EMNLP 2023

Robust Prompt Optimization for Large Language Models Against Distribution Shifts

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, Tat-Seng Chua

Background

Nowadays, many powerful Large Language Models(LLM) are in the form of black-box API.

Background

Nowadays, many powerful Large Language Models(LLM) are in the form of black-box API.

Writing the best prompt for downstream task is important but difficult and laborious.

Background

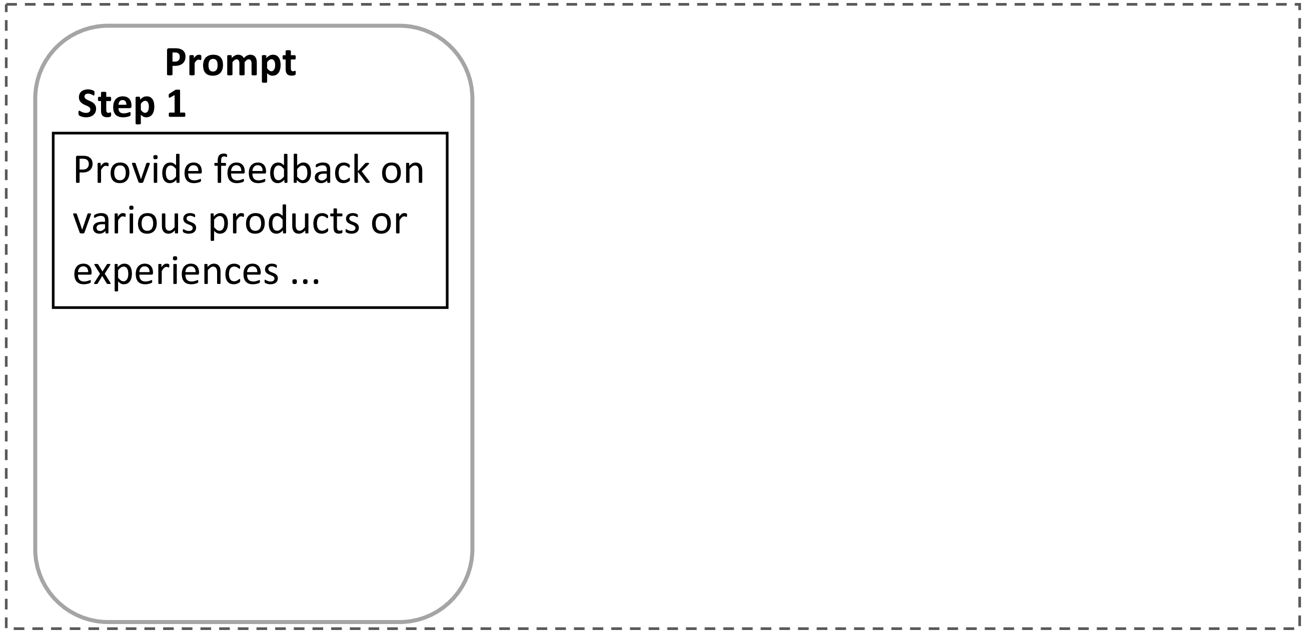
Nowadays, many powerful Large Language Models(LLM) are in the form of black-box API.

Writing the best prompt for downstream task is important but difficult and laborious.

To automatically obtain good prompts for certain tasks on black-box API LLMs?

Existing solution: gradient-free prompt optimization

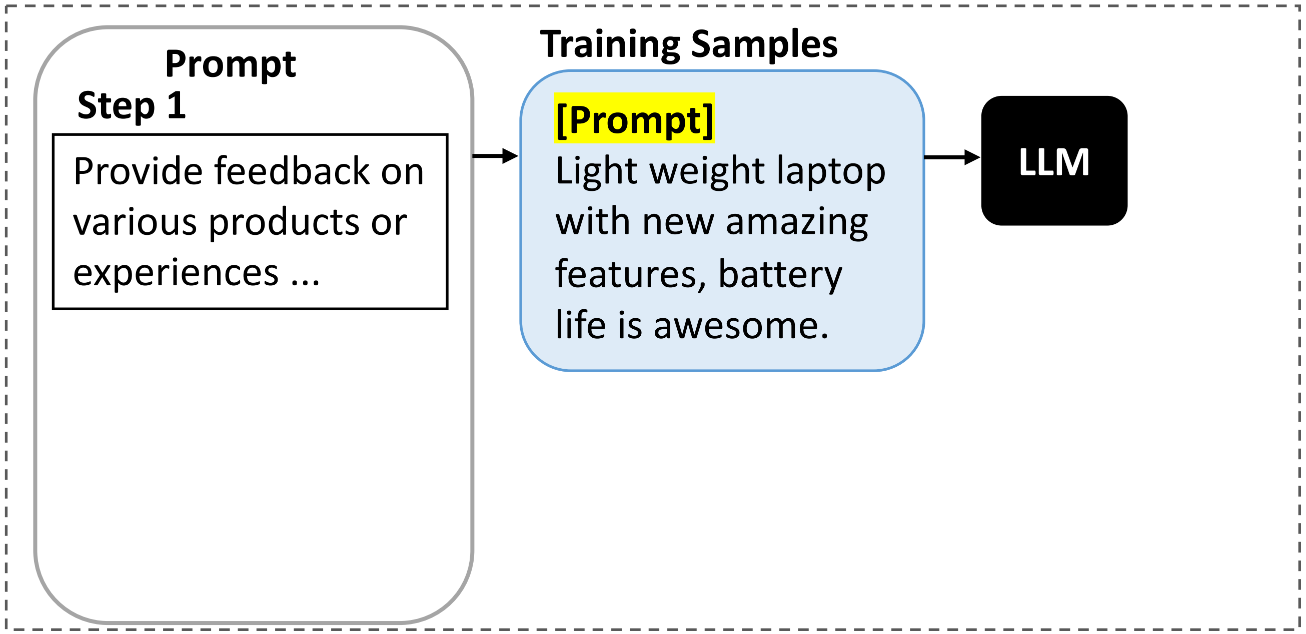
Background



An example pipeline of gradient-free prompt optimization for black-box API LLM.

Representative approaches: APE (Zhou et al., 2023) , APO (Pryzant et al., 2023).

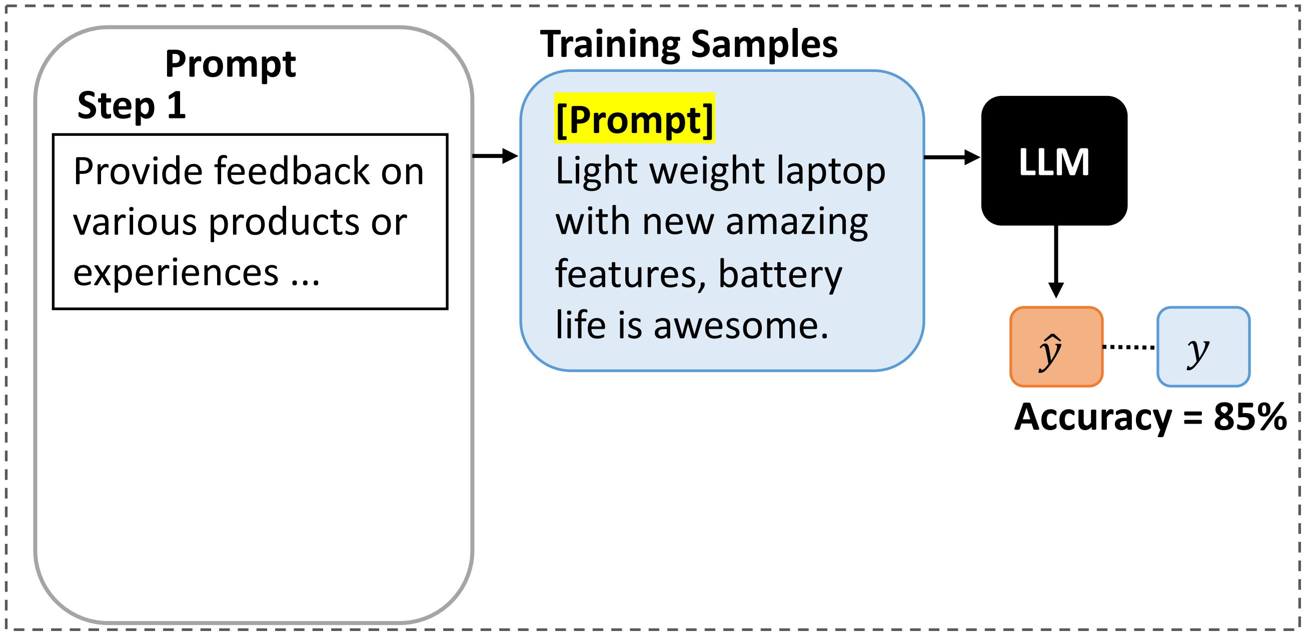
Background



An example pipeline of gradient-free prompt optimization for black-box API LLM.

Representative approaches: APE (Zhou et al., 2023) , APO (Pryzant et al., 2023).

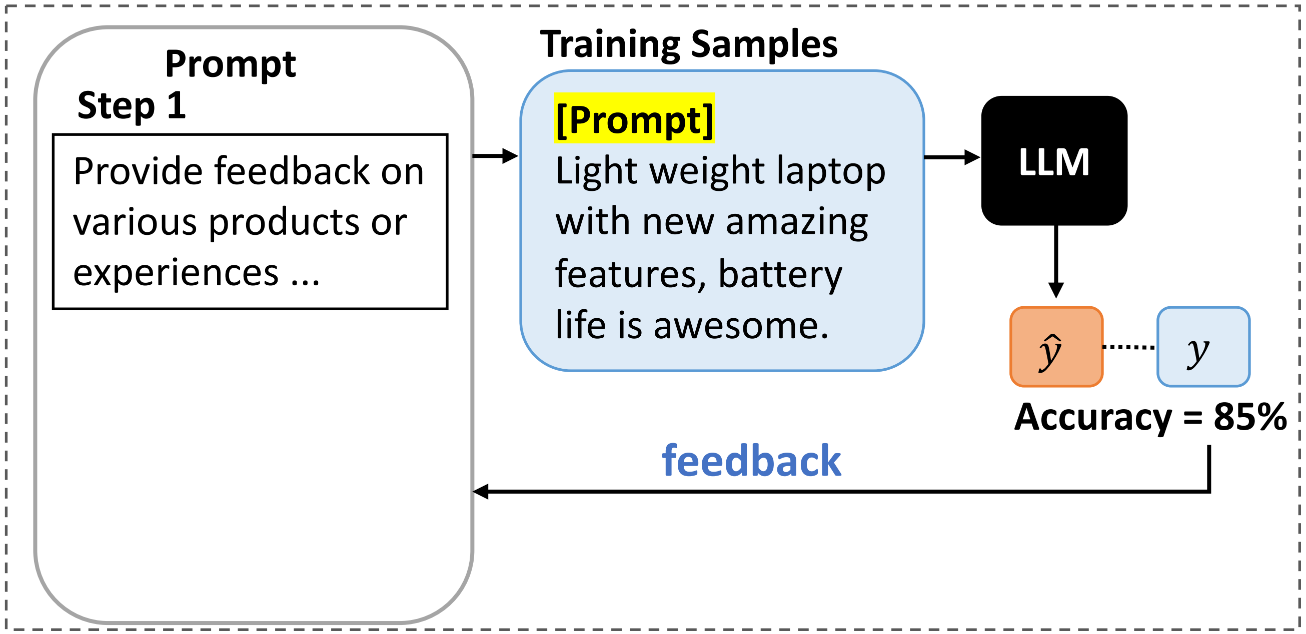
Background



An example pipeline of gradient-free prompt optimization for black-box API LLM.

Representative approaches: APE (Zhou et al., 2023) , APO (Pryzant et al., 2023).

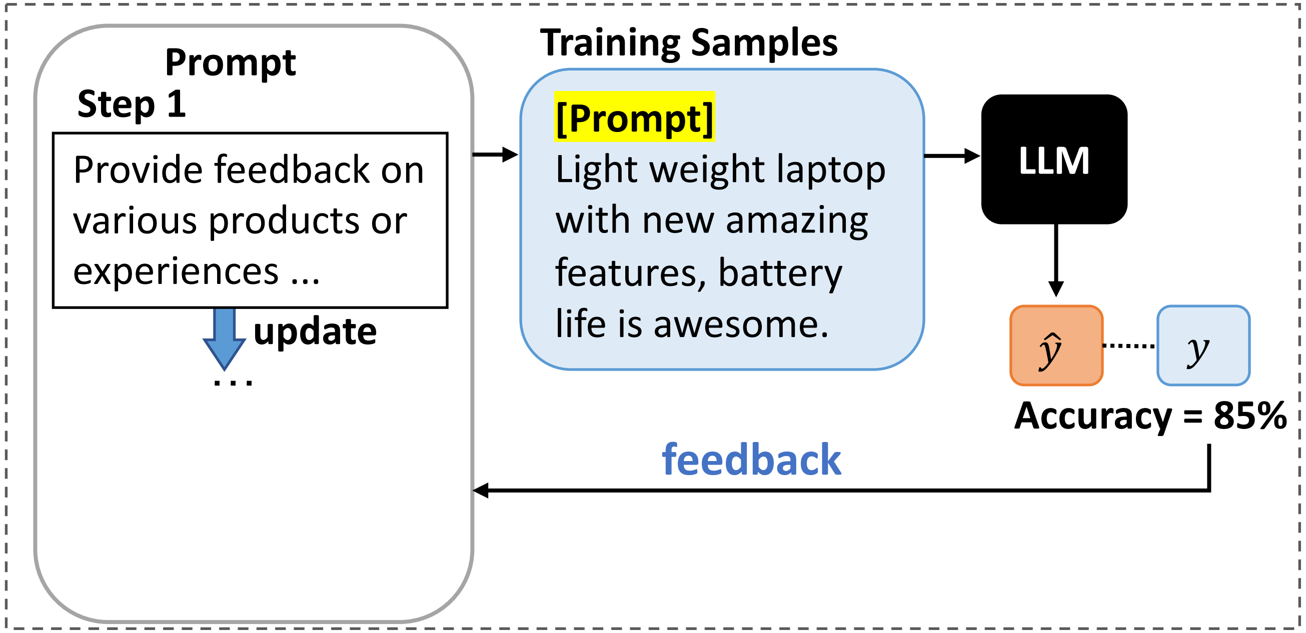
Background



An example pipeline of gradient-free prompt optimization for black-box API LLM.

Representative approaches: APE (Zhou et al., 2023) , APO (Pryzant et al., 2023).

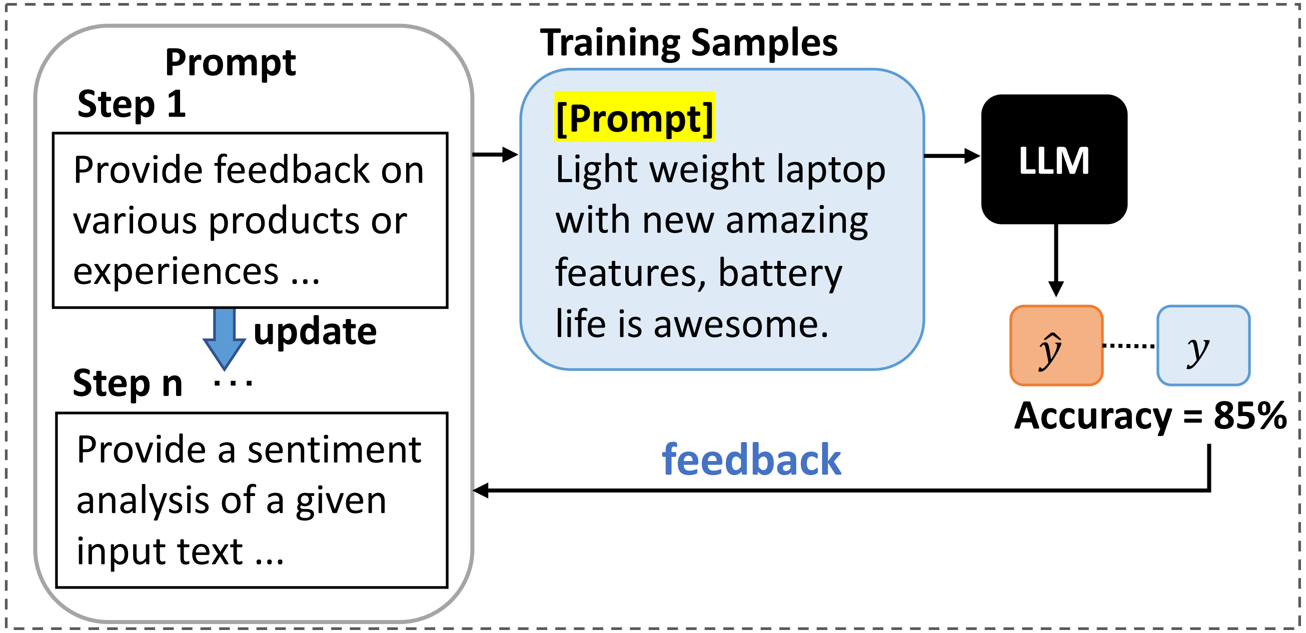
Background



An example pipeline of gradient-free prompt optimization for black-box API LLM.

Representative approaches: APE (Zhou et al., 2023) , APO (Pryzant et al., 2023).

Background

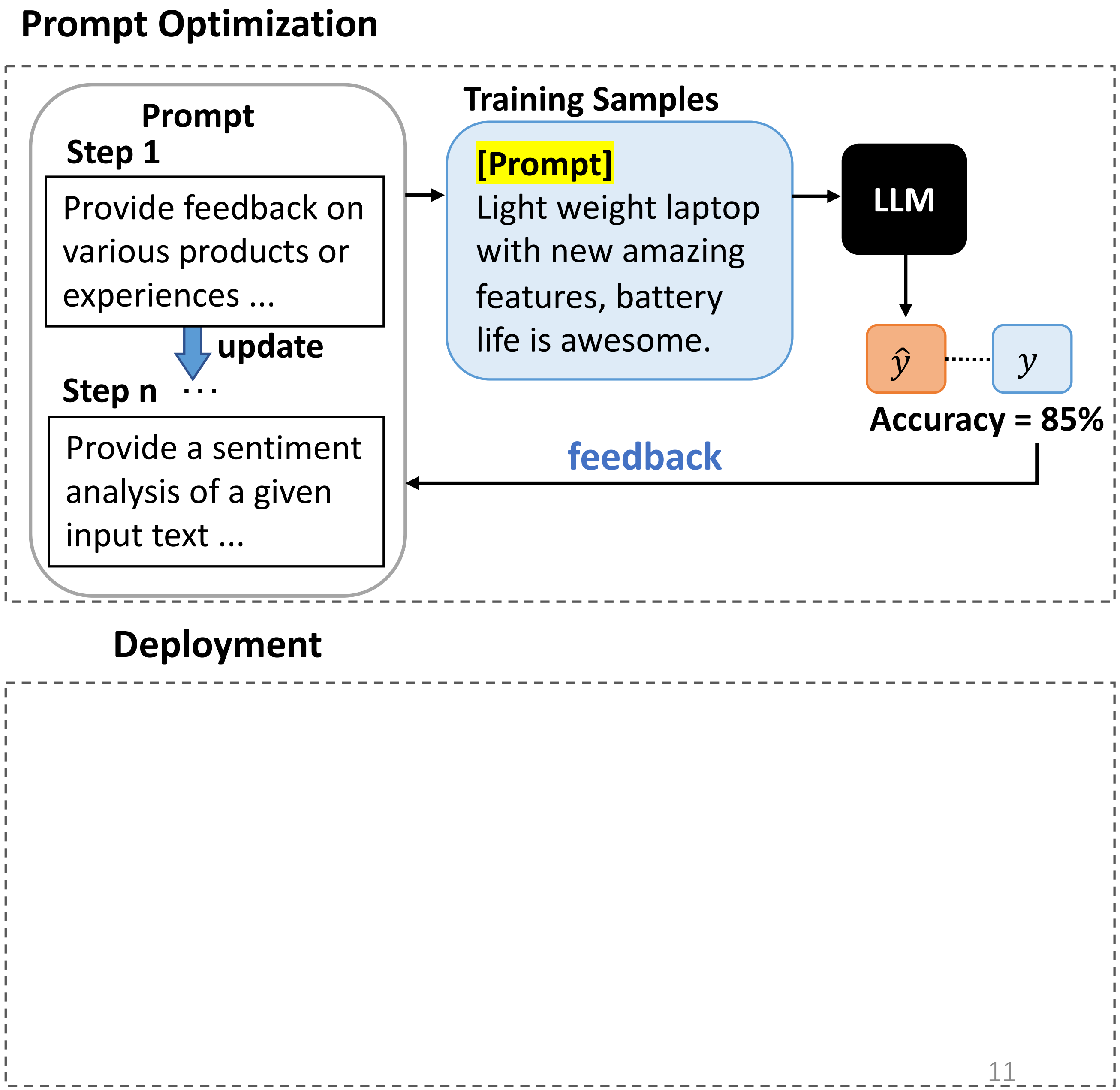


An example pipeline of gradient-free prompt optimization for black-box API LLM.

Representative approaches: APE (Zhou et al., 2023) , APO (Pryzant et al., 2023).

Motivation

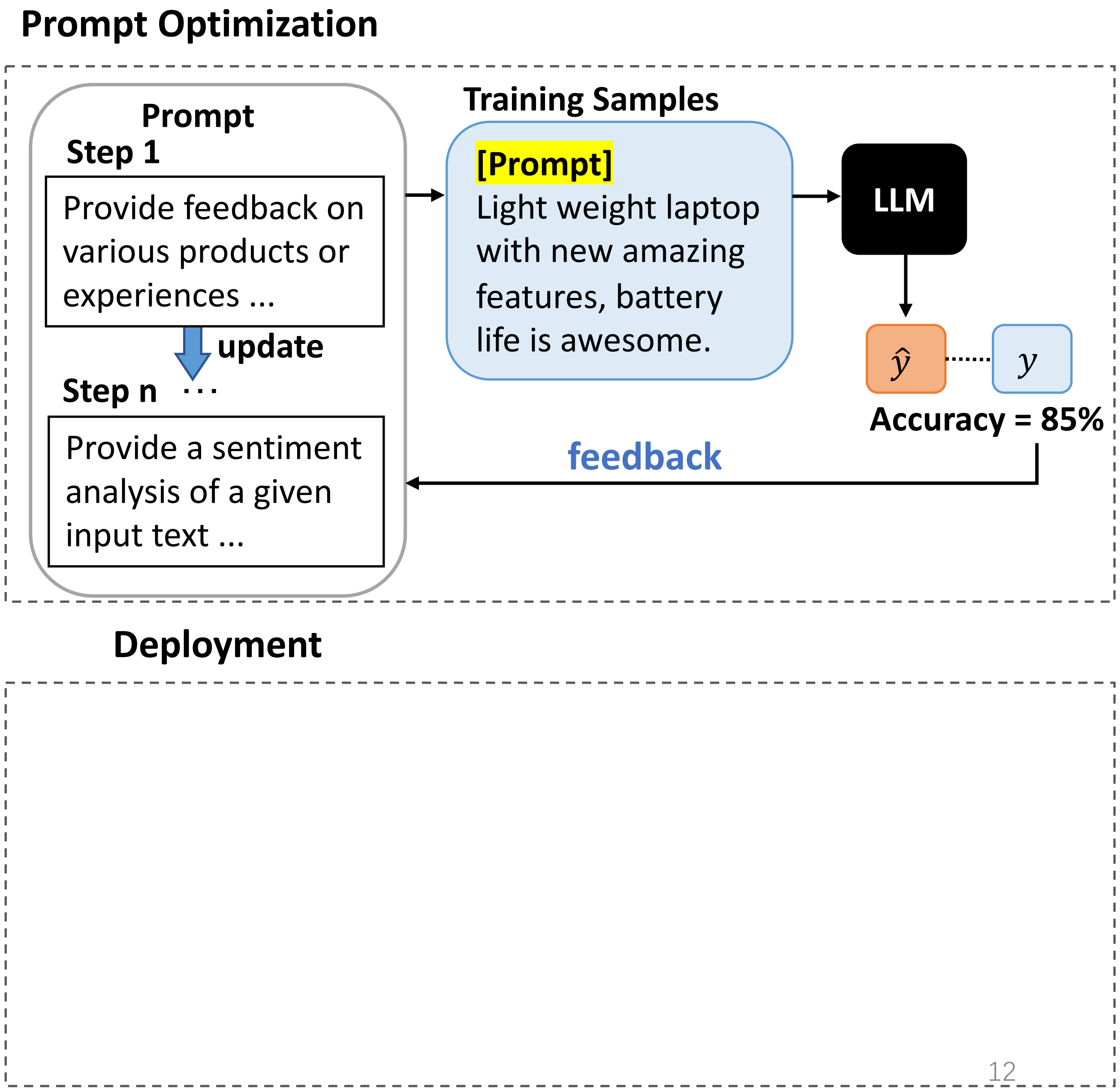
Current gradient-free prompt optimization methods ignore distribution shifts.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

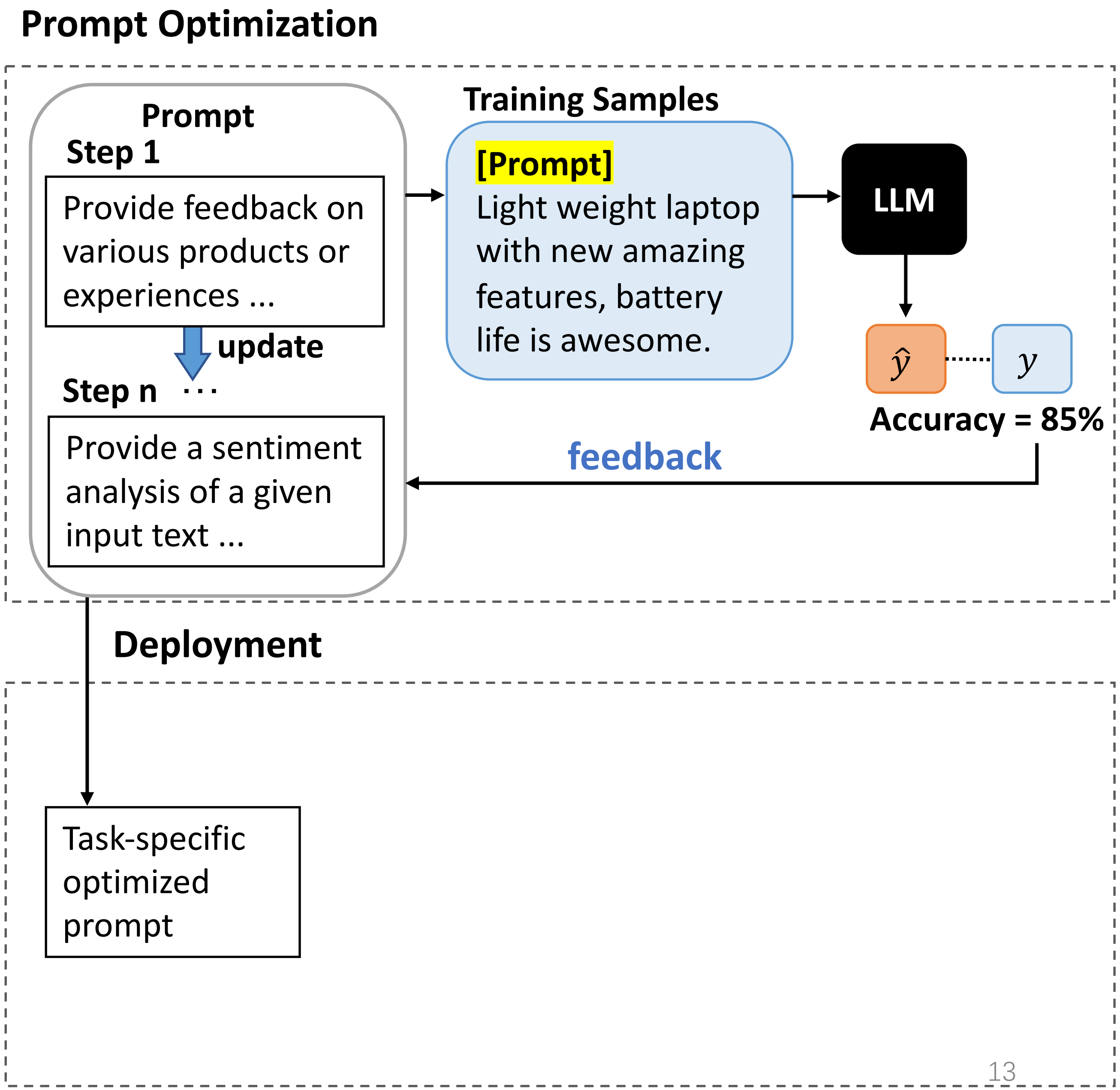
The data LLM serves may differ from the labeled data for prompt optimization.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

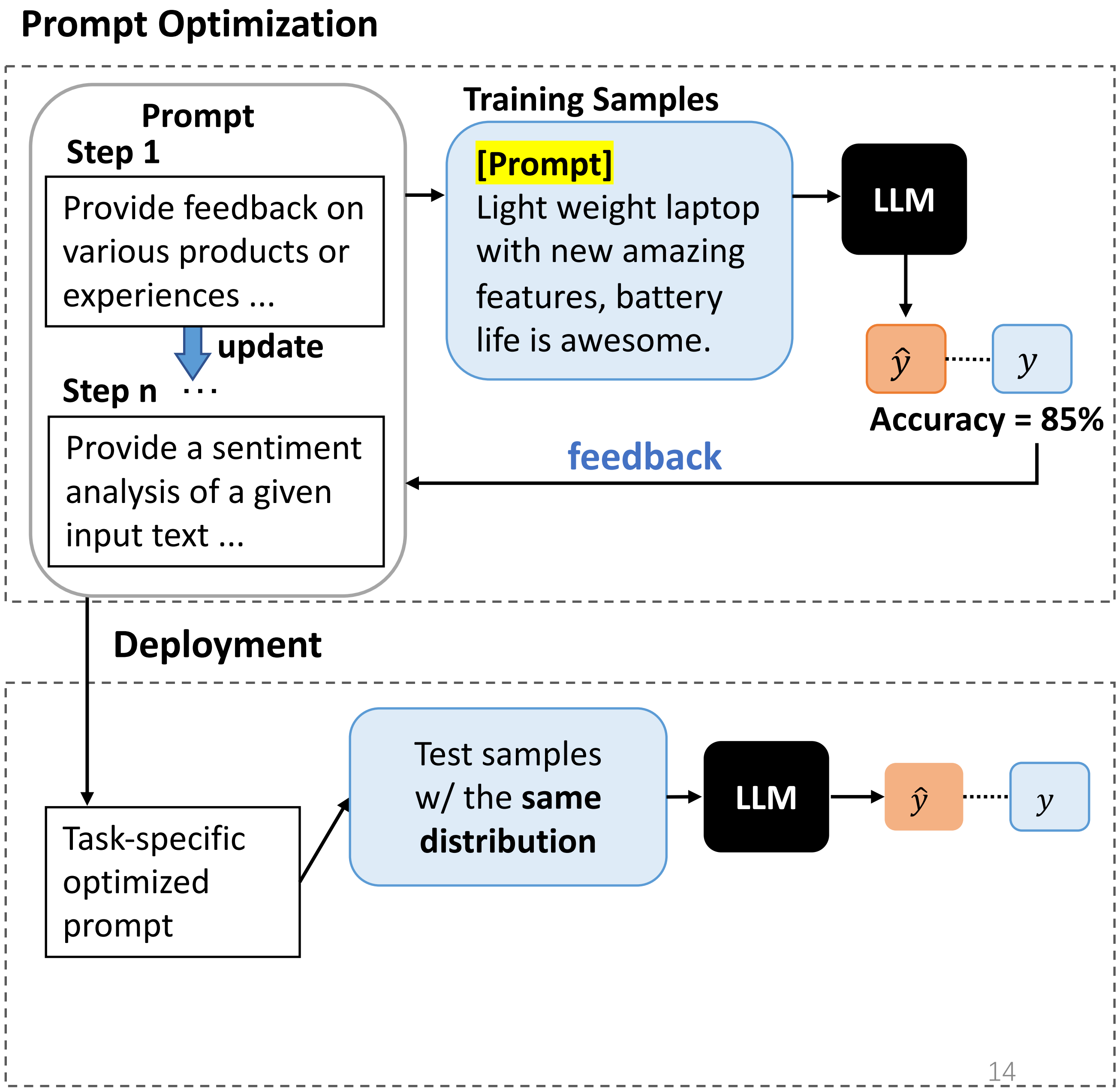
The data LLM serves may differ from the labeled data for prompt optimization.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

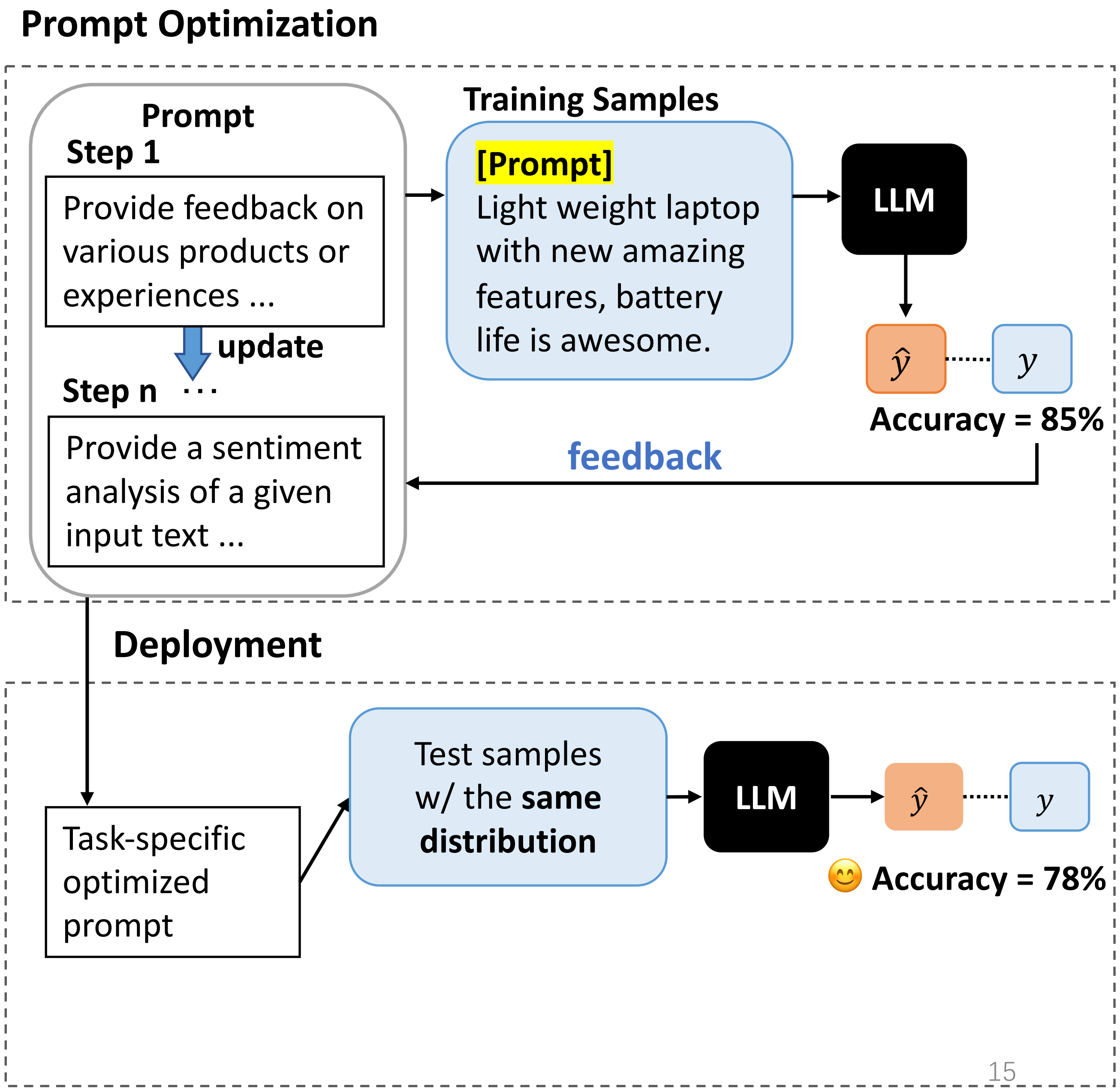
The data LLM serves may differ from the labeled data for prompt optimization.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

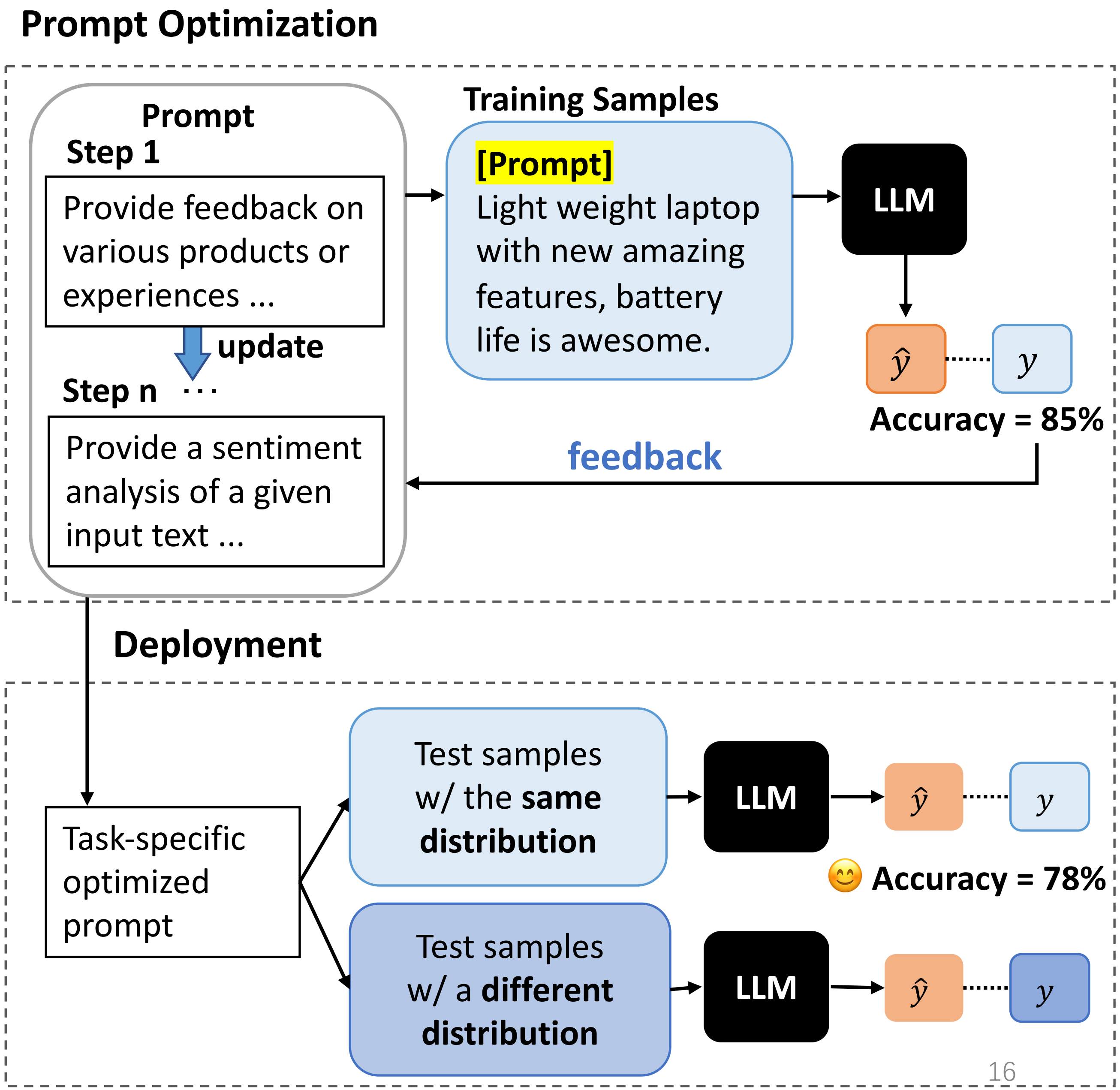
The data LLM serves may differ from the labeled data for prompt optimization.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

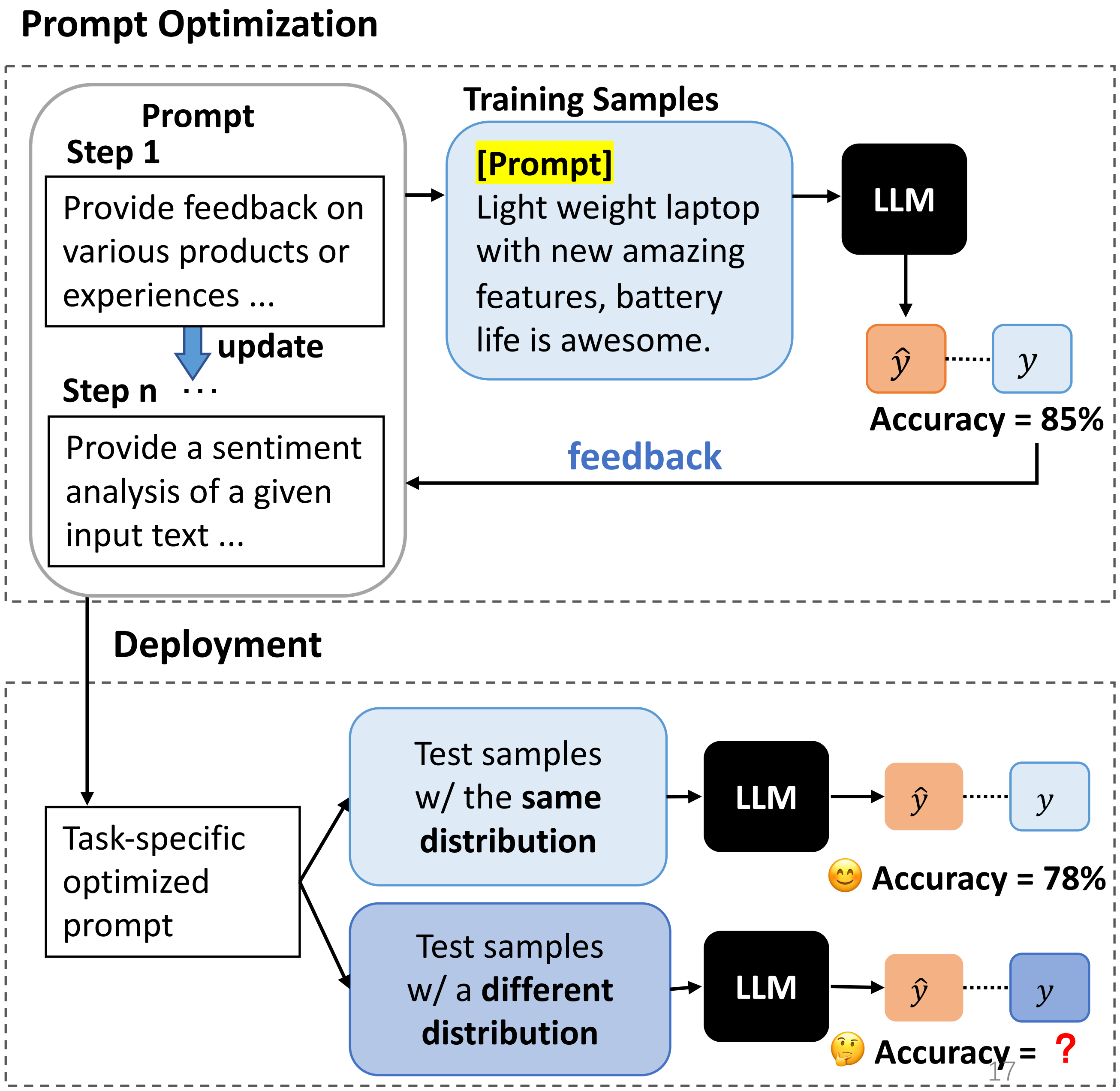
The data LLM serves may differ from the labeled data for prompt optimization.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

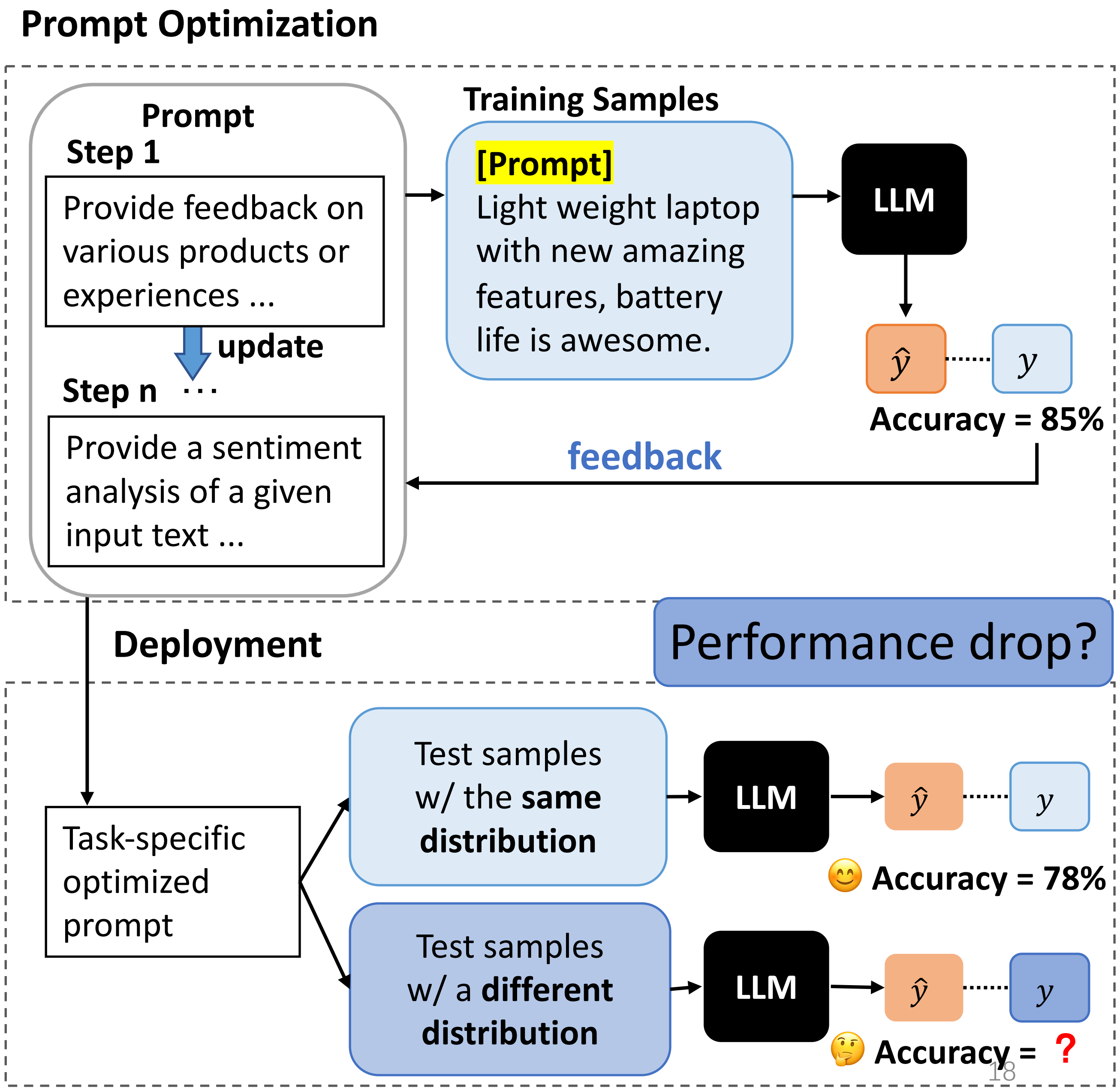
The data LLM serves may differ from the labeled data for prompt optimization.



Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

The data LLM serves may differ from the labeled data for prompt optimization.

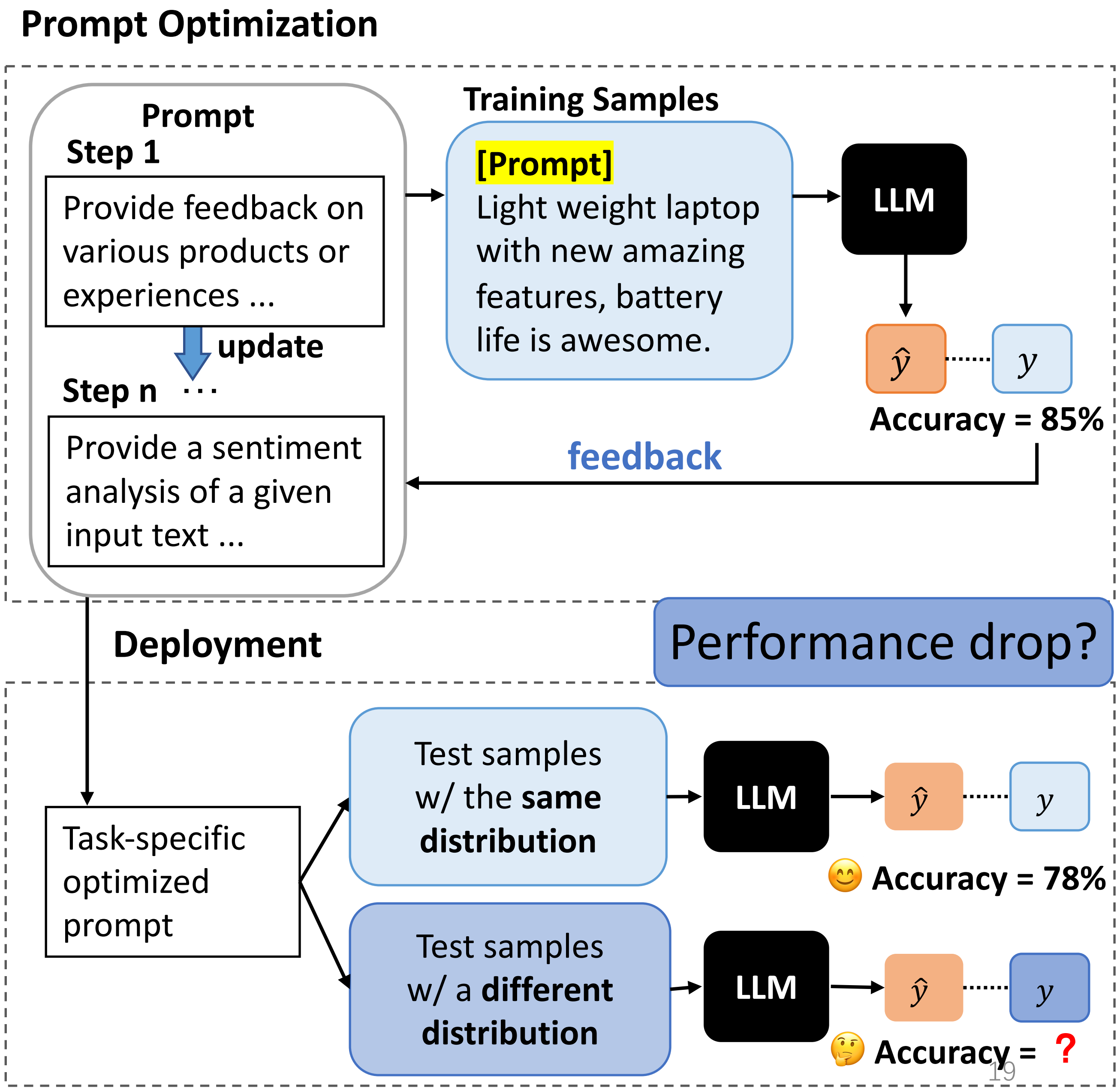


Motivation

Current gradient-free prompt optimization methods ignore distribution shifts.

The data LLM serves may differ from the labeled data for prompt optimization.

Contribution 1: We reveal the robustness issue of prompt optimization against distribution shifts.



Motivation

Real world NLP applications often encounter distribution shifts between collected and served data.

e.g., new user groups with distinct linguistic habits in customer review analysis.

Motivation

Real world NLP applications often encounter distribution shifts between collected and served data.

e.g., new user groups with distinct linguistic habits in customer review analysis.

Prompt optimization in consideration of distribution shifts,

A: Keep labeling new data and obtaining new prompts for different groups?

B: A robust task-specific prompt (this work).

Motivation

Real world NLP applications often encounter distribution shifts between collected and served data.

e.g., new user groups with distinct linguistic habits in customer review analysis.

Prompt optimization in consideration of distribution shifts,

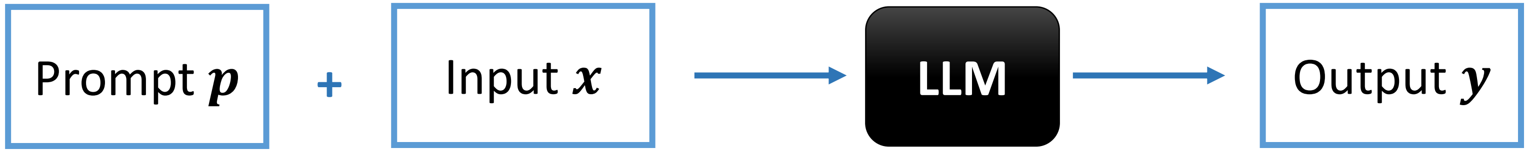
A: Keep labeling new data and obtaining new prompts for different groups?

B: A robust task-specific prompt (this work).

Contribution 2: We propose a new **robust prompt optimization problem**, and a **generalized prompt optimization framework** to solve the problem.

Preliminary Experiments

In the zero-shot setting,



Preliminary Experiments

In the zero-shot setting,

Prompt $\boldsymbol{p}$ + Input $\boldsymbol{x}$ → **LLM** → Output $\boldsymbol{y}$

Given a dataset $\{(\boldsymbol{x}, \boldsymbol{y})\}$ following distribution P, the goal of prompt optimization is to obtain

Preliminary Experiments

In the zero-shot setting,

Prompt $\boldsymbol{p}$ + Input $\boldsymbol{x}$ → LLM → Output $\boldsymbol{y}$

Given a dataset $\{(\boldsymbol{x}, \boldsymbol{y})\}$ following distribution P, the goal of prompt optimization is to obtain

$$\boxed{\boldsymbol{p}^o = argmax_{\boldsymbol{p} \in \mathcal{Z}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P}[r(LLM(\boldsymbol{p}, \boldsymbol{x}), \boldsymbol{y})]}$$

$\mathcal{Z}$: prompt optimization space; r: evaluation metric

Preliminary Experiments

Source group
$\{(\boldsymbol{x}_S, \boldsymbol{y}_S)\} \sim P_s$

Target group
$\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t \neq P_s$

Preliminary Experiments

Prompt optimization

Source group
$\{(\boldsymbol{x}_s, \boldsymbol{y}_s)\} \sim P_s$ $\longrightarrow$

Prompt optimization

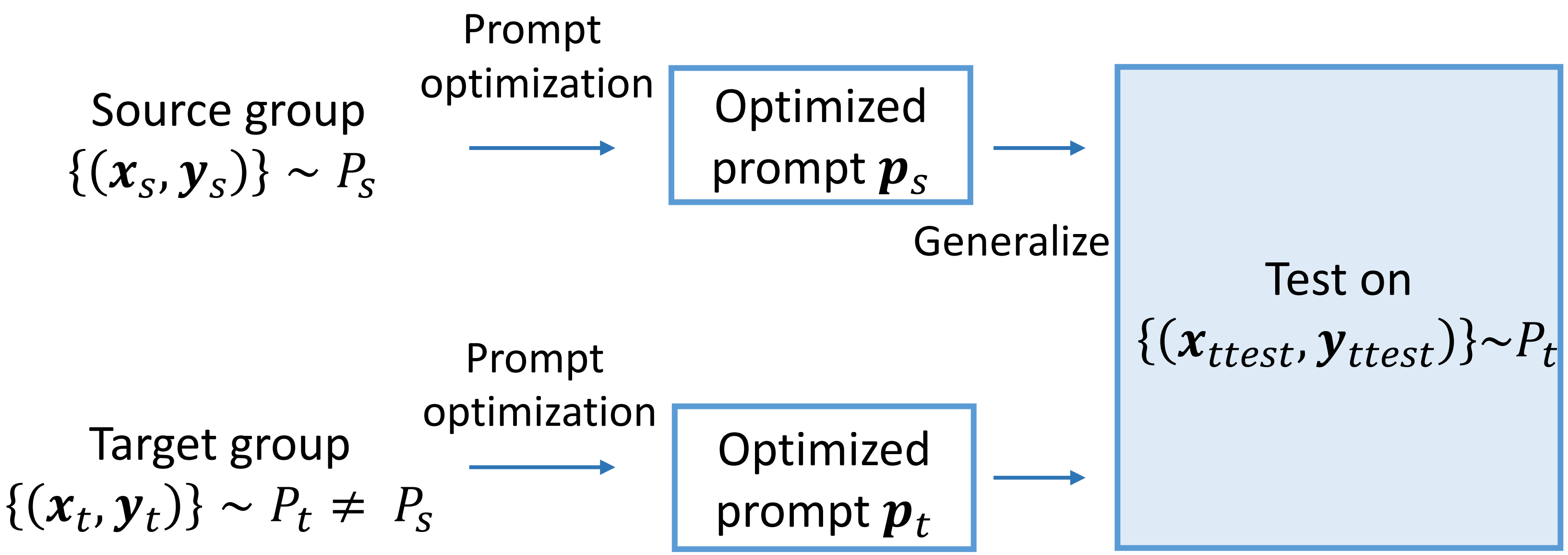
Target group
$\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t \neq P_s$ $\longrightarrow$

Preliminary Experiments

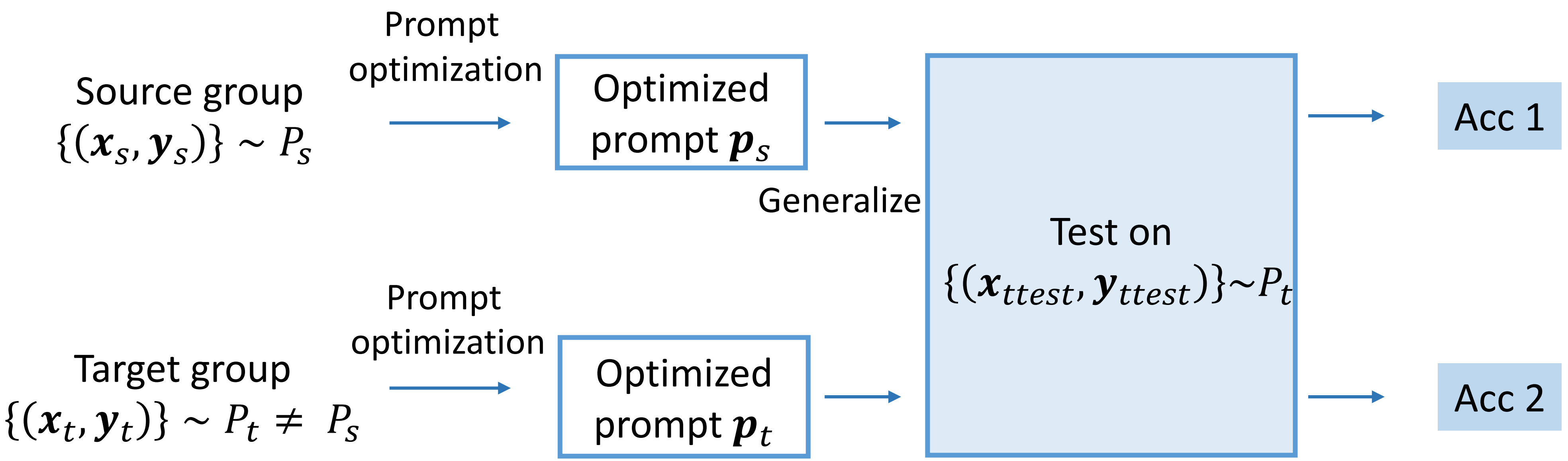
Source group $\{(\boldsymbol{x}_s, \boldsymbol{y}_s)\} \sim P_s$ $\xrightarrow{\text{Prompt optimization}}$ $\boxed{\text{Optimized prompt } \boldsymbol{p}_s}$

Target group $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t \neq P_s$ $\xrightarrow{\text{Prompt optimization}}$ $\boxed{\text{Optimized prompt } \boldsymbol{p}_t}$

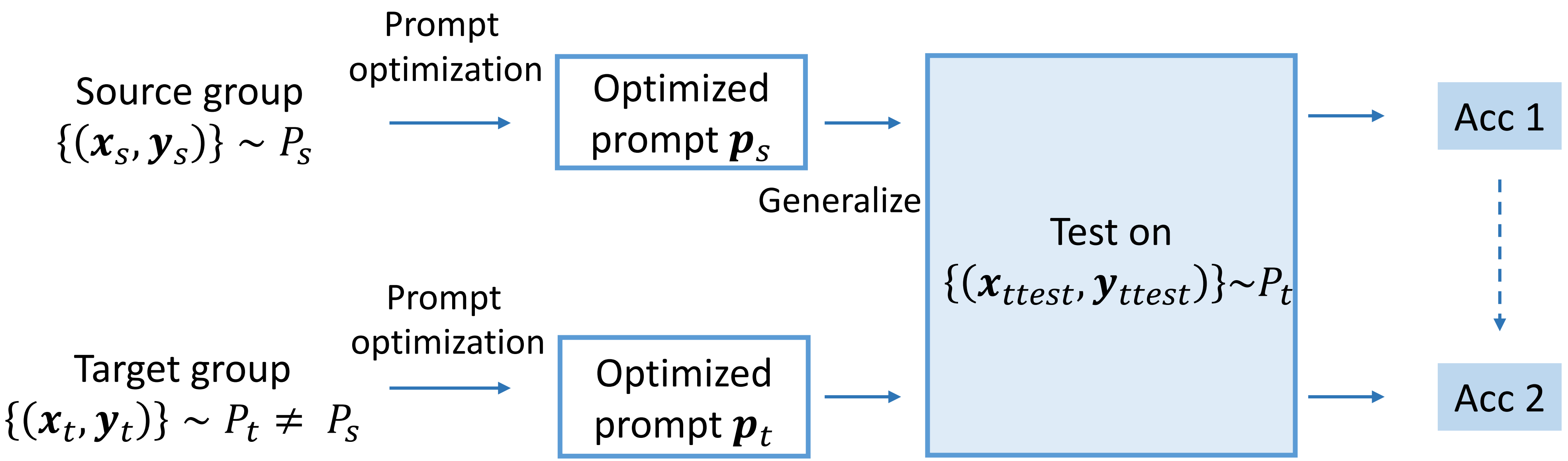
Preliminary Experiments



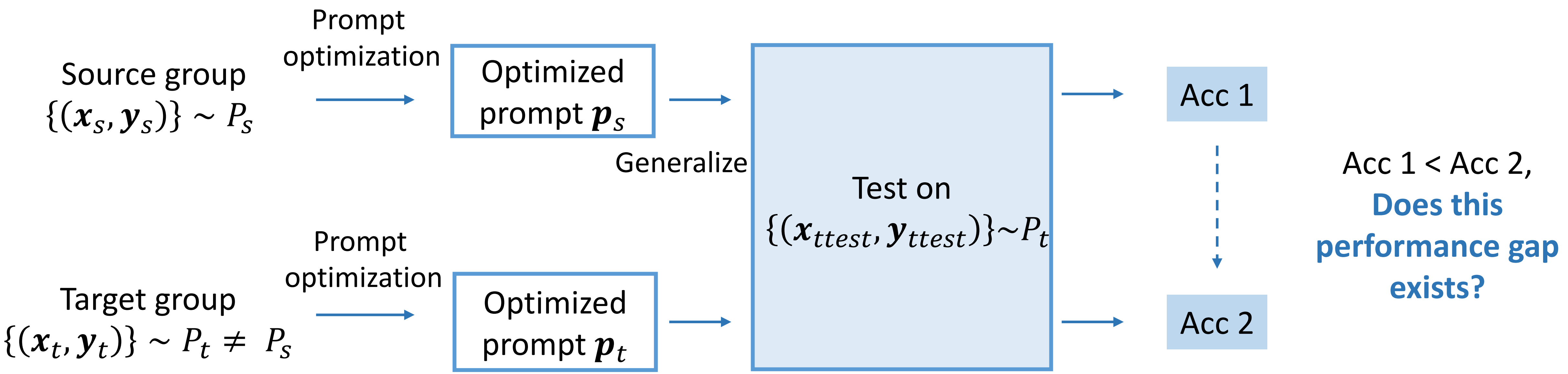
Preliminary Experiments



Preliminary Experiments



Preliminary Experiments



Preliminary Experiments

- 16 datasets from 6 NLP tasks
- **Different datasets of the same task as source and target groups**.
- APE (Zhou et al. 2023) as prompt optimization method
- *gpt-3.5-turbo-0301*
- Zero-shot
- Average with five runs

Dataset	Task	Label Example	Distribution shifts	Evaluation Metric
Yelp	Sentiment Analysis	Positive, Neutral, Negative	Different topics	Acc
Flipkart				
IMDB				
Amazon				
MNLI	NLI	entailment, neutral, contradiction	Adversarial dataset	
ANLI				
RTE	Textual Entailment	entailment, non−entailment	OOD dataset	
HANS				
SocialIQA	Commonsense QA	A, B, C, D	Different topics	
PIQA				
OpenbookQA				
DSTC7	Multi-turn dialog response selection	A, B, C, D	Different topics	
Ubunt				
MuTual				
DROP Spans	Numerical QA	e.g., 78.9	Different answer types	F1
DROP number		e.g., 3 March 1992		

Preliminary Experiments

- 16 datasets from 6 NLP tasks
- **Different datasets of the same task as source and target groups**.
- APE (Zhou et al. 2023) as prompt optimization method
- *gpt-3.5-turbo-0301*
- Zero-shot
- Average with five runs

Dataset	Task	Label Example	Distribution shifts	Evaluation Metric
Yelp	Sentiment Analysis	Positive, Neutral, Negative	Different topics	Acc
Flipkart				
IMDB				
Amazon				
MNLI	NLI	entailment, neutral, contradiction	Adversarial dataset	
ANLI				
RTE	Textual Entailment	entailment, non−entailment	OOD dataset	
HANS				
SocialIQA	Commonsense QA	A, B, C, D	Different topics	
PIQA				
OpenbookQA				
DSTC7	Multi-turn dialog response selection	A, B, C, D	Different topics	
Ubunt				
MuTual				
DROP Spans	Numerical QA	e.g., 78.9	Different answer types	F1
DROP number		e.g., 3 March 1992		

Preliminary Experiments

Significant generalization performance gaps between some data groups.

Source \ Target	Yelp	Flipkart	IMDB	Amazon
Yelp	**79.7 ± 0.7**	78.4 ± 1.9	87.1 ± 1.9	88.4 ± 1.9
Flipkart	69.1 ± 8.7	**85.1 ± 2.9**	85.2 ± 9.4	85.9 ± 12.5
IMDB	71.1 ± 8.2	76.9 ± 13.4	**91.9 ± 0.9**	90.4 ± 5.2
Amazon	75.5 ± 1.5	**85.6 ± 2.1**	**91.5 ± 0.8**	**93.5 ± 1.4**

(a) Sentiment analysis

Source \ Target	SocialIQA	PIQA	OpenbookQA
SocialIQA	75.6 ± 1.4	82.0 ± 6.0	71.2 ± 5.2
PIQA	68.9 ± 6.9	83.6 ± 2.9	69.2 ± 5.1
OpenbookQA	**79.9 ± 1.0**	**84.5 ± 1.6**	**80.1 ± 2.4**

(b) Commonsense QA

Source \ Target	Number	Spans
Number	51.9 ± 2.8	20.1 ± 1.3
Spans	**57.7 ± 2.9**	**63.1 ± 2.2**

(c) DROP

Preliminary Experiments

Significant generalization performance gaps between some data groups.

Compare each column

Source \ Target	Yelp	Flipkart	IMDB	Amazon
Yelp	**79.7 ± 0.7**	78.4 ± 1.9	87.1 ± 1.9	88.4 ± 1.9
Flipkart	69.1 ± 8.7	**85.1 ± 2.9**	85.2 ± 9.4	85.9 ± 12.5
IMDB	71.1 ± 8.2	76.9 ± 13.4	**91.9 ± 0.9**	90.4 ± 5.2
Amazon	75.5 ± 1.5	**85.6 ± 2.1**	**91.5 ± 0.8**	**93.5 ± 1.4**

(a) Sentiment analysis

Source \ Target	SocialIQA	PIQA	OpenbookQA
SocialIQA	75.6 ± 1.4	82.0 ± 6.0	71.2 ± 5.2
PIQA	68.9 ± 6.9	83.6 ± 2.9	69.2 ± 5.1
OpenbookQA	**79.9 ± 1.0**	**84.5 ± 1.6**	**80.1 ± 2.4**

(b) Commonsense QA

Source \ Target	Number	Spans
Number	51.9 ± 2.8	20.1 ± 1.3
Spans	**57.7 ± 2.9**	**63.1 ± 2.2**

(c) DROP

Preliminary Experiments

Significant generalization performance gaps between some data groups.

Are prompts optimized by existing gradient-free methods robust to distribution shifts?

Compare each column

Source \ Target	Yelp	Flipkart	IMDB	Amazon
Yelp	**79.7 ± 0.7**	78.4 ± 1.9	87.1 ± 1.9	88.4 ± 1.9
Flipkart	69.1 ± 8.7	**85.1 ± 2.9**	85.2 ± 9.4	85.9 ± 12.5
IMDB	71.1 ± 8.2	76.9 ± 13.4	**91.9 ± 0.9**	90.4 ± 5.2
Amazon	75.5 ± 1.5	**85.6 ± 2.1**	**91.5 ± 0.8**	**93.5 ± 1.4**

(a) Sentiment analysis

Source \ Target	SocialIQA	PIQA	OpenbookQA
SocialIQA	75.6 ± 1.4	82.0 ± 6.0	71.2 ± 5.2
PIQA	68.9 ± 6.9	83.6 ± 2.9	69.2 ± 5.1
OpenbookQA	**79.9 ± 1.0**	**84.5 ± 1.6**	**80.1 ± 2.4**

(b) Commonsense QA

Source \ Target	Number	Spans
Number	51.9 ± 2.8	20.1 ± 1.3
Spans	**57.7 ± 2.9**	**63.1 ± 2.2**

(c) DROP

Preliminary Experiments

Significant generalization performance gaps between some data groups.

Are prompts optimized by existing gradient-free methods robust to distribution shifts?

Under some distribution shifts, No.

Compare each column

Source \ Target	Yelp	Flipkart	IMDB	Amazon
Yelp	**79.7** $\pm$ **0.7**	78.4 $\pm$ 1.9	87.1 $\pm$ 1.9	88.4 $\pm$ 1.9
Flipkart	69.1 $\pm$ 8.7	**85.1** $\pm$ **2.9**	85.2 $\pm$ 9.4	85.9 $\pm$ 12.5
IMDB	71.1 $\pm$ 8.2	76.9 $\pm$ 13.4	**91.9** $\pm$ **0.9**	90.4 $\pm$ 5.2
Amazon	75.5 $\pm$ 1.5	**85.6** $\pm$ **2.1**	**91.5** $\pm$ **0.8**	**93.5** $\pm$ **1.4**

(a) Sentiment analysis

Source \ Target	SocialIQA	PIQA	OpenbookQA
SocialIQA	75.6 $\pm$ 1.4	82.0 $\pm$ 6.0	71.2 $\pm$ 5.2
PIQA	68.9 $\pm$ 6.9	83.6 $\pm$ 2.9	69.2 $\pm$ 5.1
OpenbookQA	**79.9** $\pm$ **1.0**	**84.5** $\pm$ **1.6**	**80.1** $\pm$ **2.4**

(b) Commonsense QA

Source \ Target	Number	Spans
Number	51.9 $\pm$ 2.8	20.1 $\pm$ 1.3
Spans	**57.7** $\pm$ **2.9**	**63.1** $\pm$ **2.2**

(c) DROP

Preliminary Experiments

Significant generalization performance gaps between some data groups.

Are prompts optimized by existing gradient-free methods robust to distribution shifts?

Under some distribution shifts, No.

Goal: achieving robust prompt optimization against distribution shifts.

Compare each column

Source \ Target	Yelp	Flipkart	IMDB	Amazon
Yelp	**79.7** $\pm$ **0.7**	78.4 $\pm$ 1.9	87.1 $\pm$ 1.9	88.4 $\pm$ 1.9
Flipkart	69.1 $\pm$ 8.7	**85.1** $\pm$ **2.9**	85.2 $\pm$ 9.4	85.9 $\pm$ 12.5
IMDB	71.1 $\pm$ 8.2	76.9 $\pm$ 13.4	**91.9** $\pm$ **0.9**	90.4 $\pm$ 5.2
Amazon	75.5 $\pm$ 1.5	**85.6** $\pm$ **2.1**	**91.5** $\pm$ **0.8**	**93.5** $\pm$ **1.4**

(a) Sentiment analysis

Source \ Target	SocialIQA	PIQA	OpenbookQA
SocialIQA	75.6 $\pm$ 1.4	82.0 $\pm$ 6.0	71.2 $\pm$ 5.2
PIQA	68.9 $\pm$ 6.9	83.6 $\pm$ 2.9	69.2 $\pm$ 5.1
OpenbookQA	**79.9** $\pm$ **1.0**	**84.5** $\pm$ **1.6**	**80.1** $\pm$ **2.4**

(b) Commonsense QA

Source \ Target	Number	Spans
Number	51.9 $\pm$ 2.8	20.1 $\pm$ 1.3
Spans	**57.7** $\pm$ **2.9**	**63.1** $\pm$ **2.2**

(c) DROP

Preliminary Experiments

Generalization gap may not exist for some OOD and adversarial groups.

More analysis in the paper.

Source \ Target	MNLI	ANLI
MNLI	73.4 ± 1.0	45.4 ± 1.9
ANLI	73.3 ± 1.3	46.0 ± 1.5

(a) Natural language inference

Source \ Target	RTE	HANS
RTE	78.3 ± 0.8	67.2 ± 1.1
HANS	79.0 ± 0.8	68.4 ± 1.8

(b) Textual entailment

Source \ Target	DSTC7	Ubuntu Dialog	MuTual
DSTC7	58.4 ± 0.8	78.9 ± 0.3	74.2 ± 2.2
Ubuntu Dialog	56.9 ± 1.3	78.7 ± 0.5	74.4 ± 2.1
MuTual	52.2 ± 4.4	74.7 ± 6.0	76.7 ± 3.4

(c) Dialog

Robust Prompt Optimization Problem

Given

Goal

Robust Prompt Optimization Problem

Given

The source group $G_s = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ following a distribution P_s.

Goal

Robust Prompt Optimization Problem

Given

The source group $G_s = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ following a distribution P_s.

$\{\boldsymbol{x}_t\}$ in an **unlabeled** target group $G_t = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t$ $(P_t \neq P_s)$, where $\boldsymbol{y}_t$ is **unseen** during prompt optimization,

Goal

Robust Prompt Optimization Problem

Given

The source group $G_s = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ following a distribution P_s.

$\{\boldsymbol{x}_t\}$ in an **unlabeled** target group $G_t = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t$ $(P_t \neq P_s)$, where $\boldsymbol{y}_t$ is **unseen** during prompt optimization,

Goal

use $\{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ and $\{\boldsymbol{x}_t\}$ to optimize a **task-specific** prompt **robust to the samples from either** P_s or P_t.

Robust Prompt Optimization Problem

Given

The source group $G_s = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ following a distribution P_s.

$\{\boldsymbol{x}_t\}$ in an **unlabeled** target group $G_t = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t$ $(P_t \neq P_s)$, where $\boldsymbol{y}_t$ is **unseen** during prompt optimization,

≈ Collecting served inputs.

Goal

use $\{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ and $\{\boldsymbol{x}_t\}$ to optimize a **task-specific** prompt **robust to the samples from either** P_s or P_t.

Robust Prompt Optimization Problem

Given

The source group $G_s = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ following a distribution P_s.

$\{\boldsymbol{x}_t\}$ in an **unlabeled** target group $G_t = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t$ $(P_t \neq P_s)$, where $\boldsymbol{y}_t$ is **unseen** during prompt optimization,

≈ Collecting served inputs.

Goal

use $\{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ and $\{\boldsymbol{x}_t\}$ to optimize a **task-specific** prompt **robust to the samples from either** P_s or P_t.

Our solution

Utilizing LLM for labeling $\{\boldsymbol{x}_t\}$
to perform joint prompt optimization with G_s

Robust Prompt Optimization Problem

Given

The source group $G_s = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ following a distribution P_s.

$\{\boldsymbol{x}_t\}$ in an **unlabeled** target group $G_t = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\} \sim P_t$ $(P_t \neq P_s)$, where $\boldsymbol{y}_t$ is **unseen** during prompt optimization,

≈ Collecting served inputs.

Goal

use $\{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}$ and $\{\boldsymbol{x}_t\}$ to optimize a **task-specific** prompt **robust to the samples from either** P_s or P_t.

Our solution

Utilizing LLM for labeling $\{\boldsymbol{x}_t\}$
to perform joint prompt optimization with G_s

Gradient-free prompt optimization needs labels.

Generalized Prompt Optimization Framework

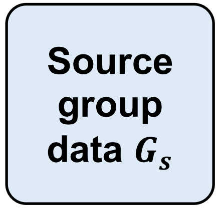
Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.

Deriving prompts from G_S.

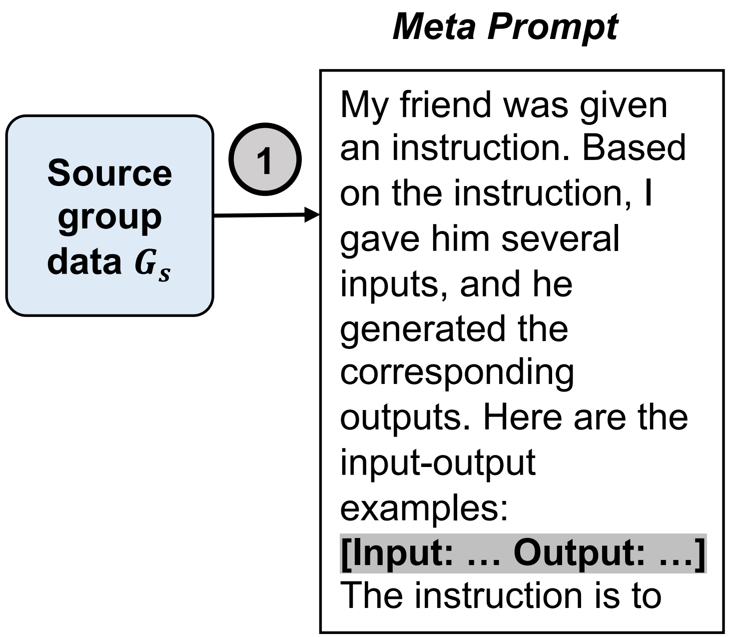
Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_S.



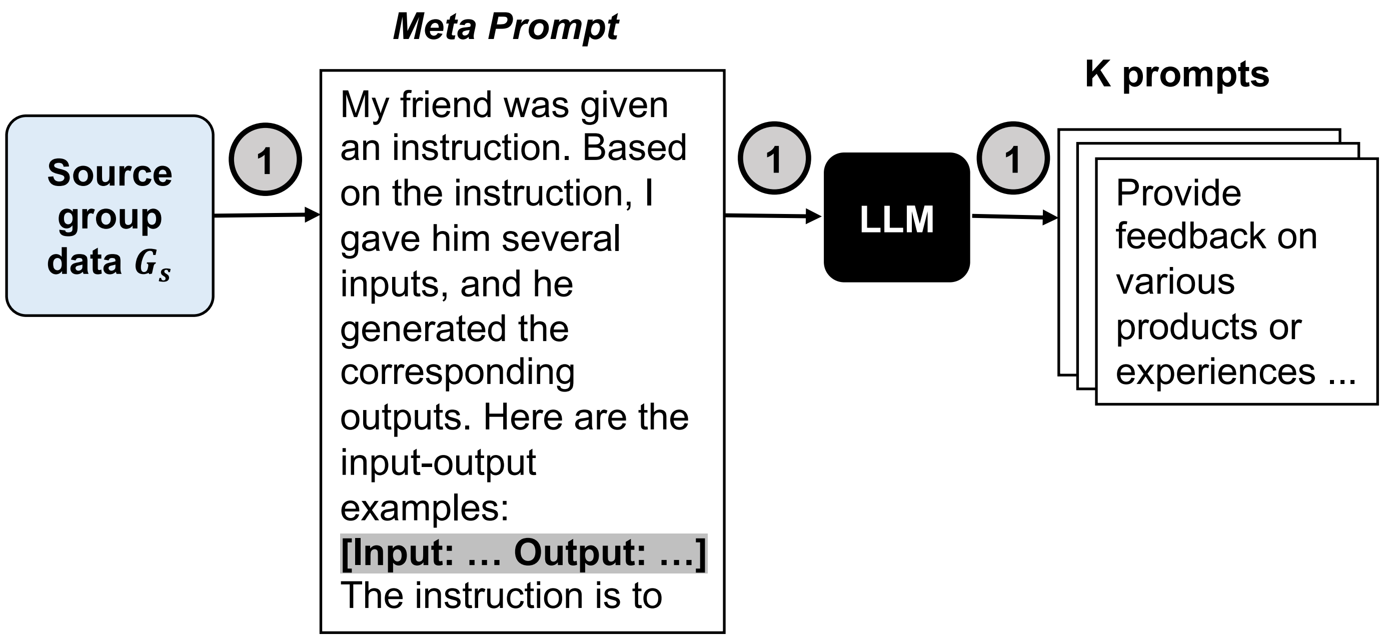
Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.



Generalized Prompt Optimization Framework

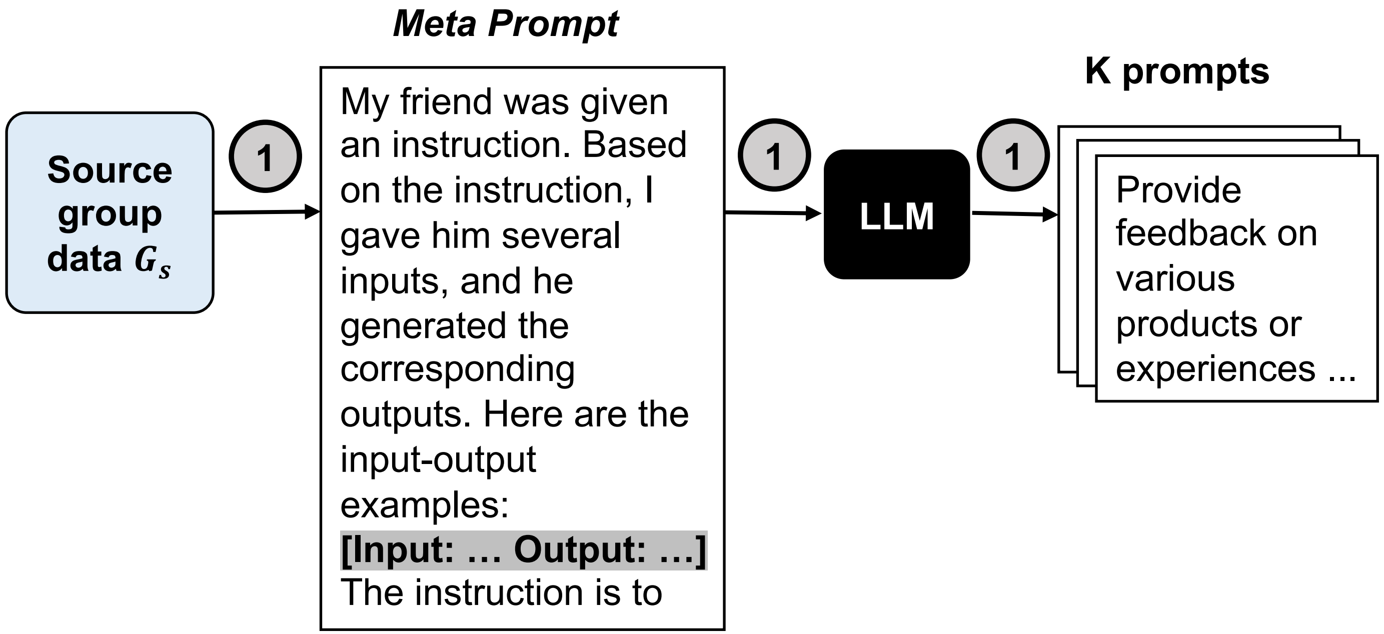
Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.



Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.

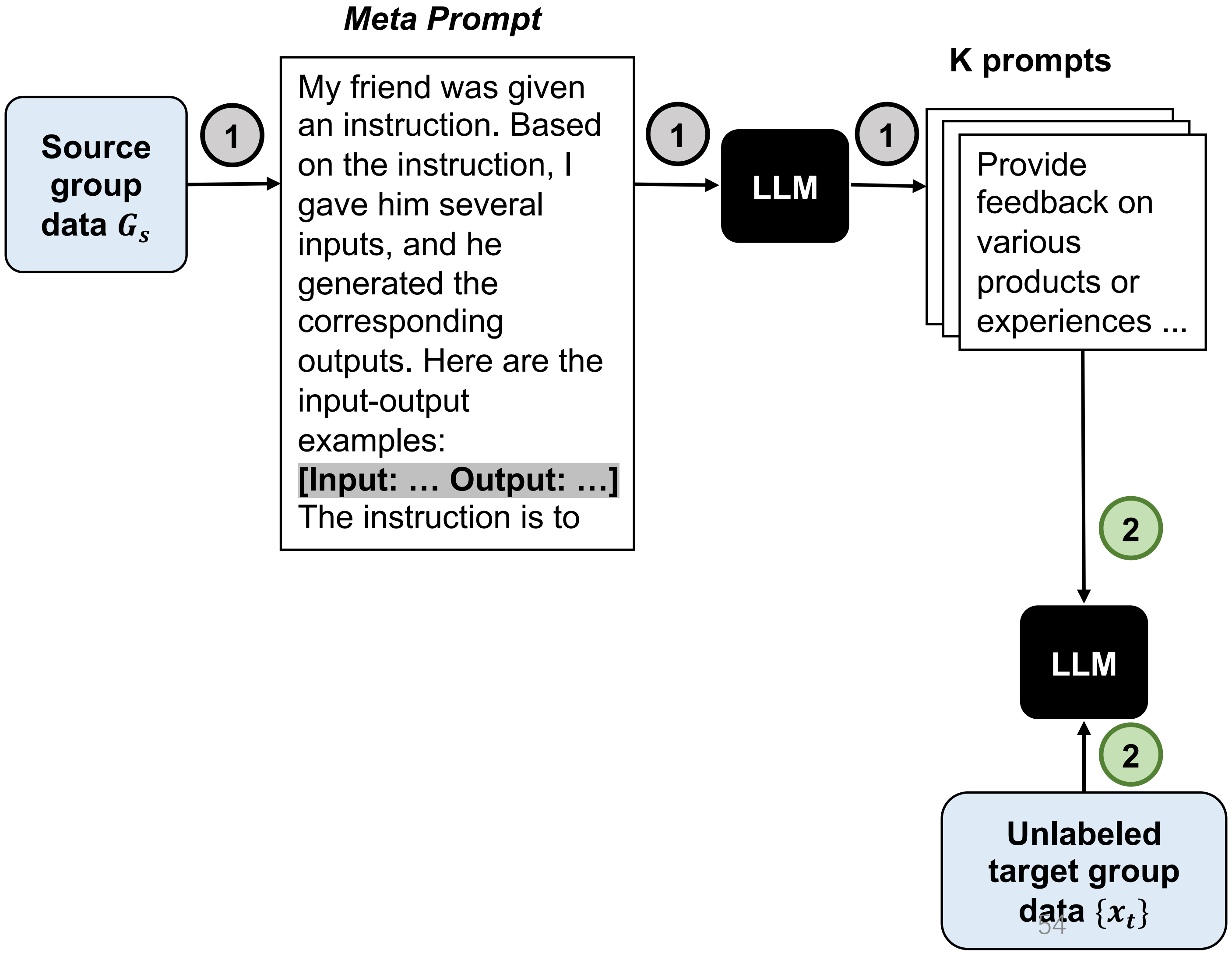
Step 2: Prompt Ensemble Labeling Strategy.
Labeling $\{x_t\}$ with ensembled prompts.



Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.

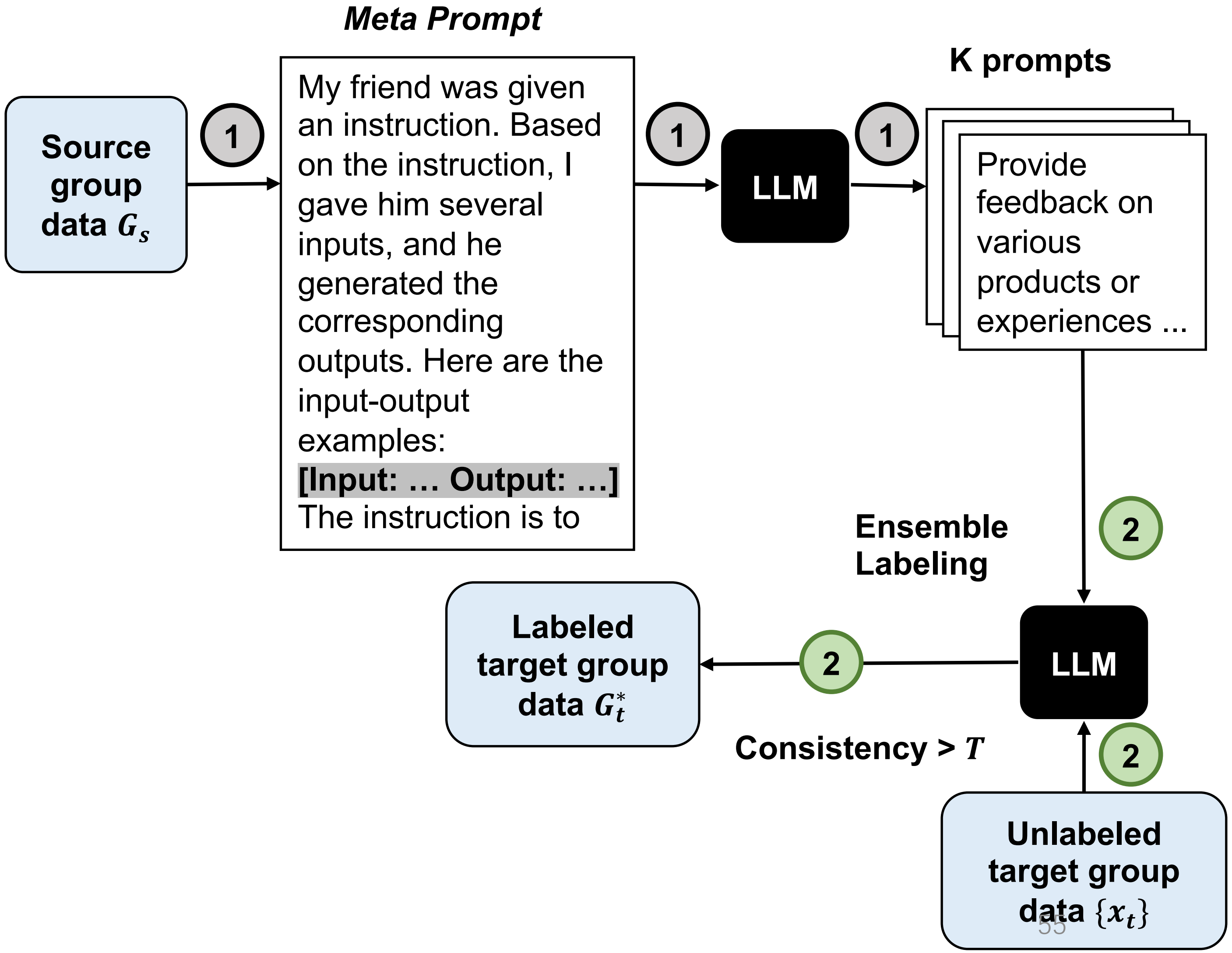
Step 2: Prompt Ensemble Labeling Strategy.
Labeling $\{x_t\}$ with ensembled prompts.



Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.

Step 2: Prompt Ensemble Labeling Strategy.
Labeling $\{x_t\}$ with ensembled prompts.

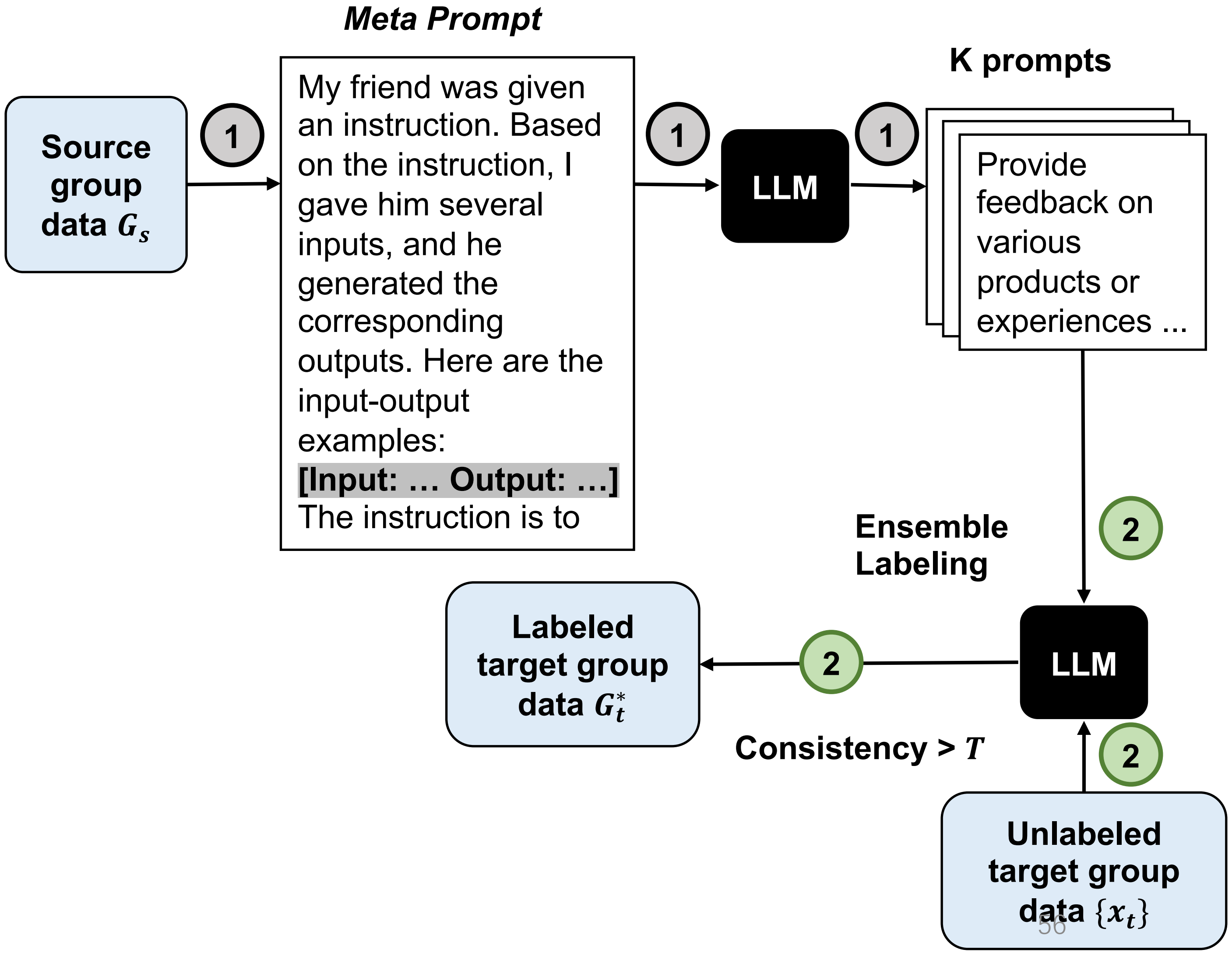


Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.

Step 2: Prompt Ensemble Labeling Strategy.
Labeling $\{x_t\}$ with ensembled prompts.

Step 3: Joint Prompt Optimization.
Mix G_s and G_t^* for prompt optimization.

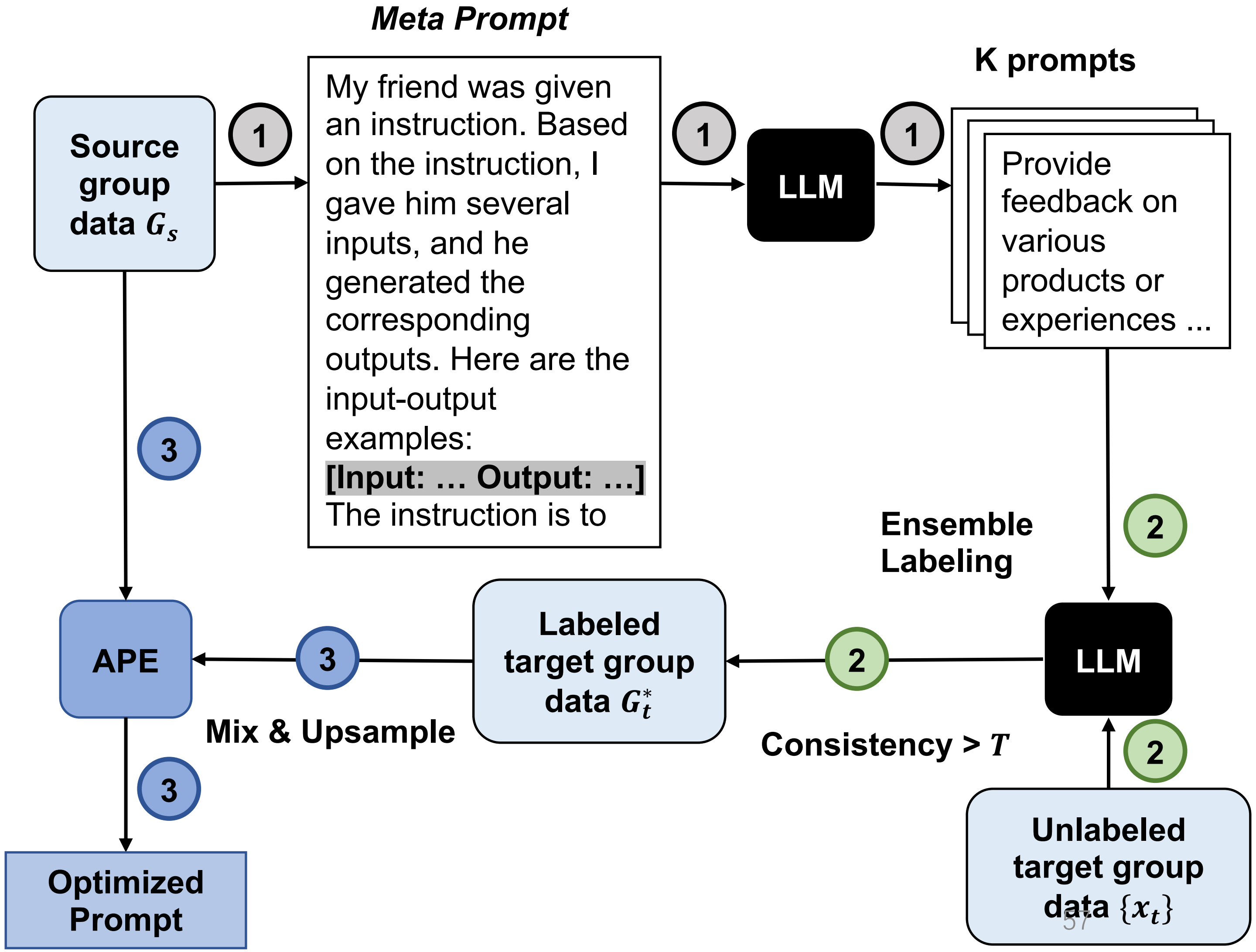


Generalized Prompt Optimization Framework

Step 1: Prompt Generation via Meta Prompt.
Deriving prompts from G_s.

Step 2: Prompt Ensemble Labeling Strategy.
Labeling $\{x_t\}$ with ensembled prompts.

Step 3: Joint Prompt Optimization.
Mix G_s and G_t^* for prompt optimization.



Experimental Setup

Experiment with 3 tasks, 6 groups with generalization gaps.

Experimental Setup

Experiment with 3 tasks, 6 groups with generalization gaps.

Compared methods:

- APE (Zhou et al. 2023)
- APO (Pryzant et al. 2023)
- APE + ut, a naïve approach to incorporate unlabeled target group data.
- Upper Bound: APE on the target group data with ground-truth labels.

Experimental Setup

Experiment with 3 tasks, 6 groups with generalization gaps.

Compared methods:

- APE (Zhou et al. 2023)
- APO (Pryzant et al. 2023)
- APE + ut, a naïve approach to incorporate unlabeled target group data.
- Upper Bound: APE on the target group data with ground-truth labels.

Testing Strategies:

- Top 1: using the single optimized prompt with top 1 validation performance.
- Ensemble: majority voting by K generated prompts.

Main Results

- GPO achieves superior performance on all target groups for both testing strategies
- But is still lower than Upper Bound.

	Yelp (Source)		Flipkart (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	**79.7 ± 0.7**	**79.7 ± 1.0**	78.4 ± 1.9	81.3 ± 1.4
APO	78.9 ± 0.5	**79.7 ± 0.8**	74.7 ± 3.0	76.4 ± 1.4
APE+ut	78.9 ± 1.4	78.8 ± 1.4	80.3 ± 2.0	80.7 ± 2.1
GPO	79.1 ± 0.7	78.7 ± 0.9	**80.5 ± 2.1**	**84.5 ± 2.0**
Upper Bound	-	-	85.1 ± 2.9	87.2 ± 0.5

(a) Sentiment analysis.

	SocialIQA (Source)		OpenbookQA (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	75.6 ± 1.4	69.6 ± 5.3	71.2 ± 5.2	74.8 ± 3.2
APO	76.1 ± 2.7	72.3 ± 2.6	72.4 ± 2.5	66.1 ± 7.2
APE+ut	**77.9 ± 1.3**	**78.9 ± 0.8**	77.5 ± 3.0	79.2 ±1.2
GPO	76.7 ± 2.0	**78.9 ± 1.2**	**78.7 ± 3.3**	**79.7 ± 0.8**
Upper Bound	-	-	80.1 ± 2.4	80.8 ± 1.1

(b) Commonsense QA.

	Number (Source)		Spans (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	51.9 ± 2.8	51.0 ± 3.2	20.1 ± 1.3	18.2 ± 0.2
APO	**55.7 ± 0.8**	**54.5 ± 2.1**	20.2 ± 2.4	20.0 ± 2.2
APE+ut	52.0 ± 1.8	53.1 ± 1.2	16.1 ± 3.5	17.7 ± 2.8
GPO	52.2 ± 6.0	53.6 ± 3.0	**27.7 ± 12.0**	**26.7 ± 4.9**
Upper Bound	-	-	63.1 ± 2.2	63.7 ± 0.8

(c) DROP.

Main Results

- GPO achieves superior performance on all target groups for both testing strategies.
- But is still lower than Upper Bound.

- GPO achieves comparable source group performance.
- Improvement on target group does not largely hinder source group.

	Yelp (Source)		Flipkart (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	**79.7 ± 0.7**	**79.7 ± 1.0**	78.4 ± 1.9	81.3 ± 1.4
APO	78.9 ± 0.5	**79.7 ± 0.8**	74.7 ± 3.0	76.4 ± 1.4
APE+ut	78.9 ± 1.4	78.8 ± 1.4	80.3 ± 2.0	80.7 ± 2.1
GPO	79.1 ± 0.7	78.7 ± 0.9	**80.5 ± 2.1**	**84.5 ± 2.0**
Upper Bound	-	-	85.1 ± 2.9	87.2 ± 0.5

(a) Sentiment analysis.

	SocialIQA (Source)		OpenbookQA (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	75.6 ± 1.4	69.6 ± 5.3	71.2 ± 5.2	74.8 ± 3.2
APO	76.1 ± 2.7	72.3 ± 2.6	72.4 ± 2.5	66.1 ± 7.2
APE+ut	**77.9 ± 1.3**	**78.9 ± 0.8**	77.5 ± 3.0	79.2 ±1.2
GPO	76.7 ± 2.0	**78.9 ± 1.2**	**78.7 ± 3.3**	**79.7 ± 0.8**
Upper Bound	-	-	80.1 ± 2.4	80.8 ± 1.1

(b) Commonsense QA.

	Number (Source)		Spans (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	51.9 ± 2.8	51.0 ± 3.2	20.1 ± 1.3	18.2 ± 0.2
APO	**55.7 ± 0.8**	**54.5 ± 2.1**	20.2 ± 2.4	20.0 ± 2.2
APE+ut	52.0 ± 1.8	53.1 ± 1.2	16.1 ± 3.5	17.7 ± 2.8
GPO	52.2 ± 6.0	53.6 ± 3.0	**27.7 ± 12.0**	**26.7 ± 4.9**
Upper Bound	-	-	63.1 ± 2.2	63.7 ± 0.8

(c) DROP.

Main Results

- GPO achieves superior performance on all target groups for both testing strategies.
- But is still lower than Upper Bound.

- GPO achieves comparable source group performance.
- Improvement on target group does not largely hinder source group.

- APE-ut: Incorporating unlabeled target input is beneficial for some tasks.
- But labeling is still important especially when labeling is challenging (Number, Spans).

	Yelp (Source)		Flipkart (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	**79.7** $\pm$ **0.7**	**79.7** $\pm$ **1.0**	78.4 $\pm$ 1.9	81.3 $\pm$ 1.4
APO	78.9 $\pm$ 0.5	**79.7** $\pm$ **0.8**	74.7 $\pm$ 3.0	76.4 $\pm$ 1.4
APE+ut	78.9 $\pm$ 1.4	78.8 $\pm$ 1.4	80.3 $\pm$ 2.0	80.7 $\pm$ 2.1
GPO	79.1 $\pm$ 0.7	78.7 $\pm$ 0.9	**80.5** $\pm$ **2.1**	**84.5** $\pm$ **2.0**
Upper Bound	-	-	85.1 $\pm$ 2.9	87.2 $\pm$ 0.5

(a) Sentiment analysis.

	SocialIQA (Source)		OpenbookQA (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	75.6 $\pm$ 1.4	69.6 $\pm$ 5.3	71.2 $\pm$ 5.2	74.8 $\pm$ 3.2
APO	76.1 $\pm$ 2.7	72.3 $\pm$ 2.6	72.4 $\pm$ 2.5	66.1 $\pm$ 7.2
APE+ut	**77.9** $\pm$ **1.3**	**78.9** $\pm$ **0.8**	77.5 $\pm$ 3.0	79.2 $\pm$1.2
GPO	76.7 $\pm$ 2.0	**78.9** $\pm$ **1.2**	**78.7** $\pm$ **3.3**	**79.7** $\pm$ **0.8**
Upper Bound	-	-	80.1 $\pm$ 2.4	80.8 $\pm$ 1.1

(b) Commonsense QA.

	Number (Source)		Spans (Target)	
	Top 1	Ensemble	Top 1	Ensemble
APE	51.9 $\pm$ 2.8	51.0 $\pm$ 3.2	20.1 $\pm$ 1.3	18.2 $\pm$ 0.2
APO	**55.7** $\pm$ **0.8**	**54.5** $\pm$ **2.1**	20.2 $\pm$ 2.4	20.0 $\pm$ 2.2
APE+ut	52.0 $\pm$ 1.8	53.1 $\pm$ 1.2	16.1 $\pm$ 3.5	17.7 $\pm$ 2.8
GPO	52.2 $\pm$ 6.0	53.6 $\pm$ 3.0	**27.7** $\pm$ **12.0**	**26.7** $\pm$ **4.9**
Upper Bound	-	-	63.1 $\pm$ 2.2	63.7 $\pm$ 0.8

(c) DROP.

Ablation Study

w/o cons
   setting the consistency threshold as 0
w/o cons + t-train
   removing the target group training data during the final prompt generation

	Yelp		Flipkart	
	Top 1	Ensemble	Top 1	Ensemble
GPO	79.1 ± 0.7	78.7 ± 0.9	80.5 ± 2.1	**84.5 ± 2.0**
w/o cons	78.8 ± 1.2	78.7 ± 0.4	**81.5 ± 1.4**	84.0 ± 0.9
w/o cons+t-train	**79.9 ± 0.8**	**79.7 ± 1.0**	80.3 ± 3.2	81.3 ± 1.4

(a) Sentiment analysis.

	SocialIQA		OpenbookQA	
	Top 1	Ensemble	Top 1	Ensemble
GPO	76.7 ± 2.0	**78.9 ± 1.2**	**78.7 ± 3.3**	**79.7 ± 0.8**
w/o cons	76.0 ± 2.8	78.1 ± 1.4	77.6 ± 3.8	78.8 ± 2.2
w/o cons+t-train	**77.9 ± 1.6**	69.6 ± 5.3	78.2 ± 2.2	74.8 ± 3.2

(b) Commonsense QA.

	Number		Spans	
	Top 1	Ensemble	Top 1	Ensemble
GPO	**52.2 ± 6.0**	**53.6 ± 3.0**	**27.7 ± 12.0**	**26.7 ± 4.9**
w/o cons	49.3 ± 2.8	51.0 ± 2.1	20.6 ± 2.1	22.2 ± 3.2
w/o cons+t-train	51.3 ± 3.6	50.9 ± 1.6	20.4 ± 1.9	18.7 ± 2.2

(c) DROP.

Ablation Study

w/o cons
- setting the consistency threshold as 0

w/o cons + t-train
- removing the target group training data during the final prompt generation

- In nearly all cases, GPO performs better than w/o cons on target groups.

	Yelp		Flipkart	
	Top 1	Ensemble	Top 1	Ensemble
GPO	79.1 ± 0.7	78.7 ± 0.9	80.5 ± 2.1	**84.5 ± 2.0**
w/o cons	78.8 ± 1.2	78.7 ± 0.4	**81.5 ± 1.4**	84.0 ± 0.9
w/o cons+t-train	**79.9 ± 0.8**	**79.7 ± 1.0**	80.3 ± 3.2	81.3 ± 1.4

(a) Sentiment analysis.

	SocialIQA		OpenbookQA	
	Top 1	Ensemble	Top 1	Ensemble
GPO	76.7 ± 2.0	**78.9 ± 1.2**	**78.7 ± 3.3**	**79.7 ± 0.8**
w/o cons	76.0 ± 2.8	78.1 ± 1.4	77.6 ± 3.8	78.8 ± 2.2
w/o cons+t-train	**77.9 ± 1.6**	69.6 ± 5.3	78.2 ± 2.2	74.8 ± 3.2

(b) Commonsense QA.

	Number		Spans	
	Top 1	Ensemble	Top 1	Ensemble
GPO	**52.2 ± 6.0**	**53.6 ± 3.0**	**27.7 ± 12.0**	**26.7 ± 4.9**
w/o cons	49.3 ± 2.8	51.0 ± 2.1	20.6 ± 2.1	22.2 ± 3.2
w/o cons+t-train	51.3 ± 3.6	50.9 ± 1.6	20.4 ± 1.9	18.7 ± 2.2

(c) DROP.

Ablation Study

w/o cons
    setting the consistency threshold as 0
w/o cons + t-train
    removing the target group training data during the final prompt generation

- In nearly all cases, GPO performs better than w/o cons on target groups.
- Removing t-train harms target group ensemble results, while has less effect on Top 1 results.

	Yelp		Flipkart	
	Top 1	Ensemble	Top 1	Ensemble
GPO	79.1 ± 0.7	78.7 ± 0.9	80.5 ± 2.1	**84.5 ± 2.0**
w/o cons	78.8 ± 1.2	78.7 ± 0.4	**81.5 ± 1.4**	84.0 ± 0.9
w/o cons+t-train	**79.9 ± 0.8**	**79.7 ± 1.0**	80.3 ± 3.2	81.3 ± 1.4

(a) Sentiment analysis.

	SocialIQA		OpenbookQA	
	Top 1	Ensemble	Top 1	Ensemble
GPO	76.7 ± 2.0	**78.9 ± 1.2**	**78.7 ± 3.3**	**79.7 ± 0.8**
w/o cons	76.0 ± 2.8	78.1 ± 1.4	77.6 ± 3.8	78.8 ± 2.2
w/o cons+t-train	**77.9 ± 1.6**	69.6 ± 5.3	78.2 ± 2.2	74.8 ± 3.2

(b) Commonsense QA.

	Number		Spans	
	Top 1	Ensemble	Top 1	Ensemble
GPO	**52.2 ± 6.0**	**53.6 ± 3.0**	**27.7 ± 12.0**	**26.7 ± 4.9**
w/o cons	49.3 ± 2.8	51.0 ± 2.1	20.6 ± 2.1	22.2 ± 3.2
w/o cons+t-train	51.3 ± 3.6	50.9 ± 1.6	20.4 ± 1.9	18.7 ± 2.2

(c) DROP.

Ablation Study

	Flipkart	OpenbookQA	Spans
w/o cons	81.9	69.8	3.6
GPO	94.2	84.3	3.7

Consistency Threshold improves labeling accuracy.

	Yelp		Flipkart	
	Top 1	Ensemble	Top 1	Ensemble
GPO	79.1 ± 0.7	78.7 ± 0.9	80.5 ± 2.1	$\mathbf{84.5 \pm 2.0}$
w/o cons	78.8 ± 1.2	78.7 ± 0.4	$\mathbf{81.5 \pm 1.4}$	84.0 ± 0.9
w/o cons+t-train	$\mathbf{79.9 \pm 0.8}$	$\mathbf{79.7 \pm 1.0}$	80.3 ± 3.2	81.3 ± 1.4

(a) Sentiment analysis.

	SocialIQA		OpenbookQA	
	Top 1	Ensemble	Top 1	Ensemble
GPO	76.7 ± 2.0	$\mathbf{78.9 \pm 1.2}$	$\mathbf{78.7 \pm 3.3}$	$\mathbf{79.7 \pm 0.8}$
w/o cons	76.0 ± 2.8	78.1 ± 1.4	77.6 ± 3.8	78.8 ± 2.2
w/o cons+t-train	$\mathbf{77.9 \pm 1.6}$	69.6 ± 5.3	78.2 ± 2.2	74.8 ± 3.2

(b) Commonsense QA.

	Number		Spans	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\mathbf{52.2 \pm 6.0}$	$\mathbf{53.6 \pm 3.0}$	$\mathbf{27.7 \pm 12.0}$	$\mathbf{26.7 \pm 4.9}$
w/o cons	49.3 ± 2.8	51.0 ± 2.1	20.6 ± 2.1	22.2 ± 3.2
w/o cons+t-train	51.3 ± 3.6	50.9 ± 1.6	20.4 ± 1.9	18.7 ± 2.2

(c) DROP.

Ablation Study

	Flipkart	OpenbookQA	Spans
w/o cons	81.9	69.8	3.6
GPO	94.2	84.3	3.7

Consistency Threshold improves labeling accuracy.

- With high labeling acc, cons is unlikely to largely improve generalization .

	Yelp		Flipkart	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\underline{79.1 \pm 0.7}$	$\underline{78.7 \pm 0.9}$	$\underline{80.5 \pm 2.1}$	$\mathbf{84.5 \pm 2.0}$
w/o cons	78.8 ± 1.2	$\underline{78.7 \pm 0.4}$	$\mathbf{81.5 \pm 1.4}$	$\underline{84.0 \pm 0.9}$
w/o cons+t-train	$\mathbf{79.9 \pm 0.8}$	$\mathbf{79.7 \pm 1.0}$	80.3 ± 3.2	81.3 ± 1.4

(a) Sentiment analysis.

	SocialIQA		OpenbookQA	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\underline{76.7 \pm 2.0}$	$\mathbf{78.9 \pm 1.2}$	$\mathbf{78.7 \pm 3.3}$	$\mathbf{79.7 \pm 0.8}$
w/o cons	76.0 ± 2.8	$\underline{78.1 \pm 1.4}$	77.6 ± 3.8	$\underline{78.8 \pm 2.2}$
w/o cons+t-train	$\mathbf{77.9 \pm 1.6}$	69.6 ± 5.3	$\underline{78.2 \pm 2.2}$	74.8 ± 3.2

(b) Commonsense QA.

	Number		Spans	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\mathbf{52.2 \pm 6.0}$	$\mathbf{53.6 \pm 3.0}$	$\mathbf{27.7 \pm 12.0}$	$\mathbf{26.7 \pm 4.9}$
w/o cons	49.3 ± 2.8	$\underline{51.0 \pm 2.1}$	$\underline{20.6 \pm 2.1}$	$\underline{22.2 \pm 3.2}$
w/o cons+t-train	$\underline{51.3 \pm 3.6}$	50.9 ± 1.6	20.4 ± 1.9	18.7 ± 2.2

(c) DROP.

Ablation Study

	Flipkart	OpenbookQA	Spans
w/o cons	81.9	69.8	3.6
GPO	94.2	84.3	3.7

Consistency Threshold improves labeling accuracy.

- With high labeling acc, cons is unlikely to largely improve generalization .
- With low labeling acc, a tiny improvement by cons can largely improve generalization.

	Yelp		Flipkart	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\underline{79.1 \pm 0.7}$	$\underline{78.7 \pm 0.9}$	$\underline{80.5 \pm 2.1}$	$\mathbf{84.5 \pm 2.0}$
w/o cons	78.8 ± 1.2	$\underline{78.7 \pm 0.4}$	$\mathbf{81.5 \pm 1.4}$	$\underline{84.0 \pm 0.9}$
w/o cons+t-train	$\mathbf{79.9 \pm 0.8}$	$\mathbf{79.7 \pm 1.0}$	80.3 ± 3.2	81.3 ± 1.4

(a) Sentiment analysis.

	SocialIQA		OpenbookQA	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\underline{76.7 \pm 2.0}$	$\mathbf{78.9 \pm 1.2}$	$\mathbf{78.7 \pm 3.3}$	$\mathbf{79.7 \pm 0.8}$
w/o cons	76.0 ± 2.8	$\underline{78.1 \pm 1.4}$	77.6 ± 3.8	$\underline{78.8 \pm 2.2}$
w/o cons+t-train	$\mathbf{77.9 \pm 1.6}$	69.6 ± 5.3	$\underline{78.2 \pm 2.2}$	74.8 ± 3.2

(b) Commonsense QA.

	Number		Spans	
	Top 1	Ensemble	Top 1	Ensemble
GPO	$\mathbf{52.2 \pm 6.0}$	$\mathbf{53.6 \pm 3.0}$	$\mathbf{27.7 \pm 12.0}$	$\mathbf{26.7 \pm 4.9}$
w/o cons	49.3 ± 2.8	$\underline{51.0 \pm 2.1}$	$\underline{20.6 \pm 2.1}$	$\underline{22.2 \pm 3.2}$
w/o cons+t-train	$\underline{51.3 \pm 3.6}$	50.9 ± 1.6	20.4 ± 1.9	18.7 ± 2.2

(c) DROP.

Different Backbone LLMs

	Top 1		Ensemble	
	APE	GPO	APE	GPO
Vicuna-7B	38.4 ± 25.3	63.5 ± 15.6	43.9 ± 21.3	71.9 ± 13.1
Vicuna-13B	66.8 ± 18.4	68.3 ± 13.7	60.7 ± 9.5	70.7 ± 10.8
GPT-3.5	**78.4 ± 1.9**	80.5 ± 2.1	81.3 ± 1.4	84.5 ± 2.0
GPT-4	77.5 ± 13.7	**85.3 ± 2.7**	**83.3 ± 0.0**	**85.4 ± 2.4**

Generalization performance on Flipkart.

- Spaces for improving generalization across different LLMs.
- GPO achieves improvement in all cases.
- GPT-4 achieves the best performance on GPO.

Case Study

Prompts contains group-specific information.

+GPO

More general prompts.

Yelp	Provide feedback on various experiences, such as **dining, shopping, and service**. The output format is a sentiment analysis, where the input is analyzed to determine whether the experience was positive, negative, or neutral. The output is a single word indicating the sentiment of the experience.
Flipkart	Provide a sentiment analysis of **customer reviews**. The input consists of a customer review of a product, and the output is a binary classification of the sentiment as either positive or negative.
GPO	provide a sentiment analysis of **a given text**. The output format is a single word indicating whether the sentiment is positive, negative, or neutral.

Number	Answer a specific question based on a given context. The output format is **a numerical value** that directly answers the question asked.
Spans	Answer a specific question based on a given context. The output format is **a single word or phrase** that directly answers the question asked.
GPO	Answer questions based on given context information. The output format is **a numerical value or a single word answer.**

Contribution

- Revealed the **robustness issue of prompt optimization against distribution shifts** and propose a new **robust prompt optimization problem**.
- Proposed the **Generalized Prompt Optimization framework**.
- Conducted extensive experiments on three NLP tasks, validating the rationality and effectiveness of our proposed framework.

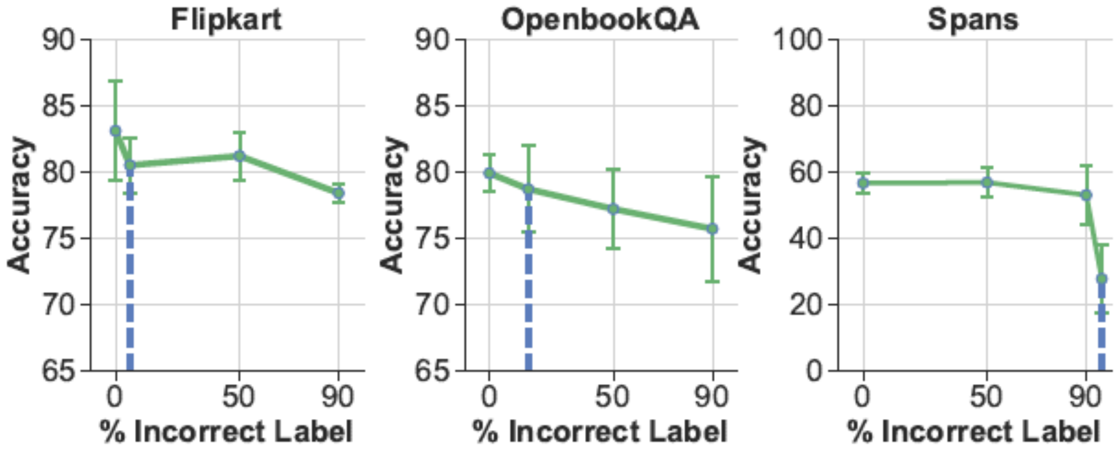
Thank You for Listening!

Code: https://github.com/li-moxin/GPO

Arxiv: https://arxiv.org/abs/2305.13954

Contact: limoxin@u.nus.edu

Analysis on the Consistency Threshold



Generalization performance under different percentage of wrongly labeled target group data.

Higher labeling accuracy -> Better generalization performance.

Compare to Human-Written Prompts

	Yelp (Source)	Flipkart (Target)	SocialIQA (Source)	OpenbookQA (Target)	Number (Source)	Spans (Target)
Human	**78.7**	80.0	71.3	60.0	**54.9**	**37.1**
PromptPerfect	77.3	83.3	74.7	64.0	54.0	26.9
GPO best	**78.7**	**84.5**	**78.9**	**79.7**	52.2	27.7

Human: a prompt written by computer science college student.
PromptPerfect: https://promptperfect.jina.ai.
GPO best: best testing strategy of GPO.

- GPO achieves best performance on the left two tasks.
- But worse performance on DROP due to inaccurate labels.