EMNLP 2023

Making Large Language Models Better **Data Creators**

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen W. White, Sujay Kumar Jauhar

Microsoft

USC INFORMATION SCIENCES INSTITUTE

LLM-based Data Creation

Create a **train data** for training a model aimed at …

LLM

LLM-based Data Creation

Create a **train data** for training a model aimed at …

Generating coherent response for instruction-following

Self-Instruct (Wang et al, 2023), Alpaca, Vicuna …

LLM

LLM-based Data Creation

Create a **train data** for training a model aimed at …

Generating coherent response for instruction-following

Generating action sequence for agents

LLM

Self-Instruct (Wang et al, 2023), Alpaca, Vicuna …

Lumos (Yin et al., 2023)

LLM-based Data Creation

Create a **train data** for training a model aimed at …

Generating coherent response for instruction-following

Generating action sequence for agents

Generating correct answer for the target task

LLM

Self-Instruct (Wang et al, 2023), Alpaca, Vicuna …

Lumos (Yin et al., 2023)

SuperGen (Meng et al., 2022), ZeroGen (Ye et al., 2022) …,

Our focus

Create a **train data** for training a model aimed at …

Generating coherent response for instruction-following

Generating action sequence for agents

Generating correct answer for the target task

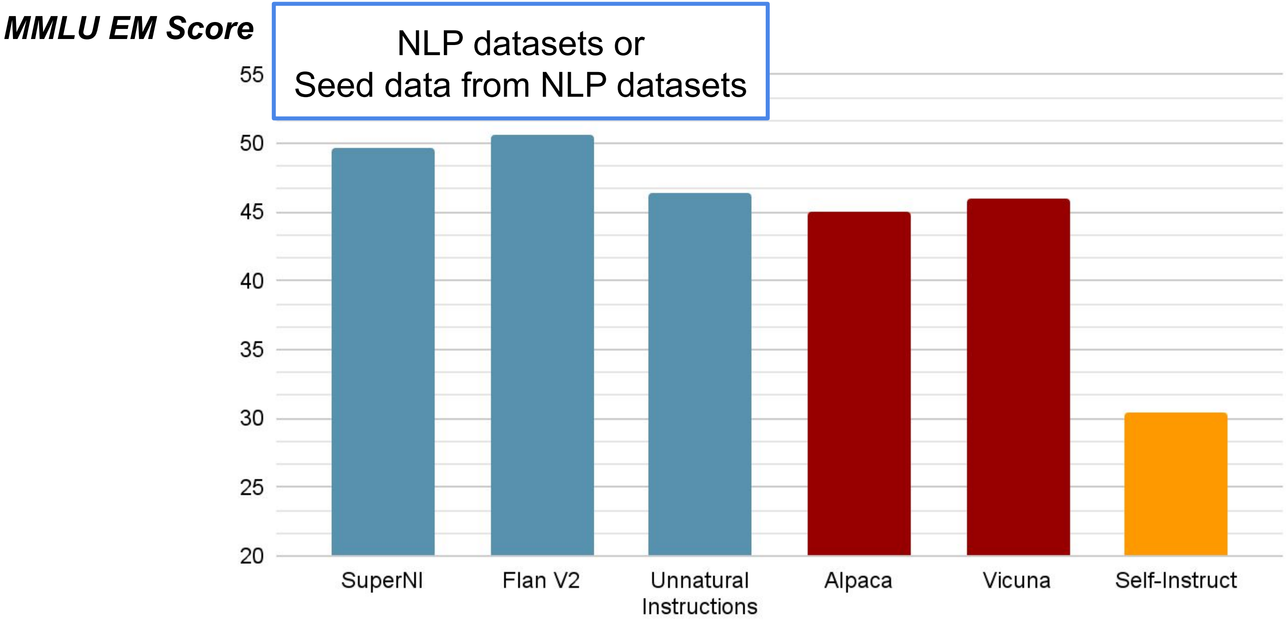
LLM

Self-Instruct (Wang et al, 2023), Alpaca, Vicuna …

Lumos (Yin et al., 2023)

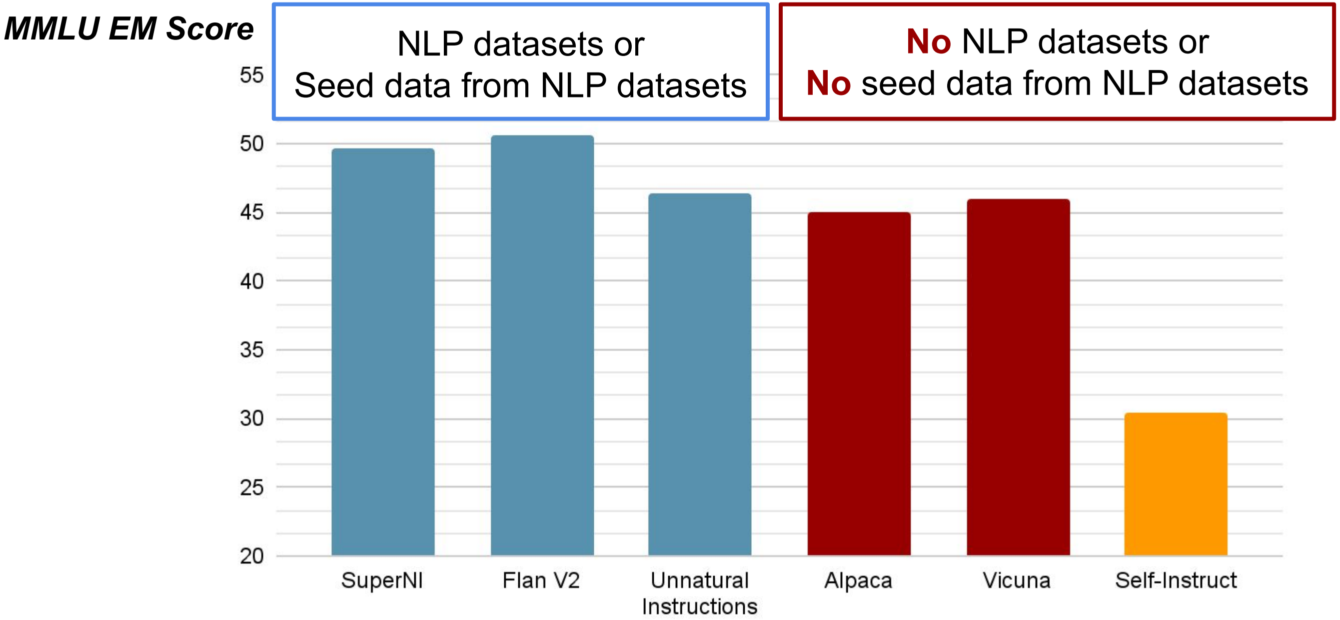
SuperGen (Meng et al., 2022), ZeroGen (Ye et al., 2022) …,

Importance of creating train data for generating correct answer for the target task



How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources, Wang et al., 2023

Importance of creating train data for generating correct answer for the target task

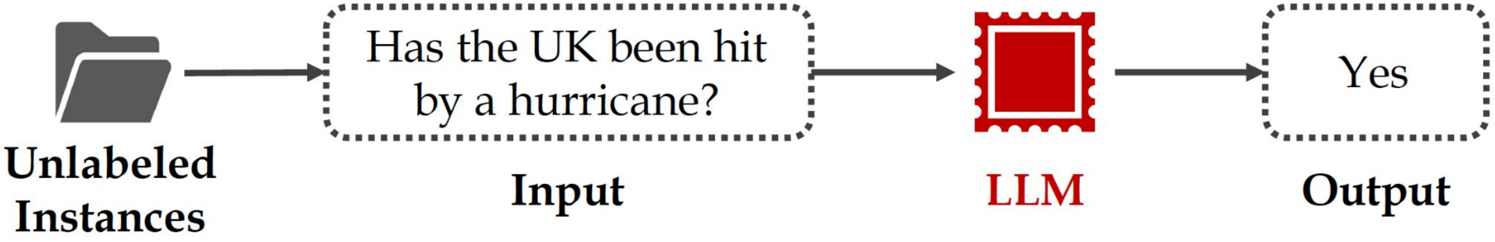


Task correctness != Coherent response

How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources, Wang et al., 2023

Existing Works

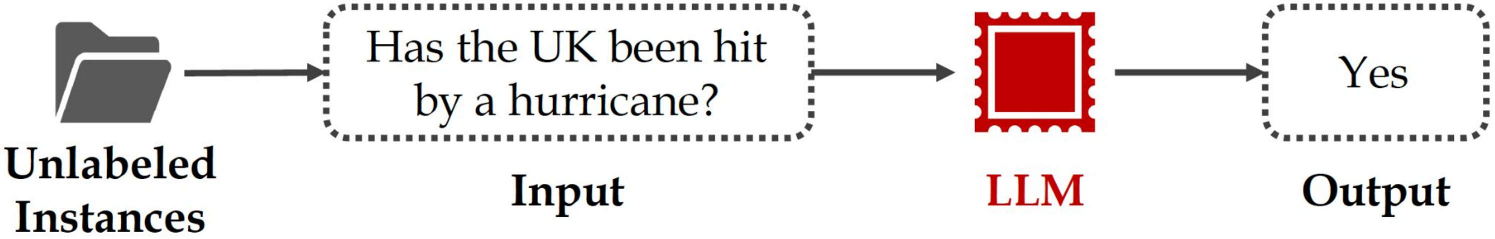
LLM as **Labelers**

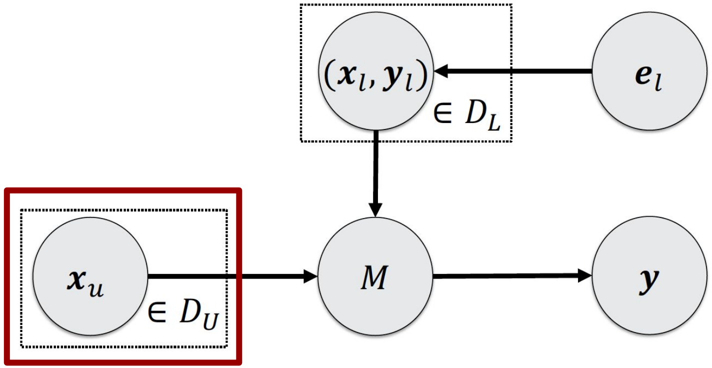


GPT3Mix: Leveraging Large-Scale Language Models for Text Augmentation., Yoo et al., 2021
Want to Reduce Labeling Cost? GPT-3 can help., Wang et al., 2021
Co-training Improves Prompt-based Learning for Large Language Models., Lang et al., 2022

Existing Works

LLM as **Labelers**





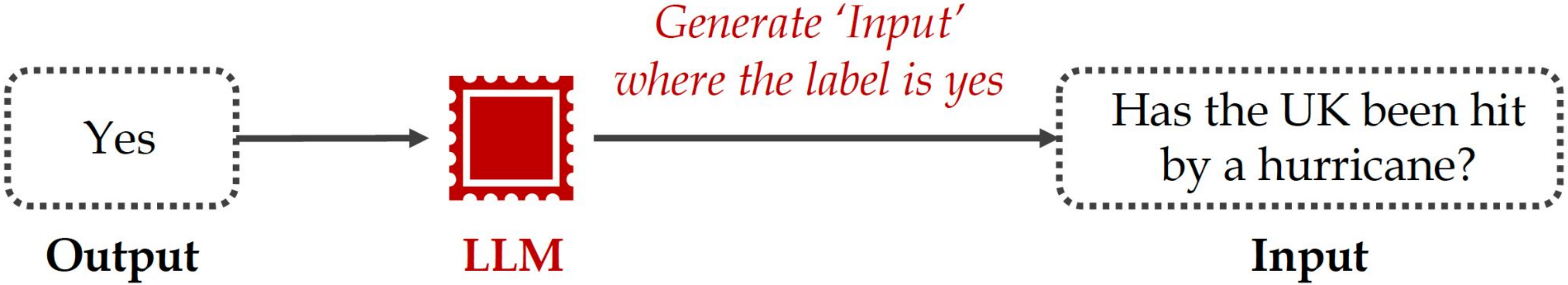
Need unlabeled examples D_U

Curating diverse and representative examples to label is challenging.

GPT3Mix: Leveraging Large-Scale Language Models for Text Augmentation., Yoo et al., 2021
Want to Reduce Labeling Cost? GPT-3 can help., Wang et al., 2021
Co-training Improves Prompt-based Learning for Large Language Models., Lang et al., 2022

Existing Works

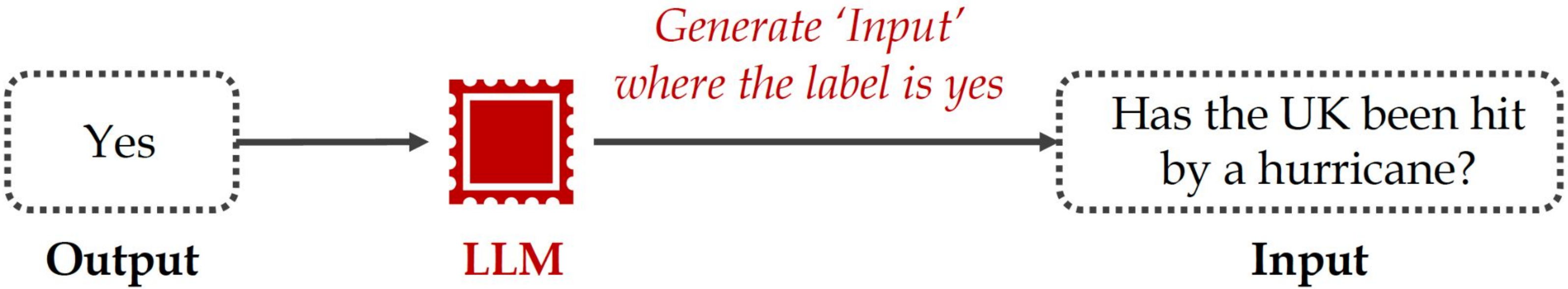
LLM as **Generators**

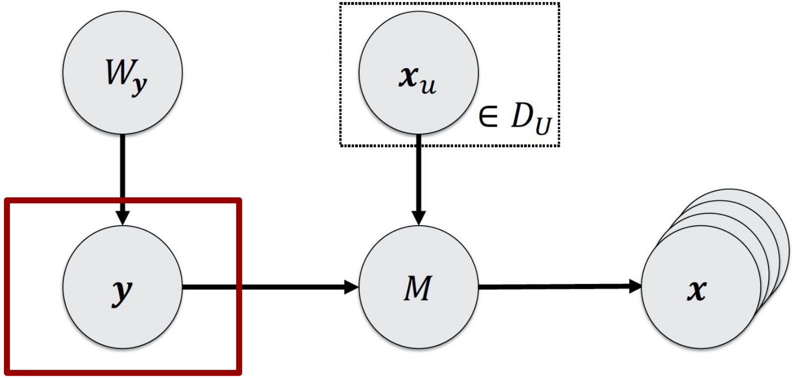


Generating Training Data with Language Models: Towards Zero-shot Language Understanding., Meng et al., 2022
ZeroGen: Efficient Zero-shot Learning via Dataset Generation., Ye et al., 2022
Self-guided Noise-Free Data Generation for Efficient Zero-shot Learning., Gao et al., 2022

Existing Works

LLM as **Generators**





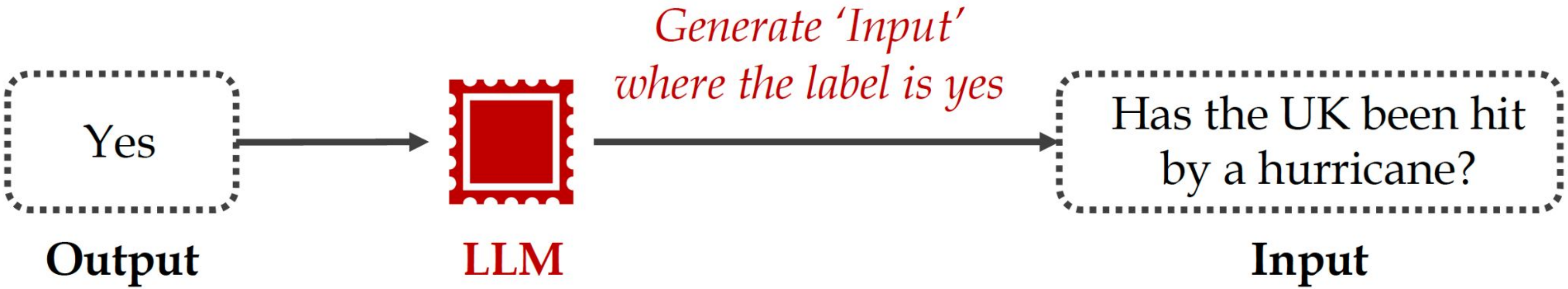
Semantic meaning of y

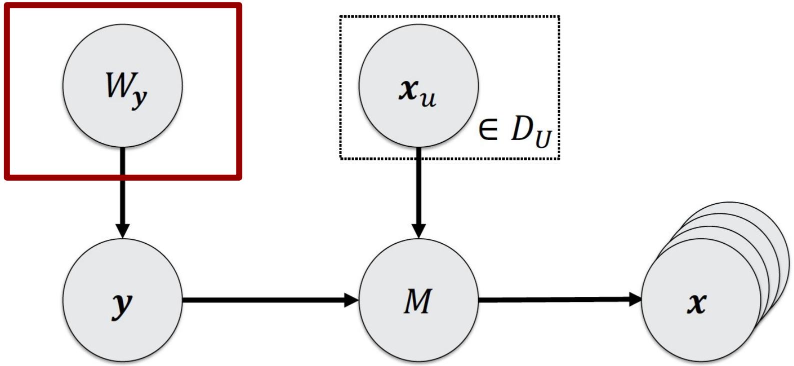
If y is an index or binary response, it is difficult to generate examples

Generating Training Data with Language Models: Towards Zero-shot Language Understanding., Meng et al., 2022
ZeroGen: Efficient Zero-shot Learning via Dataset Generation., Ye et al., 2022
Self-guided Noise-Free Data Generation for Efficient Zero-shot Learning., Gao et al., 2022

Existing Works

LLM as **Generators**



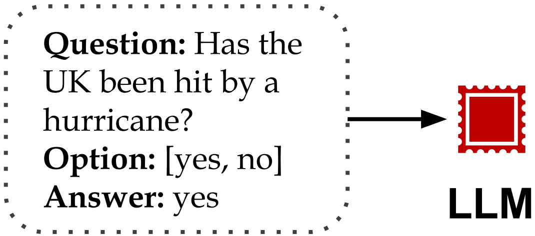


Need specific prompt W_y

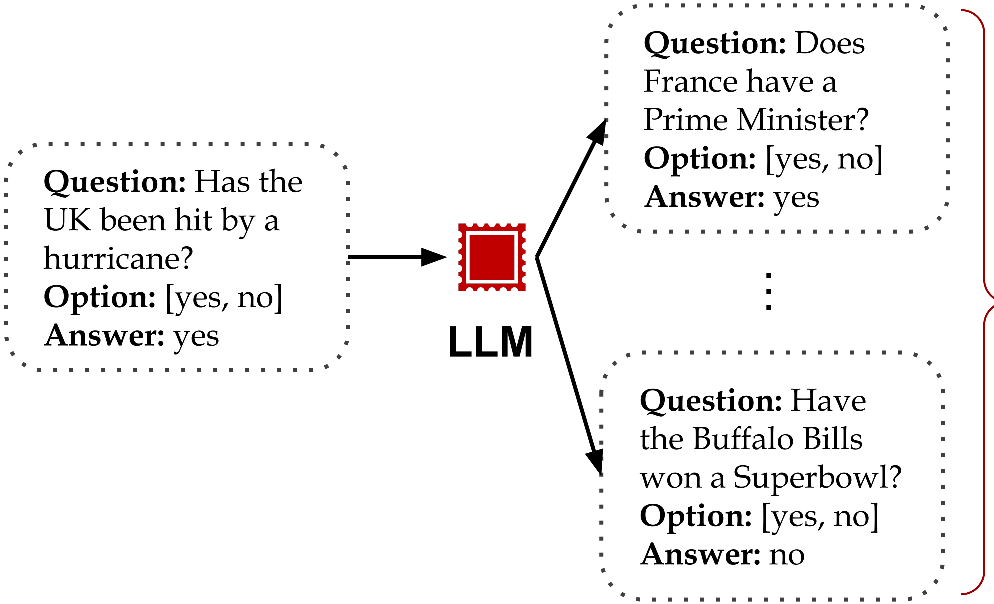
If the label is "entailment" for NLI, prompt should include "s1, in other words"

Generating Training Data with Language Models: Towards Zero-shot Language Understanding., Meng et al., 2022
ZeroGen: Efficient Zero-shot Learning via Dataset Generation., Ye et al., 2022
Self-guided Noise-Free Data Generation for Efficient Zero-shot Learning., Gao et al., 2022

Example-based Data Creation



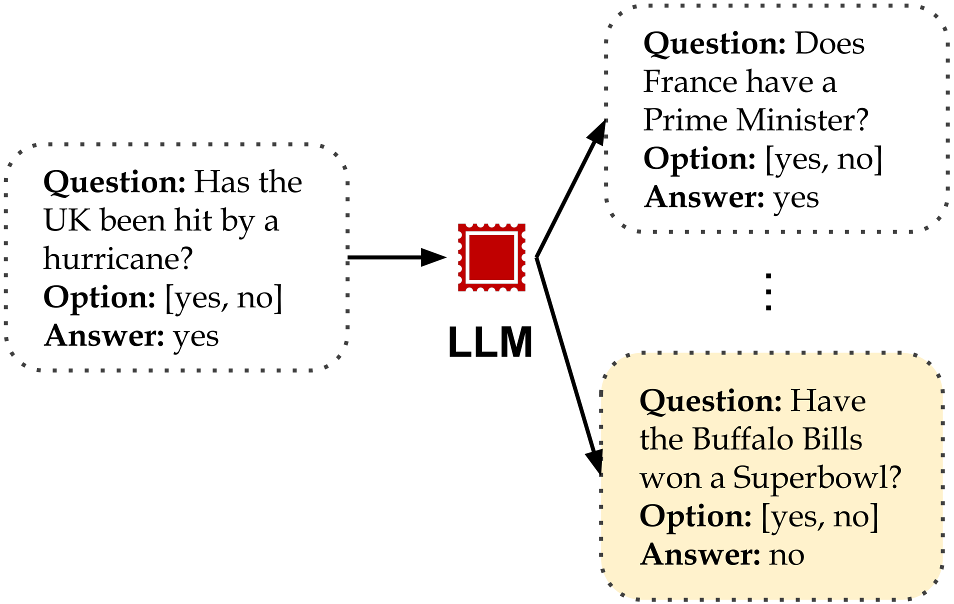
Example-based Data Creation



Instruction
`- You are creating {number_of_examples} examples that follow the format of the example provided, but with a different content.`
`- The created examples **must** all have different answers.`
`- The output **must** be in unnumbered JSON format.`

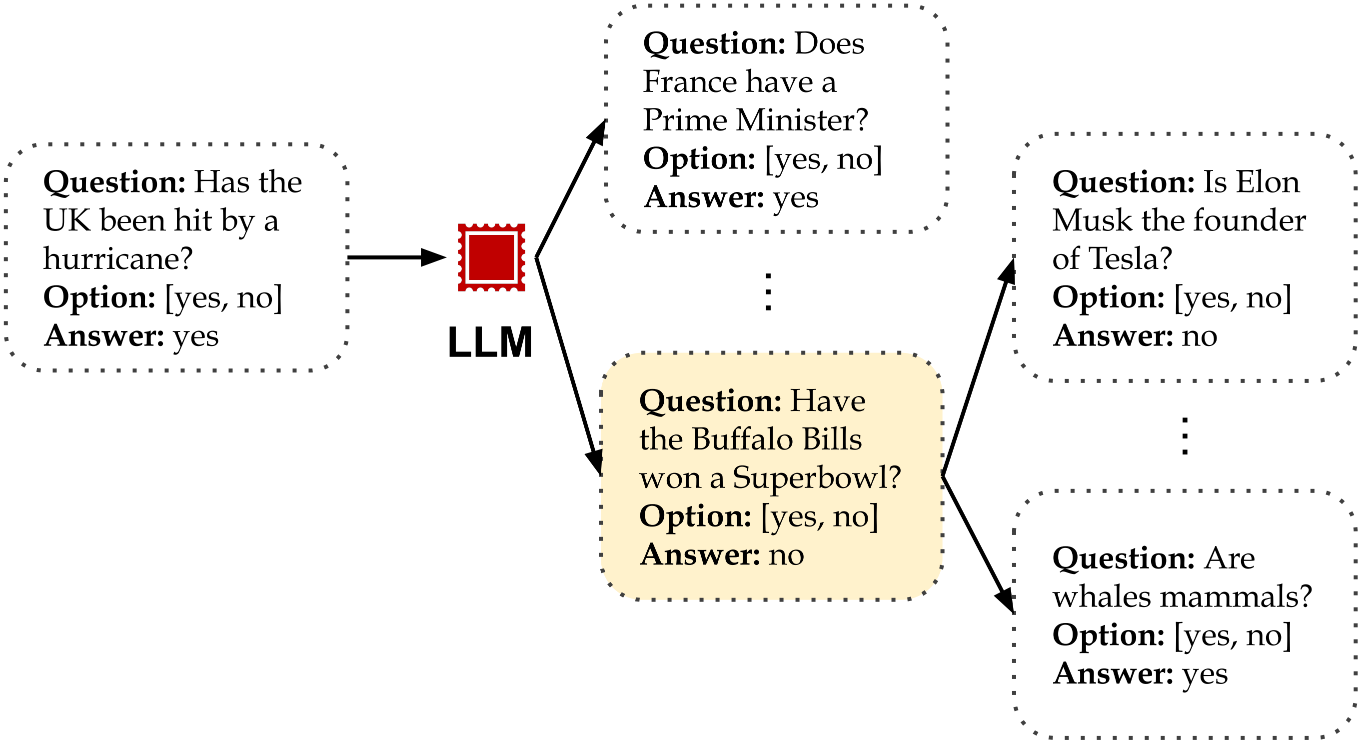
- *N* different examples
- Should provide **as many different answers as** possible

Example-based Data Creation

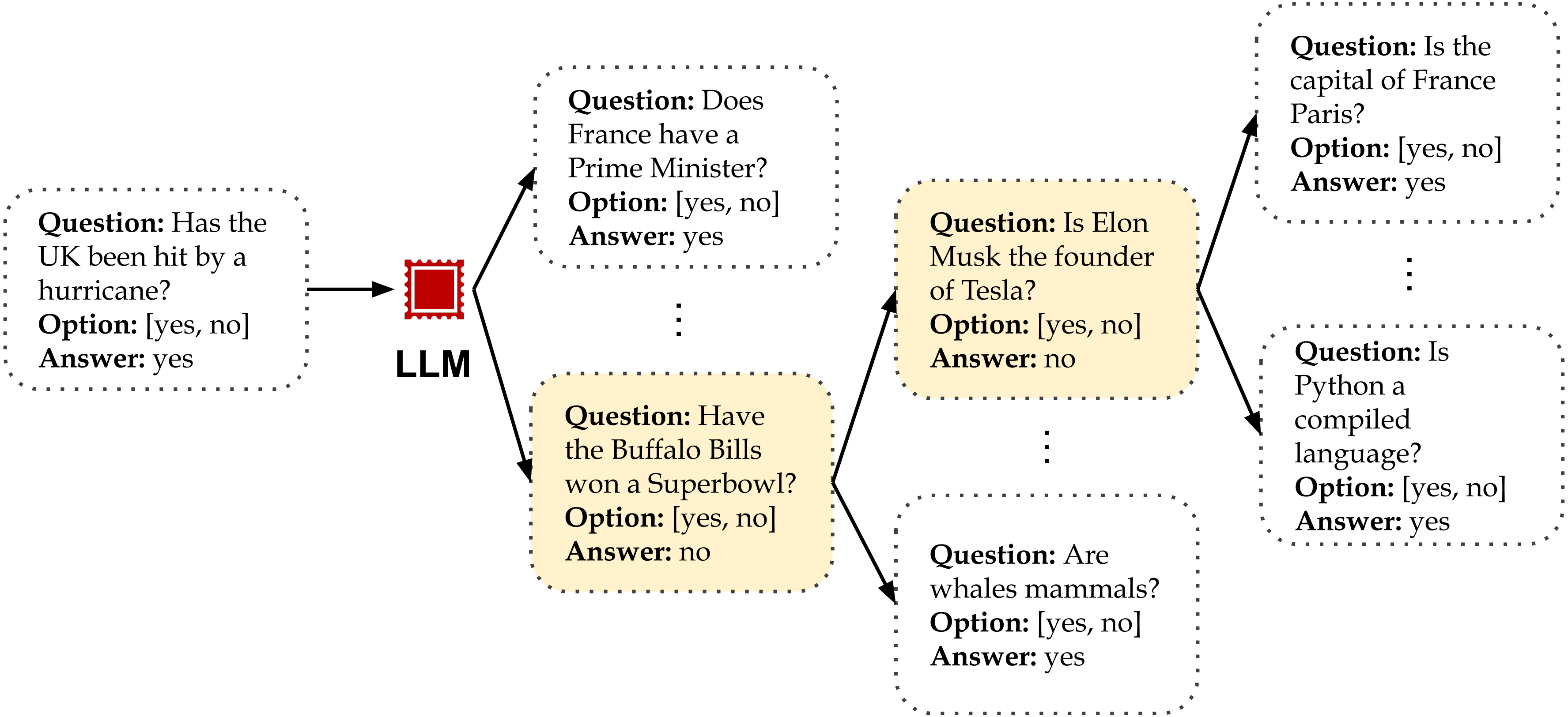


Select **seed data** for the next generation step.

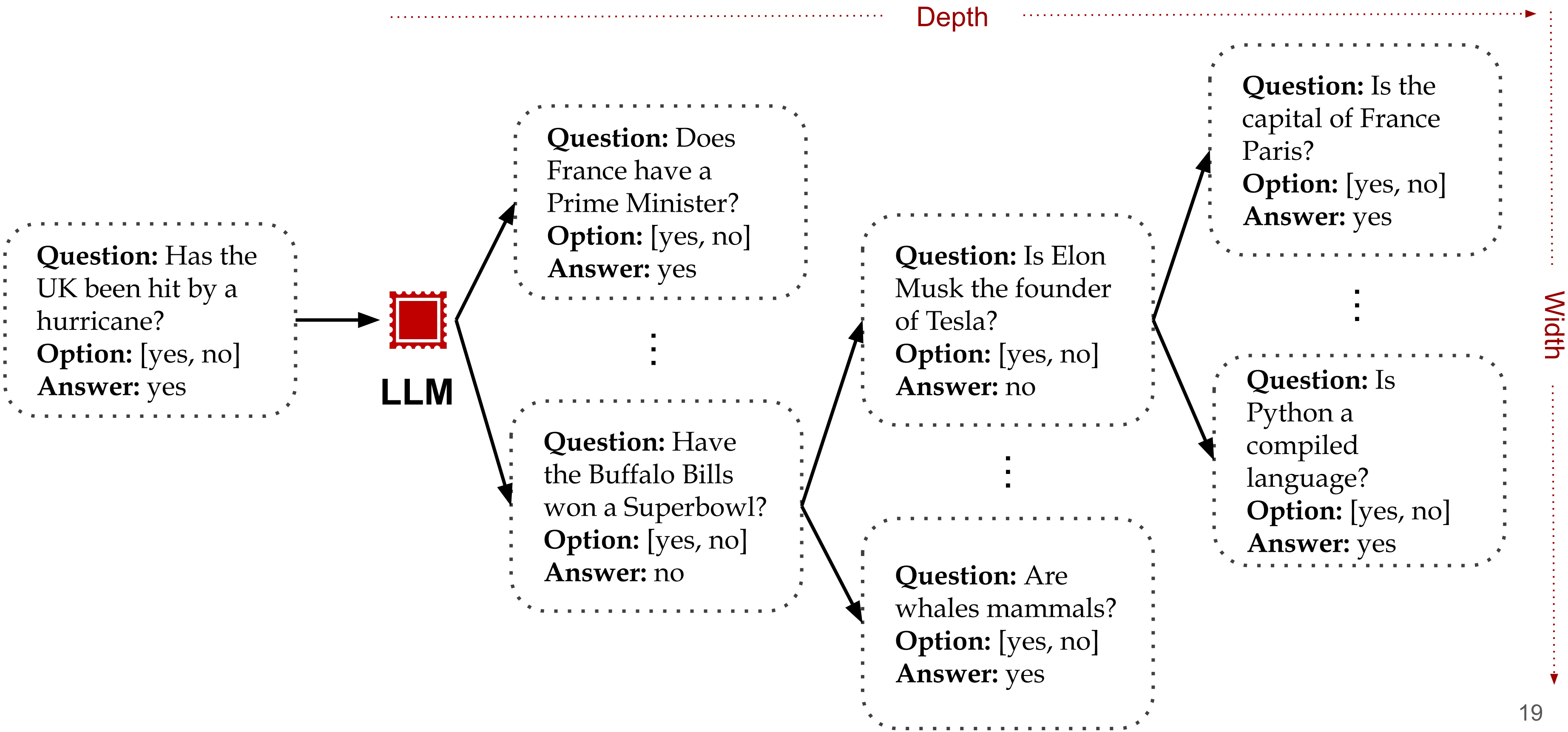
Example-based Data Creation



Example-based Data Creation



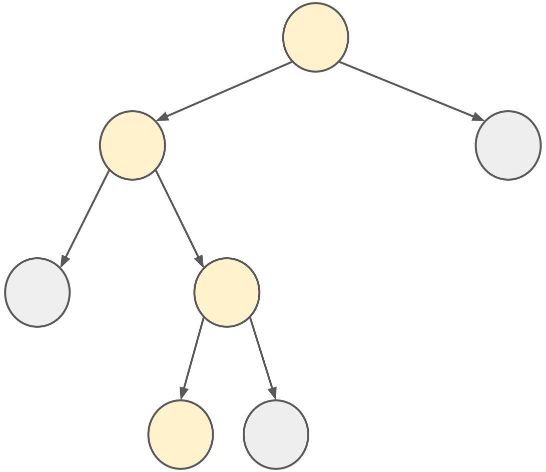
Example-based Data Creation



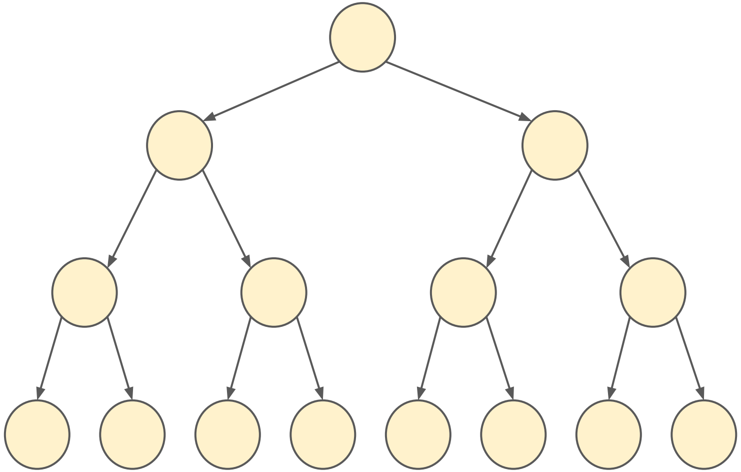
Selection Strategy

DFS vs. BFS

DFS



BFS

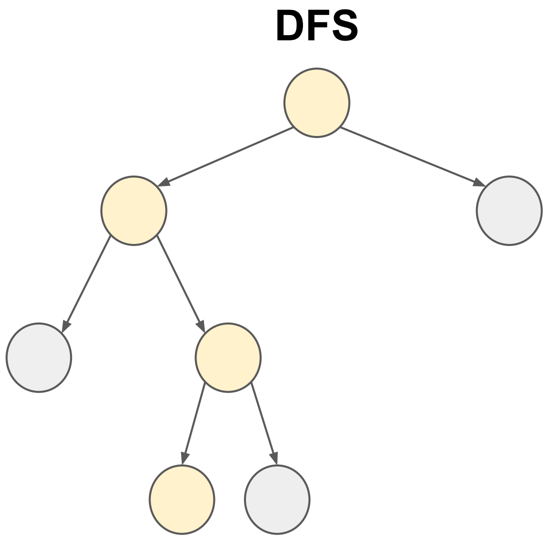


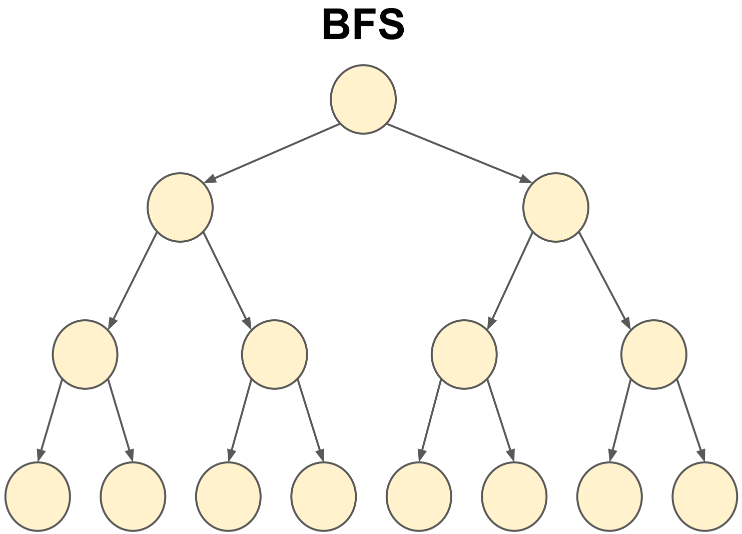
Random selection: randomly select the seed data
Similar / Contrastive selection: select the seed data by computing sentence similarity between the pervious seed and the generated examples.

Tree selection: All the generated examples become the seed data for the next generation

Selection Strategy

DFS vs. BFS





Large depth / Small width: Huge distance between the <u>initial formatting example</u> and generated examples
Large diversity, but small task consistency

Small depth / Large width: Small distance between the <u>initial formatting example</u> and generated examples
Large task consistency, but small diversity

Experimental Setup

Category	**Data** Name	**# Train data**
Multiple-choice QA (# Choices: 2)	PIQA	14,113
	WinoGrande	160
Multiple-choice QA (# Choices: 5)	CommonsenseQA	8,500
	RiddleSense	3,510
Open-book Yes/No QA	BoolQ	9,427
	PubMedQA	450
	BioASQ	670
Closed-book Yes/No QA	BoolQ	9,427
	StrategyQA	2,061
	CREAK	10,176

For each **data**,

1. Select one **formatting example** from the **train data**.

Experimental Setup

Category	Data Name	# Train data
Multiple-choice QA (# Choices: 2)	PIQA	14,113
	WinoGrande	160
Multiple-choice QA (# Choices: 5)	CommonsenseQA	8,500
	RiddleSense	3,510
Open-book Yes/No QA	BoolQ	9,427
	PubMedQA	450
	BioASQ	670
Closed-book Yes/No QA	BoolQ	9,427
	StrategyQA	2,061
	CREAK	10,176

For each **data**,

1. Select one **formatting example** from the **train data**.
2. generate data until the number of generated data reaches the same **number of train data**.
   - Discard ill-formatted data automatically with *json.loads()*

Experimental Setup

Category	Data Name	# Train data
Multiple-choice QA (# Choices: 2)	PIQA	14,113
	WinoGrande	160
Multiple-choice QA (# Choices: 5)	CommonsenseQA	8,500
	RiddleSense	3,510
Open-book Yes/No QA	BoolQ	9,427
	PubMedQA	450
	BioASQ	670
Closed-book Yes/No QA	BoolQ	9,427
	StrategyQA	2,061
	CREAK	10,176

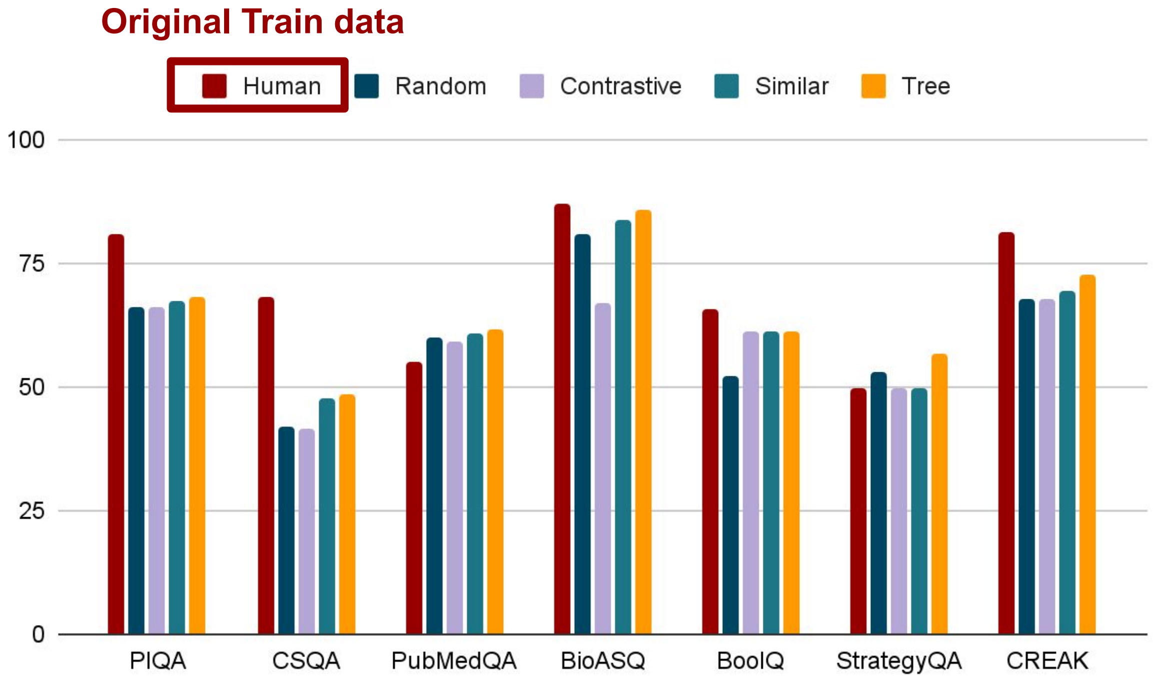
For each **data**,

1. Select one **formatting example** from the **train data**.
2. generate data until the number of generated data reaches the same **number of train data**.
   - Discard ill-formatted data automatically with *json.loads()*

Train & Test with **RoBERTa-large**.

Experimental Results

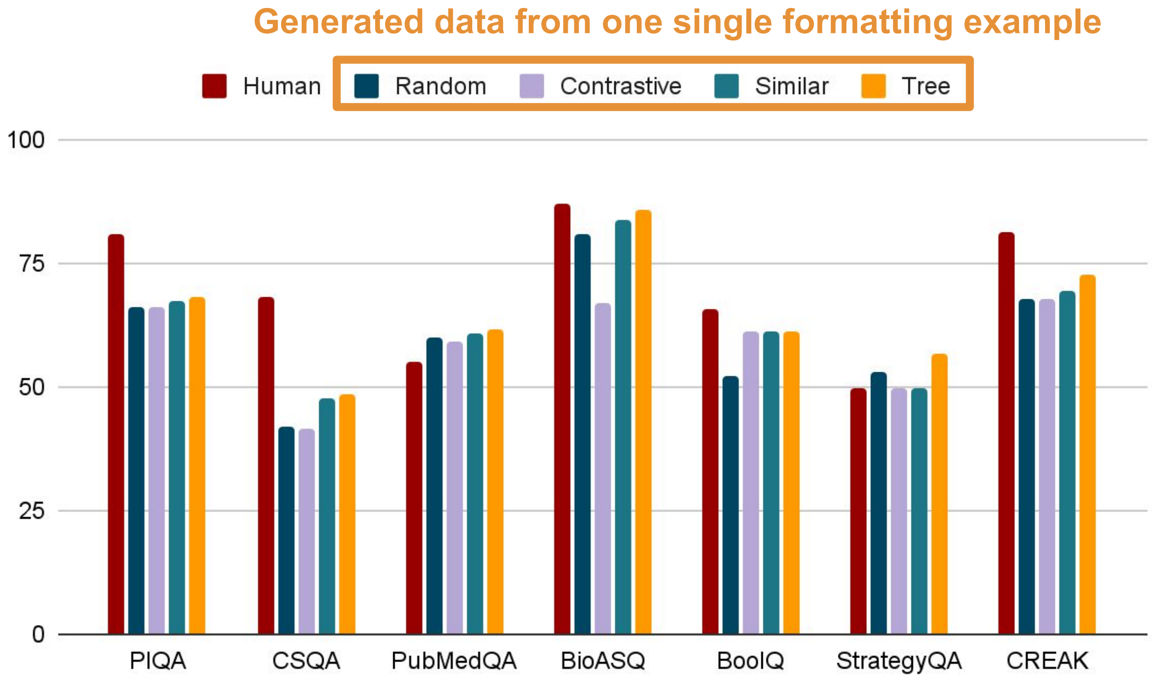
In-distribution Performance



Human-labeled train data always shows the best performance.

Experimental Results

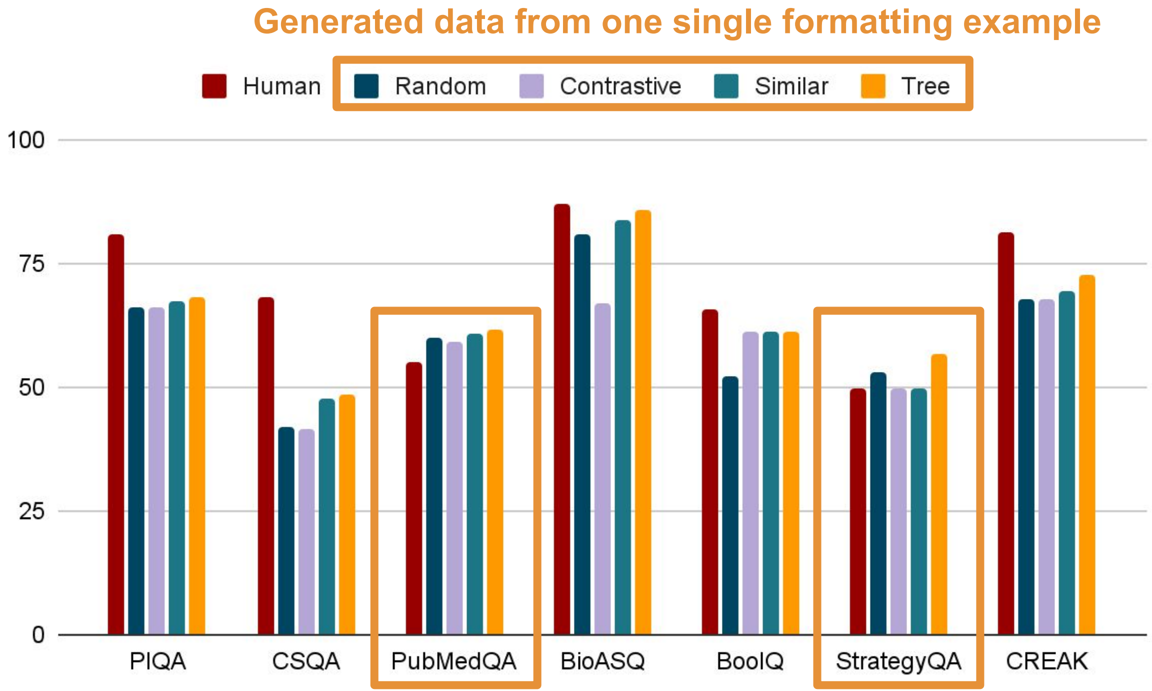
In-distribution Performance



Tree (BFS) shows the best performance among the data generation selection strategies.

Experimental Results

In-distribution Performance



Sometimes **Tree (BFS)** shows better than **Human-labels**

Experimental Setup

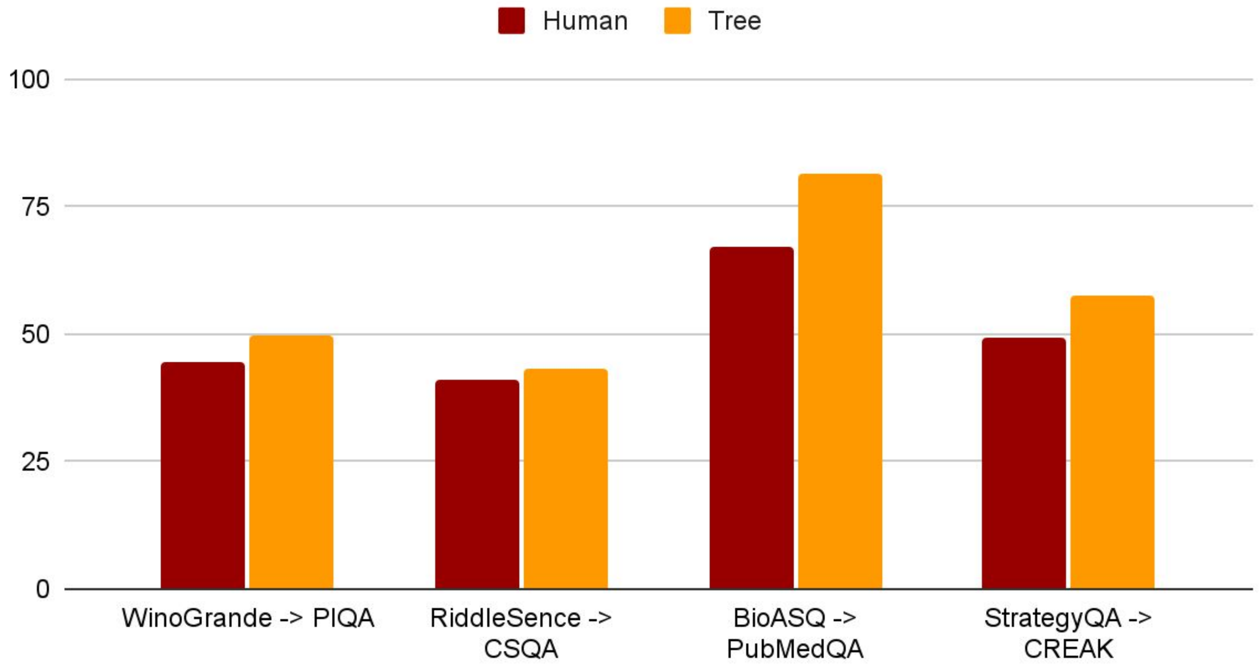
Category	Data Name	# Train data
Multiple-choice QA (# Choices: 2)	PIQA	14,113
	WinoGrande	160
Multiple-choice QA (# Choices: 5)	CommonsenseQA	8,500
	RiddleSense	3,510
Open-book Yes/No QA	BoolQ	9,427
	PubMedQA	450
	BioASQ	670
Closed-book Yes/No QA	BoolQ	9,427
	StrategyQA	2,061
	CREAK	10,176

How about the **out-of-distribution** performance?

1. Generate data starting from the seed in **CommonsenseQA**
2. Train the model
3. Test on **RiddleSense**

Experimental Results

Out-of-distribution Performance



LLM-created data has more generalizability than **human-labeled data**

Conclusion

1. Example-based data creation has a **flexibility** to create a task-specific data.

2. BFS-style data creation is better than DFS-style data creation.
   a. **low semantic distance** between the seed and created instances is important.

3. Models trained on LLM-created data are showing **strong generalizability**.
   a. Real-world systems often deal with inputs that are very different from carefully curated academic datasets

Q&A





Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen W. White, Sujay Kumar Jauhar

https://github.com/microsoft/llm-data-creation