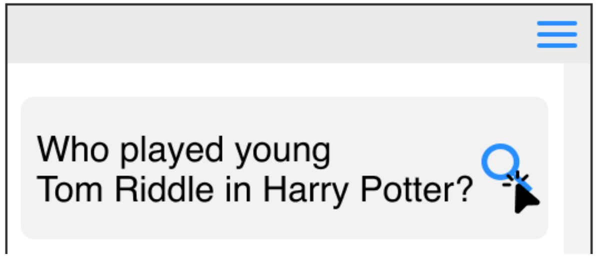
Asking Clarification Questions to Handle Ambiguity in Open-Domain QA

Dongryeol Lee*, Segwang Kim*, Minwoo Lee,
Hwanhee Lee, Joonsuk Park, Sang-woo Lee, Kyomin Jun

Machine Intelligence Lab
Seoul National University

Introduction: Ambiguity in Open-Domain QA



Ambiguity

- Ambiguity arises when there exist **multiple plausible answers** for the given Ambiguous Question (AQ).

Introduction: Format of Clarification Questions

AQ: Who played young Tom Riddle in Harry Potter?

CQ: Which version: young in series 2, child in series 6, or teenager in series 6?

: Category summarize the options

: Option represent single interpretation of AQ.

Introduction: Format of Clarification Questions

AQ: Who played young Tom Riddle in Harry Potter?

CQ: Which version: young in series 2, child in series 6, or teenager in series 6?

DQ_1: Who played young Tom Riddle in Harry Potter and the Chamber of Secrets?

Introduction: Format of Clarification Questions

AQ: Who played young Tom Riddle in Harry Potter?

CQ: Which version: young in series 2, **child in series 6,** or teenager in series 6?

DQ_2: Who played child version of Tom Riddle in Harry Potter and the Half Blood Prince?

Introduction: Format of Clarification Questions

AQ: Who played young Tom Riddle in Harry Potter?

CQ: Which version: young in series 2, child in series 6, or teenager in series 6?

DQ_3: Who played the teenage version of Tom Riddle in Harry Potter and the Half Blood Prince?

Dataset: CAMBIGNQ

Step 1: Generation via Instruct GPT

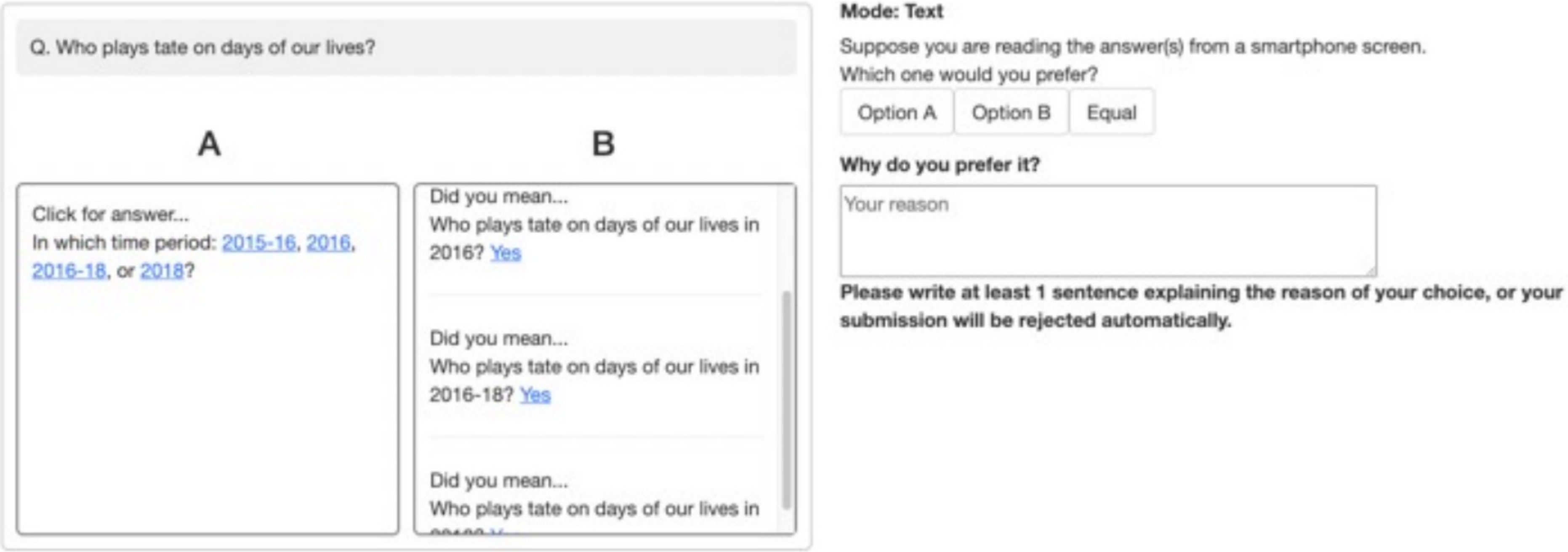
[Instruction], AQ_1, DQs_1, CQ_1,
… AQ_n, DQs_n, CQ_n,
AQ_{target}, DQ_{target} → **InstructGPT** → Pseudo CQ_{target}

Step 2: Manual Inspection and Revision by human annotators

Pseudo CQ_{target} →

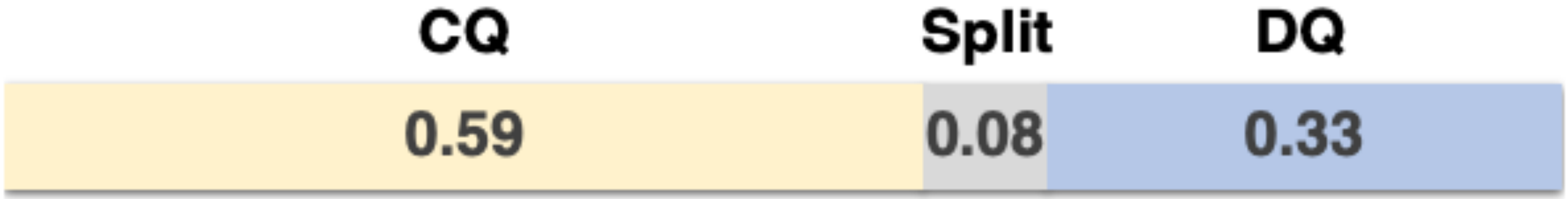
→ Revision (71.8%)
→ Pass (26.7%)

Human Preference Test



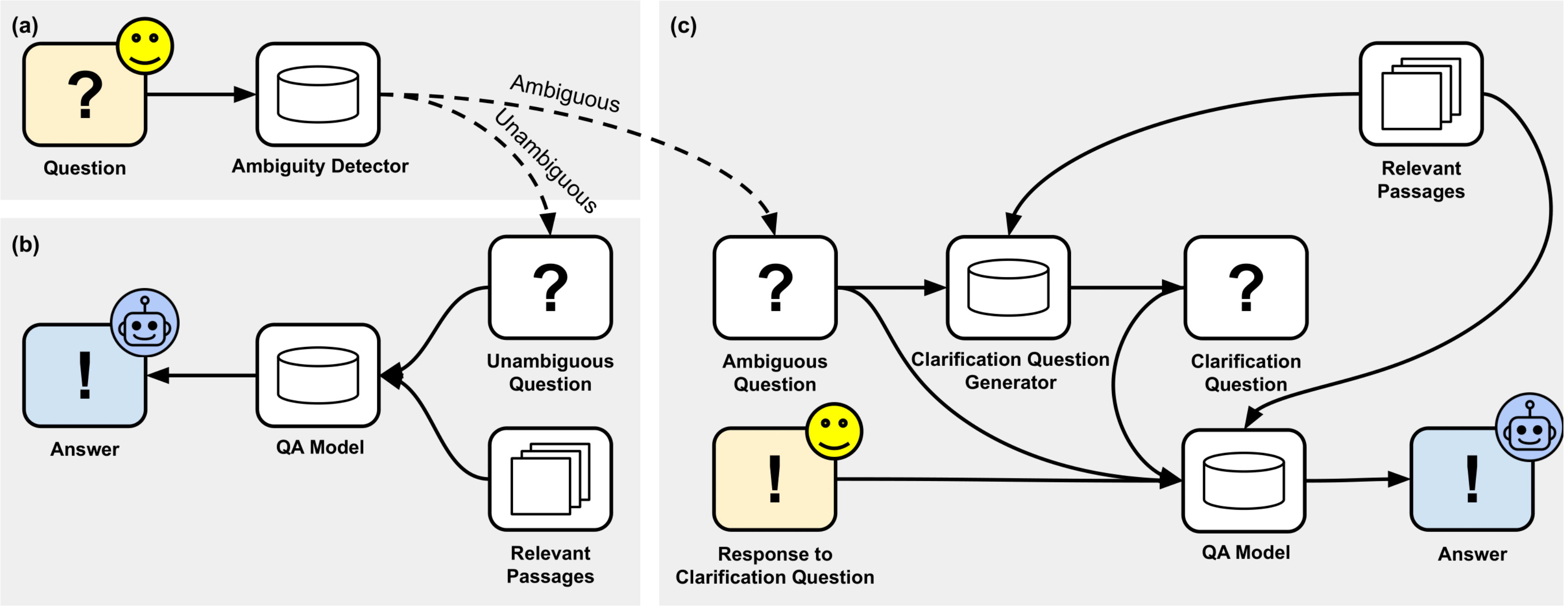
- Asked 200 mturkers for preference

Human Preference Test



- Our **proposed method CQ (59%)** is **preferred over DQ (33%).**
- The **prominent reasons for choice** was its **ease of use**, **conciseness**, **interactivity**, and **ability to provide clear guidance.**
- As the **number of option becomes larger**, the **preference rate of CQ increased**

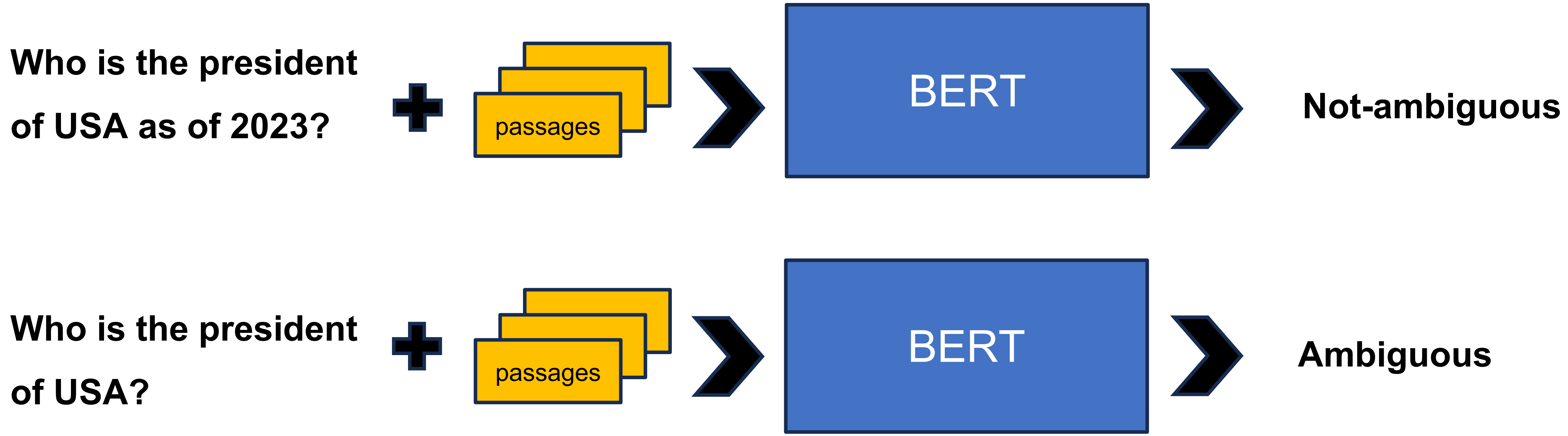
Task composition



- **Ambiguity Detection: Given a question q, classify whether q is ambiguous or not (binary classification)**
- **Clarification Questions Generation: Given AQ and relevant passages, generate a CQ**
- **Clarification-based QA: Given AQ, relevant passages, and a CQ, generate a unique answer for each option**

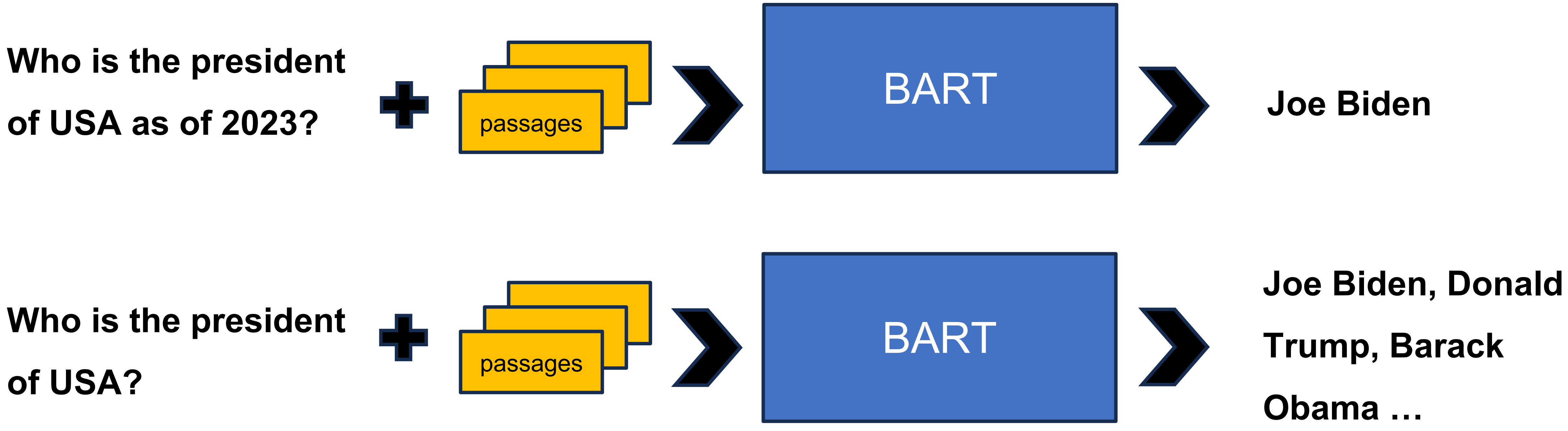
Task 1: Ambiguity Detection

- **Direct Classification using BERT**



Task 1: Ambiguity Detection

- **Generation-based classification using BART**



Task 1: Ambiguity Detection

Input in addition to AQ	Acc.	Pre.	Rec.	F1
No Answers for AQ	**63.9**	**61.9**	**60.7**	**61.3**
Predicted Answers for AQ	56.5	59.7	24.1	34.3

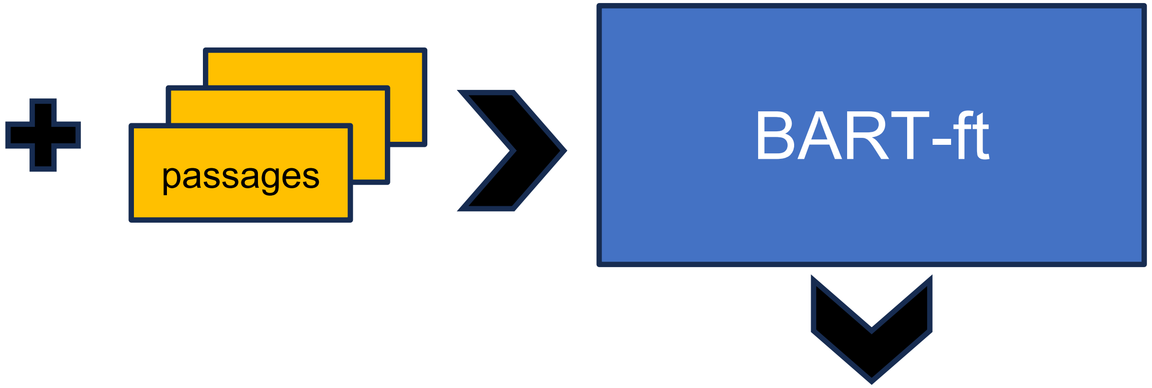
- **Ambiguity Detection: Given a question q, classify whether q is ambiguous or not (binary classification)**

- **Direct Classification** (No Answers for AQ) shows **higher F1** compared to **Generation-based Classification** (Predicted Answers for AQ) because **average answers generated AQ is 1.24**, **resulting in low recall.**

Task 2: Clarification Questions Generation

- **CQ generation by BART**

Who is the president of USA?



Model generated CQ: In which year: 2023, 2019, 2015?

Automatic metrics (BLEU, EM, BERTSCORE …)

Gold CQ: In which period: 2021-present, 2017-2021, 2009-2017?

Task 2: Clarification Questions Generation

Input in addition to AQ and RPs	CQ		Category		Options			
	BLEU-4	BERTSCORE	EM	BLEU-1	Pre.	Rec.	F1	Avg. #
No Answers for AQ	**7.9**	**88.9**	20.2	**47.3**	**37.4**	18.2	24.5	2.0
Predicted Answers for AQ	**7.9**	**88.9**	**22.8**	44.0	36.9	**19.0**	**25.1**	2.0
Ground Truth Answers for AQ	15.4	89.6	25.2	46.9	34.3	34.4	34.3	3.7

- **Clarification Questions Generation: Given AQ and relevant passages, generate a CQ**

- Evaluating generated CQs against gold CQs **using automatic metrics can not capture semantic similarity.**

Task 3: Clarification-based QA

- **Clarification-based QA**

Who is the president of USA? + **In which year: 2023** + passages → QA model

Model generated answer: Joe Biden

Evaluation using Precision, Recall, F1

Gold answer: Joe Biden

Task 3: Clarification-based QA

CQ used to clarify the AQ	NQ-pretrained BART				CQ-finetuned BART			
	Pre.	Rec.	F1	# Ans.	Pre.	Rec.	F1	# Ans.
CQ generated with No Answers for AQ	47.9	25.2	33.0	1.5	54.4	31.1	39.6	1.6
CQ generated with Predicted Answers for AQ	**49.6**	**26.2**	**34.3**	1.5	**55.4**	**32.0**	**40.5**	1.6
CQ generated with Ground Truth Answers for AQ	39.7	37.5	38.6	2.0	47.5	49.5	48.5	2.5
Ground Truth CQ	47.5	39.8	43.3	2.0	58.0	53.8	55.8	2.5

- **Clarification-based QA: Given AQ, relevant passages, and a CQ, generate a unique answer for each option**

- The result shows **insufficient performance** across different settings because the QA model produce **“Same Answer”** for the different questions.

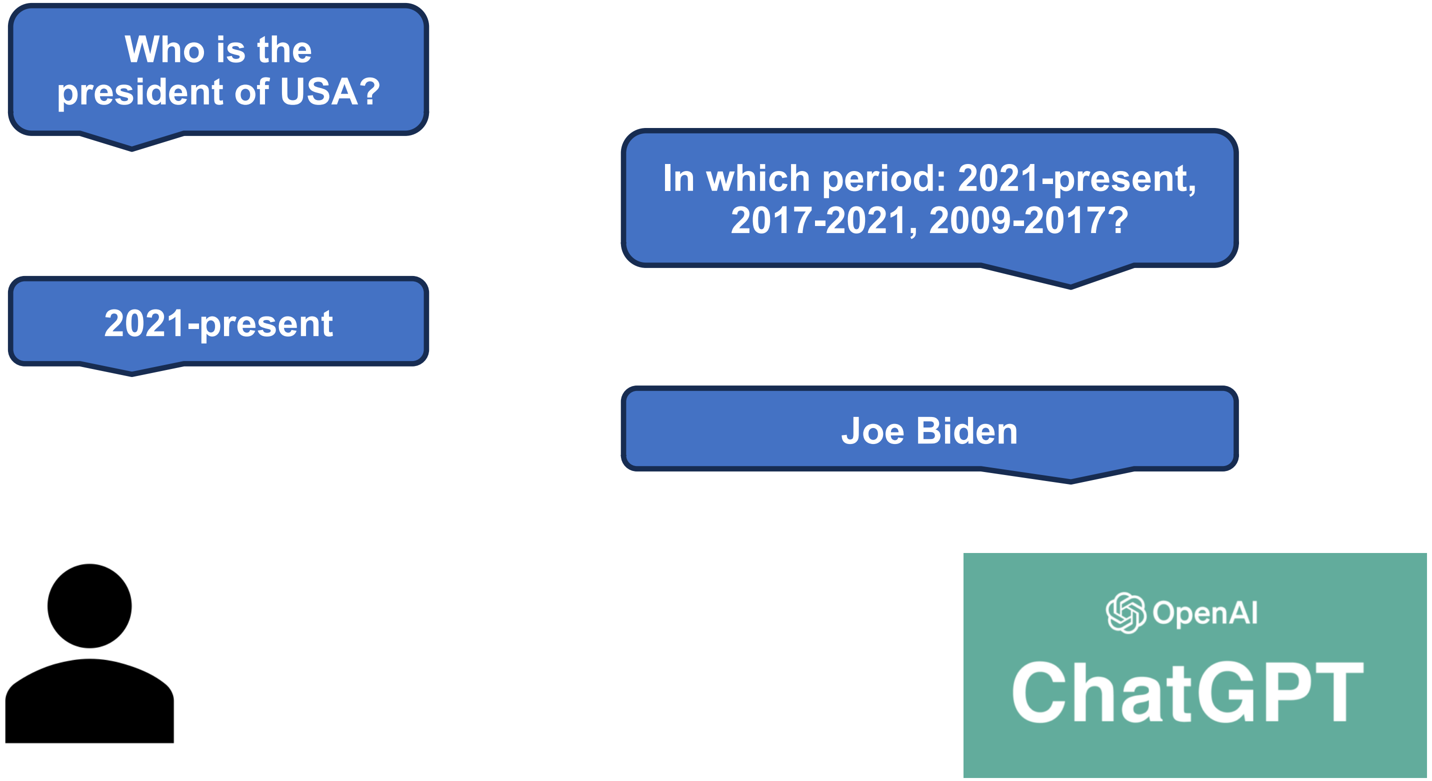
Task 3: Clarification-based QA

Reader Model	Pre.	Rec.	F1	Acc.
CQ finetuned BART	**58.0**	53.8	**55.8**	35.8
InstructGPT	7.4	**60.0**	13.1	**43.2**

- **Clarification-based QA: Given AQ, relevant passages, and a CQ, generate a unique answer for each option**

- Result using LLM (InstructGPT) as a reader model still shows **insufficient performance**

Task 3: Clarification-based QA



- We also test the conversational setting using **ChatGPT**.

Task 3: Clarification-based QA

ChatGPT	Pre.	Rec.	F1	Accuracy
Zero-shot	8.0	**64.5**	14.3	**50.8**
Four-shot	**11.3**	64.0	**19.2**	49.9

- **Clarification-based QA: Given AQ, relevant passages, and a CQ, generate a unique answer for each option**

- Result using ChatGPT with conversational setting, still shows **insufficient performance**

Contributions

Contributions

- We propose to use **CQs as a practical means to handle AQs in Open-Domain QA.**
- We present **CAMBIGNQ**, a dataset to support CQ-based handling of AQs in Open-Domain QA. It was built efficiently by leveraging a well-curated resource, **AMBIGNQ**, as well as the **power of InstructGPT** and **human annotators.**
- We **define a pipeline of tasks** and **appropriate evaluation metrics** for handling AQs in Open-Domain QA.

Thank you