

GPT Perdetry Test: Generating new meanings for new words

Nikolay Malkin¹
Yale University
New Haven, CT

Sameera Lanka²
Microsoft
Redmond, WA

Pranav Goel³
University of Maryland
College Park, MD

Sudha Rao²
Microsoft Research
Redmond, WA

Nebojsa Jojic²
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.

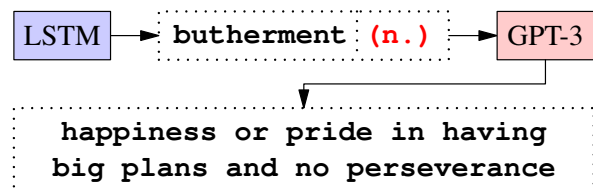
1 Introduction

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

¹We use the term 'nonce word' for a new word not used in English. It becomes a 'neologism' once it acquires a meaning.



per|detry (n.) – an instance of inventing words, esp. as a hobby
har|bole|mic (adj.) – tending to babble; talking nonsense
sh|out|ze (v.) – to laugh through half-open teeth

Figure 1: **Above:** The neologism generation pipeline. **Below:** Definitions created by GPT-3 for three nonce words. Prompts **bold**, subword tokens separated by |.

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.

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
<input type="radio"/> Option A is much better <input type="radio"/> Option A is better <input type="radio"/> Option A is a little better <input type="radio"/> Option B is much better <input type="radio"/> Option B is better <input type="radio"/> Option B is a little better	

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 (phonemes) 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 phonesthemes, 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 words² 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.² Definitions for these words were taken from a dictionary.³ 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-

²github.com/dwyl/english-words

³en.wiktionary.org

		n.	v.	adj.
human	fake-fake	70.7%	59.8%	64.3%
	fake-rare	72.6	60.9	65.3
	rare-rare	79.6	65.3	69.8
GPT-3	fake-fake	92.4	83.2	87.3
	fake-rare	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’).

dices **A** and **C** for the full lexicon and generation details, including all points of human input.

4 Evaluation

4.1 Machine-generated neologisms

We 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.⁴ 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 (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 $65 \cdot \frac{220}{2} = 7150$ 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 $(w_1 - d_2, w_2 - d_1)$:

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

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 $\text{LLD}(w_1, w_2)$ 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

⁴Users were not told that some of the definitions were machine-generated. The full study details are in Appendix B.

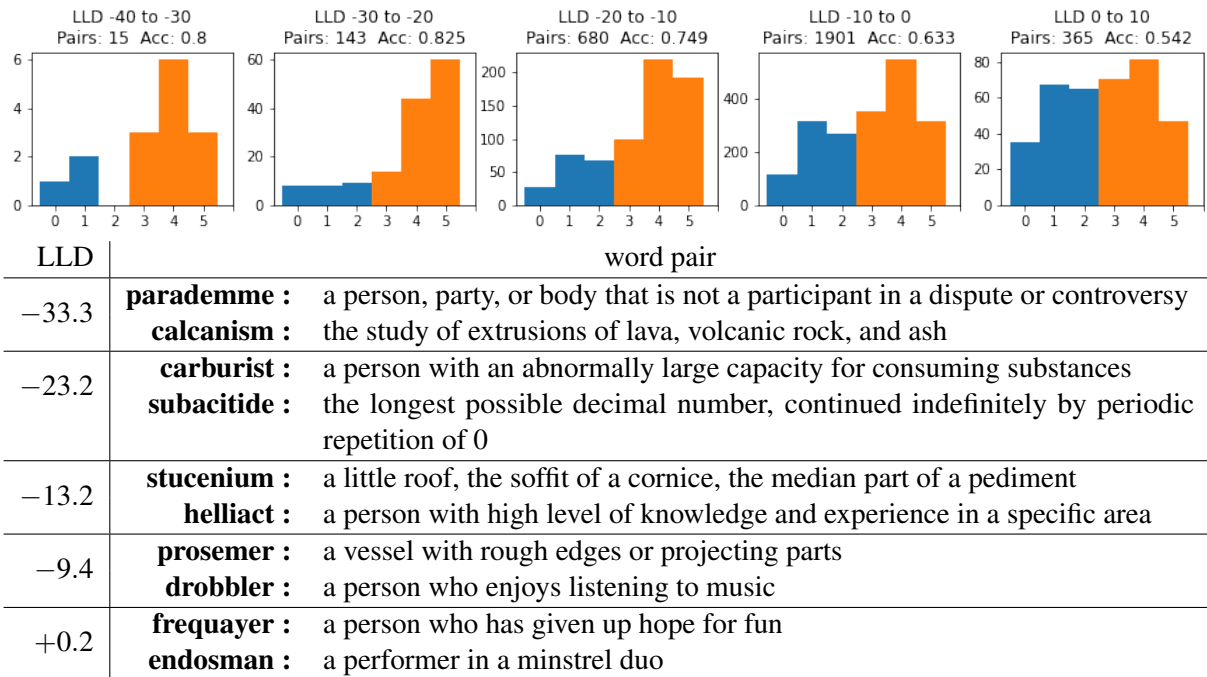


Table 2: **Above:** Histograms of confidence scores for fake-fake pairs with LLD in each of the five ranges defined in the text. Values of 3, 4, or 5 (orange) indicate preference for the correct matching.

Below: Randomly sampled pairs of fake nouns with LLD falling in each of the five ranges. A negative number with large magnitude indicates that GPT-3 – and usually human subjects – strongly prefers this pairing to the reverse. A number close to 0 indices that the pairing and the reverse have similar likelihood.

the correct matches for fake-fake pairs, since the definitions of fake words were sampled from the same model of likelihood. Indeed, we see this in the bottom rows of Table 1. GPT-3’s imperfect performance on fake-fake pairs is a byproduct of the sampling used in the generation and perhaps of the edits made in postprocessing. To maximize total likelihood of the lexicon, GPT-3 would prefer to enact some post-factum swaps of definitions.

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 $LLD(w_1, w_2)$, 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: $[-40, -30), [-30, -20), \dots, [0, 10)$. 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.⁵ Many of these words were created out of real English morphemes. We sampled 20 words from this set, got GPT-3 defini-

⁵dictionaryofobscuresorrows.com/

- 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.)

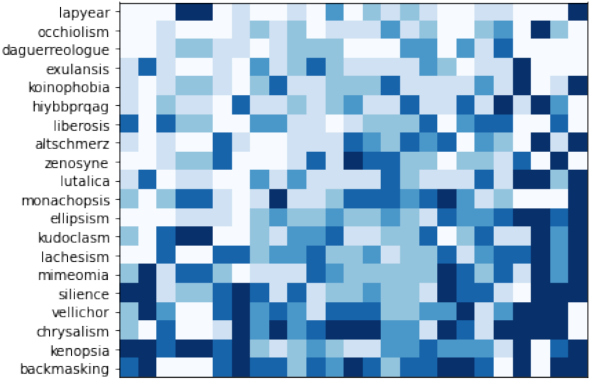


Figure 4: Human subjects’ preference for GPT-3-generated definitions (bluer) or human-generated definitions (whiter). Each column represents a single user. The rows and columns have been sorted by their means. The full set of definitions can be found in Appendix C.

tions for them using the same procedure as above, and also extracted human definitions (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 Fazleh 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., "frequayer (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 buttohole. (*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.

85	allote	an outstanding opera singer
86	anemord	the killing of a she-mule
89	annyiss	a bad smell, stink, stench, fetor
69	barthoon	a cheap, inferior, or fraudulent imitation
91	bellamen	a strip of land that juts up from the surrounding land
19	blossard	a garment made of cloth or leather
76	butherford	happiness or pride in having big plans and no perseverance
25	calcanism	the study of extrusions of lava, volcanic rock, and ash
40	cantah	a reindeer parka
26	carabacity	weakness; feebleness; bad luck
102	caraber	a male witch; a wizard; a warlock
46	carabert	one who constructs or repairs catamarans or other Polynesian sailing craft
41	carburist	a person with an abnormally large capacity for consuming substances
27	carcention	a movement of the muscles of the nose
99	carebock	an edible species of seaweed common on the Atlantic coast of Europe
49	decoment	a chemical used to control odors
28	demotence	a sense of powerlessness
74	drawch	a large, low area that is blackened
31	drobbler	a person who enjoys listening to music
20	endosman	a performer in a minstrel duo
60	flambuna	a stove-pipe
95	flaudite	a military salute
59	fleccon	a piece of cloth
53	frequayer	a person who has given up hope for fun
8	garalism	the practice of eating animals
11	harbonet	a barbed anchor
73	helliact	a person with high level of knowledge and experience in a specific area
98	hooddar	a person who is not a blood relative, but who is considered a part of the family
37	humity	ill temper, peevishness
42	hurran	a warm, gentle Mediterranean wind
72	maidentry	an alms-box in a church porch
7	malemetry	the study of geometric figures, esp. for purpose of measuring or comparing them
78	mandrope	a mystical union of opposites in the same entity
14	manicity	the extreme mental instability commonly occurring prior to a homicidal or suicidal act
45	marchasis	a female friend, a female companion, an associate, a close friend
4	mercartist	one who advocates realism
101	nonacy	one who secretly profits from another's dishonesty
64	notterin	a woman who rides in a basket attached to a rope, for amusement
34	opister	a giant spider
80	parademme	a person, party, or body that is not a participant in a dispute or controversy
79	parascound	a shallow canoe or raft
47	patabasity	an act of showing off
35	perdetry	an instance of inventing words, especially as a hobby
75	persecole	a small dome-shaped structure resembling a thimble on the top of an ear of corn
48	pinilet	a metal contraption used to clear a clogged sink
83	prexicule	a person who is very shy, especially in new social situations
97	pronocule	a person who is not wearing pants while riding a bicycle
87	prosemer	a vessel with rough edges or projecting parts
96	punguel	a sudden, brief forward movement of a body part
90	recommor	a female dwarf
57	sacrel	the small of the back
39	scourism	recreation that includes outdoor activities that combat blight
2	septanis	a seven-year plan
58	serodom	a person who is sacrificed to the gods
66	shrifful	a hodgepodge of things or persons
70	sigatch	a short, vertical stroke placed above a letter or word to indicate that it is a Hebrew abbreviation
84	silicily	British theater jargon for a comic actor
88	steenfook	a sleeping place; a place where one goes for enjoyment or rest
44	steffice	the place where a bishop has his seat or throne
52	sterpon	guardian, warden, protector, keeper, caretaker
61	stucenium	a little roof, the soffit of a cornice, the median part of a pediment
50	subactide	the longest possible decimal number, continued indefinitely by periodic repetition of 0
56	subtree	a base or source of artistic production
29	tagabism	a tendency to be trapped in a view or a way of thinking
51	transpanity	excessive appetite for salt
36	undergrism	the controversial practice of participating in retrograde activities within a group
94	wairl	an Anglo-Saxon stanzaic poem which imitates the stress patterns of an earlier poem
93	aeolipyle	a steam engine powered by rocket propulsion due to escaping steam
100	aroba	a horse-drawn carriage once used for transportation
21	boll	the rounded seed-bearing capsule of a cotton or flax plant
92	chott	a dry salt lake that stays dry in the summer but receives some water in the winter
3	cirrhopod	any barnacle or similar crustacean
22	codicil	an addition or supplement that explains, modifies, or revokes a will
30	concinnity	the harmonious reinforcement of the various parts of a work of art
17	crudite	a type of salad, usually put in sandwiches
12	dicrotism	the condition of each beat of the pulse consisting of two waves
63	estover	an allowance provided from an estate for a person's support
6	flagellum	a long, whiplike membrane-enclosed organelle used for locomotion or feeding
77	geat	the channel or spout through which molten metal runs into a mould in casting
32	hobgoblin	a source of dread, fear or apprehension
1	impala	an antelope noted for its leaping ability
43	jabot	a cascading or ornamental frill down the front of a blouse, shirt, etc.
71	loggia	a roofed, open gallery, usually on an upper level
23	maschie	a metal-headed golf club with a moderate loft
68	muntin	one of the separators between panes of glass in a composite window
33	nephalism	the practice of abstaining completely from the drinking of alcohol
24	pemmican	a food made from meat which has been dried and beaten into a paste
5	potentate	a powerful leader; a monarch, a ruler
67	poule	a girl, a young woman, especially seen as promiscuous
15	repartee	a swift, witty reply, especially one that is amusing
9	rhonchus	a dry rattling sound heard during breathing
18	rundle	a round, a step of a ladder, a rung
82	scumble	an opaque kind of glaze, layer of paint
65	sedum	any of various succulent plants native to temperate zones
55	staphyle	the uvula of the soft palate
13	swankpot	a show-off
10	sybarite	a person devoted to pleasure and luxury
38	toccatà	a piece of music designed to emphasize the dexterity of the performer
81	tuskar	a type of spade used for cutting peat
62	vimana	a mythological flying palace or chariot
16	wherry	a light ship used to navigate inland waterways
54	xoanon	a wooden statue used as a cult image

Table 3: The list of nouns and their definitions (fake words above the line, rare words below).

68	account	to underestimate
31	batherize	to talk up, boast of, brag on
36	beckain	to touch gently
41	bedeak	to plant or sow seeds; to place in the ground
70	bedrame	to augment a story or allegation with further details
25	belail	to miss, be lacking in, to need
13	belithe	to be endowed with, possess
7	bestrowe	to observe, behold, see, gape, stare, look
15	chestermate	to go in quest of a wife
18	disapplase	to become insubordinate or rebellious
32	dischall	to deny the truth of
17	dolonize	to renew, to resanctify the earth
52	dreed	to be in two minds; to be undecided
11	encraim	to be in love with ; desire intensely
63	encreen	to draw attention to oneself with a display of bravery
48	enfrone	to begin, to launch
14	excease	to decrease, fall off, grow smaller, lessen, diminish
59	fedulate	to award, grant
65	forfine	to administer physical punishment for a transgression
40	glongate	to break or deform in any part of the body
30	hoand	to strike or lash one's foot against the ground
53	infleen	to drench in blood
35	intersove	to move or walk with one's eyes closed, guided only by the sense of touch
64	marricate	to shoot or fire a weapon
26	meliserate	to pay special attention to someone for selfish reasons
8	misdeint	to be wrong; to be deceived
20	misdown	to dissipate, squander
67	mistrude	to hurry
9	percear	to look on the bright side
61	phreen	to be blind to
29	phrumb	to mix up, muddle
50	purpoom	to crush the game; be on point
21	regimble	to work out the details of a complex matter
66	reprine	to move about from excitement, to fluster or to bustle
57	respord	to dress in the latest fashion
42	sangalize	to spin or whirl, to turn swiftly
22	scolerize	to turn to stone
34	shoutze	to laugh through half-open teeth
49	squirse	to be indecisive, to have second thoughts
16	subcoint	to pool funds into a single account
37	subtice	to draw an inference; to derive meaning from context
44	superpate	to bring to heel, subdue, subjugate
10	travent	to cause to deviate or turn aside from a course
51	trionize	to speak with great eloquence or style
60	trovel	to be in a state of confusion or disarray
19	unimple	to break up, to disunite
62	warl	to torment or coerce someone
47	apocopate	to remove the final sound or syllable
28	bedizen	to ornament something in showy, tasteless, or gaudy finery
45	chine	to crack, split, fissure, break
38	constate	to relay information in a statement and say whether it is true or false
39	dislimn	to remove the outlines of; to efface
33	dizen	to dress with clothes; attire; deck
2	doff	to remove or tip a hat, as in greeting
5	flense	to strip the blubber or skin from
55	intumesce	to swell or enlarge
12	inveigle	to convert, convince, or win over with flattery or wiles
3	kyanize	to preserve wood from decay
69	laveer	to beat against the wind
54	lucubrate	to work diligently by artificial light, to study at night
43	marage	to make tough and malleable by means of heat treatment
56	moider	to perplex or bewilder
27	nidificate	to make or build a nest
24	oppilate	to block, to stop up, to obstruct
58	peculate	to steal or misappropriate money that one has been trusted with
6	rootle	to search for something from a drawer, closet, etc.
23	supererogate	to do more than God commands
46	taigle	to muddle, confuse
1	traipse	to walk in a messy or unattractively casual way
4	vacillate	sway unsteadily from one side to the other

Table 4: The list of verbs and their definitions (fake words above the line, rare words below).

22	antisard	unrelievedly unpleasant in taste
34	bedduline	friendly, genial
28	brawbly	rough and lacking in clarity
47	carbodent	keenly careful, attentive, painstaking
14	chariocious	deserving or inciting ridicule
13	chartish	short and easy to understand
37	choriant	likely to catch on fire
7	climyrean	bluish, azure
27	conceded	decayed with much use or age; shabby
6	confical	connective (e.g., tissues that support hearts or lungs)
32	despious	loudly satirical or mocking
36	finitrotic	having a tendency to destroy
16	foreal	arising from a mental vision, having visionary qualities
19	harbolemic	tending to babble; talking nonsense
5	immortaunt	having a high risk of immediate death
18	imperful	critical, unyielding
48	importical	lukewarm, unenthusiastic
44	memolessive	able to endure much suffering, hardened
41	oristitious	having a feeling of foreboding
33	paranory	unsympathetically aggrieved by other people's problems
20	perpagant	mutually involved; of or involving both parties.
11	reprimonic	having cells that some can multiply into a variety of cell types
23	scrappic	charmingly disheveled
25	spriative	driven by the need for independence
42	sterebous	bad-tempered
30	storie	afflicted with a disease of the bones
10	strideless	untouched by anxiety
17	suspeptory	never requiring reconsideration, settlement, or revision; irrevocable
40	syncrant	never satisfied
24	trapescent	sexually mature but not yet adult
45	tricy	containing light, of the nature of light
4	trigropose	alternating groups of three consecutive notes in rhythm
21	acescent	turning sour; readily becoming tart or acid
35	cantic	oblique, slanting
43	costive	miserly, parsimonious
8	decoctible	able to be boiled down
26	gremial	of or pertaining to the lap
9	grum	morose, stern, surly, sullen
3	lissom	flexible and graceful in movement
15	matutine	of or relating to early morning
2	piscatory	of or pertaining to fishermen or fishing
29	rakehell	immoral; dissolute
31	reasty	rusty and rancid
38	saprogenic	causing or resulting from putrefaction
12	serotinal	occurring in late summer)
1	stodgy	dull, old-fashioned
39	stridulant	making a high-pitched chirping, grating, hissing, or squeaking sound
46	tiliaceous	of, pertaining to, or resembling the linden

Table 5: The list of adjectives and their definitions (fake words above the line, rare words below).

rank	word	GPT-3 definition	human definition
6	altschmerz	the pain of childbirth	weariness with the same old issues that you've always had
20	backmasking	the act of disguising messages within recordings via sound effects	the instinctive tendency to see someone as you knew them in their youth
8	chrysalism	a state of suspended development	the amniotic tranquility of being indoors during a thunderstorm
1	daguerreologue	a clumsy fortune teller	an imaginary interview with an old photograph of yourself
18	ellipsism	the quality of being only capable of feeling extreme empathy	sadness that you'll never be able to know how history will turn out
11	exulansis	a pointy piece of space rock	the tendency to give up trying to talk about an experience because people are unable to relate to it
15	hiybbprqag	someone who bursts into tears easily	the feeling that everything original has already been done
5	kenopsia	the creepy feeling that something is wrong, when everything is actually fine	the eerie, forlorn atmosphere of a place that's usually bustling with people but is now abandoned and quiet
3	koinophobia	the irrational fear of going without pants	the fear that you've lived an ordinary life
19	kudoclasm	the organized resistance to the (generally computer-based) changes that come with the digital world	the destruction of lifelong dreams
10	lachesism	excessive taste for symmetry	the desire to be struck by disaster
14	lapyear	a lazy person; someone of a low-energy lifestyle	the age at which you become older than your parents were when you were born
9	liberosis	a slow lapse into apathy and a subsequent excuse for drinking all day	the desire to care less about things
12	lutalica	mixture of yodelling and headbanging	the part of your identity that doesn't fit into categories
2	mimeomia	the act of using too large a word when a small one will do	the frustration of knowing how easily you fit into a stereotype
17	monachopsis	the act of remembering a smell (especially an unpleasant one) years later	the subtle but persistent feeling of being out of place
7	occhiolism	a belief that personal power increases proportionally with one's height	the awareness of the smallness of one's perspective
4	silience	the illusion that noises we hear are different from those heard by others	the unnoticed creative excellence that happens around us every day
13	vellichor	the dim lightness on the brink of darkness; twilight on the universe	the strange wistfulness of used bookstores
16	zenosyne	a classical name for iodide of potassium	the sense that time keeps going faster

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 parasound : a shallow canoe or raft
+0.9	parasound : 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	infleen : 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 accont : to underestimate
-5.4	importical : lukewarm, unenthusiastic spriative : driven by the need for independence
-4.7	despious : loudly satirical or mocking bedduline : 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.