

Word Definitions from Large Language Models

Anonymous ACL submission

Abstract

Dictionary definitions are historically the arbitrator of what words mean, but this primacy has come under threat by recent progress in NLP, including word embeddings and generative models like ChatGPT. We present an exploratory study of the degree of alignment between word definitions from classical dictionaries and these newer computational artifacts. Specifically, we compare definitions from three published dictionaries to those generated from variants of ChatGPT. We show that (i) definitions from different traditional dictionaries exhibit more surface form similarity than do model-generated definitions, (ii) that the ChatGPT definitions are highly accurate, comparable to traditional dictionaries, and (iii) ChatGPT-based embedding definitions retain their accuracy even on low frequency words, much better than GloVe and FastText word embeddings.

1 Introduction

Many generations of readers have relied on the notion that the meaning of a word is what it says in the dictionary. Definitions of the primary and alternate senses of words have appeared in printed books well before Samuel Johnson’s famous *A Dictionary of the English Language* in 1755. Early western lexicography evolved from glosses and glossaries. American English was later standardized by Noah Webster in *American Dictionary of the English Language* in 1828.

But the primacy of dictionary definitions as the arbitrator of word meanings has come under threat by recent progress in natural language processing. Vector representations (word embeddings) have proven more valuable computationally than handcrafted semantics or definitions in describing the effective meaning of vocabulary words in a given language. And generative dialog systems like ChatGPT will happily produce full text definitions of every word when asked; indeed multiple versions of these definitions in response to different prompts.

This paper is an exploratory study of the degree of alignment between word definitions from classical dictionaries and these newer computational artifacts. There are several distinct questions we address here:

- How consistent is the representation of semantics from classical dictionary definitions with those of word embeddings and generative language models?
- How quickly are the definitions generated by LLMs improving as the technology advances?
- LLMs have been described as “stochastic parrots” (Bender et al., 2021). To what extent are the definitions they produce mere repetitions or even plagiarisms of classical dictionary definitions?

Our motivation for the paper is to understand the impact of generative models on the understanding and interpretation of word definitions. The findings can potentially benefit downstream tasks in several ways:

- Demonstrating that GPT models can provide accurate definitions even for low-frequency words suggests generative models can lead to better semantic understanding in various downstream tasks.
- Enhancing the credibility of educational tools and resources based on generative models, especially in language learning, where accurate and comprehensive definitions are crucial.

Our methodology is as follows. For each of over 2,500 carefully selected words, we extracted their handcrafted definitions in each of three different dictionaries (WordNet, Merriam-Webster, and Random House/Dictionary.com), and also vector representations from two prominent collections of word

078 embeddings (FastText and GloVe). We also asked
079 two versions of ChatGPT (3.5 and 4.0) for the def-
080 initions of every word, each with two different
081 forms of prompts. The first prompt is "what is the
082 meaning of this word?", while the second prompt
083 is "define this word." We measure the distances
084 between these representations using several tech-
085 niques, including vector distance/similarity, edit
086 distance, and neighborhood correlations.

087 Our primary observations include:

- 088 • *The surface form of generated definitions dif-*
089 *fer substantially from those of published dic-*
090 *tionaries* – Although ChatGPT presumably
091 trained in part on the handcrafted definitions
092 in our study, little-to-no unexpected trace of
093 these texts remain in the generated text. In-
094 deed, we find roughly twice as many long ex-
095 act matches between traditionally published
096 dictionary definitions than we observe be-
097 tween published and generated texts.
- 098 • *ChatGPT generated word definitions are*
099 *highly accurate, consistent with published dic-*
100 *tionaries* – We evaluate SBERT definitions
101 embeddings to identify candidates for mis-
102 matching definitions across sources. Of the 50
103 most distant pairs between Merriam-Webster
104 and GPT4 definitions, human evaluation con-
105 firms that GPT4 recognized 48 compared to
106 34 for GPT3, and that all but one of these
107 definitions matched the primary sense of the
108 published dictionary.
- 109 • *The consistency of LLM-generated definitions*
110 *is relatively independent of frequency, unlike*
111 *word embeddings* – We propose a new average
112 distance correlation metric to compare word
113 embeddings in different dimensional spaces,
114 and use it to study both traditional word em-
115 beddings (GloVe and FastText) and sentence
116 embeddings of both dictionary and generated
117 definitions. Our results suggest that SBERT
118 embeddings of definitions (synthesized or pub-
119 lished) may create more accurate word embed-
120 dings for low frequency words than traditional
121 methods.

122 This paper is organized as follows. Section 2
123 presents related work on dictionaries, definitions,
124 and semantic representation. The dictionaries and
125 generative models we use to formulate our dataset
126 are described in Section 3. The question of just

127 how original generated definitions are is consid-
128 ered in Section 4. We use vector representations to
129 quantify the similarity of definitions by source in
130 Section 5, and between definitions and word em-
131 beddings in Section 6. We conclude with directions
132 for future work in Section 7.

133 2 Related Work

134 Large language models (LLM) recently burst into
135 popular consciousness through the phenomenon of
136 ChatGPT, which reached 100 million users in only
137 three months (Milmo, 2023), offering the prospect
138 of knowledge systems that can write professional-
139 quality documents. Generative Pre-trained Trans-
140 former (GPT) models are a series of large language
141 models (LLMs) developed by OpenAI. Each model
142 in the series (Radford and Narasimhan, 2018; Rad-
143 ford et al., 2019; Brown et al., 2020; OpenAI, 2023)
144 is trained on a larger corpus of text and achieves
145 better results on natural language processing (NLP)
146 tasks compared to its predecessors.

147 ChatGPT promises to have a major impact
148 on many fields, including healthcare (Scerri and
149 Morin) and education (Han et al., 2023; Mogali).
150 The models have also been used to assess transla-
151 tion quality (Kocmi and Federmann, 2023) and the
152 nature of human personalities (Rao et al., 2023).

153 2.1 LLMs and Prompt Engineering

154 One aspect of our work is assessing how different
155 prompts to LLMs (here, for word definitions) yield
156 responses of different quality. Prompt engineering
157 has been studied for Human-Computer Interaction
158 (Oppenlaender, 2022), NLP (Liu et al., 2021; Wang
159 et al., 2023), and tool learning based on foundation
160 models (Qin et al., 2023). Prompt design helps
161 overcome model bias (Rao et al., 2023). Chen et al.
162 identify the best prompt to induce correct answer
163 from pre-trained language models at a low cost
164 (Chen et al., 2023).

165 Chain-of-thought prompting (Wei et al., 2023)
166 decomposes a task into a sequence of steps. It has
167 been used to allow LLMs to use tools such as a
168 calculator and a chemical reaction predictor (Inaba
169 et al., 2023), to improve reasoning over long input
170 documents (Sun et al., 2023), to perform better on
171 web navigation tasks (Sridhar et al., 2023), and
172 to achieve more correct answers through dynamic
173 sampling (Aggarwal et al., 2023).

2.2 Dictionaries and Definitions

Lexicographical works can be dated back to ancient Mesopotamia (Veldhuis, 2019) and ancient China (Creamer, 1992). The ancient Greeks also use glossaries to understand religious and literature works (Lee, 2022; Dolezal and Velde, 2021). Medieval lexicographers use derivation and alphabetization techniques (Weijers, 1989), which lay the foundations for early modern dictionaries (Considine, 2008).

The English language was standardized through dictionaries by Samuel Johnson and Noah Webster (Sokolowski, 2020). The Oxford English Dictionary represents another milestone, with multiple editions created in the 19th and 20th centuries. Most dictionaries in the digital age now appear in electronic form (Zimmer, 2014).

Researchers have shown that typical native speakers of American English know an average of 42,000 words and 4,200 multiword expressions. This number ranges from 27,000 words for the lowest 5% to 52,000 words for the highest 5% (Brybaert et al., 2016).

2.3 Definitions and Word Embeddings

Word embeddings are a multi-dimensional vector representation of a word that conveys semantic and syntactic information. Words with similar meanings have embeddings closer to each other in the vector space. Popular word embeddings like Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) have been used for many NLP tasks.

Definition modeling seeks to generate word definitions given a word and its embedding (Noraset et al., 2016). Definition generation solves the WSD problem by creating definitions instead of choosing a correct one (Bevilacqua et al., 2021). Information captured in contextualized word representations have been analyzed and shown that they can be interpreted in a human-readable way (Chang and Chen, 2019). Huang et al. add a re-ranking mechanism to the decoder-encoder model to control specificity (Huang et al., 2021). Liu et al use capsule networks first proposed in (Sabour et al., 2017) to decompose word embeddings into morpheme-like vectors and reconstruct them into sense embeddings. Word embeddings have also been extracted from Urban dictionary (Wilson et al., 2020a) and Twitter (Wilson et al., 2020b).

	Common	Less common	Rare	Total
NOUN	749	475	182	1406
VERB	179	149	53	381
ADJ	153	77	19	249
ADV	83	32	16	131
Other	55	0	290	345
Total	1219	733	560	2512

Table 1: Word frequency and POS tag summary.

3 Dataset

Rate limitations on using the ChatGPT API prevented us from studying the definitions of all English-language words, so we were selective in what words we study. Starting from a lexicon containing 246,591 English words sorted by their frequency, we sample 3000 words representing 500 words in each of six frequencies tiers:

- *Common*: rank 1 to 500, and 501 to 1000.
- *Less Common*: 5001-5500, and 10001-10500.
- *Rare*: 20001-20500, and 50001-50500.

We cleaned the dataset by removing any word whose definition is missing in any source. This leaves 2512 words with complete definitions remaining in our final dataset. Table 1 presents a breakdown of these words by frequency tier and part of speech (POS) tag, demonstrating that our sampled words are well distributed by both features. Nouns are the most common type of words in the dataset.

For each selected lexicon word X, we collect answers from both GPT3.5 and GPT4 using two different prompts, identified throughout the text as type 1 or 2 prompts:

- Type 1: "What is the meaning of the word X?"
- Type 2: "Define the word X."

This experimental design enables us to make a variety of comparisons between definitions produced (a) between different prompts on the same model, (b) two different models, namely GPT3.5 and GPT4.0, (c) comparisons between models and published dictionaries, (d) between common, moderate, and rare words, and (e) between different parts of speech.

To provide fair grounds for comparison, we extracted published definitions for each of these words from three online dictionaries, namely:

Source	N	Avg Length	Min	Max	σ
GPT3-1	2459	35.97	3	253	20.21
GPT3-2	2488	43.09	3	248	28.65
GPT4-1	2488	132.31	20	386	64.97
GPT4-2	2489	133.77	28	466	65.26
Wordnet	2247	37.25	1	416	45.36
Merriam-W.	2231	132.43	2	1957	179.59
Dict.com	2398	223.11	3	3049	328.89

Table 2: Length summary for collected dictionary datasets, showing the mean, min, max and standard deviation for each source, plus the number of queried lexicon terms defined within each sources (out of 3000).

- **WordNet** (Miller, 1995) (<http://wordnetweb.princeton.edu/perl/webwn>): This manually constructed database of word definitions and semantic relationships among words. It is highly accurate and serves as a useful tool in the field of NLP and linguistics.
- **Merriam-Webster dictionary** (Merriam-Webster Online, 2023) (<https://www.merriam-webster.com>): The Merriam-Webster.com site is a commercial digital dictionary which is regularly updated.
- **Dictionary.com** (<https://www.dictionary.com>): An online dictionary based on the *Random House Unabridged Dictionary*, and is updated regularly by editors.

The length distribution of these sources is summarized in Table 2, including mean and min/max lengths and standard deviation. Observe that GPT4 definitions are generally a little longer than those produced by GPT3.5, but within a given model the two prompts produce similar length texts. The correlation of the lengths of the two GPT3 definitions is 0.61, and that of the two GPT4 definitions is 0.74.

Two of the published dictionaries (Merriam-Webster and Dictionary.com) produce substantially longer definitions than WordNet and GPT3.5.

4 Does ChatGPT Plagiarize?

Although neural network-based generative models synthesize response texts instead of explicitly cutting-and-pasting from training texts, this does not mean they cannot plagiarize, say through overfitting model parameters. Dictionary definitions

provide an interesting domain to assess this possibility: they are precisely written texts that are readily available in training data. There is a long history of dictionaries plagiarizing each other (Cooper, 1962), and commonly include definitions of fake words (such as Mountweazel or jungftak) in order to catch wholesale pirating by other works (Williams, 2016).

Various plagiarism detection techniques (Foltýnek et al., 2020; Zhou et al., 2022) can be used to check if GPT models are plagiarizing. The originality of essays produced by ChatGPT can be evaluated using plagiarism detection tools (Khalil and Er, 2023). The result indicates that ChatGPT output is complex enough to avoid being caught, and that ChatGPT itself performs better on determining whether the text is generated by the model or not.

We investigate this issue by identifying the longest common substring between the ChatGPT-generated definition for a given word and the corresponding definition in a published dictionary. Table 3 reports the words which contain the longest case-insensitive match between each model/dictionary pair. These examples are quite compelling, with common sequences as long as 17 words capturing the heart of the definition.

These examples are provocative, but not necessarily damning. Determining a meaningful expected longest common substring length in our instance is challenging, for several reasons. For randomly generated sequences, statisticians have proven this quantity grows logarithmically in the length of the sequences (Arratia and Waterman, 1985; Barros et al., 2019). But natural language text is far from random, and uses a large vocabulary as opposed to a constant-sized alphabet. Further, text with a high semantic similarity (two definitions of the same word) should share greater surface similarity than mismatched definitions. The length of the definitions matter, in a non-trivial way: longer, more detailed definitions should be expected to contain longer matches than briefer descriptions.

To assess whether these common phrases reflect untoward plagiarism or instead naturally constrained word choices in precise definitions, we compare the degree of borrowing between published dictionaries with what one sees generated by models. Standards for acceptable borrowing in published dictionaries are established by social convention and copyright law. Do LLMs borrow more from published dictionaries than dictionaries

Source	GPT	Word	Longest Common Substring
Wordnet	3-1	clinic	a medical establishment run by a group of medical
	3-2	minutes	a unit of time equal to 60 seconds or 1/60th of an hour
	4-1	forerunner	that precedes and indicates the approach of something or someone
	4-2	nodule	ally harder than the surrounding rock or sediment
Merriam-Webster	3-1	kelvin	the base unit of temperature in the International System of Units
	3-2	town	ally larger than a village but smaller than a city
	4-1	atheism	a philosophical or religious position characterized by disbelief in the existence of a god or any gods
	4-2	econometric	tical methods to the study of economic data and problems
Dictionary.com	3-1	letter	addressed to a person or organization and usually transmitted by mail
	3-2	delta	the fourth letter of the Greek alphabet (Δ , δ)
	4-1	back	he rear part of the human body, extending from the neck to the lower end of the spine
	4-2	compared	o examine (two or more objects, ideas, people, etc.) in order to note similarities and differences

Table 3: The longest common substring between GPT-model generated definitions and published dictionaries. These matching phrases often capture the primary sense of the underlying word.

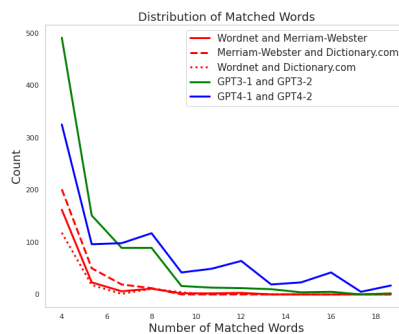
borrow from each other?

Figure 1 presents the frequency distributions of match length between all three pairs of published dictionaries, and the two prompts for each model. They show that each pair of dictionaries share a small tail of long common substrings in their definitions, with substantially greater sharing between the two definitions from GPT-3 and GPT-4. There is nothing untoward about this, just that the generative models repeat language in the two equivalent definitions. Surprisingly, GPT-4 repeats itself substantially more frequently than GPT-3, perhaps an artifact from efforts to reduce model hallucination.

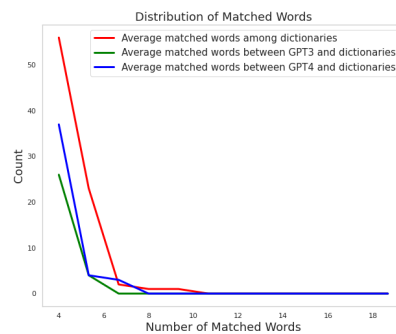
The average length of common strings, and frequencies of long matches reported in Table 4 shows more aggressive borrowing between Merriam-Webster and Dictionary.com than any other pairing of published dictionaries or dictionary-model pairs. In particular, the 155 long matches (of length ≥ 5) observed between Merriam-Webster and Dictionary.com were roughly twice as many as between either of these dictionaries and any model. There seems no real evidence that GPT models unfairly replicate published training data, despite the long matches reported in Table 3.

5 Are GPT Definitions Consistent with Dictionaries?

Model-generated word definitions aim to capture the same essential meaning as published dictionaries, while avoiding textual equivalence. In order to obtain quantitative information about the relative quality of published and generated definitions, we consider three distinct distance or similarity functions between short texts—here definitions of the same word from different sources. Two of these are based on the distance between embeddings, while the third works in the full text space of the defini-



a) Match length distributions for dictionaries and GPT models.



b) Average match length distributions against dictionaries.

Figure 1: Frequency distribution of number of matched words, among dictionary pairs and GPT models (top), and the average number of matched words among dictionaries and between GPT model and dictionaries (bottom). GPT models exhibit substantially more borrowing than published dictionaries.

Matched Pair	Average longest match (words)	Count matches $\text{len} \geq 5$
Wordnet / Merriam-Webster	1.32	98
Wordnet / Dictionary.com	1.13	77
Merriam-Webster / Dictionary.com	1.54	155
GPT3-1 / Wordnet	1.55	99
GPT3-2 / Wordnet	1.59	97
GPT4-1 / Wordnet	1.65	93
GPT4-2 / Wordnet	1.62	91
GPT3-1 / Merriam-Webster	1.23	80
GPT3-2 / Merriam-Webster	1.25	75
GPT4-1 / Merriam-Webster	1.20	77
GPT4-2 / Merriam-Webster	1.18	70
GPT3-1 / Dictionary.com	1.00	69
GPT3-2 / Dictionary.com	1.00	71
GPT4-1 / Dictionary.com	1.04	65
GPT4-2 / Dictionary.com	1.04	75
GPT3-1 / GPT3-2	3.09	592
GPT4-1 / GPT4-2	4.00	711

Table 4: Average length and number of long matches (≥ 5 words) between definition sources.

tions:

- **BERT** (Reimers and Gurevych, 2019) – a modification of BERT that uses Siamese network structures to extract sentence-level representation using triplet loss.
- **MPnet** (Song et al., 2020) – a pre-trained model that combines permuted language modeling and the use of auxiliary position information to derive feature embedding.
- **Edit Distance** – a string metric that represents the minimum number of insertions, deletions, or replacements required to transform one string into another. Here the edit distance is normalized by dividing itself by the maximum length of the two strings.

We calculate the distance between definitions provided by GPT models and online dictionaries. Table 5 shows the cosine distance between each pair of online dictionaries based on sentence embeddings and edit distance. The definitions from ChatGPT appear closer to Wordnet and Merriam-Webster, while definitions from GPT4 are closer to those on Dictionary.com.

Similarly, Table 6 reports the distance between each pair of GPT model definitions. For both GPT-3.5 and GPT-4, the definitions produced by the same model with different prompts are closest by all three distance metrics. Both BERT and MPNet-based distances perform well in capturing the similarity of definition pairs. We use cosine distance of

Dictionary	GPT model	BERT	MPNet	Edit Dist.
Wordnet	GPT3-1	0.33	0.43	4.72
	GPT3-2	0.33	0.44	4.81
	GPT4-1	0.37	0.49	5.01
	GPT4-2	0.35	0.45	5.08
Merriam-Webster	GPT3-1	0.31	0.38	4.49
	GPT3-2	0.30	0.39	4.50
	GPT4-1	0.33	0.43	4.53
	GPT4-2	0.31	0.40	4.59
Dictionary.com	GPT3-1	0.36	0.31	4.92
	GPT3-2	0.35	0.31	4.86
	GPT4-1	0.33	0.32	4.59
	GPT4-2	0.31	0.28	4.63

Table 5: Distance between online dictionaries and definitions provided by GPT models (cosine).

Model 1	Model 2	BERT	MPNet	Edit Dist.
GPT3-1	GPT3-2	0.18	0.18	3.96
	GPT4-1	0.28	0.24	4.79
	GPT4-2	0.25	0.19	4.86
GPT3-2	GPT3-1	0.18	0.18	3.96
	GPT4-1	0.26	0.22	4.61
	GPT4-2	0.22	0.17	4.65
GPT4-1	GPT3-1	0.28	0.24	4.79
	GPT3-2	0.26	0.22	4.61
	GPT4-2	0.17	0.12	3.84
GPT4-2	GPT3-1	0.25	0.19	4.86
	GPT3-2	0.22	0.17	4.65
	GPT4-1	0.17	0.12	3.84

Table 6: Distance between definitions generated from GPT models (cosine). The two distinct definitions produced by the same version of GPT are closer than those generated by different models.

BERT embeddings as our distance function in the rest of this paper, although similar results follow from MPNet.

5.1 POS Analysis

A natural hypothesis is that certain classes of words are easier to generate reliable definitions for than others. Nouns represent objects, which might appear to be easier to precisely than descriptive words like adjectives and adverbs. We obtain a dominant POS tag on each word regardless of its context using the NLKT library (Bird et al., 2009). Table 7 presents the average cosine distance error between generated and published definitions, for each model and POS type. Curiously, GPT3.5 proves most accurate on descriptive words, which GPT4 produces its best definitions for nouns. The results are consistent across all three published dictionaries, and indistinguishable for Type 1 and Type 2 prompts.

Model	POS	WordNet	Merriam-Webster	Dictionary.com
GPT3-1	NOUN	0.32	0.30	0.35
	VERB	0.32	0.34	0.42
	ADJ	0.33	0.28	0.35
	ADV	0.29	0.28	0.41
GPT3-2	NOUN	0.33	0.30	0.34
	VERB	0.32	0.33	0.40
	ADJ	0.36	0.27	0.34
	ADV	0.29	0.27	0.39
GPT4-1	NOUN	0.36	0.32	0.31
	VERB	0.37	0.35	0.35
	ADJ	0.40	0.33	0.33
	ADV	0.45	0.40	0.37
GPT4-2	NOUN	0.35	0.30	0.30
	VERB	0.35	0.33	0.34
	ADJ	0.38	0.31	0.31
	ADV	0.41	0.35	0.35

Table 7: Cosine distance based on SBERT for different POS tags.

5.2 Word Frequency Analysis

There are natural but contradictory hypotheses to govern how the quality of LLM-generated definitions should vary as a function of the relative frequency of each word. One may speculate that the most common words are hardest to define, because of functional forms like prepositions, and that they are more likely to be enriched with multiple senses. But it is equally reasonable to think that low frequency words will be most difficult for generative models to understand, as they are seen least frequently in training data.

To resolve this debate, we compare the average cosine distance between SBERT-encoded dictionary and generated definitions, partitioned by class into words of high frequency, moderate frequency, and rare words. The results in Table 8 show words of middle frequency produce the best dictionary-generated for all models, across all dictionaries. The differences between frequency tiers is generally quite modest, generally less than the impact of model version. GPT-3.5 appears to generate slightly better definitions than the later GPT-4 by this metric, while the choice of prompt has little effect on the accuracy of result from either model.

5.3 Are GPT-Generated Definitions Accurate?

Researchers have found that ChatGPT can easily generate answers with complete assurance, even sometimes the answer is wrong (Lehnert, 2023). To evaluate the correctness of GPT-generated definitions, we manually compare the definitions of the 50 words with the largest Euclidean distance between embedding definitions between Merriam-

Model	Frequency	Word Net	Merriam-Webster	Dictionary.com
GPT3-1	high	0.35	0.34	0.42
	medium	0.28	0.28	0.31
	low	0.32	0.28	0.30
GPT3-2	high	0.35	0.33	0.41
	medium	0.29	0.27	0.30
	low	0.34	0.29	0.30
GPT4-1	high	0.40	0.37	0.37
	medium	0.34	0.29	0.28
	low	0.37	0.29	0.28
GPT4-2	high	0.37	0.34	0.35
	medium	0.33	0.28	0.28
	low	0.39	0.30	0.28

Table 8: Cosine distance based on SBERT for different word frequency.

Webster and GPT-4.

We tabulate the words with the largest such distances in Table 9. Of the fifty words we manually evaluate, the biggest distances occur when models admit they do not know the definitions of generally obscure words (often proper names and abbreviations) that appeared in Merriam-Webster. Only two of the fifty words (cordier and imon) were unknown to GPT-4, compared to 16 words GPT-3 did not understand. Of the 48 words for which GPT-4 ventured a definition, we deemed all of them similar except for “Acton”, which denoted a place in GPT-4 and a person in Merriam-Webster. We conclude that the model generated definitions are generally of high quality, consistent with those from published dictionaries.

6 Word Embeddings and Definitions

Word embeddings are vector representations that capture the semantics of word usage. What dictionary definitions are for people, word embeddings are for NLP models: an easily-understood representation of the meaning of a vocabulary word. In this section we will directly compare traditional word embeddings to text embeddings of explicit dictionary embeddings, to help establish the level of correspondence between these representations.

Specifically, we obtain GloVe and 300-dimensional FastText word embeddings on all the words, and compare the closest neighbors of a word based on definition and word embeddings. We cannot compare these embeddings directly, because the underlying spaces and even dimensionality are incomparable.

Hence, we propose a new technique to measure the consistency of word and definition embeddings over a common vocabulary. For each word w in

Word	Definition from GPT3	Definition from GPT4	Definition in Merriam-Webster
paget	I'm sorry, as an AI language model, I cannot define the word "paget" accurately without any context.	"Paget" is primarily known as a surname of Norman origin and can be seen in various contexts.	Sir James 1814–1899 English surgeon and pathologist.
acton	I'm sorry, but "acton" is not a word in the English language.	"Acton" is primarily known as a place name, particularly in England. It's a district in the West London, in the Borough of Ealing.	1st Baron 1834–1902 John Emerich Edward Dalberg-Acton, English historian.
nov	I'm sorry, I cannot define the word "nov" as it does not exist or have any clear meaning in the English language.	"Nov" is most commonly recognized as an abbreviation for the month of November.	November.
cordier	As an AI language model, I cannot determine the context of the word "cordier".	I'm sorry, but as of my knowledge cutoff in September 2021, "cordier" doesn't have a specific definition in English. It seems like it could be a surname.	comparative of cordy.
inion	There is no word "inion" in the English language. It may be a misspelling of "onion," which is a type of vegetable.	The inion is a term used in anatomy that refers to the most prominent projection of the occipital bone at the lower rear part of the skull in humans and other hominids.	the external occipital protuberance of the skull.

Table 9: Representative examples of the words whose GPT definition embeddings are far from online dictionary definition embedding.

the lexicon, we calculate the distance from w to all other words, separately in word embedding and definition space. The correlation coefficient r between these distances defines the agreement between the spaces from the perspective of word w . To get a full vocabulary metric on space similarity, we average these correlations for each word in the lexicon.

6.1 Frequency Analysis

Table 10 compares published and generated definitions with popular word embeddings, GloVe and FastText, using the correlation measure defined above. We distinguish between our three tranches of words, high-frequency, moderate, and rare.

The results are consistent across all sources and embeddings: high-frequency words show better consistency between definition-word embeddings than moderate-frequency words, which are substantially better than low frequency words. As our results in Section 5.2 show the definitions of roughly equal quality across frequency tranches, the difference must be due to the word embeddings themselves becoming less accurate with diminishing usage frequency. This presumably results from embeddings that are trained on insufficient amounts of data.

7 Conclusion and Future Work

The relationship between full text descriptions and embeddings is an interesting one. Our work demonstrates that model-generated definitions are accurate and lead to other useful representations.

Our experiments to detect long common phrases in alternate definitions of a particular word motivates questions of how long such phrases must be in generated text. How just how far apart can two texts be in surface form while still meaning the

Model	Frequency	GloVe	FastText
GPT3-1	high	0.27	0.32
	medium	0.21	0.28
	low	0.08	0.22
GPT3-2	high	0.28	0.35
	medium	0.22	0.30
	low	0.13	0.24
GPT4-1	high	0.28	0.32
	medium	0.23	0.30
	low	0.11	0.25
GPT4-2	high	0.30	0.38
	medium	0.26	0.35
	low	0.12	0.29
Wordnet	high	0.22	0.26
	medium	0.19	0.22
	low	0.11	0.16
Merriam-Webster	high	0.25	0.30
	medium	0.18	0.23
	low	0.06	0.18
Dictionary.com	high	0.33	0.31
	medium	0.21	0.25
	low	0.08	0.20

Table 10: Pearson correlation between word embedding and definition embedding based on cosine distance for different word frequency.

same thing? A study of situations where LLMs are highly constrained to specific word/phrase choices may send interesting light on the situation.

539 Limitations

540 The generative models powering ChatGPT are un-
541 dergoing rapid evolution, so the observations in
542 this paper cannot be guaranteed to hold in future re-
543 leases. However, we do anticipate that the models
544 will only continue to improve, and since our results
545 show that they already exhibit similar accuracy and
546 originality as published dictionaries we would be
547 surprised of our observations do not continue to
548 hold with future releases.

549 Having established basic agreement between
550 publishing and generated definitions, as we have
551 done in this paper, is a less demanding task than
552 validating the completeness of multiple senses or
553 alternate usages. We also have restricted our at-
554 tention to English words and definitions. It would
555 be interesting to evaluate model fidelity in other
556 languages, or translation (e.g. "Define the French
557 word rouge.").

558 References

559 Pranjali Aggarwal, Aman Madaan, Yiming Yang, and
560 Mausam. 2023. [Let's sample step by step: Adaptive-](#)
561 [consistency for efficient reasoning with llms.](#)

562 Richard Arratia and Michael S Waterman. 1985. An
563 erdős-rényi law with shifts. *Advances in mathemat-*
564 *ics*, 55(1):13–23.

565 Vanessa Barros, Lingmin Liao, and Jérôme Rousseau.
566 2019. On the shortest distance between orbits and
567 the longest common substring problem. *Advances in*
568 *Mathematics*, 344:311–339.

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

575 Michele Bevilacqua, Tommaso Pasini, Alessandro Ra-
576 ganato, and Roberto Navigli. 2021. [Recent trends](#)
577 [in word sense disambiguation: A survey.](#) In *Pro-*
578 *ceedings of the Thirtieth International Joint Confer-*
579 *ence on Artificial Intelligence, IJCAI-21*, pages 4330–
580 4338. International Joint Conferences on Artificial
581 Intelligence Organization. Survey Track.

582 Steven Bird, Ewan Klein, and Edward Loper. 2009. *Nat-*
583 *ural language processing with Python: analyzing text*
584 *with the natural language toolkit.* " O'Reilly Media,
585 Inc."

586 Piotr Bojanowski, Edouard Grave, Armand Joulin, and
587 Tomas Mikolov. 2017. [Enriching word vectors with](#)
588 [subword information.](#)

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie
Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind
Neelakantan, Pranav Shyam, Girish Sastry, Amanda
Askell, Sandhini Agarwal, Ariel Herbert-Voss,
Gretchen Krueger, Tom Henighan, Rewon Child,
Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,
Clemens Winter, Christopher Hesse, Mark Chen, Eric
Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,
Jack Clark, Christopher Berner, Sam McCandlish,
Alec Radford, Ilya Sutskever, and Dario Amodei.
2020. [Language models are few-shot learners.](#)

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and
Emmanuel Keuleers. 2016. [How many words do we](#)
[know? practical estimates of vocabulary size depen-](#)
[dent on word definition, the degree of language input](#)
[and the participant's age.](#) *Frontiers in Psychology*, 7.

Ting-Yun Chang and Yun-Nung Chen. 2019. [What does](#)
[this word mean? explaining contextualized embed-](#)
[dings with natural language definition.](#) In *Proceed-*
ings of the 2019 Conference on Empirical Methods
in Natural Language Processing and the 9th Inter-
national Joint Conference on Natural Language Pro-
cessing (EMNLP-IJCNLP), pages 6064–6070, Hong
Kong, China. Association for Computational Linguis-
tics.

Yulin Chen, Ning Ding, Xiaobin Wang, Shengding Hu,
Hai-Tao Zheng, Zhiyuan Liu, and Pengjun Xie. 2023. [Exploring lottery prompts for pre-trained language](#)
[models.](#)

J. Considine. 2008. *Dictionaries in Early Modern Eu-*
rope: Lexicography and the Making of Heritage.
Cambridge University Press.

Louis Cooper. 1962. Plagiarism in spanish dictionaries
of the xvith and xviiith centuries. *Hispania*, pages
717–720.

Thomas B. I. Creamer. 1992. *LEXICOGRAPHY*
AND THE HISTORY OF THE CHINESE LAN-
GUAGE, pages 105–136. Max Niemeyer Verlag,
Berlin, Boston.

Fredric Dolezal and H. Velde. 2021. *Broadening Per-*
spectives in the History of Dictionaries and Word
Studies. Table of Contents and Introductory Essay.,
pages vii–xxiii.

Tomáš Foltýnek, Dita Dlabolová, Alla Anohina-
Naumeca, Salim Razi, Július Kravjar, Laima Kam-
zola, Jean Guerrero-Dib, Özgür Çelik, and Debora
Weber-Wulff. 2020. [Testing of support tools for pla-](#)
[giarism detection.](#)

Zhiyong Han, Fortunato Battaglia, Abinav Udaiyar,
Allen Fooks, and Stanley R. Terlecky. 2023. [An ex-](#)
[plorative assessment of chatgpt as an aid in medical](#)
[education: Use it with caution.](#) *medRxiv.*

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. [Definition modelling for appropriate specificity.](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

645	2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	696
646		697
647	Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng, and Sadao Kurohashi. 2023. Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting.	698
648		699
649		700
650		701
651	Mohammad Khalil and Erkan Er. 2023. Will chatgpt get you caught? rethinking of plagiarism detection.	702
652		703
653		704
654	Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.	705
655		706
656		707
657	John A. L. Lee. 2022. <i>Collected Essays on the Greek Bible and Greek Lexicography</i> , volume 112. Peeters Publishers.	708
658		709
659		710
660	Kay Lehnert. 2023. Ai insights into theoretical physics and the swampland program: A journey through the cosmos with chatgpt.	711
661		712
662		713
663	Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.	714
664		715
665		716
666	Merriam-Webster Online. 2023. Merriam-Webster Online Dictionary.	717
667		718
668		719
669	Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.	720
670		721
671		722
672	George A. Miller. 1995. Wordnet: A lexical database for english. <i>Commun. ACM</i> , 38(11):39–41.	723
673		724
674		725
675	Dan Milmo. 2023. Chatgpt reaches 100 million users two months after launch. <i>The Guardian</i> .	726
676		727
677		728
678	Sreenivasulu Reddy Mogali. Initial impressions of chatgpt for anatomy education. <i>Anatomical sciences education</i> .	729
679		730
680		731
681	Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2016. Definition modeling: Learning to define word embeddings in natural language.	732
682		733
683		734
684		735
685	OpenAI. 2023. Gpt-4 technical report.	736
686		737
687	Jonas Oppenlaender. 2022. A taxonomy of prompt modifiers for text-to-image generation.	738
688		739
689		740
690	Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In <i>Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.	741
691		742
692		743
693		744
694		745
695	Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi,	746
	Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models.	747
		748
		749
	Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.	750
		751
		752
	Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.	753
		754
		755
	Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework.	756
		757
		758
	Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing</i> . Association for Computational Linguistics.	759
		760
		761
	Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In <i>Advances in Neural Information Processing Systems</i> , volume 30. Curran Associates, Inc.	762
		763
		764
	Anthony Scerri and Karen H Morin. Using chatbots like chatgpt to support nursing practice. <i>Journal of clinical nursing</i> .	765
		766
		767
	Peter Sokolowski. 2020. 168C10The Invention of the Modern Dictionary: Webster’s Unabridged of 1864. In <i>The Whole World in a Book: Dictionaries in the Nineteenth Century</i> . Oxford University Press.	768
		769
		770
	Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.	771
		772
		773
	Abishek Sridhar, Robert Lo, Frank F. Xu, Hao Zhu, and Shuyan Zhou. 2023. Hierarchical prompting assists large language model on web navigation.	774
		775
		776
	Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, and Mohit Iyyer. 2023. Pearl: Prompting large language models to plan and execute actions over long documents.	777
		778
		779
	Niek Veldhuis. 2019. <i>Ancient Mesopotamia</i> , page 11–35. Cambridge University Press.	780
		781
		782
	Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, Qingqing Zhu, Zhenzhu Yang, Adam Nik, Qi Liu, Chenghua Lin, Shi Wang, Ruibo Liu, Wenhu Chen, Ke Xu, Dayiheng Liu, Yike Guo, and Jie Fu. 2023. Interactive natural language processing.	783
		784
		785
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.	786
		787
		788

- 750 Olga Weijers. 1989. [Lexicography in the middle ages](#).
751 *Viator*, 20:139–154.
- 752 Eleanor Williams. 2016. *Unclear Definitions: Investi-*
753 *gating Dictionaries' Fictitious Entries Through Cre-*
754 *ative and Critical Writing*. Ph.D. thesis, Royal Hol-
755 loway, University of London.
- 756 Steven Wilson, Walid Magdy, Barbara McGillivray, Ki-
757 ran Garimella, and Gareth Tyson. 2020a. [Urban dic-](#)
758 [tionary embeddings for slang NLP applications](#). In
759 *Proceedings of the Twelfth Language Resources and*
760 *Evaluation Conference*, pages 4764–4773, Marseille,
761 France. European Language Resources Association.
- 762 Steven R. Wilson, Walid Magdy, Barbara McGillivray,
763 and Gareth Tyson. 2020b. [Analyzing temporal rela-](#)
764 [tionships between trending terms on twitter and urban](#)
765 [dictionary activity](#). In *12th ACM Conference on Web*
766 *Science*, WebSci '20, page 155–163, New York, NY,
767 USA. Association for Computing Machinery.
- 768 Chao Zhou, Cheng Qiu, and Daniel E. Acuna. 2022.
769 [Paraphrase identification with deep learning: A re-](#)
770 [view of datasets and methods](#).
- 771 Ben Zimmer. 2014. [Lexicography 2.0: Reimagining](#)
772 [dictionaries for the digital age](#). *Dictionaries: Journal*
773 *of the Dictionary Society of North America*, 35:275–
774 286.

A Appendix

Model	Definition of "wasp"
GPT3-1	a winged insect with a slender, often brightly colored body, the females of which have a poisonous sting
GPT3-2	A wasp is a predatory flying insect with a narrow waist and a sting, belonging to one of several families of the order Hymenoptera, such as the Vespidae, which includes the social wasps and hornets, or the Ichneumonidae, which includes parasitic wasps.
GPT4-1	"Wasp" has a couple of different meanings, depending on the context: 1. In entomology (the study of insects), a wasp is a type of flying insect, known for its black and yellow body, narrow waist, and ability to sting. There are many thousands of species of wasps, and they are found all over the world. 2. In social contexts, especially in the United States, WASP is an acronym for White Anglo-Saxon Protestant. It refers to a socio-economic and cultural group that has traditionally been dominant in American society, particularly among the upper class. This usage can sometimes carry implications about privilege and exclusivity. Remember that the context usually makes clear which meaning is intended.
GPT4-2	A wasp is a type of insect that is part of the order Hymenoptera and the suborder Apocrita, which is not a true ant or a bee. There are many different species of wasps, which can vary significantly in terms of size, color, and behavior. Most wasps are known for their ability to sting in self-defense. Unlike bees, which typically can sting only once before dying, many wasps can sting multiple times. Wasp stings are painful and can be dangerous to people who are allergic to them. Wasps can be categorized as either social or solitary. Social wasps, like the common yellowjacket, live in large colonies, and their nests can be found in trees, roof eaves, or underground. Solitary wasps, on the other hand, live alone and do not form colonies. Wasps play an important role in the ecosystem, as they prey on other pests, helping to control their populations. Some species are also important pollinators.
WordNet	white Anglo-Saxon Protestant (a white person of Anglo-Saxon ancestry who belongs to a Protestant denomination); (social or solitary hymenopterans typically having a slender body with the abdomen attached by a narrow stalk and having a formidable sting)
Merriam-Webster	any of numerous social or solitary winged hymenopterous insects...; any of various hymenopterous insects...; an American of Northern European and especially British ancestry and of Protestant background, especially : a member of the dominant and the most privileged class of people in the U.S.
Dictionary.com	any of numerous social or solitary hymenopterous insects...; a person who is snappish or petulant; a member of the Women's Air Force Service Pilots, an organization disbanded in 1944; a white Anglo-Saxon Protestant; a member of the privileged, established white upper middle class in the U.S.

Table 11: Definitions of the word "wasp" from GPT models and online dictionaries.