Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers

Tsinghua University

Qingyan Guo*, Rui Wang*, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, Yujiu Yang

Microsoft Research

Abstract

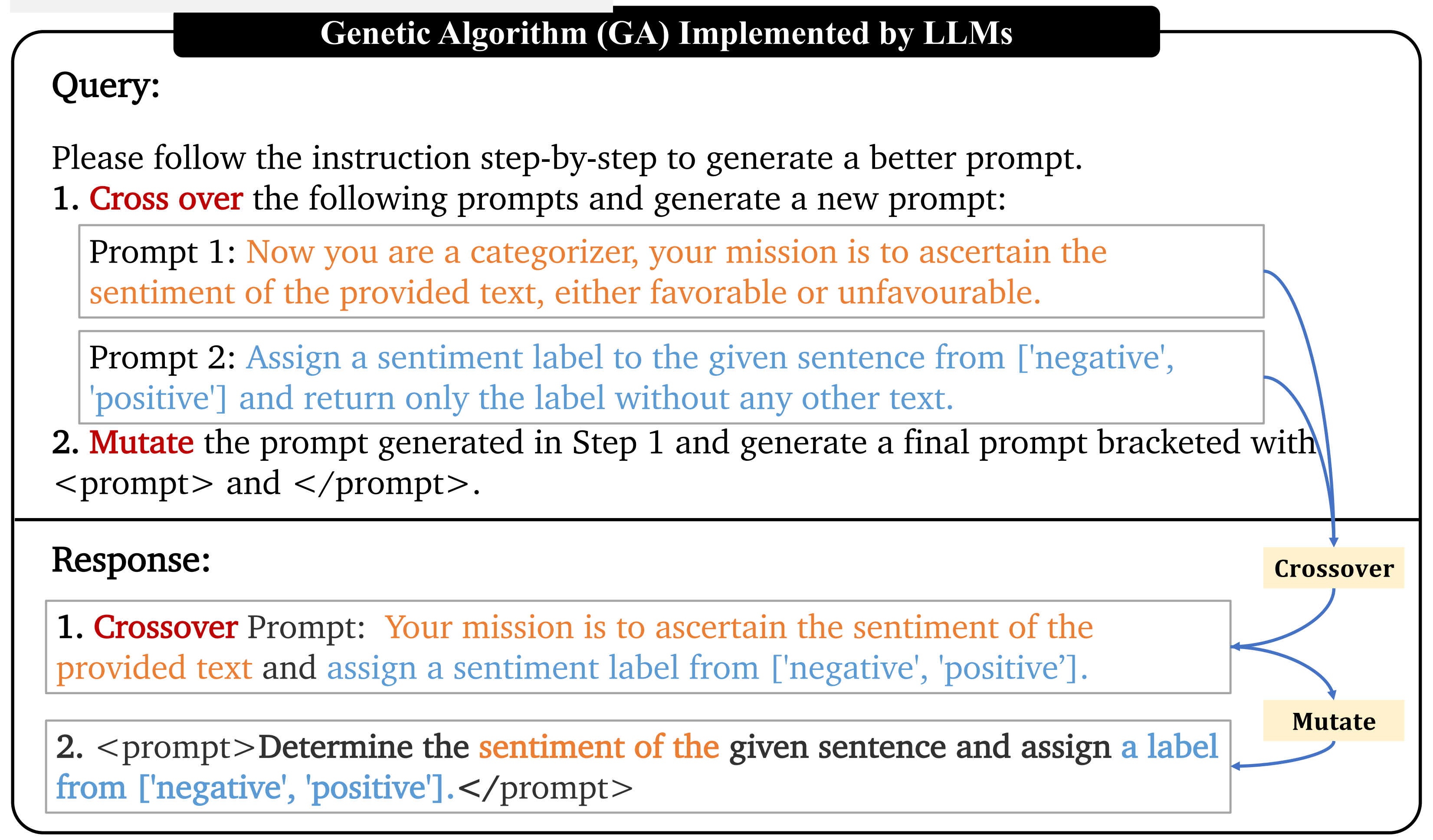
- LLMs, often interacted by **black-box** APIs , rely on carefully **crafted prompts** that often demand substantial **human effort**. We propose **EvoPrompt**, connecting LLMs with Evolutionary Algorithms, which are famous for fast convergence and striking a balance between **exploration** and **exploitation**, to generate human-readable prompts. Experiments on **31** datasets demonstrate the effectiveness of **EvoPrompt** compared with crafted prompts, as well as existing methods.

Framework

- **Initial population:** Introduce prompts written by humans and LLMs to achieve ***diversity***, avoid local optimum
- **Evolution:** Use LLMs as evolutionary operators (***mutation*** and ***crossover***) to generate a new prompt based on parent prompts from the current population
- **Update:** ***evaluation*** on a dev set and ***selection***

EvoPrompt (Genetic Algorithm)

Mutation & Crossover



Genetic Algorithm (GA) Implemented by LLMs

Query:

Please follow the instruction step-by-step to generate a better prompt.
1. Cross over the following prompts and generate a new prompt:

Prompt 1: Now you are a categorizer, your mission is to ascertain the sentiment of the provided text, either favorable or unfavourable.

Prompt 2: Assign a sentiment label to the given sentence from ['negative', 'positive'] and return only the label without any other text.

2. Mutate the prompt generated in Step 1 and generate a final prompt bracketed with <prompt> and </prompt>.

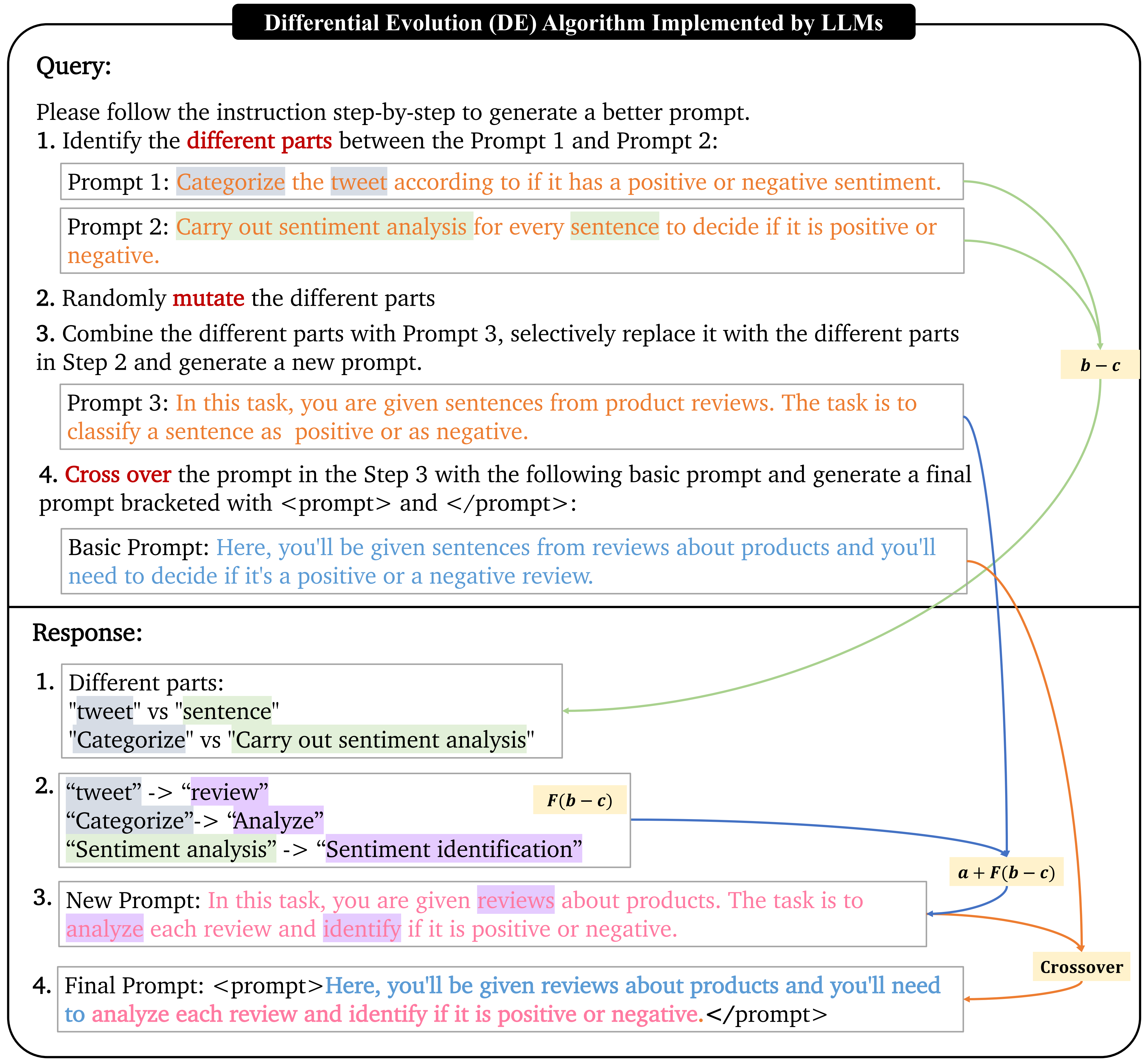
Response:

1. Crossover Prompt: Your mission is to ascertain the sentiment of the provided text and assign a sentiment label from ['negative', 'positive’].

2. <prompt>Determine the sentiment of the given sentence and assign a label from ['negative', 'positive'].</prompt>

EvoPrompt (Differential Evolution)

- For each prompt, select two parental prompts and identify the different parts: $b - c$
- Mutate on the different parts: $F(b - c)$
- Combination with current best prompt: $a + F(b - c)$
- Crossover with current prompt
- Replace the old one if performing better



Differential Evolution (DE) Algorithm Implemented by LLMs

Query:

Please follow the instruction step-by-step to generate a better prompt.
1. Identify the different parts between the Prompt 1 and Prompt 2:

Prompt 1: Categorize the tweet according to if it has a positive or negative sentiment.

Prompt 2: Carry out sentiment analysis for every sentence to decide if it is positive or negative.

2. Randomly mutate the different parts
3. Combine the different parts with Prompt 3, selectively replace it with the different parts in Step 2 and generate a new prompt.

Prompt 3: In this task, you are given sentences from product reviews. The task is to classify a sentence as positive or as negative.

4. Cross over the prompt in the Step 3 with the following basic prompt and generate a final prompt bracketed with <prompt> and </prompt>:

Basic Prompt: Here, you'll be given sentences from reviews about products and you'll need to decide if it's a positive or a negative review.

Response:

1. Different parts:
"tweet" vs "sentence"
"Categorize" vs "Carry out sentiment analysis"

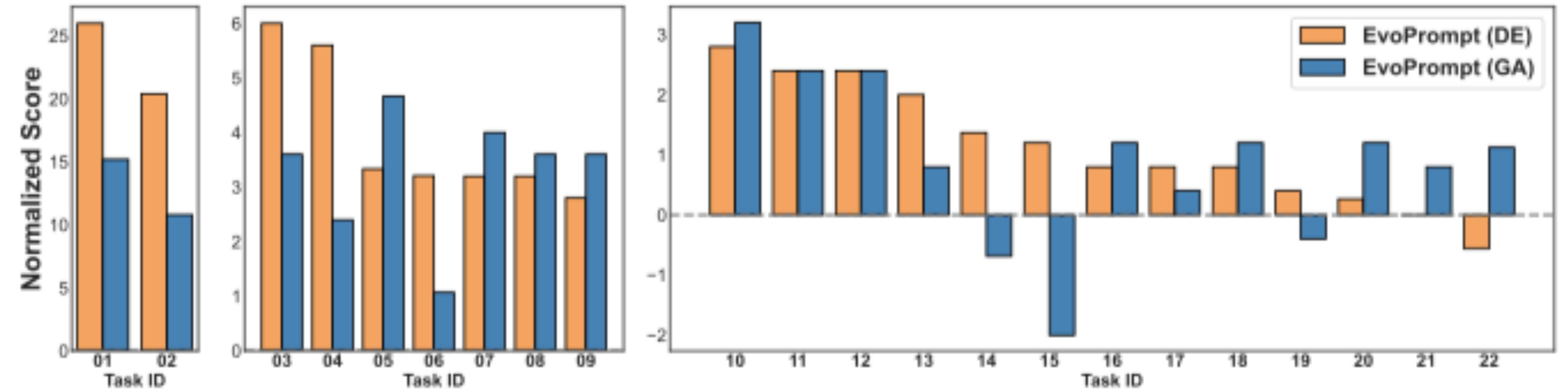
2. “tweet” -> “review”
“Categorize”-> “Analyze”
“Sentiment analysis” -> “Sentiment identification”

3. New Prompt: In this task, you are given reviews about products. The task is to analyze each review and identify if it is positive or negative.

4. Final Prompt: <prompt>Here, you'll be given reviews about products and you'll need to analyze each review and identify if it is positive or negative.</prompt>

Big-Bench Hard

- EvoPrompt obtains better prompts for all 22 tasks.
- DE version is generally a good choice for challenging tasks.



NLU tasks

- Compared with previous works and human written instructions, EvoPrompt (GA and DE) delivers significantly better results.
- When the initial prompts are not of high quality, DE evades local optima

Method	SST-2	CR	MR	SST-5	AG’s News	TREC	Subj	Avg.
MI (Zhang et al., 2023b)	93.68	**91.40**	88.75	42.90	70.63	50.60	49.75	71.07
NI (Mishra et al., 2022c)	92.86	90.90	89.60	48.64	48.89	55.00	52.55	68.21
PromptSource (Bach et al., 2022)	93.03	-	-	-	45.43	36.20	-	-
APE (Zhou et al., 2022)	93.45(0.14)	91.13(0.45)	89.98(0.29)	46.32(0.49)	71.76(2.81)	58.73(1.37)	64.18(0.59)	73.80
APO (Pryzant et al., 2023)	93.87(0.39)	91.20(0.04)	89.85(0.35)	-	-	-	70.55(1.02)	-
EvoPrompt (GA)	**95.13**(0.21)	91.27(0.06)	90.07(0.25)	**49.91**(0.61)	72.81(0.61)	**64.00**(0.16)	70.55(2.58)	76.25
EvoPrompt (DE)	94.75(0.21)	**91.40**(0.04)	**90.22**(0.09)	49.89(1.73)	**73.82**(0.35)	63.73(1.54)	**75.55**(2.26)	**77.05**

Analysis

- **Importance** of **Prompt 3** in **DE:** the best prompt as Prompt 3 is more effective
- **DE or GA?**
  - When starting from top-performing initialization, GA is better.
  - When the initialization is poor, DE is a better choice when the available manual prompts are not of high quality.

Mutation	Prompt 3	Subj	ASSET
Diff	best	**75.55**(2.26)	**46.21**(0.27)
All	best	69.87(0.82)	45.73(0.45)
Diff	random	69.82(2.47)	45.89(0.37)
Diff	eliminate	69.07(4.21)	45.90(0.23)

Initialization	GA	DE
bottom-10	47.80(0.92)	48.64(0.15)
random-10	49.34(0.53)	**50.03**(1.08)
random-5 + var-5	49.84(1.49)	49.53(1.04)
top-10	49.62(1.00)	49.61(2.30)
top-5 + var-5	**49.91**(0.61)	49.89(1.73)

Paper

Code

ICLR