Natural Language Processing Journal 7 (2024) 100064

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Utilization of generative AI for the characterization and identification of visual unknowns

Kara Combs [a,*], Trevor J. Bihl [a], Subhashini Ganapathy [b]

[a] *Sensors Directorate, Air Force Research Laboratory, 2241 Avionics Cr., Wright-Patterson Air Force Base, OH, 45433, USA*

[b] *Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 3640 Col. Glenn Hwy., Dayton, OH 45435, USA*

ARTICLE INFO

Keywords:
Generative AI
Computer vision
Natural language processing
Analogical reasoning

ABSTRACT

Current state-of-the-art artificial intelligence (AI) struggles with accurate interpretation of out-of-library objects. One method proposed remedy is analogical reasoning (AR), which utilizes abductive reasoning to draw inferences on an unfamiliar scenario given knowledge about a similar familiar scenario. Currently, applications of visual AR gravitate toward analogy-formatted image problems rather than real-world computer vision data sets. This paper proposes the Image Recognition Through Analogical Reasoning Algorithm (IRTARA) and its "generative AI" version called "GIRTARA" which describes and predicts out-of-library visual objects. IRTARA characterizes the out-of-library object through a list of words called the "term frequency list". GIRTARA uses the term frequency list to predict what the out-of-library object is. To evaluate the quality of the results of IRTARA, both quantitative and qualitative assessments are used, including a baseline to compare the automated methods with human-generated results. The accuracy of GIRTARA's predictions is calculated through a cosine similarity analysis. This study observed that IRTARA had consistent results in the term frequency list based on the three evaluation methods for the high-quality results and GIRTARA was able to obtain up to 65% match in terms of cosine similarity when compared to the out-of-library object's true labels.

1. Introduction

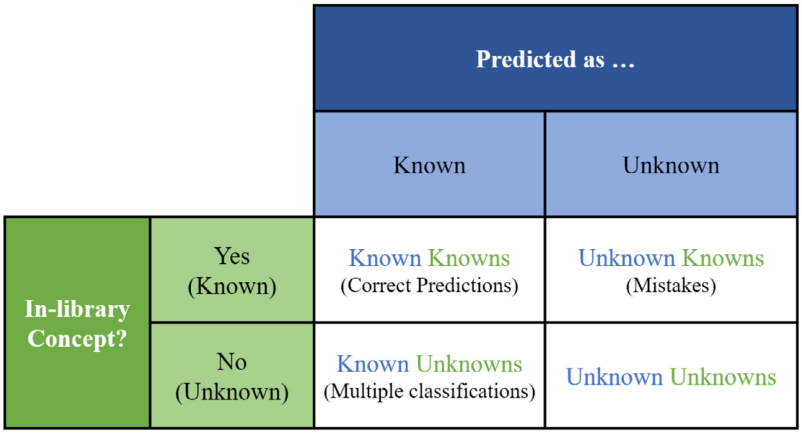
Artificial intelligence (AI) is often believed to be computers that can mimic human thought processes exactly; however, this is largely untrue among current-day AI (McCarthy, 2004). Most AI technology available today is "weak", which means it is limited to the tasks and datasets that it was originally trained on (IBM, 2024). When placed into practice, the AI can observe or come into contact with something (a situation, object, etc.) that it either knows or does not know. The result is that the AI interaction involves one of four categories of possible results as shown in Fig. 1 based on the algorithm's prediction (blue) and whether it has been pre-trained on the concept (green) (Situ et al., 2016; Combs, 2021). A known known is a correct prediction that the algorithm has been pre-trained to recognize (i.e., an in-library concept). An unknown known is a concept that the algorithm predicts as being unknown but it was an in-library concept, hence a mistake. A known unknown is a concept predicted as a known entity, but the algorithm was not pretrained to recognize it (an out-of-library concept). This suggests that there are multiple classifications of the concept. Finally, there are unknown unknowns, which are concepts the algorithm was not previously trained to recognize (out-of-library) and the algorithm recognizes that by predicting them as unknown concepts. Since unknown unknowns, called "true unknowns" from this point on, are outside of the scope of training data understanding how to correctly evaluate them is a critical step in the direction of "strong" AI, which can provide generalization in perception and cognition (IBM, 2024).

AI is a broad domain with many applications and methods, including those in healthcare, defense, and business. Due to the rise in popularity of applications such as handwriting recognition, depth perception, and augmented reality, the ability to accurately identify and describe images is of great importance (Google, 2021). In this context, true unknown handling would be considered images of objects the AI has not been previously trained on such as explored in zero-shot learning (see Socher et al., 2013; Pourpanah et al., 2022; Sun et al., 2021). One way to assist the transition to "strong" AI for computer vision is through integrating various types of machine learning algorithms and techniques.

One proposed method to quickly and accurately evaluate unknown unknown is analogical reasoning (Mitchell, 2021; Antic, 2022). Learning by analogies is a concept from cognitive science based on using information from the familiar "base" and extending this information onto an unfamiliar "target" (Gentner and Maravilla, 2018). The success of analogical reasoning in solving analogy problems has been proven in both the visual/pictorial (Polya, 1990; Zhang et al., 2019) and text/verbal spaces (French, 2002; Rogers et al., 2017). However, the vast majority of visual analogical reasoning has been focused on

* Corresponding author.
E-mail address: kara.combs.1@us.af.mil (K. Combs).

https://doi.org/10.1016/j.nlp.2024.100064
Received 11 December 2023; Received in revised form 4 March 2024; Accepted 15 March 2024



		Predicted as …	
		Known	Unknown
In-library Concept?	Yes (Known)	Known Knowns (Correct Predictions)	Unknown Knowns (Mistakes)
	No (Unknown)	Known Unknowns (Multiple classifications)	Unknown Unknowns

Fig. 1. Known and unknown data matrix.
Source: Modified from Combs (2021).

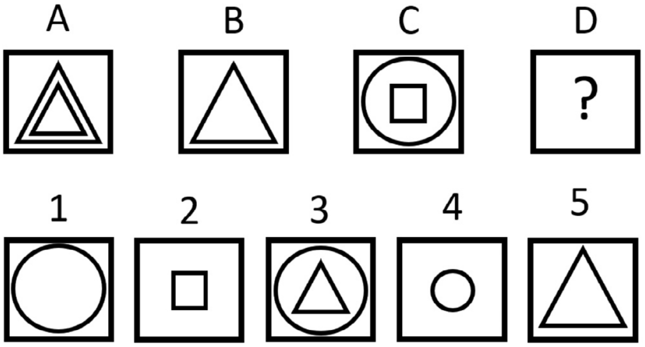


Fig. 2. Geometric visual analogy problem.
Source: Modified from Evans (1964).

novel geometric problems rather than real-world images exemplified by Fig. 2. Similar to what may be seen in an IQ test, when presented with the arrangement of geometric figures posed as "*A* is to *B* as *C* is to what numeric option?" and corresponding Options 1–5, the participant is tasked with selecting the correct option.

This paper aims to address the following research questions:

- RQ1: What is the current state of research on image-based analogical reasoning, generative AI, and their intersection?
- RQ2: How can analogical reasoning be leveraged in an unknown unknown computer vision scenario via an automated, repeatable process?
- RQ3: How can generative AI be leveraged to assist in the identification of objects from text-only descriptions?

RQ1 is addressed through a review of the background of computer vision methods and analogical reasoning algorithms and a discussion of their general capabilities (functional and algorithmic). Limited prior work exists in the image-based and image-to-text AR (Lu et al., 2019a; Sadeghi et al., 2015; Doumas & Hummel, 2010; Reed, Zhang, Yuting, & Lee, 2015; Hwang, Grauman, & Sha, 2013), with the vast majority of literature focusing on text-based AR (Gentner, 1983; Holyoak and Thagard, 1989; Hofstadter and Mitchell, 1995; Mikolov et al., 2013a,b; Pennington et al., 2014; Bojanowski et al., 2017; Hummel and Holyoak, 1997; Wilson et al., 2001). RQ2 is answered through the proposed Image Recognition Through Analogical Reasoning Algorithm (IRTARA), which provides descriptions of out-of-library objects (unknown unknowns described in Fig. 1. An analysis of IRTARA on classification image data with results and their interpretation through qualitative and quantitative measures is then presented and followed by conclusions. Our proposed solution to RQ3 is the generative AI expansion of IRTARA (dubbed "GIRTARA") which leverages generative AI to provide contextual information for a bag of words produced by IRTARA. Finally, we are interested in the connection between data policy and generative AI given the rise and overlap between both. The solutions proposed throughout this article are unique since they approach a computer vision problem using the cognitive science concept of analogical reasoning to derive new information without the need for additional training data and/or computational power.

This paper first begins with a background on analogical reasoning (2.1) and generative AI (2.2). Next, in Section 3, we discuss the common processes of IRTARA and GIRTARA (originally discussed in Combs et al. (2023a,b)). In Section 4, we describe the specific modules and parameters selected in our experiment for IRTARA and GIRTARA. Next, in Section 5, we present the metrics and results used to evaluate the results of IRTARA (the term frequency list) and the prediction of GIRTARA. Section 5 is split into two subsections first focusing on the evaluation methods and metrics for IRTARA via the definition, analogical reasoning, and human factors evaluations, and then, considering the cosine similarity metric used for evaluating the GIRTARA results. Finally, we end with future work and conclusions in Section 6.

2. Background

Presently, considerable AI utility has been seen in image data through convolutional neural networks (CNN), a form of an artificial neural network (see LeCun et al., 1989; He et al., 2016a,b; Liu et al., 2018). Even the most advanced deep CNN, unless integrated with another process, can only produce results that it was pre-trained on and aware of, i.e., not true unknown objects. One method with proven success in extrapolating new information is analogical reasoning which has seen limited image-based applications.

2.1. Analogical reasoning

Considerable emphasis has been on the development of analogical reasoning for text-based analogies with many algorithms developed to address the wide range of text-based analogy problems (Combs et al., 2022). These text-based problems range from novel word problems (e.g., *king:queen::man:woman*) to mapping sentence elements (e.g., "She is growing like a weed") to drawing parallels between stories (Ichien et al., 2020). Initially, analogical reasoning started as psychologically-based algorithms (see Gentner, 1983; Holyoak and Thagard, 1989; Hofstadter and Mitchell, 1995) but recently, with the rise of natural language processing, vector space models and artificial neural network approaches have increased in popularity (Combs et al., 2022). To date, the most prominent vector space models include Word2Vec (Mikolov et al., 2013a,b), Global Vectors (GloVe) (Pennington et al., 2014), and fastText (Bojanowski et al., 2017). Within the artificial neural network scope, models include Learning and Inference with Schemas and Analogies (Hummel and Holyoak, 1997) and Structured Tensor Analogical Reasoning 2 (Wilson et al., 2001). A handful of these algorithms were selected for an apples-to-apples comparison which showed the advantages and disadvantages of each (Combs et al., 2022).

The exploitation of these approaches in the image space has been limited to simply drawing visual analogies rather than applying analogical reasoning as a methodology. ANALOGY was arguably the first analogical reasoning algorithm that was designed to solve geometric analogy problems (Evans, 1964). The visual analogical reasoning space has largely been dominated by similar geometric-based problems such as Raven's Progressive Matrices (see Raven and Court, 1938). Of the remaining visual analogy algorithms, only a handful apply analogical reasoning to a computer vision-like problem. One example is the Visalogy, which can solve visual analogy problems that would be phrased as "A red car is to a blue car as a red bike is to what?" (Sadeghi et al., 2015). Though successful in its application, Visalogy is limited in regards to most analogies focused on the action, attribute, and/or orientation of an object (Sadeghi et al., 2015). Another visual analogy application is demonstrated by utilizing the semantic and visual aspects of an image from a visual analogy data set to solve *A:B::C:D*-like problems (Lu et al., 2019a). Despite the limitations of visual analogy data sets, there has been some research in utilizing generative AI to remedy this gap (Li et al., 2023; Ichien et al., 2023; Webb et al., 2023; Combs and Bihl, 2024). When evaluating text- and image-based AR methods, the former is a very thoroughly explored field compared to the latter. Challenges the latter faces include significant computational resources, additional processing and interpretation, and the tendency to be catered to an analogy-formatted image data set (meaning where the problem(s) to be solved are stated such that "Image A is to Image B as Image C is to what?"). A clear gap is the lack of image-based analogical reasoning applications using a general image data set, but also integrating it with a textual analogy to alleviate the computational and processing expectations typically associated with computer vision applications.

2.2. Generative AI

There has been a long history of generative-AI-like products starting with the first chat box, ELIZA, in 1966 to significant improvements in the 1995 ALICE (Artificial Linguistic Internet Computer Entity) system, all the way up to the recently popular ChatGPT (Rajarman, 2023). Concerning the data, information, knowledge, and wisdom hierarchy, where legacy systems were able to generate and interpret data and information, generative AI can include context, escalating its ability to "knowledge" (Bihl and Talbert, 2020). Generative AI has already had a significant impact on multiple domains including healthcare (Doshi et al., 2023; Singh, 2023; Raimondi et al., 2023; Eggmann et al., 2023), education (Rudolph et al., 2023; Fisk, 2023; Whalen and Mouza, 2023), and academic research (Crawford et al., 2023; Thorp, 2023; Huang and Tan, 2023). In the visual realm, AI-generated art has also become extremely popular and in some cases caused controversy (Boden and Edmonds, 2009; Galanter, 2016; Srinivasan and Uchino, 2021). Furthermore, generative AI is being explored to understand how it can be leveraged in programming tasks (Chen et al., 2021). To show the versatility of GPT-3, it was fine-tuned for coding and turned into Codex, which powered GitHub Copilot (later updated to utilize GPT-4 in its "X" version see Dohmke, 2023). Codex itself has since been depreciated as of March 2023; however, with the release of ChatGPT, it has been able to reproduce and improve on original results by Codex (OpenAI, 2023a,b,c,d,e,f). The analogical reasoning abilities of large language models (LLMs), or lack thereof, have already been explored in the literature (Webb et al., 2023; Yu et al., 2023; Mitchell et al., 2023; Petersen and van der Plas, 2023). Some of the most popular generative AI models for a variety of tasks are shown in Table 1. Generative AI's abilities have made it an all-around tool for day-to-day and even advanced tasks; however, there has been some discussion about limiting its seemingly unbounded and growing capabilities.

Taking a closer look at the generative AI language models, the most famous of these applications is OpenAI's ChatGPT, which reached over 100 million users two months after its initial launch in November 2022 (Hu, 2023) . The uniqueness of ChatGPT was its free availability to the public (with some limitations of daily use) in an easy-to-use chat-like interface. Later in February 2023, OpenAI released a premium version, ChatGPT Plus, for $20/mo. that gives subscribers more stable access during busy times, quicker response rates, and the ability to experience new features first (OpenAI, 2023a,b,c,d,e,f). One of the benefits of Plus is the ability to use the highly anticipated GPT-4 model; whereas, free-tier users are limited to GPT-3.5-Turbo (as of March 2023). ChatGPT uses a custom version of GPT-3.5-Turbo fine-tuned for conversations (OpenAI, 2022a,b). The standalone GPT-3.5 model is older and also 9/10th more expensive than its Turbo counterpart (OpenAI, 2023a,b,c,d,e,f). There are three non-Turbo versions of GPT-3.5: (1) text-davinci-003 (trained via reinforcement learning), (2) text-davinci-002 (trained via supervised fine-tuning), and (3) code-davinci-002 (fine-tuned for code-related tasks) (OpenAI, 2023a,b,c,d,e,f). One of the limitations of all the GPT-3.5 models (ChatGPT included) is the fact they are pretrained on data up to a certain date and unaware of anything that occurred afterward unless connected to a browser through an extension. However, that is not an issue for the other models being considered. Microsoft Bing utilizes a custom version of GPT-4 optimized for web searches (Mehdi, 2023; OpenAI, 2023a,b,c,d,e,f). More similar to ChatGPT Microsoft Bing Chat can be accessed through the Bing search engine and provides three conversation styles based on varying priorities: creativity, precision, and a balance between the two. Outside of OpenAI products, Google has also entered the market with its AI chatbot, Bard (Pichai, 2023). Upon its release in February 2023, Bard originally used Google's Language Model for Dialogue Applications (LaMDA) (see Collins and Ghahramani, 2021, but has since switched to versions of the Pathways Language Model (PaLM) (see Narang and Chowdhery, 2022; Google, 2023). You.com is a lesser-known search engine but has had an AI Chat component since December 2022 called "YouChat" (You.com, 2022, 2023a,b). YouChat runs off of multiple large language models, collectively called Chat, Apps, and Links (C-A-L) (Eliacik, 2023).

3. Methodology

Proposed to remedy this gap is the Generative-AI version of the Image Recognition through Analogical Reasoning Algorithm (GIRTARA), which is an expansion of the earlier Image Recognition through Analogical Reasoning Algorithm, IRTARA (Combs et al., 2023a,b). IRTARA integrated a computer vision algorithm that outputs declarations based on known classes and an analogical reasoning algorithm; however, GIRTARA takes it a step further by including a generative AI model that takes these declarations and searches for the meaning of the visual unknowns. One of the limitations of IRTARA is that the result is a bag of words called the "term frequency list", which describes the true unknown but never exactly identifies it (Combs et al., 2023a,b). However, it is difficult to derive meaningful text from words without knowing the relationship between them. This task has become easier with the rise of generative AI as discussed in Section 2. GIRTARA fills a gap for analogical reasoning applied to true unknown computer vision scenarios and has an additional advantage over former image-based methods by leveraging matured text-based analogical reasoning applications. IRTARA and GIRTARA have the same underlying framework where their input is raw image data and its output is a term frequency list as shown in Fig. 3. The first three steps to both algorithms are image classification, creation of class name word vectors, and application of analogical reasoning & knowledge extraction, which produces the

Table 1
Types of generative AI models.

Type	Creator (& model family)	Model name	Release date	Source(s)
Language Models	OpenAI Generative Pre-Trained (GPT) Model	GPT-1	Jun. 2018	Vaswani et al. (2017), Radford et al. (2018)
		GPT-2	Nov. 2019	Radford et al. (2019)
		GPT-3	May 2020	Brown et al. (2020)
		GPT-3.5	Mar. 2022	OpenAI (2022a,b)
		GPT-4	Mar. 2023	OpenAI (2023a,b,c,d,e,f)
	Google Language Model for Dialogue Applications (LaMDA)	LaMDA	May 2021	Collins and Ghahramani (2021)
		LaMDA 2	May 2022	Pichai (2022)
	Google Pathways Language Model (PaLM)	PaLM	Mar. 2023	Narang and Chowdhery (2022), Chowdhery et al. (2022)
		PaLM 2	May 2023	Google (2023), Ghahramani (2023)
	Meta Large Language Model Meta AI (LLaMA)	LLaMA 1	Feb. 2023	Meta AI (2023), Touvron et al. (2023)
		LLaMA 2	Jul. 2023	Meta (2023)
	You	$C - A - L$	May 2023	You.com (2022, 2023a,b)
	Infection	Infection-1	Jun. 2023	Inflection AI (2023)
Image Creators	OpenAI DALL-E	DALL-E	Jan. 2021	OpenAI (2021), Ramesh et al. (2022), Radford et al. (2021)
		DALL-E 2	Apr. 2022	OpenAI (2022a,b)
		DALL-E 3	Oct. 2023	OpenAI (2023a,b,c,d,e,f)
	Craiyon	Craiyon[a]	Jul. 2021	Dayma et al. (2021, 2022), Dayma and Cuenca (2023)
	Midjourney	Midjourney	Feb. 2022	Midjourney (2022)
	Stability AI	Stable Diffusion	Aug. 2022	Stability.AI (2022), Rombach et al. (2022))
Speech Recognition	OpenAI	Whisper	Sept. 2022	Radford et al. (2022)
Code Generators	OpenAI	Codex	Aug. 2021	Zaremba and Brockman (2021)
Multimodal	OpenAI ChatGPT	Pro	Oct. 2023	OpenAI (2023a,b,c,d,e,f)
	Google Bard with Gemini	Ultra Pro Nano	May 2023	Gemini Team (2023)

[a] Craiyon was originally named "DALL-E Mini" until June 2022.

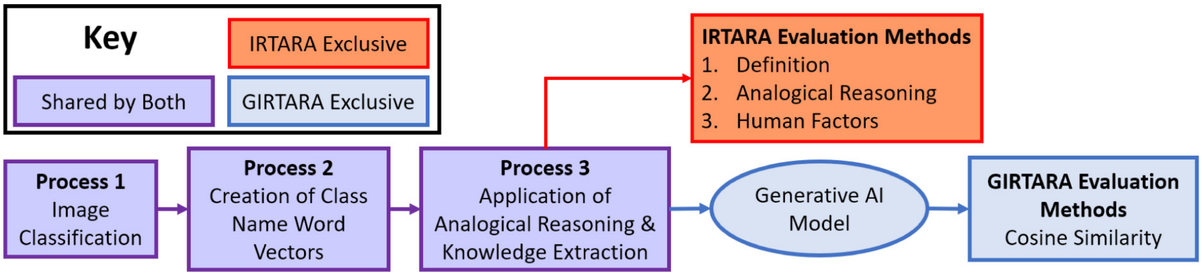


Fig. 3. GIRTARA and IRTARA framework.

"term frequency list". The term frequency list, a collection of words that describe the visual unknown, is immediately evaluated through 3 methods for IRTARA; however, GIRTARA sends the top-5 words to a generative AI model before evaluation via cosine similarity.

The remainder of this section covers the shared processes of IRTARA and GIRTARA, which are shown in Fig. 4. The first process involves taking an image data set with class labels and training a CNN as would typically happen in a computer vision problem. Later this CNN is used on an unknown class of images, which can be from the data set (by removing one of the classes as demonstrated later in this application) or can be sourced from an external data set. IRTARA is modular and can utilize any CNN architecture that can produce probabilities associated with how likely that unknown class image belongs in the number of classes in which it is trained, |C|. IRTARA is only interested in the top-k classes where $k \in C$.

In the second process, the pre-existing class names are converted into their respective word embedding vectors via the analogical reasoning algorithm. The algorithm sends the class names to the analogical reasoning algorithm, which retrieves the pre-trained word embedding vectors. In the case where the class name is not within the analogical reasoning algorithm's vocabulary, it may need to be altered into a "representative" version that is recognizable by the algorithm. These class name word vector representations are used with the image classification's class predictions in the next process.

In the third process, an "unknown word vector" needs to be created for each unknown class image to ideally "represent" the unknown class. For a given class, c, if the probability of the unknown class image belongs to said class, p_c, is greater than the threshold, α, its word vector representation is retrieved (from the immediate previous process), and it influences the unknown word vector. Mathematically stated, note that $c \in C$ and consider the set $I = \{c : 1 \leq c \leq k \text{ and } p_c \geq \alpha\}$, which consists of the top-k class indexes where the probability of a given class, p_c, is greater than the threshold, α. Given the scenario where none of the probabilities, p_c, are greater than α (and hence $I = \emptyset$), the class vector with the highest probability, denoted by v_i, will be the unknown word vector's only influence. In this scenario, i is equivalent

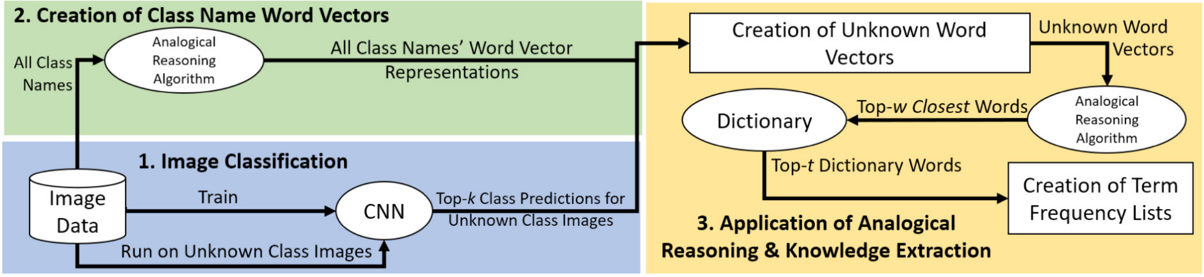

2. Creation of Class Name Word Vectors
All Class Names
Analogical Reasoning Algorithm
All Class Names' Word Vector Representations
1. Image Classification
Image Data
Train
CNN
Top-k Class Predictions for Unknown Class Images
Run on Unknown Class Images
Creation of Unknown Word Vectors
Unknown Word Vectors
Analogical Reasoning Algorithm
Top-w Closest Words
Dictionary
Top-t Dictionary Words
Creation of Term Frequency Lists
3. Application of Analogical Reasoning & Knowledge Extraction

Fig. 4. In-depth visualization of processes 1–3.

to the index of the class with the largest probability. The corresponding formula describes how the unknown word vector is calculated:

$$\text{Unknown Word Vector} = \begin{cases} \sum_{i\in I} v_i p_i \text{for } i \in I, I \neq \varnothing \\ v_i \text{ for } i \in \operatorname{argmax}\left\{p_1, p_2, \dots, p_C\right\}, I = \varnothing \ \forall c \in C. \end{cases} \tag{1}$$

The unknown word vector is sent back to the analogical reasoning algorithm, which is tasked with identifying the closest-w words whose word embeddings best match the unknown word vector. The algorithm is interested in the closest-w words, where w can range from one word to the entire vocabulary of the VSM, w_{max}.

Given the closest-w words, they are each sent to the selected dictionary, which retrieves each word's definition(s). In the case that the word identified by the analogical reasoning algorithm does not exist in the dictionary's vocabulary, it is skipped. The definitions are then modified to remove any "stop words" (as defined by Bird et al., 2009) in addition to other words that lack significant semantic meaning which is at the discretion of the experimenter. The remaining words, dubbed "definition words", are compiled. Starting with the creation of the unknown word vector, this entire process is repeated for each image within the unknown class. This yields a large list of words, which are filtered down to the top-t words based on how frequently they occur. Theoretically, t can range from 1 to the total number of unique words identified; however, t has a direct relationship with computation time. This list of the top-t words is the term frequency list and is the final product of the third process.

4. Module and parameter selection

IRTARA/GIRTARA, simultaneously referred to as "the algorithm" throughout the remainder of this section, requires four modules which are the data set, CNN, analogical reasoning algorithm, and dictionary. "Module" is a term referring to any portion of the algorithm that can be replaced with another data set, algorithm, or application. Briefly, and described below, the algorithm selected the following modular components, the Caltech-256 data set (Griffin et al., 2007), a shallow 11-layer CNN, the GloVe (Pennington et al., 2014) analogical reasoning algorithm, and the PyDictionary (Bora, 2020) and Lexico (Oxford University Press, 2021) dictionaries. Six parameters were selected by the authors in addition to one determined by the data set selected shown in Table 2. Only the number of CNN classes, p, is dictated outside of the user's control since it is reliant on the input image data set. The remaining parameters were selected based on brief, informal experimentation on a range of varying values.

Though not explicitly part of the process shown in Fig. 4, before the running of the algorithm, an appropriate data set needed to be selected. Ideal characteristics of a data set include (1) variety and scope of classes, (2) focus on single objects in an image, and (3) availability of baseline results. Based on these various needs, Caltech-256 (see Griffin et al., 2007) was selected to test this framework based on its variety of concrete classes and depth of samples per class. Caltech-256 has 257 classes, but for this study, the 257th "clutter" class was not considered, leaving 256 classes. In each iteration tested, the unknown class is taken from Caltech-256 classes so it is "known" to the experimenter, but "unknown" to the algorithm. This means that the CNN is trained on 255 classes, and then attempts to classify the remaining "unknown" class accordingly. For example, if the unknown class in the first iteration is "coffee mug", the coffee mug images would be set aside and the CNN would train on the remaining 255 classes. In the next iteration, if the unknown class was "American flag", the CNN would train on the remaining 255 classes, coffee mug included, and then attempt to classify the American flag images. Caltech-256 also included images of varying sizes and thus all were resized within the IRTARA algorithm to 128 x 128 pixels grayscale images.

Table 2
IRTARA parameters.

Parameter	Variable	Value
Number of CNN classes	C	256
Number of top classes per image	k	5
Threshold for influencing unknown word vector	α	0.05
Number of closest words per unknown word vector	w	5
Number words in the term frequency list	t	100
Number of primary words	u	20
Number of secondary words	v	10
Number of words in the modified term frequency list	b	21

The CNN is the primary concern of the image classification section (enumerated process 1) Fig. 4. To avoid the computational demands required in applying deep CNN architectures, an 11-layer CNN that balanced accuracy and computational performance was developed for algorithm demonstration purposes. The CNN had the following architecture shown in Fig. 5. When trained on all 256 classes, this architecture had an average of 22.5% classification accuracy across 10 runs (compared to 38% for Griffin et al. (2007) across 40 runs), where *optimizer* = Adam(), *batch_size* = 32, *epochs* = 10, and *validation_split* = 0.1. Despite a CNN with higher accuracy, e.g., ResNet (He et al., 2016a,b) or VGGNet (Simonyan and Zisserman, 2015), being likely to yield better results, such CNNs are computationally costly and the algorithm of (3) can be rapidly retrained to assess IRTARA. Thus, the development and demonstration of the algorithm focused on the whole process and used this simple CNN which trained quickly on standard desktop hardware.

To select the analogical reasoning algorithm (involved in processes 2 and 3 of Fig. 4) the review of analogical reasoning algorithms from Combs et al. (2022) was used and the following methods were considered: Bayesian Analogy with Relational Transformations (BART) 1.0 (Lu et al., 2012) and 2.0 (Lu et al., Emergence of analogy from relation learning, 2019b), 3 Cosine Average (3CosAvg) (Drozd et al., 2016), Distributed Representation Analogy MApper (DRAMA) (Eliasmith and Thagard, 2001), Linear Regression Cosine (LRCos) (Drozd et al., 2016), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov

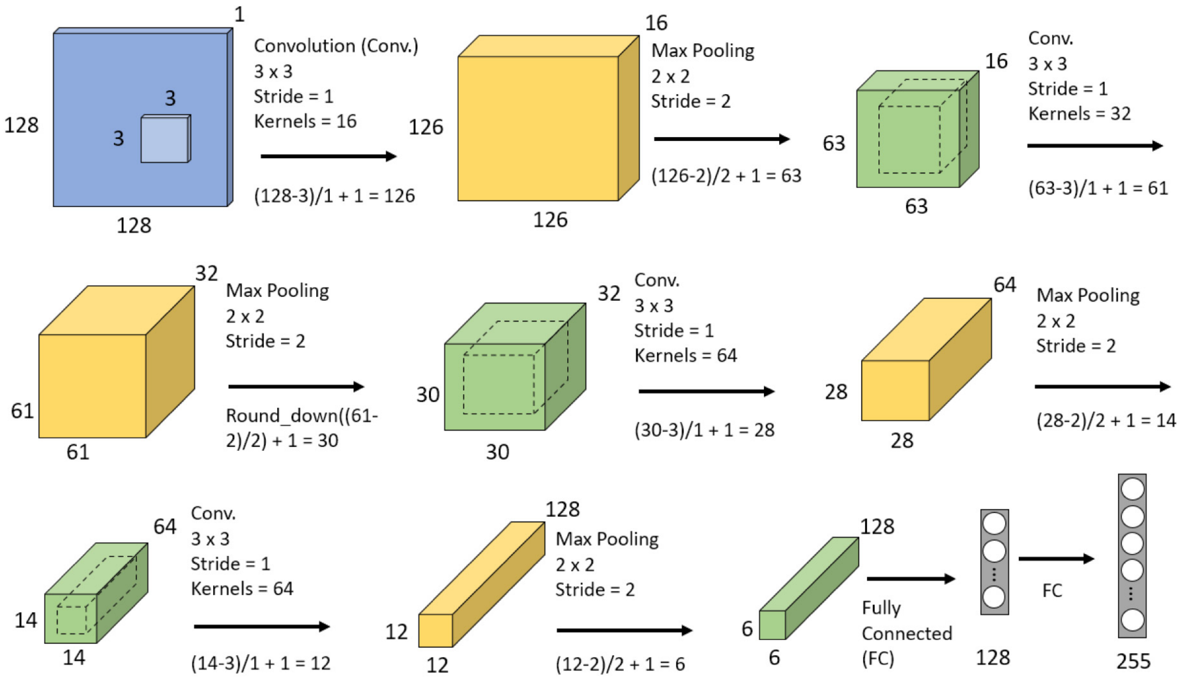


Fig. 5. IRTARA's Convolutional neural network architecture.

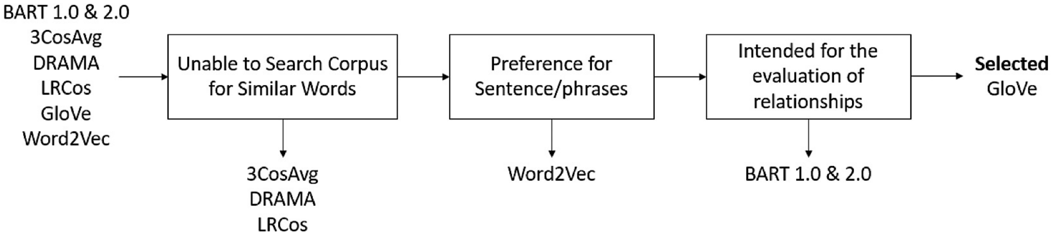

BART 1.0 & 2.0
3CosAvg
DRAMA
LRCos
GloVe
Word2Vec
Unable to Search Corpus for Similar Words
Preference for Sentence/phrases
Intended for the evaluation of relationships
Selected GloVe
3CosAvg
DRAMA
LRCos
Word2Vec
BART 1.0 & 2.0

Fig. 6. Analogical reasoning algorithm selection justification.
Source: Modified from Combs (2021)

et al., 2013a,b). To select an AR method for IRTARA, the adjusted correctness (based selection of the correct answer) and goodness (how close to an "ideal" analogy the correct answer is according to the algorithm) (Combs et al., 2022). This produced the ranking of methods seen in Fig. 6. From this process and other constraints, GloVe was selected based on that it does not utilize analogy relationships, it does use singular-word embeddings (as opposed to multi-word phrases), and its ease of implementation is visually shown in Fig. 6. Specifically, the Glove-wiki-gigaword-300 model was used, which was pre-trained on 2014 Wikipedia and Gigaword 5 textual data to create the word vector for 400,000 words, each with 300 dimensions (Pennington et al., 2014). In certain cases, GloVe was unaware of the original class name, so a substitute representation was used. These representations typically were a simplification (e.g., "American flag" to "flag") or a merge between two-word vectors representing different words (e.g., "baseball bat" to "baseball" & "bat") so as long they would not be confused with another class. This verification and revision process would need to be repeated whenever different word embeddings are used as well as when a different image data set is used.

The dictionary is important in process 3 of Fig. 4, application of analogical reasoning and knowledge extraction. The primary dictionary used was the external Python library, PyDictionary, based on ease of integration with the algorithm and standardized definition format. This dictionary was used in the third process of Fig. 4 and the algorithm evaluation is discussed in Section 3. PyDictionary uses WordNet (see Princeton University, 2010) for its definitions which were created in 1995. Considering how language has evolved and changed since then, some of the definitions came from an alternative dictionary, Lexico, which affected 16 classes. In addition to using an alternative dictionary, some of the definitions were created by simplifying the original class name (e.g., "self-propelled lawn mower" to "mower") and merging individual words' definitions (e.g., "cowboy hat"). Like with the GloVe word vector representations, the definitions would be verified and modified accordingly whenever a new dictionary is used and/or a different image data set.

Exclusive to GIRTARA, upon the creation of the term frequency lists in the third process, the top five words are sent to the generative AI algorithm, which predicts the unknown class. The analogical reasoning algorithm identifies the vector representation of the unknown class's name and the prediction and compares them via the cosine similarity as the final step. The five popular Generative AI algorithms were selected to be queried and evaluated: ChatGPT (OpenAI, 2022a,b), GPT-3.5 (Brown et al., 2020), Microsoft Bing (Microsoft, 2023), Google Bard (Pichai, An import next step on our AI journey, 2023), and YouChat (You.com, 2023a,b). Table 3 displays the generative AI algorithm used and its underlying model and version. Note that many of these algorithms have had significant updates over time and that the model size and version release date shown in Table 3 are specific to the one used in this study (conducted June 2023). Also, the underlying models may have a different number of parameters (model size) compared to

Table 3
Key generative AI algorithm comparison.

Generative AI algorithm	Language model & Version	Model size	Version-release date	End of training data
ChatGPT	GPT-3.5-Turbo	1.56B	Mar. 2023	Sept. 2021
GPT-3.5	GPT-3.5 text-davinci-003	1.37B	Nov. 2020	Jun. 2021
Microsoft Bing	GPT-4	(est.) 1T	May 2023	Aug. 2022[a]
Google Bard	PaLM 2	770M	May 2023	Mar. 2023[a]
YouChat (3.0)	$C-A-L$	Unknown	May 2023	Unknown

[a] Underlying models finished training on the designated date, but the algorithm is still able to learn through its interaction with humans and internet connection.

the version utilized in the algorithm to explain any seemingly obvious discrepancies.

5. Metrics & results

Given the differences in the output of IRTARA and GIRTARA, different evaluation methods and metrics are needed. The output of IRTARA is a list of words, which is evaluated through 3 methods: the definition evaluation, analogical reasoning evaluation, and the human factors evaluation. The definition and analogical reasoning evaluations are automated methods that are two different approaches that answer "How well does the term frequency list describe the visual unknown?" The definition evaluation automatically identifies words that are directly related to the unknown class; whereas, the analogical reasoning evaluation automatically detects associated words to the unknown class. The human factors evaluation method still answers the same question but also acts as a validation for the automated methods to ensure their results resemble human judgment regarding the term frequency list quality. GIRTARA takes a portion of the term frequency list and asks a generative AI model to identify the object the list describes. Therefore, instead of the list, the result is one word (or phrase in some cases), which was compared to the label of the unknown class via cosine similarity.

5.1. IRTARA evaluation processes

Fifteen Caltech-256 classes were chosen at random to be the unknown class, which yielded a wide range of results. Table 4 shows the results for each of the three evaluation metrics as well as the overall rank of the classes in terms of performance across all three (see Appendix C and E of Combs (2021) for full results). Broadly looking at the lists many words repeat across lists, which may be due to the dictionary's tendency to frequently use those words in its definitions. Each evaluation method yielded two metrics and a rank assigned to each class ranging from 1 to 15. In case of a tie in rank, the average of the places was used, i.e., if two classes were tied for third and fourth place, they would both receive a score of 3.5. Each evaluation method's rank contributed equally to the overall rank.

5.1.1. Definition method

This evaluation method compared the term frequency list to the words found in the unknown class's definition(s). This analysis was able to determine directly related words used to describe the unknown class. This method utilized two different metrics due to many words having multiple meanings. To establish a baseline, each unknown class's definition was retrieved from the same dictionary used in the previous process. Similar to the definition words described earlier, the same stop words and words deemed to lack semantic meaning were also unknown from each unknown class's definition(s). The remaining words are called "all words" since they are the words found in all the definitions for the unknown class. If multiple definitions, the true definition is identified and the words found within are called "true words". In this case, there is only one definition, the all and true word lists are identical. The term frequency list is compared against the unknown class's true and all words. The "best-case scenario" is for the term frequency list to overlap with the true words because these accurately describe the unknown class. However, the "all words" are also considered in case IRTARA can pick up other meanings of the unknown class (if applicable). Since each unknown class's definition has varying word length, the two metrics from this method are expressed as percentages, namely the true words and all words percentages.

The definition evaluation produced two metrics that looked at the percentage of words in the term frequency list that also appear "true" definition, the True Words Metric, and those that appear in any definition of the unknown class, the All Words Metric. The percentage of true words ranges from 0%–50%; whereas, the percentage of all words range was slightly lower, between 0% to 33.3%. The top three performing classes were skyscraper, t-shirt, and iguanas. The classes were ordered from highest to lowest based on the true words and all words percentage and assigned a rank between 1 and 15. An average of these rankings was taken, ordered, and ranked again for the definition evaluation rank shown in Table 4.

5.1.2. Analogical reasoning method

The analogical reasoning evaluation method seeks to identify associated words to the unknown class that may not appear directly in the definition. This evaluation method has two metrics looking at the overlap between the term frequency list and the unknown class's primary and secondary words as determined by the analogical reasoning algorithm. Primary words are the top-u words closest to the unknown class based on cosine similarity. Furthermore, the top-v words closest to the top-u primary words based on cosine similarity are considered the number of secondary-per-primary words. There can be up to $u{\cdot}v_{max}$ secondary words; however, after excluding word variations and duplicates, it is usually less than $u{\cdot}v_{max}$. The term frequency list ideally overlaps with primary words; however, secondary words are also related to the unknown class to a lesser degree. The metrics that use this information are the percentage of primary words and the percentage of secondary words which looks at how many words in the term frequency list are also primary or secondary words, respectively.

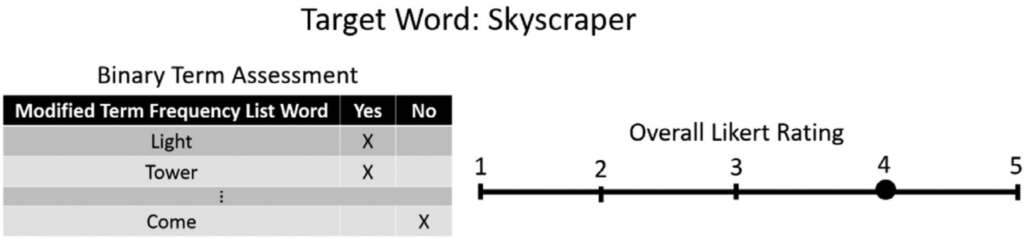
The analogical reasoning evaluation also produced two metrics that looked at the percentage of words in the term frequency list that also appeared as primary or secondary words, called "Pri. Words" and "Sec. Words" respectively in Table 4. Around the board, most classes had low scores for both metrics; however, the top 3 in both were Mars (15%; 5.4%), galaxy (15%; 3.8%), and skyscraper (15%; 2.7%). Similarly, to the definition evaluation rank, there was a rank assigned to each class based on its percentage of primary and secondary words. An average of these ranks was used to order and rank the classes for the analogical reasoning rank shown in Table 4.

5.1.3. Human experiment

The final and only non-automated method looks at how a human judges the quality of the term frequency list. The human experiment only considers a modified term frequency list, which consists of the top-b words of the original term frequency list. The human experiment yields two metrics: descriptive words the number of words the majority of respondents deem descriptive for an unknown class (based on the "binary term assessment" portion of the survey), and a quality score which reflects how well the descriptive words describe the unknown class (based on the "overall Likert rating" portion of the survey). Fig. 7

Table 4
IRTARA evaluation results.

True unknown	Definition evaluation			Analogical reasoning evaluation			Human factors evaluation			Overall rank
	Tue words	All words	Rank	Pri. words	Sec. words	Rank	Desc. words	Quality score (Avg ±Stdev)	Rank	
Ak-47	0%	0%	14	0%	0%	14	7	2.04 ± 0.77	14.5	15
Chandelier	0%	0%	14	0%	0%	14	5	2.52 ± 0.95	13	14
Fireworks	12.5%	12.5%	7	0%	1.7%	9	14	3.62 ± 0.85	2	5
Floppy Disk	0%	6.7%	11.5	0%	0.9%	11	9	2.81 ± 1.02	10	11.5
Frog	10%	15%	8	0%	2.4%	6	11	2.92 ± 0.9	7.5	7
Galaxy	0%	0%	14	15%	3.8%	2	18	4.5 ± 0.81	1	4
Iguanas	31.3%	31.3%	3	0%	2.1%	7	13	3.15 ± 0.89	4	3
Mars	23.5%	20.6%	5	15%	5.2%	1	4	2.58 ± 1.14	11	6
People	0%	16.7%	10	5%	2.2%	4	9	2.69 ± 1.05	7.5	9
Rainbow	11.1%	9.1%	9	0%	1.8%	8	9	3.42 ± 0.81	5	8
Sheet Music	0%	6.7%	11.5	0%	1%	10	7	2.08 ± 1.06	15	13
Skyscraper	50%	50%	1	15%	2.7%	3	8	3.35 ± 0.89	6	2
Swiss Army Knife	16.7%	16.7%	6	0%	0%	14	9	2.65 ± 0.85	9	10
T-shirt	40%	33.3%	2	5%	1.5%	5	12	3.27 ± 1.12	3	1
Waterfall	25%	25%	4	0%	0.8%	12	5	2.27 ± 0.96	12	11.5
Average	13%	14%		3%	2%		9.24	2.91 ± 0.93		


Target Word: Skyscraper

Binary Term Assessment

Modified Term Frequency List Word	Yes	No
Light	X	
Tower	X	
⋮		
Come		X

Fig. 7. Example survey section, modified from Combs et al. (2022).

is an example of how the questions for both portions were presented to the respondents.

The Binary Term Assessment determines how many of the modified term frequency list words "describe" the target word, which is the class man. Looking at the example in Fig. 7's Binary Term Assessment section, the respondent is presented with the target word, "skyscraper", and words from the modified term frequency list. The respondent would go through the latter words one-by-one and answer "yes" or "no" to whether, "by itself, in combination of another listed word or its characteristics, [would the given modified term frequency list word] describe or could be associated with [the unknown class]?" Returning to the example, the respondent would ask the aforementioned question for each word "light", "tower", etc., and mark their respective answers in the column like what is shown in Fig. 7. The number of "Yes's" was compiled for each of the b-word(s) and then, the number of modified term frequency list words where a majority (defined as 50% + 1) of respondents said "Yes" were summed as the unknown class's descriptive words. This process was dubbed the "binary term assessment".

The quality score had the respondents look at the descriptive words identified in the previous step and give a ranking between 1 and 5 on a Likert scale regarding how well they, as a whole, described the unknown class. As shown on the right side of Fig. 7, the respondents were given a slider that could accept values between 1 and 5 inclusive. This value is averaged across all respondents and the standard deviation was also calculated. This process was called the "overall Likert rating".

These results were derived from a homework assignment given to a mix of 25 undergraduate/graduate students enrolled in a Midwestern university's introductory human factors engineering class. The class consisted of 10 graduate and 15 undergraduate students all of whom were pursuing a degree within the biomedical, industrial, and human factors engineering department. Of the 25 subjects, 7 were male and 18 were female. The results reflect a 26th respondent, which are the opinions of the first author who was also a female graduate student in the department. The descriptive words metric ("Desc. Words" column in Table 4), looking at the number of words at least (50% + 1) respondents thought were relevant, ranged from 4–18 words (out of 21 total words in the modified term frequency list). This metric showed the top classes to be galaxy (18), fireworks (14), and iguanas (13). The second metric from the human experiment, the quality score, ranged from 1 to 5, which was averaged across all 26 respondents' responses. Most of the classes are statistically similar to one another, with the exception being the galaxy class with a score between 3.69–5. A rank was calculated for both metrics like in the previous methods. The quality score rank was based only on the average quality score. These rankings were averaged and ranked to calculate the human experiment rank for this method. The top three classes saw galaxy rank first followed by the tied classes, fireworks, and t-shirt. These results are elaborated in Combs (2021), Combs et al. (2023a,b) The top three classes, t-shirt, skyscraper, and iguanas, consistently scored in the top 50% of the rankings for all three evaluation ranks. Whereas, at the bottom, floppy disk (11.5), sheet music (13), chandelier (14), and AK-47 (15) consistently ranked in the bottom 50% for all three evaluation methods. The best example with varying results is galaxy, which performed poorly on the definition evaluation (14), but well on the analogical reasoning evaluation (2) and the human experiment (1). These results suggest that IRTARA performs consistently at the extreme ends but to a lesser extent with mid-range results. Spearman's rank coefficient was calculated for both automated methods in comparison with the human experiment. The definition-human evaluation yielded a ρ = 0.168 and *p-value* = 0.549 and the analogical-reasoning-human evaluation yielded a ρ = 0.434 and *p-value* = 0.082.

5.2. GIRTARA analysis & results

The first 5 words of IRTARA's term frequency lists for each true unknown (shown in Table 5) were used to query each generative AI algorithm. The first four algorithms in Table 3 were presented with the prompt, "What object does [word 1], [word 2], [word 3], [word 4], and

Table 5
Top-5 words produced by IRTARA for each true unknown.

	AK-47	**Chandelier**	**Fireworks**	**Floppy disk**	**Frog**
1	Long	Small	Large	Small	Large
2	Small	Observe	Small	Ball	Body
3	Move	Person	Long	Body	Fungi
4	Person	Determine	Cloud	Long	Small
5	Played	Light	Light	Device	Edible
	Galaxy	**Iguana**	**Mars**	**People**	**Rainbow**
1	Planet	Long	Brain	Large	Light
2	Sun	Small	Skull	Body	Little
3	Mythology	Large	Nervous	Ball	Illumination
4	Th[a]	Coat	Ability	Move	Fire
5	Small	Genus	Planet	Small	United
	Sheet Music	**Skyscraper**	**Swiss Army Knife**	**T-shirt**	**Waterfall**
1	Small	Light	Small	Ball	Fungi
2	Rectangular	Tower	Ball	Light	Large
3	Area	Small	Instrument	Game	Fleshy
4	Glass	Building	Body	Face	Body
5	Box	Little	Device	Small	Edible

[a] "Th" is about numbering such as 4th; however, all numerical characters were removed in preprocessing.

[word 5] describe?" YouChat often insisted that the words had multiple common links, so a more specific prompt was used for it, that is "What is your best guess of an object that is described by [word 1], [word 2], [word 3], [word 4], and [word 5]?" The results of each algorithm are shown in Table 6 in the corresponding "Prediction" column for each algorithm.

Under each algorithm is also a similarity percentage to determine how similar the true unknown is to the predicted word. This analysis used the word vector representation created by the GloVe-wiki-gigaword-300 model, which was already used as the algorithm's analogical reasoning algorithm. GloVe represents each word as a 300-dimensional vector (Pennington et al., 2014). If the prediction was a multi-word phrase, the most representative word (according to the authors) was used, which is denoted by the non-bracketed words in Table 6. In some cases, concerning the true unknowns, a completely different word was used (e.g., "Kalashnikov" instead of "AK-47") due to GloVe not being aware of the true unknown. The cosine similarity is a common metric for evaluating how alike text is (see Combs et al., 2022; Rogers et al., 2017). For two general vectors, v_1 and v_2, their cosine similarity is given by the following equation:

$$CosineSimilarity\left(v_1, v_2\right) = \frac{v_1 \cdot v_2}{\left\|v_1\right\| \left\|v_2\right\|}. \tag{2}$$

A cosine similarity of 0 means that the two vectors, v_1 and v_2, are dissimilar, and a cosine similarity of 1 means that v_1 and v_2 are exactly alike. Ideally, we would like the prediction word and the true unknown to have a cosine similarity of 1. However, that is not always the case as seen in Table 6. The highest similarity percentage (shortened to "Sim. (%)") of any algorithm obtained was 65% achieved by Google Bard which predicted "comet" for the true unknown, "galaxy". However, the vast majority of the other predictions fell quite short of this. The averages across all the true unknowns evaluated are shown in the bottom row of Table 6 with Microsoft Bing in the lead with 27%, followed by Google Bard (24%), ChatGPT (21.8%), GPT-3.5 (18.5%), and YouChat (17.1%), respectively.

However, the results are not statistically different from one another given the large standard deviations ranging between 10.3–18.5%. A box and whisker plot of each algorithm's results are shown in Fig. 8. The average for each algorithm is marked with a dashed line with the median shown via the solid line.

Each of the Generative AI algorithms produced similar results which spanned a variety of similarity scores, but the average similarity score across all the true unknowns and algorithms was 21.7%. There was a total of 6 predictions that had a similarity score equal to or above 50% shown in Table 7. Interestingly for the true unknown, iguana, both snake and panther yielded the same similarity score despite the *CosineSimilarity(Snake, Panther)* = 26.4%. Another interesting aspect is that similarity does not necessarily represent truth. Returning to the iguana example, an iguana is not a snake nor a panther, but an iguana is a type of animal. However, the animal received a similarity score of 4% lower than the snake/panther.

Specific to the algorithms, some of the predictions were more specific than others. YouChat in particular stood out in this regard exemplified by its prediction of "Homoneanderthalensis fossil" for "Mars" and "torsion balance apparatus" for "chandelier". On some occasions, best shown by the t-shirt true unknown, the predictions would include alternative names for the same object such as "table tennis ball" and "ping pong ball" or specific types of objects such as "medicine ball" and "light-up ball". These made the similarity analysis difficult because GloVe has a limitation to one word (which was initially identified as an advantage in Fig. 6). Therefore, each multi-word prediction had to be reduced to a more representative word at the discretion of the experimenters. This often caused two different predictions to have the same similarity score such as "glass display box" and "glass jewelry box" both being reduced to simply "box" (see sheet music row in Table 5. This also was difficult for the true unknowns, particularly "sheet music". Proposed and implemented in Combs (2021), Combs et al. (2023a,b) one remedy would be to combine the word vector for "sheet" and "music;" however, this is not always representative of the phrase as a whole.

6. Conclusions & future work

Of significant interest to artificial intelligence (AI) research is to accurately interpret and describe out-of-library (true unknown) objects (Situ et al., 2016). Analogical reasoning has been proposed to assist with this end goal in both textual and visual scenarios; however, there has been limited research conducted regarding its application in computer vision problems. To answer our first research question, we present a literature review of image-based analogical reasoning and generative AI in Section 2. Related to research question #2, we present the Image Recognition Through Analogical Reasoning Algorithm (IRTARA), which integrates standard image classification methods from computer vision with the semantic meaning and interpretation from an analogical reasoning algorithm and dictionary initially described in Combs et al. (2023a,b). To answer research question #3, we proposed the "generative AI" version of the (GIRTARA), which is built upon the earlier IRTARA model but adds a generative AI module that takes the term frequency list and identifies a single object that the words describe.

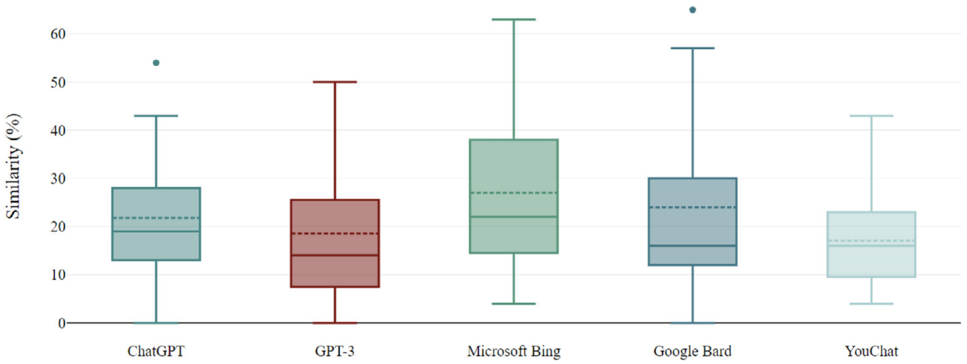


Fig. 8. Similarity box and whisker plots.

Table 6
Generative AI IRTARA results.

True unknown	OpenAI ChatGPT		OpenAI GPT-3.5		Microsoft Bing		Google Bard		YouChat	
	Prediction	Sim. (%)	Prediction	Sim. (%)	Prediction	Sim. (%)	Prediction	Sim. (%)	Prediction	Sim. (%)
[AK-47] Kalashnikov	Flute	11	Yo-yo	6	Violin	14	Snake	10	Toy [Car]	19
Chandelier	Microscope	14	Microscope	14	Microscope	14	Microscope	14	Microscope	14
Fireworks	Lightning [bolt]	15	Rainbow	26	Rainbow	26	Contrail	15	Lightning [bolt]	15
[Floppy] Disk	Smartphone	24	Eyewear [stylus]	11	[Mini-exercise] ball	15	Slinky	8	Torsion [balance apparatus]	16
Frog	Mushroom	43	Mushroom	43	Mushroom	43	[Puffball] mushroom	43	Mushroom	43
Galaxy	Mercury	17	[Solar system] model	9	Planet	63	Comet	65	Mercury	17
Iguana	Snake	54	Animal [or plant species]	50	[Black] panther	54	Mammoth	16	[Sloth] bear	24
Mars	Human	19	[Human] body	2	Human	19	Earth	57	[Homo neanderthalensis] fossils	27
People	Earth	9	Soccer [ball]	6	[Medicine] ball	4	Planet	9	[Medicine] ball	4
Rainbow	Candle	25	Candle	25	Candle	25	Candle	25	Candle	25
[Sheet] Music	[Picture] frame	0	[Picture] frame	0	[Glass display] box	6	[Photo] frame	0	[Glass jewelry] box	6
Skyscraper	Lighthouse	31	Lighthouse	31	Cupola	33	Lighthouse	31	[Small tower] light	5
[Swiss Army Knife] Penknife	Maraca	31	[Massage/ fitness] ball	21	[Pilates] ball	21	Stethoscope	29	[Ball-]bearing	7
T-shirt	[Table tennis/] ping pong [ball]	12	Ping pong [ball]	12	[Light up/ small solar] ball	16	[Light-up] ball	16	Ping pong [ball]	12
Waterfall	Mushroom	22	Mushroom	22	Mushroom	22	Mushroom	22	Mushroom	22
Avg. Similarity	21.8%		18.5%		27%		24%		17.1%	
Similarity Std. Dev.	13.8%		14.6%		17.8%		18.5%		10.3%	

Table 7
Top similarity scores.

Rank	True unknown	Prediction	Similarity score
1	Galaxy	Comet	65%
2	Galaxy	Planet	63%
3	Mars	Earth	57%
T-4	Iguana	Snake	54%
T-4	Iguana	[Black] Panther	54%
6	Iguana	Animal	50%

Both IRTARA and GIRTARA results were evaluated through different evaluation methods given their outputs. For IRTARA, the quality of the term frequency list was measured through two automated methods, the definition, and analogical reasoning evaluations, both to be compared to the results from the human experiment. The definition evaluation method identifies directly related words as found in the unknown class's definition; whereas, the analogical reasoning considers associated words related to the unknown class's label. The human evaluation exists to create a baseline for how a human might judge the term frequency list compared to the two automated methods. Since GIRTARA produces a prediction of the unknown class name, it is compared to the true class name via a cosine similarity analysis. IRTARA and GIRTARA were tested with the Caltech-256 dataset (Griffin et al., 2007), with parameters consistent with those found in Table 2. Overall, the three evaluation methods show consistency with unknown classes that have performed on the extreme ends (exceptionally well or poorly). Using the Spearman rank coefficient to compare how well the automated methods match the human experiment ranks, it was determined that the analogical reasoning evaluation ranks had a higher correlation ($\rho = 0.43$; *p-value* $= 0.08$) compared to the definition evaluation ranks ($\rho = 0.17$; *p-value* $= 0.55$). In regards to GIRTARA five popular generative AI models were selected: ChatGPT, GPT-3.5, Microsoft Bing, Google Bard, and YouChat. The highest cosine similarity between a prediction and a true unknown was 65% (prediction of "comet" for "galaxy"). In summary, generative AI is a powerful tool that is useful for its ability to create context and effectively turn data and knowledge into insights. Unlike current methods and algorithms, the advantages of IRTARA and GIRTARA are that they provide a semantic understanding of visual unknowns (via analogical reasoning) without prior training on the observed concepts.

There are several directions for future studies to better take advantage of the abilities of generative AI. Generative AI is stochastic in the sense that the same prompt may generate different answers. Doing an experiment considering multiple runs to see what the most occurring prediction is rather than a one-time take may provide a more accurate overview. Different prompts will also yield different results. For example, several of the generative AI models would attempt to avoid a definite answer. For example, Google Bard would say, "They all have multiple meanings, depending on the context in which they are used". YouChat would often answer "[The words] do not appear to describe a common word or concept" hence why the prompt was edited for that particular algorithm. Therefore, it is important to put more consideration into the prompt being used to ensure the best possible results are being generated. It would also be of value to expand the generative AI algorithms selected for analysis as the number is growing each day and current ones are improving. Finally, tweaking the original IRTARA/GIRTARA parameters and modules to potentially produce more representative term frequency lists would likely improve the predictions made by generative AI. Generative AI models are often stochastic meaning that the same input may derive a different output, which is not always ideal. A quantification analysis looking at how to measure this would be a good foundation for future work. Another avenue of interest is how the original visual aspect may help IRTARA/GIRTARA in deciphering the image through the lens of being a multi-model problem. Most of this work was done through text and natural language processing, but there may be benefits to using a vision model to help guide the textual descriptions (in the case of IRTARA) and/or the textual prediction (for GIRTARA).

CRediT authorship contribution statement

Kara Combs: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Trevor J. Bihl:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Subhashini Ganapathy:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily represent the views of any part of the US Government, US Department of Defense, nor the US Air Force. This work was cleared for unlimited release under: SAF/PA-2023-0745/AFRL-2023-3583.

Funding statement

This research was supported by the Air Force Research Laboratory, USA through the Sensing, Leaning, Autonomy, and Knowledge Engineering (SLAKE) contract (FA8650-19-C-1692). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Antic, C., 2022. Analogical proportions. Ann. Math. Artif. Intell. 90, 595–644.

Bihl, T.J., Talbert, M., 2020. Analytics for autonomous C4ISR within e-Government: A research agenda. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. HICSS, Maui, pp. 2218–2227.

Boden, M.A., Edmonds, E.A., 2009. What is generative art? Digit. Creativity 20 (102), 21–46.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Bora, P., 2020. Pydictionary 2.0.1. Retrieved from PyPi: https://pypi.org/project/PyDictionary/.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al., 2020. Language models are few-shot learners. In: Advances in Neural Information Processing Systems. Virtual.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Ponde de Oliveira Pinto, H., Kaplan, J., et al., 2021. Evaluating large language models trained on code. pp. 1–35, arXiv preprint arXiv:2107.03374.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., et al., 2022. PaLM: Scaling language modeling with pathways. pp. 1–87, arXiv preprint arXiv:2204.02311.

Collins, E., Ghahramani, Z., 2021. LaMDA: Our breakthrough conversation technology. Retrieved from Google The Keyword: https://blog.google/technology/ai/lamda/.

Combs, K.L., 2021. Application of Analogical Reasoning for Use in Visual Knowledge Extraction. Wright State University. OhioLINK: Electronic Theses and Dissertation Center.

Combs, K., Bihl, T.J., 2024. A preliminary look at generative AI for the curation and evaluation of verbal-to-visual analogies. In: Proceedings of the 57th Hawaii International Conference on System Sciences. Honolulu: HICSS.

Combs, K., Bihl, T.J., Ganapathy, S., 2023a. Integration of computer vision and semantics for characterizing unknowns. In: Proceedings of the 56th Hawaii International Conference on System Sciences. Maui.

Combs, K., Bihl, T.J., Ganapathy, S., Staples, D., 2022. Analogical reasoning: An algorithm comparison for natural language processing. In: Proceedings of the 55th Hawaii International Conference on System Sciences. HICSS.

Combs, K., Ganapathy, S., Bihl, T., 2023b. Human factors evaluation of automated semantic characterization of out-of-library images for the design for explainable AI systems. In: IISE Annual Conference & Expo. IISE, New Orleans.

Crawford, J., Cowling, M., Aston-Hay, S., Kelder, J.-A., Middleton, R., 2023. Artificial intelligence and authorship editor policy: ChatGPT, bard, bing AI, and beyond. J. Univ. Teach. Learn. Pract. 20 (5), 1–13.

Dayma, B., Cuenca, P., 2023. DALL-e mini - Generative images from any text prompt. Retrieved from Weights & Biases: https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-mini-Generate-images-from-any-text-prompt--VmlldzoyMDE4NDAy.

Dayma, B., Patil, S., Cuenca, P., Saifullah, A.T., Le, P., Melas, L., Ghosh, R., 2022. DALL-E mini explained. Weights & Biases.

Dayma, B., Patril, S., Cuenca, P., Saifullah, K., Abraham, T., Le Khac, P., et al., 2021. DALL-E mini. http://dx.doi.org/10.5281/zenodo.5146400.

Dohmke, T., 2023. GitHub Copilot X: The AI-powered developer experience. Retrieved from GitHub Blog: https://github.blog/2023-03-22-github-copilot-x-the-ai-powered-developer-experience/.

Doshi, R., Amin, K., Khosla, P., Bajaj, S., Chheange, S., Forman, H.P., 2023. Utilizing large language models to simplify radiology reports: A comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. pp. 1–23, medRxiv.

Drozd, A., Gladkova, A., Matsuoka, S., 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. Osaka.

Eggmann, F., Weiger, R., Zitzmann, N.U., Blatz, M.B., 2023. Implications of large language models such as ChatGPT for dental medicine. J. Esthet. Restor. Dent. 1–5.

Eliacik, E., 2023. You.com's AI-powered features are already in use, unlike Google. Retrieved from Dataconomy: https://dataconomy.com/2023/02/21/you-com-ai-chatbot-ai-image-generator-how/.

Eliasmith, C., Thagard, P., 2001. Integrating structure and meaning: A distributed model of analogical mapping. Cogn. Sci. 25 (2), 245–286.

Evans, T.G., 1964. A heuristic program to solve geometric-analogy problems. In: Proceedings of the April 21–23, 1964, Spring Joint Computer Conference. New York City.

Fisk, R., 2023. The rise of ChatGPT and generative A.I. and what it means for schools. AASA J. Scholarsh. Pract. 20 (1).

French, R.M., 2002. The computational modeling of analogy-making. Trends Cogn. Sci. 6 (5), 200–205.

Galanter, P., 2016. Generative art theory. In: Paul, C. (Ed.), A Companion To Digital Art. John Wiley & Sons, Oxford, pp. 146–180.

Gemini Team, 2023. Gemini: A Family of Highly Capable Multi-Modal Models. Google.

Gentner, D., 1983. Structure-mapping: A theoretical framework for analogy. Cogn. Sci. 10 (3), 277–300.

Gentner, D., Maravilla, F., 2018. Analogical reasoning. In: Ball, L.J., Thompson, V.A. (Eds.), International Handbook of Thinking & Reasoning. Psychology Press, New York, pp. 186–203.

Ghahramani, Z., 2023. Introducing PaLM 2. Retrieved from Google The Keyword: https://blog.google/technology/ai/google-palm-2-ai-large-language-model/.

Google, 2021. Google research. Retrieved from Perception: https://research.google/teams/perception/.

Google, 2023. PaLM 2 Technical Report. Google, Retrieved from Google AI: https://ai.google/discover/palm2.

Griffin, G., Holub, A., Perona, P., 2007. Caltech-256 Object Category Dataset. California Institute of Technology, Pasadena.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas.

Hofstadter, D.R., Mitchell, M., 1995. The copycat project: A model of mental fluidity and analogy-making. Adv. Connect. Neural Comput. Theory 2, 205–267.

Holyoak, K.J., Thagard, P., 1989. Analogical mapping by constraint satisfaction. Cogn. Sci. 29, 5–355.

Hu, K., 2023. ChatGPT sets record for fastest-growing user base - analyst note. Retrieved from Reuters: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Huang, J., Tan, M., 2023. The role of ChatGPT is scientific communication: Writing better scientific review articles. Am. J. Cancer Res. 13 (4), 1148–1154.

Hummel, J.E., Holyoak, K.J., 1997. Distributed representations of structure: A theory of analogical access and mapping. Psychol. Rev. 104 (3), 427.

IBM, 2024. What is artificial intelligence (AI)? Retrieved from IBM: https://www.ibm.com/topics/artificial-intelligence.

Ichien, N., Liu, Q., Fu, S., Holyoak, K.J., Yuille, A.L., Lu, H., 2023. Two computational approaches to visual analogy: Task-specific models versus domain-general mapping. Cogn. Sci. 47 (9), e13347.

Ichien, N., Lu, H., Holyoak, K.J., 2020. Verbal analogy problem sets: An inventory of testing materials. Behav. Res. Methods 52 (5), 1803–1816.

Inflection AI, 2023. Inflection-1. Inflection.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1 (4), 541–551.

Li, J., Dong, S., Gong, Y., He, Y., Wei, X., 2023. Analogical learning-based few-shot class-incremental learning. IEEE Trans. Circuits Syst. Video Technol.

Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., et al., 2018. Progressive neural architecture search. In: Proceedings of the European Conference on Computer Vision. ECCV, Springer, Munich, pp. 19–34.

Lu, H., Chen, D., Holyoak, K.J., 2012. Bayesian analogy with relational transformations. Psychol. Rev. 119 (3).

Lu, H., Liu, Q., Ichien, N.Y., Holyoak, K.J., 2019a. Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. In: Proceedings of the 41st Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, pp. 1–7.

Lu, H., Wu, Y.N., Holyoak, K.J., 2019b. Emergence of analogy from relation learning. Proc. Natl. Acad. Sci. 116 (10), 4176–4181.

Mehdi, Y., 2023. Confirmed: The new Bing runs on OpenAI's GPT-4. Microsoft Bing Blogs. Retrieved from https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4?ref=upstract.com.

Meta, 2023. Meta and microsoft introduce next generation of llama. Retrieved from Meta Newsroom: https://about.fb.com/news/2023/07/llama-2/.

Meta AI, 2023. Introducing LLaMA: A foundational, 65-billion-parameter large language model. Retrieved from Meta AI Blog: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/.

Microsoft, 2023. Bing chat. Retrieved from Microsoft Edge: https://www.microsoft.com/en-us/edge/features/bing-chat.

Midjourney, 2022. Midjourney. Retrieved from Midjourney: https://www.midjourney.com/.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013a. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119.

Mikolov, Tomas, Yih, W.-t., Zweig, G., 2013b. Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Atlanta, pp. 746–751.

Mitchell, M., 2021. Abstraction and analogy-making in artificial intelligence. Ann. New York Acad. Sci. 1505 (1), 79–101.

Mitchell, M., Palmarini, A.B., Moskvichev, A., 2023. Comparing humans, GPT-4, and GPT-4V on abstract and reasoning tasks. arXiv:2311.09247.

Narang, S., Chowdhery, A., 2022. Pathways Language Model (PaLM): Scaling to 540 billion parameters for breakthrough performance. Retrieved from Google Research Blog: https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html.

OpenAI, 2021. DALL-E: Creating images from text. Retrieved from OpenAI: https://openai.com/research/dall-e.

OpenAI, 2022a. DALL-E 2. Retrieved from OpenAI: https://openai.com/dall-e-2.

OpenAI, 2022b. Introducing ChatGPT. Retrieved from OpenAI: https://openai.com/blog/chatgpt.

OpenAI, 2023a. Code completion. Retrieved from OpenAI Documentation: https://platform.openai.com/docs/guides/code.

OpenAI, 2023b. DALL-E 3. Retrieved from OpenAI: https://openai.com/dall-e-3.

OpenAI, 2023c. DALL-E 3 is now available in ChatGPT Plus and Enterprise. Retrieved from OpenAI Blog: https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise.

OpenAI, 2023d. GPT-4 technical report. pp. 1–100, arXiv.

OpenAI, 2023e. Introducing ChatGPT plus. Retrieved from OpenAI: https://openai.com/blog/chatgpt-plus.

OpenAI, 2023f. Models. Retrieved from OpenAI Docs: https://platform.openai.com/docs/models/gpt-3-5.

Oxford University Press, 2021. Lexico. Retrieved from Lexico: https://www.lexico.com/.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, Doha, pp. 1532–1543.

Petersen, M.R., van der Plas, L., 2023. Can language models learn analogical reasoning? Investigating training objectives and comparisons to human performance. arXiv:2310.05597.

Pichai, S., 2022. Google I/O 2022: Advancing knowledge and computing. Retrieved from Google The Keyword: https://blog.google/technology/developers/io-2022-keynote/.

Pichai, S., 2023. An important next step on our AI journey. Retrieved from Google The Keyword: https://blog.google/technology/ai/bard-google-ai-search-updates/.

Polya, G., 1990. Mathematics and Plausible Reasoning: Induction and Analogy in Mathematics, vol. 1, Princeton University Press, Princeton.

Pourpanah, F., Zbdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., et al., 2022. A review of generalized zero-shot learning methods. IEEE Trans. Pattern Anal. Mach. Intell. 1–20.

Princeton University, 2010. About WordnNet. Princeton University, Retrieved April 14 2022, from https://wordnet.princeton.edu/citing-wordnet.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 8748–8763, Virtual.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training.

Radford, A., Wook, K.J., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision. pp. 1–28, arXiv preprint arXiv:2212.04356.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. 1, 8. OpenAI.

Raimondi, R., Tzoumas, N., Salisbury, T., Simplicio, S.D., Romano, M.R., 2023. Comparative analysis of large language models in the royal college of ophthalmologists fellowship exams. Eye 1–4.

Rajarman, V., 2023. From ELIZA to ChatGPT: History of human–computer conversation. Resonance 889–905.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latent. pp. 1–27, arXiv preprint arXiv:2204.06125.

Raven, J.C., Court, J.H., 1938. Raven's Progressive Matrices. Western Psychological Services, Los Angeles.

Rogers, A., Drozd, A., Li, B., 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, Association for Computational Linguistics, Vancouver, pp. 135–148.

Rombach, R., Blattmann, A., Lorenz, D., Essert, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, New Orleans, pp. 10684–10695.

Rudolph, J., Tan, S., Tan, S., 2023. War of the chatbots: Bard, bing chat, ChatGPT, ernie, and beyond. The new AI gold rush and its impact on higher education. J. Appl. Learn. Teach. 6 (1), 1–26.

Sadeghi, F., Zitnick, C.L., Farhadi, A., 2015. Visalogy: Answering visual analogy questions. In: Advances in Neural Information Processing Systems. Montreal.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the Third International Conference on Learning Representations. San Diego.

Singh, O.P., 2023. Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. Indian J. Psychiatry 65 (3), 297–298.

Situ, J.X., Friend, M.A., Bauer, K.W., Bihl, T.J., 2016. Contextual features and Bayesian belief networks for improved synthetic aperture radar combat identification. Mil. Oper. Res. 21 (1), 89–106.

Socher, R., Ganjoo, M., Manning, C.D., Ng, A., 2013. Zero-shot learning through cross-modal transfer. Adv. Neural Inf. Process. Syst. 26, 935–943.

Srinivasan, R., Uchino, K., 2021. Biases in generative art: A casual look from the lens of art history. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, pp. 41–51, Virtual.

Stability.AI, 2022. Stable difussion launch announcement. Retrieved from Stability AI Blog: https://stability.ai/blog/stable-diffusion-announcement.

Sun, X., Gu, J., Sun, H., 2021. Research process of zero-shot learning. Appl. Intell. 360, 0–3614.

Thorp, H.H., 2023. ChatGPT is fun, but not an author. Science 379 (6630), 313.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al., 2023. LLaMA: Open and efficient foundation language model. arXiv:2302.13971.

Vaswani, A., Shazeer, N., Parmer, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Long Beach, pp. 5999–6009.

Webb, T., Holyoak, K.J., Lu, H., 2023. Emergent analogical reasoning in large language models. Nat. Hum. Behav. 7 (9), 1526–1541.

Whalen, J., Mouza, C., 2023. ChatGPT: Challenges, opportunities, and implications for teacher education. Contemp. Issues Technol. Teach. Educ. 23 (1), 1–23.

Wilson, W.H., Halford, G.S., Gray, B., Phillips, S., 2001. The STAR-2 model for mapping hierarchically structured analogs. In: Gentner, D., Holyoak, K.J., Kokinov, B.N. (Eds.), The Analogical Mind. MIT Press, Cambridge, pp. 125–160.

You.com, 2022. What is YouChat? - Unlock the power of AI with the search assistant that works for you. Retrieved from You.com: https://about.you.com/introducing-youchat-the-ai-search-assistant-that-lives-in-your-search-engine-eff7badcd655/.

You.com, 2023a. Introducing YouChat 2.0 - Unlock the power of AI with the search assistant that works for you. Retrieved from You.com: https://about.you.com/introducing-youchat-2-0-unlock-the-power-of-ai-with-the-search-assistant-that-works-for-you-4b18aa3007bf-2/.

You.com, 2023b. YouChat 3.0 is here - experience the ultimate chat search. Retrieved from About You.com: https://about.you.com/youchat-3-0-is-here-experience-the-ultimate-chat-search-6862c41166cc/.

Yu, J., He, R., Ying, R., 2023. Thought propagation: An analogical approach to complex reasoning with large language models. arXiv:2310.03965.

Zaremba, W., Brockman, G., 2021. OpenAI codex. Retrieved from OpenAI: https://openai.com/blog/openai-codex.

Zhang, C., Gao, F., Baoxiong, J., Zhu, Y., Song-Chun, Z., 2019. RAVEN: A dataset for relational and analogical visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Long Beach, pp. 5317–5327.