2023 Summer Seminar

이승윤

1. **Generate rather than Retrieve:**
   Large Language Models are Strong Context Generators

2. **Guess The Instruction!**
   Flipped Learning Makes Language Models Strong Zero-Shot Learners

3. **Leveraging Large Language Models For**
   Multiple Choice Question Answering

Generate rather than Retrieve: Large Language Models are Strong Context Generators

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju,
Soumya Sanyal, Chenguang Zhu, Michael Zeng, Meng Jiang

ICLR2023

Objective

Limitations In retriever-based Open-domain QA system

1. Noisy Information can be contained

: The retrieved documents might contain noisy information that is irrelevant to the question

2. Shallow interactions captured between question & documents

: Representations of questions and documents are obtained independently in modern two-tower dens e retrieval models

3. Heavy resources and computation

: Document retrieval over a large corpus requires the retriever model to first encode all candidate do cuments and store representations for each document

Method

Generate-then-Read

Zero-Shot setting

⇒ First prompts a large language model to generate contextual documents based on a given question, and then reads the generated documents to produce the final answer.

< Open-Domain QA >	< Fact checking >	< Open-domain Dialogue System >
“Generate a background document from Wikipedia to answer the given question. \ n\ n {query} \ n\ n”	“Generate a background document from Wikipedia to support or refute the statement. \ n\ n Statement: {claim} \ n\ n”	“Generate a background document from Wikipedia to answer the given question. \ n\ n {utterance} \ n\ n”
Refer to the passage below and answer the following question with just a few words. Passage: {background} Question: {query} The answer is	{background} claim: {claim} Is the claim true or false?	{background} utterance

Method

Generate-then-Read

Supervised Setting

- Leverage a small reader model such as FiD

No.	Prompts	Validation
#1	Generate a background document from Wikipedia to answer the given question.	66.0
#2	Provide a background document from Wikipedia to answer the given question.	65.0
#3	Generate a background document from web to answer the given question.	64.0
#4	Generate a Wikipedia document to support the given question.	63.5
#5	Provide a background document for the given question.	63.0
#6	Prepare a background document to support the given question.	63.0
#7	To support the given question, prepare a background document.	62.5
#8	Create a background document that supports the given question.	61.5
#9	Retrieve a document from Wikipedia to answer the given question.	60.5
#10	Retrieve a Wikipedia article to address the posed question.	59.5

Table 21: Top-10 human prompts, evaluated on merged validation set of NQ, TriviaQA and WebQ.

1. **Diverse Human Prompts**

- Ask human annotators to provide different prompts , to make the generated document diverse

⇒ Requires human annotators

⇒ Different large language models, different pr ompt words

Method

Generate-then-Read

Supervised Setting

2. Clustering-Based Prompts

GET ONE INITIAL DOCUMENT PER QUESTION

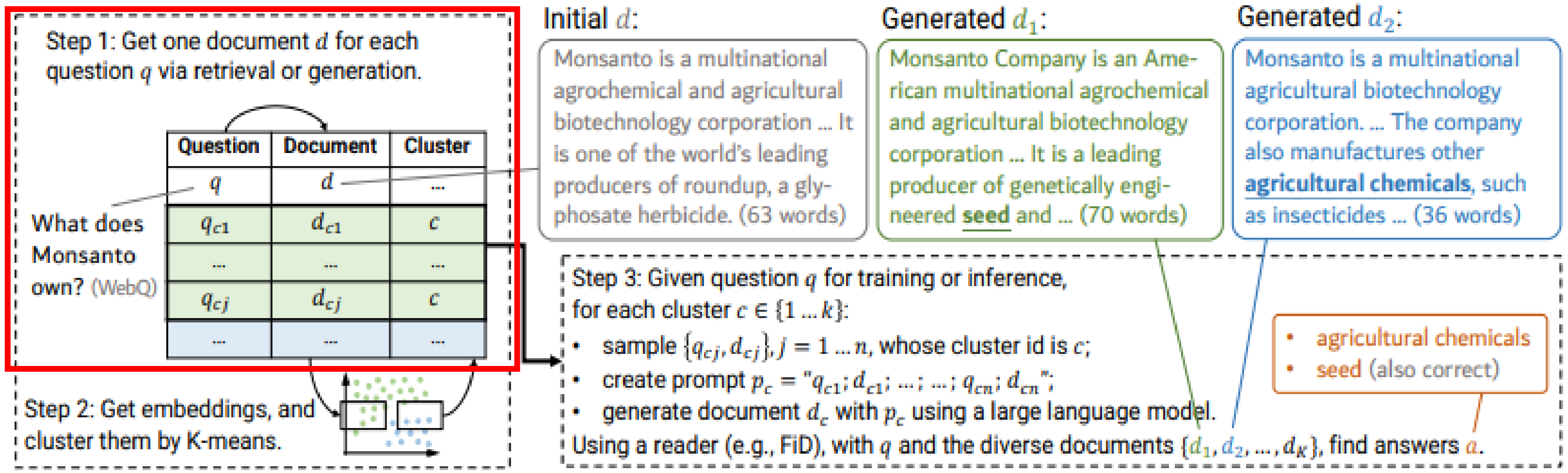


Figure 1: An overall framework of clustering-based prompting method. It leverages distinct question-document pairs sampled from each embedding cluster as in-context demonstrations to prompt a large language model to generate diverse documents, then read the documents to predict an answer.

Step-1

: Ask a large language model to generate one contextual document for each question

Method

Generate-then-Read

Supervised Setting

2. Clustering-Based Prompts

ENCODE EACH DOCUMENT, DO K-MEANS CLUSTERING

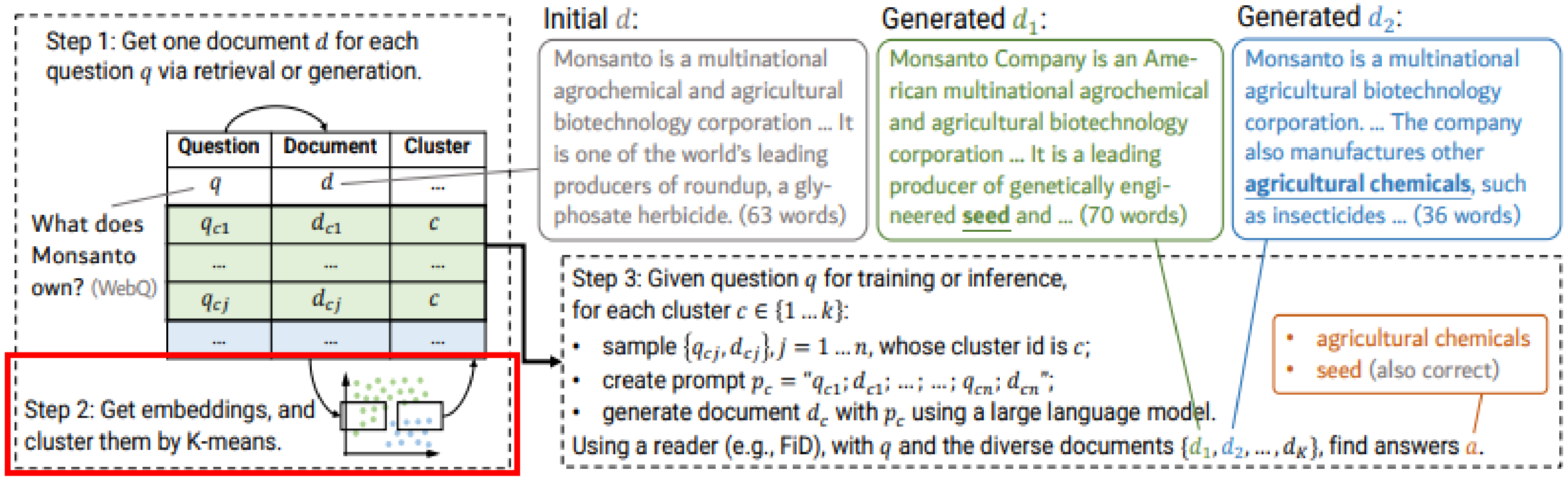


Figure 1: An overall framework of clustering-based prompting method. It leverages distinct question-document pairs sampled from each embedding cluster as in-context demonstrations to prompt a large language model to generate diverse documents, then read the documents to predict an answer.

Step-2

: Use LLM to encode each question-document pair
⇒ 12,288-dimensional vector per document

: And Do K-means to cluster all embedding vectors

Method

Generate-then-Read

Supervised Setting

2. Clustering-Based Prompts

SAMPLE AND GENERATE K DOCUMENTS

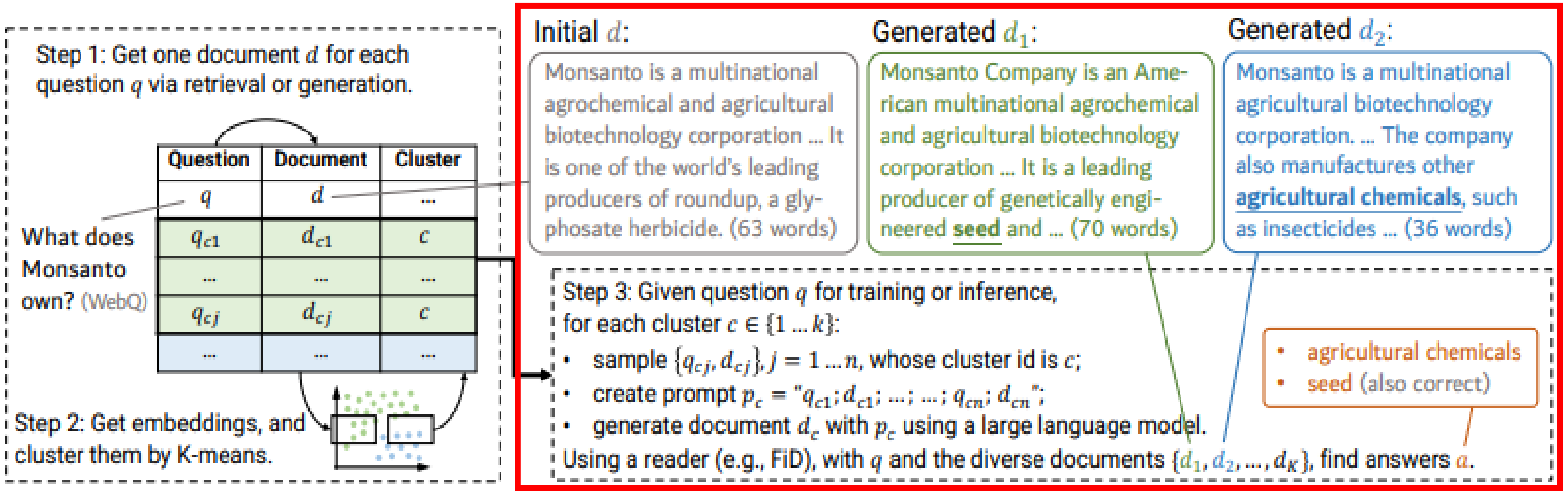


Figure 1: An overall framework of clustering-based prompting method. It leverages distinct question-document pairs sampled from each embedding cluster as in-context demonstrations to prompt a large language model to generate diverse documents, then read the documents to predict an answer.

Step-3

: Sample n question-document pairs from each cluster c

⇒ Finally get K-generated documents

Experiments

Zero-Shot Setting

Models	Open-domain QA NQ	TriviaQA	WebQ	Fact Checking FEVER	FM2	Dialogue System WoW (F1 / R-L)	
**with retriever, AND directly trained on these datasets*							
DPR + InstructGPT*	29.1	53.8	20.2	79.8	65.9	15.4	13.7
**with retriever, BUT NOT trained on these datasets*							
BM25 + InstructGPT	19.7	52.2	15.8	78.7	65.2	<u>15.7</u>	13.7
Contriever + InstructGPT	18.0	51.3	16.6	80.4	**66.6**	15.5	<u>14.0</u>
Google + InstructGPT	**28.8**	<u>58.8</u>	<u>20.4</u>	**82.9**	<u>66.0</u>	14.8	13.2
**without retriever, and not using external documents*							
Previous SoTA methods	24.7[1]	56.7[2]	19.0[1]	-	-	-	-
InstructGPT (no docs.)	20.9	57.5	18.6	77.6	59.4	15.4	13.8
GENREAD (InstructGPT)	<u>28.0</u>	**59.0**	**24.6**	<u>80.4</u>	65.5	**15.8**	**14.2**

Table 1: Zero-shot open-domain QA performance. Our proposed GENREAD with the InstructGPT reader (named GENREAD (InstructGPT)) can significantly outperform the original InstructGPT, achieving new state-of-the-art performance on three open-domain QA benchmarks (previous SoTA: [1]GLaM (Du et al., 2022), [2]FLAN (Wei et al., 2021)) under this setting without using any external document. Our GENREAD can achieve comparable or even better performance than zero-shot *retrieve-then-read* models that use a retriever or search engine to first obtain contextual documents. To ensure reproducibility, we use greedy search in decoding. All prompts used are shown in the §B.1. Note: fix numbers in v2 by adding average performance of different prompts, see details in Table 20.

Experiments

Supervised Setting

Open Domain QA Performance

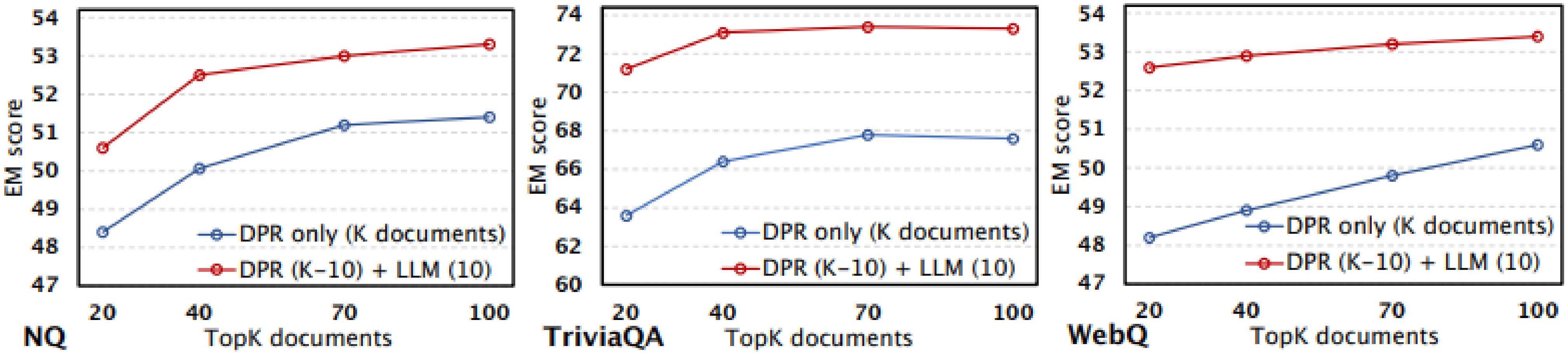
Models	# reader parameters	# documents	TriviaQA open test	WebQ open test	NQ open test	Avg.
**baselines with retrieving from Wikipedia; all numbers reported by existing papers*						
DPR (Karpukhin et al., 2020)	110M	100	56.8	41.1	41.5	46.5
RAG (Lewis et al., 2020)	400M	10	56.1	45.2	44.5	48.6
FiD (Izacard & Grave, 2021)	770M	100	67.6	50.5	<u>51.4</u>	56.5
**baselines with retrieving from Wikipedia or Google; all numbers from our experiments*						
FiD-l (DPR, Wikipedia)	770M	10	61.9	48.1	46.7	52.2
FiD-xl (DPR, Wikipedia)	3B	10	66.3	50.8	50.1	55.7
FiD-xl (Google search)	3B	10	70.1	53.6	45.0	56.2
**our proposed method by leveraging a large language model to generate documents*						
GENREAD (FiD-l) (sampling)	770M	10	67.8	51.5	40.3	53.2
GENREAD (FiD-l) (clustering)	770M	10	70.2	53.3	43.5	55.6
GENREAD (FiD-xl) (sampling)	3B	10	69.6	52.6	42.6	54.9
GENREAD (FiD-xl) (clustering)	3B	10	<u>71.6</u>	<u>54.4</u>	45.6	<u>57.1</u>
⊢ merge retrieved documents with generated documents			**74.3**	**56.2**	**54.0**	**61.5**

Table 2: Supervised open-domain QA performance. By only using generated documents from InstructGPT, our GENREAD with FiD reader (named GENREAD (FiD)) can achieve better performance than baseline methods on TriviaQA and WebQ. Through our detailed analysis of NQ, we found the performance gap mainly due to the temporality issue, which will be elaborated in §A.7.

- FiD model performs the best among all baseline models
- GENREAD can outperform Google search on all benchmarks
- Clustering-based prompt method is effectively increasing the knowledge coverage

Experiments

Supervised Setting



- Increasing the number of documents can lead to better model performance and achieve state-of-the-art when using 100 documents
- DPR retrieved documents with large language model (LLM) generated documents can achieve significantly better performance than using DPR retrieved documents only

Experiments

Supervised Setting

On Other Tasks

Models	FEVER Acc.	FM2 Acc.	WoW F1 / R-L
RAG (Lewis et al., 2020)	86.3	71.1	13.1 / 11.6
FiD (Izacard & Grave, 2021)	90.2	77.6	17.5 / 16.1
GENREAD (FiD-xl) (sampling)	89.0	76.3	18.9 / 16.7
GENREAD (FiD-xl) (clustering)	89.6	77.8	19.1 / 16.8
⊢ merge two source docs.	**91.8**	**78.9**	**20.1 / 17.9**

Table 3: Supervised performance on fact checking (FEVER and FM2) and open-domain dialogue system (WoW).

- GENREAD can achieve on par performance on the fact checking task a nd superior performance on the dialogue system task

⇒ **Large language model can be seen as a strong knowledge generator**

Coverage Analysis

Documents obtained by ↓	NQ	TriviaQA		WebQ
	-	w. alias	w/o alias	-
BM25 (Robertson et al., 2009)	48.4	17.1	63.8	41.2
Google search engine[3]	57.9	18.9	72.0	54.2
DPR (Karpukhin et al., 2020)	**67.9**	17.9	67.3	58.8
GENREAD (nucleus sampling)	56.6	19.6	74.5	59.8
GENREAD (10 human prompts)	57.4	20.1	74.8	61.1
GENREAD (clustering prompts)	61.7	**20.4**	**76.5**	**62.1**

Table 4: Answer coverage (%) over 10 retrieved or generated documents. Case studies are provided in Tables 16-19 in Appendix.

- Generated documents tends to have little diversity compared to retrieved documents

⇒ Generated text tends to have lower coverage than retrieved documents

⇒ **GENREAD with clustering improves coverage**

Experiments

Examples

Original question	NQ labels	Correct labels
Q: When is the last time the philadelphia won the superbowl?	Super Bowl LII; 2017	2018; February 4, 2018
DPR: 2017 ✘; **Google search:** 2018 ✔; **GENREAD :** February 4, 2018 ✔		
Q: Who has the most big ten championships in football?	Michigan	Ohio State
DPR: Michigan ✘; **Google search:** Ohio State ✔; **GENREAD :** Ohio State ✔		
Q: Who has the most super bowls in nfl history?	Pittsburgh Steelers	Pittsburgh Steelers; New England Patriots
DPR: Pittsburgh Steelers ✔; **Google search:** New England Patriots ✔; **GENREAD :** New England Patriots ✔		
Q: How many casinos are in atlantic city new jersey?	11; eleven	9; nine
DPR: eleven ✘; **Google search:** nine ✔; **GENREAD :** nine ✔		
Q: When did the us not go to the olympics?	1980	1980; 1984
DPR: 1980 ✔; **Google search:** 1980 and 1984 ✔; **GENREAD :** 1984 ✔		
Q: Largest cities in the world by population?	Beijing	Tokyo
DPR: Beijing ✘; **Google search:** Tokyo ✔; **GENREAD :** Tokyo ✔		
Q: Who has most followers on instagram in world?	Selena Gomez	Cristiano Ronaldo
DPR: Instagram ✘; **Google search:** Cristiano Ronaldo ✔; **GENREAD :** Cristiano Ronaldo ✔		
Q: Who is the no. 1 ranked tennis player in the world?	Rafael Nadal	Novak Djokovic
DPR: Rafael Nadal ✘; **Google search:** Novak Djokovic ✔; **GENREAD :** Novak Djokovic ✔		

Table 13: Case studies of temporality issues of the NQ dataset. All these questions are drawn from Zhang & Choi (2021), which contains a subset of NQ data examples with time-dependent questions.

Experiments

Examples

Original question	*DPR-FiD* predictions	GENREAD predictions
Q: Who played lionel in as time goes by?	Geoffrey Dyson Palmer	Geoffrey Palmer
Explanation: The labeled answer is "Geoffrey Dyson Palmer", however, "Geoffrey Palmer" is also correct. **DPR retrieved documents:** Geoffrey Dyson Palmer, (born 4 June 1927) is an English actor known for his roles in British television sitcoms playing Jimmy Anderson in "The Fall and Rise of Reginald Perrin", Ben Parkinson in "Butterflies" and Lionel Hardcastle in "As Time Goes By". His film appearances include "A Fish Called Wanda", "The Madness of King George", "Mrs. Brown", and "Tomorrow Never Dies". **GPT generated documents:** As Time Goes By is a British sitcom that aired on BBC One from 1992 to 2005. The show starred Geoffrey Palmer and Judi Dench as Lionel and Jean Pargetter, a middle-aged couple who reunite after many years apart. Lionel was played by Palmer, who was also a writer on the show.		
Q: How many cracker barrels in the united states?	645	over 630
Explanation: The labled answer is "639" or "over 600", so "over 630" is also a reasonable answer.		
Q: Where do the greasers live in the outsiders?	Tulsa, Oklahoma	Oklahoma
Explanation: The labled answer is "Tulsa, Oklahoma", but "Oklahoma" is also a correct answer. **DPR retrieved documents:** The movie received mostly positive reviews from critics, and performed well at the box office, grossing 33 million on a 10 million budget. In Tulsa, Oklahoma, greasers are a gang of tough, low-income working-class teens. They include Ponyboy Curtis and his two older brothers, Sodapop and Darrel, as well as Johnny. **GPT generated documents:** The Outsiders is a novel by S.E. Hinton. It is about a gang of greasers in Oklahoma in the 1960s. The greasers are from the poor side of town and constantly in trouble with the law.		
Q: Where are unipolar neurons found in spinal cord?	the granule region	dorsal root ganglia
Explanation: The labled answer is "the distal dorsal root", but the output "dorsal root ganglia" is the same.		

Table 14: Case studies of the incomplete answers issue of the NQ dataset. Since the labels in NQ are spans identified from Wikipedia passages, it is easier for *DPR-FiD* to predict correct answers.

Query: What does Monsanto own?
Answers: agricultural chemicals, seed, agriculture, chemical industry

DPR #1: In September 2016, Monsanto agreed to be acquired by Bayer for US$66 billion. In an effort to receive regulatory clearance for the deal, Bayer announced the sale of significant portions of its current agriculture businesses, including its seed and herbicide businesses, to BASF. The deal was approved by the European Union on March 21, 2018, and approved in the United States on May 29, 2018. The sale closed on June 7, 2018; Bayer announced its intent to discontinue the Monsanto name, with the combined company operating solely under the Bayer brand. ... (Wikipedia id: 935126 – "Monsanto")

DPR #2: Behind it all is of course an evil corporation: Monsanto. GMO conspiracy theories GMO conspiracy theories are conspiracy theories related to the production and sale of genetically modified crops and genetically modified food (also referred to as genetically modified organisms or "GMOs"). These conspiracy theories include claims that agribusinesses, especially Monsanto, have suppressed data showing that GMOs cause harm, deliberately cause food shortages to promote the use of GM food, or have co-opted government agencies ... (Wikipedia id: 19304897 – "GMO conspiracy theories")

Among top-10 retrieved documents: agriculture: 1 time seed: 2 times

GPT nucleus #1: Monsanto is an American agricultural biotechnology corporation founded in 1901. It is headquartered in Creve Coeur, Missouri. Monsanto is a leading producer of genetically engineered seed and Roundup, a glyphosate-based herbicide.

GPT nucleus #2: Monsanto is a multinational agrochemical and agricultural biotechnology corporation headquartered in Creve Coeur, Missouri. It is one of the world's leading producers of Roundup, a glyphosate-based herbicide. The company also manufactures genetically engineered seed and Roundup Ready crops, including corn, soybeans, and cotton. As of 2018, Monsanto is the world's largest producer of glyphosate. In 2018, Bayer acquired Monsanto for $63 billion.

Among top-10 generated documents: seed: 6 times

GPT cluster #1: Monsanto Company is an American multinational agrochemical and agricultural biotechnology corporation headquartered in Creve Coeur, St. Louis, Missouri. It is a leading producer of genetically engineered (GE) seed and of the herbicide glyphosate, which it markets under the Roundup brand. Founded in 1901 by John Francis Queeny, Monsanto spent most of the 20th century as a chemical company, becoming one of the world's leading manufacturers of plastics, resins, and other chemicals.

GPT cluster #2: Monsanto is a biotechnology and agriculture company that specializes in genetically modified organisms (GMOs). The company owns a number of patents on GMOs, and it produces a range of products that contain GMOs, including herbicides, pesticides, and seeds.

GPT cluster #3: Monsanto is a multinational agricultural biotechnology corporation. It is one of the world's leading producers of genetically engineered seed and Roundup, a glyphosate-based herbicide. The company also manufactures other agricultural chemicals, such as insecticides and fungicides.

Among top-10 generated documents: agriculture: 2 times seed: 5 times agricultural chemicals: 1 time

Table 17: Case study of retrieved documents and generated documents from language models (2/4).

Conclusion

Contribution

1. Present a novel perspective for solving knowledge-intensive tasks

: by replacing document retrievers with large language model generators

2. Propose a novel clustering-based prompting method

: that selects distinct prompts, in order to generate diverse documents that cover different perspectiv es

3. Conduct extensive experiments on three different knowledge-intensive tasks

: including open-domain QA, fact checking, and dialogue system.

Guess The Instruction! Flipped Learning Makes Language Models Strong Zero-Shot Learners

Seonghyeon Ye, Doyoung Kim, Joel Jang, Joongbo Shin, Minjoon Seo

ICLR2023

Objective

Meta-training

- **Fine-tunes the language model (LM) on various downstream given the task instruction and input instance**

⇒ Leads to significant improvement in zero-shot task generalization

- **LMs meta-trained through this standard approach are sensitive to different label words**

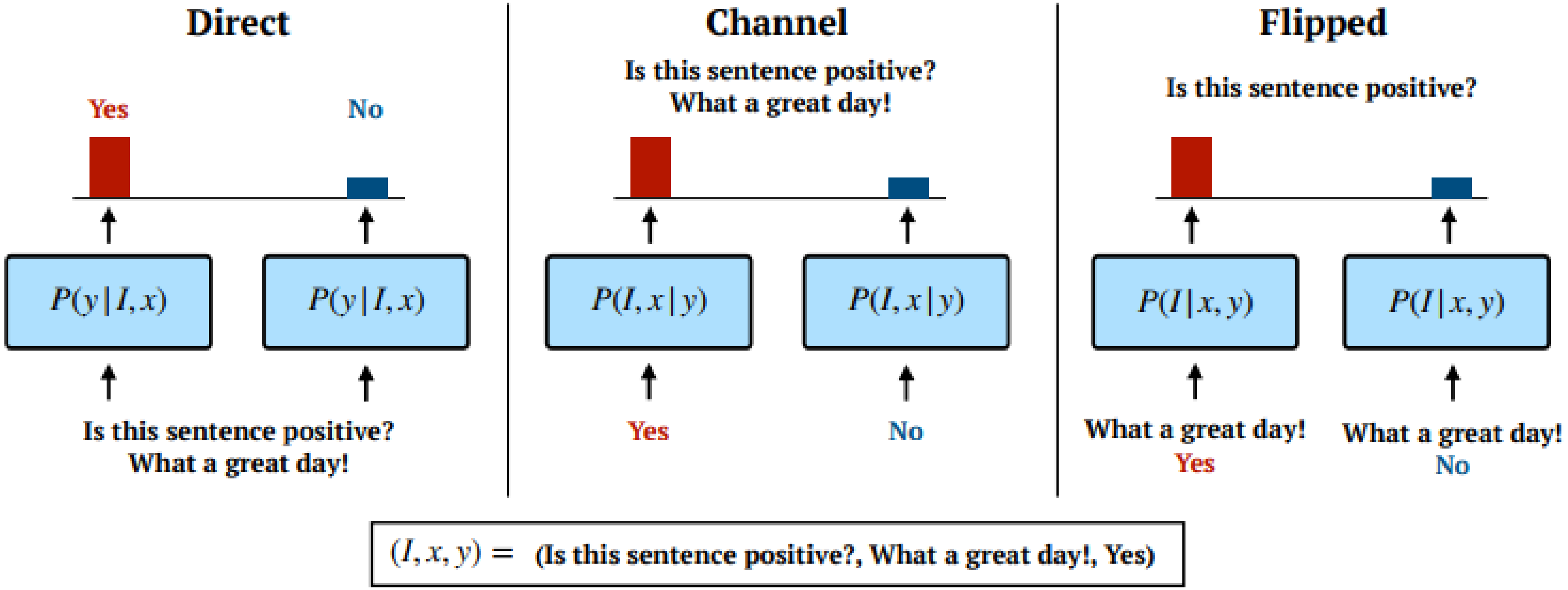
⇒ Fail to generalize to tasks that contain novel labels

=> " FLIPPED LEARNING "

Method

Flipped Learning

INFERENCE OF PROBABILISTIC LMS



Method

Flipped Learning

META-TRAINING USING FLIPPED LEARNING

$$\arg\max_{l_i} P(l_i|I,x) = \arg\max_{l_i} \frac{P(I|x,l_i)P(l_i,x)}{P(I,x)} = \arg\max_{l_i} P(I|x,l_i)P(l_i|x) \approx \arg\max_{l_i} P(I|x,l_i)$$

- Computes the conditional probability of the task instruction given an input instance and a label
- Allow the LM to put more focus on the task instruction

< Hypothesize >

FLIPPED shows strong zero-shot generalization ability on unseen tasks
because of the improved generalization capability to unseen labels

Method

Flipped Learning

$$L = L_{LM} + \lambda L_{UL}$$

Unlikelihood Loss

- Meta-training ignoring the correspondence between the input instance and label

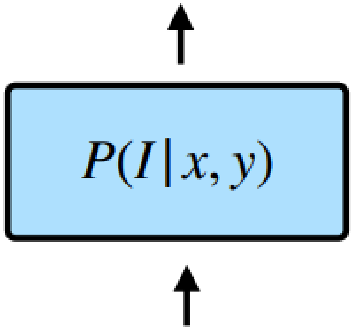
⇒ meta-trained LM generates task instruction regardless of the correspondence of the label option

$$L_{UL} = -\sum_{t=1}^{T} \log(1 - P(I_t|x, l_{c'}, I_{<t}))$$

- Unlikelihood loss term allows the LM to not generate the task instruction if the label option does not correspond to the input instance

⇒ **Strengthening the correspondence**

<extra_id_0> Using only the above description and what you know about the world, is "<extra_id_1> " definitely correct? Yes or no?



input: The girl was found in Drummondville. Drummondville contains the girl.
output: Yes

Experiments

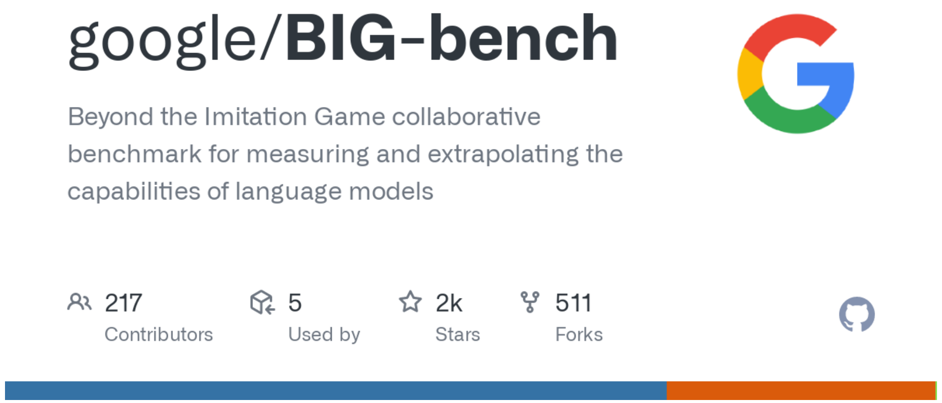
Setup

Training

- Utilize the subset of T0 (Sanh et al., 2021) meta-training datasets
- 4 task clusters (sentiment classification, paraphrase detection, topic classification, multi-choice QA), which are 20 datasets in tot al

Evaluation

- Measure unseen task generalization performance on 14 tasks of **BIG-bench**
- 14 English NLP unseen tasks, consisting of 7 classification and 7 multi-choice datasets



- **2 seen** datasets during meta-training (IMDB, PAWS) and **3 unseen** datasets (RTE, CB, WiC)

Experiments

Main Results

Dataset (metric)	T0 3B	DIR. 3B	CHAN. 3B	FLIP. 3B	T0 11B	FLIP. 11B	GPT-3 175B	PALM 540B	GPT-3 (3) 175B	PALM (1) 540B
	Zero-shot								*Few-shot*	
Known Un.	47.83	63.04	52.17	71.74	58.70	**86.96**	60.87	56.52	50.00	67.39
Logic Grid	41.10	35.90	30.90	41.70	38.30	**42.50**	31.20	32.10	31.10	42.20
Strategy.	52.79	53.28	53.01	53.19	52.75	53.23	52.30	**64.00**	57.10	69.00
Hindu Kn.	25.71	50.29	16.57	47.43	29.71	52.57	32.57	**56.00**	58.29	94.86
Movie D.	52.85	47.15	51.06	47.93	**53.69**	48.49	51.40	49.10	49.40	57.20
Code D.	46.67	33.33	**71.67**	45.00	43.33	60.00	31.67	25.00	31.67	61.67
Concept	45.52	58.14	35.67	61.64	**69.29**	64.93	26.78	59.26	35.75	80.02
Language	14.84	22.01	11.55	19.01	20.20	**26.87**	15.90	20.10	10.90	37.30
Vitamin	58.89	63.83	15.73	57.07	64.73	**65.57**	12.30	14.10	52.70	70.40
Syllogism	**52.94**	49.85	50.43	50.56	51.81	50.39	50.50	49.90	52.80	52.20
Misconcept.	50.23	50.23	47.79	46.58	50.00	**54.34**	47.95	47.49	60.27	77.63
Logical	46.64	38.06	25.73	59.82	54.86	**64.56**	23.42	24.22	33.93	34.42
Winowhy	44.29	44.33	**55.36**	53.33	52.11	55.08	51.50	45.30	56.50	47.50
Novel Con.	15.63	3.13	15.63	25.00	15.63	**46.88**	**46.88**	**46.88**	56.25	59.38
BIG-bench AVG	42.56	43.75	38.07	48.57	46.79	**55.17**	38.23	42.14	45.48	60.80

14 datasets in BigBench

- **DIRECT outperforms T0-3B**
- **CHANNEL is not effective for task generalization**
- **FLIPPED outperforms baselines**
- **FLIP 3B > T0 11B**

FLIPPED is effective for generalizing to unseen tasks that are challenging

Experiments

Main Results

Seen | Unseen

Dataset (metric)	T0 3B	DIR. 3B	CHAN. 3B	FLIP. 3B	T0 11B	FLIP. 11B	GPT-3 175B
RTE (F1)	61.89	72.83	36.62	71.03	**80.91**	72.20	40.68
CB (F1)	30.94	49.81	22.35	52.27	53.82	**61.51**	29.72
ANLI R1 (F1)	24.39	30.17	21.30	33.92	34.72	**34.93**	20.90
ANLI R2 (F1)	23.73	28.23	21.44	**32.62**	31.25	32.59	22.50
ANLI R3 (F1)	23.45	30.41	22.50	34.65	33.84	**34.77**	23.77
WSC (F1)	54.64	50.35	46.38	52.82	**58.36**	49.88	26.24
WiC (F1)	38.53	36.42	38.69	37.36	**51.64**	39.26	45.36
COPA	75.88	89.63	50.13	89.88	**91.50**	90.75	91.00
Hellaswag	27.43	31.61	20.82	41.64	33.05	41.97	**78.90**
StoryCloze	84.03	94.24	57.84	95.88	92.40	**96.12**	83.20
Winogrande	50.97	55.96	50.99	58.56	59.94	66.57	**70.20**
PIQA	56.63	62.60	47.08	67.32	67.67	71.65	**81.00**
ARC-Chall	51.10	49.30	29.23	49.63	56.99	**64.62**	51.40
OpenbookQA	42.66	54.00	38.57	62.11	59.11	**72.54**	68.80
En NLP AVG	46.16	52.54	36.00	55.69	57.51	**59.24**	52.41
En NLP STD (↓)	4.74	4.36	4.58	3.29	5.24	**3.11**	-

- Direct show strong performance for seen task
- FLIPPED shows strong performance on unseen task

FLIPPED is not only effective for zero-shot task generalization but also robust to different surface forms of the instruction

Experiments

Main Results

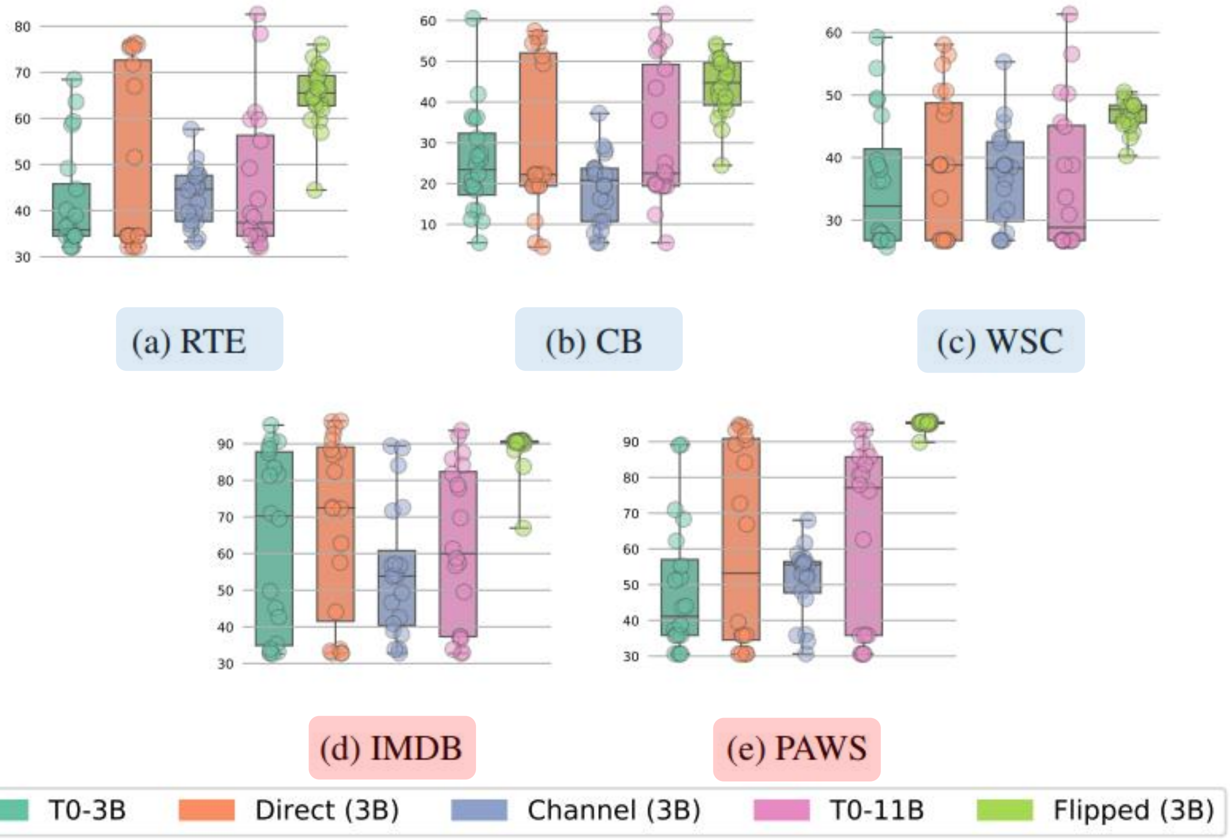


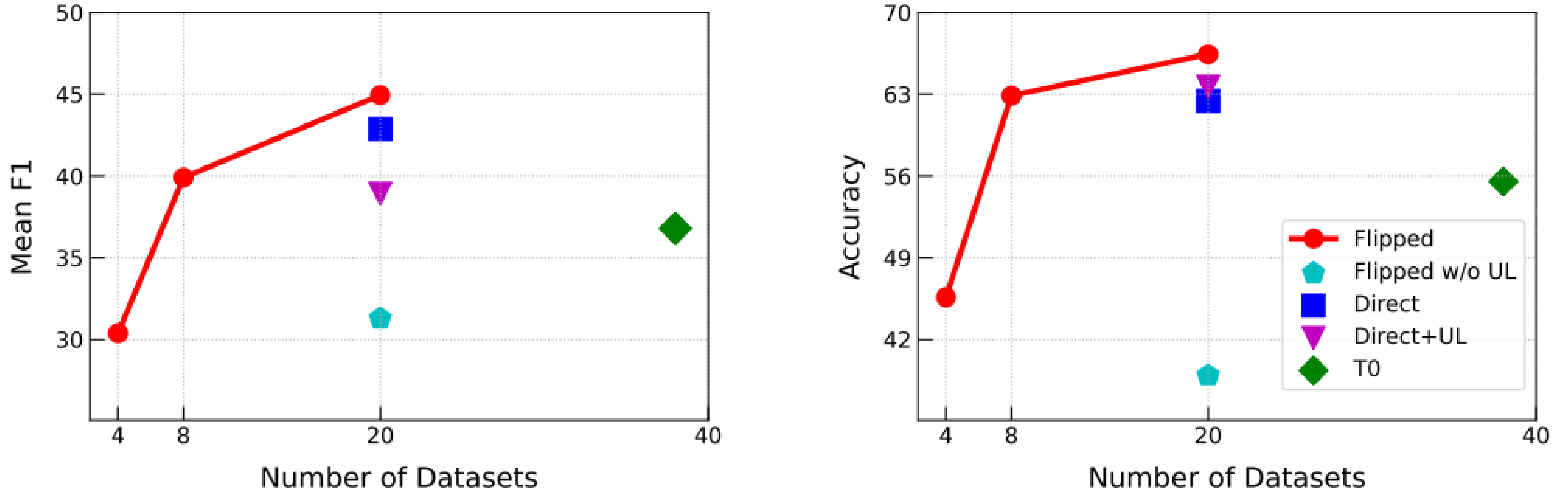
Figure 3: Label generalization performance on 3 unseen and 2 seen datasets during meta-training. We evaluate on 20 different label pairs including many unseen labels. Result shows that FLIPPED significantly outperforms other baseline models.

- **Standard meta-training leads to label overfitting**

FLIPPED avoids this by conditioning on the label option instead of generating it

Experiments

Ablation



- DIRECT+UL < DIRECT

⇒ Effectiveness of FLIPPED is not coming from unlikelihood training itself

- 8 datasets with FLIPPED > 20 datasets with DIRECT on multi-choice tasks

⇒ FLIPPED not only effective but also efficient zero-shot learners

Conclusion

Contribution

1. Propose FLIPPED LEARNING

: a novel meta-training method that computes the likelihood of the task instruction given the concaten ation of input instance and label

2. 11B-sized FLIPPED outperforms not only meta-trained T0-11B, but also 16x larger 3-shot GPT-3

: FLIPPED outperforms all baseline models on average

3. FLIPPED is effective on generalization to labels that are unseen during meta-training

: not only effective but also efficient / avoids label overfitting

Leveraging Large Language Models For Multiple Choice Question Answering

Joshua Robinson, Christopher Michael Rytting, David Wingate

ICLR2023

Objective

Limitations In Prompting MCQA

- **While LLMs have achieved SOTA results on many tasks, they generally fall short on MCQA**

⇒ MCQA ability of LLMs has been previously underestimated

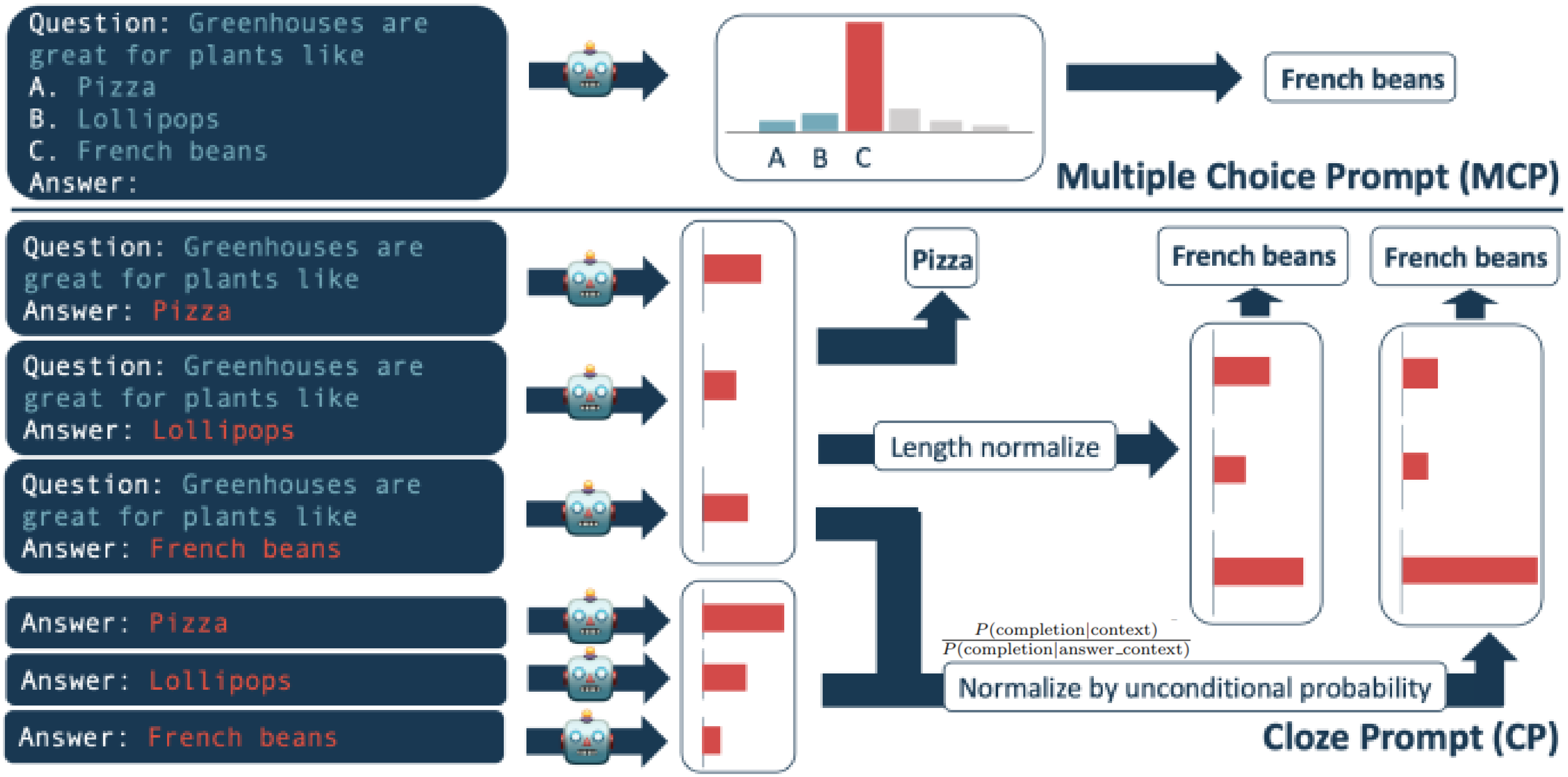
- **Cloze Prompting:**
  1. Conflation of likelihood as answer and likelihood as natural language
  2. Computational expense of scoring multiple candidate answers
  3. No direct comparison between answers
  4. Reliance on normalization procedures

⇒ **MCP can substantially improve LLM accuracy across a diverse set of tasks**

Method

Multiple Choice Prompting(MCP)

Multiple Choice Prompting vs Cloze Prompting



MCP

- Question and its symbol-enumerated candidate answers are all passed to LLM as a single prompt
- Symbols serve as a proxy for each answer's probability

Method

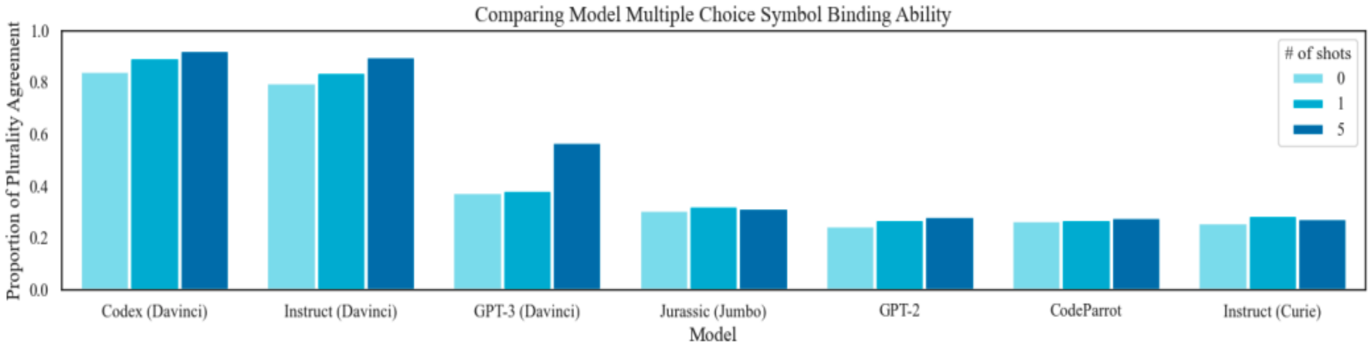
Multiple Choice Prompting(MCP)

Multiple Choice Symbol Binding

- **Problem:**
  - Humans' answers to such questions are generally order-invariant
  - Simply changing the order of the candidate answers changes the model's answer

Proportion of Plurality Agreement (PPA):

- N answer options => N! combination
- The proportion of orderings that chose the plurality answer among all orderings



⇒ If the model is highly reliable, the PPA should be high

⇒ Codex and Instruct significantly outperform the other models

Experiments

Setup

Models

- Codex / InstructGPT / GPT-3
- Zero-shot / One-shot / Few-shot
- Not to maximize accuracy by extensive prompt engineering => **Simple Prompt**
- K is always chosen to be as high as possible while respecting Codex's 4,000 token context limit

Task

- Multiple choice prompts across a set of 20 diverse datasets
- Common Sense Reasoning / Natural Language Inference / Cloze and Completion / Text Classification / Winograd-style / Reading Comprehension

Experiments

Main Results

Model Performance Across Prompting Strategies

Dataset	GPT-3				Instruct				Codex			
	Raw	LN	UN	MCP	Raw	LN	UN	MCP	Raw	LN	UN	MCP
OpenBookQA	35.0	46.8	**57.4**	41.4	41.8	49.6	58.4	**77.4**	43.0	51.4	65.6	**83.0**
StoryCloze	75.2	**76.4**	75.6	70.8	78.0	78.8	82.4	**97.6**	80.8	83.6	84.0	**97.4**
RACE-m	55.6	**57.2**	56.6	50.2	63.2	64.8	66.8	**89.6**	63.4	67.0	63.8	**89.2**

Table 1: Comparison of large language model performance across prompting strategies. The three cloze prompting normalization strategies are described in Section 3. MCP is multiple choice prompting. The best accuracy for each model and dataset is bolded.

- CP differs largely by normalization strategy
- MCP always performs best for Instruct and Codex

⇒ **High multiple choice symbol binding ability => effectively leverage MCP prompts across tasks**

Experiments

Main Results

Dataset	N	K	Zero-Shot		One-Shot		Few-Shot		Server	SOTA
			CP	MCP	CP	MCP	CP	MCP		
AG News	4	38	68.2	**83.5**	77.6	**87.1**	**90.1**	89.4		95.6[a]
ANLI R1	3	27	**45.3**	33.2	35.6	**61.7**	58.4	**64.2**		75.5[b]
ANLI R2	3	26	**39.2**	33.6	35.7	**53.0**	51.8	**55.2**		58.6[c]
ANLI R3	3	26	**37.8**	34.3	35.5	**47.8**	54.2	**54.5**		53.4[c]
ARC (Challenge)	4	50	58.9	**81.7**	64.1	**82.8**	66.6	**86.1**		86.5[d]
ARC (Easy)	4	57	84.2	**93.1**	85.9	**93.5**	87.8	**94.7**		94.8[d]
CODAH	4	63	56.8	**76.0**	65.4	**87.8**	73.6	**91.9**		84.3[e]
CommonsenseQA	5	79	68.5	**72.0**	73.1	**78.9**	78.6	**83.2**	76.6	79.1[f]
COPA	2	113	**92.0**	89.0	95.0	**99.0**	96.0	**100.0**	—	99.2[d]
Cosmos QA	4	24	43.0	**75.5**	44.0	**81.8**	38.1	**82.4**	83.5	91.8[g]
DREAM	3	7	72.7	**91.3**	82.5	**93.3**	84.3	**94.1**		92.6[h]
Fig-QA	2	99	79.6	**84.7**	82.4	**86.7**	82.5	**94.0**	93.1	90.3[i]
HellaSwag	4	16	—	71.0	—	75.1	—	73.6	—	93.9[g]
LogiQA	4	16	36.6	**44.5**	37.5	**45.3**	37.8	**47.3**		42.5[j]
MedMCQA	4	58	37.8	**52.1**	42.1	**53.9**	41.2	**54.4**	58.0	41.0[k]
MMLU	4	5	49.5	**62.1**	—	68.2	—	69.5		67.5[l]
OpenBookQA	4	83	63.2	**72.0**	64.0	**81.6**	71.2	**87.0**		87.2[f]
PIQA	2	35	**83.7**	73.7	**84.1**	81.8	**86.1**	84.5	—	90.1[g]
RACE-h	4	4	52.3	**82.1**	53.2	**85.1**	55.2	**86.2**		89.8[m]
RACE-m	4	8	67.5	**85.4**	70.5	**89.3**	71.7	**90.3**		92.8[m]
RiddleSense	5	59	**79.8**	67.6	**89.1**	77.1	**91.3**	83.9	80.0	68.8[f]
Social IQa	3	72	52.1	**64.4**	58.1	**72.2**	62.4	**74.9**	76.0	83.2[g]
StoryCloze	2	44	80.3	**97.5**	83.4	**98.3**	88.2	**98.5**		89.0[n]
Winogrande (XL)	2	102	62.5	**64.5**	**71.6**	**71.6**	**75.5**	72.1	72.3	91.3[g]
Winogrande (XS)	2	102	63.0	**64.8**	71.0	**71.3**	**76.2**	73.6	73.8	79.2[g]

MCP vs CP with Codex

- Without reliance on normalization and with 4.3x less API calls than the chosen CP strategies

- AG News, Winogrande, RiddleSense tend to have short, often one word answers

⇒ CP is acting more like MCP

- Cosmos QA have somewhat irregular spacing

⇒ No issue for MCPs, but serious issue for CPs

Experiments

Main Results

Answer Choice Corruption

Corruption	OpenBookQA				StoryCloze				RACE-m			
	Raw	LN	UN	MCP	Raw	LN	UN	MCP	Raw	LN	UN	MCP
None	43.0	51.4	65.2	**82.4**	81.0	83.6	83.8	**97.4**	63.2	66.4	64.0	**89.4**
Caps	31.4	43.0	49.6	**79.8**	63.6	71.4	70.4	**96.8**	50.6	57.0	52.6	**88.8**
Space	32.2	43.4	44.4	**80.6**	71.6	78.2	71.2	**98.0**	53.0	63.2	51.2	**89.0**

Table 3: Comparison of Codex accuracy under different answer choice corruptions. The three cloze prompting normalization strategies are described in Section 3. MCP is multiple choice prompting. The best accuracy for each dataset and corruption type is bolded.

- Caps: randomly uppercase or lowercase each character
- Space: randomly add a space before, after, or within each word

⇒ **Benefits from direct comparison between answer choices**

⇒ **Benefits from separating likelihood of answer choices and their likelihoods in terms of natural language**

Experiments

Examples

Passage: [header] How to get around london easily [title] Know how you're going to travel. [step] The easiest method of travel in london is the tube. For this, it is easiest to buy what is called an' oyster card' or a get a travelcard for all zones from one of the automated machines in a tube station.
Question: Which choice best continues the passage?
A. People take an oyster card (this is a permanent, digital card) for optimal services and there are a number of reputable card companies that buy oyster cards. [title] Firstly, when considering destination, are you travelling with a package? [step] Do you want to surprise your friends and family at london.
B. These cover buses, tubes, trams and overground trains throughout the city. This is usually the best option, especially for tourists, as you can travel as much as you'd like in one day with one flat fare.
C. [title] Know the locations of the railway stations you are going to. [step] Look for normal bus lines around london.
D. The card lets you ride on the tube without the added cost of any rail, bus, or train stops. You can also travel by car (train makes easier to return for rides in london if you're travelling as non-railway cars), train from the station, or post office.
Answer:

Passage: (Kayaking) Man is kayaking in a calm river. Man is standing in te seasore talking to the camera and showing the kayak.
Question: Which choice best continues the passage?
A. man is getting in the sea and sits in a kayak.
B. man is kayaking in rafts and going through mountains.
C. man is kayaking on a snowy river.
D. man is returning in a river with a land trail and a shop.
Answer:

Figure 12: Prompt examples for the HellaSwag dataset. We include a WikiHow example (top) and an ActivityNet example (bottom) because they are formatted slightly differently.

Premise: press release: Did you know that Marquette University owns the original manuscripts for J. R. R. Tolkien's The Hobbit and The Lord of the Rings? William Fliss, Archivist in Marquette's Department of Special Collections and University Archives, will share the remarkable tale of how these literary treasures came to Wisconsin, and he will explain what these manuscripts can tell us about one of the most iconic authors of the twentieth century. Cost: Suggested donation of $3/person
Hypothesis: Attendees will pay $3.
A. Hypothesis is definitely true given premise
B. Hypothesis might be true given premise
C. Hypothesis is definitely not true given premise
Answer:

Figure 4: Prompt example for the ANLI dataset. Wording was taken from Gururangan et al. (2018).

Question: What adaptation is necessary in intertidal ecosystems but not in reef ecosystems?
A. the ability to live in salt water
B. the ability to use oxygen in respiration
C. the ability to cope with daily dry periods
D. the ability to blend into the surroundings
Answer:

Figure 5: Prompt example for the ARC dataset.

Conclusion

Contribution

1. LLM has enough ability for MCQA

: Simple change to prompting leads to drastic improvement

2. Formally define multiple choice symbol binding (MCSB): Required ability for an LLM to benefit from MCP

: High MCSB ability like OpenAI Codex leads to high performance in MCQA

⇒ Not all LLMs are equally skilled in this regard

3. Models most capable of MCSB can approach or beat SOTA

: Most capable of MCSB can individually approach or beat SOTA

: Solve many problems with CP

Thank you

Q&A