AnaDE1.0: A Novel Data Set for Benchmarking Analogy Detection and Extraction

Bhavya Bhavya[1], **Shradha Sehgal**[1], **Jinjun Xiong**[2], and **ChengXiang Zhai**[1]

[1]University of Illinois at Urbana-Champaign
{bhavya2, sshegal4, czhai}@illinois.edu
[2]University at Buffalo
jinjun@buffalo.edu

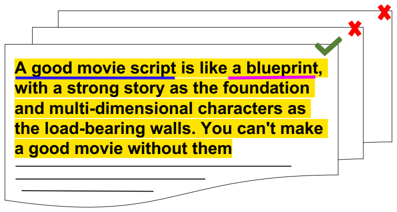
Abstract

Textual analogies that make comparisons between two concepts are often used for explaining complex ideas, creative writing, and scientific discovery. In this paper, we propose and study a new task, called Analogy Detection and Extraction (AnaDE), which includes three synergistic sub-tasks: 1) detecting documents containing analogies, 2) extracting text segments that make up the analogy, and 3) identifying the source and target concepts being compared. To facilitate the study of this new task, we create a benchmark dataset by scraping Metamia.com and investigate the performances of state-of-the-art models on all sub-tasks to establish the first-generation benchmark results for this new task. We find that the Longformer model achieves the best performance on all three sub-tasks demonstrating its effectiveness for handling long texts. Moreover, smaller models fine-tuned on our dataset perform better than non-fine-tuned ChatGPT, suggesting high task difficulty. Overall, the models achieve a high performance on document detection suggesting that it could be used to develop applications like analogy search engines. Further, there is a large room for improvement on the segment and concept extraction tasks[1].

1 Introduction

By mapping a complex or an unfamiliar concept, called the target, onto a more familiar concept, called the source, analogies aid in explaining educational concepts (Gray and Holyoak, 2021), inspiring creativity and scientific discovery (Ayele and Juell-Skielse, 2021; Gentner, 2002).

While analogy has been studied from the perspective of NLP for a long time, most work has studied analogy-finding either based on semantic similarity between texts (e.g., (Wijesiriwardene et al., 2023)), or by detecting shorter nominal metaphors (e.g., time is money) in documents. (Su et al., 2017). Detecting longer and more detailed analogies (we call them as **long form analogy**) at the level beyond a sentence that explains the similarities between the source and target concepts has not been well studied. Such analogies, however, are abundant on the Web and, if detected and extracted, would enable a wide spectrum of applications, such as search engines and dialog agents that can retrieve and generate analogies for education, research, and writing assistance.



A good movie script is like a blueprint, with a strong story as the foundation and multi-dimensional characters as the load-bearing walls. You can't make a good movie without them

Figure 1: AnaDE sub-tasks: (i) Analogy Document Detection to identify documents containing analogies (✓) vs. not (✗), (ii) Analogy Segment Extraction to extract text describing the analogy, (iii) Analogy Concept Extraction to extract the target and source concepts.

In this paper, we conduct the first systematic study of such long-form analogies in documents and propose a new task, called **Ana**logy **D**etection and **E**xtraction (**AnaDE**), which includes three synergistic sub-tasks: 1) **Ana**logy **D**ocument **Det**ection (**AnaDet**): detecting documents containing analogies, 2) **Ana**logy **S**egment **E**xtraction (**AnaSE**): extracting text segments that make up the analogy, and 3) **Ana**logy **Con**cept **E**xtraction (**AnaConE**): identifying the source and target concepts being compared in an analogy. The three subtasks are illustrated in Figure 1. The rationale behind the AnaDE task is that effective algorithms for it would enable running the algorithms on large amounts of online information (even the entire Web) to "harvest" all the analogies, which can then be used to power many downstream applications, such as a search engine for analogies.

[1] Data and code available at https://github.com/Bhaavya/analogy_classification

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1723–1737
March 17-22, 2024 ©2024 Association for Computational Linguistics

Although recent generative models like ChatGPT could already be quite powerful for some of these downstream tasks, e.g., generating analogies for education (Bhavya et al., 2022, 2023), our proposed tasks and datasets are still quite useful because of the following reasons: (1) A reference dataset of analogies is required to quantitatively evaluate the analogy-generation capabilities of ChatGPT and more powerful LLMs in future. (2) Using ChatGPT every time for generating analogies is computationally and monetarily expensive and unreliable, e.g., prone to hallucinations (Bender et al., 2021), compared to searching over the extracted analogies. (3) Our dataset and tasks would enable training or tuning models like ChatGPT to potentially generate better analogies, in addition to detection and extraction. Thus, these methods can be regarded as being complementary to ChatGPT.

Technically, the proposed task AnaDE is closely related to many well-studied NLP tasks, such as text categorization and information extraction. However, there are also significant differences between the AnaDE subtasks and those existing tasks and the AnaDE tasks present interesting new technical challenges for NLP research.

For example, the AnaDet problem is essentially a special case of the general text categorization problem, where special attention needs to be given to how an analogy is discussed in a document (e.g., use of local, analogy-indicating phrases). More importantly, to the best of our knowledge, there is only one work on AnaDet (Kumar et al., 2014, 2015), which only studied the special form of explanatory analogies for explaining educational concepts using a small data set that is not publicly available.

To the best of our knowledge, the AnaSE and AnaConE problems have never been studied, and naturally, there are no data sets available for studying them either. Although they are related to extraction tasks like NER (Yadav and Bethard, 2019), they are clearly different in nature given that the criteria to identify them are specific to analogies.

To facilitate the study of the new task AnaDE, we create the first benchmark dataset (called AnaDE1.0) by scraping Metamia.com and investigate the performances of state-of-the-art, transformer-based models on all sub-tasks to establish the first-generation benchmark results for this new task. We find that the Longformer model achieves the best performance on all the tasks, confirming its effectiveness in processing longer texts.

Overall, our findings suggest that the trained models already have a good performance on the AnaDet task and could potentially be used to develop useful applications, such as analogy search engines. Additionally, the AnaSE and AnaConE tasks were found to be challenging for all the models, leaving much room for future research as indicated by our error analyses. Finally, we envision that once AnaDE1.0 is released to the public, its quality and size can both be further improved continuously by the research community, leading to additional versions of the data set to better support the study of the proposed AnaDE task in the long run.

Specifically, our contributions include the following: 1) Our work advances our understanding of how to create a data set for studying long-form analogy in text, revealing challenges that can help future work on data set construction. 2) We create a new data set that, for the first time, enables quantitative evaluation of algorithms for long-form analogy detection and extraction. 3) We compare multiple representative state-of-the-art algorithms and establish the very first benchmark for the new computational problems we defined for long-form analogy. Although the reliability of this data can be further improved, this benchmark, including the data set and performance of SOTA methods, is the necessary initial step toward advancing research on long-form analogy detection and extraction, which have widespread applications (e.g., help students understand complex concepts using analogies).

2 Related Work

We now describe related work on identifying analogies, text categorization and information extraction.

2.1 Analogy Identification

Analogies have been computationally modeled for a long time (Mitchell, 2021) but our work is the first that has studied the effectiveness of representative state-of-the-art algorithms for detection and extraction in the case of the long-form analogy. Below we give an overview of the related work.

Some of the earlier work, such as Structural Mapping Engine (SME) (Forbus et al., 2017), used a rule-based method to match analogous concepts based on a structural representation of their attributes and relations. More recently, several methods including deep-learning ones have been developed for identifying proportional analogies (e.g., man:woman:: king:queen), where the analogous concepts share a single relation (Ushio et al., 2021; Turney et al., 2003; Boteanu and Chernova, 2015). Compared to all this work, our input consists of documents (and not concepts) and the outputs are the analogy segments and concept pairs extracted from them, which is required to develop new algorithms and advance technology for supporting many important applications especially to explain complex concepts to students to facilitate learning.

Another direction is finding pairs of documents (e.g., research papers (Chan et al., 2018)) based on their semantic similarity. Again, our task is formulated differently in that we aim to identify analogies present within a document, not to compute cross-document analogical similarity.

Large language models have also been recently used to generate longer analogies (Bhavya et al., 2022), similar to the ones we aim to detect and extract from documents. Our dataset and proposed tasks (that would enable crawling an even larger dataset) could be used to evaluate the analogy generation capabilities of models and also train or fine-tune even better generative models.

Closest to our work, there has been limited work that aims to classify a small dataset (300 webpages) of explanatory analogies using classical machine learning with features such as linguistic markers (e.g., "is similar") (Kumar et al., 2014, 2015). Compared to their work, we construct and investigate SOTA models on a larger dataset with all kinds of analogies, not just explanatory analogies.

2.2 Text Categorization

Text and document categorization (Hasan and Ng, 2014; Sebastiani, 2002) have been studied for a long time. The Analogy Document Detection is a special type of document categorization task with some unique properties, such as identifying local analogical features within long documents.

The problem of long document classification is challenging in itself, and recent work has investigated the performance of transformer-based models on this task (Park et al., 2022). In our work, we benchmark the performance of such models on the proposed AnaDet task that can help us increase our understanding of the behavior of these models.

Finally, there is also work on classifying other forms of figurative and comparative texts (e.g., metaphors (Su et al., 2017) and similes (Liu et al., 2018)). While these are much shorter (no more than a sentence), analogies are typically much longer with detailed explanations for how the concepts are related as indicated in Figure 2.

2.3 Information Extraction

Broadly speaking, both AnaSE and AnaConE tasks can be regarded as new types of information extraction (IE) (Grishman, 2015) tasks. AnaConE aims to extract shorter text segments, similar to entities, relations, and events (Yadav and Bethard, 2019; Gurulingappa et al., 2012; Xiang and Wang, 2019; Doddington et al., 2004), and keyphrases (Hasan and Ng, 2014). However, the context and nature of the full task is new as it involves extracting *pairs* (unlike individual units in NER) of *analogous* (a special kind of relation) concepts.

AnaSE can also be considered a special text segmentation (Pak and Teh, 2018) problem. Similar to research on other specialized segment extraction tasks (e.g., argument mining(Lawrence et al., 2014; Sardianos et al., 2015)), our new IE tasks offer an opportunity for designing and leveraging unique features and knowledge about analogies.

Another related work is textual analogy parsing (Lamm et al., 2018), which created deeper representations of only quantitative comparisons.

3 AnaDE1.0 Dataset

In AnaDE, the input is a set of documents (e.g., webpages) and the output is a set of analogy segments along with the source and target concepts used to form the analogy. To our knowledge, there is no publicly available dataset for studying AnaDE. A major contribution of our work is to construct the first dataset for studying AnaDE.

Specifically, to facilitate research on AnaDE with quantitative evaluation, we construct a dataset based on Metamia.com [2], called AnaDE1.0. Metamia is a crowdsourced website, where general web users submit analogies found on the web with the following information: a brief description of the analogy, the two concepts that are being compared to each other —a target (usually a more unfamiliar concept) and a source concept (usually the more familiar concept), and a reference link to the webpage where the analogy was found.

3.1 Dataset Construction

In this section, we describe the construction of the analogy dataset, negative samples for AnaDet task,

[2] http://www.metamia.com/

and summary statistics for all three tasks.

3.1.1 Analogy dataset

By scraping Metamia, we collected $9k$ records. The smaller number of records (as compared to all the records on the website) is because we collected only those records having reference webpage urls, which we then used for downloading the full webpages. Since webpages are constantly updated, we also downloaded their old versions from Internet Archive[3] to potentially get the version accessed by the crowdworker. Further, as a sanity check to ensure that the analogy is present in the webpage, we filtered out pages that did not contain the given target and source concepts. In this way, we collected 3.6k analogy webpages. This steep reduction (60%) comes from several issues, such as webpages being behind paywalls, the exact source and target concepts missing from the webpages either due to errors or paraphrasing by crowd workers or because the specific version of the webpage accessed by the crowd worker could not be found. This also highlights the challenge of collecting a large dataset for this task.

Additionally, for the analogy and concept extraction tasks respectively, we directly used the analogy texts and concepts from Metamia and performed the following processing to ensure data quality. Firstly, we want to ensure that the submitted analogy text is found exactly in the webpage and that the submitted source and target concepts are found exactly in the analogy. So, we automatically checked whether the full analogy text could be found in the downloaded webpage after removing special characters and spaces. Further, we manually made minor updates to 10% of the analogies (e.g., removed certain characters) to match the exact text in the webpages. Finally, 5% (200) of the total records were discarded due to issues, such as no analogy or irrelevant texts submitted. For concept extraction, we discarded an additional 10.5% (360) records because the submitted source or target concept could not be exactly found in the analogy text (e.g., they were paraphrased).

3.1.2 Negative Samples for AnaDet

To be able to distinguish between webpages containing analogies vs. not, negative samples are required. One solution is to use a random sample of documents from the Web as negative examples (which might be appropriate since most pages do not contain an analogy). However, such a data set might allow an algorithm to overfit the topics to detect analogy documents. Thus to make our data set more useful for studying AnaDet algorithms, we need to collect “harder" negative samples, i.e., webpages about similar topics as the positive samples but that do not contain any analogies. To this end, we retrieved 2-3 webpages for each analogy in the positive sample using its target concept and source concept each as a query with Microsoft Bing API [4]. In this way, we collected 11.5k negative samples.

3.1.3 Summary statistics

Table 1 shows the overall statistics for the AnaDet task. The documents are generally quite long, containing 3.6k words on average. For AnaSE, there are 3410 total analogies after discarding some analogy webpages as described in section 3.1.1. On average, there are 46 words per analogy. For AnaConE, there are 6102 total concepts (2 concepts per analogy) and 4864 unique concepts after discarding some analogies as described in section 3.1.1. Average number of words per concept is 2.53.

For all experiments, we perform 3-fold cross-validation, using a 70/30 train/test split.

Table 1: Dataset statistics for AnaDet

	# of samples	# words/sample
Positive	3.6k	4.5k
Negative	11.5k	3.3k
Total	15.1k	3.6k

3.2 Dataset Analysis

We now describe the overall dataset characteristics and results of label validation for the AnaDet task.

3.2.1 Dataset Characteristics

To better understand the analogies in our dataset, we first analyze their distribution based on their lengths. Figure 3 shows the results. We observe that about half of the analogies have >=40 words, or are longer than 2 sentences, unlike existing datasets on shorter analogies and metaphors. Similarly, the analogies have several nouns (average=10) and verbs (average=5) (Figures 6 and 7, Appendix A) suggesting that several analogies have detailed explanations of the comparison between concepts.

[3] https://archive.org/

[4] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

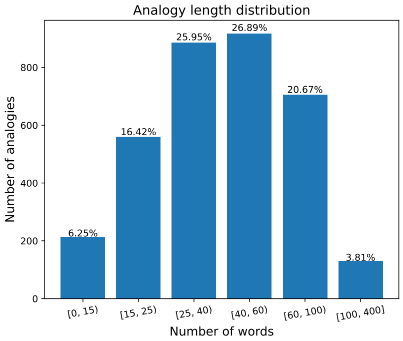


Figure 2: Analogy length distribution

Further, Figure 3 shows the positional distribution of analogies based on their starting character in documents, indicating that analogies can be located anywhere in them.

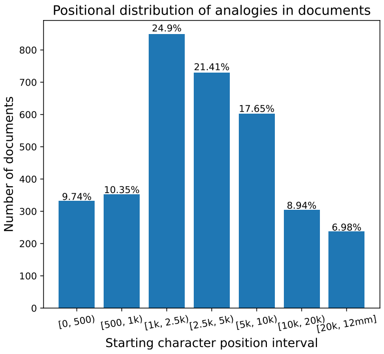


Figure 3: Positional distribution of analogies

Next, we look at the distribution of common indicators of analogical relation and comparative phrases in analogies in figure 4. We also show the top n-grams identified from the analogies in Appendix A, figure 5 and notice similar phrases. All analogies are in English. We note that ‘is like’ is present in a large majority of the analogies. Further, 177/3410 analogies did not contain any of these phrases and typically compare concepts directly using ‘is’, and ‘are’, e.g., “software evolution is the fruit fly of artificial systems”.

Finally, to help understand the types of concepts in the dataset, we clustered them into 50 clusters using k-means clustering (Ahmed et al., 2020) of their Sentence-BERT (Reimers and Gurevych, 2019)-based embeddings, and manually assigned a descriptive label to each cluster. Table 2 shows the top ten largest clusters and the remaining clusters can be found in Appendix A, table 7. We observe that there is a broad range of concepts, including the following: (1) academic or abstract topics (e.g., finance, chemistry, astronomy, etc.) – that are generally the “target” concepts being explained, (2) more familiar topics (e.g. common objects, animals, food, etc.) that could be the “source” concepts used to explain the target. Note that this dataset does not have labels to distinguish between source and target concepts but it would be interesting to add such fine-grained labels in the future.

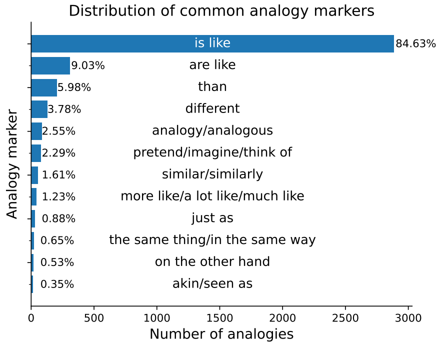


Figure 4: Distribution of common analogy markers

3.2.2 AnaDet label validation

We manually validated 50 positive and negative samples each from both Metamia. We found all positive samples to contain analogies (100% accurate). 10% of the negative samples also contained analogies (i.e., 90% accuracy). Unfortunately, we did not have the resources to manually judge all the 11.5k negative samples, we thus ended up using this data as is, leaving further improvement of the dataset as the future work. Although the dataset is not perfect, for comparison of different models (our main goal of research), it is still useful since the noise in the negative set (i.e., a small number of them are positive) would unlikely favor any particular method, thus any conclusions we draw about which method performs better are still meaningful.

4 AnaDE Benchmarking

In this section, we present our benchmarking results. As our data set is the first dataset for studying AnaDE, our goal in evaluation is to compare a set of representative state-of-the-art methods in terms of their performances on all the 3 subtasks so that we can establish the very first-generation benchmark results to enable further improvement of the models for solving the AnaDE problem. We report all implementation details in the Appendix A.

Table 2: Largest 10 clusters based on clustering the concepts into 50 clusters

Cluster label	# of concepts	Examples
human body organs	216	human heart; pancreas; small intestine
common objects	214	snow globe; cup of coffee; vacuum cleaners or hoovers
mental health	203	suicide; embarrassment; emotional bullying
miscellaneous	198	wolves; gold; home
grammar	173	sentence; semicolon; grammar
government/ideology	170	catholicism; socialism; communism
water	166	walk to a waterfall; heavy water; wave in the ocean
IT	160	computer; cache; oracle database
cells	159	cell wall; rna - binding proteins; white blood cells
cities/countries	156	scotland; new york; afghanistan

4.1 Analogy Document Detection

We frame the problem as a binary classification problem and investigate the performance of transformer-based neural models and classical ML models for long document classification.

Inspired by the recent success of transformer-based models shown in text classification, we benchmark them on our task. As mentioned in Section 3, the webpages in our dataset are quite long, while, transformer models like BERT (Devlin et al., 2018) have a limitation on the input sequence length due to the computationally expensive self-attention mechanism. Following previous work on comparative analysis of such models for long document classification (Park et al., 2022), we investigate the performance of the following models on our dataset: **BERT** (Devlin et al., 2018) processes the first 512 tokens from a document; **BERT+Random** processes randomly selected sentences constituting 512 tokens; **BERT+TextRank** processes 512 tokens from document summary based on TextRank (Mihalcea and Tarau, 2004); **ToBERT** (Pappagari et al., 2019) hierarchically processes blocks of 200 tokens; **CogLTX** (Ding et al., 2020) jointly trains BERT models for text classification and key sentence selection; and **Longformer** (Beltagy et al., 2020) processes first 4096 tokens.

As baselines, we compare the performance of standard text classification models: Naive Bayes, Logistic Regression, and Random Forest (Kamath et al., 2018) with TF-IDF vectors.

4.1.1 Results

Tables 3 and 4 show the results of 3-fold cross-validation (average and standard deviation) using classical and neural models. Among the classical models, we observe that the Random Forest model has the best overall performance, which is consistent with the observations made in most previous studies about such methods. We also experimented with different input token lengths for the classical models (Appendix A, table 9) and found that as the number of tokens increases, the overall performance generally improves but Random Forest seems quite robust. As expected, the neural models are typically better than the classical ML models although the gap is relatively small for some models including Bert. The overall accuracy on this task is high, suggesting the feasibility of using these methods to create a large collection of analogies.

Longformer model achieves the overall best performance. This suggests that it is effective in addressing the limitation of BERT in handling long docs. Based on the positional distribution (figure 3), the analogy starts within the first 2500 characters (or about 500 tokens) position for roughly half the documents. The higher accuracy of models with 512 tokens likely comes from either multiple analogies or other characteristics of pages with analogies (e.g., certain document types, like blogs, might generally contain analogies). While in an application scenario of identifying webpages containing analogies, this may or may not be an issue, depending upon how well the models or patterns generalize to the full web, it is certainly also interesting to create and study more challenging datasets that account for and remove any "spurious" features.

Unlike the observation in previous work (Park et al., 2022), we find that Bert+Random and Bert+TextRank do not achieve comparable performance to the more complex models like ToBERT and Longformer. This is expected to some extent because the sentence with analogies may not be part of the summary or randomly selected sentences. Although the CogLTX model is meant to identify key sentences (e.g., those containing analogies), we did not observe good performance. More hyperparameter tuning could be explored in future although it is time-intensive to train this model.

4.1.2 Error Analysis

Some examples of hard cases where all the models failed to identify the analogy webpages include:
(1) *Missing lexical indicator "is like"*: As discussed in Section 3, "is like" is a very strong analogy indicator. So, even when another lexical indicator was present, e.g., the word "analogy" in "...a common analogy is a water tank. In this analogy, charge...", all the models failed. This suggests that the trained models might have low generalizability over analogies expressed in different ways.
(2) *Presence of foreign characters surrounding the analogy:* For example, a Hebrew proverb surrounding the analogy likely caused the model to fail because of the unusual context.

Future work is needed to both create larger datasets covering a wider variety of analogies and surrounding contexts and develop more general models from smaller datasets. For example, one possibility is to automatically paraphrase the analogies or generate controlled analogical text comparing the source and target concepts from our dataset.

4.2 Analogy Segment Extraction

Next, we investigate the performance of algorithms on AnaSE. Since no previous work has studied this task, there are no obvious choices of models for comparison. We thus considered two representative families of models for extraction, i.e., transformer-based extractive and generative models.

4.2.1 Extractive Models

We frame the problem as an answer span extraction task. As observed in the analogy document detection task, the Longformer model is able to handle long webpages the best. Thus, we investigate its performance compared to Bert and follow the standard architecture of adding an answer span classification layer on top of the underlying transformer model (Devlin et al., 2018).

Imbalanced vs. Balanced Data: Due to the input token limits of these models, our long documents cannot be processed as a whole. In the standard approach [5], documents are split into overlapping chunks and the model is run on each chunk. The predicted span having the highest probability out of all predictions from all the chunks is selected as the predicted span for the entire document. In our case, with long documents having only a single analogy, this approach would create an imbalanced dataset with many chunks having no actual analogy spans (negative samples). Thus, we investigate how training on a balanced dataset (by undersampling the negative samples) impacts performance.

4.2.2 Generative Model

Given the impressive performance of large language models, particularly ChatGPT (GPT3.5) (Ouyang et al., 2022), on several NLP tasks recently, we also investigate its performance by designing a zero-shot prompt for this with precise instructions to extract the exact analogy (if any) without paraphrasing (Appendix A, Table 8). Similar to the chunking methodology used for extractive models, the API [6] was called on document chunks.

4.2.3 Results

Table 5 shows the main results on this task based on 3-fold cross-validation (average and standard deviation from the best hyperparameter run) based on Exact Match (EM) and F1 scores. We observe that the Longformer model obtains superior performance compared to Bert by a large margin, further demonstrating its effectiveness at handling long text. Moreover, training the models on balanced datasets helps improve the performance substantially, particularly in the case of Bert (+30.5% absolute improvement), likely due to a larger imbalance with Bert tokenization because of a smaller token limit. Moreover, zero-shot prompting ChatGPT obtains a relatively poor performance on this task when compared to the fine-tuned extractive models.

The overall lower performance of all models on this task clearly suggests the challenging nature of this new task that requires further research to solve.

4.2.4 Error Analysis

We conducted an in-depth error analysis for the AnaSE task since the results suggest that it is the most challenging task and it also touches on several issues relevant to the other two sub-tasks. In order to identify the common errors made by the models, which would be the most critical to address, we sampled 100 analogies where both the Bert and Longformer (balanced) models performed

[5] https://huggingface.co/docs/transformers/tasks/question_answering

[6] https://platform.openai.com/docs/api-reference/chat

Table 3: Classical ML Models Performance on Analogy Detection

Model	Acc.	P	R	F1
Naive Bayes	0.7978 ± 0.005	0.5754 ± 0.01	0.5655 ± 0.026	0.5703 ± 0.018
Log. Reg.	0.8866 ± 0.001	0.8821 ± 0.003	0.6031 ± 0.006	0.7163 ± 0.004
Random Forest	**0.9044 ± 0.002**	**0.8899 ± 0.006**	**0.682 ± 0.013**	**0.7721 ± 0.006**

Table 4: Transformer-based Analogy Document Detection Performance

Model	Acc.	P	R	F1
Bert+Random	0.8847 ± 0.058	0.7560 ± 0.137	0.7990 ± 0.048	0.7738 ± 0.097
CogLTX	0.9012 ± 0.006	0.7847 ± 0.027	0.8071 ± 0.014	0.7953 ± 0.009
Bert	0.9153 ± 0.017	0.8547 ± 0.022	0.7802 ± 0.049	0.8154 ± 0.037
Bert+TextRank	0.9239 ± 0.007	0.8568 ± 0.009	0.8438 ± 0.028	0.85 ± 0.02
ToBERT	0.9711 ± 0.003	0.9367 ± 0.01	**0.9426 ± 0.027**	0.9393 ± 0.009
Longformer	**0.9722 ± 0.002**	**0.9469 ± 0.01**	0.9359 ± 0.01	**0.9413 ± 0.005**

poorly (i.e., F1<10). Based on manual analysis, we identified the following major error types.

(1) *Multiple analogies:* (43% cases) In case of multiple analogies in a document, the model may sometimes extract an analogy that was not annotated in the ground truth.

(2) *Analogy span boundary issues*: (22% cases) These include cases where either the complete analogy is not labeled in the ground truth, or it has more statements labeled in addition to the analogy.

(3) *No analogy*: (5% cases) In this case, the ground truth analogy text did not contain nor was part of any analogy. For example, "see the link…"

(4) *Missing common analogy indicators*: (34% cases) Based on the distribution of common analogy indicators described in section 3.2., a majority of the training set has such indicators. Thus, analogies without them would be challenging cases. For example, "the human brain is a transsonic plane and the doctors studying it are engineers from 1900."

The first three error categories indicate that there is some noise in the data set. However, we note that this number could be disproportionately higher because it is representative of particularly hard samples and not a random sample of the full dataset, which could be done in future. Since manually inspecting the full dataset to detect such issues could be expensive, one possibility is to leverage our trained models for this. For example, to check for the presence of multiple analogies in a document, run the extraction model on small chunks of a document and check how many of those chunks had analogies. This would be useful future work for validating AnaDE1.0 data quality

To check for the impact of any noise, we conducted a paired t-test and found Longformer$_{bal}$ to be significantly better than Bert$_{bal}$ (p<.0001). These results suggest that the data is sufficient for discriminating different methods, allowing us to see statistically significant differences between them.

Table 5: Analogy Segment Extraction Performance

Model	EM	F1
Bert	8.77 ± 1.55	18.19 ± 3.37
GPT3.5	6.92 ± 0.63	29.21 ± 0.43
Bert$_{bal}$	21.17 ± 0.33	48.69 ± 0.49
Longformer	26.42 ± 0.73	59.58 ± 1.30
Longformer$_{bal}$	**28.59 ± 0.61**	**63.82 ± 0.52**

4.3 Analogy Concept Extraction

Finally, we investigate the performance of SoTA approaches for analogy concept extraction. We again considered two representative approaches, i.e., extractive models and generative models.

For the extractive models, we frame the task as a token classification problem, i.e., identifying tokens from the given analogy that belong to the label 'Concept'. For the generative ChatGPT model, we design a one-shot prompt with precise instruction to extract the exact source and target concept without paraphrasing (Appendix A, Table 8).

4.3.1 Evaluation

As AnaConE is a new task, it is also not immediately clear how we should evaluate it. We addressed this challenge by using the following method. Since each sample contains two ground truth concepts and potentially several predicted concepts, we align each predicted concept to its best matching ground truth based on the respective word-level Exact Match or F1 scores. Next, we take an average over all the aligned predicted concepts for each ground truth concept.

4.3.2 Results

Table 6 shows the main results (average and standard deviation over 3-fold cross-validation from best hyperparameter run). Again, we notice a similar pattern where smaller, fine-tuned models on our dataset perform better than one-shot prompting the larger GPT3.5 model. Moreover, the performance of the Longformer model is a bit better than Bert although the gap is not as large as observed on the previous two tasks. This is likely because the inputs for the previous two tasks are the full webpages, which are much longer than the inputs for this task (analogy segments). Thus, processing longer contexts, something that Longformer does better than Bert, is not as essential for this task.

Table 6: Analogy Concept Extraction Performance

Model	EM	F1
GPT3.5	67.10 ± 1.06	79.81 ± 0.90
Bert	76.11 ± 0.83	86.43 ± 0.43
Longformer	**79.19 ± 0.30**	**89.08 ± 0.16**

4.3.3 Error Analysis

We manually investigated errors made by the Longformer model and identified the following common issues apart from completely wrong extractions.
(1) *Minor dataset imperfections:* Similar dataset issues were observed for the AnaConE task as seen in the AnaSE task. For example, multiple sub-analogies can be present within an analogy, which should all be in the ground truth.
(2) *Analogous concept pair mismatch:* There could be several concepts within an analogy. Identifying which concept pairs are analogous might require an understanding of the sentence structure and semantic similarities between concepts.

5 Conclusion and Future Work

Automated detection and extraction of analogies from large amounts of online text has many applications such as search engines and dialog agents based on analogies. To this end, we proposed and studied a new task, called Analogy Detection and Extraction (AnaDE), which includes three synergistic sub-tasks: 1) Analogy Document Detection (AnaDet), 2) Analogy Segment Extraction (AnaSE), and 3) Analogy Concept Extraction (AnaConE). To facilitate the study of this new task, we created the first dataset, called AnaDE1.0. We systematically investigated the performances of state-of-the-art transformer-based models on all subtasks and established the first-generation benchmark results for this new task. We found that the Longformer model achieves high performance at the AnaDet task, suggesting that we can already use it to crawl many analogy documents from the Web to build an analogy search engine. At the same time, AnaSE and AnaConE are not only novel but also challenging, suggesting much room for future research based on our data set as suggested by our error analyses. Further, since the Longformer model can process a longer input context (e.g., compared to Bert), it has much better performance, especially on tasks where the input is the full web page that is typically very long. This indicates that, in future, models that are better at processing longer input contexts would likely perform even better. Moreover, smaller models fine-tuned on our dataset perform better than the non-finetuned ChatGPT model, further suggesting the value of our dataset for training and the difficulty of our proposed tasks.

Overall, our work also paves the way for interesting applications and research in this area. For example, for analogy segment extraction simile detection methods (Liu et al., 2018) could be leveraged to first identify the core comparative phrase (e.g., A is like B) and then boundaries of the full analogy around this phrase could be identified. Further, a sentence-level formulation of analogy segment extraction could also be explored to potentially alleviate some of the issues due to subjectivity in segment boundary annotation. Our dataset would enable studying such a formulation too after additional pre-processing, such as sentence-tokenizing the documents and analogy spans.

Finally, one limitation of the current version of our data set is the existence of some noise that we couldn't remove due to limited resources. Naturally, an important future task is to further improve the quality of AnaDE1.0 (e.g., by human-in-the-loop annotation of errors made by our trained models) and further verify our findings.

6 Limitations

Although the dataset is of significant size when it comes to analogies, it is smaller as compared to datasets available for other extraction use cases. Additionally, the submitted analogies on Metamia.com are all crowdsourced and may not be completely accurate. The process of extracting and labeling analogies introduces the potential for accuracy errors as analogies can be subjective and open to interpretation, making the task of labeling inherently challenging. Finally, the dataset in this study covers only the English language. Future work can focus on addressing these challenges to study and improve the robustness, scalability, and domain-specificity of analogy extraction.

7 Ethics Statement

We scraped publicly available content from Metamia.com. The intent as stated on the website is that the database is a collection of metaphors and analogies that will help with understanding the topic in question, and copyrights, if they exist, are retained by those doing the submitting. Although the rule list on Metamia.com [7] suggests that copyrighted material should not be submitted and obscenities should be avoided unless essential to the context, the dataset may still contain some offensive or copyrighted content.

To study the distribution of potentially offensive or inappropriate analogies, we employed the combination of a word-search-based and model-based method. For word search, we compiled a lexicon of terms related to gender, race, nationality, religion, body parts, etc. We seeded this lexicon with the terms related to identity found in the Jigsaw Bias in Text Classification dataset (Borkan et al., 2019) along with their synonyms from WordNet (Fellbaum, 2010). For the model-based method, we used two RoBERTa-based classification models for offensive and hate speech classification (Camacho-Collados et al., 2022).

We ran these two models and the word-search method across all analogies in our dataset and filtered the analogies that any of these methods labeled as offensive / hate / inappropriate. A total of 466 out of 3410 analogies were returned from combining these methods in an OR format. To manually evaluate if the analogies were indeed inappropriate, we sampled 100 of these 466 filtered analogies. We found 10 of these filtered analogies to be truly offensive. This may suggest that there exists a small number of inappropriate analogies in the dataset, and hence, it must be used with caution.

We have removed the identified offensive analogies from our released data and included disclaimers in the README about the potential existence of additional offensive data. We will also allow users of our data set to report offensive content so that we can further filter out the offensive content over time as part of the future plan.

This research also acknowledges the environmental and financial costs associated with using large language models. Furthermore, in real-world scenarios such as education, where analogies are useful, there is a need for robust validation mechanisms. While generative models have demonstrated impressive capabilities in generating text, any such models trained on this dataset should be thoroughly vetted before deployment as there is a risk of producing analogies that may be misleading, biased, or potentially dangerous.

8 Acknowledgement

The authors would like to thank National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, through Award # 2229873 (National AI Institute for Exceptional Education) and Award # 2229612 (National AI Institute for Inclusive Intelligent Technologies for Education). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295.

Workneh Y Ayele and Gustaf Juell-Skielse. 2021. A systematic literature review about idea mining: The use of machine-driven analytics to generate ideas. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 744–762. Springer.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

[7] http://www.metamia.com/the-rules-of-the-house

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2022. Analogy generation by prompting large language models: A case study of instructgpt. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 298–312.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2023. Cam: A large language model-based creative analogy mining framework. In *Proceedings of the ACM Web Conference 2023*, pages 3903–3914.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

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.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Christiane Fellbaum. 2010. Princeton university: About wordnet.

Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. 2017. Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, 41(5):1152–1201.

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

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Maureen E Gray and Keith J Holyoak. 2021. Teaching by analogy: From theory to practice. *Mind, Brain, and Education*, 15(3):250–263.

Ralph Grishman. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273.

Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. 2018. Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–11.

Varun Kumar, Savita Bhat, and Niranjan Pedanekar. 2014. Automatically retrieving explanatory analogies from webpages. In *European Conference on Information Retrieval*, pages 481–486. Springer.

Varun Kumar, Savita Bhat, and Niranjan Pedanekar. 2015. Stickipedia: A search engine and repository for explanatory analogies. In *2015 IEEE 15th International Conference on Advanced Learning Technologies*, pages 280–284. IEEE.

Matthew Lamm, Arun Tejasvi Chaganty, Christopher D Manning, Dan Jurafsky, and Percy Liang. 2018. Textual analogy parsing: What's shared and what's compared among analogous facts. *arXiv preprint arXiv:1809.02700*.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

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

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*.

Irina Pak and Phoey Lee Teh. 2018. Text segmentation techniques: a critical review. *Innovative Computing, Optimization and Its Applications: Modelling and Simulations*, pages 167–181.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 838–844. IEEE.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *arXiv preprint arXiv:2203.11258*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

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, 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? *arXiv preprint arXiv:2105.04949*.

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.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

A Appendix

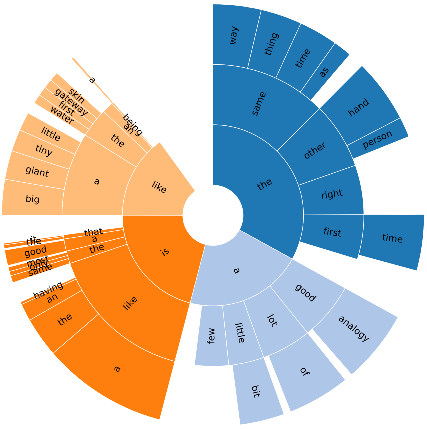


Figure 5: Top ngrams in the analogies scraped from Metamia.com. It shows common analogy indicators, e.g., ‘is like’, ‘good analogy’, and comparative phrases, e.g., ‘the same thing’.

A.1 Implementation details

A.1.1 Detection

For the baseline models, we use scikit-learn’s[8] default parameters. All the transformer-based models were trained for 10 epochs. For the CogLTX model, we used the implementation provided by the model developers [9] with an effective batch size of 6 on two NVIDIA GeForce 1080 GPUs. Introspector model learning rate was set to 0.75e-05 and reasoner model learning rate was set to 0.2e-04.

For all other models, we used the implementation provided by Park et al. [10]. The batch size of all BERT model variants was set to 8, Longformer was set to 12, both on single GPU. An effective batch size of 4 on four NVIDIA RTX A500 GPUs was

[8] https://scikit-learn.org/stable/

[9] https://github.com/Sleepychord/CogLTX

[10] https://github.com/amazon-science/efficient-longdoc-classification/

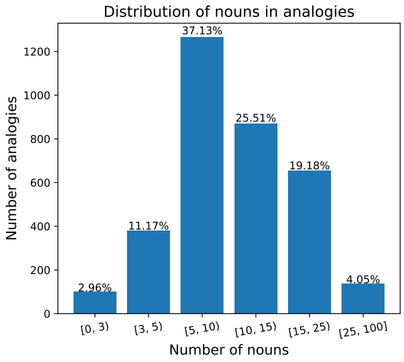


Figure 6: Distribution of nouns in analogies. The following is an example of an analogy with few nouns (=4): “The vacuole is like a rain barrel because it collects and holds water until it is needed by the cell.” The following is an example of an analogy with several nouns (=23): “In many ways, the interior of a eukaryotic cell is like a teeming metropolis. The nucleus, which is the repository of genetic information, mirrors the city hall, being a seat of legislative power while also doubling as the public library. The mitochondrion, which generates most of the cell’s supply of fuel (atp) is the power-station of the city, while the golgi apparatus that is responsible for packaging and processing proteins and lipids functions as the post-office.”

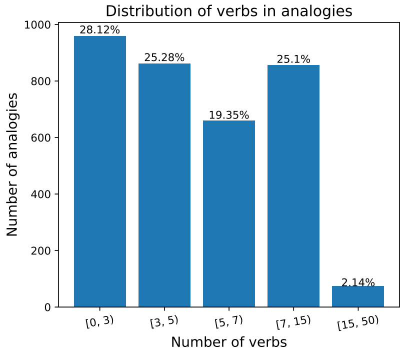


Figure 7: Distribution of verbs in analogies

used for the ToBERT model based on the memory and training time requirements. Learning rate for BERT and BERT+Random and ToBERT was set to 5e-05 for a batch size of 8 and adjusted accordingly for other batch sizes (multiply learning rate by k when batch size is scaled by k (Goyal et al., 2017)). In case of the BERT+TextRank model, a learning rate of 3e-05 was used because the model performance quickly dropped with 5e-05. For Longformer, learning rate was set to 1e-05, following previous work (Park et al., 2022). Training eac model on a single data split took a few hours to days and CogLTX took the longest time of about a week.

All models using bert-base have 110 million parameters, CogLTX has two RoBERTa-base models with 125 million parameters each, longformer-base has about 149M.

A.1.2 Analogy Segment and Concept Extraction

For the extractive models (Bert and Longformer), we report the best results from the following learning rates: {2e-5, 3e-5, 5e-5}. The batch size was set to 8 for Bert and 6 for Longformer. For Analogy Segment Extraction, the models were trained for 5 epochs, and 3 epochs for Analogy Concept Extraction, which took a few hours. The Longformer models were trained on a single NVIDIA A40 GPU and Bert model on a single NVIDIA GeForce 1080 GPU.

We used ‘gpt-3.5-turbo-16k‘ model i.e. GPT-3.5 model with the context of 16,384 tokens [11] as the generative model. We performed prompt engineering to ensure that the request returns the exact analogy from the provided text and also returns the source and target concept. To deal with long documents that do not fit within the token limit, we iterated in chunks of 9216 words (≈12,288 tokens), with a stride of 3072 words (≈ 4,096 tokens). We used the default parameters for the rest of our experiments. We spaced out each API request to adhere to the rate limits.

Bert-base model has 110 million parameters, longformer-base has about 149M, and gpt-3.5 has over 175 billion parameters.

[11] https://platform.openai.com/docs/models/gpt-3-5

Table 7: Smallest 40 clusters based on clustering the concepts into 50 clusters

Cluster label	# of concepts	Examples
fluids	150	espresso; gasoline; liquid nitrogen
materials and machinery	145	welding; plywood; plastic insulation
animals	142	white frail bird; octopus; elephant
genes	141	genetic material; genetics; dna fingerprinting
people	140	ron paul; reagan; orson welles
sex	136	sex; losing your virginity; sex education
finance	134	stocks; low interest rates; bitcoin
chemistry and chemical processes	130	oxygen; photosynthesis; methylation
building structures	128	fence; strong beams and columns; brick wall
atoms and nuclear processes	128	atom; neutrinos; beta decay
work/profession	127	unpaid internship; teacher; interviewing
vegetation	126	roots of a plant; succulents; tree
cognition	126	concentration; memory; mind
food	123	hamburger; water in the chunky noodle soup; thai food
relationships	120	bad boyfriend; critical friend or relative; my love
geology and scientific principles	115	geology; seismograph; weber's law
energy	113	little power plant; electric charge; free energy
region	113	gated community; habitat; candy store
car parts	111	airbags; the muffler; turbocharged engine
spirituality	110	holy spirit; spiritual light; old shamans
miscellaneous activities	109	breathing; cupping your hands; going to the supermarket
nervous system	106	prefrontal cortex; myelin; cerebrospinal fluid
movies	106	the lord of the rings; the academy awards; mean girls
nutrition	106	metabolism; exercise; nutrients
virus	104	virus; immunity; herpes
society and culture	101	sociology; western culture; racism
music	98	beethoven; orchestra; violin
transport	98	bus timetables; train travel; two - lane highway
diseases	95	eating disorder; hypertension; coronary heart disease
literature	91	shakespeare; poetry; reading hemingway
weather	87	spring; snowflakes;greenhouse effect
physics	82	torque; laws of motion; rotational force
combustion	79	grenade going off; using gasoline to light your charcoal grill; lava waiting to burst
astronomy	77	dark matter; tiny solar system; the milky way
games	75	jenga; football; golf
family	75	single family; little kids; christian parents
art	67	the mona lisa; tapestry; sculpture
drugs	66	antidepressants; nicotine; lsd
signals and communication technology	54	radio wave; higher frequency; network of phone lines
colors	53	food coloring; yellow; red hair

Table 8: GPT3.5 Prompts

Analogy Extraction	Find the exact analogy and all the sentences that explain it from the "document" below. The following is an example of an analogy: "My mom said life is like a box of chocolates. You never know what youŕe gonna get." Do not paraphrase, return the exact substring / sentences containing the analogy. Return this information in the following JSON format: {"analogy": <analogy>}. Return only one analogy even if there are multiple analogies present. In case no analogy is found in the text, explicitly return the string "No analogy found." Do not return any other string if no analogy is found. ===== Document: [document]
Concept Extraction	Find the source and target concepts in the analogy below. For example, "My mom said life is like a box of chocolates. You never know what you're gonna get" has the source "life" and "box of chocolates" as the target. Return this information in the following JSON format: {"source": <source concept>, "target": <target concept>}. In case no source and target concept is found in the text, explicitly return the string "No concept found." Do not return any other string if no concept is found. ===== Analogy: [analogy]

Table 9: Classical ML Models Performance on Analogy Detection with Truncated Documents

# Tokens	Model	Acc.	P	R	F1
512	Naive Bayes	0.6652 ± 0.002	0.4048 ± 0.002	**0.8707 ± 0.005**	0.5526 ± 0.002
	Log. Reg.	0.879 ± 0.001	**0.8565 ± 0.006**	0.5893 ± 0.012	0.6981 ± 0.007
	Random Forest	**0.9016 ± 0.003**	0.8374 ± 0.007	0.7268 ± 0.006	**0.7782 ± 0.006**
4096	Naive Bayes	0.7817 ± 0.001	0.5308 ± 0.002	0.6943 ± 0.017	0.6015 ± 0.006
	Log. Reg.	0.8868 ± 0.001	0.8804 ± 0.007	0.6055 ± 0.006	0.7175 ± 0.003
	Random Forest	**0.908 ± 0.002**	**0.8913 ± 0.004**	**0.6977 ± 0.013**	**0.7826 ± 0.007**