CriSPO: Multi-Aspect Critique-Suggestion-guided Automatic Prompt Optimization for Text Generation

Han He*, Qianchu Liu*, Lei Xu* [Equal Contribution], Chaitanya Shivade, Yi Zhang, Sundararajan Srinivasan, Katrin Kirchhoff
hankcs, liufqian, leixx @ amazon.com Amazon AWS AI Labs

amazon | science

Background / Motivation

Writing good prompts for text generation is challenging

Example of naïve prompt:

Please write a detailed clinical note summary for the input conversation.

Example of a good prompt:

Write a structured patient summary from the conversation. Your 300-350 word summary should reconstruct the clinical rationale by comprehensively addressing in each section:
- <Chief Complaint>: Presenting symptom in 2-3 precisely worded sentences
- <HPI>: …

Manually crafting good prompts is tedious, requires expertise, and doesn't scale well.

Current automated methods rely on numerical metrics to guide a search, but these approaches are less effective and lack robustness.

Our Novelty

←Code available!

LLM generated Critiques and Suggestions

Use LLM to compare reference with generation, and provide multi-dimensional critiques and actionable suggestions.

Example:

Dimensions: length, professionalism

Critiques: (1) The generated text is shorter than reference.

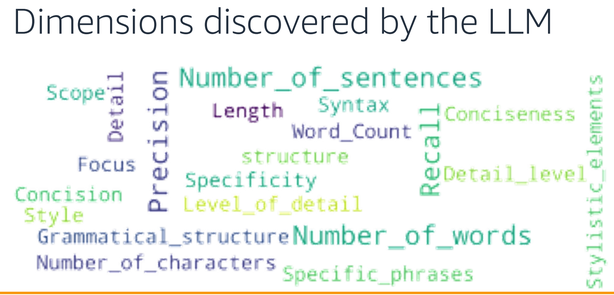
(2) The generated text is written in layman words.

Suggestions:

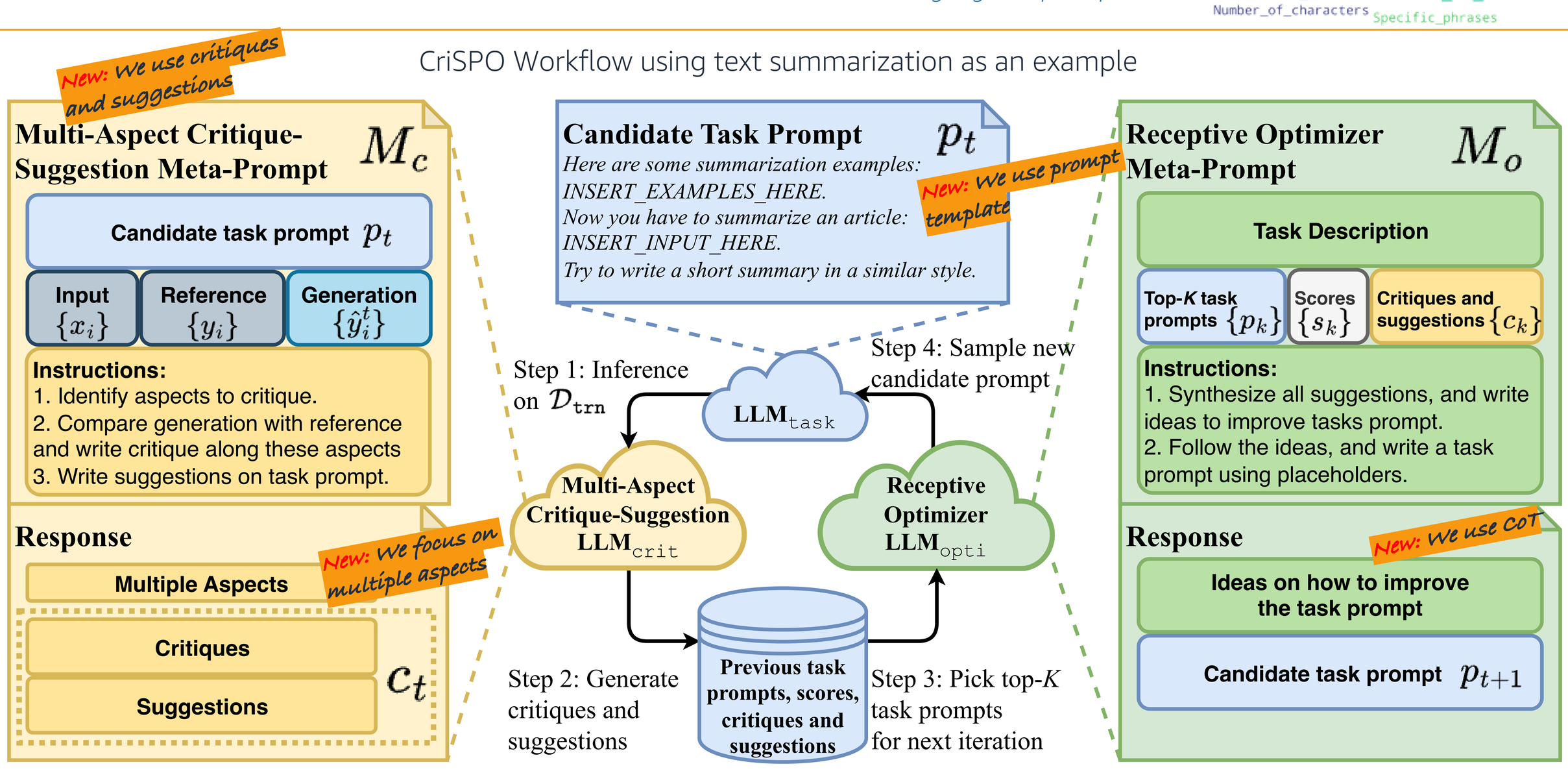
(1) Add "300-350 words" in prompt.

(2) Add "use medical language" in prompt.

Dimensions discovered by the LLM



CriSPO Workflow using text summarization as an example



Experiments

We conduct thorough experiments on 4 summarization datasets, 5 QA/RAG datasets, and 2 other natural language generation (NLG) tasks to demonstrate the effectiveness. We achieve new SoTA on medical note generation dataset (ACI-Bench).

Extension – multi-metric

Text generation usually involves optimizing multiple metrics. We first optimize for a primary metric, then use CriSPO to optimize a suffix to improve secondary metrics.

Best prompt found by CriSPO and OPRO (baseline) on SAMSum dataset

OPRO [Best Prompt]: Generate a **one to two sentence** summary within the ⟨summary⟩ tags that concisely describes the key details of the conversation and **any conclusions** reached. INPUT_DOC

CriSPO [Best Prompt]: The text below contains a discussion expressing several **key facts and events**. Your concise **1-sentence** summary should relate only the **2 most** important pieces of information stated, **without assumptions or extra context**. INPUT_DOC Write the summary within ⟨summary⟩ tags.

OPRO [Example Output]: Ralph asked Andrew if he heard a Polish joke, then told a joke about sinking a Polish battleship by putting it in water. Andrew responded that the joke was terrible and so unfunny that it made his mouth dry, requiring a sip of water.

CriSPO [Example Output]: Ralph tells Andrew a Polish battleship joke that Andrew finds unfunny.

[Reference]: Ralph told Andrew a joke.

CriSPO shows high prompt diversity

Dataset	Length↑	Vocab↑	ROUGE-L↓	Cosine↓
CNN				
OPRO	41±6	36±5	57.5	0.93
CriSPO	**149±24**	**96±12**	**50.3**	**0.90**
MeetingBank				
OPRO	31±5	28±4	44.9	0.84
CriSPO	**216±41**	**135±19**	**39.7**	**0.80**
SAMSum				
OPRO	34±6	30±5	57.0	0.94
CriSPO	**172±22**	**112±12**	**46.0**	**0.88**
D2Note				
OPRO	58±11	46±8	62.7	0.95
CriSPO	**247±40**	**117±13**	**54.3**	**0.93**

We achieve an average of 10% ROUGE-1 and ROUGE-L improvement. Check our paper for details!