Can Language Models Learn Analogical Reasoning? Investigating Training Objectives and Comparisons to Human Performance

Molly R. Petersen[1,2] and **Lonneke van der Plas[1,3]**

[1] Computation, Cognition & Language Group, Idiap Research Institute, Martigny, Switzerland
[2] NLP Lab, EPFL, Lausanne, Switzerland
[3] Institute of Linguistics and Language Technology, University of Malta, Malta
`molly.petersen@idiap.ch, lonneke.vanderplas@idiap.ch`

Abstract

While analogies are a common way to evaluate word embeddings in NLP, it is also of interest to investigate whether or not analogical reasoning is a task in itself that can be learned. In this paper, we test several ways to learn basic analogical reasoning, specifically focusing on analogies that are more typical of what is used to evaluate analogical reasoning in humans than those in commonly used NLP benchmarks. Our experiments find that models are able to learn analogical reasoning, even with a small amount of data. We additionally compare our models to a dataset with a human baseline, and find that after training, models approach human performance.

1 Introduction

Solving proportional analogies $a : b :: c : d$ (ex. *Paris:France::Tokyo:Japan*) (Mikolov et al., 2013a,b; Rogers et al., 2017) with embeddings has become an iconic way to demonstrate that they encode semantic information regarding the relationships between word pairs. Proportional analogies have been formalized with word embeddings by means of a simple arithmetic equation (often referred to as the vector offset method) $b - a + c = d$ where the choice of d is solved by maximizing $cos(b-a+c, d)$, or other slightly more complicated similarity measures (Mikolov et al., 2013a; Rogers et al., 2017; Levy and Goldberg, 2014; Ushio et al., 2021b).

This task has developed into many datasets such as the Google Analogy Test Set and the Bigger Analogy Test Set (BATS) (Mikolov et al., 2013a; Gladkova et al., 2016). Examples of relation types included in these datasets are morphological in nature ($entertain : entertainer :: examine : examiner$), based on encyclopedic knowledge ($Paris : France :: Lima : Peru$), or lexicographic relations, such as hypernyms and hyponyms ($candy : lollipop :: color : white$). Notably, the relations between entities that form each analogy are explicitly verbalized for each example, and are grouped into collections of many pairs with equivalent relations. While the ability of word embeddings to solve these sorts of analogies is interesting, the analogies contained in these datasets are different from what is typically used to test analogical reasoning in humans. Many would be trivially easy to solve, as is the case with morphological relations and a lot of encyclopedic knowledge (Ushio et al., 2021b). Even the task of completing an analogy by filling in a correct d term is perhaps not well suited to humans. Rogers et al. (2017) noted that analogy questions are often given to people as multiple choice questions, as people would likely fill in d with a variety of words where a single correct answer is not always the case. It has also been pointed out that model performance on analogy tests tend to rely on how semantically related all the entities in the analogies are to each other, suggesting these tests might in fact not be measuring analogical reasoning (Rogers et al., 2017). When using the vector offset method, the a, b, and c terms are generally excluded from candidates for d. If they are not excluded b or c is often the closest neighbor to the estimate of d and will be selected (Rogers et al., 2017; Linzen, 2016).

Analogical reasoning has the potential to extend beyond just solving proportional analogies. Analogies and analogical reasoning have value regarding scientific discovery, problem solving, human creative endeavors such as metaphor generation, as well as in the field of education (Boger et al., 2019; Gentner, 2002; Gentner et al., 2016; Clement, 1988; Gick and Holyoak, 1980; Gentner and Markman, 1997). In this work, we do not focus on the semantic/morphological analogies in datasets such as BATS, but on more "complex" analogies that are closer to what is used to test analogical reasoning in humans, what Ushio et al. (2021b) referred to as "Psychometric Analogy Tests". Most of the data we use has been developed precisely to test humans.

Furthermore, many of these analogies will have no verbalized relation between entities available, and there are likely no ways to group instances by type of relation without over-generalizing the relation. Unlike most previous studies, we want to go beyond just exploring what pretrained language models know about analogical reasoning. Our main contributions are as follows: we propose a training objective that allows language models to learn analogies. We find that learning is effective even with a fairly small amount of data, and approaches human performance on an unseen test set constructed for testing human analogical reasoning. Lastly, we find that fine-tuning models with this training objective does not deteriorate performance on external, but related tasks.

2   Related Work

While a lot of work tends to focus on performance on datasets such as BATS, there exist some datasets which were developed to test humans. Ushio et al. (2021b) explored the ability of word embeddings from various transformer models as well as some pretrained word embeddings (such as GloVe) to solve a variety of analogies originally designed to test humans in addition to the BATS and Google analogy test set. They tested three different scoring methods that aggregated across all valid analogy permutations. Their experiments found that the analogies designed for humans were much harder for all word embeddings to solve, with the better performing models' accuracy being less than 60% on datasets designed for humans, while the highest accuracy attained on Google and BATS was 96.6% and 81.2% respectively.

Czinczoll et al. (2022) introduced a dataset composed of scientific and metaphorical analogies, called SCAN, also testing performance on transformer based embeddings by using a natural language prompt, and filling in an ending [MASK] token. They experimented with a zero and one-shot learning setting, and additionally fine-tuned models on the BATS dataset. Overall they found that performance on the SCAN was low, and that models which had been fine-tuned on BATS lead to a decreased performance on the SCAN dataset, which the authors attribute to the types of analogies being inherently different and that datasets such as the BATS did not well represent how humans utilize analogies.

In this work, we test a novel training objective

to explore whether analogical reasoning in itself is a task that can be learned, and later test to unseen analogies with non-equivalent relation types. There have been a few works in addition to Czinczoll et al. (2022) that attempted at actively learning analogies or relations between words (Ushio et al., 2021a; Drozd et al., 2016) using language models. Czinczoll et al. (2022) and Drozd et al. (2016), handle the analogy problem in the more common way of predicting a word to complete the analogy. Our training focuses more on similarity in relations between entities, as opposed to similarity between entities in the analogy themselves, and uses this objective to fine-tune a pretrained BERT-base model to solve analogies as a classification problem. Ushio et al. (2021a) proposed RelBERT, whose motivation and training scheme is most similar to our set up in that they are focused on relations and relational similarity, even if their training task is not formulated as an analogy problem specifically. However their methods involve prompting and their training data was much larger (SemEval 2012 Task 2) and involved only lexical relations. Our work involves an arguably simpler training scheme, with a much smaller training dataset that includes a wider variety of analogical relations.

Additionally, there has been a lot of work in knowledge graph representation to learn embeddings for relations between entities, some of these which explicitly utilize analogical structure to learn embeddings (Liu et al., 2017). However, this area of research is out of scope as we are focusing specifically on learning to represent knowledge and relations in contextual language models.

3   Methods

All our experiments used a pretrained BERT-base uncased model (110M parameters)[1] (Devlin et al., 2019) unless stated otherwise. Models are trained on a binary classification task- when given an analogy, they must label it as either positive or negative.

We evaluate our models in three ways: 1) We evaluate the models' ability to learn analogy classification using two variants of a classification task, the binary classification task used during training and a ranking task, where we present them with both a positive analogy and its negative counterpart (how these are created is described in the *Datasets* section below), much like a multiple choice question with two possible answers. Whichever pair the

---

[1]https://huggingface.co/

model scores as more likely an analogy is the one the model "chooses". We chose to include this task as this is how it was formulated for the human baseline we use to compare our models, and is closer to how analogy tasks are often presented to humans (Rogers et al., 2017; Jones et al., 2022). Unlike the binary classification task, which will rely on a cutoff chosen to differentiate positive from negative, the decision made using ranking is relative to what a specific analogy is being compared to. 2) We compare the models' ability to solve analogies to a human baseline on an unseen dataset (also using ranking as described in the previous paragraph). 3) We evaluate performance on some external but related semantic similarity tasks. Statistical significance was determined using a two-sided z-test for proportions[2].

3.1 SBERT-modifications

Many of our experiments are a modification of the SBERT[3] architecture presented by Reimers and Gurevych (2019). SBERT creates sentence level embeddings, using the sentence representations- as opposed to token level representations- to solve a variety of tasks such as sentence similarity. Sentence-level representations are created by feeding the full sentence through a BERT or BERT-variant model, then pooling the embeddings from the final hidden layer.

In the original SBERT setup, sentences are fed through a model and the token embeddings are pooled to create a single representation. The main modification we do is that instead of feeding SBERT sentences to create a single sentence-level embedding, we feed SBERT word pairs and create two word-level embeddings- one for each word in the pair. In case a word is separated into word pieces, this still requires a pooling layer. We mean-pool over the token level embeddings to create each word embedding.

Additionally, in the original paper, sentences were fed through individually. So, for example, in a sentence similarity task which rates the similarity between two individual sentences (no attending between the sentences), the representations are determined separately. Their two individual representations are then compared. In our specific case with the $a : b :: c : d$ analogy structure, the meaning of a is dependent on b and vice versa. However

the meaning of c is not necessarily related to a, as they are just related to each other by their relation to their partner in their pair. Therefore we send through "*a [SEP] b*" separately from "*c [SEP] d*". This is only the case when the SBERT architecture is used. In the case that it is not, it will be explicitly stated.

3.2 Models

All models are trained[4] for three epochs, with a batch size of 32, and were trained using the Adam optimizer with an initial learning rate of $2e^-5$ (Loshchilov and Hutter, 2019). Each model took between 2.5-5 hours to train on a single CPU. We use BERT-base in our main experiments, however reproduce several experiments with other BERT architectures in the appendix.

3.2.1 Model 1: Simple Classifier

As a first model, we trained BERT (not SBERT) to classify a proportional analogy as being an analogy or not. The model is fed the analogy in the form '*a [SEP] b [SEP] c [SEP] d*', and outputs a 0 for not analogy or a 1 for true analogy. We chose to include one experiment with a simple classification head as this is the generic classification training paradigm for transformer models.

3.2.2 Model 2: BERT a-b

The rest of our models maintained formulating the task as a binary classification task, however instead of having a final classification head, we incorporated a novel training objective. We base our training off the idea that in a proportional analogy, the relationship between *a* and *b* can be framed as *a-b* (Mikolov et al., 2013a). Given that the pairs that make up an analogy should have the same relation expressed between the entities, it means *(a-b)=(c-d)*. We train our model to maximize the cosine similarity *cos((a-b),(c-d))* for positive examples and minimize for negative. We use a cosine embedding loss to train the model, with a margin of 0.

3.2.3 Model 3: BERT a-c

Since the above experiments using cosine distance as a similarity measure do not take into account $sim((a-c),(b-d))$, we test the model using $cos((a-c),(b-d))$ as the training objective. This reordering is of interest because in these analogy

---

[2]Code available at https://github.com/idiap/analogy-learning
[3]https://www.sbert.net/

[4]We did not do hyperparameter tuning as our goal is not to find the best model or beat any benchmarks, our goal is to test whether or not analogical reasoning is something that can be explicitly learned by comparing across models.

pairs, *a* and *b* are semantically related to each other in some way (ex: *nucleus* and *electron*), but *a* and *c* are generally not (ex: *nucleus* and *sun*). This means that, in theory, *a-c* is a much larger distance in semantic space. Previous studies have shown that when it comes to analogies, word embeddings often fail to capture more long-distance relations (Rogers et al., 2017). Long-distance relations (connecting two seemingly unrelated concepts) is of interest to the topic of computational creativity.

3.2.4 Baselines: No fine-tuning and FastText

We have three non-finetuned baselines, which we use to test the impact of our model training. The first is FastText[5], which was used because of its large vocabulary and ability to handle out of vocabulary words (Bojanowski et al., 2017). Given an analogy pair $a : b :: c : d$, the cosine similarity *cos((a-b),(c-d))* was calculated. In addition to FastText, we also used two pretrained SBERT baselines: *BERT a-b non-tuned* and *BERT a-c non-tuned*. These last two use the same cosine similarity measures as *BERT a-b* and *BERT a-c*, which will allow a more direct comparison to evaluate learning capabilities.

3.3 Datasets

For fine-tuning, we use a combination of four datasets (described below). This data was balanced, with half of the data points being true analogies and half false. We generated one negative analogy from every positive analogy, where for every true $a : b :: c : d$, a negative analogy $a : b :: c' : d'$ was created, which resultsed in a total of 4930 analogies. The choice of *c',d'* depended on the dataset and is described in their respective subsections. Examples of analogies contained in the datasets used, as well as other characteristics of the data, are in Table 7 in appendix A. The data was split into ten parts (each equal to ≈10% of the data). When creating each test set, we randomly selected positive analogies, and included their negative counterpart so that the model would not see either. Each model was trained ten times with one of the portions held out. Results shown are averaged across all ten runs.

3.3.1 SAT Dataset

The first dataset is composed of 374 analogies taken from a portion of the SAT college entrance exam in the US (this section has since been discontinued),

and has been used in several NLP publications (Turney et al., 2003; Ushio et al., 2021a,b). The original format was multiple choice, where, given a pair of words, the student had to choose another pair of words out of five provided that best formed an analogy with the given pair. Each question has one valid pair. One incorrect pair from the remaining four incorrect choices was chosen for each analogy to create negative samples, creating a total of 748 SAT analogies. One beneficial quality about these negative edges is that the questions were originally developed to be challenging to humans, therefore the incorrect option is not a pair of two random entities, but instead two entities that likely were chosen to be tricky.

3.3.2 U2 and U4

These analogies come from an English language learning website[6], used in some previous NLP analogy publications (Ushio et al., 2021a,b; Boteanu and Chernova, 2015). They were made for approximately ages 9-18, therefore comprising of a range of difficulty for humans. These questions were originally formatted as multiple choice as well, so negative instances were created in the same way as with the SAT data. The U2 dataset is a subset of the U4 dataset, so we removed all analogies that were present in the U2 from the U4. These two datasets contributed a total of 1208 analogies.

3.3.3 Scientific and Creative Analogy dataset (SCAN)

The final dataset used in training is the Scientific and Creative Analogy dataset (SCAN), and is made up of analogies found in science, as well as commonly used metaphors in English literature (Czinczoll et al., 2022). Instead of forming pairs of pairs (with 4 entities), each analogy pair is composed of multi-entity relations. For example, in the analogy comparing the *solar system* to an *atom*, *solar system* includes the entities, $sun, planet, gravity$, while *atom* includes the analogous entities $nucleus, electron, electromagnetism$. Each analogy pair provides $C(n, 2)$ total analogies in the format $a : b :: c : d$, where n is the number of entities in a topic. So, for example, in an analogy where each topic in a pair contains four entities, there are $C(4, 2) = 6$ total analogies. While analogies that make up the evaluation test were not seen during training, the entities in the pairs that make up the evaluation analogies may have been seen,

---

| | BASELINES | | | | | | | | |
| | FastText | | | BERT a-b non-tuned | | | BERT a-c non-tuned | | |
Category	Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.
OVERALL	0.51	0.02	1.00	0.52	0.79	0.24	0.49	0.56	0.42
SAT	0.50	0.01	1.00	0.52	0.66	0.39	0.47	0.40	0.54
U2	0.51	0.02	1.00	0.51	0.69	0.33	0.48	0.54	0.42
U4	0.50	0.01	1.00	0.49	0.68	0.30	0.50	0.55	0.44
SCAN	0.51	0.03	1.00	0.52	0.87	0.17	0.49	0.60	0.38
SCAN - *Science*	0.57	0.13	1.00	0.51	0.84	0.19	0.43	0.51	0.35
SCAN - *Metaphor*	0.50	0.01	1.00	0.52	0.88	0.16	0.50	0.62	0.39
	TRAINED MODELS								
	Simple Classification			BERT a-b			BERT a-c		
Category	Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.
Overall	0.66	0.73	0.58	**0.72**↑	**0.71**↓	**0.73**↑	0.51↑	**0.39**↓	**0.62**↑
SAT	0.57	0.78	0.37	0.59↑	**0.64**↓	**0.53**↑	0.50↑	0.38↓	**0.63**↑
U2	0.57	0.66	0.48	0.58↑	**0.60**↓	**0.56**↑	0.53↑	0.50↓	**0.57**↑
U4	0.60	0.70	0.50	**0.56**↑	**0.60**↓	**0.52**↑	0.52↑	0.50↓	**0.53**↑
SCAN	0.71	0.74	0.68	**0.81**↑	**0.77**↓	**0.86**↑	0.50↑	**0.35**↓	**0.65**↑
SCAN - *Science*	0.77	0.85	0.68	**0.87**↑	**0.86**↑	**0.88**↑	0.46↑	**0.29**↓	**0.63**↑
SCAN - *Metaphor*	0.69	0.71	0.68	**0.80**↑	**0.75**↓	**0.85**↑	0.51↑	**0.36**↓	**0.66**↑

Table 1: Average Accuracy on Analogy Classification Task. Arrows next to *BERT a-b* are labeling whether accuracy went up↑, down↓, or stayed the same→, as compared to *BERT a-b non-tuned*. *BERT a-c* is compared to *BERT a-c non-tuned*. Boldface indicates a statistically significant change (p<0.05) using a z-test for proportions.

given the multi-entity nature of this dataset. This allows us to test the model's ability to learn to infer analogical relations when it is given other relations the entities do and do not have. Negative edges were created by randomly shuffling the *c,d* terms in the dataset. Given that the analogies were formed from combinations, random shuffling may accidentally result in a true analogy. All negative analogies were checked to make sure that they were not actually present in the positive analogy group. This created a total of 3102 analogies from this dataset - this represents about 63% of the total data.

3.3.4 Human Baseline Comparison: Distractor Dataset

These analogies were compiled by researchers in a university psychology department, where they tested whether semantic similarity affected an adult human's ability to correctly solve analogies (Jones et al., 2022). This dataset was not used in our model training, and recently has been used to probe large language model for analogical reasoning ability (Webb et al., 2023). In the original paper, the human subjects are presented with an incomplete analogy $a : b :: c :?$, where they must choose between two options for the *d* term. There are two levels of semantic similarity the authors ex-

plored. First they test human's abilities to solve analogies with regards to how related the *c,d* term is to the *a,b* term. Analogies are grouped into near analogies, where the *a,b* entities are semantically similar to the *c,d* entities, and far analogies, where the *a,b* entities are not semantically similar to the *c,d* pair. Then within each of these groups, they come up with two incorrect *d* options, which they refer to as distractors (the incorrect choices for *d*). One of the incorrect *d* entities is more semantically similar to the *c* term than the true *d* term is to the *c* term, which they refer to as a high distractor salience. For example, a true analogy they use is $tarantula : spider :: bee : insect$. They replace *insect* with *hive* as it is more related to *bee*. They measured semantic distance using LSA. The second incorrect *d* term that was chosen was less semantically similar (ex: replacing *insect* with *yellow*), which they refer to as low distractor salience. They also test three types of analogical relations: categorical, compositional and causal. Definitions and examples of these relations can be found in Jones et al. (2022).

3.3.5 External Tasks: Semantic Similarity

In order to see if our training scheme affects performance on external tasks, as can be the case

	BASELINES			FINE-TUNED MODELS		
Category	FastText	BERT a-b non-tuned	BERT a-c non-tuned	Simple Classification	BERT a-b	BERT a-c
OVERALL	0.81	0.69	0.46	0.54	**0.84**↑	**0.50**↑
SAT	0.87	0.63	0.47	0.52	**0.82**↑	0.49↑
U2	0.76	0.71	0.49	0.52	**0.84**↑	0.52↑
U4	0.75	0.71	0.43	0.56	**0.84**↑	**0.51**↑
SCAN	0.82	0.70	0.46	0.55	**0.85**↑	**0.50**↑
SCAN - *Science*	0.91	0.69	0.46	0.56	**0.81**↑	0.49↑
SCAN - *Metaphor*	0.80	0.70	0.46	0.55	**0.86**↑	**0.50**↑

Table 2: Average Accuracy on the Analogy Ranking Task. Arrows next to *BERT a-b* are labeling whether accuracy went up↑, down↓, or stayed the same→, as compared to *BERT a-b non-tuned*. *BERT a-c* is compared to *BERT a-c non-tuned*. Boldface indicates a statistically significant change ($p<0.05$) using a z-test for proportions.

with catastrophic forgetting, we test our models on three word-level, non-contextual semantic similarity datasets: SimLex-999, MEN, and WordSim353 (WS353) (Goodfellow et al., 2014; Kirkpatrick et al., 2017; Hill et al., 2015; Bruni et al., 2014; Finkelstein et al., 2001; Kemker et al., 2018)[7]. All these datasets contain words pairs with a similarity measure, however Hill et al. (2015) details some key features of how these dataset differ when they introduced SimLex-999; namely that both the MEN and WS353 tended to measure word relatedness/association as opposed to word similarity (not that these are mutually exclusive), and MEN's tendency to focus on less abstract concepts, such as nouns. The similarity measure within these datasets range from 0 to 10, while we use cosine similarity (details described in the next section), which gives a similarity measure in the range -1 to 1. We chose these tasks because it is a word level similarity task, which is related to our analogy task. Ideally the performance on these tasks would improve with our training, or minimally not decrease.

4 Results

4.1 Proportional Analogies as a Learnable Objective for Neural Networks

Table 1 shows accuracy on classifying the testset with both the baselines and trained models. FastText has a tendency to label all analogies as negative given the cosine similarity measure, while BERT models have a tendency to classify all analogies as positive. This is perhaps unsurprising, as it has been demonstrated that word embeddings

---

[7]We used some code from https://github.com/kudkudak/word-embeddings-benchmarks

exhibit anisotropy, and that anisotropy is higher with contextual word embeddings, resulting in any two random words having high cosine similarity to each other, perhaps translating into the distances between words being similar to each other (Ethayarajh, 2019). *BERT a-b* seems to be less likely to be biased towards one label as compared to the other baselines, with the relatively large SCAN dataset having the greatest tendency to be classified as positive.

The *a-b* training schemes improved overall accuracy on analogy classification over the previously discussed baselines, with most positive changes in accuracy being statistically significant. The largest gains were with the SCAN dataset, mostly due to an increased ability to correctly classify negative analogies. Performance was generally better with the science analogies than metaphors, with the *a-b* model reaching over 0.84 accuracy overall. Czinczoll et al. (2022) found that models performed better on science analogies than on metaphor analogies, which they attributed to metaphors being more abstract. As mentioned before, while the model would have never seen the examples from the evaluation set, it would have seen the entities in the pairs that make up the samples in the evaluation set as parts of other analogies. There was improvement on the other datasets with the *a-b* model, however the overall improvements were less dramatic, ranging from +0.07-0.08 when compared to *BERT a-b non-tuned*. We cannot directly compare our results to Czinczoll et al. (2022) and Ushio et al. (2021b), as Czinczoll et al. (2022) does not use a classification or ranking multiple choice task while Ushio et al. (2021b) used the entire list of negative analogies for the ranking task. Moreover, the main goals

	Before finetuning			After finetuning		
	Model Positive	Model Negative		Model Positive	Model Negative	
True Positive	205376	108667		187234	181359	185524
True Negative	207319	141762		170619	199694	191815
	206328	126385		182640	194465	

Table 3: Average # of times entities seen in pre-training data by true label and model guess before(B)/after(A) fine tuning (*BERT a-b*)

	Before finetuning			After finetuning		
	No OOV entity	1+ OOV entity	Total	No OOV entity	1+ OOV entity	Total
True Positive	0.98	0.54	0.79	0.75	0.65	0.71
True Negative	0.03	0.53	0.24	0.73	0.73	0.73
	0.50	0.53		0.74	0.69	

Table 4: Accuracy among analogies with no OOV entities and those with at least one before/after fine tuning (*BERT a-b*). No OOV n=2889, OOV n= 2041

of these papers were not to test training schemes.

The fine-tuned and non-fine-tuned models were generally able to perform an analogy ranking task better than the classification task, as shown in Table 2. The performance of each model between SCAN and the other datasets was less variable with the ranking task as opposed to the classification task. Again, the fine-tuned models outperformed their respective baselines with statistically significant improvements, with improvements being much greater with the *a-b* scheme.

4.2 Exploring Accuracy in Relation to Word Frequency and Subwords

Inspired by Zhou et al. (2022a,b), we explored whether there were any trends in classification associated with entity frequency in BERT's pre-training data, as well as subword tokenization. They found that cosine similarity between BERT embeddings tends to be under-estimated among frequent words (Zhou et al., 2022a). They also found that countries who were out-of vocabulary (OOV) in BERT's vocabulary were more likely to be judged as similar to other countries, and being OOV was related to being mentioned less in BERT's pre-training data (Zhou et al., 2022b). Keep in mind that we classified analogies using the cosine similarity between the distances between entities, and not the cosine between the entities themselves, which differentiates our results from theirs.

In order to approximate word frequency in the training data for our experiments, we use the estimates released by (Zhou et al., 2022a). Like Zhou et al. (2022b) found, the less common a word in

the training data is, the more likely it is to be out-of vocabulary (OOV) (Figure 1 in appendix A). Words tokenized into two or more subwords have generally been seen <10,000 times in the training data. Table 8 in appendix A shows the percent of each dataset that contains OOV words, as well as average times an entity is seen in the training data. The SCAN dataset contained < 10% OOV entities, while the SAT dataset contained almost 30% OOV words. Entities in the SCAN were seen on average twice as much in the pre-training data as entities in the SAT dataset.

Table 3 shows average word frequency by true and predicted label, while Table 4 shows classification accuracy by whether an analogy had at least one OOV entity. The entities contained in false analogies tended to be observed in the pre-training data more frequently than those in true analogies. However, analogies predicted as being true analogies contained entities that were seen a little over 60% more on average than those contained in analogies predicted as false before training. Additionally, analogies that contained no OOV entities were almost always predicted as true before training (Table 4). After training, the average frequency among predicted labels closely matches that among the true labels, and accuracy improved greatly among negative analogies with no OOV entities, as well as among analogies with OOV entities. It appears that before fine-tuning, the model overestimated the similarity in relations between analogy pairs with in-vocabulary words, and a bulk of the learning affected the ability to correctly identify lack of analogy. Similar trends can be seen when looking

		BASELINES								
		FastText			BERT a-b non-tuned			BERT a-c non-tuned		
Semantic Distance	Relation Type	Distractor Salience			Distractor Salience			Distractor Salience		
		Overall	High	Low	Overall	High	Low	Overall	High	Low
	OVERALL	0.70	0.63	0.77	0.53	0.52	0.53	0.34	0.27	0.42
Near	Overall	0.82	0.77	0.87	0.53	0.50	0.57	0.32	0.27	0.37
	Categorical	0.75	0.70	0.80	0.55	0.50	0.60	0.15	0.20	0.10
	Causal	0.70	0.60	0.80	0.55	0.50	0.60	0.35	0.20	0.50
	Compositional	1.00	1.00	1.00	0.50	0.50	0.50	0.45	0.40	0.50
Far	Overall	0.58	0.50	0.67	0.52	0.53	0.50	0.37	0.27	0.47
	Categorical	0.75	0.70	0.80	0.65	0.60	0.70	0.30	0.20	0.40
	Causal	0.55	0.50	0.60	0.45	0.50	0.40	0.30	0.30	0.30
	Compositional	0.45	0.30	0.60	0.45	0.50	0.40	0.50	0.30	0.70
		TRAINED MODELS								
		Simple			BERT a-b			BERT a-c		
		Distractor Salience			Distractor Salience			Distractor Salience		
		Overall	High	Low	Overall	High	Low	Overall	High	Low
	OVERALL	0.52	0.51	0.54	**0.70**↑	0.69↑	0.71↑	0.43↑	0.33↑	0.53↑
Near	Overall	0.63	0.56	0.70	**0.76**↑	0.72↑	0.80↑	0.39↑	0.31↑	0.47↑
	Categorical	0.64	0.57	0.70	0.76↑	0.74↑	0.77↑	**0.44**↑	0.34↑	**0.53**↑
	Causal	0.63	0.58	0.67	0.72↑	0.66↑	0.77↑	0.38↑	0.27↑	0.48↑
	Compositional	0.63	0.53	0.73	**0.81**↑	0.77↑	0.85↑	0.36↓	0.32↓	0.39↓
Far	Overall	0.42	0.46	0.37	0.63↑	0.65↑	0.61↑	0.47↑	0.34↑	0.60↑
	Categorical	0.46	0.55	0.37	0.68↑	0.74↑	0.61↓	0.49↑	0.32↑	0.66↑
	Causal	0.36	0.38	0.34	0.68↑	0.69↑	0.66↑	0.45↑	0.40↑	0.49↑
	Compositional	0.44	0.46	0.41	0.55↑	0.52↑	0.57↑	0.48↓	0.31↑	0.65↓

Table 5: Average Accuracy on Distractor Dataset. Arrows next to *BERT a-b* are labeling whether accuracy went up↑, down↓, or stayed the same→, as compared to *BERT a-b non-tuned*. *BERT a-c* is compared to *BERT a-c non-tuned*. Boldface indicates a statistically significant change (p<0.05) using a z-test for proportions.

within each dataset (Tables 9-16 in appendix A).

4.3 Neural Networks as Compared to Humans

Table 5 shows the results of testing our methods on an unseen testset that was previously tested on college students in the US by Jones et al. (2022). As a summary of what the original paper found - humans overall did well on solving these analogies (≈84% accuracy overall). They found that humans were better at solving near analogies than far analogies, and that humans had a harder time correctly choosing *d* then when there was high distractor salience as compared to a low distractor salience. When looking at relation categories, human performance was highest on the categorical analogies, and lowest on the causal analogies. To see results in detail please refer to Jones et al. (2022).

In our experiments, the best performing model was *BERT a-b*, with a 0.70 overall accuracy, up from 0.53 with *BERT a-b non-tuned*, and ≈0.14 worse than human performance. Accuracy for

BERT a-b mostly increased with training over the baseline, however most increases among subgroups were not statistically significant, although notably the sample size was small. When looking at subgroups, the same trends observed in humans were not present, nor were there any obvious trends among subgroups between the models, with the exception that near analogies were easier to solve than far analogies for our best model.

Results for additional BERT models and architectural designs are in the appendices C and D.

4.4 Performance on External Tasks

Finally, we tested on an external task to find out whether fine-tuning on the task of analogical reasoning might have a (negative) effect on semantic similarity tasks. Table 6 shows the Spearman's rank-order correlation coefficient for the three Semantic Similarity tasks. *BERT a-b* improved over *BERT non-tuned*, showing that training actually improved performance, even if performance is still

	Baselines		Fine-tuned Models		
Tasks	FastText	BERT non-tuned	Simple Classification	BERT a-b	BERT a-c
SimLex-999	0.44	0.17	0.21	**0.33**↑	**0.22**↑
MEN	0.81	0.27	0.27	**0.31**↑	**0.38**↑
WS535	0.69	0.27	0.27	0.27→	**0.36**↑

Table 6: Average Performance on Several Semantic Similarity Tasks. Arrows next to *BERT a-b* are labeling whether accuracy went up↑, down↓, or stayed the same→, as compared to *BERT non-tuned*. *BERT a-c* is also compared to *BERT non-tuned*, since this task compares the similarity between two words, not between the differences between two words. Boldface indicates a statistically significant change (p<0.05) using a z-test for proportions.

overall low compared to FastText. FastText outperformed all transformer models on the external tasks, which is unsurprising. Ethayarajh (2019) found that FastText and embeddings from lower layers of BERT outperformed final layer hidden representations from BERT, although they used the first principal component of the embeddings so the results are not directly comparable. Given the tasks are non-contextual, perhaps the contextual nature of BERT that allows it to perform well on certain tasks hinders it in others. Interestingly *BERT a-b* performed better on the SimLex-999 task than the other two tasks, unlike the baseline models presented. Hill et al. (2015) had found the SimLex-999 task was harder for neural embeddings to solve than MEN or WS353, which they attributed to these models being better at identifying word association than similarity. However they did not test BERT-like models.

5   Conclusion, Limitations and Future Work

In this paper, we aimed to move from testing relatively simple analogical relations in pretrained language models to testing the ability to learn more complex relations used for testing human analogical reasoning, with a tailored training objective. We found that overall, analogies are something that can be learned. We reach an accuracy of 0.70 up from 0.53 while being 0.14 below the human upper bound, on an unseen test set constructed to test human analogical reasoning. Lastly, we find that fine-tuning models with certain training objectives generally does not deteriorate their performance on external, but related tasks. In fact, on some tasks we observed improved accuracy.

Our experiments involve several limitations. For one, the dataset is small, making claims that analogical reasoning is something that for sure can or can-

not be learned with language models is not possible. Another important consideration is that analogies are permutable (Ushio et al., 2021b; Marquer et al., 2022). Given an analogy (1) $a : b :: c : d$, the following analogies also hold: (2) $b : a :: d : c$, (3) $c : d :: a : b$, (4) $d : c :: b : a$, (5) $a : c :: b : d$, (6) $c : a :: d : b$, (7) $b : d :: a : c$, (8) $d : b :: c : a$. Our *a-b* models account for 1-4, while our *a-c* models account for 5-8. These permutations are not without criticism - specifically (5) $a : c :: b : d$ and all its derivatives (6-8) (Marquer et al., 2022). Consider the analogy ($electron : nucleus :: planet : sun$). In our *a-b* models, we are making the assumption *cos((a-b),(c-d))*. In natural language, the relation on either side could be verbalized as *revolves around*. In the measure *cos((a-c),(b-d))*, there is no verbalizer that can describe both the *a,c* and *b,d* equivalently. The question is if that corresponds to no vector transformation that is equivalent between the two pairs. Lastly, Czinczoll et al. (2022) mentioned some of the metaphorical analogies contained antiquated gender roles, which could be potentially harmful.

One direction for future work is to address limitations of word embeddings or their representation power, such as the anisotropy that exists among contextual word embeddings (Ethayarajh, 2019). Perhaps a distance measure between two entities that is not subtraction would be a better way to represent their relation. Additionally, since prior work has suggested that analogies where the entities are close to each other in space are generally easier to solve with the vector offset method, perhaps focusing on incorporating knowledge regarding semantic distance between entities during training would be helpful (Rogers et al., 2017). We would also like to explore whether augmenting LM training with analogy learning for other common NLP benchmarks affects performance on these benchmarks.

Acknowledgements

References

Mark Boger, Antonio Laverghetta, Nikolai Fetisov, and John Licato. 2019. Generating near and far analogies for educational applications: Progress and challenges. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1968–1975.

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.

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

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. volume 49, page 1–47, El Segundo, CA, USA. AI Access Foundation.

John Clement. 1988. Observed methods for generating analogies in scientific problem solving. *Cognitive Science*, 12(4):563–586.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates. 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.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Dedre Gentner. 2002. Analogy in scientific discovery: The case of johannes kepler. *Model-based reasoning: Science, technology, values*, pages 21–39.

Dedre Gentner, Susan C Levine, Raedy Ping, Ashley Isaia, Sonica Dhillon, Claire Bradley, and Garrett Honke. 2016. Rapid learning in a children's museum via analogical comparison. *Cognitive science*, 40(1):224–240.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology*, 12(3):306–355.

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.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Lara L Jones, Matthew J Kmiecik, Jessica L Irwin, and Robert G Morrison. 2022. Differential effects of semantic distance, distractor salience, and relations in verbal analogy. *Psychonomic bulletin & review*, 29(4):1480–1491.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

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.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical inference for multi-relational embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2168–2178. PMLR.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Esteban Marquer, Pierre-Alexandre Murena, and Miguel Couceiro. 2022. Transferring learned models of morphological analogy. In *ATA@ ICCBR2022-Analogies: from Theory to Applications (ATA@ ICCBR2022)*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

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.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*

(*SEM 2017), pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Peter D Turney, Michael L Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules in lexical multiple-choice problems. In *RANLP*, pages 101–110.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. 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.

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

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022a. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022b. Richer countries and richer representations. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2074–2085. Association for Computational Linguistics.

A Characteristics of our Dataset

Tables 7 and 8 as well as Figure 1 characterize the training data that we used to reach a better understanding of what the model actually 'sees' during training. For explanations on how we created the training data from the original datasets we refer the reader to the section on Datasets. Table 7 gives examples of analogies and their negative counterparts used in the data. Table 8 shows some characteristics of the entities that make up each dataset we used for training our models.

	n	Positive	Negative
SCAN	2974	nucleus:electron::sun:planet	nucleus:electron::traveler:station
SAT	748	amalgam:metals::coalition:factions	amalgam:metals::car:payments
U2	504	permanent:temporary::skeptical:trusting	permanent:temporary::ordinary:plain
U4	704	order:chaos::unity:division	order:chaos::culture:feeling

Table 7: Example of Analogies in the Datasets used for Training

	% OOV	Ave. # of times entity seen in pre-training data
SCAN	9.4%	217127
SAT	29.4%	126132
U2	21.6%	158047
U4	23.8%	156826

Table 8: Characteristics of Entities in Datasets
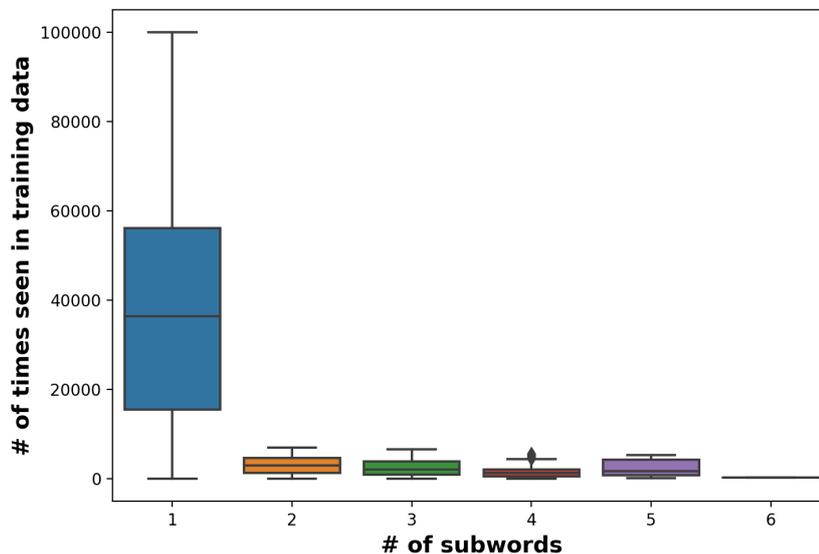
Figure 1: Distribution of estimated word frequency seen in pre-training data, by number of token per word (among words seen <100,000 times)

B Accuracy in Relation to Word Frequency and Subwords: Within Datasets

Tables 9 - 16 reproduce tables 3 and 4 in the manuscript but for each dataset individually. Similar trends seen in the overall data are seen within each dataset, specifically the improvements among being able to classify false analogies correctly that contain no OOV entities, as well as the model becoming less biased with regards to the number of times entities were seen and the label it chose.

	A - SCAN			B - SCAN		
	Model Positive	Model Negative		Model Positive	Model Negative	
True Positive	225457	158858		209101	244098	217127
True Negative	222250	191652		230867	214842	217127
	223893	177630		212499	221015	

Table 9: SCAN dataset: Average # of times entities seen in pre-training data by true label and model guess before(B)/after(A) fine tuning (*BERT a-b*)

	A - SCAN			B - SCAN		
	No OOV entity	1+ OOV entity	Total	No OOV entity	1+ OOV entity	Total
True Positive	0.97	0.62	0.87	0.77	0.83	0.77
True Negative	0.03	0.47	0.17	0.78	0.92	0.86
	0.51	0.54		0.80	0.85	

Table 10: SCAN dataset: Accuracy among analogies with no OOV entities and those with at least one before(A)/after(B) fine tuning (*BERT a-b*)

	A - SAT			B - SAT		
	Model Positive	Model Negative		Model Positive	Model Negative	
True Positive	138714	83755		122769	114774	119904
True Negative	143643	114737		119850	143477	132359
	141085	100264		121534	131892	

Table 11: SAT dataset: Average # of times entities seen in pre-training data by true label and model guess before(A)/after(B) fine tuning (*BERT a-b*)

	A - SAT			B - SAT		
	No OOV entity	1+ OOV entity	Total	No OOV entity	1+ OOV entity	Total
True Positive	0.97	0.53	0.66	0.77	0.59	0.64
True Negative	0.02	0.56	0.39	0.47	0.55	0.53
	0.47	0.55		0.62	0.57	

Table 12: SAT dataset: Accuracy among analogies with no OOV entities and those with at least one before(A)/after(B) fine tuning (*BERT a-b*)

	A - U2			B - U2		
	Model Positive	Model Negative		Model Positive	Model Negative	
True Positive	187756	78472		178890	116153	153497
True Negative	191153	104450		167793	158504	162596
	189435	91782		174171	140727	

Table 13: U2 dataset: Average # of times entities seen in pre-training data by true label and model guess before(A)/after(B) fine tuning (*BERT a-b*)

	A - U2			B - U2		
	No OOV entity	1+ OOV entity	Total	No OOV entity	1+ OOV entity	Total
True Positive	0.97	0.44	0.67	0.68	0.52	0.60
True Negative	0.03	0.62	0.33	0.48	0.63	0.56
	0.49	0.53		0.58	0.57	

Table 14: U2 dataset: Accuracy among analogies with no OOV entities and those with at least one before(A)/after(B) fine tuning (*BERT a-b*)

	A - U4			B - U4		
	Model Positive	Model Negative		Model Positive	Model Negative	
True Positive	177438	75383		147907	139783	144676
True Negative	202324	90524		149768	186713	168975
	190086	82676		148732	166372	

Table 15: U4 dataset: Average # of times entities seen in pre-training data by true label and model guess before(A)/after(B) fine tuning (*BERT a-b*)

	A - U4			B - U4		
	No OOV entity	1+ OOV entity	Total	No OOV entity	1+ OOV entity	Total
True Positive	0.99	0.45	0.68	0.69	0.53	0.60
True Negative	0.00	0.58	0.30	0.46	0.58	0.52
	0.47	0.51		0.57	0.55	

Table 16: U4 dataset: Accuracy among analogies with no OOV entities and those with at least one before(A)/after(B) fine tuning (*BERT a-b*)

C   Results from architectures: Single layer

We tested a model where we added a single feed forward layer on top of a pretrained BERT model, before the pooling layer. We kept the pretrained BERT weights frozen and trained only the additional layer using the *a-b* training scheme. Training with a single layer saw minimal change over the *BERT a-b non-tuned* baseline presented in the main text.

	Classification			Ranking
Category	Overall	Positive	Negative	
OVERALL	0.53	0.72	0.33	0.69
SAT	0.53	0.61	0.45	0.63
U2	0.52	0.63	0.40	0.72
U4	0.48	0.62	0.34	0.70
SCAN	0.54	0.79	0.28	0.70
SCAN *-Science*	0.54	0.76	0.31	0.72
SCAN *-Metaphor*	0.54	0.79	0.28	0.70

Table 17: Single Layer BERT a-b: Accuracy on Analogy Classification and Ranking Task

Semantic Distance	Relation Type	Distractor Salience		
		Overall	High	Low
	OVERALL	0.54	0.52	0.56
Near	Overall	0.55	0.53	0.57
	Categorical	0.57	0.51	0.62
	Causal	0.55	0.55	0.55
	Compositional	0.54	0.53	0.55
Far	Overall	0.53	0.51	0.55
	Categorical	0.66	0.6	0.72
	Causal	0.44	0.41	0.46
	Compositional	0.49	0.51	0.46

Table 18: Single Layer BERT a-b: Accuracy on Distractor Dataset

D Results from other architectures: Different Models

We ran a subset of our experiments with different BERT models: BERT-cased (110M parameters), BERT-large-uncased (340M parameters), and RoBERTa-base (125M parameters). Specifically the models we reproduced were *BERT a-b non-tuned*, *BERT a-b*, and *Simple Classifier*.

On the classification task, all the models saw some increase in overall accuracy as compared to the baseline, with RoBERTa achieving the highest overall accuracy (Tables 19-21). RoBERTa slightly outperformed BERT-base on the classification and ranking tasks (Table 21). The untrained RoBERTa *a-b* started out with a more extreme difference in accuracy between positive and negative samples in the classification task as compared to the *BERT a-b*, and while training closed the gap between the accuracy (while improving overall accuracy), *BERT a-b* was closer to achieving parity between the two groups. RoBERTa's performance did not transfer as well to the Distractor Dataset as compared to *BERT a-b* (Table 27). BERT-large-uncased did not see as much improvement with fine-tuning, likely due to the model being too large to learn from the relatively small dataset (Table 20,23,26).

Category	BERT a-b non-tuned			Simple Classifier			BERT a-b		
	Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.
OVERALL	0.52	0.85	0.19	0.51	0.97	0.06	0.66	0.83	0.48
SAT	0.52	0.90	0.13	0.50	0.98	0.02	0.53	0.76	0.29
U2	0.54	0.90	0.18	0.51	0.96	0.06	0.54	0.71	0.37
U4	0.51	0.90	0.12	0.50	0.96	0.05	0.54	0.73	0.35
SCAN	0.51	0.76	0.26	0.52	0.97	0.07	0.74	0.89	0.58
SCAN - *Science*	0.55	0.81	0.29	0.53	0.99	0.06	0.75	0.94	0.56
SCAN - *Metaphor*	0.49	0.73	0.26	0.52	0.97	0.07	0.74	0.88	0.59

Table 19: BERT-cased: Accuracy on Analogy Classification Task

	BERT a-b non-tuned			Simple Classifier			BERT a-b		
Category	Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.
OVERALL	0.54	0.39	0.69	0.50	1.00	0.00	0.59	0.47	0.71
SAT	0.52	0.03	0.74	0.50	1.00	0.00	0.54	0.45	0.64
U2	0.50	0.34	0.65	0.50	1.00	0.00	0.53	0.41	0.65
U4	0.49	0.32	0.65	0.50	1.00	0.00	0.52	0.42	0.63
SCAN	0.56	0.44	0.69	0.50	1.00	0.00	0.59	0.47	0.71
SCAN - *Science*	0.62	0.60	0.64	0.50	1.00	0.00	0.63	0.49	0.76
SCAN - *Metaphor*	0.55	0.40	0.70	0.50	1.00	0.00	0.68	0.57	0.80

Table 20: BERT-large: Accuracy on Analogy Classification Task

	RoBERTa a-b non-tuned			Simple Classifier			RoBERTa a-b		
Category	Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.
OVERALL	0.51	0.92	0.10	0.56	0.54	0.58	0.73	0.84	0.62
SAT	0.51	0.95	0.08	0.52	0.54	0.51	0.61	0.76	0.47
U2	0.51	0.90	0.12	0.51	0.48	0.55	0.59	0.76	0.42
U4	0.51	0.96	0.07	0.53	0.55	0.52	0.59	0.77	0.42
SCAN	0.51	0.88	0.14	0.58	0.55	0.61	0.82	0.90	0.74
SCAN - *Science*	0.50	0.88	0.13	0.60	0.61	0.60	0.88	0.98	0.78
SCAN - *Metaphor*	0.50	0.87	0.12	0.58	0.54	0.62	0.80	0.88	0.73

Table 21: RoBERTa: Accuracy on Analogy Classification Task

	BERT a-b non-tuned	Simple Classifier	BERT a-b
OVERALL	0.63	0.50	0.81
SAT	0.60	0.51	0.81
U2	0.62	0.48	0.83
U4	0.63	0.53	0.80
SCAN	0.63	0.49	0.81
SCAN - *Science*	0.66	0.49	0.84
SCAN - *Metaphor*	0.63	0.49	0.81

Table 22: BERT-cased: Accuracy on the Analogy Ranking Task

	BERT a-b non-tuned	Simple Classifier	BERT a-b
OVERALL	0.66	0.53	0.68
SAT	0.65	0.55	0.64
U2	0.67	0.53	0.67
U4	0.67	0.52	0.67
SCAN	0.65	0.53	0.69
SCAN - *Science*	0.64	0.54	0.72
SCAN - *Metaphor*	0.65	0.53	0.68

Table 23: BERT-large: Accuracy on the Analogy Ranking Task

	RoBERTa a-b non-tuned	Simple Classifier	RoBERTa a-b
OVERALL	0.60	0.54	0.88
SAT	0.59	0.53	0.86
U2	0.65	0.55	0.86
U4	0.61	0.56	0.89
SCAN	0.60	0.53	0.88
SCAN - *Science*	0.60	0.56	0.89
SCAN - *Metaphor*	0.60	0.52	0.88

Table 24: RoBERTa: Accuracy on the Analogy Ranking Task

Semantic Distance	Relation Type	BERT a-b non-tuned Distractor Salience			Simple Classifier Distractor Salience			BERT a-b Distractor Salience		
		Overall	High	Low	Overall	High	Low	Overall	High	Low
	OVERALL	0.58	0.52	0.63	0.48	0.51	0.45	0.69	0.70	0.68
Near	Overall	0.62	0.57	0.67	0.47	0.52	0.41	0.73	0.75	0.71
	Categorical	0.70	0.70	0.70	0.61	0.77	0.44	0.70	0.65	0.75
	Causal	0.40	0.30	0.50	0.52	0.55	0.48	0.66	0.68	0.64
	Compositional	0.75	0.70	0.80	0.28	0.25	0.30	0.82	0.91	0.73
Far	Overall	0.53	0.47	0.60	0.50	0.50	0.49	0.65	0.65	0.64
	Categorical	0.60	0.60	0.60	0.48	0.52	0.43	0.68	0.63	0.72
	Causal	0.45	0.30	0.60	0.59	0.59	0.58	0.56	0.64	0.48
	Compositional	0.55	0.50	0.60	0.44	0.40	0.47	0.70	0.67	0.73

Table 25: BERT-cased: Accuracy on Distractor Dataset

Semantic Distance	Relation Type	BERT a-b non-tuned Distractor Salience			Simple Classifier Distractor Salience			BERT a-b Distractor Salience		
		Overall	High	Low	Overall	High	Low	Overall	High	Low
	OVERALL	0.58	0.55	0.60	0.57	0.60	0.54	0.57	0.58	0.56
Near	Overall	0.63	0.57	0.70	0.59	0.59	0.59	0.61	0.60	0.61
	Categorical	0.80	0.70	0.90	0.67	0.69	0.64	0.59	0.60	0.57
	Causal	0.45	0.40	0.50	0.61	0.66	0.55	0.58	0.57	0.58
	Compositional	0.65	0.60	0.70	0.50	0.42	0.57	0.66	0.63	0.68
Far	Overall	0.52	0.53	0.50	0.55	0.60	0.49	0.53	0.55	0.51
	Categorical	0.65	0.50	0.80	0.58	0.60	0.56	0.60	0.62	0.58
	Causal	0.50	0.50	0.50	0.53	0.64	0.41	0.57	0.58	0.55
	Compositional	0.40	0.60	0.20	0.54	0.57	0.50	0.42	0.45	0.39

Table 26: BERT-large: Accuracy on Distractor Dataset

Semantic Distance	Relation Type	RoBERTa a-b non-tuned Distractor Salience			Simple Classifier Distractor Salience			RoBERTa a-b Distractor Salience		
		Overall	High	Low	Overall	High	Low	Overall	High	Low
	OVERALL	0.58	0.58	0.57	0.50	0.49	0.51	0.64	0.63	0.65
Near	Overall	0.55	0.53	0.57	0.58	0.58	0.57	0.67	0.63	0.70
	Categorical	0.55	0.60	0.50	0.60	0.59	0.61	0.60	0.60	0.60
	Causal	0.65	0.60	0.70	0.60	0.61	0.58	0.80	0.70	0.90
	Compositional	0.45	0.40	0.50	0.53	0.55	0.51	0.60	0.60	0.60
Far	Overall	0.60	0.63	0.57	0.43	0.39	0.46	0.62	0.63	0.60
	Categorical	0.45	0.60	0.30	0.42	0.45	0.39	0.70	0.60	0.80
	Causal	0.75	0.70	0.80	0.41	0.37	0.44	0.50	0.60	0.40
	Compositional	0.60	0.60	0.60	0.45	0.36	0.54	0.65	0.70	0.60

Table 27: RoBERTa: Accuracy on Distractor Dataset