GPT Perdetry Test: Generating new meanings for new words

Nikolay Malkin
Yale University
New Haven, CT

Sameera Lanka
Microsoft
Redmond, WA

Pranav Goel
University of Maryland
College Park, MD

Sudha Rao
Microsoft Research
Redmond, WA

Nebojsa Jojic
Microsoft Research
Redmond, WA

1kolya.malkin@yale.edu 2{sameera.lanka,sudha.rao,jojic}@microsoft.com 3pgoel1@umd.edu

Abstract

Human innovation in language, such as inventing new words, is a challenge for pretrained language models. We assess the ability of one large model, GPT-3, to process new words and decide on their meaning. We create a set of nonce words and prompt GPT-3 to generate their dictionary definitions. We find GPT-3 produces plausible definitions that align with human judgments. Moreover, GPT-3’s definitions are sometimes preferred to those invented by humans, signaling its intriguing ability not just to adapt, but to add to the evolving vocabulary of the English language.

1 Introduction

Humans are constantly expanding languages with new words. How are artificial language models, which are increasingly deployed ‘in the wild’, to handle the stream of neologisms that are appearing in slang or on social media (Grieve et al., 2018)?

Today’s most advanced language models, including GPT-3 (Brown et al., 2020), use a subword tokenization of input text, rather than consuming it word by word. This allows them to process words never seen in their training data. For example, the word ‘perdetry’, which has never been used in English, is treated by GPT-3 as a sequence of two tokens (Fig. 1). The subword tokenization algorithm is designed for text compression and does not respect the natural morpheme boundaries.

We explore GPT-3’s understanding of English at the subword level by prompting it to give definitions of nonce words1 (Fig. 1). We find in human studies that not only does GPT-3 generate realistic, original meanings for new words, but its definitions are sometimes preferred to those invented by humans. This finding sheds light on GPT-3’s ability to adapt to and even extend a changing vocabulary.

While we cannot ascertain GPT-3’s exact reasons for assigning meanings to nonce words, our results prove that these reasons are not limited to morphology: many neologisms have no clear roots or derivational origin. The meanings of words may be imported by their phonological qualities – more precisely, their orthographic realizations – or by clues to their membership in certain lexical strata. Thus, at a high level, our findings suggest that GPT-3 has learned not only its world knowledge and capacity for long-range reasoning in text (Brown et al., 2020), but also the nuances of etymology and the correspondences of sound and meaning that lie at the very base of language understanding.

1We use the term ‘nonce word’ for a new word not used in English. It becomes a ‘neologism’ once it acquires a meaning.
Below are some pairs of words together with their definitions. The goal is to guess, for each pair, which word goes with which definition. We will show you two options, and you will decide which of them is a better match. The words you’ll get are rare, and we do not expect you to know many, or indeed any, of them. Make your best guess. For some pairs, there is no correct answer. We’ll show you the expected answers at the end. Do not look up the words while doing the task: we are really interested in your gut feeling, right or wrong.

| A. recommor : a female dwarf | caraber : a male witch; a wizard; a warlock |
| B. recommor : a male witch; a wizard; a warlock | caraber : a female dwarf |

Figure 2: The word-definition matching task instructions and a typical question. (GPT-3 happens to have generated the definitions in Option A. In the tests, the assignments of definitions to words in each pair were randomized.)

2 Related work

The notion that some subword elements (phonethemes) carry meaning, but, unlike morphemes, do not play a part in word formation has caused controversy in linguistics for over a century (Nuckolls, 1999; Feist, 2013). In his seminal work, de Saussure (1916) rejected this notion. Yet, later work identified a large set of English phonethemes, such as the cluster /gl/ in ‘glow’, ‘glitter’, ‘gloss’, etc. meaning “light”; a notable list was compiled by Marchand (1959a,b). Recent studies found phonosemantic patterns that are common to many languages (Blasi et al., 2016). In practice, words are even engineered for subconscious reactions: certain sounds in brand names are correlated with associations such as size (of a gadget) or speed (of a courier) (Klink, 2000). Our study suggests that GPT-3 may understand such patterns as well.

There is a body of work on joint modeling of (orthographic or phonological) word forms and grammatical classes such as noun gender and inflection pattern. In a recent study, Williams et al. (2020) used neural models to measure mutual information between meanings and inflection classes of Czech and German nouns, which, for borrowed words, often depend on the language of origin. It is plausible that GPT-3 implicitly uses likely source languages of nonce words to generate meanings associated with some lexical strata, e.g., abstract nouns from Norman French, concrete nouns from the Germanic substrate, and artificially constructed terms with Greek or Latinate elements. (We direct the interested reader to the lexicon in Appendix C.)

Work on neologisms in NLP includes tracking their emergence and spread on the Internet (Grieve et al., 2018; Würschinger et al., 2016), mapping them into embedding spaces (Bojanowski et al., 2017; Zalmout et al., 2019; Ryskina et al., 2020), and codifying and predicting etymologies (Melo, 2014; Wu and Yarowsky, 2020). Others have studied definition generation (Noraset et al., 2017) and the reverse task of mapping definitions to words (Hill et al., 2015), albeit with pretrained embeddings. Limited examples of a pretrained model’s use of nonce words appear in Brown et al. (2020). In this work, we study GPT-3’s ability to define words never seen in context.

3 Creation of new words and meanings

We trained a LSTM model (Hochreiter and Schmidhuber, 1997) on a corpus of English words with a standard character-level objective, then sampled strings from the LSTM to create nonce words. The words were lemmatized and assigned parts of speech (POS): noun (n.), verb (v.), or adjective (adj.).

To produce definitions for these words, we generated text from GPT-3, primed with input in the format “word (POS.) –”. Usually, GPT-3’s outputs had the style of a dictionary definition (Fig. 1). The definitions were filtered by common-sense criteria and lightly edited for consistency, as explained in Appendix A. By this procedure, we obtained 146 word-definition pairs (67 n., 47 v., 32 adj.).

For comparison in our study, we also sampled a set of real but rare English words from a corpus. Definitions for these words were taken from a dictionary. This resulted in a combined set of 220 words (102 n., 70 v., 48 adj.), with a 2:1 ratio of fake to rare words in each POS. See Appen-
We collected 65 sets of annotations for each POS, performed a study in which human subjects were presented with pairs of words of the same part of speech together with their definitions (generated by GPT-3, for fake words, or extracted from the dictionary, for rare words), but not told which definition matches with which word. Some questions contained two fake words, some two rare words, and some one fake and one rare word. Users were asked to decide which assignment of definitions to words is a better fit and to rate their confidence (Fig. 2); the choices were converted to a scale of 0 to 5 (confident in the incorrect match) to 5 (confident in the correct match). Each user received a random pairing of words, but saw each word exactly once. We collected 65 sets of annotations for each POS, for a total of $65 \times 220 = 7150$ data points.

### Results
Humans prefer the pairing from our lexicon in 68% of cases. The scores by the POS and the kind of pair (fake-fake, fake-rare, or rare-rare) are shown in the top rows of Table 1. GPT-3’s definitions align with human judgments far better than random choice ($p$-values below floating-point epsilon). Notably, humans’ performance on pairs containing a fake and a rare word was about the same as on pairs of fake words.

Correlation in performance between different parts of the word-definition matching task is high. Considering only the fake-fake pairs, the score (number of correctly matched pairs) on the noun portion of the task is correlated with the score on verb and adjective pairs with Spearman $\rho \approx 0.42$; a permutation test on rank correlation gives $p \approx 0.01$. The verb and adjective portions are similarly predictive of the other two ($p \approx 0.05$ for both). The correlation is even stronger ($p < 0.0001$ for nouns) when all pairs, not just fake-fake, are considered. This indicates that some users can be identified as ‘better’ at the task, perhaps due to their personal vocabulary, education, or effort. (For example, the average score on the fake-fake noun pairs is 70.7%. However, the average score on fake-fake noun pairs among users who scored above median on the fake-fake adjective pairs is 74.2%). This is strong evidence that the values in Table 1 would be higher with a better selection of users.

There was significant agreement between annotators. In cases when the same pair of words was shown to two users, the mean difference between the two users’ choices on the 0-5 scale was 1.5, and in 61% of cases the two users preferred the same assignment. Remarkably, the latter number is the same for rare-rare, rare-fake, and fake-fake pairs.

It is possible that the subjects knew some of the rare words – and the tables in Appendix C do suggest this. However, assuming that a subject will choose the correct match if they know the meaning of at least one word in a pair, and will do no worse than random guessing on pairs where they know neither word, the last ‘human’ row is consistent with less than a quarter of the rare words, on average, being known to the subjects.

### Likelihood analysis
For each word $w$ and definition $d$ in the lexicon (where $d$ may be the definition of a word different from $w$), we compute the likelihood under GPT-3 of the definition $d$ to follow word $w$, $p(d|w)$. For each pair of words $(w_1, w_2)$ of the same POS, with definitions $(d_1, d_2)$, we compute the difference in log-likelihood between the proper match $(w_1 - d_1, w_2 - d_2)$ and the inverted assignment $(w_1 - d_2, w_2 - d_1)$:

$$LLD(w_1, w_2) = \log \frac{p(d_2|w_1)p(d_1|w_2)}{p(d_1|w_1)p(d_2|w_2)}.$$

If GPT-3 were to perform the matching task done by our human subjects, it would choose the option with higher total likelihood. In other words, it would prefer the correct pairing if $LLD(w_1, w_2)$ is negative and the inverted pairing if it is positive.

Assuming that GPT-3 has seen the rare words in training, we expect it to score very well on rare-rare and rare-fake pairs. We also expect it to prefer

|        | n.    | v.    | adj.  |
|--------|-------|-------|-------|
| fake-fake | 70.7% | 59.8% | 64.3% |
| human   | 72.6  | 60.9  | 65.3  |
| rare-rare| 79.6  | 65.3  | 69.8  |
| fake-fake| 92.4  | 83.2  | 87.3  |
| GPT-3   | 98.5  | 95.5  | 97.3  |
| rare-rare| 99.4  | 98.4  | 100.0 |

Table 1: Accuracies on the task of matching real and machine-generated words with definitions (Fig. 2), performed by study participants (‘human’) or the language model that created the fake definitions (‘GPT-3’).
LLD and human confidence. LLD is a good predictor of human judgments: confidence in the correct pairing for fake-fake pairs $(w_1, w_2)$ is strongly correlated with $\text{LLD}(w_1, w_2)$, a rank correlation test giving $p < 0.001$ for all POS.

One may object that this correlation – and indeed much of humans’ performance – is due to the presence of simple disambiguating markers: for example, a word with suffix ‘-ist’ is likely to denote a person, while an ‘-ism’ is probably an abstract noun. However, examination of log-likelihood differences shows that this is not the case. We stratify the pairs of fake words by LLD and consider the distribution of humans’ confidences for pairs with LLD falling in five ranges: $[-40, -30), [-30, -20), \ldots, [0, 10)$. Confidence in the correct matching is inversely correlated with LLD, but humans tend to choose the correct assignment for pairs in all five strata (Table 2). For pairs with LLD in the ranges $[-10, 0)$ and $[0, 10)$, which form a majority, there tend to be no revealing morphological markers. (Table 2 shows pairs of words with LLD falling into these ranges; Table 7 in the appendix shows more examples.)

Conclusion. Finally, we observe that many of GPT-3’s definitions are original: we are not aware of English words that describe the same concepts (see Table 2 and Appendix C). Some of the innovated meanings fill plausible lexical gaps (‘drobbler’), while others require a degree of creativity (‘subacitide’). This shows that GPT-3 is not simply aligning new words with existing words as in Zalmout et al. (2019), but inventing new meanings.

4.2 Human-generated neologisms

We test GPT-3’s ability to define new words on a set of human-proposed neologisms from the Dictionary of Obscure Sorrows. Many of these words were created out of real English morphemes. We sampled 20 words from this set, got GPT-3 defini-
A. **occhiolism**: a belief that personal power increases proportionally with one’s height

B. **occhiolism**: the awareness of the smallness of one’s perspective

Figure 3: A typical question in the definition choice task. The instructions were similar to those in Fig. 2; the answer choices were identical. (In this case, Option A was generated by GPT-3, Option B by a human.)

Figure 4: Human subjects’ preference for GPT-3-generated definitions (bluer) or human-generated definitions (whiter). Each column represents a single user. The rows and columns have been sorted by their means. The full set of definitions can be found in Appendix C.

We then ran a study with 25 users, in which each user was given words and both definitions (in random order, without being told how each definition was generated) and asked to pick the better match. The responses were converted to a scale of 0 (human-generated is much better) to 5 (GPT-3-generated is much better). Each user marked their definition preference for all 20 words (Fig. 3).

**Results.** Remarkably, users preferred GPT-3’s definitions in 40% of cases, despite the fact that a human thought up each of these word-meaning pairs. This is not simply the result of random guessing by the workers: the result matrix (Fig. 4) shows a significant amount of structure. There are words on which most users agree that the better definition is the one generated either by the human inventor (top rows) or by GPT-3 (bottom rows).

Users most prefer GPT-3’s definition for **backmasking**: “the act of disguising messages within recordings via sound effects” to the human definition “the instinctive tendency to see someone as you knew them in their youth”, while the human definition of **lapyear**: “the age at which you become older than your parents were when you were born” is preferred to GPT-3’s “a lazy person; someone of a low-energy lifestyle”.

**User clusters.** These human-coined neologisms have a bias towards meanings with an existential slant, which results in additional structure in our results, reflecting the population structure of the subjects. Indeed, some workers prefer human-made definitions and others prefer GPT-3’s definitions, which reflect a mixture of meanings seen in a crawl of the Internet.

To analyze the significance such preferences, we perform a randomization test. We define the polarization of a user as the absolute difference between the number of words for which they prefer the human-generated definition and the number for which they prefer GPT-3’s definition. The average polarization over users is greater than that seen in 99% of random preference matrices, indicating that there may indeed be two types of users, with different preferences for the types of meanings they see in words.6

5 Conclusion

A character-level model of English words composed with GPT-3 is a complete scheme for generating new words and innovative meanings. GPT-3 invents definitions for words it has not seen in training that are seen as reasonable by humans. These results have implications for language models’ ability to adapt and even add to an evolving vocabulary. They can inspire future work on machine understanding of new slang, optimization of words and acronyms, creation of fictitious entries, and automatically generating word games.

6A similar test could be performed taking the confidence into account. Here we define polarization as the absolute difference between a user’s mean confidence and 2.5. In each random sample, we flip a random subset of the entries in the confidence matrix to the opposite preference, while keeping the level of uncertainty the same: 0 ↔ 5, 1 ↔ 4, 2 ↔ 3. This results in a p-value around 0.04.
Ethics statement

The authors see no immediate negative societal consequences arising from this work.

As explained in Appendix B, we followed data privacy and anonymization procedures to the greatest extent possible and fairly compensated human subjects.

References

Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences, 113(39):10818–10823.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

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. arXiv:2005:14165.

Ferdinand de Saussure. 1916. Cours de linguistique générale. Payot, Paris.

Jim Feist. 2013. “Sound symbolism” in English. Journal of Pragmatics, 45(1):104–118.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on American social media. Journal of English Linguistics, 46(4):293–319.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2015. Learning to understand phrases by embedding the dictionary. Transactions of the Association for Computational Linguistics, 4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735—1780.

Richard Klink. 2000. Creating brand names with meaning: The use of sound symbolism. Marketing Letters, 11:5–20.

Hans Marchand. 1959a. Phonetic symbolism in English word-formation. Indogermanische Forschungen, 64:146–168.

Hans Marchand. 1959b. Phonetic symbolism in English word-formation. Indogermanische Forschungen, 64:256–277.

Gerard De Melo. 2014. Etymological wordnet: Tracing the history of words. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pages 1148–1154.

Thanapon Noraset, Chen Liang, Lawrence A. Birnbaum, and Douglas C. Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In 31st AAAI Conference on Artificial Intelligence (AAAI), pages 3259–3266.

Janis Nuckolls. 1999. The case for sound symbolism. Annual Review of Anthropology, 28:225–252.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In Proceedings of the Society for Computation in Linguistics 2020, pages 367–376, New York, New York. Association for Computational Linguistics.

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6682–6695, Online. Association for Computational Linguistics.

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3252–3259, Marseille, France. European Language Resources Association.

Quirin Würschinger, Mohammad Fazleb Elahi, Desislava Zhekova, and Hans-Jörg Schmid. 2016. Using the web and social media as corpora for monitoring the spread of neologisms. the case of ‘rapefugee’, ‘rapeugee’, and ‘rapugee’. In Proceedings of the 10th Web as Corpus Workshop, pages 35–43, Berlin. Association for Computational Linguistics.

Nasser Zalmout, Kapil Thadani, and Aasish Pappu. 2019. Unsupervised neologism normalization using embedding space mapping. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 425–430, Hong Kong, China. Association for Computational Linguistics.
A Lexicon creation details

The one-layer LSTM that generated nonce words had a 27-symbol alphabet (‘a’-‘z’ and the start/stop token), embedding dimension 8, and hidden size 128. It was trained on ASCII lowercase words with a standard autoregressive modeling objective of maximizing next-token log-likelihood. Samples from this model were generated greedily with temperature 0.5. Samples that contained extremely rare (deemed unpronounceable) consonant clusters, those bearing a resemblance to obscene words, and those attested in English (as checked by a Web search) were removed. The part-of-speech assignments were chosen based on the word form by a fluent English speaker and cross-checked by three other fluent English speakers.

For all experiments with GPT-3 (generation and likelihood scoring), we used the davinci variant of the model, accessed through the free beta version of the API. In the generation of definitions from GPT-3, samples were taken with temperature 1 and truncated at the first line break or period. Because GPT-3 encountered dictionaries in its training data, acceptable samples were obtained for nearly all words. However, the following kinds of outputs were rejected:

- Outputs that were not in the form of dictionary definitions (were not lists of items that can be parsed as syntactic constituents of the appropriate type), e.g., “sacrel (n.) – the dwarf sighed”.
- Definitions with words that differed from the word being defined by an affix, e.g., “fre-quayer (n.) – one who frequays”.
- Definitions containing sexual or violent content.

If three samples for a word failed the filters above, the word was discarded.

To clean the definitions, we removed technical (parenthesized) comments and alternative senses and made minor edits for consistent syntax and punctuation, such as making all definitions of verbs begin with ‘to’. A similar processing was applied to the definitions from the Dictionary of Obscure Sorrows: the compendium contains extended explanations for each word, composed of the definition and examples of experiential manifestations of the given sorrow. We present only the definition to users.

For example, here are five samples for the fake word ‘perbroil’, which appears to trigger questionable content:

1. perbroil (v.) – Desperate to "lance all boils" on the ass of a superior (Rejected for obscene language and for not being a verb phrase. Otherwise, postprocessing would lowercase the ‘D.’)
2. perbroil (v.) – to have sex three times in one night (Rejected for sexual content.)
3. perbroil (v.) – To cook and then serve an already dead animal as though it were still alive. Results should include blood oozing from the creature’s cherry butthole. (Rejected for violent and sexual content. Otherwise, postprocessing would lowercase the ‘T.’)
4. perbroil (v.) – broil on a plancha (type of griddle) (Rejected for presence of ‘broil’. Otherwise, postprocessing would insert ‘to’.)
5. perbroil (v.) – fry something gently on one side only [early 1990s.] (This would be accepted, and postprocessing would insert ‘to’ and remove the bracketed comment.)

This word would have been rejected, since the first three samples were unsuitable, but we would have found a suitable definition after five tries.

The rare words were randomly sampled and agreed upon as little-known by four fluent English speakers with postgraduate education. Words whose dictionary definitions did not meet the above criteria were rejected. Most of the words were known by none of them. 36 of the 74 words do not appear in the top 2 million words of the Common Crawl corpus, according to the GloVe embedding matrix (Pennington et al., 2014), and the median rank of the other words is 656565. The most common word is ‘impala’ (rank 89578).

B Human study details

The studies with human subjects were performed on Amazon Mechanical Turk, with workers from the pool of native English speakers with at least 95% approval rating. Users were paid an average of US$0.07 per pair in the word-definition matching task and $0.10 per word in the definition choice task, equivalent to a wage of about $20/hour at
the average speed of labeling. All data was collected anonymously and no information was retained other than the answers and time taken to complete the study.

C Fake and rare word lexicon

The full set of 220 words used in our definition-matching experiments can be found in Tables 3, 4, and 5. For each word, we computed the average confidence in the correct matching for all pairs containing the word shown to humans in the study. The first column in each table shows the rank of this average confidence (a lower number indicating that the word’s definition disambiguates it well).

The set of words from the Dictionary of Obscure Sorrows is shown in Tab. 6.
Table 3: The list of nouns and their definitions (fake words above the line, rare words below).
Table 4: The list of verbs and their definitions (fake words above the line, rare words below).
Table 5: The list of adjectives and their definitions (fake words above the line, rare words below).

| rank | word          | GPT-3 definition                                      | human definition                                                                 |
|------|---------------|-------------------------------------------------------|----------------------------------------------------------------------------------|
| 1    | allocism       | the pain of childbirth                                |                                                                                   |
| 2    | alacritious    | the pungency of space rock                            |                                                                                   |
| 3    | atavism        | the quality of being only capable of feeling extreme empathy |                                                                                   |
| 4    | atricydolous   | the quality of being only capable of feeling extreme empathy |                                                                                   |
| 5    | atrocity       | a state of suspended development                      |                                                                                   |
| 6    | atriandian     | a clumsy fortune teller                               |                                                                                   |
| 7    | atriandian     | a clumsy fortune teller                               |                                                                                   |
| 8    | atrophobia      | the irrational fear of going without paws             |                                                                                   |
| 9    | atropony       | the pungency of space rock                            |                                                                                   |
| 10   | atrophy        | the pungency of space rock                            |                                                                                   |
| 11   | atrophy        | the pungency of space rock                            |                                                                                   |
| 12   | atrophy        | the pungency of space rock                            |                                                                                   |
| 13   | atrophy        | the pungency of space rock                            |                                                                                   |
| 14   | atrophy        | the pungency of space rock                            |                                                                                   |
| 15   | atrophy        | the pungency of space rock                            |                                                                                   |
| 16   | atrophy        | the pungency of space rock                            |                                                                                   |
| 17   | atrophy        | the pungency of space rock                            |                                                                                   |
| 18   | atrophy        | the pungency of space rock                            |                                                                                   |
| 19   | atrophy        | the pungency of space rock                            |                                                                                   |
| 20   | atrophy        | the pungency of space rock                            |                                                                                   |
| 21   | atrophy        | the pungency of space rock                            |                                                                                   |
| 22   | atrophy        | the pungency of space rock                            |                                                                                   |
| 23   | atrophy        | the pungency of space rock                            |                                                                                   |
| 24   | atrophy        | the pungency of space rock                            |                                                                                   |
| 25   | atrophy        | the pungency of space rock                            |                                                                                   |
| 26   | atrophy        | the pungency of space rock                            |                                                                                   |
| 27   | atrophy        | the pungency of space rock                            |                                                                                   |
| 28   | atrophy        | the pungency of space rock                            |                                                                                   |
| 29   | atrophy        | the pungency of space rock                            |                                                                                   |
| 30   | atrophy        | the pungency of space rock                            |                                                                                   |
| 31   | atrophy        | the pungency of space rock                            |                                                                                   |
| 32   | atrophy        | the pungency of space rock                            |                                                                                   |
| 33   | atrophy        | the pungency of space rock                            |                                                                                   |
| 34   | atrophy        | the pungency of space rock                            |                                                                                   |
| 35   | atrophy        | the pungency of space rock                            |                                                                                   |
| 36   | atrophy        | the pungency of space rock                            |                                                                                   |
| 37   | atrophy        | the pungency of space rock                            |                                                                                   |
| 38   | atrophy        | the pungency of space rock                            |                                                                                   |
| 39   | atrophy        | the pungency of space rock                            |                                                                                   |
| 40   | atrophy        | the pungency of space rock                            |                                                                                   |
| 41   | atrophy        | the pungency of space rock                            |                                                                                   |
| 42   | atrophy        | the pungency of space rock                            |                                                                                   |
| 43   | atrophy        | the pungency of space rock                            |                                                                                   |
| 44   | atrophy        | the pungency of space rock                            |                                                                                   |
| 45   | atrophy        | the pungency of space rock                            |                                                                                   |
| 46   | atrophy        | the pungency of space rock                            |                                                                                   |
| 47   | atrophy        | the pungency of space rock                            |                                                                                   |
| 48   | atrophy        | the pungency of space rock                            |                                                                                   |
| 49   | atrophy        | the pungency of space rock                            |                                                                                   |
| 50   | atrophy        | the pungency of space rock                            |                                                                                   |
| 51   | atrophy        | the pungency of space rock                            |                                                                                   |
| 52   | atrophy        | the pungency of space rock                            |                                                                                   |

Table 6: The list of words from the Dictionary of Obscure Sorrows. The first column is the rank of the frequency with which GPT-3's definition was preferred.
| LLD | word pair                                                                                           |
|-----|-----------------------------------------------------------------------------------------------------|
| −4.24 | **bellamen** : a strip of land that juts up from the surrounding land                             |
|      | **blossard** : a garment made of cloth or leather                                                 |
| −6.7 | **stucenium** : a little roof, the soffit of a cornice, the median part of a pediment             |
|      | **persecole** : a small dome-shaped structure resembling a thimble on the top of an ear of corn   |
| −8.0 | **flambuna** : a stove-pipe                                                                       |
|      | **carcention** : a movement of the muscles of the nose                                             |
| +4.9 | **bellamen** : a strip of land that juts up from the surrounding land                             |
|      | **silicily** : British theater jargon for a comic actor                                           |
| +0.6 | **bellamen** : a strip of land that juts up from the surrounding land                             |
|      | **parascound** : a shallow canoe or raft                                                           |
| +0.9 | **parascound** : a shallow canoe or raft                                                            |
|      | **cantah** : a reindeer parka                                                                     |
| −0.2 | **shoutze** : to laugh through half-open teeth                                                        |
|      | **batherize** : to talk up, boast of, brag on                                                      |
| −5.7 | **encreen** : to draw attention to oneself with a display of bravery                               |
|      | **bedrame** : to augment a story or allegation with further details                                 |
| −4.3 | **batherize** : to talk up, boast of, brag on                                                       |
|      | **bedeak** : to plant or sow seeds; to place in the ground                                        |
| +1.0 | **inflleen** : to drench in blood                                                                  |
|      | **batherize** : to talk up, boast of, brag on                                                      |
| +2.0 | **disapplase** : to become insubordinate or rebellious                                              |
|      | **dreed** : to be in two minds; to be undecided                                                    |
| +0.2 | **beckain** : to touch gently                                                                      |
|      | **account** : to underestimate                                                                    |
| −5.4 | **importical** : lukewarm, unenthusiastic                                                          |
|      | **spiative** : driven by the need for independence                                                 |
| −4.7 | **despious** : loudly satirical or mocking                                                         |
|      | **bedduine** : friendly, genial                                                                    |
| −3.7 | **perpagant** : mutually involved; of or involving both parties.                                   |
|      | **carabodent** : keenly careful, attentive, painstaking                                            |
| +1.7 | **tricy** : containing light, of the nature of light                                              |
|      | **despious** : loudly satirical or mocking                                                         |
| +4.5 | **foreal** : arising from a mental vision, having visionary qualities                               |
|      | **paranory** : unsympathetically aggrieved by other people’s problems                              |
| +6.4 | **sterebous** : bad-tempered                                                                      |
|      | **despious** : loudly satirical or mocking                                                         |

Table 7: Additional random samples of word pairs with LLD between −10 and 10.