AutoPrompt: An Automated Prompt-Tuning Tool

Nicholas Derby
ServiceNow; University of Colorado, Boulder

Project Motivation

- Prompt tuning is one of the most important, yet time consuming, steps of developing any system that incorporates LLMs.
- This project aims to automate the prompt tuning process by iteratively modifying a prompt using an LLM.
- The resulting system, AutoPrompt, is built into ServiceNow's existing LLM evaluation framework, allowing seamless prompt tuning across many tasks and evaluation models.

Existing Evaluation Framework

- ServiceNow's model evaluation framework tests LLM performance on a variety of tasks, including:
  - Generating a regular expression based on a natural language description.
  - Creating a survey when given detailed instructions on survey questions and their desired formats.
  - Summarizing documents.
  - Planning an agentic workflow when provided a goal and available tools.
- The LLM testing framework supports a variety of open source and API-based LLMs, including GPT, Llama, NowLLM (ServiceNow's own LLM), Claude Sonnet, etc.
- To run the framework, a user selects the task and the model to be tested.
- The testing framework consists of the following pipeline. All prompts and data are stored within the framework's codebase for easy replicability:

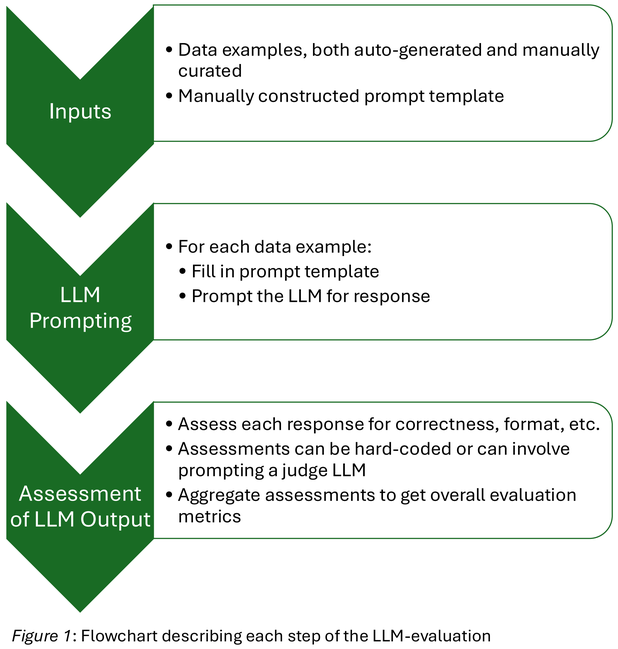

Figure 1: Flowchart describing each step of the LLM-evaluation framework

System Overview

- AutoPrompt directly modifies the stored prompts within the evaluation framework.
- To use AutoPrompt, a user selects a task for prompt tuning from a list of available tasks within the framework. The user also selects a model on which to run the task. The prompt tuning model is always GPT-4.1.
- AutoPrompt can be optionally run with a human in the loop, prompting the user for additional prompt tuning instructions with each iteration.
- The user is provided with a full report of the evaluation metrics and a copy of each tuned prompt after each iteration, allowing the user to select the best prompt after all iterations have completed.

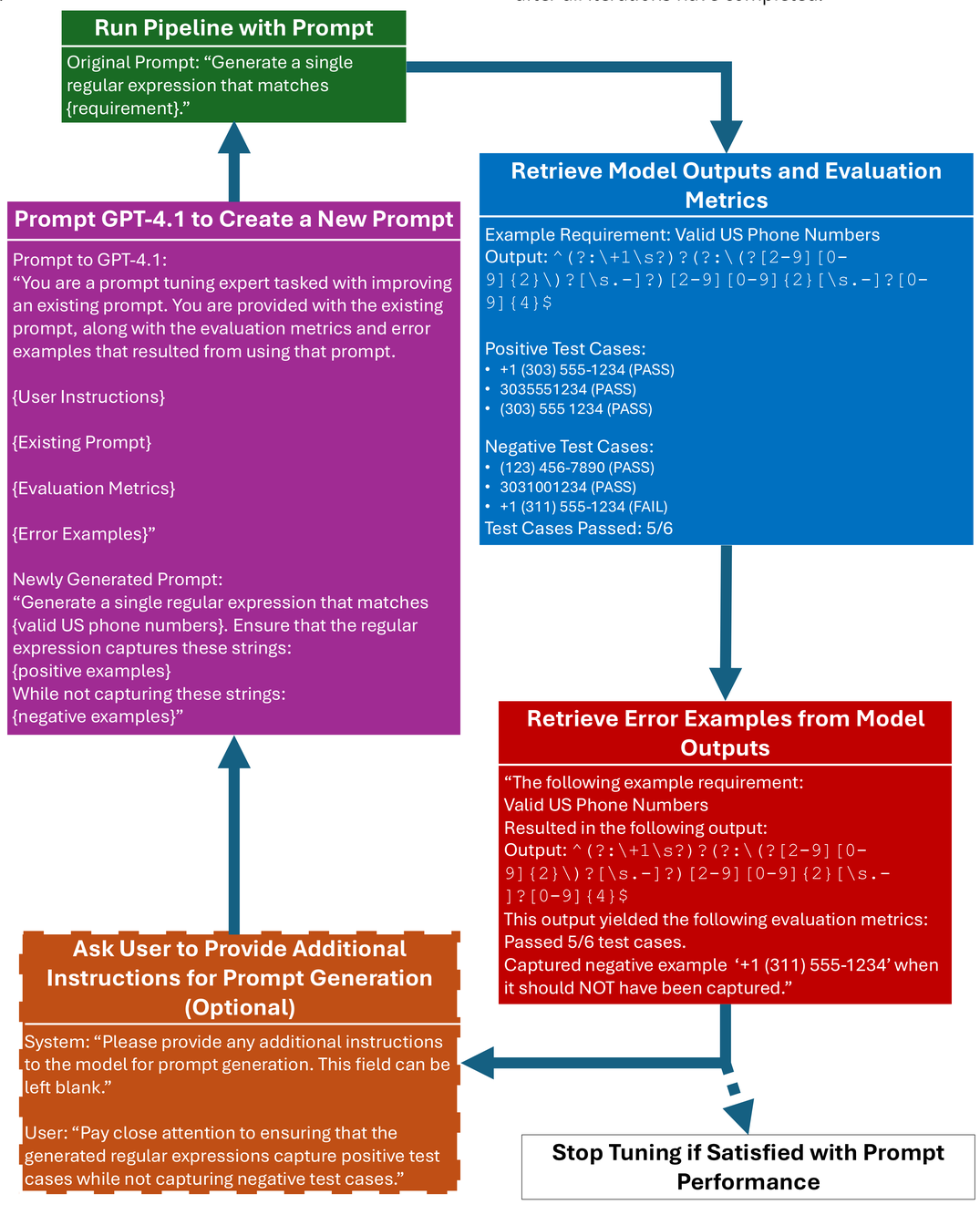

Figure 2: Cyclical chart describing each step of AutoPrompt. This uses a toy RegEx Generation example; these are not the actual prompts/data examples used in the evaluation framework or by AutoPrompt.

Experiments & Results

- AutoPrompt tuned the existing prompts for 3 tasks:
  - Regular Expression Generation
  - Survey Creation
  - Agentic Planning
- All prompt tuning runs were performed 5 times with the following settings:
  - Tuning Model: GPT-4.1
  - Evaluation Model: NowLLM
  - Number of Iterations: 10
  - Automatic Mode (no human in the loop)
- Results reported below are from the best run after 5 iterations, as performance tended to plateau or degrade for subsequent iterations.

Table 1: Comparison of Accuracy Metrics of Original Prompts and Prompts Tuned with AutoPrompt after 5 Iterations

Task	Original Accuracy	AutoPrompt Accuracy
RegEx Generation	0.72	**0.94**
Survey Creation	0.41	**0.79**
Agentic Planning	0.28	**0.31**

Conclusions

- AutoPrompt is most effective on tasks that were clearly defined in scope, struggling to prompt-tune on a very open-ended task such as Agentic Planning.
- AutoPrompt can handle more complex tasks to a certain limit, as evidenced by its excellent performance on survey generation.
- This project demonstrates LLMs' efficacy in prompt tuning for themselves on clearly scoped tasks, enabling self-improvement of LLM systems with minimal human intervention.

Future Work

- AutoPrompt does not currently allow the user to manually edit a generated prompt between iterations; this feature would be a welcome addition to the system.
- The error example retrieval struggles to adapt to new tasks, since each task has its own set of evaluation metrics. Therefore, adding robustness to this step, possibly with an additional LLM call, could help here.