Distractor Generation for Multiple-Choice Questions: A Survey of Methods, Datasets, and Evaluation

Anonymous ACL submission

Abstract

Distractors as part of multiple-choice question (MCQ) are vital in learning evaluation and are commonly used in education across a variety of domains such as Science, English, and Mathematics. The advancement of artificial intelligence (AI) has enabled the *Distractor Generation* (DG) problem to progress from traditional methods into advanced neural networks and pre-trained models. This survey paper reviews DG tasks using English MCQ datasets for textual and multi-modal contexts. In particular, this paper presents a thorough literature review of the recent methods on DG tasks, discusses multiple choice components and their characteristics, analyzes the related datasets, summarizes the evaluation metrics, reveals current findings discovered from exiting benchmarks and methods, and highlights the challenges and open issues.

1 Introduction

Distractor Generation (DG) (Chen et al., 2022), the process of generating an erroneous plausible candidate answer in MCQ (Zhang et al., 2021b), is essential in education and assessment due to its objectivity and usability. Objective questions (Das et al., 2021) such as fill-in-the-blank, true-false and multiple-choice questions require an examinee to select one valid answer from a set of invalid options (Kurdi et al., 2020). It contributes into fair evaluation in several domains (e.g., Science (Liang et al., 2018), English (Panda et al., 2022), Math (McNichols et al., 2023), and Medicine (Yaneva et al., 2018)). It is also beneficial for educators in assessing large capacity of students with non-bias results.

Creating MCQs manually is one of the most labor-intensive task for educators (Ch and Saha, 2018), since questions need to include plausible false options, known as *distractors*, that are able to confuse the examinee. To generate distractors, various approaches are utilized, including similarity-based methods (Guo et al., 2016), learning-based approaches (Liang et al., 2018) that rank options according to a set of features, advanced deep neural networks (Maurya and Desarkar, 2020), transformer-based models (Chiang et al., 2022), and recently prompting methods (Bitew et al., 2023) in a large language models. These methods are applied to distractors in multiple-choice questions, reading comprehension (Gao et al., 2019) and multi-modal domains (Lu et al., 2022a).

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 MCQ (Ch and Saha, 2018; Kurdi et al., 2020; Das et al., 2021; Zhang et al., 2021b) without discussing DG tasks. A recent work (Chen et al., 2022) discussed DG as a subtask of natural language generation (NLG) in the text abbreviation tasks, rather than MCQ task. We aim to fill the gap and conduct the first survey for DG in MCQ. To this end, we collected over 100 high-quality papers from top conferences such as ACL, AAAI, IJCAI, ICLR, NAACL, and EMNLP 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.

From these collected papers, we explored English DG tasks, taxonomies, datasets and evaluation metrics to provide a comprehensive understanding of text and multi-modal research studies. Our main contributions include: (1) conducting a detailed review of the DG tasks and the recent related studies; (2) examining the existing datasets and multiple choice components used on each task to assist in choosing between selecting an available dataset or creating a new one; (3) presenting a comprehensive comparison of MCQ datasets used in DG tasks; (4) summarizing the evaluation metrics; and (5) discussing the main findings and open issues in DG methods.

The rest of this paper is organized as follows. Section 2 overviews the recent studies on DG. Section 3 introduces multiple choice components and their characteristics. Section 4 discusses MCQ datasets in DG. Section 5 summarizes the evaluation metrics and Section 6 discusses the findings in DG benchmarks and methods. Finally, Section 7 offers some concluding remarks.

2 An Overview of Distractor Generation

Distractor Generation (Ch and Saha, 2018) is time-consuming and non-trivial in MCQ, yet promising in text generation (Chen et al., 2022). Recent major developments in text fields include *multiple-choice distractor generation* (MC-DG) and *reading comprehension distractor generation* (RC-DG). *Multimodal distractor generation* (M-DG) is also proposed as a novel task to generate textual distractors in visual question answering (VQA). This section will provide an overview of research methods and Table 1 summarizes recent studies, methods and datasets[1].

2.1 Multiple-Choice DG

Generating MCQ, including cloze queries (e.g., fill-in-the-blank) (Das and Majumder, 2017) and Wh-questions (e.g., who, when, what) (Das et al., 2021) with distractors has been of interest in the community for decades (Miller, 1995; Mitkov et al., 2003; Agarwal and Mannem, 2011). Several methodologies are used for generating plausible yet incorrect distractors, including *similarity-based methods*, *ranking-based approaches*, *transformer-based models*, *candidate generation ranking framework* and *prompt-based methods*.

Similarity-based methods select distractors based on their similarity to the answer, using WordNet (Miller, 1995). This graph-based method is used in studies (Pino et al., 2008; Mitkov et al., 2009; Kumar et al., 2023), while ontology-based strategies (Stasaski and Hearst, 2017; Yaneva et al., 2018; Faizan and Lohmann, 2018) are used in domain-specific (e.g., biology and medicine). Corpus-based methods demonstrate similarity as part-of-speech (Coniam, 1997), high co-occurrence likelihood (Hill and Simha, 2016), phonetic and morphological features (Pino and Eskenazi, 2009), context sensitive inference (Zesch and Melamud, 2014), syntactic similarity (Chen et al., 2006), and semantic similarity based on embedding models like word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), and fasttext (Bojanowski et al., 2017), which are common in several studies (Jiang and Lee, 2017; Guo et al., 2016; Kumar et al., 2015; Susanti et al., 2018). These techniques however lack contextual information support and long sentence distractors.

Ranking-based approaches use learning-based models (Liu et al., 2016) to rate the existing distractor candidate pool, allowing high-quality distractors to receive high ranking scores. Liang et al. (2018) compared feature-based machine learning classifiers to neural generative adversarial networks (Goodfellow et al., 2014; Liang et al., 2017) for distractors ranking, Sinha et al. (2020) proposed a semantically aware single-encoder ranking model, and Wang et al. (2023c) used dual-encoder framework to improve the ranking models.

Transformer-based models, including large-scale pre-trained language models (PLMs) (Zhang et al., 2023a) with fine-tuning abilities, are also used in ranking models. Gao et al. (2020) improved model performance by combining hand-crafted and context-sensitive features using the masked language modelling (MLM) task in BERT (Kenton and Toutanova, 2019) and ELMo (Peters et al., 2018). Bitew et al. (2022) also utilized context-aware multilingual BERT to score distractors by reusing comparable question contexts across subjects and languages. These studies support context-awareness in distractor ranking, but they are not used to generate new distractors.

Some transformer-based models are text-to-text (Text2Text) models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). Several studies (Vachev et al., 2022; Rodriguez-Torrealba et al., 2022; Lelkes et al., 2021; Foucher et al., 2022) utilized T5 in MCQ generation. Also, Wang et al. (2023a) used Text2Text generative models (e.g., T5 and BART), data augmentation technique, and Kullback Leibler Divergence (PKL) to generate distractors with different levels of difficulty.

The candidate generation and ranking framework (CGR) involves two key steps: candidate set generator (CSG) and distractor selector (DS). Ren and Zhu (2021) introduced a context-based candidate generator using general-purpose knowledge base (KB) (Leacock, 1998; Wu et al., 2012), context-dependent conceptualisation (Kim et al., 2013; Blei et al., 2003), and ranking models to

[1]We provide Table 5 in Appendix B to summarize all the collected studies in this survey.

Models	Approaches	Datasets
Multiple-Choice DG		
(Liang et al., 2018)	Feature-based vs Neural-based Ranking Models	SciQ, MCQL
(Yeung et al., 2019)	Word Embedding + MLM (BERT)	Small-Scale
(Ren and Zhu, 2021)	Knowledge Base + LDA + Ranking Model	MCQ
(Chiang et al., 2022)	MLM (BERT) + Ranking Features	CLOTH, MCQ
(Panda et al., 2022)	Machine Translation (MT) and Alignment	Small-Scale
(Bitew et al., 2022)	MLM (mBERT)	Televic
(Gomez et al., 2023)	Machine Translation (MT) and Alignment	Small-Scale
(Yoshimi et al., 2023)	Word Embedding + MT + MLM (BERT)	Small-Scale
(Wang et al., 2023a)	Transformer (T5, BART) + MLM (BERT) + PKL	CLOTH, MCQ
(Bitew et al., 2023)	LLM (ChatGPT) + Prompting	Televic
(Doughty et al., 2024)	LLM (GPT-4) + Prompting	Small-Scale
Reading Comprehension DG		
(Sutskever et al., 2014)	RNN + Global Attention	RACE
(Gao et al., 2019)	RNN (HRED + Static Attention)	RACE
(Zhou et al., 2020)	RNN (HRED + Co-Attention)	RACE
(Maurya and Desarkar, 2020)	RNN (HRED(s) + SoftSel + Gate) + Transformer	RACE, RACE-C
(Qiu et al., 2020)	RNN + Attention + Fusion + Reforming Module	RACE
(Chung et al., 2020)	MLM (BERT)	RACE
(Offerijns et al., 2020)	Transformer (GPT2)	RACE
(Shuai et al., 2021)	RNN (HRED + TMCA + LDA + Static Attention)	RACE, DREAM
(Xie et al., 2021)	Transformer (T5)	RACE, CosmosQA
(Shuai et al., 2023)	GCN + RNN (HRED + Fusion-Attention Layers)	RACE
Multi-Modal DG		
(Lu et al., 2022a)	Generative Adversarial Network (GAN) + RL	Visual7w

Table 1: Distractor Generation (DG) tasks with recent studies and the datasets.

select distractors. Chiang et al. (2022) explored BERT in candidate generation step, while Yeung et al. (2019) proposed to re-rank distractors achieved through similarity-based approach (Jiang and Lee, 2017) with BERT language model in attempts to achieve plausible and difficult distractors.

In English test, round trip machine translation methods (Panda et al., 2022; Gomez et al., 2023) with alignment computation (Sabet et al., 2020) can detect a variety of distractors. Yoshimi et al. (2023) applied different methods (Jiang and Lee, 2017; Yeung et al., 2019; Panda et al., 2022) on different cloze English questions.

Large language models (LLMs) (Chang et al., 2023), including GPT-models (Floridi and Chiriatti, 2020; Ouyang et al., 2022) with prompt-based methods, have demonstrated promising results (Tan et al., 2022; Sarsa et al., 2022). In DG, Zu et al. (2023) proposed cloze DG using a fine-tuned GPT-2 model and five different prompts. Bitew et al. (2023) compared the ChatGPT model with zero-shot and few-shot examples to a ranking-based model (Bitew et al., 2022). McNichols et al. (2023) explored DG and feedback message generation in math using few-shot prompting. Doughty et al. (2024) compared GPT-4 for programming questions to human-crafted ones. Maity et al. (2024) explored multi-stage prompting approach using GPT-4 in multiple languages. Other studies discussed the efficacy and errors of LLMs. For instance, Tran et al. (2023) showed that GPT-4 generates isomorphic MCQs better than GPT-3, yet both models can provide many correct answers. Olney (2023) evaluated other LLMs like Macaw and Bing Chat.

2.2 Reading Comprehension DG

Reading Comprehension (Zhang et al., 2021b) is a dual-task in natural language understanding and answering, yet RC-DG is a NLG task (Chen et al., 2022) introduced by (Gao et al., 2019). In terms of approaches, *deep neural networks* and *transformer-based models* are recently proposed.

In deep neural networks, basic sequence-to-sequence model (Sutskever et al., 2014) with attention mechanism (Luong et al., 2015) can generate distractors, but not for long passage inputs as in the RACE (Lai et al., 2017) datasets. Thus, Gao et al. (2019) used a hierarchical encoder decoder framework (HRED) (Li et al., 2015) and static attention (Chen et al., 2019) to generate plausible and incorrect n-distractors through beam search and Jaccard distance during decoding. The attention helps the framework learn passage sentence distribution relevant to the query but not relevant to the answer.

Zhou et al. (2020) used HRED with co-attention mechanism (Seo et al., 2016) to allow encoder better learn the semantic interaction between article and query. Shuai et al. (2021) proposed topic information (Zhang et al., 2021a), using latent Dirichlet allocation (LDA) to enhance topic multi-head co-attention network (TMCA).

Without beam search and Jaccard distance in decoding, Maurya and Desarkar (2020) used HRED with three decoders and SoftSel operation (Tang et al., 2019) to find suitable sentence candidates for DG. Without HRED, Qiu et al. (2020) proposed a model recurrent neural network (RNN) (Rumelhart et al., 1986) (i.e., Bi-LSTM), attention and fusion layer for reforming modules to guarantee incorrectness and plausibility in generating distractors. Shuai et al. (2023), as a first work, proposed an end-to-end question-distractor joint generation framework, using graph convolutional network (GCN) (De Cao et al., 2019) and attention mechanisms (Cao et al., 2019; Gao et al., 2019).

In transformer-based models, Chung et al. (2020); Offerijns et al. (2020) used BERT (Kenton and Toutanova, 2019) and GPT2 (Radford et al.; Raffel et al., 2020) for RC-DG. Also, Xie et al. (2018) suggested a multi-selector generation network (Cho et al., 2019) with the T5 model for diverse distractor generation. The network uses question and answer aware mechanisms to ensure plausibility (i.e., options related to article and query) and reliability (i.e., options not similar to answer).

2.3 Multimodal DG

Visual Question Answering (VQA) (Zellers et al., 2019; Zhu et al., 2016) has gained recent interest, leading to textual distractor generation (DG-VQA) (Lu et al., 2022a). This task generates contextual distractors based on image, question, and correct answer, inspired by reinforcement learning (LR) and adversarial generation (Moosavi-Dezfooli et al., 2016; Goodfellow et al., 2014). It is also related to reasoning in real-world videos (Wu et al., 2021; Wang et al., 2023e).

3 Multiple Choice Components

The fundamental elements of a multiple-choice data item consist of: a *stem*, the query or question, an *answer*, the only true option, and a set of *distractors*, the set of false options. A *supported content* can be a passage (i.e., mainly in reading comprehension), a sentence, an image or a video.

3.1 Stem

A stem, known as a query or question, 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 multi-modal domain.

Fill-in-the-Blank (FITB): selecting an appropriate word, sentence or an image to complete a given content is known as cloze or FITB. In textual data, CLOTH (Xie et al., 2018) shown in (1) describes stem passage, and DGen (Ren and Zhu, 2021) in (2) indicates stem sentence. In multi-modal data, RecipeQA (Yagcioglu et al., 2018) outlines a stem image where one image is missing from the required set to complete the recipe.

(1) **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 ...*
Options -1-: *(A) excited, (B) depressed, (C) encouraged, (D) surprised*
Options -2-: *(A) turned, (B) pushed, (C) knocked, (D) forced*
Answer: *(1-A) (2-B)*

(2) **Stem:** *the main organs of the respiratory system are* _____
Options: *(A) carbon, (B) oxygen, (C) hydrogen, (D) nitrogen*
Answer: *(B)*

Multiple Choice Question (MCQ): forming a question as Wh-Q or declarative sentence is common in MCQ. SciQ (Welbl et al., 2017) in (3) and MCQL (Liang et al., 2018) in (4) illustrate textual factoid questions and declarative sentence stems, respectively.

(3) **Passage:** *All radioactive decay is dangerous to living things, but alpha decay is the least dangerous.*
Stem: *What is the least dangerous radioactive decay?*
Options: *(A) zeta decay, (B) beta decay, (C) gamma decay, (D) alpha decay*
Answer: *(D)*

(4) **Stem:** *Light energy is absorbed by pigment molecules present on*
Options: *(A) carbon, (B) oxygen, (C) sugar, (D) hydrogen*
Answer: *(C)*

3.2 Answer

An answer, also known as correct option, must be unique for each query to satisfy reliability. It can

be formed as textual short phrase or sentence. It can also be extractive from given passage or free-form generated from supported passage or prior knowledge. In multi-modal DG, it can be form as an image as shown in RecipeQA.

Short or Long Phrase: MCQL in (4) describes word level answer, while RACE (Lai et al., 2017) in (5) describes long sentence answer.

> (5) **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 ___ .*
> **Choices (AD):***(A) how they really feel when they are learning, (B) whether mood affects their learning ability, (C) what methods are easy for kids to learn, (D) the relationship between sadness and happiness*
> **Answer:** *(B)*

Extractive or Free-Form: SciQ in (3) describes extractive answer type as the answer is span on the supported content, while MCQL in (4) is free form.

3.3 Option

All options, also known as distractors or false candidates, must be incorrect candidates to satisfy objectivity. Similar to answer, options may be formed as word or sentence, mostly are separated with each query but SCDE (Kong et al., 2020) introduced shared options across all queries. candidates can also be a set of images as mentioned in RecipeQA.

Separated or Shared: CLOTH in (1) describes separated options, while SCDE in (6) shows shared options.

> (6) **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 allergic reaction. – 2 – You can follow them even when you are applying dye on your hair at home. – 3 – ...*
> **Options:** *(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...*
> **Answer:** *(1-A) (2-B) (3-D)...*

Figure 1: An Example of Visual7W in Multimodal DG.

3.4 Supported Content

A supported content can take the textual form (e.g., sentence or passage) or visual form (e.g., image or video). Passages are essential in the reading comprehension task, or optional in question answering.

Textual Form: OpenBookQA (Mihaylov et al., 2018) in (7) describes supported sentence text while RACE (Lai et al., 2017) in (5) describes passage content.

> (7) **Sentence:** *the sun is the source of energy for physical cycles on Earth*
> **Stem:** *The sun is responsible for*
> **Options:***(A) puppies learning new tricks, (B) children growing up and getting old, (C) flowers wilting in a vase, (D) plants sprouting, blooming and wilting*
> **Answer:** *(D)*

Visual Form: Visual7W in Figure 1 shows image as supported content and MovieQA (Tapaswi et al., 2016) describes movie as supported content.

4 Datasets

Table 2 describes dataset properties and classifications[2], including related domain, generation method, source of data, corpus size with unit, multiple choice components such as passage, query, and options. We determine the most common query type for each dataset, using heuristic rules[3]. We also list the average length and vocabulary size for each component, and provide an overview of each dataset availability and usability in terms of DG.

[2]We count sub-datasets (CLOTH, RACE, ARC, MCTest).
[3]https://github.com/Distractor-Generation/DG_Survey

Dataset	Domain	Source	Creation	Corpus	Unit	SC	#Q	#O	P_{avg}	Q_{avg}	O_{avg}	P_{vcb}	Q_{vcb}	O_{vcb}	MFQ	Used	AV
FITB Datasets																	
CLOTH (Xie et al., 2018)	English exam	ER	Expert	7,131	passage	✔	99,433	4	329.8	—	1	22,360	—	7,455	B	Yes	✔
CLOTH-M (Xie et al., 2018)	English exam	ER	Expert	3,031	passage	✔	28,527	4	246.3	—	1	9,478	—	3,330	B	Yes	✔
CLOTH-H (Xie et al., 2018)	English exam	ER	Expert	4,100	passage	✔	70,906	4	391.5	—	1	19,428	—	6,922	B	Yes	✔
SCDE (Kong et al., 2020)	English exam	ER	Expert	5,959	passage	✔	29,731	7	248.6	—	13.3	21,410	—	12,693	B	—	✉
DGen (Ren and Zhu, 2021)	Multi domain	M.S	Auto	2,880	sentence	✗	2,880	4	—	19.5	1	—	4,527	3,630	B	Yes	✔
CELA (Zhang et al., 2023b)	English Exam	M.S	Auto	150	passage	✔	3,000	4	408.5	—	1.3	3,500	—	3,716	B	Yes	✔
MCQ Datasets																	
SciQ (Welbl et al., 2017)	Science exam	ER	Crowd	28	book	✻	13,679	4	78	14.5	1.5	20,409	7,615	9,499	Q	Yes	✔
AQUA-RAT (Ling et al., 2017)	Math	Web	Crowd	97,975	problem	✻	97,975	5	52.7	37.2	1.6	127,404	31,406	76,115	Q	—	✔
OpenBookQA (Mihaylov et al., 2018)	Science exam	ER, WT	Crowd	1,326	WT fact	✻	5,957	4	9.4	11.5	2.9	1,416	4,295	6,989	S	—	✔
ARC (Clark et al., 2018)	Science exam	ER, Web	Expert	14M	sentence	✗	7,787	4	—	22.5	4.6	—	6,079	6,164	Q	—	✔
ARC-Challange (Clark et al., 2018)	Science exam	ER, Web	Expert	14M	sentence	✗	2590	4	—	24.7	5.5	—	4,057	4,245	Q	—	✔
ARC-Easy (Clark et al., 2018)	Science exam	ER, Web	Expert	14M	sentence	✗	5197	4	—	21.4	4.1	—	4,998	5,021	Q	—	✔
MCQL (Liang et al., 2018)	Science exam	ER, Web	Crawl	7,116	query	✗	7,116	4	—	9.4	1.2	—	5,703	7,108	S	Yes	✔
CommonSenseQA (Talmor et al., 2019)	Narrative	CN	Crowd	236,208	triplets	✗	12,102	5	—	15.1	1.5	—	6,844	6,921	Q	—	✔
MathQA (Amini et al., 2019)	Math	Web	Crowd	37,297	problem	✻	37,297	5	63.3	38.2	1.7	16,324	10,629	11,573	Q	—	✔
QASC (Khot et al., 2020)	Science exam	ER, WT	Crowd	17M	sentence	✗	9,980	8	—	9.1	1.7	—	3,886	6,407	Q	Yes	✔
MedMCQA (Pal et al., 2022)	Medicine exam	ER	Expert	2.4K	topics	✻	193,155	4	92.7	14.3	2.8	370,658	53,010	65,773	S	Yes	✔
Televic (Bitew et al., 2022)	Multi domain	ER	Expert	62,858	query	✗	62,858	>2	—	*	*	—	*	*	*	Yes	✔
EduQG (Hadifar et al., 2023)	Education	ER	Expert	13/283	book/chapter	✻	3,397	4	209.3	16.3	4.2	21,077	5,311	8,632	MF	Yes	✔
Reading Comprehension-FITB Datasets																	
ChildrenBookTest (Hill et al., 2016)	Story	PG	Auto	108	book	✔	687,343	10	474.2	31.6	1	34,611	32,912	23,253	B	Yes	✔
Who Did What (Onishi et al., 2016)	News	GIG	Auto	10,507	book	✔	205,978	2..5	*	31.4	2.1	*	70,198	82,397	B	—	✉
Reading Comprehension-MCQ Datasets																	
MCTest-160 (Richardson et al., 2013)	Story	FS	Crowd	160	story	✔	640	4	241.8	9.2	3.7	1,991	802	1,481	Q	Yes	✔
MCTest-500 (Richardson et al., 2013)	Story	FS	Crowd	500	story	✔	2,000	4	251.6	8.9	3.8	3,079	1,436	23,34	Q	Yes	✔
RACE (Lai et al., 2017)	English exam	ER	Expert	27,933	passage	✔	97,687	4	352.8	12.3	6.7	88,851	20,179	32,899	B	Yes	✔
RACE-M (Lai et al., 2017)	English exam	ER	Expert	7,139	passage	✔	28,293	4	236	11.1	5	21,566	6,929	11,379	B	Yes	✔
RACE-H (Lai et al., 2017)	English exam	ER	Expert	20,784	passage	✔	69,394	4	361.9	12.4	6.9	81,887	18,318	29,491	B	Yes	✔
RACE-C (Liang et al., 2019)	English exam	ER	Expert	4,275	passage	✔	14,122	4	424.1	13.8	7.4	34,165	10,196	15,144	B	Yes	✔
DREAM (Sun et al., 2019)	English exam	ER	Expert	6,444	dialogue	✔	10,197	3	86.4	8.8	5.3	8,449	2,791	5,864	Q	Yes	✔
CosmosQA (Huang et al., 2019)	Narratives	Blog	Crowd	21,866	narrative	✔	35,588	4	70.4	10.6	8.1	36,970	10,685	18,173	Q	Yes	✔
ReClor (Yu et al., 2019)	Standard exam	ER	Expert	6,138	passage	✔	6,138	4	75.1	17	20.8	15,095	3,370	13,592	Q	—	✔
QuAIL (Rogers et al., 2020)	Multi domain	M.S	Crowd	800	passage	✔	12966	4	395.4	9.7	4.4	13,750	6,341	9,955	Q	—	✔
Multi-Modal Dataset																	
MovieQA (Tapaswi et al., 2016)	Movie	Movies	Crowd	408	movie	✻	14,944	5	—	10.7	5.6	—	7,440	15,242	Q	—	✉
Visual7W (Zhu et al., 2016)	Visual	Images	Crowd	47,300	image	✻	327,939	4	—	8	2.9	—	12,168	15,430	Q	Yes	✔
TQA (Kembhavi et al., 2017)	Science exam	ER	Expert	1,076	lesson	✻	26,260	2..7	241.1	10.5	2.3	8,304	7,204	9,265	Q	—	✔
RecipeQA (Yagcioglu et al., 2018)	Cooking	Recipes	Auto	19,779	recipe	✔	36,786	4	575.1	10.8	5.7	78,089	5,587	71,369	B	—	✔
ScienceQA (Lu et al., 2022b)	Science exam	ER	Expert	21,208	query	✻	21,208	>2	—	14.2	4.9	—	7,373	7,638	Q	—	✔

Table 2: Multiple choice datasets. SC: supported content availability (✔: passage, ✗: no, ✻: text, image or video); (P | Q | O | B | S): (passage, question, option, blank, sentence); $(P|Q|O)_{avg}$: (P | Q | O) average token; $(P|Q|O)_{vcb}$: (P | Q | O) vocabulary size; MFQ: most frequent query; Used: usability in distractor generation (**Yes**, —: suitable); (MF | ER | MS | FS): (multi-format, educational source, multi sources, fiction story); (WT | CN | PG | GIG): (WorldTree, CONCEPTNET, Project Gutenberg, Gigaword); K/M: thousands/millions; AV: public available (✔: yes, ✉: upon request); *: require licence; —: not found.

4.1 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 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.

4.2 Data Sources

22 datasets are from educational materials such as educational websites, textbooks or WorldTree corpus (Jansen et al., 2018), and 14 are from blogs, stories, movies, images, or recipes sources.

4.3 Data Creation

30 out of 36 datasets are mostly created by human. For these 30 datasets, 18 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.

4.4 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 CONCEPTNET triplet (i.e., knowledge graph with commonsense relationships).

5 Evaluation Methods

DG evaluation focuses on: *plausibility* (i.e., options that are semantically comparable to the answer, grammatically correct with query), *reliability* (i.e., options that are incorrect), and *diversity* (i.e., options with varying difficulty levels) under two main categories: *automatic* and *manual* (Haladyna et al., 2002; Pho et al., 2014).

5.1 Automatic Evaluation

Automatic evaluation metrics are divided into *ranking-based metrics* (Valcarce et al., 2020) and *natural language generation (NLG) based metrics* (Sai et al., 2022).

5.1.1 Ranking-based Metrics

Ranking-based models in several studies as shown in Table 3 use ranking metrics to assess the performance of the models in retrieving relevant distractors to the question and the answer. The used metrics are divided into *order-unaware metrics*, *order-aware metrics*, and *semantic-based metrics*.

Order-unaware metrics, including precision (P@k), recall (R@k), and F1 score (F1@k), evaluate a model in retrieving relevant distracters across k-top locations without considering order, yet they cannot explain the location of relevant distractors. **Precision (P@k)** calculates the ratio of correctly identified relevant distractors to the total number of options ranked within top-k positions.

$$P@k = \frac{\text{Total } \textbf{Correct} \text{ Distractors in Top-k}}{\text{Number of Options in k}}$$

Recall (R@k) calculates the ratio of correctly identified relevant distractors to the total number of relevant distractors in the ground truth.

$$R@k = \frac{\text{Total } \textbf{Correct} \text{ Distractors in Top-k}}{\text{Number of Distractors in Ground-Truth}}$$

F1-score(F1@k) is the harmonic mean of precision and recall

$$F1@k = \frac{2 \times P@k \times R@k}{P@k + R@k}$$

On the other hand, order-aware metrics, including mean reciprocal rank (MRR) (Craswell, 2009), normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002), and mean average precision (MAP) (Baeza-Yates and Ribeiro-Neto, 1999), evaluate a model in retrieving relevant distractor across k-top positions with considering order.

MRR@k considers the first relevant item and is calculated by taking the average of the reciprocal ranks across all instances where N is the total number of queries and $rank_i$ is reciprocal rank of the first correct distractor in query u:

$$MRR@k = \frac{1}{N} \sum_{u=1}^{N} \frac{1}{rank_i}$$

NDCG@k compares rankings to an ideal order where all relevant items are at the top of the list and is calculated by dividing discounted cumulative gain (DCG) by ideal discounted cumulative gain (IDCG) where $DCG@k$ measures the quality of the ranked list up to position k:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

MAP@k considers the number of relevant distractors and their positions in the list (at the top) and is calculated by taking the mean of average precision ($AP@k$) at k across all queries, where $AP@k$ is computed as an average precision for given ranking list, and N is the total number of queries:

$$MAP@k = \frac{1}{N} \sum_{i=1}^{N} AP_@k_i$$

In semantic-based metric, cosine similarity is used (Ren and Zhu, 2021) to evaluate plausibility. Cosine similarity measures the similarity between two vectors A and B in a multidimensional space, by using the dot product of the vectors divided by the product of their magnitudes:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

5.1.2 NLG-based Metrics

Several studies as shown in Table 4 use NLG model metrics that are based on *word-based (n-gram)* metrics and *static-embedding metrics*.

In word-based (n-gram) metrics, bilingual evaluation understudy (BLEU) (Papineni et al., 2002) (i.e., machine translation precision-based metric) computes n-gram overlap between the reference and the hypothesis. ROUGE (Lin, 2004) computes n-gram matching based on recall. ROUGE-L measures the longest common subsequence (LCS) between a pair of sentences. These metrics have limitations in covering semantic similarity and commonsense reasoning (Novikova et al., 2017; Sulem

Model	Recall	Precision	F1	MAP	NDCG	MRR	CS
(Sinha et al., 2020)	✗	✔	✗	✔	✔	✔	✗
(Gao et al., 2020)	✔	✔	✔	✗	✗	✗	✗
(Ren and Zhu, 2021)	✔	✔	✔	✗	✔	✔	✔
(Bitew et al., 2022)	✔	✔	✗	✔	✗	✔	✗
(Chiang et al., 2022)	✔	✔	✗	✔	✔	✗	✗
(Yoshimi et al., 2023)	✗	✗	✔	✗	✗	✗	✗
(Wang et al., 2023c)	✔	✔	✗	✔	✔	✔	✗
(Wang et al., 2023a)	✔	✔	✔	✗	✔	✔	✗
(Zu et al., 2023)	✗	✗	✗	✗	✗	✗	✔
(McNichols et al., 2023)	✗	✗	✗	✗	✗	✗	✔

Table 3: Evaluation metrics on ranking-based models **MAP**: mean average precision; **NDCG**: normalized discounted cumulative gain; **MRR**: mean reciprocal rank; **CS**: cosine similarity; ✔: used; ✗: not used.

Model	BLUE	ROUGE	ROUGE_L	METEOR	E.A	G.M	V.E	CS	S.B
(Gao et al., 2019)	✔	✔	✔	✗	✗	✗	✗	✗	✗
(Zhou et al., 2020)	✔	✔	✔	✗	✗	✗	✗	✗	✗
(Maurya and Desarkar, 2020)	✔	✔	✔	✔	✔	✔	✔	✔	✗
(Qiu et al., 2020)	✔	✔	✔	✗	✗	✗	✗	✗	✗
(Chung et al., 2020)	✔	✗	✔	✗	✗	✗	✗	✗	✗
(Offerijns et al., 2020)	✔	✗	✔	✗	✗	✗	✗	✗	✗
(Xie et al., 2021)	✔	✗	✔	✔	✗	✗	✗	✗	✔
(Shuai et al., 2021)	✔	✗	✗	✗	✗	✗	✗	✗	✗
(Rodriguez-Torrealba et al., 2022)	✔	✗	✔	✗	✗	✗	✗	✔	✗
(Shuai et al., 2023)	✔	✗	✔	✗	✗	✗	✗	✗	✗
(Maity et al., 2024)	✔	✗	✔	✗	✗	✗	✗	✔	✗
(Wang et al., 2023b)	✔	✗	✗	✗	✗	✗	✗	✗	✗

Table 4: Evaluation metrics on NLG-based models **E.A**: embedding average; **G.M**: greedy matching; **V.E**: vector extrema; **CS**: cosine similarity; ✔: used; ✗: not used.

et al., 2018; Xie et al., 2021). Two studies (Maurya and Desarkar, 2020; Xie et al., 2021) used METEOR (Lavie and Denkowski, 2009) F-score based metric. It is unigram lexical-similarity-based machine translation metric that captures lexical overlap, semantic and relevance. Static-embedding metrics such as Greedy Matching (Rus and Lintean, 2012), Embedding Average (John et al., 2016) and Vector Extrema (Forgues et al., 2014) are also used for semantic similarity evaluation. Self-BLEU (Caccia et al., 2019) is used for measuring the diversity (Xie et al., 2021) of distractors by calculating the average BLEU score

In the RC-DG task, accuracy score (Lan et al., 2019a) is applied (Xie et al., 2021; Shuai et al., 2021) with BERT and ALBERT (Lan et al., 2019b) as reading comprehension systems.

5.2 Human Evaluation

Human judgments (Ghanem and Fyshe, 2023) are essential in DG. In multiple-choice studies, *reliability* and *plausibility* are the most common metrics (Ren and Zhu, 2021). Participants use a 3-point scale for plausibility, and a binary mode for reliability of given generated and ground-truth distractors.

In reading comprehension studies, *Comparative* methods (Gao et al., 2019) involve the selection of distractors based on specific objectives: *Confusion* assesses the number of times a distractor being chosen as the best option without providing the correct answer and *Non-error* metric measures the number of correct answers to a question via distractors.

Quantitative (Maurya and Desarkar, 2020) methods rely on numerical scales within specific ranges: *Fluency* assesses if the distractor follows proper English grammar, human logic, and common sense, *Coherence* evaluates distractors key phrases for relevance to the article and question, *Distractibility* measures the likelihood of a candidate being chosen as a distractor in real exams, *Diversity* measures semantic difference between multiple distractors, and finally *Difference* measures the proportion of distractors and answer with the same semantics.

Bitew et al. (2022) proposed two metircs, good distractor rate (GDR@K) and nonsense distractor rate (NDR@K), to calculate the percentage of good rated distractors and nonsense (i.e., completely out of context) distractors at top K positions.

6 Discussion and Findings

After reviewing these DG studies, we discovered several interesting findings. Specifically, multimodal DG is limited but becoming a novel task in VQA. Open domain datasets are crucial in understanding the LLM performance. Although LLMs have shown comparable performance to human-crafted texts, these models still face challenges in generating good distractors (i.e., plausible but incorrect). Exploring the effectiveness of auto-generated questions by large models in schools may be beneficial in education assessments. The detailed discussion of each key finding can be found in Appendix A and a comprehensive survey literature tree is presented in Figure 2 at the end of Appendix.

7 Conclusion

Distractor Generation (DG) is critical in education assessment and has recently received significant attention in the research community. This paper surveys the current research activities on DG tasks, the related DG datasets, and the evaluation methods. We classify the recent DG developments into multiple choice, reading comprehension, and multi-modal tasks. We analyze the characteristics of the multiple choice components, compare the DG datasets, collect the evaluation methods, and discuss the main findings. Reading list is on https://github.com/Distractor-Generation/DG_Survey.

8 Limitations

We report the following limitations for our DG survey. The survey covers contemporary research in advanced neural networks and LLM, but it may not cover the field history. Despite a concise review of the methods, datasets, and evaluation metrics, we did not cover qualitative and quantitative comparison between models and benchmark datasets. Our discussion did not address whether existing approaches are suitable in deployment for education. Despite these limitations that could consider as future work, our survey is the first contribution in exploring distractor generation tasks and gives a concise summary in main findings and challenges, which can serve as a valuable resource for scholars working in this field.

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.

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.

R Baeza-Yates and B Ribeiro-Neto. 1999. Modern information retrieval addison-wesley longman. *Reading MA*.

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*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

Kevin Burton, Akshay Java, Ian Soboroff, et al. 2009. The icwsm 2009 spinn3r dataset. In *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 *International Conference on Learning Representations*.

Yu Cao, Meng Fang, and Dacheng Tao. 2019. Bag: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. In *Proceedings of NAACL-HLT*, pages 357–362.

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

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.

D Chen, L Yinghui, G Haifan, et al. 2022. A survey of natural language generation. *ACM Computing Survey (CSUR)*.

Wang Chen, Yifan Gao, Jiani Zhang, Irwin King, and Michael R Lyu. 2019. -guided encoding for keyphrase generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6268–6275.

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.

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.

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.

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*.

David Coniam. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *Calico Journal*, pages 15–33.

Nick Craswell. 2009. Mean reciprocal rank. *Encyclopedia of database systems*, 1703.

Bidyut Das and Mukta Majumder. 2017. Factual open cloze question generation for assessment of learner's knowledge. *International Journal of Educational Technology in Higher Education*, 14:1–12.

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.

Neisarg Dave, Riley Bakes, Barton Pursel, and C Lee Giles. 2021. Math multiple choice question solving and distractor generation with attentional gru networks. *International Educational Data Mining Society*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of NAACL-HLT*, pages 2306–2317.

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.

Ainuddin Faizan and Steffen Lohmann. 2018. Automatic generation of multiple choice questions from slide content using linked data. In *Proceedings of the 8th International Conference on web intelligence, mining and semantics*, pages 1–8.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

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.

Lingyu Gao, Kevin Gimpel, and Arnar Jensson. 2020. Distractor analysis and selection for multiple-choice cloze questions for second-language learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 102–114.

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*.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *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 *IJCAI-16: Proceedings of the AAAI Twenty-Fifth International Joint Conference on Artificial Intelligence*.

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.

Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th 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.

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.

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)*.

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 SY 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.

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*, 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.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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.

Dongwoo Kim, Haixun Wang, and Alice Haeyun Oh. 2013. Context-dependent conceptualization. In *international joint conference on Artificial Intelligence*, pages 2654–2661. International Joint Conferences on Artificial Intelligence Organization (IJCAI).

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.

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 E 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.

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.

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*. Association for Computational Linguistics.

Yunshi Lan, Shuohang Wang, and Jing Jiang. 2019a. Knowledge base question answering with a matching-aggregation model and question-specific contextual relations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1629–1638.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019b. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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

Claudia Leacock. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: A Lexical Reference System and its Application*, pages 265–283.

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.

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.

Jiwei Li, Minh-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.

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.

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 *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.

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.

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.

Ming Liu, Vasile Rus, and Li Liu. 2016. Automatic chinese factual question generation. *IEEE Transactions on Learning Technologies*, 10(2):194–204.

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*, pages 4921–4930.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021. 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.

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 *International Conference on Learning Representations (ICLR)*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. 2024. A novel multi-stage prompting approach for language agnostic mcq generation using gpt. *arXiv preprint arXiv:2401.07098*.

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*.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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

Ruslan Mitkov, Andrea Varga, Luz Rello, et al. 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.

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.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

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

Andrew M Olney. 2023. Generating multiple choice questions from a textbook: Llms match human performance on most metrics.

Takashi 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.

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 *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

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.

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.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Van-Minh Pho, Thibault André, Anne-Laure Ligozat, Brigitte Grau, Gabriel Illouz, and Thomas François. 2014. Multiple choice question corpus analysis for distractor characterization. In *International conference on language resources and evaluation*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

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.

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 JC 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.

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.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

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.

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.

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

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.

Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43.

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

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.

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.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 738. 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.

Yicheng Sun and Jie Wang. 2023. Constructing cloze questions generatively. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. Msp: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142.

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.

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 *2023 IEEE 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 *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.

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.

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 *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.

Zhecan Wang, Long Chen, Haoxuan You, Keyang Xu, Yicheng He, Wenhao Li, Noal Codella, Kai-Wei Chang, and Shih-Fu Chang. 2023e. Dataset bias mitigation in multiple-choice visual question answering and beyond. *arXiv preprint arXiv:2310.14670*.

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.

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.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

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.

Victoria Yaneva et al. 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.

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

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.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2019. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

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.

Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2021a. Modeling topical relevance for multi-turn dialogue generation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3737–3743.

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. 2021b. 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.

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 Discussion and Findings

Multimodal distractor generation is limited but becoming a novel task in visual question answering (Lu et al., 2022a).

Benchmark datasets for DG typically focus on textual-only modality, with datasets mainly focusing on Science (Welbl et al., 2017) and English (Xie et al., 2018), with recent datasets (Bitew et al., 2022; Hadifar et al., 2023) focusing on multi-domain fields. In education, textual questions and supported content, such as figures (Wang et al., 2021; Lu et al., 2021), charts (Kafle et al., 2018), and tables (Lu et al., 2023), are used in real assessments, including science (Kembhavi et al., 2017; Lu et al., 2022b) and mathematics (Verschaffel et al., 2020). Despite the availability of datasets, there is limited work in multi-modal distractor generation.

Open-domain datasets are crucial for understanding the performance of pre-trained large language models in distractor generation.

SciQ, CLOTH, and RACE are recent DG datasets that are mostly used. SciQ handles textual science questions, while CLOTH and RACE handle textual English questions. These datasets have been significantly used to compare DG methods in specific domains, but recent open-domain datasets like Televic (Bitew et al., 2022) and EduQG (Hadifar et al., 2023) have significantly contributed to understanding the performance of advanced pre-train large language models (Bitew et al., 2023).

Pre-trained large language models still face concerns about generating nonsense distractors, which are the same as answers or previous generated distractors.

Generating repeated incoherent or factual inconsistent results are commonly concerns in NLG (Ji et al., 2023). In generative models such as T5, Wang et al. (2023a) emphasized that candidate augmentation strategy with generative models contributes in reducing the occurrence of generated nonsense distractors, yet still a problem. In ranking-based approach (Bitew et al., 2022), two metrics were used to measure the number of correct distractors compared to non-correct ones in manual evaluation stage. The models with prompting showed greater performance in reducing nonsense distractors compared to fine-tuning models (Bitew et al., 2023). In particular, nonsense distractor rate reduced from 50% in ranking-based model (Bitew et al., 2022) to 16%. GPT-4 with prompting in programming domain (Doughty et al., 2024) demonstrated superior performance in generating MCQ with quality language and logical syntax comparable to human-crafted once.

One of the most important goals in distractor generation is how to generate diverse distractors.

In cloze questions, Wang et al. (2023a) proposed using the pseudo Kullback-Leibler Divergence (PKL) technique to regulate the inter-correlation between generated distactors in attempt to generate

difficult distractors like human-created exam (i.e., one may be easily eliminated while the other two put a greater challenge in identifying the correct answer), though it does not perform well. Two studies (Panda et al., 2022; Gomez et al., 2023) discussed diversity of distractors based on back neural machine translation (MT) systems which showed noticeable results with some drawbacks, including producing inadequate candidates and being computationally more expensive than previous methods. Also, high-quality machine translation systems are crucial for language pairs without English pivot, and Gomez et al. (2023) raises a question about MT benefits for first language learners.

In reading comprehension, early RNN models (Gao et al., 2019; Zhou et al., 2020; Qiu et al., 2020; Shuai et al., 2021) had limitations in beam search for generating n-distractors using Jaccard distance, treating the problem as a one-to-many task (i.e., same input generates multiple distractors). Two studies proposed diversity as characteristic for DG. Maurya and Desarkar (2020) adopted mixture of decoders and Xie et al. (2021) adopted mixture content selection (Cho et al., 2019) with T5 model. Mixture of decoders still produced lexically diverse but semantically similar distractors. Mixture content selection produced diverse distractors from different sentences in passage, coherent with the question and not equivalent to the answer.

One of the current tasks in distractor generation by large models is assessing the effectiveness of auto-generated multiple-choice questions in evaluating students learning.

Because the quality of multiple choice questions generated by large language models regardless of nonsense problems is similar to human-generated questions (Doughty et al., 2024), it is essential to explore the effectiveness of using auto-generated in the field of education and pedagogical studies. Tran et al. (2023) proposed a future study to assess students by conducting an experiment using human-created MCQs and auto-generated questions. The experiment is expected to contribute in understanding the effectiveness of evaluating students by auto-generated questions and the possibility of using LLMs in real educational assessments.

B Datasets

Table 2 describes dataset properties and classifications.

B.1 Data Sources

Out of 36 datasets, 22 are from educational materials and 14 are from blogs, stories, movies, images, or recipes 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 WT and CK-12. ReClor is generated from open websites and books. EduQG, Televic, and MedMCQA are collected from 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.2 Passage/Query/Options Components

The only dataset introduced as multi-format by labeling and forming a query as cloze and normal is EduQG. Thus, to find the most common query types (i.e., blank, sentence or question) as shown in Table 2, we utilized dataset analysis as proposed by Dzendzik et al. (2021) to process our heuristic rules and statistics. Using spaCy[4] tokenizer we

[4] https://spacy.io/.

determined the average length and vocabulary size of queries, passages, and options.

- **Passage**: all datasets contain supported text except DGen, ARC, CommonSenseQA, MCQL, QASC and Televic in textual datasets. In multi-modal, some datasets such as RecipeQA and TQA contain supported text and images. Other datasets such as MovieQA contains movies as supported content 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 is SciQ (14K) and in math dataset is AQUA-RAT (98K). Televic contains (63K) questions, yet it is multilingual datasets (i.e., 50% in Dutch then French and English next common) and the authors provided sample test with 198 questions (Q_{avg}14.9,O_{avg} 1.9) token in GitHub. In reading comprehension, the most usable dataset is RACE (98K). The largest number of questions in multi-model is Visual7W (327.9K).

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

- **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).

- **Vocabulary Size**: The vocabulary for passages ranges from 1.4K to 371K based on the number of unique lowercased token lemmas for each component in MCQs. The vocabulary for the queries spans from 802 to 70.2K, and the options span from 1.5K to 82.4K.

B.3 Data Usability and Availability

CLOTH, DGen, SciQ, and MCQL are benchmark datasets used in several studies (Ren and Zhu, 2021; Chiang et al., 2022; Wang et al., 2023a; Liang et al., 2018; Sinha et al., 2020; Sun and Wang, 2023). Televic (Bitew et al., 2022), EduQG (Hadifar et al., 2023), and MedMCQA (Wang et al., 2023d) are studied in the DG task. RACE is also a benchmark dataset in the RC-DG task. CosmosQA and DREAM are used in recent studies (Shuai et al., 2021; Xie et al., 2021). MCTest (Wang et al., 2023b) is utilized for data augmentation. CBT, QuAIL and ReClor (Sharma Mittal et al., 2018; Ghanem and Fyshe, 2023) are also studied in part of DG and Visual7W (Lu et al., 2022a) is used for textual DG in visual question answering.

The majority of datasets are public except upon request datasets (i.e., SCDE, MovieQA) and upon payment of a licence fee to access part of dataset (i.e., WDW) or whole dataset (i.e., Televic the whole dataset requires IEEE DataPort subscription).

Model	Task	DG Approach	Dataset	Query	Domain	Automatic	Manual
(Mitkov et al., 2003)	Multiple-choice DG	Graph-based	Small-scale	MCQ	English	✗	✔
(Chen et al., 2006)	Multiple-choice DG	Corpus-based	Web Crawl	Cloze	English	✗	✔
(Pino et al., 2008)	Multiple-choice DG	Graph-based	REAP database	Cloze	English	✗	✔
(Pino and Eskenazi, 2009)	Multiple-choice DG	Corpus-based	Small-scale	Cloze	English	✗	✔
(Mitkov et al., 2009)	Multiple-choice DG	Graph-based	Small-scale	MCQ	English	✗	✔
(Zesch and Melamud, 2014)	Multiple-choice DG	Corpus-based	Small-scale	Cloze	English	✗	✔
(Sutskever et al., 2014)	Reading Comprehension DG	RNN and Attention	RACE	MCQ	English	✔	✔
(Kumar et al., 2015)	Multiple-choice DG	Semantic-Similarity	Text-book	Cloze	Biology	✗	✔
(Hill and Simha, 2016)	Multiple-choice DG	Corpus-based	Small-scale	Cloze	English	✗	✔
(Guo et al., 2016)	Multiple-choice DG	Semantic-Similarity	Wikipedia	MCQ	Wikipedia	✗	✔
(Stasaski and Hearst, 2017)	Multiple-choice DG	Graph-based	K-12 Biology concepts	MCQ	Biology	✗	✔
(Jiang and Lee, 2017)	Multiple-choice DG	Similarity-based	Chinese Data	MCQ	Chinese	✗	✔
(Liang et al., 2017)	Multiple-choice DG	Learning-based	Small-scale	Cloze	Biology	✗	✔
(Faizan and Lohmann, 2018)	Multiple-choice DG	Graph-based	SlideWiki	MCQ	Open-Domain	✗	✔
(Susanti et al., 2018)	Multiple-choice DG	Semantic-Similarity	English Wikipedia	MCQ	English	✗	✔
(Liang et al., 2018)	Multiple-choice DG	Ranking-based	SciQ, MCQL	MCQ	Science	✔	✔
(Yeung et al., 2019)	Multiple-choice DG	CGR	Small-scale	Cloze	English	✗	✔
(Gao et al., 2019)	Reading Comprehension DG	HRED	RACE	MCQ	English	✔	✔
(Zhou et al., 2020)	Reading Comprehension DG	HRED	RACE	MCQ	English	✔	✔
(Maurya and Desarkar, 2020)	Reading Comprehension DG	HRED(s)	RACE, RACE-C	MCQ	English	✔	✔
(Qiu et al., 2020)	Reading Comprehension DG	RNN	RACE	MCQ	English	✔	✔
(Chung et al., 2020)	Reading Comprehension DG	Transformer-based	RACE	MCQ	English	✔	✗
(Sinha et al., 2020)	Multiple-choice DG	Ranking-based	SciQ, RACE	MCQ	Open-domain	✔	✗
(Offerijns et al., 2020)	Reading Comprehension DG	Transformer-based	RACE	MCQ	English	✔	✔
(Xie et al., 2021)	Reading Comprehension DG	Transformer-based	RACE , CosmosQA	MCQ	Open-Domain	✔	✔
(Shuai et al., 2021)	Reading Comprehension DG	HRED	RACE, DREAM	MCQ	English	✔	✔
(Dave et al., 2021)	Multiple-choice DG	RNN	Mathematics Dataset	MCQ	Math	✔	✔
(Lelkes et al., 2021)	Multiple-choice DG	Transformer-based	NewsQuizQA	MCQ	News	✔	✔
(Ren and Zhu, 2021)	Multiple-choice DG	CGR	DGen	Cloze	Open-Domain	✔	✔
(Chiang et al., 2022)	Multiple-choice DG	CGR	CLOTH, DGen	Cloze	Open-Domain	✔	✔
(Panda et al., 2022)	Multiple-choice DG	MT	ESL Lounge Website	Cloze	English	✔	✔
(Vachev et al., 2022)	Multiple-choice DG	Transformer-based	RACE	MCQ	News	✔	✗
(Rodriguez-Torrealba et al., 2022)	Multiple-choice DG	Transformer-based	RACE	MCQ	Wikipedia	✔	✔
(Foucher et al., 2022)	Multiple-choice DG	Transformer-based	Small-scale	MCQ	Open-domain	✗	✔
(Lu et al., 2022a)	Multi-modal DG	GAN and RL	Visual7w	MCQ	Multi-modal	✔	✗
(Bitew et al., 2022)	Multiple-choice DG	Transformer-based	Televic , WeZooz	MCQ	Open-domain	✔	✔
(Shuai et al., 2023)	Reading Comprehension DG	GCN and HRED	RACE	MCQ	English	✔	✔
(Kumar et al., 2023)	Multiple-choice DG	Similarity based	Text-book	MCQ	Computing	✗	✔
(Wang et al., 2023b)	Multiple-choice DG	Ranking-based	RACE, MCTest	MCQ	Open-domain	✔	✗
(Wang et al., 2023c)	Multiple-choice DG	Ranking-based	RACE, MCTest	MCQ	Open-domain	✔	✔
(Wang et al., 2023a)	Multiple-choice DG	Transformer-based	CLOTH, DGen	Cloze	Open-domain	✔	✔
(Gomez et al., 2023)	Multiple-choice DG	MT	Small-scale	Cloze	English	✔	✔
(Yoshimi et al., 2023)	Multiple-choice DG	Multiple Methods	Small-scale	Cloze	English	✔	✔
(Zu et al., 2023)	Multiple-choice DG	LLM and Prompt	Large-scale	Cloze	English	✔	✔
(Bitew et al., 2023)	Multiple-choice DG	LLM and Prompt	Televic, Wezooz	MCQ	Open-domain	✗	✔
(Tran et al., 2023)	Multiple-choice DG	LLM and Prompt	Small-scale	MCQ	Computing	✗	✔
(McNichols et al., 2023)	Multiple-choice DG	LLM and Prompt	Eedi Repository	MCQ	Math	✔	✔
(Olney, 2023)	Multiple-choice DG	LLM and Prompt	Small-scale	MCQ	Anatomy and Physiology	✗	✔
(Hadifar et al., 2023)	Multiple-choice DG	Transformer-based	EduQG	Multi-format	Open-domain	✗	✔
(Doughty et al., 2024)	Multiple-choice DG	LLM and Prompt	Small-scale	MCQ	Computing	✗	✔
(Maity et al., 2024)	Multiple-choice DG	LLM and Prompt	SQuAD and others	MCQ	Multi-lingual	✔	✔

Table 5: A summary of the studies in DG tasks. DG: distractor generation; MCQ: multiple choice question; CGR: candidate generation and ranking framework; RNN: recurrent neural network; HRED: hierarchical encoder decoder framework; MT: round-trip machine translation; GAN: generative adversarial network; RL: reinforcement learning; GCN: graph convolutional network; LLM: pre-trained large language model; ✔: used in evaluation; ✗: not used in evaluation.

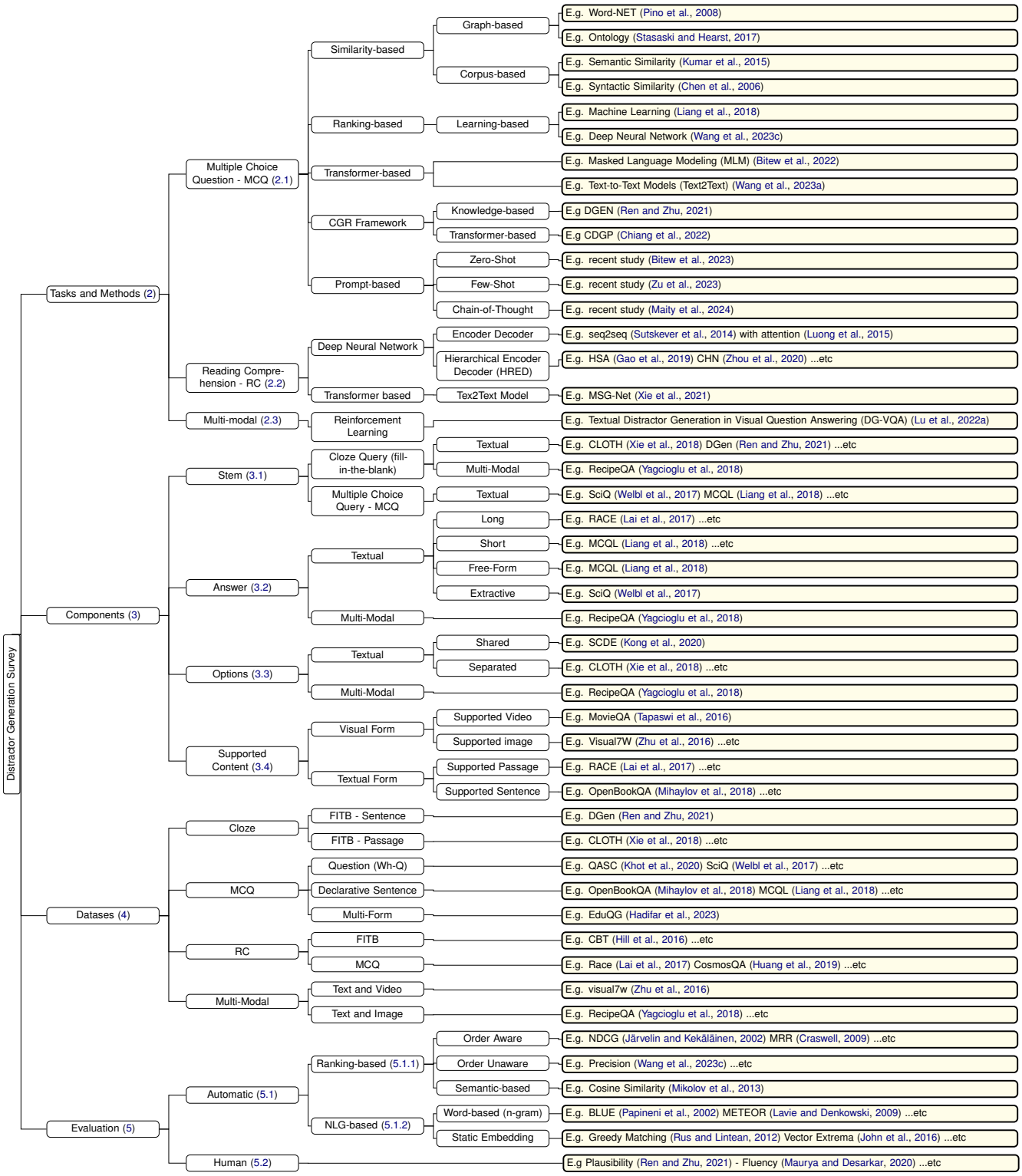

Figure 2: The literature survey tree.