ANALOGYKB: Unlocking Analogical Reasoning of Language Models with A Million-scale Knowledge Base

Anonymous ACL submission

Abstract

Analogical reasoning is a fundamental cognitive ability of humans. However, current language models (LMs) still struggle to achieve human-like performance in analogical reasoning tasks due to a lack of resources for model training. In this work, we address this gap by proposing ANALOGYKB, a million-scale analogy knowledge base (KB) derived from existing knowledge graphs (KGs). ANALOGYKB identifies two types of analogies from the KGs: 1) analogies of *the same relations*, which can be directly extracted from the KGs, and 2) analogies of *analogous relations*, which are identified with a selection and filtering pipeline enabled by large language models (LLMs), followed by minor human efforts for data quality control. Evaluations on a series of datasets of two analogical reasoning tasks (analogy recognition and generation) demonstrate that ANALOGYKB successfully enables both smaller LMs and LLMs to gain better analogical reasoning capabilities.[1]

1 Introduction

Making analogies requires identifying and mapping a familiar domain (*i.e.*, source domain) to a less familiar domain (*i.e.*, target domain) (Hofstadter and Sander, 2013). As shown in Figure 1, utilizing the analogy of the solar system can facilitate comprehension of the complex structure of atoms. Analogical reasoning is an important aspect of the cognitive intelligence of humans, allowing us to quickly adapt our knowledge to new domains (Hofstadter, 2001; Ding et al., 2023), make decisions (Hansen-Estruch et al., 2022), and solve problems (Yasunaga et al., 2023). As a result, the topic of analogy has been drawing significant research attention in the community.

However, resources for analogical reasoning are rather limited in scale (Mikolov et al., 2013b; Gladkova et al., 2016; Chen et al., 2022), which usually consist of only hundreds or thousands of data samples. As a result, these datasets do not support effective training of language models to gain analogical reasoning abilities. Although large language models (LLMs) can make some reasonable analogies without requiring gradient update, their performance still lies behind humans (Bhavya et al., 2022; Jiayang et al., 2023). Therefore, larger-scale data sources are needed to facilitate the research in this area. With richer analogies, we can train specialized analogy-making models and retrieve high-quality examples to assist LLMs. Therefore, the research question is: *How to acquire large-scale analogies at a moderate cost?*

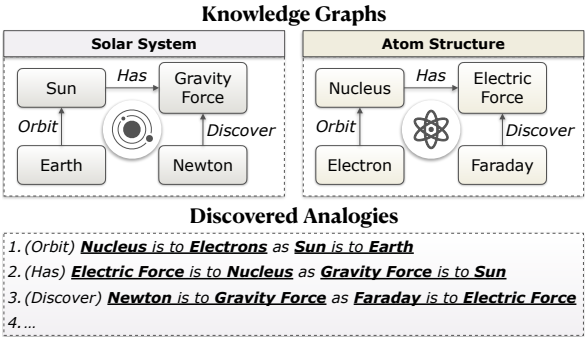


Figure 1: An example of acquiring analogies from KGs. Based on the relational knowledge triples from KGs, *i.e.*, facts about the *solar system* and an *atom structure*, we can discover new analogies using the corresponding relations between concepts.

An analogy is determined by the *relational structure* (Bartha, 2013), *e.g.*, A:B::C:D (*i.e.*, A is to B as C is to D), where the relation between A and B is analogous to the relation between C and D. The *concepts* A, B, C, and D can be entities and events. As shown in Figure 1, the "solar system" and an "atom" share a similar structure, allowing us to quickly grasp the relation between an "electron" and a "nucleus" in concepts of their source domain counterparts. Such relational structure can be de-

[1] Resources of this paper will be released upon publication.

rived from the triplet knowledge, *e.g.* (electron, orbit, nucleus) and (earth, orbit, sun), in knowledge graphs (KGs) (Wu et al., 2012). Therefore, such structure knowledge can be utilized and reorganized to create new analogy knowledge, supporting large-scale knowledge acquisition.

In this work, we aim to build a knowledge base (KB) for storing analogies derived from existing KGs to improve analogical reasoning. However, due to the complicated relational structures, discovering analogies from KGs is not a trivial task. Although two pairs of concepts with the same relation can form a valid analogy (*e.g.*, *lion, isA, animal* and *apple, isA, fruit*), interesting and diverse analogies are implicit in the KGs, with more complex relations. Concepts under two distinct but similar relations in KGs can also form a reasonable analogy (Hesse, 1959). For example, *chief executive officer* and *head of state* can both be abstracted into a *meta relation* (Hesse, 1959; Gentner and Maravilla, 2017), *i.e.*, *head of organization*. Therefore, they are analogous relations under a meta relation. It is important to generalize the finding of implicit analogies beyond the same relations within KGs.

We present ANALOGYKB, which is a large-scale analogy KB. We use Wikidata (Vrandečić and Krötzsch, 2014) and ConceptNet (Speer et al., 2017) as our seed KGs and discover two types of analogies from these KGs: analogies of 1) *same relations* and 2) *analogous relations*. Analogies of the same relations can be directly extracted from existing KGs. In contrast, analogies of analogous relations are more implicit, requiring the finding of relation pairs from the KGs that can form valid analogies. However, it is costly to manually select analogous relation pairs. Therefore, we use InstructGPT$_{003}$ (Ouyang et al., 2022), a LLM of great capabilities in NLP tasks, for finding and deciding the analogical semantics of relations. To eliminate the noise from the outputs of InstructGPT$_{003}$ (§ 3.5), we devise two filtering rules based on 1) the symmetry of analogy and 2) *meta relation* summarization, which generalizes two relations into a more abstract meta relation. Then, we *manually* review the filtered results to further ensure data quality.

Our ANALOGYKB comprises over 1 million analogies with 943 relations, including 103 analogous relations. Smaller LMs trained on ANALOGYKB gain significant improvements over the previous methods, even rivaling human performance on some analogy recognition tasks. Furthermore, we prove that ANALOGYKB can endow both smaller LMs and LLMs with satisfactory analogy-making capabilities. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to construct an analogy KB (ANALOGYKB) with a million scale and diverse relational structures.
- We propose a novel framework with LLMs to discover more interesting and implicit analogies of analogous relations;
- We conduct extensive experiments to evaluate the effectiveness of ANALOGYKB, which significantly improves the analogical reasoning performance of both smaller LMs and LLMs.

2 Related Work

Analogy Acquisition Early studies mainly acquire analogy knowledge via linguists (Turney et al., 2003; Boteanu and Chernova, 2015), which is costly and inefficient. Recent studies consider exploiting relations in KGs to build analogies (Speer et al., 2008; Allen and Hospedales, 2019; Ulčar et al., 2020), which can be divided into two lines of work: *1) Acquiring from commonsense KGs*, which leverages semantic and morphological relations from WordNet (Miller, 1995), ConceptNet (Speer et al., 2017), etc. However, some of these datasets are large-scale but of poor quality (Li et al., 2018, 2020), while others are of high quality but limited in size (Mikolov et al., 2013b; Gladkova et al., 2016). *2) Acquiring from encyclopedia KGs* (Si and Carlson, 2017; Zhang et al., 2022; Ilievski et al., 2022), which utilizes the relations from DBpedia (Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014), but their empirical experiments are relatively small in size.

Analogical Reasoning Analogical reasoning aims to identify a relational structure between two domains (Bartha, 2013; Chen et al., 2022). Previous work adopts the word analogy task to investigate the analogical reasoning capability of LMs (Mikolov et al., 2013a,b; Levy and Goldberg, 2014; Gladkova et al., 2016; Schluter, 2018; Fournier et al., 2020; Ushio et al., 2021). Recent work demonstrates that LLMs can generate some reasonable abstract (Mitchell, 2021; Hu et al., 2023; Webb et al., 2023) and natural language-based

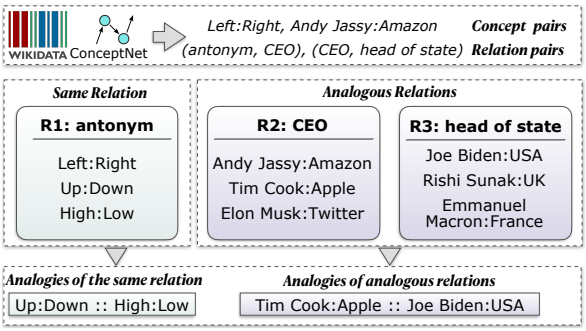


Figure 2: The relations with concept pairs are stored in ANALOGYKB. We define two types of analogies, *i.e.*, analogies of *the same relation* and analogies of *analogous relations*, and derive them from existing KGs.

analogies (Bhavya et al., 2022; Wijesiriwardene et al., 2023; Jiayang et al., 2023) but still lay behind humans in some cases, and smaller LMs struggle to learn analogical reasoning ability due to a lack of training data.

Knowledge Base Construction Knowledge base (KB) consists of structured knowledge to support various applications. The approaches to constructing KBs can be divided into three categories: *1) Manual construction* (Miller, 1995; Speer et al., 2017), which creates the KBs with specialized knowledge written by experts, and thus is labor-intensive; *2) Automatic construction* (Wu et al., 2012; Martinez-Rodriguez et al., 2018), which leverages models to extract knowledge from unstructured corpora, may lead to low data quality; *3) Semi-automatic construction* (Dalvi Mishra et al., 2017; Romero and Razniewski, 2020), which involves manual curation and annotation. Our work is based on automatic approaches with LLMs only requiring small-scale human checking efforts.

3 ANALOGYKB Construction

This section details the framework for building ANALOGYKB. We first define the schema of ANALOGYKB (§ 3.1). Then, we collect relations with concept pairs from existing KGs (§ 3.2, Step 1) and directly obtain analogies of the same relations from KGs (§ 3.3, Step 2). We propose adopting LLMs (Ouyang et al., 2022) followed by minor human efforts to acquire analogies of analogous relations (§ 3.4, Step 3).

3.1 Schema for Analogies in ANALOGYKB

This paper focuses on the analogy formed as A:B::C:D, where *concepts* as A, B, C and D can be entities or events. The *concept pair* A:B is analogous to C:D based on an underlying relational structure. Since ANALOGYKB is built on existing KGs, we define two types of that relational structure based on KG semantics: 1) *analogies of the same relation* and 2) *analogies of analogous relations*. Data in ANALOGYKB is organized as in Figure 2, where each relation R contains subject-object *concept pairs* $s : o$. Within each relation, analogies of the same relation can be naturally formed, *e.g.*, "*Up is to Down as High is to Low*". Also, the concept pairs between two relations can form analogies, as long as the *relation pair* have analogous structures (Hesse, 1959). For example, "*Tim Cook is to Apple as Joe Biden is to USA*", where R2 (CEO) is analogous to R3 (head of state). Therefore, ANALOGYKB only has to store concept pairs of each relation and analogous relation pairs, from which analogies can be easily derived. We list the definitions of each terminology with examples in Appendix A.1 for better understanding.

3.2 Source Data Collection

We choose the two most-used KGs, *i.e.*, ConceptNet and Wikidata consisting of high-quality concept pairs with relations, as our data sources. For ConceptNet, we select the concept pairs with weights bigger than 2.0 to improve the data quality and collect 100,000 concept pairs with 27 relations. Due to the vast amount of Wikidata, we randomly sample 5 million concepts with 813 relations from Wikidata, resulting in 20 million concept pairs.

3.3 Acquiring Analogies of the Same Relation

We can directly utilize the concept pairs in the KGs to generate analogies of the same relations. An important perspective is that humans usually draw upon familiar domains and situations to better understand unfamiliar ones. To make our analogy KB more applicable to real-world scenarios, we rank the concept pairs according to their popularity scores, reflected by pageview times (in Wikidata) and concept weights (in ConceptNet).

3.4 Acquiring Analogies of Analogous Relations

As defined in § 3.1, analogies of analogous relations consist of two concept pairs with analogous relations R_1 and R_2. However, it is difficult to automatically check whether R_1 and R_2 are analogous and manual annotation is costly. Recently, LLMs (Ouyang et al., 2022; OpenAI, 2022) have

I: Analogous Relations Generation

/ I: Task prompt */*
Choose the relations from the relation candidates that can form an analogy with the given relation.
/ Examples */*
Given relation: written by
Relation candidates: [lyrics by, composed by, ...]
Answer: lyrics by, composed by, ...
/ Auto selection of analogical relations */*
Given relation: chief executive officer
Relation candidates: [head of state, ...]
Answer: *head of state, head of government, ...*

II: Meta Relation Summarization

/ Task prompt */*
Induce two relations into a higher-level relation and explain why they can form an analogy.
/ Examples */*
The relation **[lyrics by]** and the relation **[composed by]** can form an analogy because both of them can be induced into a relation: **[created by]**.
The relation **[written by]** and the relation **[written system]** can form an analogy because both of them can be induced into a relation: None.
/ Auto-completion for meta relation */*
The relation **[chief executive officer]** and the relation **[head of government]** can form an analogy because both of them can be induced into a relation: *head of organization*.

Table 1: Examples of prompt for InstructGPT$_{003}$ for analogous relations generation and meta relation summarization. *Green texts* are generated by InstructGPT$_{003}$.

shown their remarkable few-shot learning abilities with in-context learning. Given a task prompt describing the task and several examples, LLMs can do the task well without training. Therefore, we propose to exploit LLMs (*e.g.*, InstructGPT$_{003}$) to acquire analogies of analogous relations.

Finding Candidate Relation Pairs We collect 840 relations, leading to a potential amount of $\binom{840}{2}$ relation pairs. The relations that are semantically similar to each other can form an analogy (Hesse, 1959). For each relation, we first narrow down the candidate set from the 840 relations to the 20-most similar ones. Specifically, we use InstructGPT embeddings (`text-embedding-ada-002`) to convert the relations into embeddings and calculate the cosine similarity between them. By identifying the top 20 relations with the highest similarity as candidate relations for the query relation, the search space is significantly reduced for filtering analogous relations.

Predicting Analogous Relation Pairs While the search space is reduced, manual annotation remains cost-prohibitive (840 $\times$ 20). Thus, we continue to adopt InstructGPT$_{003}$ to predict analogous relation pairs. An example in Table 1 (I) shows the acquisition of analogous relation pairs. Given examples and the query ("*chief executive officer*"), InstructGPT$_{003}$ selects the relations "*head of state*" and "*head of organization*" from the candidates to form analogies. Finally, InstructGPT$_{003}$ obtains 284 relation pairs. However, we find that InstructGPT$_{003}$ struggles to filter out similar but wrong relations that cannot form analogies with queries, *e.g.*, "*operator*" for "*chief executive officer*", which requires further filtering.

Filtering for High-quality Relation Pairs In the examination process of 284 acquired relation pairs, we further implement two automatic filtering rules before conducting manual filtering to reduce human labor:

1. *Rule 1*: if two relations can form an analogy, InstructGPT$_{003}$ should simultaneously select R_1 for R_2 and R_2 for R_1.
2. *Rule 2* (Hesse, 1959): The second rule is using a more abstract *meta relation* to decide if two relations can form an analogy.

The rationale behind the *Rule 2* is that if two relations are analogous, then they can be generalized into a more abstract meta relation. For example, in Table 1 (II), *written by* and *composed by* are analogous since they can be induced to a meta relation *created by*. To acquire meta relations, we prompt InstructGPT$_{003}$ with a task prompt with some examples, as shown in Table 1 (II). If InstructGPT$_{003}$ returns "*None*", we discard this case.

After filtering, 103 relation pairs remain. To further improve data quality, we adopt a *third* filtering by recruiting two volunteers to manually examine the remaining results, including **deleting** relation pairs that fail to form analogies or **adding** previously unchosen relation pairs that can form analogies from candidates. Finally, we sort the concept pairs by pageview (Wikidata) and weight (ConceptNet).

3.5 Analysis of ANALOGYKB

As shown in Table 2, ANALOGYKB is massive, consisting of over 1 million concept pairs and 943 relations, which can form even more pairs of analogies. Since ANALOGYKB provides a more comprehensive range of relations than previous datasets, it allows users to select their preferred analogies within each relation (pair). To evaluate the quality

Source	# Concept Pair	# Rel(s)	Analogy Acc.
Analogies of the Same Relation			
ConceptNet	75,019	27	98.50%
Wikidata	563,024	813	98.00%
Analogies of Analogous Relations			
ConceptNet	11,829	5	95.50%
Wikidata	382,168	98	96.00%
Total	1,032,040	943	97.00%

Table 2: The statistics of ANALOGYKB. We report the number of concept pairs (**# Concept Pair**) and relations (pairs if for analogous relations) (**# Rel(s)**), manually evaluated the accuracy of randomly selected 200 analogies (**Analogy Acc.**) and the source KB (**Source**).

Data	# Analogy	# Rel	Language
SAT	374	-	En
Google	550	15	En
UNIT 2	252	-	En
UNIT 4	480	-	En
BATS	1,998	4	En
E-KAR	1251	28	En
E-KAR	1655	28	Zh
ANALOGYKB	$\geq$1,032,040	943	En

Table 3: Comparison between ANALOGYKB and previous analogy data source: numbers of analogies (*i.e.*, A:B::C:D), number of relations and language.

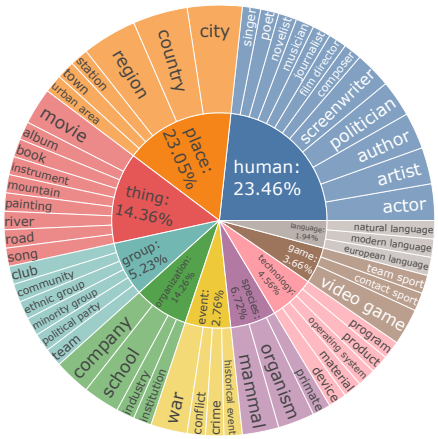


Figure 3: Distribution of concept categories in our ANALOGYKB.

Method	# Total	# Correct
ChatGPT	299	74
InstructGPT$_{003}$	284	97
+ Rule 1	139	97
+ Rule 1 & Rule 2	103	97
+ Rule 1 & Rule 2 & Human	103	103

Table 4: Ablated evaluation results of the analogous relation pairs. We record the total number of analogous relation pairs (**# Total**) the model selects and correct ones (**# Correct**). Note that "Human" denotes manual modifications, including adding missing relations or deleting incorrect ones, so the results are already correct (103→103).

of ANALOGYKB, we randomly sample 200 analogies from each data type, *i.e.*, two concept pairs of the same or analogous relations, in the form of A:B::C:D. The data is annotated by two annotators with Fleiss's $\kappa = 0.86$ (Fleiss et al., 1981). Results show that ANALOGYKB is of high quality. Even for analogies of analogous relations, analogies are still of over 95% accuracy.

We further compare ANALOGYKB with the resources related to the analogy, as reported in Table 3. We find that ANALOGYKB is much larger than previous data sources, with more analogies and relations. To better present the fabric of ANALOGYKB, we present the distribution of the categories of concepts covered in ANALOGYKB in Figure 3. The categories are obtained from the hypernym of concepts from Probase (Wu et al., 2012). We find that ANALOGYKB exhibits high diversity.

Are the filtering techniques for analogous relations useful? We evaluate the usefulness of the filtering components, *i.e.*, symmetry (*Rule 1*) and meta relation summarization (*Rule 2*), and manual correction. We also adopt ChatGPT (OpenAI, 2022) as an ablated variant. We record the total number of analogous relation pairs output by models (**# Total**) and then employ annotators to report the number of correct ones out of them (**# Correct**). In this process, the annotators need to review these relation pairs but no need to correct them. Each pair is examined by two annotators with Fleiss's $\kappa = 0.86$. The results in Table 4 show that: *1)* InstructGPT$_{003}$ is superior to ChatGPT but it still cannot filter out similar but wrong relation pairs, indicating the need for further filtering; *2)* We find the rule-based filtering technique to be rather effective, as there are not many manual corrections based on human annotations. This overcomes the labor-intensiveness of traditional KB construction methods and reveals the potential of this approach to be extended to the construction of other KBs.

4 ANALOGYKB Evaluation

4.1 Analogy Recognition Evaluation

Analogy recognition task aims to recognize the most analogous candidate to the query, formulated

Method	E-KAR	BATS	UNIT 2	UNIT 4	Google	SAT	Mean
Word Embedding from RoBERTa-Large	28.20	72.00	58.30	57.40	96.60	56.70	61.53
Word Embedding from InstructGPT	33.41	78.30	65.39	62.60	98.70	55.38	65.63
Sentence Embedding from SentenceBERT	25.40	68.00	53.40	46.00	90.45	47.70	55.16
Sentence Embedding from SimCSE	23.50	66.54	54.29	50.32	92.32	45.10	55.35
T5-Large	40.08	77.37	34.65	31.25	75.60	31.45	48.40
BERT-Large	36.64	70.10	32.89	34.49	90.40	41.30	50.97
ERNIE	40.83	82.54	34.21	36.80	82.40	34.92	51.95
LUKE	40.45	82.82	34.64	39.12	88.40	30.26	52.62
RoBERTa-Large	46.70	78.20	46.05	40.04	96.90	51.60	59.92
+ ANALOGYKB	**53.43**	90.93	87.28	**76.15**	97.80	59.05	77.44
+ ANALOGYKB (w/o check)	45.34	80.30	44.20	39.25	96.01	43.38	58.08
DeBERTa-v3	47.18	79.54	50.00	46.99	96.20	52.26	62.03
+ ANALOGYKB	53.05	**92.42**	**88.32**	75.30	**98.80**	**60.78**	**78.11**
+ ANALOGYKB (w/o check)	43.89	78.82	45.18	45.60	96.00	48.36	59.64
Human	77.80	84.85	87.50	66.66	99.41	57.00	78.87

Table 5: Accuracy on the analogy recognition task. We compare models and human performance on different benchmarks under different settings. The human performance values are obtained from the original papers of these analogy datasets. The best results are **bolded** and the second best ones are underlined.

as multiple-choice question-answering.

Can models trained on ANALOGYKB acquire better analogy recognition abilities? We adopt six analogy benchmarks, *i.e.*, E-KAR (Chen et al., 2022), BATS (Gladkova et al., 2016), UNIT 2 and UNIT 4 (Boteanu and Chernova, 2015), Google (Mikolov et al., 2013b) and SAT (Turney et al., 2003) for evaluation. Compared to BATS and Google, E-KAR, UNIT 2, UNIT 4, and SAT contain more abstract and complex analogies and thus more difficult for humans.

For the backbone model, we use the RoBERTa-Large (Liu et al., 2019) and randomly sample 10,000 data points from ANALOGYKB to train the model in a multiple-choice question-answer format. We first train the model on the data from ANALOGYKB and then further fine-tune it on benchmarks.[2] For baselines, we adopt pre-trained word embeddings (Ushio et al., 2021; Ouyang et al., 2022), pre-trained sentence embeddings (Reimers and Gurevych, 2019; Gao et al., 2021), pre-trained language models (Raffel et al., 2022; Devlin et al., 2019; Liu et al., 2019; He et al., 2023). To rule out the confounder in ANALOGYKB, we also add knowledge-enhanced models, ERNIE (Zhang et al., 2019) and LUKE (Yamada et al., 2020) which contain the relational knowledge between entities. Moreover, we also randomly sampled 10,000 data points from the ANALOGYKB without checking and filtering, *i.e.*, + ANALOGYKB (w/o check), to prove the necessity of filtering. After human examination, nearly about 63% of data points do not form analogies. Previous benchmarks, except E-KAR, do not have a training set. Thus, we fine-tune LMs on their small development set.[3]

The results presented in Table 5 show that: *1)* The performance of sentence embeddings is inferior to word embeddings. The rationale is that such word analogy is based on relational rather than semantic similarity between two sentences. Therefore, taking the difference between two word embeddings is a more reasonable yet still problematic approach for finding word analogies. *2)* Incorporating entity knowledge cannot improve model performance on analogy recognition; *3)* The training data without checking brings noise and even degrades model performance, further emphasizing the importance of high-quality data in ANALOGYKB. *4)* Training models on ANALOGYKB can significantly improve the model performance on analogy recognition by a large margin.

How much do analogies of analogous relations in ANALOGYKB contribute to performance? We create two ablated variants from ANALOGYKB to train the models: *1) Analogies of the same relations*, denoted as $\text{Data}_{\texttt{same}}$: we randomly sampled 10,000 data of the same relations as an ablated variant. *2) Pseudo analogies*, denoted as $\text{Data}_{\texttt{pseudo}}$:

[2]Detailed information on the benchmarks is shown in Appendix B and the construction of ANALOGYKB sample data is shown in Appendix C.1.

[3]Details about the baselines and training process are shown in Appendix C.2 and C.3. The statistical significant test is shown in Appendix C.5

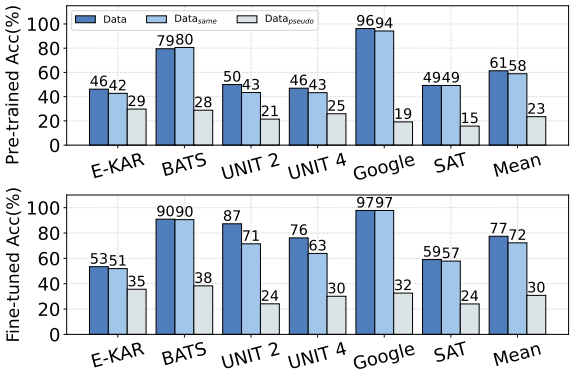
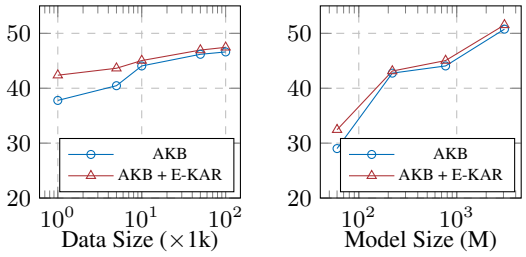


Figure 4: The accuracy of RoBERTa-Large trained on different data subsets on the analogy recognition task. **Data** denotes the dataset sampled directly from ANALOGYKB, **Data$_{same}$** denotes the dataset that only has same-relation analogies, and **Data$_{pseudo}$** denotes the dataset with concept pairs that do not form analogies. All the datasets have the same size.



(a) Different data sizes. (b) Different model sizes.

Figure 5: Performance change (Accuracy %) for T5 on E-KAR test set with increasing training data (1K, 5K, 10K, 50K, 100K) from ANALOGYKB and model size (60M, 220M, 770M, 3B). T5 is either trained on ANALOGYKB (AKB) or both ANALOGYKB and E-KAR (AKB + E-KAR).

Model	E-KAR	UNIT 4	SAT
vanilla T5	13.00	17.00	8.00
AnalogyT5$_{same}$	42.00	63.00	37.00
AnalogyT5	**57.00**	**80.00**	**64.00**
InstructGPT$_{003}$	61.00	70.00	60.00
+ Human	68.00	76.00	74.00
+ ANALOGYKB$_{same}$	64.00	77.00	77.00
+ ANALOGYKB	**75.00**	80.00	85.00
ChatGPT	58.00	76.00	78.00
+ Human	64.00	81.00	80.00
+ ANALOGYKB$_{same}$	64.00	80.00	81.00
+ ANALOGYKB	69.00	**92.00**	**91.00**

Table 6: Accuracy on analogy generation. For LLMs, we compare LLMs with 0-shot and human-written examples (+ Human) vs. ANALOGYKB-retrieved examples (+ ANALOGYKB). For smaller LMs, we compare AnalogyT5 with vanilla T5. AnalogyT5$_{same}$ and ANALOGYKB$_{same}$ are the ablation variants with analogies of the same relations from ANALOGYKB.

for each data point, we randomly sample 5 concept pairs from the ANALOGYKB and choose one as the query, one as the answer, and the remaining three as distractions. This makes sure that ANALOGYKB indeed imposes analogical reasoning ability on the model rather than simply data augmentation. We adopt two settings: only train RoBERTa-Large on 10,000 data (*i.e.*, **Pre-trained**) and first train RoBERTa-Large on 10,000 data and then fine-tune it on the specific benchmarks (*i.e.*, **Fine-tuned**).

The results in Figure 4 show that: *1)* Analogies of analogous relations in ANALOGYKB are rather important for models to comprehend analogies with more abstract and complex relations. *2)* Training models on randomly constructed analogy-style data even drags down model performance, further emphasizing the importance of ANALOGYKB.[4]

How do data sizes and model sizes affect performance? We use T5-Large as the base model to examine the effects of training data size on model performance. We first train the model on data from ANALOGYKB, and fine-tune it on E-KAR. As illustrated in Figure 5(a), increasing the amount of training data from ANALOGYKB improves model performance. Figure 5(b) shows the results of different-sized T5 models on 10,000 data points from ANALOGYKB. We find that the larger models get less of a performance gain from E-KAR, indicating that they learn more from ANALOGYKB and can better generalize to E-KAR.

[4]We also compare the data from different KB sources, which is shown in Appendix C.4.

4.2 Analogy Generation Evaluation

This task can be formulated as a text generation task: completing the *D* given *A*, *B*, *C* to form a plausible analogy *A is to B as C is to D*. Analogy generation is of more practical use, since the generation of familiar analogies could be helpful to comprehend the source problem.

Does ANALOGYKB support analogy generation? To answer this question, we investigate two settings: For smaller LMs, we randomly sample 1 million data points from ANALOGYKB. Then we fine-tune T5-Large on ANALOGYKB (named AnalogyT5) to compare vanilla T5. For LLMs, we convert the query and analogies from ANALOGYKB into InstructGPT embeddings, retrieve the top-8 most similar analogies based on cosine em-

Model	Acc.	MRR	Rec@5
GPT-2	2.00	5.70	10.70
+ BATS	1.10	2.20	3.60
+ E-KAR	2.39	5.44	10.46
+ ANALOGYKB$_{\text{same}}$	4.00	5.49	11.45
+ ANALOGYKB	5.12	6.41	12.77
BERT	0.90	4.40	8.00
+ BATS	0.40	1.90	3.10
+ E-KAR	1.50	4.32	7.92
+ ANALOGYKB$_{\text{same}}$	4.01	7.44	10.89
+ ANALOGYKB	6.24	10.36	14.07
InstructGPT$_{003}$	3.32	15.75	34.58
+ BATS	5.07	21.40	32.37
+ E-KAR	9.12	25.00	36.27
+ ANALOGYKB$_{\text{same}}$	6.91	25.32	33.42
+ ANALOGYKB	**15.30**	**32.80**	**38.46**

Table 7: Analogy generation results on SCAN. For LLMs, we compare LLMs with 0-shot and examples retrieved from BATS (+ BATS) and E-KAR (+ E-KAR) vs. retrieved from ANALOGYKB (+ ANALOGYKB). For smaller LMs, we pre-train the models on BATS (+ BATS) or E-KAR (+ E-KAR) or data sampled from ANALOGYKB (+ ANALOGYKB).

bedding similarity, and use them as examples in the prompt. We test models on 100 test data sampled from three challenging benchmarks, which are not found in the training set.[5] Each generation is evaluated by three annotators with Fleiss's $\kappa = 0.93$. The results in Table 6 show that, in both pre-training and in-context learning, ANALOGYKB enables better analogy generation, and the analogies of analogous relations prove significantly valuable to the performance of models.

Does ANALOGYKB help LMs generalize to out-of-domain analogies? Despite its high coverage of common concepts (§ 3.5), ANALOGYKB contains few analogies related to metaphor and science which are not common in the KGs and thus out-of-domain. To examine whether ANALOGYKB can generalize the ability of LMs to reason about these analogies, we test AnalogyT5 on the SCAN dataset (Czinczoll et al., 2022), which has 449 analogies of metaphor and science domains. For smaller LMs, we follow the original experimental setup and compare the models trained on ANALOGYKB (see Appendix D.5 for details). For LLMs, we retrieve the top-8 most similar analogies from ANALOGYKB as examples, in contrast to zero-shot settings, retrieving from BATS and E-KAR. The results shown in Table 7 reveal that *1)* For smaller LMs, training on BATS even worsens performance on SCAN. However, training on E-KAR with complex analogies can indeed improve the model performance on SCAN. *2)* Compared to E-KAR, ANALOGYKB can further help both LMs and smaller models generalize to out-of-domain analogies.

[5]Detailed information on the training process and the results on six benchmarks are shown in Appendix D.1 and D.3. We also conduct the impact of data and model sizes and case studies for further analysis in Appendix D.2 and D.4.

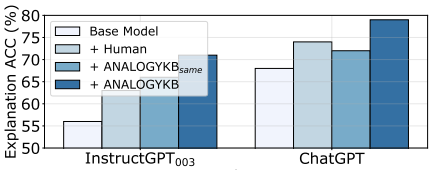


Figure 6: The accuracy of LLMs on the analogy explanation task. We compare LLMs with 0-shot (Base Model) and human-written examples (+ Human) vs. ANALOGYKB-retrieved examples (+ ANALOGYKB).

Can ANALOGYKB better support analogy explanation for LLMs? Analogy explanation needs LLMs to provide a reasonable explanation for a given analogy, which more closely simulates the process of human reasoning and knowledge explanation. In this setting, we first retrieve top-8 most similar analogies based on cosine embedding similarity. Then, we ask GPT-4 to generate explanations for the analogies given relations and use them as examples in the prompt. We test InstructGPT$_{003}$ and ChatGPT on 100 data samples from E-KAR, and employ two annotators to judge whether the explanations are correct with Fleiss's $\kappa = 0.97$). The results in Figure 6 are consistent with Table 6, demonstrating that ANALOGYKB can facilitate better analogy explanation for LLMs, and the analogies of analogous relations are significantly valuable for performance.

5 Conclusion

In this paper, we introduce ANALOGYKB, a million-scale analogy KB to improve model performance in analogical reasoning tasks. We identify two types of analogies in existing KGs, *i.e.*, analogies of the same and analogous relations, and utilize LLMs with minor human examinations to find them. ANALOGYKB demonstrates its great value in assisting both smaller LMs and LLMs with the resolution of analogy recognition and generation tasks, especially with analogies of analogous relations in ANALOGYKB.

Limitations

First, this paper only considers analogies involving one or two relations and primarily concentrates on analogies in the form of "A is to B as C is to D". However, analogies may involve the combination of multiple relations of multiple entities or even events. For example, an engineer can learn the eye cross-section by taking the analogy of the camera structure. Here, the analogy involves multiple entities and relations in the two systems (camera and eye): Aperture should be analogous to pupil since both are channels for light to enter and black paint should be analogous to choroid since both absorb light to prevent it from bouncing and reflecting.

Second, our ANALOGYKB is constructed using data from Wikidata and ConceptNet, which do not include analogies in other domains such as the scientific domain. For example, it would be challenging for LMs trained on ANALOGYKB to reason about an analogy such as Protein synthesis in a cell is like a factory assembly line as it would require a deep understanding of biological and industrial processes, which is not well-covered in our data sources. Also, ANALOGYKB is stored in the form of tuples, but in practice, some analogy situations may not be easily converted to this format. Future research should address how to bridge this gap.

Due to the limited computational resources, we only use a subset of ANALOGYKB. Assuming unlimited computational resources, the far-stretching goal of this project is to enable the discovery of new, better analogies for applications such as explanation (e.g., science popularization), text polishing, and case-based reasoning. So, with the full scale of the data, we can train a specialized open-source large language model (e.g., Llama 2) in such related tasks with data from ANALOGYKB so that these models can discover novel analogies and understand new concepts and knowledge with analogical reasoning ability.

Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

Use of Human Annotations The annotations of relation pairs in ANALOGYKB are implemented by annotators recruited by our institution. The construction team remains anonymous to the authors, and the annotation quality is ensured by using a double-check strategy as described in Section 3. We ensure that the privacy rights of all annotators are respected throughout the annotation process. All annotators are compensated above the local minimum wage and consent to the use of ANALOGYKB for research purposes, as described in our paper. The annotation details are shown in Appendix A.2.

Risks The database is sourced from publicly available sources, Wikidata and ConceptNet. However, we cannot guarantee that it is free of socially harmful or toxic language. Additionally, analogy evaluation relies on commonsense, and different individuals with diverse backgrounds may have varying perspectives.

References

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Paul Bartha. 2013. Analogy and analogical reasoning.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2022. Analogy generation by prompting large language models: A case study of instructgpt. *arXiv preprint arXiv:2210.04186*.

Adrian Boteanu and Sonia Chernova. 2015. Solving and explaining analogy questions using semantic networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. *arXiv preprint arXiv:2211.15268*.

Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. 2017. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246.

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.

Zijian Ding, Arvind Srinivasan, Stephen Macneil, and Joel Chan. 2023. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. In *Proceedings of the 15th Conference on Creativity and Cognition*, C&C '23, page 489–505, New York, NY, USA. Association for Computing Machinery.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. 2020. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dedre Gentner and Francisco Maravilla. 2017. Analogical reasoning. In *The Routledge International Handbook of Thinking and Reasoning*, pages 186–203. Routledge.

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.

Philippe Hansen-Estruch, Amy Zhang, Ashvin Nair, Patrick Yin, and Sergey Levine. 2022. Bisimulation makes analogies in goal-conditioned reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8407–8426. PMLR.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Mary B Hesse. 1959. On defining analogy. In *Proceedings of the Aristotelian Society*, volume 60, pages 79–100. JSTOR.

Douglas R Hofstadter. 2001. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538.

Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books.

Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. In-context analogical reasoning with pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.

Filip Ilievski, Jay Pujara, and Kartik Shenoy. 2022. Does wikidata support analogical reasoning? In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 178–191. Springer.

Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding.

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.

Peng-Hsuan Li, Tsan-Yu Yang, and Wei-Yun Ma. 2020. CA-EHN: Commonsense analogy from E-HowNet. In *Proceedings of the Twelfth 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. 2018. 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.

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*.

Jose L. Martinez-Rodriguez, Ivan Lopez-Arevalo, and Ana B. Rios-Alvarado. 2018. Openie-based approach for knowledge graph construction from text. *Expert Systems with Applications*, 113:339–355.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

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.

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

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

OpenAI. 2022. Chatgpt.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julien Romero and Simon Razniewski. 2020. Inside quasimodo: Exploring construction and usage of commonsense knowledge. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3445–3448, New York, NY, USA. Association for Computing Machinery.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana. Association for Computational Linguistics.

Mei Si and Craig Carlson. 2017. A data-driven approach for making analogies. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Robyn Speer, Catherine Havasi, and Henry Lieberman. 2008. Analogyspace: Reducing the dimensionality of common sense knowledge. In *AAAI Conference on Artificial Intelligence*.

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.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pages 1–16.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 481–492, New York, NY, USA. Association for Computing Machinery.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. 2022. Multimodal analogical reasoning over knowledge graphs. *arXiv preprint arXiv:2210.00312*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Details of ANALOGYKB

A.1 Terminology Definition in ANALOGYKB

To better understand the schema for analogies in ANALOGYKB, we list the terminologies in Table 9.

A.2 Crowd-sourcing Details

We have recruited a team of two undergraduates. We pay each annotator $8/h, exceeding the local minimum wage. The screenshots of the instructions and annotation interface are shown in Figure 8.

B Benchmark

We compare our methods with baselines and human performance in 6 different benchmarks. An example of these benchmarks is given in Table 8 For benchmarks without training sets, we only fine-tune models on their validation sets.

- **E-KAR** (Chen et al., 2022): a **E**xplainable **K**nowledge-intensive **A**nalogical **R**easoning benchmark sourced from the publicly available Civil Service Examinations (CSE) of China, which contains linguistic, commonsense, encyclopedic, and cultural (e.g., idiom and historical) knowledge. This dataset contains 870 training data, 119 validation data, and 262 test data. The SOTA model on this benchmark is proposed by Chen et al. (2022).

- **BATS** (Gladkova et al., 2016): is **B**igger **A**nalogy **T**est **S**et containing more than 1,000 analogies. The analogies can be divided into four categories: lexicographic, encyclopedic, derivational and inflectional morphology. This dataset contains 199 validation data and 1799 test data. The SOTA model on this benchmark is proposed by Ushio et al. (2021).

- **UNIT 2** (Boteanu and Chernova, 2015): a benchmark using word analogy problems from an educational resource. This dataset contains 24 validation data and 228 test data. The SOTA model on this benchmark is proposed by Ushio et al. (2021).

- **UNIT 4** (Boteanu and Chernova, 2015): this benchmark also comes from an educational resource but is harder than U2. This dataset contains 48 validation data and 432 test data. The SOTA model on this benchmark is proposed by Ushio et al. (2021)

- **Google** (Mikolov et al., 2013b): a benchmark for intrinsic evaluation of word embeddings proposed by Google, which contains semantic and morphological relations. This dataset consists of 50 validation data and 500 test data. The SOTA model on this benchmark is proposed by Chen et al. (2022)

- **SAT** (Turney et al., 2003): a benchmark constructed from SAT exams in the US college admission test consisting of 374 word analogy problems. This dataset contains 37 validation data and 337 test data. The SOTA model on this benchmark is proposed by Ushio et al. (2021).

As shown in Table 10, We list the overlap rates of ANALOGYKB with other analogy datasets. The overlap rates are calculated as (Data in ANALOGYKB Data in Other Datasets) / (Data in Other Datasets). Specifically, one data sample, i.e., "A is to B as C is to D" can be changed into two tuples (A, R1, B) and (C, R2, D), where R1 and R2 can be exactly the same or analogous. If both tuples are present in ANALOGYKB, the overlap rate for this data instance is considered greater than 0. The results indicate that ANALOGYKB contains a portion of the data from other analogy benchmarks, exhibiting high coverage. However, after our checking, we confirm that the training data sampled from ANALOGYKB, which is used to train LMs, does not contain the test data from other analogy benchmarks. This confirms the absence of data leakage, underscoring that LMs on ANALOGYKB can significantly improve the model performance on analogy recognition and generation tasks.

Query	army:order
Candidates:	(A) volunteer:summon
	(B) band:band leader
	(C) tourist:guide
	(D) students:instruction

Table 8: An example of analogy recognition task. The true answers are ***highlighted***.

C Analogy Recognition Task

C.1 Data Construction

To pre-train RoBERTa-Large on ANALOGYKB, we randomly sample 5,000 analogies of the same relation and 5,000 analogies of analogous relations from ANALOGYKB and formulate them into the multiple-choice question-answering format. Specifically, for each instance, we randomly sample a concept pair from ANALOGYKB as a query and select another concept pair from the analogous relation as the answer. Then, we randomly sample 3 concept pairs from the relations that can not be analogous to the relation of the query as distractions.

We also randomly sample 10,000 data of the same relations as an ablated variant to show the effectiveness of analogies of analogous relations (denoted as $\textbf{Data}_{\texttt{same}}$). The construction method is similar, except that the query and answer are derived from the same relation. Additionally, we randomly sample 10,000 data points from ANALOGYKB and construct analogy-style data (denoted as $\textbf{Data}_{\texttt{pseudo}}$). Specifically, we randomly sample 50,000 concept pairs without considering analogous relations from ANALOGYKB as the data pool. For each data point, we randomly sample 5 concept pairs from the data pool and choose one as the query, one as the answer, and the remaining three as distractions.

C.2 Details of Baselines

Word Embedding and Sentence Embedding For the method of pre-trained word embeddings, we follow the method proposed by Ushio et al. (2021). And represent word pairs by taking the difference between their embeddings. Then, we choose the answer candidate with the highest cosine similarity to the query in terms of this vector difference. For the method of sentence embedding, we convert query A:B to "A is to B" and choose the answer candidate ("C is to D") with the highest cosine similarity to the query.

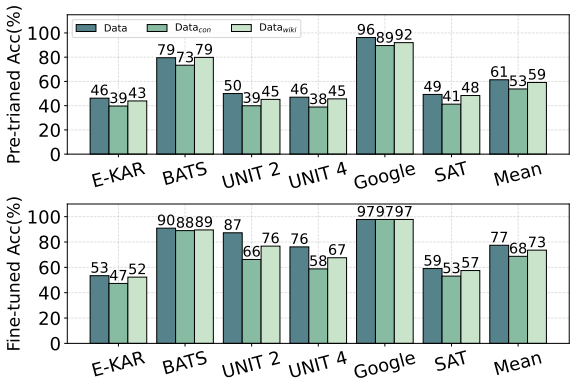


Figure 7: The accuracy of RoBERTa-Large trained on different data subsets on the analogy recognition task. **Data** denotes the dataset sampled directly from ANALOGYKB and $\textbf{Data}_{\texttt{con}}$ (or $\textbf{Data}_{\texttt{wiki}}$) denotes the analogies only from ConceptNet (or Wikidata). All the datasets have the same size.

C.3 Training Process

To pre-train language models on the sample data from ANALOGYKB, we follow the code from Huggingface [6]. Since previous benchmarks, except E-KAR, do not have a training set, we fine-tune LMs on their small development set (about 300 samples). To achieve hyperparameter search, we maximize performance on the development set of E-KAR (119 data samples) as a compromise. The training settings are: batch size = 64, learning rate = 3e-5, dropout rate = 0.1 and training epoch = 10.

C.4 Comparison with Different KB Sources

We also create two ablated variants to train the models to evaluate the necessity of ConceptNet and Wikidata: *1) Analogies from ConceptNet*, denoted as $\textbf{Data}_{\texttt{con}}$: we randomly sampled 10,000 (the same size as before) data of the relations only in ConceptNet as an ablated variant. *2) Analogies from Wikidata*, denoted as $\textbf{Data}_{\texttt{wiki}}$: we randomly sampled 10,000 data of the relations only in Wikidata as an ablated variant. The results in Figure 7 show that ANALOGYKB can combine the commonsense knowledge of ConceptNet and the entity knowledge of Wikidata and thus exhibits superior performance in improving the analogy-making ability of models compared to utilizing a single data source.

[6] https://huggingface.co/docs/transformers/tasks/multiple_choice

Category	Definition	Example
Analogies	A:B::C:D (A is to B as C is to D)	Up:Down::High:Low, Tim Cook:Apple::Joe Biden:USA
Concept pairs	A:B or C:D	Left:Right, Tim Cook:Apple
Relation pairs	Two relations	(antonym, CEO), (CEO, head of state)
Analogous relations	Two relations that can form analogies	(CEO, head of state)
Analogies of analogous relations	A:B::C:D where the relation of A:B is different but analogous to the relation of C:D	Tim Cook:Apple::Joe Biden:USA

Table 9: The definitions of terminologies with examples in the schema for ANALOGYKB

Dataset	Overlap Rate
E-KAR	28.98%
BATS	78.25%
UNIT 2	52.32%
UNIT 4	41.48%
Google	98.52%
SAT	34.70%

Table 10: The overlap rates of ANALOGYKB with other analogy datasets.

/ Task prompt */*
Please make analogies.
/ Examples */*
input: artist is to paintbrush as magician is to
output: wand
input: razor is to shave as knife is to
output: cut
...
/ Test data */*
input: classroom is to desk as church is to
output: *pew*

Table 11: Prompt for LLMs for analogy generation task. Generated texts by LLMs are *highlighted*.

Data Size	Hit@k	E-KAR	UNIT 4	SAT
100K	1	30.00	38.00	25.00
	3	33.00	44.00	25.00
	5	33.00	44.00	26.00
500K	1	39.00	53.00	38.00
	3	42.00	58.00	38.00
	5	42.00	63.00	41.00
1M	1	57.00	80.00	64.00
	3	62.00	86.00	76.00
	5	**66.00**	**91.00**	**84.00**

Table 12: The model trained on data with different sizes is T5-Large (770M).

Model Size	Hit@k	E-KAR	UNIT 4	SAT
T5-small (60M)	1	18.00	18.00	14.00
	3	18.00	21.00	15.00
	5	18.00	22.00	15.00
T5-base (220M)	1	22.00	31.00	28.00
	3	23.00	31.00	34.00
	5	25.00	31.00	34.00
T5-large (770M)	1	57.00	80.00	64.00
	3	62.00	86.00	76.00
	5	**66.00**	**91.00**	**84.00**

Table 13: The model trained on data with different sizes is T5-Large (770M).

C.5 Significant Test

For the results in Table 5, we demonstrate that the random sampling of data does not greatly impact the accuracy through the statistical significance test. Specifically, we sample the training data of ANALOGYKB twice with different random seeds and run our method on these benchmarks in Table 5. Then, we implement a t-test on the two results with a 0.05 significance level. The result is not significant (p-value: 0.208), and thus we can not reject the null hypothesis (H0: $r_1 - r_2 = 0$, where r_i=(Acc. of E-KAR, Acc. of BATS, Acc. of UNIT 2, Acc. of UNIT 4, Acc. of Google, Acc. of SAT)); Furthermore, we fix the training data of ANALOGYKB and run our method on the benchmarks in Table 5 twice with different random seeds. The result is insignificant (p-value: 0.250), and thus, we can not reject the null hypothesis (H0: $r_1 - r_2 = 0$).

For the results in Figure 4, we conduct a statistical significance test on Data and $\text{Data}_{\texttt{same}}$. We average the accuracy of the two settings and implement a t-test with a 0.05 significance level. The null hypothesis H0 is $r_1 - r_2 = 0$, and the H1 is $r_1 - r_2 > 0$, where r_1 and r_2 are the lists of benchmarks' average accuracy of Data and $\text{Data}_{\texttt{same}}$ in Pre-trained and Fine-tuned settings. The result is significant (p-value: 0.012), and we can reject the null hypothesis H0. Thus, we can conclude that analogies of analogous relations in ANALO-

Model	E-KAR	UNIT 4	SAT	BATS	UNIT 2	Google
vanilla T5	13.00	17.00	8.00	38.00	35.00	45.00
AnalogyT5$_{\text{same}}$	42.00	63.00	37.00	75.00	73.00	94.00
AnalogyT5	**57.00**	**80.00**	**64.00**	**80.00**	**84.00**	**95.00**
InstructGPT$_{003}$	61.00	70.00	60.00	82.00	79.00	94.00
+ Human	68.00	76.00	74.00	85.00	83.00	98.00
+ ANALOGYKB$_{\text{same}}$	64.00	77.00	77.00	83.00	85.00	**100.00**
+ ANALOGYKB	**75.00**	80.00	85.00	88.00	88.00	**100.00**
ChatGPT	58.00	76.00	78.00	84.00	84.00	96.00
+ Human	64.00	81.00	80.00	88.00	88.00	**100.00**
+ ANALOGYKB$_{\text{same}}$	64.00	80.00	81.00	92.00	91.00	**100.00**
+ ANALOGYKB	69.00	**92.00**	**91.00**	**96.00**	**94.00**	**100.00**

Table 14: The accuracy of different methods on the six analogy benchmark tasks in the analogy generation task.

Input	Completion
Mcdonald is to *America* as *Samsung* is to	south korea
oxygen is to *breathe* as *brain* is to	thinking
terrestrial is to *land* as *aquatic* is to	water
meticulous is to *careful* as *ascetic* is to	asceticism
triangle is to *area* as *cube* is to	volume
electron is to *nucleus* as *earth* is to	sun
electron is to *electric force* as *earth* is to	gravity
electron is to *atom* as *earth* is to	solar system

Table 15: Randomly selected and novel analogy generated from the AnalogyT5. Novel generations are concept pairs not found in the training set of AnalogyT5. Whether the analogy is considered plausible or not is decided by human annotators.

Target	Source	Attribute	mapping
Argument	War	Debater	Combatant
		Topic	Battleground
		Claim	Position
		Criticize	Attack
		Rhetoric	Maneuver

Table 16: Example mappings in SCAN. For a source concept, multiple related attributes are mapped to corresponding attributes of the target concept.

GYKB are rather important for models in the analogy recognition task.

D Analogy Generation Task

D.1 Training Process

To construct the training data, we convert A:B::C:D to "*A is to B as C is to D*" and let T5-Large generate the concept D given the input text "*A is to B as C is to*". The training settings are: batch size = 32, learning rate = 3e-5, dropout rate = 0.1 and training epoch = 20.

D.2 The impact of data sizes and model sizes

For the analogy generation task, we have examined the effects of training data size and model size on model performance. The results in Fugure 12 and Fugure 13 show that: *1)* By incorporating a larger volume of data from ANALOGYKB, we observe a gradual improvement in model performance, revealing the essential role of ANALOGYKB. *2)* Only larger models with enough training data can boost the ability to generate reasonable analogies.

D.3 Results on Six Benchmarks in Analogy Generation Tasks

We expanded the experiments in Table 6 to six analogy benchmark tasks. The results in Table 14 indicate that compared to analogies with simple and same relations, ANALOGYKB is more crucial for models to understand analogies with more abstract and complex relations, such as E-KAR, UNIT 4, and SAT.

D.4 Case Study

We are curious whether LMs trained on ANALOGYKB can generalize to novel analogies. After manual inspection, we observe from Table 15 that, AnalogyT5 can generate a reasonable concept D for the input. AnalogyT5 also generates reasonable analogies of analogous relations, such as "*triangle*" is to "*area*" as "*cube*" is to "*volume*". However, analogies about adjectives are more error-prone, possibly due to the paucity of adjectives in ANALOGYKB. We also discover that training on ANALOGYKB enables LMs to generate reasonable analogies by changing concept B while holding fixed A (*i.e.*, *electron*) and C (*i.e.*, *earth*).

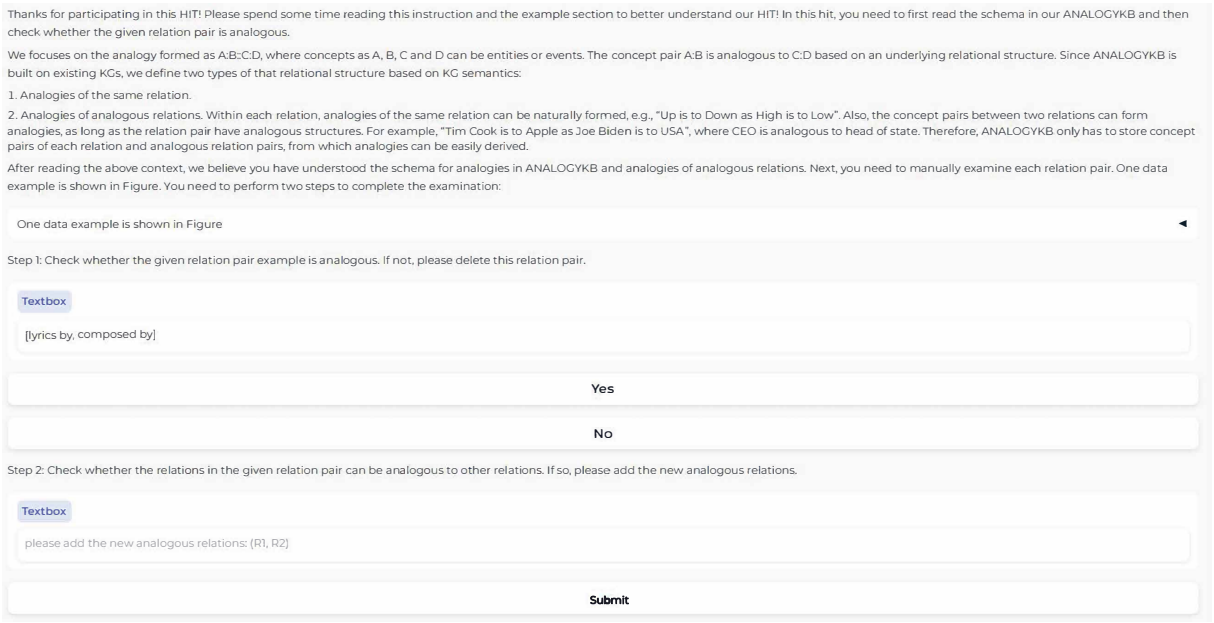


Figure 8: The screenshots of the instructions and annotation interface.

D.5 Out-of-domain Analogy

Dataset SCAN (Czinczoll et al., 2022) is an analogy dataset consisting of 449 analogy instances clustered into 65 full-concept mappings. The overlap rate of ANALOGYKB with SCAN is only 2.67%. An example mapping in SCAN is shown in Table 16. Unlike the previous analogy dataset, SCAN mainly contains metaphorical and scientific analogies, which are abstract and thus rarely appear in the corpus and are difficult for LMs. In addition, each concept in SCAN only has one token and SCAN is not confined to the word analogy task due to its full-concept mappings.

Baseline The original paper evaluates the analogical capabilities of GPT-2 and BERT on the SCAN dataset. The authors convert the analogy instance to "*If A is like B, then C is like D*", and force the models to predict the last token of the sentence. For GPT-2, the model needs to generate the last token given the input text "*If A is like B, then C is like*". For BERT, the authors first mask D as "*If A is like B, then C is like* `[MASK]`" and let the model predict word D.

In addition, the authors fine-tune the LMs on the 1,500-sized set of BATS (*i.e.*, + BATS) and investigate whether the models learn about analogical reasoning in general after training on BATS. We follow this setting and randomly sample 1,500 data from ANALOGYKB and fine-tune the LMs on the sample data (*i.e.*, + ANALOGYKB). To prove the necessity of analogies of analogous relations, we randomly sample 1,500 analogies of the same relations as an ablated variant (*i.e.*, + ANALOGYKB$_{\texttt{same}}$). We also added LMs trained on the 800 data points of E-KAR (*i.e.*, + E-KAR).

We further explore the performance of LLMs on the SCAN dataset. Specifically, we also adopt the prompt in Table 11 to let LLMs generate the word D. Since each concept in SCAN has only one token, we can obtain the top 5 results from InstructGPT through the OpenAI API.

Evaluation Metrics Following Czinczoll et al. (2022), we report accuracy, recall@5 and the mean reciprocal rank (MRR) to compare the performance of models. To reduce computing, we only consider the MRR of the first token of the target word among the top 10 predicted tokens. The RR of a label is 0 if it is not in the top 10 tokens.

Training Process The training settings of GPT-2 and BERT are: batch size = 128, learning rate = 3e-5, dropout rate = 0.1 and training epoch = 10.