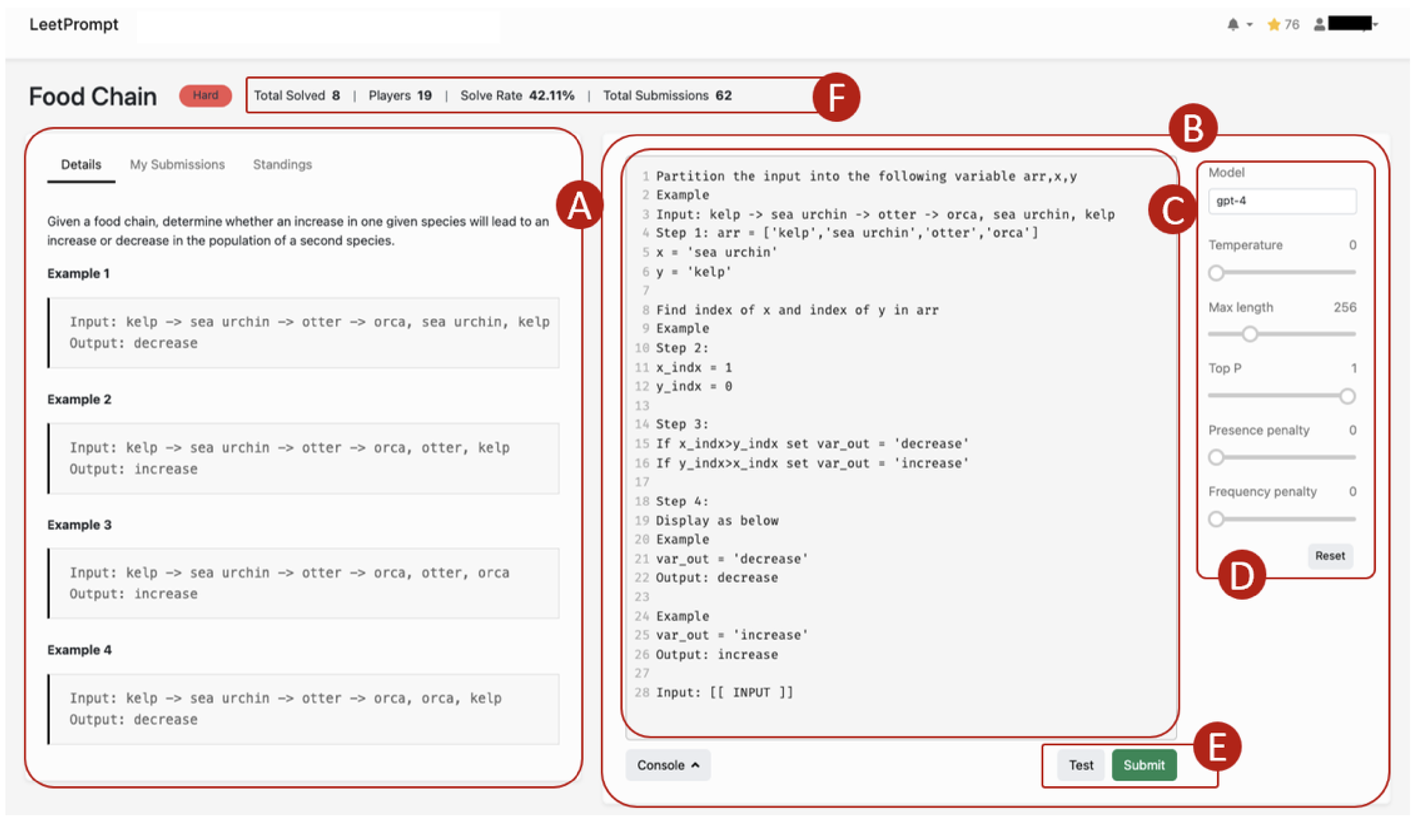
(3) Lexical diversity of instructions is significantly correlated with people’s ability to solve the problem, highlighting the need for diverse instruction strategies.

(4) Many instruction strategies are unsuccessful, highlighting the misalignment between the participant’s conceptual model of the LLM and its functionality.



Visualizing the space of Auto-CoT & LeetPrompt Instructions: Auto-CoT solutions (purple) lack both lexical and solvability diversity, as visualized in each problem presented to participants. Participants also struggled with debugging, with entire solution clusters that don't solve the problem. Finally, clusters appear to follow similar patterns between each column (representing a domain), implying that people use similar strategies in a given domain.

LeetPrompt Interface



(A) Problem description **(B)** Interaction interface **(C)** Writing instructions **(D)** Model hyper-parameters **(E)** Test and submit **(F)** Problem details and submissions

Discussion & Future Work

(1) Potential to train or ”instruction-tune” language models by leveraging collective creative through collecting diverse prompts

(2) Micro-analysis of LLMs, allowing researchers to focus on individual problems for analysis

(3) Future possibilities: Tutorials on how to use LLMs; Audit released models by the public; Micro-analysis evaluation of individual inputs; Track interaction patterns between people and LLMs over time.