

# E-KAR : A Benchmark for Rationalizing Natural Language Analogical Reasoning

Jiangjie Chen<sup>♣,♦,\*</sup>, Rui Xu<sup>♣</sup>, Ziquan Fu<sup>♡</sup>, Wei Shi<sup>♣</sup>, Zhongqiao Li<sup>♣</sup>,  
Xinbo Zhang<sup>◇</sup>, Changzhi Sun<sup>◇,†</sup>, Lei Li<sup>¶</sup>, Yanghua Xiao<sup>♣,§,†</sup>, Hao Zhou<sup>◇</sup>

<sup>♣</sup>Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University

<sup>◇</sup>ByteDance AI Lab <sup>♡</sup>Brain Technologies, Inc.

<sup>♣</sup>South China University of Technology <sup>¶</sup>University of California Santa Barbara

<sup>§</sup>Fudan-Aishu Cognitive Intelligence Joint Research Center

{jjchen19, shawyh}@fudan.edu.cn, sunchangzhi@bytedance.com

## Abstract

The ability to recognize analogies is fundamental to human cognition. Existing benchmarks to test word analogy do not reveal the underneath process of analogical reasoning of neural models. Holding the belief that models capable of reasoning should be right for the right reasons, we propose a first-of-its-kind Explainable Knowledge-intensive Analogical Reasoning benchmark (**E-KAR**). Our benchmark consists of 1,655 (in Chinese) and 1,251 (in English) problems sourced from the Civil Service Exams, which require intensive background knowledge to solve. More importantly, we design a free-text explanation scheme to explain whether an analogy should be drawn, and manually annotate them for each and every question and candidate answer. Empirical results suggest that this benchmark is very challenging for some state-of-the-art models for both explanation generation and analogical question answering tasks, which invites further research in this area. Project page of **E-KAR** can be found at <https://ekar-leaderboard.github.io>.

## 1 Introduction

Analogy holds a vital place in human cognition, driving the discovery of new insights and the justification of everyday reasoning (Johnson-Laird, 2006; Gentner and Smith, 2012; Bartha, 2013; Benio et al., 2021). Due to their unique value in many fields such as creativity (Goel, 1997) and education (Thagard, 1992), analogy and analogical reasoning have become a focus in AI research. The grand question is, are artificial neural networks also capable of recognizing analogies?

Relatively little attention has been paid in NLP to answer this question. The problem of recognizing analogies is mainly benchmarked in the form

\*Work is done during internship at ByteDance AI Lab.

†Corresponding authors.

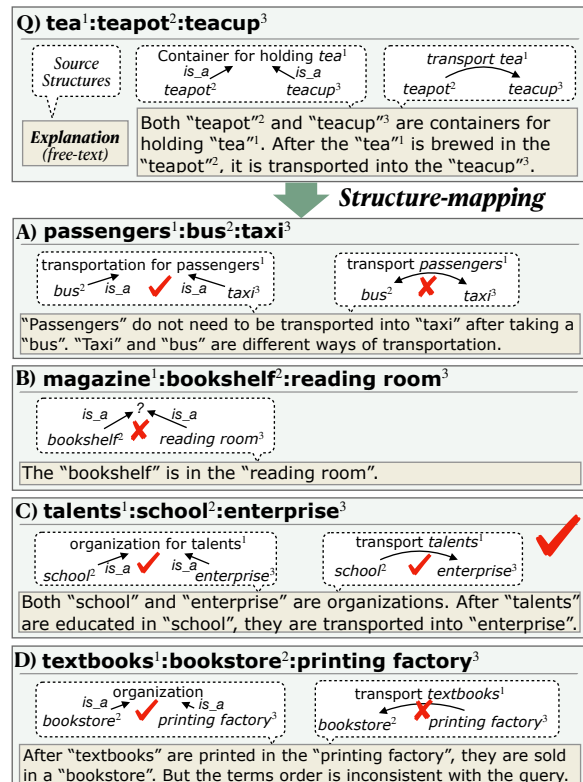


Figure 1: An example in **E-KAR**. The explanations in **E-KAR** explain the *structure-mapping* process for analogical reasoning, where source structures are drawn from the query and mapped onto each candidate answer for decision-making.

of (A:B::C:D) (Turney et al., 2003; Mikolov et al., 2013b; Gladkova et al., 2016; Li et al., 2018a) and targeted for testing the ability of pre-trained word embeddings. Given a tuple of terms as *query* (e.g., tea:teapot:teacup) and a list of *candidate answers* as in Figure 1, a model needs to find the most analogous candidate to the query, which is C in the example since it matches the relations inherent in the query better than others.

Most methods (Mikolov et al., 2013a; Levy and Goldberg, 2014; Pennington et al., 2014) hold a

connectionist assumption (Feldman and Ballard, 1982) of *linear analogy* (Ethayarajh et al., 2019), that the relation between two words can be estimated by vector arithmetic of word embeddings. For example,  $\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} = \vec{\text{queen}}$ . However, current benchmarks focus on the recognition of binary analogies such as syntactic, morphological and direct semantic (e.g., *is\_a* and *synonym\_of*) relations. And the analogical reasoning procedure behind them is far beyond the scope of this line of research.

In addition, how to explain and rationalize analogical reasoning remains to be the major challenge. Psychological literature (Gick and Holyoak, 1983; Gentner, 1983; Minnameier, 2010) suggests that analogical reasoning follows the *structure-mapping* process. That is, a target (the domain where a problem must be solved, i.e., candidates) and a source (the domain where the analogy is drawn, i.e., the query) are matched, and the relevant features of the source have to be mapped onto the target. In Figure 1, source structures are drawn (or *abducted*) from the query and mapped onto candidates, and candidates A, B, D all fail at certain structures. We argue that such a process can be verbalized into natural language to explain analogical reasoning.

Moving from simply recognizing analogies to exploring human-like reasoning for neural models, we emphasize the importance of a new kind of analogical reasoning benchmark. To fill in this blank, we propose a first-of-its-kind benchmark for **Explainable Knowledge-intensive Analogical Reasoning (E-KAR)**. We collect 1,655 analogical reasoning problems sourced from the publicly available Civil Service Examinations (CSE) of China. These CSE problems are challenging multiple-choice problems designed by human experts, thus solving them requires the intensive involvement of linguistic, commonsense, encyclopedic, and cultural (e.g., idiom and historical) knowledge.

To justify the reasoning process, we follow the aforementioned guidelines from psychological theories and manually annotate free-text explanations for each query and candidate answers in **E-KAR**. Since the annotation requires intensive involvement of knowledge and reasoning, we carefully design a *double-check* procedure for quality control. We also translate this dataset into an English version, resulting in 1,251 problems after discarding language and cultural specific cases.

In summary, our contributions include:

- We advance the traditional setting of word analogy recognition by introducing a knowledge-intensive analogical reasoning benchmark (**E-KAR**) in Chinese and English, which is first-of-its-kind and challenging.
- To justify the analogical reasoning process, we design free-text explanations according to theories on human cognition, and manually annotate them.
- In **E-KAR**, we define two tasks (analogical QA and explanation generation) in two modes (EASY and HARD) and report the performance of some state-of-the-art language models. We discuss the potentials of this benchmark and hope it facilitates future research on analogical reasoning.

## 2 Related Work

**Word Analogy Recognition in NLP** Benchmarks for word analogy recognition (Turney et al., 2003; Mikolov et al., 2013b; Gladkova et al., 2016; Li et al., 2018a) examine mostly linear relations between words (Ethayarajh et al., 2019). Such analogies can often be effectively solved by vector arithmetic for neural word embeddings, such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014). Recent studies (Brown et al., 2020; Ushio et al., 2021) also test such ability of pre-trained language models (PLMs) (Radford et al., 2019; Devlin et al., 2019; Brown et al., 2020) on these benchmarks. An exceptional benchmark is Li et al. (2020), where they build a knowledge-enhanced analogy benchmark that leverages word sense definitions in a commonsense knowledge base (Ma and Shih, 2018). However, these benchmarks are mainly set up for evaluating learned representations, and few of them ever investigated the analogical reasoning skills for neural models. Thus, the goal of this work largely differs from this line of research, as we aim to build a knowledge-intensive benchmark to teach neural models analogical reasoning for correct thinking.

**Reasoning Benchmarks from Examinations** There are abundant benchmarks derived from human examinations to facilitate the study of machine reasoning (Clark et al., 2016; Schoenick et al., 2017). For example, RACE (Lai et al., 2017) is collected from the English exams for middle and high school students, focusing on skills of passage summarization and attitude analysis. ARC (Clark et al., 2018) contains natural, grade-school science

questions authored for human tests. MCQA (Guo et al., 2017), GeoSQA (Huang et al., 2019) and GCRC (Tan et al., 2021) are sourced from national college entrance exams of China, measuring a comprehensive set of reasoning abilities. LogiQA (Liu et al., 2020a) consists of logical reading comprehension problems from Civil Service Exams of China, which is also our source of analogical problems. ReClor (Yu et al., 2020) and LR-LSAT (Wang et al., 2021), collected from Law School Admission Test, aim for testing logical reasoning abilities. In our work, we focus on analogical reasoning skills for machines and additionally equip **E-KAR** with annotated explanations to rationalize reasoning.

**Explainable NLP Datasets** One of the most prominent objectives in machine reasoning is giving reasons for a prediction. In current datasets for explainable NLP, such reasons can be categorized into three classes (Wiegrefe and Marasović, 2021): 1) *highlights explanations* (Camburu et al., 2018; Yang et al., 2018; Thorne et al., 2018; Kwiatkowski et al., 2019), which are subsets of the input elements to explain a prediction, e.g., words or sentences; 2) *free-text explanations* (Camburu et al., 2018; Zellers et al., 2019; Aggarwal et al., 2021) that are textual explanations for justification; 3) *structured explanations* (Mihaylov et al., 2018; Khot et al., 2020; Clark et al., 2020; Jhamtani and Clark, 2020; Geva et al., 2021), which are not fully free-text and generally follow certain structures such as a chain of facts. The explanations can be utilized to augment (Rajani et al., 2019), supervise (Camburu et al., 2020) and evaluate (DeYoung et al., 2020) model predictions. In this work, we phrase analogical reasoning itself as an instance of machine reasoning tasks with free-text rationales, advancing the research on analogical reasoning from the perspectives of data collection.

### 3 Explainable Analogical Reasoning

In this work, we consider a classic setting of analogical reasoning within NLP: recognizing word/term analogies.<sup>1</sup> This task can be formulated as multiple-choice question-answering. Given a query tuple  $Q$  with  $k$  (two or three) terms, and  $m$  candidate answer tuples  $A = \{A_i\}_{i=1}^m$ , the goal is to find the most analogous one in the candidates to the query.

We advocate that reasoning is about giving reasons explaining a prediction. In order to teach

<sup>1</sup>Here, “term” corresponds to “word” in previous analogy benchmarks, but allows for multiple words.

machines to analogize as humans do, we draw inspiration from theories in cognitive psychology to design the forms of explanations.

#### 3.1 Analogical Reasoning: A Psychological Perspective

Before designing suitable forms of explanations, we introduce some important theories from cognitive psychology for a better understanding of analogical reasoning. In the psychological literature, analogical reasoning is described as a *schema-induction* (Gick and Holyoak, 1983) or *structure-mapping* (Gentner, 1983) process. Peirce (1896) claimed that analogy is a combination of abductive and inductive reasoning. Minnameier (2010) further developed the inferential process of analogy into three steps, which we take as the guidelines for designing explanations:

1. A possibly suitable structure in the source domain is abduced from the target domain, which might also work for the target;
2. The specific concepts of the source structure have to be replaced by suitable target concepts (by an inductive inference);
3. The validity of the transformation is judged w.r.t. solving the target problem.

Take Figure 1 for example: Source structures can be abduced that both term 2 (teapot) and term 3 (teacup) belong to a concept, and term 1 (tea) can be transported from term 2 to term 3. The mapping naturally reveals the validity, for example, candidate A is wrong because `passengers` do not follow a unidirectional transportation (i.e., from `bus` to `taxi`) but a bidirectional one.

#### 3.2 Explanations for Analogical Reasoning

Following the above guidelines, the explanations for the analogical reasoning task should also include three parts:

1. *Abduction*: description of suitable structures for the query;
2. *Mapping*: how the structure is mapped onto candidates, analogous to template-filling;
3. *Validation*: justification for the correctness of the counterfactual mapping.

To this end, we define *free-text explanation* for analogical reasoning, which is one of the most expressive and commonly-used explanations (Wiegrefe and Marasović, 2021). We ensure the free-text explanations are self-contained, knowledge-rich, and

sufficient to solve the problem as a substitute for the original input.

Specifically, for each query ( $Q$ ) and candidate ( $A_i$ ), we define free-text explanations  $\mathcal{E}_Q$  and  $\mathcal{E}_{A_i}$ . Following the guidelines in §3.1,  $\mathcal{E}_Q$  should describe the best suited inherent structure of a query abduced from the problem.  $\mathcal{E}_{A_i}$  should decide the correctness in mapping the counterfactual  $A_i$  into structure expressed in  $\mathcal{E}_Q$ , while providing facts as support evidence.

## 4 The E-KAR Benchmark

### 4.1 Dataset Collection

We build our dataset upon the publicly available problems of Civil Service Exams of China (CSE), which is a comprehensive test for candidates’ critical thinking and problem-solving abilities. CSE consists of problems that test various types of reasoning skills, such as graphical reasoning, logical reasoning and comprehension (Liu et al., 2020b), analogical reasoning, etc.

We collect in total 1,655 Chinese analogical reasoning problems from CSE over the years, each of them consisting of a query term tuple and *four* candidate answer tuples of terms (as shown in Figure 1). One of the prominent features in CSE problems is the intensive involvement of commonsense, encyclopedic, and idiom knowledge. For example, one needs to be aware of the fact that “the tide is caused by both Lunar gravity and Solar gravity”. More importantly, one needs to know a *negated fact* (Barker and Jago, 2012; Hossain et al., 2020; Hosseini et al., 2021) in order to reject a candidate, such as the fact that “husband is *not* a job” or “a car is *not* made of tires”. We keep mainly those requiring knowledge and reasoning skills. The rest is manually removed, such as the ones testing mathematics, morphology, and phonics, as well as the problems with the number of terms larger than three.

### 4.2 Manual Annotation of Explanations

We work with a private company for annotating the explanations defined in §3.2. Before annotation starts, we conduct a training session for all annotators to fully understand the requirements and pick the capable ones based on a selection test. The selected workers are allocated into two teams, a team of explanation constructors and a team of checkers, where the checkers achieves better scores in the test. All of them are paid above the local minimum


Dataset	Lang.	Data Size (train / val / test)	# of Terms in Cand.	Has Expl.
SAT	En	0 / 37 / 337	2	✗
Google	En	0 / 50 / 500	2	✗
BATS	En	0 / 199 / 1,799	2	✗
 E-KAR	Zh	1,155 / 165 / 335	2 <sup>(64.5%)</sup> , 3 <sup>(35.5%)</sup>	✓
	En	870 / 119 / 262	2 <sup>(60.5%)</sup> , 3 <sup>(39.5%)</sup>	✓

Table 1: Comparison between **E-KAR** and previous analogy benchmarks: language, data sizes in different splits, number of terms in a query or candidate answer, and whether the benchmark has explanations.

wage. The annotation consists of two stages: 1) the construction stage for writing explanations, and 2) the double-check stage for quality control.

**Construction** During annotation, each problem is assigned to a constructor to build five sentences of explanations: one for query and four for candidate answers. The explanations are required to be: 1) fluent and factually correct, 2) able to solve the problem on their own, and 3) knowledge-rich. To reduce the labeling difficulty, we allow them to use the search engine for querying the Internet.

**First-round Checking** Afterward, a problem with five annotated explanations is fed to a checker for a first-round checking. The checker decides whether to accept an explanation sentence according to the criteria in the construction stage. The rejected ones are sent back to the construction team for revision along with reasons to reject, which serve to re-train the construction team. The process repeats until a batch reaches 90% accuracy (i.e., decided to be correct according to the checker). Then, a second-round checking initiates.

**Second-round Checking** A verified batch is presented to authors for double-checking. Authors conduct random inspections for 50% samples of a batch, and unqualified annotations are sent back with reasons to the check team to fine-tune their checking criteria, which in turn regularize the construction team. The process also repeats until a batch reaches 95% accuracy.

In the end, the authors manually calibrate every explanation and acquire 1,655 analogical problems and a total number of 8,275 ( $5 \times 1,655$ ) free-text explanations, with an average of 31.9 Chinese characters per sentence.



### 4.3 Bilingual E-KAR: English and Chinese

For a broader impact of this work, we also build an English version of E-KAR via translation.

To translate the Chinese E-KAR into English, we ask three Chinese undergraduate students majoring in English to post-edit the machine-translated results of E-KAR by Google. Besides translation fluency, we also make sure that 1) terms in options and explanations have the same word stems; 2) the parts of speech of terms in a query or candidate answer are encouraged to be the same.

However, in practice, we notice that some samples in the Chinese dataset can not be accurately translated into English, such as ones involving idioms, poems, and other knowledge of Chinese culture. Such samples could be hard for non-Chinese people and models to understand without culture-specific knowledge. Therefore, in the English E-KAR, we manually remove or rewrite these samples, resulting in 1,251 problems and 6,255 ( $5 \times 1,251$ ) explanations that would require mostly commonsense and factual knowledge and reasoning skills that are universal across cultures and languages. Nevertheless, those removed samples are valid ones, and the cultural knowledge within them could be of unique value to the Chinese NLP community. Thus, we keep all samples in the Chinese E-KAR to encourage the research of Chinese NLP.

In the end, we have a bilingual E-KAR for rationalizing analogical reasoning. Both versions of E-KAR are randomly split into training, development, and test set at the ratio of 7:1:2. The statistics of E-KAR as well as comparison between previous benchmarks are reported in Table 1, including SAT (Turney et al., 2003), Google (Mikolov et al., 2013b) and BATS (Gladkova et al., 2016). There are 35.5%/39.5% problems with three terms in E-KAR, whereas previous ones only consist of two, making E-KAR even more challenging.

### 4.4 Shared Tasks in E-KAR

Given input  $\mathcal{X} = (Q, A)$ , the ultimate goal is to make the correct choice  $\mathcal{Y}$ , while producing rational explanations  $\mathcal{E} = \{\mathcal{E}_Q, \mathcal{E}_A = \{\mathcal{E}_{A_i}\}_i\}$ . To this end, we define two shared tasks, *multiple-choice question-answering* (QA) and *explanation generation* (EG), for teaching models how to analogize.

Moreover, to reduce the difficulty of this task as well as follow the structure-mapping process (as in §3), we propose an easier task form of the shared tasks by adding  $\mathcal{E}_Q$  into input  $\mathcal{X}$ . Next, we will

elaborate on these settings.

**Task 1: Analogical QA** The analogical QA task is formulated as  $P_{QA}(\mathcal{Y}|\mathcal{X})$ . The QA task requires an understanding of the relationship between the query and each of the candidates to find the correct answer. For evaluation, we directly use the *accuracy* of multiple-choice QA.

Note that all candidates may be related to the query tuple from certain perspectives. The challenge lies in finding the *most* related one, i.e., to identify the inherent connections and relations between terms in the query and candidates, considering properties such as linguistic features, order of terms, commonsense knowledge, etc. For example, the error for candidate D in Figure 1 can be attributed to the incorrect term order, though three terms follow similar relations as in the query. Hence, the best choice is C.

**Task 2: Explanation Generation** This task aims to produce a *pipelined rationalization* for analogical reasoning, formulated as  $P_{EG}(\mathcal{E}|\mathcal{X})$ . The generated explanations  $\mathcal{E}$  can be further utilized for the analogical QA, i.e.,  $P_{QA}(\mathcal{Y}|\mathcal{X}, \mathcal{E})$ . Note that the EG task does *not* generate post-hoc explanations for the QA task, therefore there will not be any predicted choice labels in the input  $\mathcal{X}$ . Rather, it indicates that the model should make implicit label predictions in explanations (Wiegrefe et al., 2021). The generated explanations can be directly evaluated the same as text generation tasks. Or, indirectly, we can follow a pipelined rationalization paradigm and see how generated explanations can help downstream QA tasks.

**Task Mode: EASY vs. HARD** The abduction of *source structure* (query explanation  $\mathcal{E}_Q$ ) is critical but difficult for making rational analogical reasoning. Therefore, we propose two task modes:

- *HARD mode*: the original setting, where only  $Q$  and  $A$  are available in  $\mathcal{X}$ ;
- *EASY mode*: in addition to  $Q$  and  $A$ ,  $\mathcal{E}_Q$  is allowed as part of the given input  $\mathcal{X}$ .

Essentially, EASY mode sets a much clearer playground for evaluating a system’s ability to validate counterfactuals (as in §3.2): *What if candidate terms follow the structures in the query instead of query terms? Will they hold logically?* Therefore, we believe it to be an important supplement for E-KAR benchmark.

## 5 Methods

In this section, we describe the baseline methods in both QA and EG tasks in EASY and HARD modes. We mainly evaluate some of the state-of-the-art language models for solving tasks in **E-KAR**. Some implementation details are reported in Appendix A.

### 5.1 Baselines for Analogical QA

**Pre-trained Methods** As pre-trained-only baselines, we adopt three static word embeddings that have shown their effectiveness in previous analogy tasks: Word2Vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017). We also test contextualized embeddings from PLMs, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). The averaged token representation is taken as the term representation. A query or a candidate is estimated by the sum of the representations of each term pair, which is represented as the embedding vector differences (Hakami and Bollegala, 2017; Ushio et al., 2021). The candidate with the highest cosine similarity to the query is chosen as the answer.

**Fine-tuned Methods** We also set up fine-tuned baselines for QA with PLMs (BERT and RoBERTa). Since previous benchmarks do not have a training set, we only fine-tune the models on their development set. The query and candidates are respectively *verbalized* into text using simple prompts, and an example prompt can be found in Appendix A.1. Each candidate is concatenated with the query into one sentence, which is fed into a PLM for contextualized representation learning. Averaged hidden states are then fed to an MLP layer and a softmax layer for classification.

**Human Evaluation** We ask three students to solve the QA task in **E-KAR**, who are undergraduate or graduate students and fluent in English and Chinese. We randomly sample 100 problems from **E-KAR** of each language. Subjects are asked to first solve them in HARD mode then in EASY mode, in order to reveal the change in performance of the same problem when prompted with the query explanation. The averaged score is reported as the human baseline.

### 5.2 Baselines for Explanation Generation

We formulate the EG task in a Seq2Seq paradigm, instantiated with state-of-the-art pre-trained lan-

guage models for Seq2Seq tasks, including BART (Lewis et al., 2020; Shao et al., 2021) and T5 (Rafael et al., 2020; Zhang et al., 2021).

Although the explanation is individually specific to each query and candidate, the generator has to take into account the whole problem for generating with the *best* source structure (as in §3.1) and thus finding the most analogous candidate. Similar to fine-tuned methods in QA task, the EG model takes as input the concatenation of the query  $Q$  and all candidate answers  $A$  (and the query explanation  $\mathcal{E}_Q$  if in EASY mode). Note that in HARD mode, we switch the prefix of input from generating for  $Q$  or  $A_i$  in order to distinguish between generating explanations for the query or candidate answer. An example prompt is presented in Appendix A.1.

**Evaluation for the EG Task** In HARD mode, both the generated explanations for query  $\mathcal{E}_Q$  and candidate answers  $\mathcal{E}_A$  should be evaluated. In EASY mode, since  $\mathcal{E}_Q$  is fed into the model as input, only  $\mathcal{E}_A$  are required for evaluation. The generated text can be evaluated with text generation metrics such as ROUGE (Lin, 2004), BERTScore<sup>2</sup> (Zhang et al., 2020), BLEURT (Sellam et al., 2020) and MoverScore (Zhao et al., 2019). However, we would like to highlight that great challenges remain for automatically evaluating semantic-rich text generation (Celikyilmaz et al., 2020).

We also follow the pipelined rationalization paradigm and calculate the gain on QA accuracy as a supplement evaluation metric, i.e., the accuracy drop of  $P_{QA}(\mathcal{Y}|\mathcal{X}, \mathcal{E})$  over  $P_{QA}(\mathcal{Y}|\mathcal{X}, \mathcal{E}_{gold})$ . This metric is denoted as **Acc** ( $\Delta$ ), where **Acc** is the QA accuracy when including generated explanations  $\mathcal{E}$  as input during inference, and  $\Delta$  reflects the accuracy drop. Here we fix a trained QA model  $P_{QA}(\cdot)$  based on a large-version RoBERTa. This model is designed to be *different* from the ones in the QA task, as it is fine-tuned by concatenating gold explanations to the corresponding query or candidates as input during training (prompt detail can be found in Appendix A.1). As an evaluation metric, we alter the input explanations to the model from *gold*  $\mathcal{E}$  to *generated*  $\mathcal{E}$ , and see their performance drops over gold. Note that the query explanation  $\mathcal{E}_Q$  is still the input for all settings in EASY mode.

<sup>2</sup>We use the code of BERTScore at [https://github.com/Tiiiger/bert\\_score](https://github.com/Tiiiger/bert_score), where English BERTScore is based on a RoBERTa (large) and Chinese one is based on a BERT (base).

## 6 Results and Analysis

In the experiments, we wish to answer two questions: *Q1*) Can models do knowledge-intensive analogical QA? *Q2*) Can models generate rational reasons for analogical thinking?

**Categorization of Problems** We first manually categorize the relational types of problems in **E-KAR** according to a pre-defined schema. Unlike free text, we are unable to induce a comprehensive set of relations that covers all candidates due to the complexity of CSE problems. As a result, we carefully assign at least one relation to each query. To facilitate analysis, we also try to assign relations to each candidate and query *in the development and test set*, ending up covering 76% of the candidates and 100% of the queries.

We refer to several sources of word analogy definitions and textbooks for analogy tests (listed in Appendix B), and categorize the relations into five *meta-relations* (as well as their coverage in the test set) and several accompanying *sub-relations*:

1. *Semantic* (R1, 8.36% for Zh, 4.12% for En), the similarity or difference in the meaning of terms, including *synonym\_of* and *antonym\_of*;
2. *Extension* (R2, 41.25% for Zh, 42.30% for En), the relation between the extension of terms, including *is\_a*, *contradictory\_to*, etc.;
3. *Intension* (R3, 37.94% for Zh, 40.21% for En), terms relate to each other by inherent properties, including *made\_of*, *has\_function*, etc.;
4. *Grammar* (R4, 6.36% for Zh, 6.72% for En), the grammatical relations between terms, including *subject-predicate*, *head-modifier*, etc.;
5. *Association* (R5, 6.08% for Zh, 6.65% for En), logical association between terms, including *result\_of*, *sufficient\_to*, etc.

Complete sub-relations are presented in Appendix B, as well as their definitions and examples.

### 6.1 Can models do knowledge-intensive analogical reasoning?

Table 2 reports the accuracy results of baseline methods on previous analogy tasks and the QA task in **E-KAR**.

**How do machines solve analogical reasoning problems?** To answer this question based on Table 2, the findings can be summarized as:

Method	SAT	Google	BATS	E-KAR (H/E)	
				Zh	En
<i>Pre-trained Word Embeddings</i>					
Word2Vec <sup>†</sup>	41.5	93.2	63.9	28.2/-	25.6/-
GloVe <sup>†</sup>	<b>47.7</b>	96.0	67.6	30.9/-	27.8/-
FastText <sup>†</sup>	47.1	<b>96.6</b>	<b>72.0</b>	<b>31.4/-</b>	<b>28.2/-</b>
<i>Pre-trained Language Models</i>					
BERT <sub>b</sub> <sup>†</sup>	32.9	80.8	61.5	34.5/-	30.4/-
RoBERTa <sub>b</sub> <sup>†</sup>	42.4	90.8	69.7	41.7/-	37.4/-
RoBERTa <sub>l</sub> <sup>†</sup>	<b>45.4</b>	<b>93.4</b>	<b>72.2</b>	<b>44.6/-</b>	<b>39.0/-</b>
<i>Fine-tuned Language Models</i>					
BERT <sub>b</sub>	38.9	86.6	68.0	41.8/46.7	37.9/42.2
RoBERTa <sub>b</sub>	47.7	93.8	75.2	46.9/51.1	42.2/48.1
RoBERTa <sub>l</sub>	<b>51.6</b>	<b>96.9</b>	<b>78.2</b>	<b>50.1/54.8</b>	<b>46.7/50.5</b>
Human	-	-	-	77.8/83.3	

Table 2: Accuracy results on previous analogy tasks and the QA task in **E-KAR**. **E-KAR** (H/E) denotes HARD or EASY mode of analogical QA. Method<sup>†</sup> is not tuned. PLM<sub>b</sub> or PLM<sub>l</sub> denote *base* or *large* version, respectively.

1) We find contextualized word embeddings from PLMs not very competitive against static word embeddings in previous analogy tasks, which is consistent with the findings in Peters et al. (2018).

2) In a more knowledge-intensive **E-KAR**, the opposite conclusion can be made, with PLMs prevailing over static word embeddings.

3) Furthermore, performance from contextualized representations can be improved in all tasks through fine-tuning, especially for **E-KAR**, where accuracy increases by roughly 5 to 6 points.

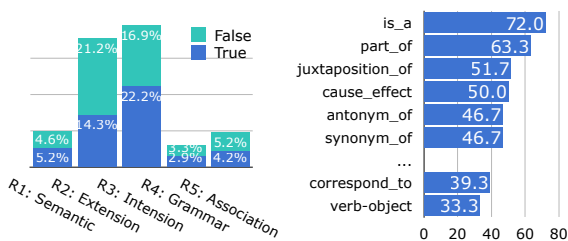
4) When incorporating gold source structure (i.e., EASY mode), the QA results significantly improve by roughly 5 points in both languages.

5) Moreover, despite our efforts to eliminate culture-specific samples in English **E-KAR**, the accuracy still falls behind its Chinese counterpart, which could be attribute to: *a*) fewer training samples, *b*) language-specific pre-training and *c*) language-specific information noise by translation.

**How do humans solve analogical reasoning problems?** In contrast to machines, humans achieve in **E-KAR** 77.8% accuracy in HARD mode and 83.3% in EASY mode, indicating the challenge of this task as well as showing that current SOTA language models still fall far behind human performance. We also find the trend of human performance is generally aligned with machines, with accuracy boost (also  $\sim 5$  points) when prompted with query explanations.

EG Method	E-KAR (Zh)					E-KAR (En)				
	ROUGE	BERT.	BLRT.	Mover.	Acc $\uparrow$ ( $\Delta$ $\downarrow$ )	ROUGE	BERT.	BLRT.	Mover.	Acc $\uparrow$ ( $\Delta$ $\downarrow$ )
None ( $\times$ )	N/A	N/A	N/A	N/A	29.1 (68.6)	N/A	N/A	N/A	N/A	25.6 (72.1)
BART <sub>b</sub> ( $\times$ )	39.85	72.68	63.43	64.72	33.0 (64.7)	17.71	91.27	54.40	59.91	29.0 (68.7)
BART <sub>i</sub> ( $\times$ )	40.39	72.67	63.60	64.57	<b>38.8 (58.9)</b>	18.34	<b>91.54</b>	<b>55.48</b>	60.13	<b>34.1 (67.6)</b>
T5 <sub>b</sub> ( $\times$ )	<b>43.37</b>	<b>83.17</b>	<b>66.34</b>	<b>75.92</b>	30.7 (67.0)	17.44	91.17	53.71	60.40	25.6 (72.1)
T5 <sub>i</sub> ( $\times$ )	-	-	-	-	-	<b>19.77</b>	91.44	55.00	<b>60.78</b>	29.4 (68.3)
None ( $\checkmark$ )	N/A	N/A	N/A	N/A	30.5 (67.2)	N/A	N/A	N/A	N/A	26.7 (71.0)
BART <sub>b</sub> ( $\checkmark$ )	39.08	72.84	62.10	65.07	33.4 (64.3)	25.14	91.85	56.16	62.16	29.8 (67.9)
BART <sub>i</sub> ( $\checkmark$ )	39.18	72.93	62.45	65.13	<b>36.1 (61.6)</b>	25.31	91.92	56.14	62.26	<b>32.4 (65.3)</b>
T5 <sub>b</sub> ( $\checkmark$ )	<b>40.04</b>	<b>82.52</b>	<b>63.54</b>	<b>74.99</b>	34.0 (63.7)	26.59	92.12	57.39	63.01	30.2 (67.5)
T5 <sub>i</sub> ( $\checkmark$ )	-	-	-	-	-	<b>28.10</b>	<b>92.38</b>	<b>58.76</b>	<b>63.64</b>	31.3 (66.4)
Gold	N/A	N/A	N/A	N/A	97.7 (0.0)	N/A	N/A	N/A	N/A	97.7 (0.0)

Table 3: Results of explanation generation models w.r.t. ROUGE-2, BERTScore, BLEURT, MoverScore and Acc ( $\Delta$ ) on the analogical QA task, where EASY mode ( $\checkmark$ ) incorporates gold  $\mathcal{E}_Q$  as part of the model input. Note that the QA model here is trained as described in §5.2, and we switch input explanations during inference.



(a) Meta-relations distributions and their error ratios. (b) Sub-relations in a sorted order of error rate.

Figure 2: Error analysis of different query relations. The results are predicted by a fine-tuned RoBERTa (large) in §5.1 on E-KAR (Zh).

**Error Analysis for QA** We further conduct an error analysis based on the results in E-KAR (Zh) predicted by a fine-tuned RoBERTa (large). The erroneous ones are classified based on the manually annotated meta-relations and sub-relations of *queries*, which is a fine-grained tool for analyzing a model’s predictions.

Figure 2(a) shows that the model performs poorly on nearly all meta-relations, with R2 (Extension) being the most error-prone one (only 40.3% accuracy, normalized) and R3 (Intension) being the least one (56.8% accuracy). One of the most prominent reasons is that R2 and R3 rely heavily on commonsense and encyclopedic knowledge and reasoning skills such as commonsense and world knowledge, at which current models easily fail.

Figure 2(b) shows the error rate of sub-relations with more than 10 samples. Consistent with Figure 2(a), the three most error-prone sub-relations (*is\_a*, *part\_of* and *juxtaposition\_of*) all belong to R2 (Extension). Besides, the model seems to do well in linguistic knowledge, with *verb-object* achieving only 33.3% error rate. These findings may shed

light on future directions for knowledge-intensive reasoning with language models.

## 6.2 Can models rationalize analogical thinking?

We report the automatic evaluation results of generated explanations in Table 3. However, such results hardly mean anything due to the incapability to evaluate the semantic-rich text of current automatic metrics. Therefore, the following analyses mainly focus on Acc ( $\Delta$ ) and human evaluation.

**Can (generated) explanations benefit analogical QA?** To start with, we highlight again that the QA model in Table 3 is different from the one in Table 2 since the training of the former involves gold explanations. When exposing gold explanations to the QA model, it achieves 97.7% accuracy on E-KAR of both languages coincidentally.

However, the QA model performs poorly when removing the explanations during inference (i.e., *None*). This is because the pipelined rationalization in training makes the QA model rely heavily on the rationales (explanations) than the problem itself, and the removal of them causes severe performance degradation. When we switch the explanations to generated ones during inference, the accuracy gap ( $\Delta$ ) between gold results slightly narrows, with the gain in EASY mode being more significant than in HARD mode. To conclude, current SOTA generative language models still fall short of rationalizing analogical reasoning, which would be a challenging but interesting future direction.

**Error Analysis for EG** We also randomly select 100 sentences generated by a BART (large) for manual inspection by the authors. Aside from



Q)	氧气 (oxygen):臭氧 (ozone)
A)	盐 (salt):氯化钠 (sodium chloride)
B)	硫酸 (sulfuric acid):硫 (sulfur)
C)	石墨 (graphite):金刚石 (diamond)
D)	石灰水 (lime water):氢氧化钙 (calcium hydroxide)
$\mathcal{E}_Q$	氧气和臭氧都只由氧元素组成。Both oxygen and ozone are made of only the oxygen element.
$\mathcal{E}_Q^\dagger$	臭氧是氧气的一种。Ozone is a kind of oxygen.
$\mathcal{E}_A$	氯化钠是盐的主要成分，盐和氯化钠不是只由一种元素组成。Sodium chloride is the main component of salt. Neither salt nor sodium chloride is made of only one element.
$\mathcal{E}_A^\dagger$	氯化钠是盐的一种。Sodium chloride is a kind of salt.

Table 4: Case study of EG in HARD mode, where  $\mathcal{E}_*$  is gold and  $\mathcal{E}_*^\dagger$  is generated by a BART (large).

the common errors in generation models such as repetition, we find that task-specific errors for generated explanations can be roughly categorized into three classes: 1) *unable to generate negated facts to refute source structure*; 2) *generating factually incorrect statements*; 3) *biasing towards common patterns*, e.g., “term 1 and term 2 have similar meanings” and “term 1 is a term 2”. For example, in Table 4, both generated  $\mathcal{E}_Q$  (only in HARD mode) and  $\mathcal{E}_A$  are factually incorrect, and the model fails to generate the negated fact that “both are not exclusively made of one component.”

We dig further into the first class of errors (w.r.t. negation), which is important to refute a candidate, as mentioned in §4.1. We find  $\sim 90\%$  gold explanations of wrong candidates contain negated statements. Yet, the number drops to 14.9% (Zh) and 22.1% (En) in the generated ones in HARD mode, and 21.3% (Zh) and 38.6% (En) in EASY mode. An interesting conclusion can be drawn that current generative models do not seem to know how to generate a negated yet truthful fact, such as “feeling can *not* guide psychological reaction.” since feeling *is* a reaction. And exposing source structure to the model (EASY mode) seems to alleviate this problem.

The fact also questions the astonishing QA performance by adding gold explanations (97.7%), as the model could be biased towards surface-level negation. To debias this, we conduct a simple ablation study by directly removing the clauses containing the negation word “不” (*not*) from the gold explanations in the test set, and still achieve 92.5% in QA accuracy. This finding indicates that the QA model with correct rationales would not be very much biased towards negation in the explanation.

## 7 Conclusion and Discussion

In this work, we propose a first-of-its-kind benchmark **E-KAR** (in both Chinese and English) for explainable analogical reasoning, which sets a concrete playground and evaluation benchmark to boost the development of human-like analogical reasoning algorithms. The **E-KAR** benchmark is featured by its rich coverage in knowledge and well-designed free-text explanations to rationalize the analogical reasoning process. Preliminary experiments show that this benchmark provides a rather difficult challenge for prevailing language models.

However, there are still many open questions to be addressed. For example, humans solve the analogy problems in a trial-and-error manner, i.e., adjusting the abduced source structure and trying to find the most suited one for all candidate answers. However, the explanation annotation process in **E-KAR** (not the EG task) is mostly post-hoc and reflects only the result of reasoning. Such explanations cannot offer supervision for intermediate reasoning, though it is an interesting question whether an intelligent model should be deeply supervised at every step (Tafjord et al., 2021). Furthermore, **E-KAR** only presents one feasible explanation for each problem, whereas there may be several.

This benchmark also invites reasoning models that can effectively interact with extra knowledge. It remains to be a great challenge to generate and evaluate factually correct explanation text. Especially, how to generate negated facts is relatively under-explored in the research community but of much importance. Finally, whether the analogical QA system can *correctly* exploit explanations and background knowledge is also worth investigating, which may intersect with research on debiasing (Tang et al., 2020; Niu et al., 2021).

We hope this work to be a valuable supplement to future research on natural language reasoning, especially for research on analogical reasoning and explainable NLP.

## Acknowledgement

We thank the anonymous reviewers for their valuable suggestions. We also thank Ruxin Yu for the logo design. This work was supported by National Key Research and Development Project (No. 2020AAA0109302), Shanghai Science and Technology Innovation Action Plan (No.19511120400) and Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

## Ethical Considerations

This paper proposes a new kind of analogical benchmark with explanations to rationalize models' predictions. The dataset is collected from Civil Service Exams of China, which is publicly available and has been used in other public datasets before, such as LogiQA (Liu et al., 2020a). The annotated explanations for each problem in our dataset are crowd-sourced by working with ByteDance. The construction team remains anonymous to the authors, and the annotation quality is guaranteed by the double-check strategy as mentioned in §4.2. We ensure that all annotators' privacy rights are respected in the annotation process. All annotators have been paid above local minimum wage and consented to use the datasets for research purposes covered in our paper.

## References

- Shourya Aggarwal, Divyanshu Mandowara, Vishwa-jeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. [Explanations for CommonsenseQA: New Dataset and Models](#). 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 3050–3065, Online. Association for Computational Linguistics.
- Stephen Barker and Mark Jago. 2012. Being positive about negative facts. *Philosophy and Phenomenological research*, pages 117–138.
- Paul Bartha. 2013. Analogy and analogical reasoning.
- Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. 2021. [Deep learning for ai](#). *Commun. ACM*, 64(7):58–65.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. [e-snli: Natural language inference with natural language explanations](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. [Make up your mind! adversarial generation of inconsistent natural language explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.
- 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*.
- Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. [Transformers as Soft Reasoners over Language](#). pages 3882–3890.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. [Revisiting pre-trained models for Chinese natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.
- 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.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. [Towards understanding linear word analogies](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.
- Jerome A Feldman and Dana H Ballard. 1982. Connectionist models and their properties. *Cognitive science*, 6(3):205–254.

- Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.
- Dedre Gentner and Linsey Smith. 2012. Analogical reasoning. *Encyclopedia of human behavior*, 2:130–136.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Mary L Gick and Keith J Holyoak. 1983. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38.
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.
- Ashok K Goel. 1997. Design, analogy, and creativity. *IEEE expert*, 12(3):62–70.
- Shangmin Guo, Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2017. Which is the effective way for gaokao: Information retrieval or neural networks? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 111–120.
- Huda Hakami and Danushka Bollegala. 2017. Compositional approaches for representing relations between words: A comparative study. *Knowledge-Based Systems*, 136:172–182.
- Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.
- Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordani, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.
- Zixian Huang, Yulin Shen, Xiao Li, Gong Cheng, Lin Zhou, Xinyu Dai, Yuzhong Qu, et al. 2019. Geosqa: A benchmark for scenario-based question answering in the geography domain at high school level. 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 5866–5871.
- Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. Association for Computational Linguistics.
- Philip Nicholas Johnson-Laird. 2006. *How we reason*. Oxford University Press, USA.
- 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.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- 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.
- Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.
- 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.
- Peng-Hsuan Li, Tsan-Yu Yang, and Wei-Yun Ma. 2020. CA-EHN: Commonsense analogy from E-HowNet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2984–2990, Marseille, France. European Language Resources Association.
- Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018a. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*



- Papers*), pages 138–143, Melbourne, Australia. Association for Computational Linguistics.
- Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018b. [Analogical reasoning on chinese morphological and semantic relations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.
- 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.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020a. [Logiqa: A challenge dataset for machine reading comprehension with logical reasoning](#). In *IJCAI*.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020b. [Logiqa: A challenge dataset for machine reading comprehension with logical reasoning](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.
- 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*.
- Wei-Yun Ma and Yueh-Yin Shih. 2018. [Extended HowNet 2.0 – an entity-relation common-sense representation model](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. [Distributed representations of words and phrases and their compositionality](#). In *Advances in neural information processing systems*, pages 3111–3119.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. [Linguistic regularities in continuous space word representations](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- Gerhard Minnameier. 2010. [Abduction, induction, and analogy](#). In *Model-based reasoning in science and technology*, pages 107–119. Springer.
- Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. [Counterfactual vqa: A cause-effect look at language bias](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.
- Charles S Peirce. 1896. [Lessons from the history of science](#). *C. Hartshorne*, 660.
- 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, Luke Zettlemoyer, and Wen-tau Yih. 2018. [Dissecting contextual word embeddings: Architecture and representation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 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](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Explain yourself! leveraging language models for commonsense reasoning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.
- Carissa Schoenick, Peter Clark, Oyvind Tafjord, Peter Turney, and Oren Etzioni. 2017. [Moving beyond the turing test with the allen ai science challenge](#). *Communications of the ACM*, 60(9):60–64.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. [Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation](#). *arXiv preprint arXiv:2109.05729*.



- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. [ProofWriter: Generating implications, proofs, and abductive statements over natural language](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.
- Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. [GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, Online. Association for Computational Linguistics.
- Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. [Long-tailed classification by keeping the good and removing the bad momentum causal effect](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1513–1524. Curran Associates, Inc.
- Paul Thagard. 1992. Analogy, explanation, and education. *Journal of Research in science Teaching*, 29(6):537–544.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Peter D Turney, Michael L Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules in lexical multiple-choice problems. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP, 2003*:101–110.
- Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. [BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies?](#) 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 3609–3624, Online. Association for Computational Linguistics.
- Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2021. From Isat: The progress and challenges of complex reasoning. *arXiv preprint arXiv:2108.00648*.
- Sarah Wiegrefe and Ana Marasović. 2021. [Teach me to explain: A review of datasets for explainable nlp](#). In *Proceedings of NeurIPS*.
- Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. 2021. [Measuring association between labels and free-text rationales](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Weihaoyu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. [Mengzi: Towards lightweight yet ingenious pre-trained models for chinese](#).
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. [MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance](#). 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 563–578, Hong Kong, China. Association for Computational Linguistics.

## A Implementation Details

The pre-trained word embeddings are provided by Li et al. (2018b), and the checkpoints for PLMs are hosted in HuggingFace (Wolf et al., 2020). Most of the parameters in the baseline models take the default values from HuggingFace’s Transformers library, and we keep the best checkpoint on the validation set for testing. The Chinese version of BERT (whole word masking) and RoBERTa (whole word masking extended) are provided by Cui et al. (2020), BART by Shao et al. (2021) and T5 by Zhang et al. (2021).<sup>3</sup> Thus the EG results of T5 in **E-KAR** (zh) can be attributed to both Raffel et al. (2020) and Zhang et al. (2021).

### A.1 Example Prompts in E-KAR

We denote terms in a query  $Q$  or a candidate  $A^* \in \{A, B, C, D\}$  as  $t_{Q/A^*}^{\{1,2\}}$ . The example prompts for the QA and EG tasks in **E-KAR** are:

- *A Prompt for the QA Task:* “(context:  $\mathcal{E}_*$ ,) question:  $t_Q^1 : t_Q^2$ , options:  $t_A^1 : t_A^2, t_B^1 : t_B^2, \dots$  or  $t_D^1 : t_D^2$ ”.
- *A Prompt for the EG Task:* “query =  $t_Q^1 : t_Q^2$  </s> (query explanation =  $\mathcal{E}_Q$ ) </s> candidate =  $t_A^1 : t_A^2$  </s> candidate =  $t_B^1 : t_B^2$  </s>  $\dots$  </s> candidate =  $t_D^1 : t_D^2$  </s> generate the explanation of  $Q/A_i$ :”.
- *A Prompt for the QA model in Acc  $\Delta$ :* concatenating explanations to the query and each candidate answer, such as “ $t_Q^1 : t_Q^2$  </s> explanation:  $\mathcal{E}_Q$ ” and “ $t_A^1 : t_A^2$  </s> explanation:  $\mathcal{E}_A$ ”.

## B Detailed Relation Definitions

To design the relation taxonomy, we refer to a number of sources that categorize types of analogy tests, including MAT<sup>4</sup>, Fibonacci<sup>5</sup>, Offcn Education (in Chinese)<sup>6</sup> and Huatu Education (in Chinese)<sup>7</sup>, etc.

The complete set of meta-relations and sub-relations are presented in Table 5.

<sup>3</sup>Note that the Chinese T5 (Mengzi) does not have large version, as they claim to be lightweight but ingenious.

<sup>4</sup>[http://www.west.net/~stewart/mat/analogy\\_types.htm](http://www.west.net/~stewart/mat/analogy_types.htm)

<sup>5</sup><https://www.fibonacci.com/verbal-reasoning/analogy-examples/>

<sup>6</sup><https://www.offcn.com>

<sup>7</sup><https://www.huatu.com>

Relation	Definition	Example	Coverage	
			Zh	En
<b>R1: Semantic</b>			<b>8.36%</b>	<b>4.12%</b>
1) <i>synonym_of</i>	The meanings of two terms are similar.	clarity : transparency	4.88%	2.37%
2) <i>antonym_of</i>	The meaning of two terms are opposite or used to express different concepts.	harmony : conflict	3.48%	1.75%
<b>R2: Extension</b>			<b>41.25%</b>	<b>42.30%</b>
1) <i>identical_to</i>	The meanings of two terms are identical.	highway : road	1.64%	0.92%
2) <i>is_a</i>	One term is the hypernym of the other.	Earth : planet	11.54%	12.38%
3) <i>part_of</i>	One term is a part of the other.	steering wheel : sedan	6.82%	7.78%
4) <i>juxtaposition_to</i>	Two terms belong to the same hypernym or have the same properties or functions.	shoes : socks	12.86%	12.62%
5) <i>contradictory_to</i>	Two term are contradictory to each other.	vowel : consonant	1.19%	1.25%
6) <i>contrary_to</i>	Two propositions cannot both be true, but can both be false.	black : white	4.36%	4.08%
7) <i>intersection_to</i>	The extension of the two terms intersects.	solo : pianolude	2.45%	2.81%
8) <i>utterly_different</i>	The extensions of terms do not overlap.	apple : nuts	0.39%	0.46%
<b>R3: Intension</b>			<b>37.94%</b>	<b>40.21%</b>
1) <i>attribute_of</i>	One term is the attribute of the other.	object : inertia	1.15%	1.17%
2) <i>probabilistic_attribute</i>	One term is probably the attribute of the other.	shoes : high heels	0.33%	0.34%
3) <i>has_function</i>	One term has the function of the other.	calculator : calculate	2.94%	3.54%
4) <i>metaphor</i>	A term is the metaphor of the other, reflecting something abstract indirectly.	pigeon : peace	1.15%	0.42%
5) <i>takes_place_in</i>	A term takes place in the other.	soldier : battlefield	0.96%	1.07%
6) <i>located_in</i>	A term is located in the other.	Rhine : Europe	2.06%	2.47%
7) <i>made_of</i>	One term is the raw material of the other.	door : wood	3.21%	3.90%
8) <i>tool_of</i>	One term is the tool of the other.	knives : murder	0.91%	1.00%
9) <i>target_of</i>	One term is the target of the other.	health : exercise	0.82%	0.72%
10) <i>corresponds_to</i>	Terms generally correspond to each other.	post office : mail bank	24.41%	25.58%
<b>R4: Grammar</b>			<b>6.36%</b>	<b>6.72%</b>
1) <i>subject-predicate</i>	The originator of the action and the action itself.	plane : take off	1.19%	1.25%
2) <i>verb-object</i>	The action and the object on which the action acts.	transfer : goods	3.14%	3.36%
3) <i>head-modifier</i>	The preceding term modifies the other.	affluence : living	0.87%	0.74%
4) <i>subject-object</i>	The originator and receiver of an action.	dairy farmer : milk	1.16%	1.37%
<b>R5: Association</b>			<b>6.08%</b>	<b>6.65%</b>
1) <i>result_of</i>	One term causes the other.	lack of water : plants wither	2.99%	2.97%
2) <i>follow</i>	The terms have a chronological or other sequential relationship, but one term does not cause the other.	sign up : take the exam	1.91%	2.19%
3) <i>sufficient_to</i>	One term is a sufficient condition for the other.	raining : wet ground	0.0%	0.0%
4) <i>necessary_to</i>	One term is a necessary condition for the other.	admission : graduation	1.18%	1.49%

Table 5: Complete set of defined sub-relations with definitions, examples and coverage in the test set of **E-KAR**.