UNIVERSITY OF CAMBRIDGE

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Distractor Generation Using Generative & Discriminative Capabilities of Transformer-based Models

Shiva Taslimipoor, **Luca Benedetto**, Mariano Felice, Paula Buttery

{name.surname}@cl.cam.ac.uk

LREC-COLING 2024

Multiple Choice Questions

- MCQs are widely used to test language learners, mostly because of ease of assessment.
- Made of three components: stem, correct answer, distractors.

Context: [...] When something goes wrong with an instrument, Charles West and Larry Jernigan do the repairs. Both men approach their work with a passion. For them, [...]

Q: What's the job of West and Jernigan at school?

A. teaching music
B. repairing musical instruments
C. teaching students to make minor repairs
D. providing musical instruments for free

Distractor generation

- Very time consuming and challenging.
- Requirements of *good* distractors: **plausibility** and **incorrectness**.
  - Unambiguously wrong.
  - Semantically and syntactically coherent with the correct answer.
  - Not obviously incorrect (too easy).
  - (Possibly) trying to capture common misconceptions and comprehension errors of students.

Automated distractor generation can help

in making better distractors and making the generation scalable

Previous works

- Some previous approaches:
  - Distractors defined as having high similarity to the correct answer (Afzal and Mitkov, 2014).
  - Encoder-decoder architecture (Gao et al. 2019).
  - T5 fine-tuned for DG (Vachev et al., 2022; Rodriguez-Torrealba et al., 2022; Manakul et al., 2023).
  - Learning to rank (Liang et al., 2018).
  - BERT, only on single-word distractors for cloze items (Chiang et al. 2022).
- Frequent issues:
  - They require the correct answer for prediction.
  - Focus on generating one distractor only.

Our two-step approach

1. Generate *plausible* correct and incorrect answers.
2. Control for "incorrectness" of distractors.

Experimental datasets

CLOTH

- Cloze tests
- Text paragraphs, up to 20 gaps for each
- Four single-word options for each gap

RACE

- Reading Comprehension MCQs
- Multiple questions for each passage
- Four answer options for each question
- We work on RACE-DG (Gao et al., 2019)

Baselines

CLOTH

- **BERT** (Chiang et al., 2022)
- Baseline **T5** (Manakul et al., 2023)

RACE

- **HSA**: Hierarchical Static Attention mechanism (Gao et al., 2019)
- **EDGE**: combination of LSTM, self-attention and gated layers (Qiu et al., 2020)
- Baseline **T5** (Manakul et al., 2023)
- **GPT-3.5**: zero-shot and one-shot (Bitew et al., 2023)

Evaluation Metrics

CLOTH

- Precision@1
- F1@3
- NDCG@10

RACE

- BLEU scores
- Similarity based evaluation
- Human evaluation

Results on single-word cloze items

Models	P@1	F1@3	NDCG@10
Baseline T5	9.22	10.29	27.5
BERT	18.50	13.80	37.82
two-step DG	**26.57**	**22.05**	**47.29**

Results on single-word cloze items

- Proposed model outperforms the two baselines

Models	P@1	F1@3	NDCG@10
Baseline T5	9.22	10.29	27.5
BERT	18.50	13.80	37.82
two-step DG	**26.57**	**22.05**	**47.29**

Results on single-word cloze items

- Proposed model outperforms the two baselines
- First generated distractor is relevant for more than 26% of questions (almost 50% improvement).

Models	P@1	F1@3	NDCG@10
Baseline T5	9.22	10.29	27.5
BERT	18.50	13.80	37.82
two-step DG	**26.57**	**22.05**	**47.29**

Results on single-word cloze items

- Proposed model outperforms the two baselines
- First generated distractor is relevant for more than 26% of questions (almost 50% improvement).
- Improvements slightly lower for the other metrics

Models	P@1	F1@3	NDCG@10
Baseline T5	9.22	10.29	27.5
BERT	18.50	13.80	37.82
two-step DG	**26.57**	**22.05**	**47.29**

Reading comprehension MCQs - BLEU scores

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

- On average, our model better than all baselines for longer sequences.

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

- On average, our model better than all baselines for longer sequences.
- EDGE second best in general, particularly good on BLEU1

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

- On average, our model better than all baselines for longer sequences.
- EDGE second best in general, particularly good on BLEU1
- Baseline T5 very good only for the first distractor

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

- On average, our model better than all baselines for longer sequences.
- EDGE second best in general, particularly good on BLEU1
- Baseline T5 very good only for the first distractor

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

- On average, our model better than all baselines for longer sequences.
- EDGE second best in general, particularly good on BLEU1
- Baseline T5 very good only for the first distractor
- One-shot GPT better than zero-shot, but still not good enough.

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - BLEU scores

- On average, our model better than all baselines for longer sequences.
- EDGE second best in general, particularly good on BLEU1
- Baseline T5 very good only for the first distractor
- One-shot GPT better than zero-shot, but still not good enough.

Generated distractor	System	BLEU1	BLEU2	BLEU3	BLEU4
1st distractor	HSA (Gao et al., 2019)	0.28	0.15	0.09	0.06
	EDGE (Qiu et al., 2020)	0.33	0.18	0.11	0.08
	Baseline T5 (Manakul et al., 2023)	**0.34**	**0.23**	**0.16**	**0.12**
	zero-shot GPT (Bitew et al., 2023)	0.25	0.13	0.07	0.04
	one-shot GPT	0.28	0.16	0.10	0.06
	two-step DG	0.31	0.20	0.15	**0.12**
2nd distractor	HSA (Gao et al., 2019)	0.28	0.13	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.1	0.06
	Baseline T5 (Manakul et al., 2023)	0.21	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.14	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.09**
3rd distractor	HSA (Gao et al., 2019)	0.27	0.13	0.07	0.05
	EDGE (Qiu et al., 2020)	**0.31**	0.16	0.09	0.06
	Baseline T5(Manakul et al., 2023)	0.04	0.02	0.02	0.01
	zero-shot GPT (Bitew et al., 2023)	0.22	0.11	0.06	0.04
	one-shot GPT	0.26	0.14	0.08	0.05
	two-step DG	0.27	**0.17**	**0.12**	**0.08**
Average	HSA (Gao et al., 2019)	0.28	0.14	0.08	0.05
	EDGE (Qiu et al., 2020)	**0.32**	0.17	0.10	0.07
	Baseline T5(Manakul et al., 2023)	0.20	0.13	0.09	0.07
	zero-shot GPT (Bitew et al., 2023)	0.23	0.12	0.06	0.04
	one-shot GPT	0.27	0.15	0.09	0.05
	two-step DG	0.29	**0.18**	**0.13**	**0.10**

Reading comprehension MCQs - Similarity metrics

	gd2c	gd2d ↑			gd2gd ↓
		d1	d2	d3	
Gold	46.79	-	-	-	33.80
Baseline T5	49.25	**56.26**	37.90	17.73	58.55
zero-shot GPT	42.21	48.06	46.79	46.29	51.99
one-shot GPT	43.00	50.07	48.81	47.49	49.87
two-step DG	42.28	54.00	**51.84**	**49.39**	**43.56**

Reading comprehension MCQs - Similarity metrics

- Two-step DG generates distractors which are not too similar to the correct option

	gd2c	gd2d ↑			gd2gd ↓
		d1	d2	d3	
Gold	46.79	-	-	-	33.80
Baseline T5	49.25	**56.26**	37.90	17.73	58.55
zero-shot GPT	42.21	48.06	46.79	46.29	51.99
one-shot GPT	43.00	50.07	48.81	47.49	49.87
two-step DG	42.28	54.00	**51.84**	**49.39**	**43.56**

Reading comprehension MCQs - Similarity metrics

- Two-step DG generates distractors which are not too similar to the correct option
- Baseline T5 very good for the first distractor

	gd2c	gd2d ↑			gd2gd ↓
		d1	d2	d3	
Gold	46.79	-	-	-	33.80
Baseline T5	49.25	**56.26**	37.90	17.73	58.55
zero-shot GPT	42.21	48.06	46.79	46.29	51.99
one-shot GPT	43.00	50.07	48.81	47.49	49.87
two-step DG	42.28	54.00	**51.84**	**49.39**	**43.56**

Reading comprehension MCQs - Similarity metrics

- Two-step DG generates distractors which are not too similar to the correct option
- Baseline T5 very good for the first distractor
- Two-step DG for the other distractors

	gd2c	gd2d ↑			gd2gd ↓
		d1	d2	d3	
Gold	46.79	-	-	-	33.80
Baseline T5	49.25	**56.26**	37.90	17.73	58.55
zero-shot GPT	42.21	48.06	46.79	46.29	51.99
one-shot GPT	43.00	50.07	48.81	47.49	49.87
two-step DG	42.28	54.00	**51.84**	**49.39**	**43.56**

Reading comprehension MCQs - Similarity metrics

- Two-step DG generates distractors which are not too similar to the correct option
- Baseline T5 very good for the first distractor
- Two-step DG for the other distractors
- Two-step DG is the best model in generating sets of diverse distractors

	gd2c	gd2d ↑			gd2gd ↓
		d1	d2	d3	
Gold	46.79	-	-	-	33.80
Baseline T5	49.25	**56.26**	37.90	17.73	58.55
zero-shot GPT	42.21	48.06	46.79	46.29	51.99
one-shot GPT	43.00	50.07	48.81	47.49	49.87
two-step DG	42.28	54.00	**51.84**	**49.39**	**43.56**

Reading comprehension MCQs - Similarity metrics

- Two-step DG generates distractors which are not too similar to the correct option
- Baseline T5 very good for the first distractor
- Two-step DG for the other distractors
- Two-step DG is the best model in generating sets of diverse distractors

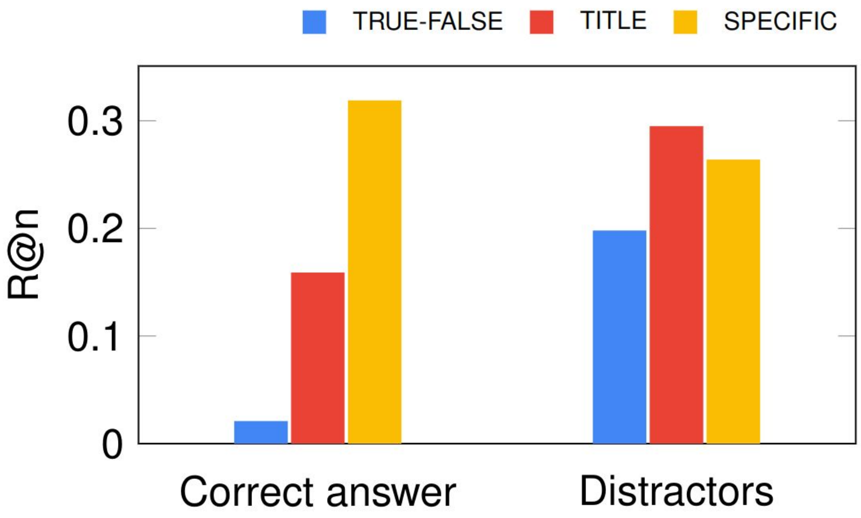
	gd2c	gd2d ↑			gd2gd ↓
		d1	d2	d3	
Gold	46.79	-	-	-	33.80
Baseline T5	49.25	**56.26**	37.90	17.73	58.55
zero-shot GPT	42.21	48.06	46.79	46.29	51.99
one-shot GPT	43.00	50.07	48.81	47.49	49.87
two-step DG	42.28	54.00	**51.84**	**49.39**	**43.56**

Reading comprehension MCQs - Analysis per question type

- **TRUE-FALSE**: ask which option is true or false according to the passage
  - *“Which of the following statements is true according to the article?”*
- **TITLE**: about the best title for the passage
  - “*What is the best title for the passage*?”
- **SPECIFIC**: related to specific information in the passage
  - *“What is Jenny doing in the park?”*

Reading comprehension MCQs - Analysis per question type

- Performance varies greatly when generating the correct answer
- Difference in performance is less significant when generating distractors.
- Performance on TRUE-FALSE questions is worse than the ones for TITLE and SPECIFIC questions.



Conclusions

- Propose a two-step Distractor Generation model which generates both distractors and correct answer options together, and leverages clustering as a way to avoid generating duplicate distractors.
- Outperforms the previous state of the art according to automatic evaluation metrics.
- Future works
    - Improve format consistency with keys
    - Leveraging the abilities of different models on different types of questions

THANK YOU!

Contacts:

luca.benedetto@cl.cam.ac.uk

@bndl22

luca-benedetto

UNIVERSITY OF CAMBRIDGE

Questions?

Examples of generic questions

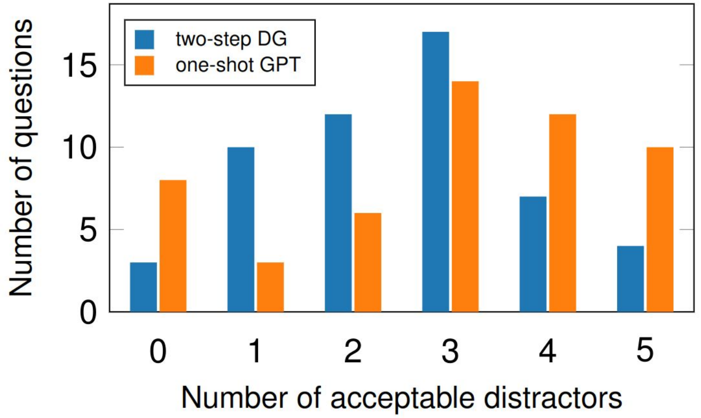
General questions
Which of the following is TRUE according to the passage ?
Which of the following is TRUE ?
Which of the following statements is TRUE ?
From the passage we can infer that
We can infer from the passage that
What can we infer from the passage ?
What might be the title of the passage ?
What is the best title of this passage ?
Which is the best title for the passage ?
What would be the best title for the passage ?
According to the passage , we can know that
What can we learn from the passage ?
What is mainly talked about in the text ?
What is the article about ?
The text is mainly about

Examples of specific questions

Specific questions
In the report , who studies hardest ?
In China , how many students fall asleep in class ?
What do American students do in their free time ?
Why did n't Chief Joseph want to leave the land ?
After some of the young men in White Bird 's group killed eleven white persons, _
Morgan invented volleyball to
What did Morgan think of basketball ?
Specific volleyball rules were formed probably because
What is included in the volleyball rules ?
What did the group of old classmates get together for ?
What cups did the old professor give to his students ?
According to the old professor , why did they have so much stress ?
What can we learn from the old professor 's words ?
Many birds travel in large groups because
Rabbits spend the cold winter by

Reading comprehension MCQs - Human evaluation

- GPT very good on “generic” questions
- Our is similarly good on specific questions.



Model	$\geq$ 3 *acceptable* Gen.	$\geq$ 3 *acceptable* Spec.	*% acceptable* Gen.	*% acceptable* Spec.
one-shot GPT	88.9%	57.1%	79.3%	50.8%
two-step DG	50.0%	54.3%	53.3%	48.6%