Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation

Anonymous ACL submission

Abstract

Distractor generation task focuses on generating incorrect but plausible options for objective questions such as fill-in-the-blank and multiple-choice questions. This task is widely utilized in educational settings across various domains and subjects The effectiveness of these questions in assessments relies on the quality of the distractors, as they challenge examinees to select the correct answer from a set of misleading options. The evolution of artificial intelligence (AI) has transitioned the task from traditional methods to the use of neural networks and pre-trained language models. This shift has established new benchmarks and expanded the use of advanced deep learning methods in generating distractors. This survey explores distractor generation tasks, datasets, methods, and current evaluation metrics for English objective questions, covering both text-based and multi-modal domains. It also evaluates existing AI models and benchmarks and discusses potential future research directions[1].

1 Introduction

Objective questions (Das et al., 2021) such as fill-in-the-blank and multiple-choice questions require an examinee to select one valid answer from a set of invalid options (Kurdi et al., 2020). These types of questions contribute to fair assessment across various domains (e.g., Science (Liang et al., 2018), English (Panda et al., 2022), Math (McNichols et al., 2023), and Medicine (Ha and Yaneva, 2018)). They are also beneficial for educators in assessing large capacity of students with unbiased results (Ch and Saha, 2018). However, creating objective questions manually is a laborious task, as it requires selecting plausible false options, known as *distractors*, that can effectively confuse the examinee.

Distractor Generation (DG) (Dong et al., 2022) is the process of generating an erroneous plausible option in objective questions. In automatic generation, various approaches are utilized, including retrieving-based methods (Ren and Zhu, 2021), learning-based approach (Liang et al., 2018) that ranks options according to a set of features, deep neural networks (Maurya and Desarkar, 2020), and pre-trained language models (Chiang et al., 2022). These methods are applied to distractors in fill-in-the-blank (Wang et al., 2023a) and multiple-choice questions, including question answering (Bitew et al., 2023), reading comprehension (Gao et al., 2019) and multi-modal (Lu et al., 2022a) domains.

Despite the emerging interest in the DG research, there is no literature review in this field, to the best of our knowledge. Existing relevant surveys focus on generating multiple-choice questions (Ch and Saha, 2018; Kurdi et al., 2020; Das et al., 2021; Zhang et al., 2021) without discussing DG tasks. A recent work (Dong et al., 2022) discussed DG as a subtask of natural language generation (NLG) in the text abbreviation tasks, rather than a subtask in objective questions. We aim to fill the gap and conduct the first survey for DG in objective type of questions. To this end, we collected over 100 high-quality papers from top conferences such as ACL, AAAI, IJCAI, ICLR, EMNLP, NAACL, and COLING and journals such as ACM Computing Surveys, ACM Transactions on Information System, IEEE Transactions on Learning Technologies and IEEE/ACM Transactions on Audio, Speech, and Language Processing.

This paper explores English DG and provides a comprehensive understanding of this research area. Figure 1 illustrates the DG survey tree. Our main contributions include: conducting a detailed review of the DG tasks (Sec. 2), related datasets, and methods (Sec. 3); summarizing the evaluation metrics (Sec. 4); discussing the main findings, including the analysis of AI models and benchmarks (Sec. 5); discussing future works (Sec. 6); and providing concluding remarks (Sec. 7).

[1]Resources are available at https://github.com/Distractor-Generation/DG_Survey.

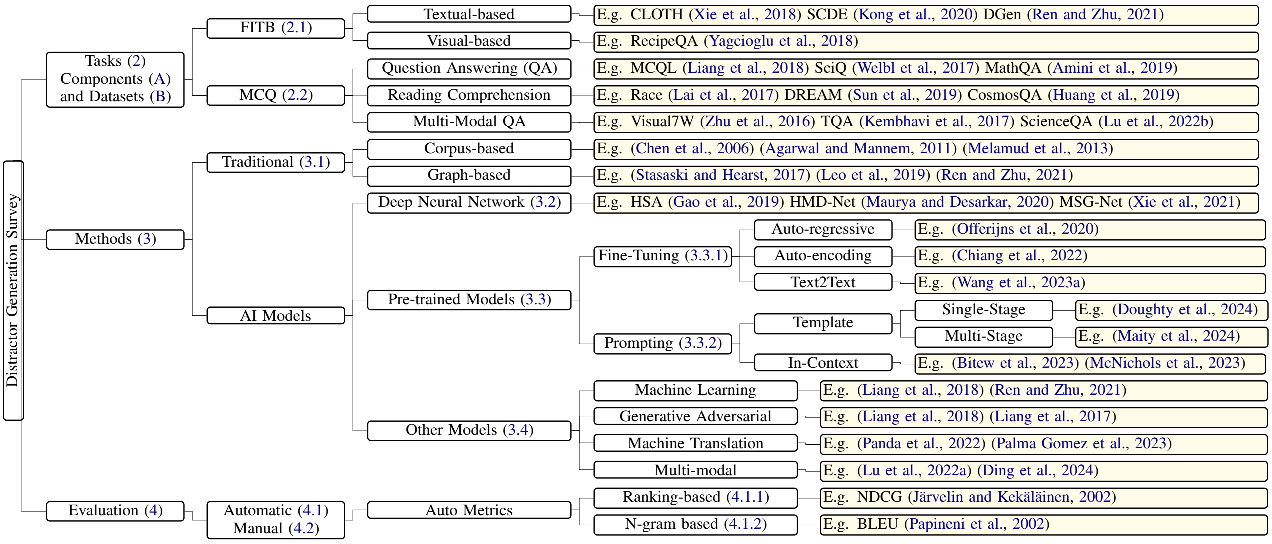

Figure 1: The Survey Tree for DG. The tasks are fill-in-the-blank (FITB) and multiple-choice questions (MCQ).

2 Tasks - Distractor Generation

The tasks are categorized into (i) *fill-in-the-blank* and (ii) *multiple-choice questions*. Table 1 summarizes the available datasets[2] and categorizes each dataset based on DG tasks. A detailed discussion and analysis of the components and datasets are outlined in (Appx A) and (Appx B), respectively.

2.1 Fill-in-the-Blank (FITB)

Cloze queries, also known as fill-in-the-blank, are available in both textual (Xie et al., 2018) and visual formats (Yagcioglu et al., 2018). An example from the DGen dataset, shown in (1), presents a stem sentence with a placeholder and a set of options intended to fill that placeholder. The challenge is to create plausible distractors yet incorrect.

(1) **Stem:** *the organs of respiratory system are __*
Distractors: *a) ovaries, b) intestines, c) kidneys*
Answer: *lungs*

2.2 Multiple-Choice Question (MCQ)

For decades, research communities have shown interest in generating distractors for MCQ (Mitkov et al., 2003; Bitew et al., 2022). MCQ is divided into (i) *question answering*, (ii) *reading comprehension*, and (iii) *multi-modal question answering*.

Question Answering: A standard example of a multiple-choice question-answering task (MC-QA) is shown in (2) from SciQ dataset. The example presents a stem question with a set of options, including one correct answer and several in-context, yet incorrect distractors.

(2) **Stem:** *What eye part allows light to enter?*
Distractors: *a) iris, b) retina, c) eyelid*
Answer: *pupil*

Reading Comprehension: A typical example of a multiple-choice reading comprehension task (MC-RC) is displayed in (3) from the RACE dataset. The challenge involves generating distractors that are relevant to the given stem question and passage, yet distinctly different from the correct answer.

(3) **Passage:** *My name's Mary. This is my family tree ... That boy is my brother. His name is Tony. This is Susan. She is my uncle's daughter.*
Stem: *Tony and Mary are Susan's _____*
Distractors: *a) brothers, b) sisters, c) friends*
Answer: *cousins*

Multi-modal Question Answering: An example of DG in the multi-modal question answering task (MM-QA) (Lu et al., 2022a) is illustrated in Figure 2. The distractors include all the options except for the correct answer, which is indicated by a green checkmark. The main challenge is to generate distractors that are relevant to the given question and image but are not correct as an answer.

Figure 2: Multi-modal Question Answering Task.

[2] We count sub-datasets (CLOTH, RACE, ARC, MCTest).

Dataset	Task	Domain	Source	Creation	Corpus (C)	C.Unit	Availability
CLOTH (Xie et al., 2018)	FITB	English exam	Educational	Expert	7,131	Passage	✔
CLOTH-M (Xie et al., 2018)	FITB	English exam	Educational	Expert	3,031	Passage	✔
CLOTH-H (Xie et al., 2018)	FITB	English exam	Educational	Expert	4,100	Passage	✔
SCDE (Kong et al., 2020)	FITB	English exam	Educational	Expert	5,959	Passage	✉
DGen (Ren and Zhu, 2021)	FITB	Multi-domain	Multi	Auto	2,880	Sentence	✔
CELA (Zhang et al., 2023b)	FITB	English exam	Multi	Auto	150	Passage	✔
SciQ (Welbl et al., 2017)	MC-QA	Science exam	Educational	Crowd	28	Book	✔
AQUA-RAT (Ling et al., 2017)	MC-QA	Math problem	Web	Crowd	97,975	Problem	✔
OpenBookQA (Mihaylov et al., 2018)	MC-QA	Science exam	Educational & WorldTree	Crowd	1,326	WorldTree fact	✔
ARC (Clark et al., 2018)	MC-QA	Science exam	Educational & Web	Expert	14M	Sentence	✔
ARC-Challange (Clark et al., 2018)	MC-QA	Science exam	Educational & Web	Expert	14M	Sentence	✔
ARC-Easy (Clark et al., 2018)	MC-QA	Science exam	Educational & Web	Expert	14M	Sentence	✔
MCQL (Liang et al., 2018)	MC-QA	Science exam	Educational & Web	Crawl	7,116	Query	✔
CommonSenseQA (Talmor et al., 2019)	MC-QA	Narrative	ConceptNet	Crowd	236,208	ConceptNet Triplets	✔
MathQA (Amini et al., 2019)	MC-QA	Math problem	Web	Crowd	37,297	Problem	✔
QASC (Khot et al., 2020)	MC-QA	Science exam	Educational & WorldTree	Crowd	17M	Sentence	✔
MedMCQA(Pal et al., 2022)	MC-QA	Medicine exam	Educational	Expert	2.4K	Topics	✔
Televic (Bitew et al., 2022)	MC-QA	Multi-domain	Educational	Expert	62,858	Query	✔
EduQG (Hadifar et al., 2023)	MC-QA	Education	Educational	Expert	13/283	Book/Chapter	✔
ChildrenBookTest (Hill et al., 2016)	MC-RC	Story	Project Gutenberg	Auto	108	Book	✔
Who Did What (Onishi et al., 2016)	MC-RC	News	Gigaword	Auto	10,507	Book	✉
MCTest-160 (Richardson et al., 2013)	MC-RC	Children story	Fiction	Crowd	160	Story	✔
MCTest-500 (Richardson et al., 2013)	MC-RC	Children story	Fiction	Crowd	500	Story	✔
RACE (Lai et al., 2017)	MC-RC	English exam	Educational	Expert	27,933	Passage	✔
RACE-M (Lai et al., 2017)	MC-RC	English exam	Educational	Expert	7,139	Passage	✔
RACE-H (Lai et al., 2017)	MC-RC	English exam	Educational	Expert	20,784	Passage	✔
RACE-C (Liang et al., 2019)	MC-RC	English exam	Educational	Expert	4,275	Passage	✔
DREAM (Sun et al., 2019)	MC-RC	English exam	Educational	Expert	6,444	Dialogue	✔
CosmosQA (Huang et al., 2019)	MC-RC	Narratives	Blog	Crowd	21,866	Narrative	✔
ReClor (Yu et al., 2020)	MC-RC	Standard exam	Educational	Expert	6,138	Passage	✔
QuAIL (Rogers et al., 2020)	MC-RC	Multi-domain	Multi	Crowd	800	Passage	✔
MovieQA (Tapaswi et al., 2016)	MM-QA	Movie	Movies	Crowd	408	Movie	✉
Visual7W (Zhu et al., 2016)	MM-QA	Visual	Images	Crowd	47,300	Image	✔
TQA (Kembhavi et al., 2017)	MM-QA	Science exam	Educational	Expert	1,076	Lesson	✔
RecipeQA (Yagcioglu et al., 2018)	MM-QA	Cooking	Recipes	Auto	19,779	Recipe	✔
ScienceQA (Lu et al., 2022b)	MM-QA	Science exam	Educational	Expert	21,208	Query	✔

Table 1: Multiple-Choice Datasets. **K** : thousand, **M** : million, ✔: public available, ✉: available upon request.

3 Methods - Distractor Generation

The methods range from traditional to advanced AI approaches, including deep neural networks and pre-trained language models.

3.1 Traditional Methods

Traditional methods propose retrieving word-level distractors similar to an answer in specific domains.

Corpus-based methods rely on corpus features and syntactic rules in selecting distractors. Chen et al. (2006) used a part-of-speech tagger to transform an answer into various grammatical distractors, such as different verb tenses, in grammar cloze tests. Pino and Eskenazi (2009) generated distractors through phonetic and morphological features. Hill and Simha (2016) utilized n-gram corpus to find potential distractors by filtering out all candidates that fit the context in cloze queries. Sakaguchi et al. (2013) extracted distractors as error-correction pairs from a large ESL corpus. Agarwal and Mannem (2011) followed part-of-speech similarity and term frequency to select distractors in biology cloze queries. Zesch and Melamud (2014) explored DG for verb cloze queries using context-sensitive inference rules (Melamud et al., 2013), as it used the rules to filter out semantically similar distractors that are out of the context. Corpus-based features are limited to simple distractors, often lacking plausibility in several domains as they fail to capture the semantic relationships required for contextually appropriate distractors.

Graph-based methods retrieve distractors from hierarchical structures representing concepts and their relationships. WordNet (Miller, 1995) and Probase (Wu et al., 2012) as knowledge-base examples are utilized to generate distractors in MC-QA (Mitkov et al., 2003, 2009) and FITB (Pino et al., 2008). Notably, Ren and Zhu (2021) proposed a framework using knowledge-base and contextual information from the question stem and key answer to construct a small set of semantically related distractors, which employs a probabilistic topic model to determine the relevance of concepts to the key within the given stem. knowledge-base contains static knowledge which may not be appropriate in specialized domains. Thus, an ontology-based method is utilized in distractor retrieving. Stasaski and Hearst (2017) used biology expert-curated concepts to select distractors that share some properties with the correct answer while differing in at least one key relationship to remain plausible but incorrect. Leo et al. (2019) utilized ontology in medical domain distractors. Kumar et al. (2023) utilized both knowledge-base and ontology as part of a generation system for collecting distractors in the technical education domain. Ontology, a static and domain-independent concept, may not cover all necessary concepts for diverse distractors. It is complex, time-consuming, and requires expert knowledge to ensure accuracy and relevance.

3.2 Deep Neural Network Models

Neural networks, including Sequence-to-Sequence (Seq2Seq) (Sutskever et al., 2014) models and attention mechanisms (Bahdanau et al., 2015), showed success in generating distractors at word and sentence levels in MC-RC task. Seq2Seq models map input sequences such as passage, question, or answer to output sequence, a distractor, through conditional log-likelihood. MC-RC task handles long input sequence (e.g., a passage average token in RACE is 352.8) and requires distractors that are (i) semantically relevant to the passage, (ii) coherent with the question, and (iii) non-equivalent to the answer.

Initially, Gao et al. (2019) proposed a hierarchical encoder-decoder (HRED) network (Li et al., 2015) with two attention mechanisms. HRED showed superior performance in handling long input sequences tasks such as head-line generation (Tan et al., 2017) and summarization (Ling and Rush, 2017). HRED encodes long given passages into word-level and sentence-level representations. A hierarchical dynamic attention allows both word-level and sentence-level attention distributions to change at each decoding time step to only focus on important sentences in the passage. A static attention is proposed to learn the distribution of the sentences that are semantically relevant to the question rather than the answer. In decoding, a special question-based initializer is used instead of encoder's last hidden state to generate a distractor that is grammatically consistent with the question.

Several studies followed HRED network with other attention mechanisms. For example, Zhou et al. (2020) utilized co-attention mechanism (Seo et al., 2016) to help the encoder better capture the rich interactions between the passage and question to generate relevant distractors. Shuai et al. (2021) explored static attention with topic-enhanced multi-head co-attention through Latent Dirichlet Allocation (LDA) to calculate the topic-level attention between question and passage sentences. Maurya and Desarkar (2020) implemented the Soft-Sel operation (Tang et al., 2019) combined with a gated mechanism to eliminate answer-revealing sentences. Notably, Shuai et al. (2023) incorporate HRED into a question-distractor joint framework while other works mainly focused on DG task.

To generate multiple n-distractors, beam search with Jaccard distance is mainly utilized in several studies while Maurya and Desarkar (2020) explored multiple decoders. Xie et al. (2021) proposed encoder-decoder multi-selector generation network (MSG-Net) based on mixture content selection (Cho et al., 2019) to generate diverse distractors based on n-sentence key selectors. The selected sentences are transformed into distractors using T5 (Raffel et al., 2020) as a generation layer.

3.3 Pre-trained Models

Pre-trained models, such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017), have revolutionized static word embedding generation. These models are commonly used in DG tasks like FITB (Kumar et al., 2015; Jiang and Lee, 2017; Yoshimi et al., 2023) and MC-QA (Guo et al., 2016) to select similar answer options using word vector cosine similarity. In the MC-RC task, Susanti et al. (2018) utilized word vector cosine similarity to select distractors for English vocabulary meaning.

Pre-trained language models (PLMs) (Min et al., 2023) based on Transformer architecture (Vaswani et al., 2017) include **(i) auto-regressive** models such as GPT-models (Radford et al., 2019; Brown et al., 2020), **(ii) auto-encoding** models such as BERT (Devlin et al., 2019), and **(iii) encoder-decoder** (Text2Text) models such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). PLMs utilize *fine-tuning* and *prompting* methods in DG.

3.3.1 PLMs with Fine-Tuning

PLMs, pre-trained on large amounts of unlabelled data, can be fine-tuned on specific tasks using small labeled datasets. Table 2 presents DG studies where PLMs with fine-tuning have been utilized.

In **auto-regressive** models, (Offerijns et al., 2020) fine-tuned GPT-2 model trained on the RACE dataset to generate three distractors for a given question and context.

In **auto-encoding** models, Chung et al. (2020) proposed BERT model as auto-regressive iterations with multi-tasking and negative answer regulariza-

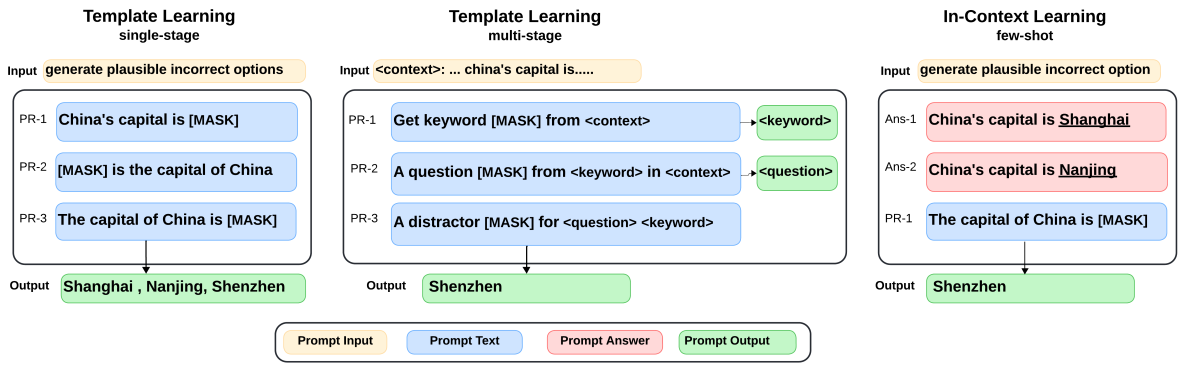

Figure 3: DG via prompting LLM. Figure is adapted from (Liu et al., 2023).

tion to generate distractors in MC-RC task. Chiang et al. (2022) explored several PLMs instead of knowledge-base methods (Ren and Zhu, 2021) for generating distractors in FITB task. The models are trained based on naive fine-tuning and answer-relating fine-tuning. (Bitew et al., 2022) explored a multilingual BERT encoder to create context-aware neural networks in MC-QA. The model ranks distractors based on relevance to the question stem and answer key through contrastive learning.

In **Text2Text** models, Wang et al. (2023a) suggested T5 and BART models for FITB task. To boost model performance, candidate augmentation strategy and multi-tasking training techniques are utilized. Taslimipoor et al. (2024) also proposed using T5 model for DG in MC-QA and MC-RC. The proposed approach utilized a two-step method: initially generating both correct and incorrect answers, and then discriminating between them with a classifier. The generated options are then clustered to remove duplicates and to ensure the diversity of the distractors. T5 has been widely used in DG for MC-QA tasks related to questionnaires (Rodriguez-Torrealba et al., 2022) and personalized exercises (Lelkes et al., 2021; Vachev et al., 2022).

3.3.2 PLMs with Prompting

Prompting (Liu et al., 2023) involves adding text to the input or output to encourage large language models (LLM) to perform specific tasks. Figure 3 illustrates prompting-based learning methods.

Template-based learning uses multiple unanswered prompts at inference time to make predictions and has shown significant capabilities in generating distractors for FITB (Zu et al., 2023) and MC-QA (Doughty et al., 2024) through single-stage prompting. Maity et al. (2024) proposed multi-stage prompting, inspired by the chain of thought method (Wei et al., 2022), to generate distractors for MC-QA based on a given text context.

In-context learning involves providing a few additional answered examples to demonstrate how the LLM should respond to the actual prompt. As shown in Table 3, in-context learning with zero and few-shot examples is also applied in MC-QA. In few-shot learning, examples are selected based on relevant questions retrieved by BERT-based ranking model (Bitew et al., 2022, 2023) and McNichols et al. (2023) used k-nearest neighbor examples for math distractor and feedback generation.

Paper	PLMS	Language	Task
(Yeung et al., 2019)	BERT (2019)	Chinese	FITB
(Chung et al., 2020)	BERT (2019)	English	MC-RC
(Offerijns et al., 2020)	GPT-2 (2019)	English	MC-RC
(Lelkes et al., 2021)	T5 (2020)	English	MC-QA
(Kalpakchi and Boye, 2021)	BERT (2019)	Swedish	MC-RC
(Chiang et al., 2022)	BERT (2019)	English	FITB
(Chiang et al., 2022)	SciBERT (2019)	English	FITB
(Chiang et al., 2022)	RoBERTa (2019)	English	FITB
(Chiang et al., 2022)	BART (2020)	English	FITB
(Vachev et al., 2022)	T5 (2020)	English	MC-QA
(Rodriguez-Torrealba et al., 2022)	T5 (2020)	English	MC-QA
(Foucher et al., 2022)	T5 (2020)	English	MC-QA
(Bitew et al., 2022)	mBERT (2019)	Multi-lingual	MC-QA
(Wang et al., 2023a)	BART (2020)	English	FITB
(Wang et al., 2023a)	T5 (2020)	English	FITB
(Hadifar et al., 2023)	T5 (2020)	English	MC-QA
(De-Fitero-Dominguez et al., 2024)	mT5 (2020)	Spanish	MC-RC
(Taslimipoor et al., 2024)	T5 (2020)	English	FITB
(Taslimipoor et al., 2024)	T5 (2020)	English	MC-RC

Table 2: Fine-tuned PLMs on DG tasks.

3.4 Other Models

Other models proposed retrieving distractors from feature-based learning models for FITB (Ren and Zhu, 2021) and MC-QA (Liang et al., 2018). Sinha et al. (2020) used a hybrid semantically aware neural network, consisting of a convolutional neural network and bidirectional LSTM, to retrieve distractors in an MC-QA task. These models have shown better performance compared to those using generative adversarial networks (Liang et al., 2017).

Paper	LLMs	Method	Prompting	Language	Domain	Task
(Bitew et al., 2022)	ChatGPT	In-Context	zero + few shots	Multi-lingual	Open-Domain	MC-QA
(Zu et al., 2023)	GPT-2	Template	single stage	English	Language proficiency	FITB
(Tran et al., 2023)	GPT-3	Template	single stage	English	Programming	MC-QA
(Tran et al., 2023)	GPT-4	Template	single stage	English	Programming	MC-QA
(McNichols et al., 2023)	Codex	In-Context	zero + few shots	English	Math	MC-QA
(McNichols et al., 2023)	ChatGPT	In-Context	zero + few shots	English	Math	MC-QA
(Doughty et al., 2024)	GPT-4	Template	single stage	English	Programming	MC-QA
(Maity et al., 2024)	GPT-4	Template	multi-stage	Multi-lingual	Open-Domain	MC-QA
(Maity et al., 2024)	Codex	Template	multi-stage	Multi-lingual	Open-Domain	MC-QA

Table 3: Prompting large language models for DG tasks. LLMs such as ChatGPT are selected based on OpenAI models such as (gpt-3.5-turbo) and Codex based on (code-davinci-002) and (text-davinci-003) (Brown et al., 2020).

In domain-specific such as English Language test, round trip machine translation methods (Panda et al., 2022; Palma Gomez et al., 2023) with alignment computation (Jalili Sabet et al., 2020) can generate a variety of distractors. In multi-modal, Lu et al. (2022a) utilized reinforcement learning for textual DG, while Ding et al. (2024) proposed framework, using encoder-decoder vision-and-language model with contrastive learning to jointly generate questions, answers, and distractors.

4 Evaluation Methods

4.1 Automatic Evaluation

The automatic metrics are *ranking-based* (Valcarce et al., 2020) and *n-gram* (Sai et al., 2022) metrics.

4.1.1 Ranking-based Metrics

Ranking-based metrics evaluate the model in retrieving relevant distractors across k-top locations.

Order-unaware metrics, which do not consider the order, include Precision (P@K), Recall (R@K), and F1-score (F1@K). (P@K) calculates the ratio of correctly identified relevant distractors to the total number of options ranked within the top k positions. (R@K) measures the ratio of correctly identified relevant distractors to the total number of relevant distractors in the ground truth, and (F1@K) is the harmonic mean of precision and recall.

Order-aware metrics, which do consider the order, include Mean Reciprocal Rank (MRR@K), Normalized Discounted Cumulative Gain (NDCG@K), and Mean Average Precision (MAP@K). MRR@K focuses on the position of the first relevant item by averaging the reciprocal ranks of this item in the top k distractors across all queries. NDCG@K compares the generated rankings to an ideal order, and MAP@K calculates the mean of average precision scores at k, considering the number and positions of relevant distractors. However, they struggle to identify semantic relatedness, multiple answers, or nonsensical distractors.

4.1.2 N-gram Metrics

N-gram metrics evaluate the word n-gram overlap between the hypothesis (i.e., generated distractors) and references (i.e., ground truth distractors). For example, BLUE (Papineni et al., 2002) is a precision-based metric calculating the ratio of n-grams between the hypothesis and references to the total n-grams in the hypothesis. Self-BLEU (Caccia et al., 2019) measures lexical diversity between hypotheses. ROUGE (Lin, 2004) is a recall-based metric calculating the ratio of n-grams between the hypothesis and references to the total n-grams in the reference. ROUGE-L uses F-score to measure the longest common subsequence between sentence pairs. METEOR (Lavie and Denkowski, 2009) is an F-score metric that applies unigram matches, performing exact word mapping, stemmed word matching, and then synonym and paraphrase matching. Lexical mismatch may fail to identify valid distractors, leading to manual evaluation methods.

4.2 Manual Evaluation

The DG evaluation primarily relies on *plausibility* to ensure distractors are semantically similar to the answer, grammatically correct within the query, and consistently relevant to the context, *reliability* to ensure incorrectness, and *diversity* to reflect the difficulty in identifying the correct answer. Thus, manual methods are utilized in this task.

Comparative method (Gao et al., 2019) selects the distractors based on specific objectives such as **confusion**, assessing the number of times a distractor being chosen as the best option without providing the correct answer, and **non-error** measuring the number of correct answers to a question.

Quantitative method (Maurya and Desarkar, 2020) relies on numerical scales within a specific range to evaluate a given objective. For instance, **reliability** and **plausibility** are the most essential metrics and participants use a 3-point scale for plausibility, and a binary mode for reliability for given generated and ground-truth distractors. Also, **fluency** assesses if a distractor follows proper language grammar, human logic, and common sense, **coherence** evaluates distractor key phrases for relevance to the article and question, **distractibility** measures the likelihood of a candidate being chosen as a distractor, **diversity** measures semantic difference between multiple distractors, and **difference** measures the proportion of distractors and answer with the same semantics.

5 Discussion and Findings

5.1 Analysis of AI Models

Do current models improve the quality of FITB and MC-QA tasks? DG studies primarily focused on plausibility, but the reliability aspect has not been thoroughly studied. Static-based word embeddings like Word2Vec (Jiang and Lee, 2017) as shown in example (1) at Table 4 are prone to generate multiple semantically correct answers, which fail to satisfy reliability. In contrast, dynamic context-based word embeddings like BERT (Devlin et al., 2019) may produce compound names as distractors that are overly technical, which leads to the answer-revealing issue and fails to satisfy diversity. Feature-based learning models (Liang et al., 2018) might predict too easy options. PLMs are still susceptible to generating nonsense distractors such as duplicate correct answers, obviously incorrect options, or previously generated distractors as shown in examples (2), and (3) through fine-tuning FITB task. Wang et al. (2023a) utilized data augmentation to reduce these issues. Few-shot examples (Bitew et al., 2023) reduced nonsense distractor rate in open-domain from 50% to 16%. Thus, the quality of DG in these tasks is still insufficient for reliable and diverse distractors.

(1) **Stem** : The main source of energy in your body is — **Answer**: carbohydrate		
Method	Distractor	Problem
EmbSim (2017)	- glucose	valid answer
BERT (2019)	- glycosaminoglycans	too technical
LR+RF (2018)	- methane	obviously wrong
(2) **Stem**: Rural area do not have school, that is not —— **Answer**: fair		
Method	Distractor	Problem
T5 (2023a)	- fair	similar to answer
BART (2023a)	- unfair	obviously wrong
(3) **Stem**: She let people —— more about Vietnam **Answer**: know		
Method	Distractor	Problem
T5 (2023a)	- think, think , think	previously generated

Table 4: DG quality in FITB and MC-QA tasks.

(1) **Passage**: Nuclear power's danger to health ... etc **Question**: Which of the following statements is true? **Answer**: Nuclear radiation can cause cancer in human beings **Method**: HMD-Net (Maurya and Desarkar, 2020)	
Distractor	Problem
- Radiation is harmless, - Radiation can't hurt all over us, - Radiation can't kill human beings.	lexically differ, but semantically similar.
(2) **Passage**: Most of the time, people wear hats to protect ...etc **Question**: which of the women would look most attractive? **Answer**: A short red-haired woman who wears a purple hat **Method**: BDG (Chung et al., 2020)	
Distractor	Problem
- young woman wears a white hat, - young woman wears a white hat, - short woman with big, round faces.	previously generated and biased options
(3) **Passage**: About a third of all common cancers ...etc **Question**: By writing the passage, the author mainly intends to **Answer**: Advice people to develop healthier lifestyle **Method**: MSG-Net (Xie et al., 2021)	
Distractor	Problem
- teach people how to prevent cancers, - advice people to stop smoking, - protect people from developing cancer.	lack difficulty control

Table 5: DG validity in the MC-RC task.

Are current models satisfied validity in MC-RC task? Despite the use of dynamic and static attentions in MC-RC models for plausibility and reliability, there are still shortcomings. The beam search methods (Gao et al., 2019; Shuai et al., 2023) in Seq2Seq models fail to generate diverse distractors. Also, multi-decoders (Maurya and Desarkar, 2020) as demonstrated in examples (1) in Table 5 used a mixture of decoders in decoding stage to generate divers distractors, but distractors are generated from the same input and have identical semantics which leads to options that are lexically diverse, but they are semantically similar. These generation methods cause an answer-revealing issue. PLMs are still vulnerable to answer copying and biased options (Chung et al., 2020), as shown in example (2). The content selection approach (Xie et al., 2021) in example (3) can generate diverse distractors from different sentences, but further exploration or implicit common sense reasoning is required for difficult controls. Thus, the validity of DG has room for improvement. Quantitative comparisons are detailed for DG tasks in (Appx C).

5.2 Analysis of Benchmarks

Are low-resource datasets explored in DG? Despite the use of English datasets, low-resource datasets remain limited in DG. Pioneering research explored DG in Spanish (De-Fitero-Dominguez et al., 2024), Swedish (Kalpakchi and Boye, 2021), Chinese (Yeung et al., 2019), Japanese (Andersson and Picazo-Sanchez, 2023) and others (Maity et al., 2024) including German, Bengali, and Hindi. Typically, small-scale datasets or translated English datasets are used to create these training data. Notably, there are efforts to build non-English multiple-choice datasets in French (Labrak et al., 2022), Chinese (Sun et al., 2020), Bulgarian (Hardalov et al., 2019), Vietnamese (Van Nguyen et al., 2020) and a multi-lingual (Bitew et al., 2022) datasets. These datasets enable low-resource DG exploration and highlight the need for non-English datasets across various domains and tasks.

Are open-domain datasets emerging in DG?. Specific domains such as Science (e.g., SciQ) or English (e.g., CLOTH) are utilized in DG, but there are limited open-domain datasets (e.g., Televic, EduQG) emerging in the field. For example, Televic, which covers multiple subjects and includes multi-lingual content, contributes significantly to DG by posing new challenges, such as generating nonsensical distractors (Bitew et al., 2022, 2023).

6 Future Work

6.1 Trustworthy Generation

AI advancements in DG are improving, but they still face challenges like hallucination (Ji et al., 2023) issues in PLMs. To control this task generation (Zhang et al., 2023a), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and few-shot examples (Bitew et al., 2023) may be utilized to improve the trustworthiness of DG. Also, pioneering works can train models to distinguish between valid and invalid distractors and manage the difficulty level between candidates.

6.2 Deployment in Education

Distractor quality is crucial in personalized learning (Vachev et al., 2022; Lelkes et al., 2021), but evaluating their effectiveness in education remains a research challenge. AI models explored LLMs ability to generate MCQs that meet course learning objectives in the programming domain (Doughty et al., 2024) and in various formats (Tran et al., 2023). Therefore, instructors must ensure the quality of DG by verifying plausibility, reliability, diversity, alignment with learning objectives, and ethical guidelines.

6.3 Multi-Modal Generation

The novel task (Lu et al., 2022a), textual DG in visual question answering, faces two potential challenges. First, there are potential needs in generating distractors for various multi-modal domains as recent studies (Ding et al., 2024) mainly used Visual7w as a visual question answering dataset. Multi-modal supported content, such as figures (Wang et al., 2021), charts (Kafle et al., 2018), and tables (Lu et al., 2023), are available and used in different domains, including science (Kembhavi et al., 2017) and mathematics (Verschaffel et al., 2020) such as math word problem (Lu et al., 2021b) and geometry problem solving (Chen et al., 2021; Lu et al., 2021a; Chen et al., 2022). Second, research should focus on visual DG , specifically images, and incorporate videos and audios for new insights.

6.4 Quality Metrics

Current automatic metrics (e.g., n-gram) showed significant limitations such as excluding acceptable candidates due to lexical mismatching. Although some metrics can perform synonym n-gram matching (e.g., greedy matching (Rus and Lintean, 2012), embedding average metrics (John et al., 2016), and vector extrema (Forgues et al., 2014)), they cannot determine if semantic similarity will cause reliability issues such as multiple-answer problems. Self-BLEU cannot ensure diversity, as it measures diversity in terms of lexical differences, which does not guarantee the difficulty of the distractors. Thus, few studies (Moon et al., 2022; Raina et al., 2023) proposed systems for the quality of DG even though generalizing quality metrics in DG is still challenging. Also, the assessing for nonsense distractors in open-domain (Bitew et al., 2022) still relies on manual metrics like nonsense distractor rate.

7 Conclusion

Distractor Generation (DG) is critical in assessment and has received significant attention with advanced AI models. This paper surveys DG tasks, including fill-in-the-blank and multiple-choice questions across text and multi-modal domains. It categorizes DG tasks within relevant datasets and discusses the associated methods and evaluation metrics. This paper also provides a detailed discussion of current methods, benchmarks, and potential future research directions.

8 Limitations

The survey focuses on contemporary research in DG using advanced AI methods, but may not cover the entire historical scope and recent advancements that have emerged around the time or after the survey was conducted due to rapid research development. However, our survey is the first to contribute to DG tasks and methods, providing detailed outlines of current datasets and evaluation methods. It also provides a concise overview of the main findings, challenges, and future research works, making it a valuable resource for scholars.

References

Manish Agarwal and Prashanth Mannem. 2011. Automatic gap-fill question generation from text books. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64, Portland, Oregon. Association for Computational Linguistics.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Tim Andersson and Pablo Picazo-Sanchez. 2023. Closing the gap: Automated distractor generation in japanese language testing. *Education Sciences*, 13(12):1203.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations, ICLR 2015*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Distractor generation for multiple-choice questions with predictive prompting and large language models. *arXiv preprint arXiv:2307.16338*.

Semere Kiros Bitew, Amir Hadifar, Lucas Sterckx, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Learning to reuse distractors to support multiple choice question generation in education. *IEEE Transactions on Learning Technologies*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Burton, Akshay Java, Ian Soboroff, et al. 2009. The icwsm 2009 spinn3r dataset. In *Proceedings of Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2019. Language gans falling short. In *Proceedings of International Conference on Learning Representations, ICLR 2019*.

Dhawaleswar Rao Ch and Sujan Kumar Saha. 2018. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25.

Chia-Yin Chen, Hsien-Chin Liou, and Jason S. Chang. 2006. FAST – an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 1–4, Sydney, Australia. Association for Computational Linguistics.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. 2022. CDGP: Automatic cloze distractor generation based on pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5835–5840, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. In *Proceedings of the 2019 Conference* on *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3121–3131, Hong Kong, China. Association for Computational Linguistics.

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4390–4400, Online. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Bidyut Das, Mukta Majumder, Santanu Phadikar, and Arif Ahmed Sekh. 2021. Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning*, 16(1):1–15.

David De-Fitero-Dominguez, Eva Garcia-Lopez, Antonio Garcia-Cabot, Jesus-Angel Del-Hoyo-Gabaldon, and Antonio Moreno-Cediel. 2024. Distractor generation through text-to-text transformer models. *IEEE Access*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. 2024. Can we learn question, answer, and distractors all from an image? a new task for multiple-choice visual question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2852–2863, Torino, Italy. ELRA and ICCL.

Chenhe Dong, Yinghui Li, Haifan Gong, et al. 2022. A survey of natural language generation. *ACM Computing Survey*.

Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, et al. 2024. A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference*, pages 114–123.

Daria Dzendzik, Jennifer Foster, and Carl Vogel. 2021. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Sébastien Foucher, Damian Pascual, Oliver Richter, and Roger Wattenhofer. 2022. Word2course: creating interactive courses from as little as a keyword. In *Proceedings of the14th International Conference on Copmputer Support Education*, pages 105–115. SCITEPRESS.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6423–6430.

Bilal Ghanem and Alona Fyshe. 2023. Disto: Evaluating textual distractors for multi-choice questions using negative sampling based approach. *arXiv preprint arXiv:2304.04881*.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Proceedings of Third international conference on weblogs and social media, data challenge workshop, San Jose, CA*, volume 46, pages 16–23.

Qi Guo, Chinmay Kulkarni, Aniket Kittur, Jeffrey P Bigham, and Emma Brunskill. 2016. Questimator: generating knowledge assessments for arbitrary topics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3726–3732.

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398, New Orleans, Louisiana. Association for Computational Linguistics.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11:20885–20896.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019. Beyond English-only reading comprehension: Experiments in zero-shot multilingual transfer for Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 447–459, Varna, Bulgaria. INCOMA Ltd.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of International Conference on Learning Representations, ICLR 2016*.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30, San Diego, CA. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Shu Jiang and John Lee. 2017. Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Wieting John, Gimpel Kevin, Livescu Karen, LeCun Yann, et al. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016*, volume 2016.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Dmytro Kalpakchi and Johan Boye. 2021. BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 387–403, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Xiang Kong, Varun Gangal, and Eduard Hovy. 2020. SCDE: Sentence cloze dataset with high quality distractors from examinations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5668–5683, Online. Association for Computational Linguistics.

Archana Praveen Kumar, Ashalatha Nayak, Manjula Shenoy, Shashank Goyal, et al. 2023. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225:120022.

Girish Kumar, Rafael Banchs, and Luis Fernando D'Haro. 2015. RevUP: Automatic gap-fill question generation from educational texts. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161, Denver, Colorado. Association for Computational Linguistics.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115.

Adam D Lelkes, Vinh Q Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021*, pages 2501–2511.

Jared Leo, Ghader Kurdi, Nicolas Matentzoglu, Bijan Parsia, Ulrike Sattler, Sophie Forge, Gina Donato, and Will Dowling. 2019. Ontology-based generation of medical, multi-term mcqs. *International Journal of Artificial Intelligence in Education*, 29:145–188.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China. Association for Computational Linguistics.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. 2018. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.

Chen Liang, Xiao Yang, Drew Wham, Bart Pursel, Rebecca Passonneaur, and C Lee Giles. 2017. Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In *Proceedings of the Knowledge Capture Conference*, pages 1–4.

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of Asian Conference on Machine Learning*, pages 742–757. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014*, 8693:740–755.

Jeffrey Ling and Alexander Rush. 2017. Coarse-to-fine attention models for document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 33–42, Copenhagen, Denmark. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiaying Lu, Xin Ye, Yi Ren, and Yezhou Yang. 2022a. Good, better, best: Textual distractors generation for multiple-choice visual question answering via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4921–4930.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *Proceedings of International Conference on Learning Representations ICLR 2023*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. 2024. A novel multi-stage prompting approach for language agnostic mcq generation using gpt. In *Proceedings of European Conference on Information Retrieval*, pages 268–277. Springer.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1115–1124.

Hunter McNichols, Wanyong Feng, Jaewook Lee, Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2023. Exploring automated distractor and feedback generation for math multiple-choice questions via in-context learning. *arXiv preprint arXiv:2308.03234*.

Oren Melamud, Jonathan Berant, Ido Dagan, Jacob Goldberger, and Idan Szpektor. 2013. A two level model for context sensitive inference rules. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1331–1340, Sofia, Bulgaria. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Ruslan Mitkov, Le An Ha, Andrea Varga, and Luz Rello. 2009. Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 49–56, Athens, Greece. Association for Computational Linguistics.

Ruslan Mitkov et al. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, pages 17–22.

Hyeongdon Moon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, Juneyoung Park, Jamin Shin, Minsam Kim, and Seungtaek Choi. 2022. Evaluating the knowledge dependency of questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10512–10526, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. 2020. Better distractions: Transformer-based distractor generation and multiple choice question filtering. *arXiv preprint arXiv:2010.09598*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Frank Palma Gomez, Subhadarshi Panda, Michael Flor, and Alla Rozovskaya. 2023. Using neural machine translation for generating diverse challenging exercises for language learner. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6115–6129, Toronto, Canada. Association for Computational Linguistics.

Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 2022. Automatic generation of distractors for fill-in-the-blank exercises with round-trip neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Juan Pino and Maxine Eskenazi. 2009. Semi-automatic generation of cloze question distractors effect of students' l1. In *International Workshop on Speech and Language Technology in Education*.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*, pages 22–32.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic distractor generation for multiple choice questions in standard tests. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2096–2106, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vatsal Raina, Adian Liusie, and Mark Gales. 2023. Assessing distractors in multiple-choice tests. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 12–22, Bali, Indonesia. Association for Computational Linguistics.

Siyu Ren and Kenny Q Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4339–4347.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Ricardo Rodriguez-Torrealba, Eva Garcia-Lopez, and Antonio Garcia-Cabot. 2022. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8722–8731.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys*, 55(2):1–39.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative approach to fill-in-the-blank quiz generation for language learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–242, Sofia, Bulgaria. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *Proceedings of International Conference on Learning Representations, ICLR 2016*.

Ruhi Sharma Mittal, Seema Nagar, Mourvi Sharma, Utkarsh Dwivedi, Prasenjit Dey, and Ravi Kokku. 2018. Using a common sense knowledge base to auto generate multi-dimensional vocabulary assessments. *International Educational Data Mining Society*.

Pengju Shuai, Li Li, Sishun Liu, and Jun Shen. 2023. Qdg: A unified model for automatic question-distractor pairs generation. *Applied Intelligence*, 53(7):8275–8285.

Pengju Shuai, Zixi Wei, Sishun Liu, Xiaofei Xu, and Li Li. 2021. Topic enhanced multi-head co-attention: Generating distractors for reading comprehension. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Damien Sileo, Kanimozhi Uma, and Marie-Francine Moens. 2024. Generating multiple-choice questions for medical question answering with distractors and cue-masking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7647–7653, Torino, Italy. ELRA and ICCL.

Manjira Sinha, Tirthankar Dasgupta, and Jatin Mandav. 2020. Ranking multiple choice question distractors using semantically informed neural networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3329–3332.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Katherine Stasaski and Marti A. Hearst. 2017. Multiple choice question generation utilizing an ontology. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312, Copenhagen, Denmark. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic distractor generation for multiple-choice english vocabulary questions. *Research and practice in technology enhanced learning*, 13:1–16.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4109–4115.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7088–7095.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Shiva Taslimipoor, Luca Benedetto, Mariano Felice, and Paula Buttery. 2024. Distractor generation using generative and discriminative capabilities of transformer-based models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5052–5063, Torino, Italia. ELRA and ICCL.

Andrew Tran, Kenneth Angelikas, Egi Rama, Chiku Okechukwu, David H Smith, and Stephen MacNeil. 2023. Generating multiple choice questions for computing courses using large language models. In *Proceedings of Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.

Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2022. Leaf: Multiple-choice question generation. In *Proceedings of European Conference on Information Retrieval*, pages 321–328. Springer.

Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-n recommendation. *Information Retrieval Journal*, 23:411–448.

Kiet Van Nguyen, Khiem Vinh Tran, Son T Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lieven Verschaffel, Stanislaw Schukajlow, Jon Star, and Wim Van Dooren. 2020. Word problems in mathematics education: A survey. *ZDM: The International Journal on Mathematics Education*, 52(1):1–16.

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023a. Distractor generation based on Text2Text language models with pseudo Kullback-Leibler divergence regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491, Toronto, Canada. Association for Computational Linguistics.

Jiayun Wang, Jun Bai, Wenge Rong, Yuanxin Ouyang, and Zhang Xiong. 2023b. Weak positive sampling and soft smooth labeling for distractor generation data augmentation. In *Proceedings of International Conference on Intelligent Computing*, pages 756–767. Springer.

Jiayun Wang, Wenge Rong, Jun Bai, Zhiwei Sun, Yuanxin Ouyang, and Zhang Xiong. 2023c. Multi-source soft labeling and hard negative sampling for retrieval distractor ranking. *IEEE Transactions on Learning Technologies*.

Yingyao Wang, Junwei Bao, Chaoqun Duan, Youzheng Wu, Xiaodong He, Conghui Zhu, and Tiejun Zhao. 2023d. An efficient confusing choices decoupling framework for multi-choice tasks over texts. *Neural Computing and Applications*, pages 1–13.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. 2021. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 Competition and Demonstration Track*, pages 191–205. PMLR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 481–492.

Jiayuan Xie, Ningxin Peng, Yi Cai, Tao Wang, and Qingbao Huang. 2021. Diverse distractor generation for constructing high-quality multiple choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:280–291.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Chak Yan Yeung, John Lee, and Benjamin Tsou. 2019. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164, Sydney, Australia. Australasian Language Technology Association.

Nana Yoshimi, Tomoyuki Kajiwara, Satoru Uchida, Yuki Arase, and Takashi Ninomiya. 2023. Distractor generation for fill-in-the-blank exercises by question type. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 276–281, Toronto, Canada. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *Proceedings of International Conference on Learning Representations, ICLR 2020*.

Torsten Zesch and Oren Melamud. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Baltimore, Maryland. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

Zizheng Zhang, Masato Mita, and Mamoru Komachi. 2023b. Cloze quality estimation for language assessment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 540–550, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2020. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9725–9732.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Jiyun Zu, Ikkyu Choi, and Jiangang Hao. 2023. Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychological Test and Assessment Modeling*, 65(1):55–75.

A Multiple Choice Components

The fundamental components of a multiple-choice data item consist of (i) a *stem*, the query or question, (ii) an *answer*, the only true option, and (iii) a set of *distractors*, the set of false options. A *supported content* can be a given text, an image, or a video.

A.1 Stem

A stem can be formed as a complete declarative sentence, a declarative sentence or passage with placeholders, a factoid query such as a deep level (why? how?) or shallow level (who? where?) in Bloom's taxonomy, or other non-factoid queries. It can also be formed as an image or a video in a multi-modal domain.

Fill-in-the-Blank (FITB): selecting an appropriate word, sentence, or an image to complete a given content or a query is known as cloze or FITB. In textual data, CLOTH (Xie et al., 2018) in example (4) describes stem passage, and DGen (Ren and Zhu, 2021) in (5) indicates stem sentence while RecipeQA (Yagcioglu et al., 2018) in Figure 4 outlines a visual stem.

> (4) **Stem:** *Nancy had just got a job as a secretary in a company. Monday was the first day she went to work, so she was very – 1 – and arrived early. She – 2 – the door open and found nobody ...*
> **Distractors -1-:** *a) depressed, b) encouraged, c) surprised*
> **Distractors -2-:** *a) turned, b) knocked, c) forced*
> **Answer -1- :** *excited*
> **Answer -2- :** *pushed*

> (5) **Stem:** *the organs of respiratory system are _*
> **Distractors:** *a) ovaries, b) intestines, c) kidneys*
> **Answer:** *lungs*

Multiple-Choice Question (MCQ): forming a question as a Wh-Q or declarative sentence is common in the MC-QA task. SciQ (Welbl et al., 2017) in (6) and MCQL (Liang et al., 2018) in (7) illustrate textual factoid and declarative sentence stems, respectively.

> (6) **Passage:** *All radioactive decay is dangerous to living things, but alpha decay is the least dangerous.*
> **Stem:** *What is the least dangerous radioactive decay?*
> **Distractors:** *a) zeta decay, b) beta decay, c) gamma decay*
> **Answer:** *alpha decay*

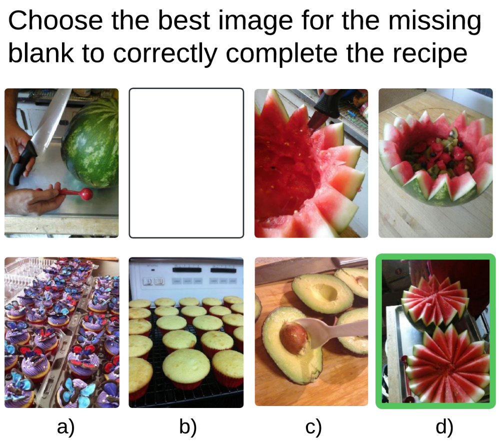

Figure 4: Visual Cloze.

> (7) **Stem:** *During dark reactions, energy is stored in molecules of*
> **Distractors:** *a) carbon, b) oxygen, c) hydrogen*
> **Answer:** *sugar*

A.2 Answer

An answer, also known as the correct option, must be unique for each query. It can be formed as a textual short phrase or a sentence. It can also be extractive from a given passage or free-form generated from a supported passage or prior knowledge. It can also be an image as indicated in Figure 4.

Short or Long Phrase: MCQL in (7) describes word level answer, while RACE (Lai et al., 2017) in (8) describes long sentence answer.

> (8) **Passage:** *Homework can put you in a bad mood ... Researchers from the University of Plymouth in England doubted whether mood might affect the way kids learn ...*
> **Stem:** *Researchers did experiments on kids in order to find out ___ .*
> **Distractors:** *a) how they really feel when they are learning, b) what methods are easy for kids to learn, c) the relationship between sadness and happiness*
> **Answer:** *whether mood affects their learning ability*

Extractive or Free-Form: SciQ in (6) describes extractive answer type as the answer is span on the supported content, while MCQL in (7) is free form.

A.3 Option

All options, also known as distractors or false candidates, must be incorrect candidates to satisfy objectivity. Similar to the answer, options may be formed as words or sentences, mostly separated with each query but SCDE (Kong et al., 2020) introduced shared options across all queries. Figure 4 shows visual options where (d) is the correct answer and others are image distractors.

Separated or Shared: CLOTH in example (4) describes separated options, while SCDE in example (9) shows shared options.

(9) **Stem:** *– 1 – Now it becomes popular and people are dyeing their hair to make it different. Dyeing hair ... Since the base of hair is the scalp, you may have an allergic reaction. – 2 – You can follow them even when you are applying dye to your hair at home. – 3 – ...*
Shared Distractors: *(A) Colorful hair speaks more about beauty, (B) While dyeing your hair it is important to take some safety measures, (C) Don't forget to treat grandparents with respect because they're an essential part of your family, (D) It is better to apply hair dye for a few minutes...*
Answers: *(1-A) (2-B) (3-D)...*

A.4 Supported Content

Supported content can take either a textual form (e.g., sentence, passage, or any form of text) or a visual form (e.g., image or video). Textual-supported content such as passage in the reading comprehension task is essential for assessing the examinee in real assessment. However, supported text content such as SciQ is not primarily provided for reading comprehension tasks, and AQUA-RAT (Ling et al., 2017) provides rationales (i.e., mathematical equation formats) to create mathematical multiple-choice datasets. Table 1 presents the classification of collected datasets in DG tasks.

Textual Form: OpenBookQA (Mihaylov et al., 2018) in (10) describes supported sentence text while RACE (Lai et al., 2017) in (8) describes passage content.

(10) **Sentence:** *the sun is the source of energy for physical cycles on Earth*
Stem: *The sun is responsible for*
Distractors: *a) puppies learning new tricks, b) children growing up and getting old, c) flowers wilting in a vase*
Answer: *plants sprouting, blooming and wilting*

Visual Form: Visual7W in Figure 2 shows image as supported content and MovieQA (Tapaswi et al., 2016) describes movie as supported content.

B Multiple- Choice Datasets

We collected multiple-choice datasets, as shown in Table 1 for DG tasks. We also summarized dataset properties, including related domain, source of data, generation method, corpus size, and unit. Table 6 presents an analysis of multiple-choice components, including average token, vocabulary size, and most frequent type of query.

B.1 Dataset Analysis

We utilized dataset analysis as proposed by Dzendzik et al. (2021) to process our heuristic rules and statistics. Using spaCy[3] tokenizer we determined the average token length and vocabulary size of queries, passages, and options. We determine the most common query type for each dataset, using our proposed heuristic rules[4].

B.1.1 Data Domains

In our collection, 10 of 36 datasets are from English exam sources and 9 from Science exam sources. ReClor is for standardized tests and 4 datasets (i.e., DGen, EduQG, QuAIL, Televic) are for multi-domain fields. One dataset from the medicine domain and 2 datasets focus on math word problems. Three datasets are designed for children stories, two datasets for narratives, and one dataset for news. Three multi-modal datasets are domain-specific such as movie, visual answering, and cooking.

B.1.2 Data Creation

30 out of 36 datasets are created by humans. 18 of them are created by experts and 12 are created by crowd workers. Some datasets are web-crawled such as MCQL and others (i.e., CBT, WDW, RecipeQA, DGen, CELA) are auto-generated.

B.1.3 Data Corpus

The corpuses of 31 datasets are text-based and 5 are multi-modal. 15 out of 36 corpuses are passages, also known as story, narratives, and dialogue. 5 datasets are based on sentence units, 2 datasets have math word problems, and 3 datasets are based on queries. 5 datasets corpuses are books, chapters, or medical topics, and 2 datasets are based on WorldTree facts. One dataset is based on the CONCEPTNET triplet (i.e., knowledge graph with commonsense relationships).

[3] https://spacy.io/.
[4] https://github.com/Distractor-Generation/DG_Survey.

Dataset	Supported Content	Most Query Type	#Passage (P)	#Query (Q)	#Option (O)	P_{avg}	Q_{avg}	O_{avg}	P_{vcb}	Q_{vcb}	O_{vcb}
CLOTH	✗	Passage-Blank	7,131	99,433	4	329.8	✗	1	22,360	✗	7,455
CLOTH-M	✗	Passage-Blank	3,031	28,527	4	246.3	✗	1	9,478	✗	3,330
CLOTH-H	✗	Passage-Blank	4,100	70,906	4	391.5	✗	1	19,428	✗	6,922
SCDE	✗	Passage-Blank	5,959	29,731	7	248.6	✗	13.3	21,410	✗	12,693
DGen	✗	Sentence-Blank	✗	2,880	4	✗	19.5	1	✗	4,527	3,630
CELA	✗	Passage-Blank	150	3,000	4	408.5	✗	1.3	3,500	✗	3,716
SciQ	Text	Question	12,252	13,679	4	78	14.5	1.5	20,409	7,615	9,499
AQUA-RAT	Text	Question	97,975	97,975	5	52.7	37.2	1.6	127,404	31,406	76,115
OpenBookQA	Text	Sentence	1,326	5,957	4	9.4	11.5	2.9	1,416	4,295	6,989
ARC	✗	Question	✗	7,787	4	✗	22.5	4.6	✗	6,079	6,164
ARC-Challange	✗	Question	✗	2590	4	✗	24.7	5.5	✗	4,057	4,245
ARC-Easy	✗	Question	✗	5197	4	✗	21.4	4.1	✗	4,998	5,021
MCQL	✗	Sentence	✗	7,116	4	✗	9.4	1.2	✗	5,703	7,108
CommonSenseQA	✗	Question	✗	12,102	5	✗	15.1	1.5	✗	6,844	6,921
MathQA	Text	Question	37,297	37,297	5	63.3	38.2	1.7	16,324	10,629	11,573
QASC	✗	Question	✗	9,980	8	✗	9.1	1.7	✗	3,886	6,407
MedMCQA	Text	Sentence	16,3075	193,155	4	92.7	14.3	2.8	370,658	53,010	65,773
Televic	✗	*	✗	62,858	>2	✗	*	*	✗	*	*
EduQG	Text	Multi-Form	3,397	3,397	4	209.3	16.3	4.2	21,077	5,311	8,632
ChildrenBookTest	Text	Sentence-Blank	687,343	687,343	10	474.2	31.6	1	34,611	32,912	23,253
Who Did What	Text	Sentence-Blank	*	205,978	2..5	*	31.4	2.1	*	70,198	82,397
MCTest-160	Text	Question	160	640	4	241.8	9.2	3.7	1,991	802	1,481
MCTest-500	Text	Question	500	2,000	4	251.6	8.9	3.8	3,079	1,436	23,34
RACE	Text	Sentence-Blank	27,933	97,687	4	352.8	12.3	6.7	88,851	20,179	32,899
RACE-M	Text	Sentence-Blank	7,139	28,293	4	236	11.1	5	21,566	6,929	11,379
RACE-H	Text	Sentence-Blank	20,784	69,394	4	361.9	12.4	6.9	81,887	18,318	29,491
RACE-C	Text	Sentence-Blank	4,275	14,122	4	424.1	13.8	7.4	34,165	10,196	15,144
DREAM	Text	Question	6,444	10,197	3	86.4	8.8	5.3	8,449	2,791	5,864
CosmosQA	Text	Question	21,866	35,588	4	70.4	10.6	8.1	36,970	10,685	18,173
ReClor	Text	Question	6,138	6,138	4	75.1	17	20.8	15,095	3,370	13,592
QuAIL	Text	Question	800	12966	4	395.4	9.7	4.4	13,750	6,341	9,955
MovieQA	Text + Video	Question	*	14,944	5	*	10.7	5.6	*	7,440	15,242
Visual7W	Image	Question	✗	327,939	4	✗	8	2.9	✗	12,168	15,430
TQA	Text + Image	Question	1,076	26,260	2..7	241.1	10.5	2.3	8,304	7,204	9,265
RecipeQA	Text + Image	Sentence-Blank	19,779	36,786	4	575.1	10.8	5.7	78,089	5,587	71,369
ScienceQA	Text + Image	Question	10,220	21,208	>2	41.3	14.2	4.9	6,233	7,373	7,638

Table 6: Dataset analysis of multiple-choice components. ✗: not available, * : available upon request

B.1.4 Data Sources

Out of 36 datasets, 22 are from educational materials and 14 are from blogs, stories, movies, images, or recipe sources.

Educational Resources: CLOTH, SCDE, RACE, RACE-C, DREAM are collected from educational public websites in China. SciQ is extracted from 28 textbooks. TQA and ScienceQA are collected from CK-12 foundation website and school science curricula, respectively. MCQL and AQUA-RAT are Web-crawled. OpenBookQA is derived from WorldTree corpus (Jansen et al., 2018). QASC has 17 million sentences from WorldTree and CK-12. ReClor is generated from open websites and books. EduQG, Televic, and MedMCQA are collected from the Openstax website, Televic education platform, and medical exam sources, respectively.

Multi-Sources: QuAIL is collected from fiction, news, blogs, and user stories. DGen contents are from SciQ, MCQL, and other websites. CELA is constructed from CLOTH dataset and four auto-generated techniques (i.e., randomized, one feature -part of speech POS (Hill et al., 2016), several features - POS, word frequency, spelling similarity (Jiang et al., 2020), and neural round trip translation (Panda et al., 2022)).

Other Sources: CBT is built based on Project Gutenberg books, MCTest is crowd sourced, and CommonSenseQA used CONCEPTNET (Speer et al., 2017). CosmosQA uses personal narratives (Gordon and Swanson, 2009) from the Spinn3r Blog Dataset (Burton et al., 2009) and crowd-sourcing to promote commonsense reasoning (Sap et al., 2019). MovieQA, Visual7W, and RecipeQA are built utilizing 408 movies, COCO images (Lin et al., 2014), and cooking websites, respectively.

B.1.5 Data Components

The only dataset introduced as multi-format by labeling and forming a query as cloze and normal is EduQG. Therefore, we used heuristic rules to find the most common query type (i.e., blank, sentence, or question). The average token length and vocabulary size of passages, queries, and options are presented in Table 6. We outline the following:

Supported Content: all datasets contain text-supported content except DGen, ARC, CommonSenseQA, MCQL, QASC, and Televic. In multi-modal, some datasets such as RecipeQA and TQA contain text and images. Other datasets such as MovieQA contain movies and (Visual7W, ScienceQA) contain images.

Query Size: CLOTH has the largest number of questions among the FITB datasets. In MCQ datasets, the largest number of science questions found in SciQ (14K) and in math dataset is AQUA-RAT (98K). Televic contains (63K) questions, covering open-domain multi-lingual dataset[5]. Only 198 questions (Q_{avg}14.9, O_{avg} 1.9 average token) are provided in the GitHub sample. The most usable dataset in the comprehension task is RACE (98K). Visual7W (327.9K) presents the largest number of questions in multi-model.

Number of Options: most datasets have 4 to 5 separated options, but the SCDE average is 7 shared options. QASC contains 8 choices. Televic and ScienceQA start with 2 choices. CBT has 10, DREAM contains 3, and TQA is ranged between 2 to 7.

Component Average Length: queries range from 8.8 to 19.5, and passages from 9.4 to 408 tokens. Word-to-phrase token options have 1 to 4, while sentence-long options have more than 4 tokens. ReClor has the longest option tokens (20.8).

Component Vocabulary Size: The vocabulary for passages ranges from 1.4K to 371K based on the number of unique lowercase token lemmas. The vocabulary for the queries spans from 802 to 70.2K, and the options span from 1.5K to 82.4K.

B.1.6 Data Usability and Availability

Table 1 showed the availability of datasets in distractor generation tasks. For example, CLOTH, DGen, SciQ, and MCQL are benchmark datasets in FITB and MC-QA tasks. Televic and EduQG are introduced specifically for distractor generation tasks. RACE is a benchmark dataset in reading comprehension while two other datasets such as CosmosQA and DREAM are utilized in recent studies. Visual7W is the only multi-modal dataset used for textual distractor generation. Other datasets such as MedMCQA, MCTest, CBT, QuAIL and ReClor are utilized in the evaluation stage (Sharma Mittal et al., 2018; Wang et al., 2023b,c,d; Ghanem and Fyshe, 2023; Sileo et al., 2024) for DG tasks.

The majority of datasets are public except upon request datasets (e.g., SCDE, MovieQA) and upon payment of a license fee to access part of the dataset (e.g., WDW) or the whole dataset (e.g., Televic).

[5] 50% Dutch then French and English comes next.

C Quantitative Results

The summary of quantitative results in DG tasks is detailed in the following sections.

C.1 Distractors in FITB and MC-QA

Table 7 summarizes the state-of-the-art (SOTA) results in DG for both FITB and MC-QA tasks, focusing on word-level distractors. The most commonly used metric, precision P@1, yielded the following observations: (i) retrieval-based methods utilizing feature-based learning outperformed neural networks based on adversarial training (Liang et al., 2018) in the SciQ and MCQL datasets; (ii) context-aware neural networks fine-tuned with BERT (Bitew et al., 2022) achieved over 40% relevant distractor retrieval in the Televic open-domain dataset; (iii) SOTA results for the DGen and CLOTH datasets showed that fine-tuning Text2Text models with data augmentation strategies generated over 22% relevant distractors.

C.2 Distractors in MC-RC

Table 8 summarizes the SOTA results in MC-RC for DG using deep neural networks, focusing on word-level to sentence-level distractors. The collected studies used a RACE-modified dataset by Gao et al. (2019), excluding samples with distractors irrelevant to the passage and questions requiring option filling at the beginning or middle. The most commonly used metric, BLUE, yielded the following observations: (i) The performance of the second and third distractors in beam search and multi-decoders showed a slight drop in BLEU-n scores due to lower likelihoods and a 0.5 Jaccard distance threshold, which enforced the use of different words. This drop was slightly less pronounced in MSG-Net due to its content selection approach. (ii) While the EDGE model achieved SOTA results in uni-gram matching for the three distractors, MSG-Net demonstrated the highest performance in bigram, trigram, and four-gram matching with the ground truth distractors.

In PLMs, Chung et al. (2020) fine-tuned the BERT model and achieved uni-gram, bigram, trigram, and four-gram matching scores of 39.81, 24.81, 17.66, and 13.56, respectively. The first distractors in fine-tuning T5 through two-step DG (Taslimipoor et al., 2024) achieved uni-gram, bigram, trigram, and four-gram matching scores of 0.31, 0.20, 0.15, and 0.12, respectively.

Paper	Task	Dataset	P@1	NDCG@10	MRR
LR+RF (2018)	MC-QA	SciQ	**36.8**	38.0	49.3
NN (2018)	MC-QA	SciQ	11.7	23.1	25.7
LR+RF (2018)	MC-QA	MCQL	**45.5**	43.8	54.8
NN (2018)	MC-QA	MCQL	22.9	34.6	36.7
DQ-SIM (2022)	MC-QA	Televic	**44.9**	—	62.8
CSG-DS (2021)	FITB	DGen	10.85	19.70	17.51
EmbSim+CF (2017)	FITB	DGen	8.10	16.33	13.86
BERT (2019)	FITB	DGen	7.72	16.21	13.60
LR+RF (2018)	FITB	DGen	8.52	19.03	15.87
multi-task (2023a)	FITB	DGen	**22.00**	—	27.15
CDGP (2022)	FITB	DGen	13.13	34.17	25.12
CDGP (2022)	FITB	CLOTH	18.50	37.82	29.96
multi-task (2023a)	FITB	CLOTH	**28.75**	—	34.46
two-step (2024)	FITB	CLOTH	26.57	47.29	—

Table 7: Ranking-based metrics for DG in FITB and MC-QA tasks.

Paper	Distractors	BLEU-1	BLEU-2	BLEU-3	BLEU-4
HSA (2019)	1st	27.32	14.69	9.29	6.47
	2nd	26.56	13.14	7.58	4.85
	3rd	26.92	12.88	7.12	4.32
CHN (2020)	1st	28.65	15.15	9.77	7.01
	2nd	27.29	13.57	8.19	5.51
	3rd	26.64	12.67	7.42	4.88
EDGE (2020)	1st	**33.03**	18.12	11.35	7.57
	2nd	**32.07**	16.75	9.88	6.27
	3rd	**31.29**	15.94	9.24	5.70
HMD-Net (2020)	1st	30.99	17.30	11.09	7.52
	2nd	30.93	16.89	10.64	7.10
	3rd	29.70	15.95	9.74	6.21
TMCA (2021)	1st	29.01	14.84	9.61	6.87
	2nd	28.26	13.79	8.68	6.10
	3rd	27.18	12.55	7.64	5.04
MSG-Net (2021)	1st	28.96	**18.15**	**12.31**	**8.87**
	2nd	27.91	**17.60**	**12.26**	**8.86**
	3rd	27.84	**17.20**	**11.81**	**8.53**

Table 8: N-gram metrics for DG using deep neural networks in MC-RC task within RACE dataset.