From *chocolate bunny* to *chocolate crocodile*: Do Language Models Understand Noun Compounds?

Jordan Coil[1] **and Vered Shwartz**[1,2]
[1] University of British Columbia [2] Vector Institute for AI
jcoil93@students.cs.ubc.ca, vshwartz@cs.ubc.ca

Abstract

Noun compound interpretation is the task of expressing a noun compound (e.g. *chocolate bunny*) in a free-text paraphrase that makes the relationship between the constituent nouns explicit (e.g. *bunny-shaped chocolate*). We propose modifications to the data and evaluation setup of the standard task (Hendrickx et al., 2013), and show that GPT-3 solves it almost perfectly. We then investigate the task of noun compound conceptualization, i.e. paraphrasing a novel or rare noun compound. E.g., *chocolate crocodile* is a crocodile-shaped chocolate. This task requires creativity, commonsense, and the ability to generalize knowledge about similar concepts. While GPT-3's performance is not perfect, it is better than that of humans—likely thanks to its access to vast amounts of knowledge, and because conceptual processing is effortful for people (Connell and Lynott, 2012). Finally, we estimate the extent to which GPT-3 is reasoning about the world vs. parroting its training data. We find that the outputs from GPT-3 often have significant overlap with a large web corpus, but that the parroting strategy is less beneficial for novel noun compounds.

1 Introduction

Noun compounds (NCs) are prevalent in English, but most individual NCs are infrequent (Kim and Baldwin, 2007). Yet, it is possible to derive the meaning of most NCs from the meanings of their constituent nouns. The task of noun compound interpretation (NCI) addresses this by explicitly uncovering the implicit semantic relation between the constituent nouns. We focus on the paraphrasing variant (Nakov and Hearst, 2006), where the goal is to generate multiple paraphrases that explicitly express the semantic relation between the constituents. For example (Figure 1), a *chocolate bunny* is a "chocolate shaped like a bunny".

Earlier methods for NCI represented NCs as a function their constituents' representations (e.g.

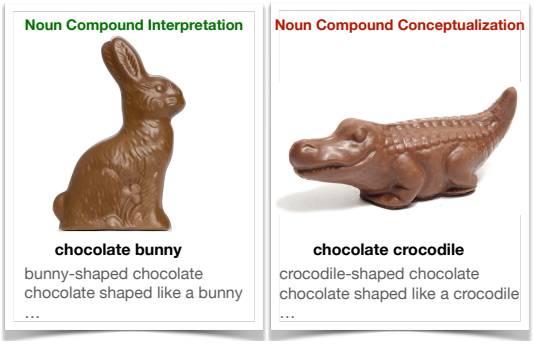

Figure 1: An example NC (input) and paraphrases (output) for each of the NCI and NCC tasks.

Van de Cruys et al., 2013; Shwartz and Dagan, 2018). In recent years, pre-trained language models (PLMs) caused a paradigm shift in NLP. Such models are based on the transformer architecture (Vaswani et al., 2017), which by design computes a word representation as a function of the representation of its context. Further, PLMs are pre-trained on vast amounts of text, which equips them with broad semantic knowledge (Rogers et al., 2020). Such knowledge may facilitate interpreting unseen NCs based on observed NCs that are semantically similar. Indeed, Ponkiya et al. (2020) showed that a masked language model is useful for this task, and Shwartz (2021) demonstrated the utility of generative language models on this task.

We formalize the experiments presented in Shwartz (2021) and evaluate generative models on NCI. We manually analyze and correct many problems with the standard SemEval 2013 task 4 dataset (Hendrickx et al., 2013), and release a cleaned version of the dataset. Following the criticism in Shwartz and Dagan (2018) on the task's dedicated evaluation metrics, we propose a more complete set of evaluation metrics including both automatic metrics and human evaluation.

Our experiments show that a few-shot model based on GPT-3 (Brown et al., 2020) achieves near-

Findings of the Association for Computational Linguistics: ACL 2023, pages 2698–2710

perfect performance on the NCI test set. The impressive performance may be due to a combination of factors. First, it tends to memorize texts seen during pre-training (Carlini et al., 2022), likely including partial or complete definitions of common NCs. Second, it has learned vast commonsense and world knowledge from its pre-training corpus, which—together with its ability to generalize—may be useful for interpreting less frequent NCs.

To test the extent that GPT-3 reasons about its knowledge as opposed to memorizes definitions, we propose a second task: noun compound conceptualization (NCC). The setup is identical to NCI, but the NCs are rare or novel (e.g., *chocolate crocodile* in Fig. 1), requiring a model to come up with a plausible interpretation based on its existing knowledge. We construct a test set for this task based on data from Dhar and van der Plas (2019). The results show that GPT-3 outperforms humans on NCC, presumably thanks to its fast access to a huge "knowledge base", and compared to the relative human slowness on this task (Connell and Lynott, 2012).

Yet, compared to its performance on NCI, GPT-3's performance on NCC shows a significant drop. We thus quantify the extent that GPT-3 copies from its pre-training corpus when generating paraphrases for either NCI or NCC. We find that the generated paraphrases have significant overlap with a large web-based corpus, but that as expected, the copying strategy is less beneficial for NCC than for NCI.

We anticipate that the cleaned dataset and proposed evaluation setup will be adopted by the research community for NCI, and hope to see further research on NCC.[1]

2 Background

2.1 Noun Compound Interpretation

Traditionally, NCI has been framed as a classification task into predefined relation labels. Datasets differed by the number of relations and their specificity level; from 8 prepositional relations (e.g. `of`, `from`, etc.; Lauer, 1995), to finer-grained inventories with dozens of relations (e.g. `contains`, `purpose`, `time of`; Kim and Baldwin, 2005; Tratz and Hovy, 2010). The classification approach is limited because even the larger relation inventories don't cover all possible relationships between nouns. In addition, each NC is classified to a single relation, although several relations may be appropriate. E.g., *business zone* is both a zone that `contains` businesses and a zone whose `purpose` is business (Shwartz and Dagan, 2018).

For these reasons, in this paper we focused on the task of interpreting noun compounds by producing multiple free-text paraphrases (Nakov and Hearst, 2006). The reference paraphrases could be any text, but in practice they typically follow a "[n_2] ... [n_1]" pattern, where n_1 and n_2 are the constituent nouns. The main dataset for this task comes from SemEval 2013 task 4 (Hendrickx et al., 2013), following a similar earlier task (Butnariu et al., 2009).

Earlier methods for this task reduced the paraphrasing task into a classification task to one of multiple paraphrase templates extracted from a corpus (Kim and Nakov, 2011; Paşca, 2015; Shwartz and Dagan, 2018). Shwartz and Dagan (2018) jointly learned to complete any item in the ([n_1], [n_2], paraphrase template) tuple, which allowed the model to generalize, predicting paraphrases for rare NCs based on similarity to other NCs.

More recently, Ponkiya et al. (2020) showed that PLMs already capture this type of knowledge from their pre-training. They used an off-the-shelf T5 model to predict the mask substitutes in templates such as "[n_2] [MASK] [n_1]", achieving a small improvement over Shwartz and Dagan (2018). Shwartz (2021) further showed that supervised seq2seq models based on PLMs and a few-shot model based on GPT-3 yielded correct paraphrases for both common and rare NCs.

2.2 Forming and Interpreting new Concepts

Research in cognitive science studied how people interpret new noun-noun combinations such as *cactus fish* (e.g. Wisniewski, 1997; Costello and Keane, 2000; Connell and Lynott, 2012). While such combinations invite various interpretations, there is usually a single preferred interpretation which is more intuitively understood. For example, a *cactus fish* would more likely mean "a fish that is spiky like a cactus" than "a fish that is green like a cactus", because "spiky" is more characteristic of cacti than "green" (Costello and Keane, 2000).

Connell and Lynott (2012) constructed a set of 27 novel NCs and asked people to (1) judge the sensibility of an NC; and (2) come up with a plausible interpretation. The short response times for the sensibility judgment task indicated that participants

[1]The code and data are available at: `https://github.com/jordancoil/noun-compound-interpretation`

relied on shallow linguistic cues as shortcuts, such as the topical relatedness between the constituent nouns. Response times in the interpretation generation task were longer, indicating that participants employed a slower process of mental simulation. Interpreting a new concept required building a detailed representation by re-experiencing or imagining the perceptual properties of the constituent nouns.

Computational work on plausibility judgement for NCs involves rare NCs (Lapata and Lascarides, 2003) and novel NCs (Dhar and van der Plas, 2019). The latter built a large-scale dataset of novel NCs by extracting positive examples from different decades in the Google Ngram corpus for training and testing. Negative examples were constructed by randomly replacing one of the constituents in the NC with another noun from the data. They proposed an LSTM-based model that estimates the plausibility of a target NC based on the pairwise similarity between the constituents of the target NC and other, existing NCs. For example, the candidate NC *glass canoe* was predicted as plausible thanks to its similarity to *glass boat*.

In this paper, we go beyond plausibility judgement to the more complicated task of interpretation. In concurrent work, Li et al. (2022) conducted similar experiments evaluating GPT-3's ability to define common and new noun compounds, as well as combinations of nonce words. They found no evidence that GPT-3 employs human-like linguistic principles when interpreting new noun compounds, and suggested it might be memorizing lexical knowledge instead. We further try to quantify the latter possibility in this work.

Similarly to novel NCs, Pinter et al. (2020b) look at novel blends from the NYTWIT corpus, collected automatically from a Twitter bot that tweets words published for the first time in the NYT (Pinter et al., 2020a). For example, *thrupple* is a blend of three and couple, used to describe "A group of three people acting as a couple". They found that PLMs struggled to separate blends into their counterparts.

In a related line of work on creativity, researchers proposed models that coin new words from existing ones. Deri and Knight (2015) generated new blends such as *frenemy* (friend + enemy). Mizrahi et al. (2020) generated new Hebrew words with an algorithm that is inspired by the human process of combining roots and patterns.

3 Noun Compound Interpretation

We first evaluate PLMs' ability to interpret existing noun compounds. We focus on the free-text paraphrasing version of NCI, as exemplified in Table 2. We use the standard dataset from SemEval 2013 Task 4 (Hendrickx et al., 2013). We identified several problems in the dataset that we address in Sec 3.1. We then trained PLM-based models on the revised dataset (Sec 3.2), and evaluated them both automatically and manually (Sec 3.3).

3.1 Data

We manually reviewed the SemEval-2013 dataset and identified several major issues with the data quality. We propose a revised version of the dataset, with the following modifications.

Train-Test Overlap. We discovered 32 NCs that appeared in both the training and test sets, and removed them from the test set.

Incorrect Paraphrases. We manually corrected paraphrases with superficial problems such as spelling or grammatical errors, redundant spaces, and superfluous punctuation. We also identified and removed NCs that were semantically incorrect. For example, *rubber glove* was paraphrased to "gloves has been made to get away from rubber", perhaps due to the annotator mistaking the word *rubber* for *robber*. Finally, we found and removed a few paraphrases that contained superfluous or subjective additions, deviating from the instructions by Hendrickx et al. (2013). For example, *tax reduction* was paraphrased as "reduction of tax *hurts the economy*", and *engineering work* as "work done *by men* in the field of engineering". Further, we discarded a total of 14 NCs from the training set and 11 NCs from the test set that had no correct paraphrases. In total, we removed 1,960 paraphrases from the training set and 5,066 paraphrases from the test set.

"Catch-All" Paraphrases. The paraphrases in Hendrickx et al. (2013) were collected from crowdsourcing workers. An issue with the crowdsourcing incentive structure, is that it indirectly encourages annotators to submit any response, even when they are uncertain about the interpretation of a given NC. In the context of this dataset, this incentive leads to what we call "catch-all" paraphrases. Such paraphrases include generic prepositional paraphrases such as "[n_2] of [n_1]" (e.g. "drawing of chalk").

	Original			Revised		
	train	**dev**	**test**	**train**	**dev**	**test**
#NCs	174	0	181	160	28	110
#paraphrases	4,256	0	8,190	5,441	1,469	4,820

Table 1: Statistics of the original SemEval 2013 dataset (Hendrickx et al., 2013) vs. our revised version (henceforth: the NCI dataset).

For verbal paraphrases, the include generic verbs, such as "[n_2] based on [n_1]", "[n_2] involving [n_1]", "[n_2] associated with [n_1]", "[n_2] concerned with [n_1]", and "[n_2] coming from [n_1]". While these paraphrases are not always incorrect, they are also not very informative of the relationship between the constituent nouns. We therefor removed such paraphrases.[2]

Data Augmentation. To increase the size of the dataset in terms of paraphrases and facilitate easier training of models, we performed semi-automatic data augmentation. Using WordNet (Fellbaum, 2010), we extended the set of paraphrases for each NC by replacing verbs with their synonyms and manually judging the correctness of the resultant paraphrase. We also identified cases were two paraphrases could be merged into additional paraphrases. For example, *steam train* contained the paraphrases "train powered by steam" and "train that operates using steam", for which we added "train operated by steam" and "train that is powered using steam". Overall, we added 3,145 paraphrases to the training set and 3,115 to the test set.

We followed the same train-test split as the original dataset, but dedicated 20% of the test set to validation. Table 1 displays the statistics of the NCI datasets.

3.2 Methods

We evaluate the performance of two representative PLM-based models on our revised version of the SemEval-2013 dataset (henceforth: the NCI dataset): a supervised seq2seq T5 model (Raffel et al., 2020) and a few-shot prompting GPT-3 model (Brown et al., 2020).

Supervised Model. We trained the seq2seq model from the Transformers package (Wolf et al., 2019), using T5-large. We split each instance in the dataset into multiple training examples, with the NC as input and a single paraphrase as output. We used the default learning rate (5×10^{-5}), batch size (16), and optimizaer (Adafactor). We stopped the training after 4 epochs when the validation loss stopped improving. During inference, we used top-p decoding (Holtzman et al., 2020) with $p = 0.9$ and a temperature of 0.7, and generated as many paraphrases as the number of references for a given NC.

Few-shot Model. We used the `text-davinci-002` GPT-3 model available through the OpenAI API. We randomly sampled 10 NCs, each with one of its paraphrases, from the training set, to build the following prompt:

```
Q: what is the meaning of <NC>?
A:<paraphrase>
```

This prompt was followed by the same question for the target NC, leaving the paraphrases to be completed by GPT-3. We used the default setup of top-p decoding with $p = 1$ and a temperature of 1.

3.3 Evaluation

We decided to deviate from the original evaluation setup of the SemEval 2013 dataset, which was criticized in Shwartz and Dagan (2018). We describe the original evaluation setup, and our proposed setup including automatic and manual evaluation.

Original Evaluation Setup. The original SemEval task was formulated as a ranking task. The paraphrases of each NC were ranked according to the number of annotators who proposed them. Hendrickx et al. (2013) introduced two dedicated evaluation metrics, an 'isomorphic' score that measured the recall, precision, and order of paraphrases predicted by the systems, and a 'non-isomorphic' score that disregarded the order. Both metrics rewarded systems for predicting shorter prepositional paraphrases (e.g. "[n_2] of [n_1]"), that were in the set of paraphrases for many NCs, and were often ranked high because many annotators proposed them. For example, for the NC *access road*, the catch-all paraphrase "road for access" was ranked higher than the more informative "road that provides access". Indeed, as noted in Shwartz and Dagan (2018), a baseline predicting a fixed set of common, generic paraphrases already achieves moderately good non-isomorphic score. In general, we do not see the benefit of the ranking system,

[2] Another factor for the quality of paraphrases is the workers' English proficiency level. Writing non-trivial paraphrases requires high proficiency, and in 2013, it wasn't possible to filter workers based on native language on Mechanical Turk.

NC	GPT-3	T5
access road	road that provides access	road for access
reflex action	a sudden, involuntary response to a stimulus	action performed to perform reflexes
sport page	a page in a publication that is devoted to sports	page dedicated to sports
computer format	the way in which a computer organizes data	format used in computers
grief process	process of grieving or mourning	process that a grief sufferer experiences

Table 2: Example paraphrases generated using GPT-3 and T5 for NCs in the revised SemEval 2013 test set.

Method	METEOR	ROUGE-L	BERTScore	Human
T5	**69.81**	**65.96**	**95.31**	65.35
GPT-3	56.27	47.31	91.94	**95.64**

Table 3: Performance of the T5 and GPT-3 models on the revised SemEval 2013 test set.

since some of the most informative paraphrases are unique and are less likely to have been proposed by many annotators. Instead, we propose to use standard evaluation metrics for generative tasks, as we describe below.

Automatic Evaluation. Table 3 (columns 2-4) displays the performance of T5 and GPT-3 on the test set using the following standard evaluation metrics for text generation tasks: the lexical overlap metrics ROUGE-L (Lin, 2004) and METEOR (Lavie and Agarwal, 2007), and the semantic-similarity metric BERT-Score (Zhang et al., 2020). These metrics compare the system generated paraphrases with the reference paraphrases, further motivating our data augmentation in Sec 3.1 (e.g., Lin (2004) found that considering multiple references improves ROUGE's correlation with human judgements). For each metric m, we compute the following score over the test set T:

$$s = \mathrm{mean}_{\mathrm{nc}\in T}\big[\,\mathrm{mean}_{\mathrm{p}\in\mathrm{system(nc)}} \max_{\mathrm{r}\in\mathrm{references(nc)}} \mathrm{m}(p, r)\big]$$

In other words, we generate a number of paraphrases equal to the number of reference paraphrases, then find the most similar reference for each of the generated paraphrases, and average across all paraphrases for each NC in the test set.

The automatic metrics show a clear preference to T5. However, upon a closer look at the outputs of each model, it seems that T5 generated paraphrases that more closely resembled the style and syntax of the references, as expected from a supervised model, but the paraphrases were not "more correct" than those outputted by GPT-3. For example, in Table 2, the paraphrase generated by GPT-3 for *reflex action* is correct but doesn't follow the syntax of the references in the training data ([n_2] ... [n_1]). The T5-generated paraphrase follows that syntax but generates the generic and inaccurate paraphrase "action performed to perform reflexes". More broadly, lexical overlap based metrics such as ROUGE and METEOR penalize models for lexical variability.

Human Evaluation. To assess the quality of predictions in a more reliable manner, we turn to human evaluation. We used Amazon Mechanical Turk (MTurk) and designed a human intelligence task (HIT) which involved displaying an NC along with 10 generated paraphrases, 5 from GPT-3 and 5 from T5, randomly shuffled. We asked workers to indicate for each paraphrase whether they deemed it acceptable or not. Each HIT was to be performed by 3 workers, and acceptability was measured using majority voting. To ensure the quality of workers, we required that workers reside in the US, Canada, or the UK, and that they had an acceptance rate of at least 99% for all prior HITs. We also required them to pass a qualification task that resembled the HIT itself. We paid each worker $0.10 per task, which yielded an approximate hourly wage $15.

The last column in Table 3 presents the results of the human evaluation in terms of percentage of paraphrases deemed acceptable by a majority of human evaluators. GPT-3 performed remarkably well with over 95% of generated paraphrases deemed acceptable by a majority of human evaluators. In contrast to the automatic metrics, T5 fared much worse on human evaluation, and human annotators judged a third of T5 outputs as incorrect.

4 Noun Compound Conceptualization

GPT-3's impressive success at interpreting existing noun compounds is related to PLMs' ability

to associate nouns with their hypernyms (Ettinger, 2020) and to generate accurate definitions for terms (Shwartz et al., 2020). Such models are trained on vast amounts of texts, including said definitions, and the target NC itself occurring alongside contexts that indicate its meaning. Humans are different in their ability to interpret NCs. We can often rely on a single context, or no context at all, to have at least an educated guess at the meaning of a new NC. We are capable of representing new concepts by "mentally manipulating old ones" (Connell and Lynott, 2012), e.g. coming up with a plausible interpretation for *chocolate crocodile* based on similar concepts such as *chocolate bunny*. Prior work on NCI simulated this by training a model to jointly predict a paraphrase as well as answer questions such as "what can chocolate be shaped like?" (Shwartz and Dagan, 2018). We are interested in learning whether PLMs already do this implicitly, or more broadly, to what extent can PLMs interpret new noun compounds?

Inspired by studies in cognitive science about "conceptual combination" (Wisniewski, 1997; Costello and Keane, 2000), we define the task of Noun Compound Conceptualization (NCC). NCC has the same setup as NCI (§3), but the inputs are rare or novel noun compounds. The task thus requires some level of creativity and the ability to make sense of the world. We first describe the creation of the NCC test set (Sec 4.1). We evaluate the best model from Sec 3.2 on the new test set, and present the results in Sec 4.2.

4.1 Data

We construct a new test set consisting of novel or rare NCs. The guidelines for adding an NC for the test set are that: (a) humans could easily make sense of it; but (b) it is infrequent in or completely absent from the web.

Noun Compounds. The main source for the test set is a dataset from Dhar and van der Plas (2019). They proposed the task of classifying an unseen sequence of two nouns to whether it can form a plausible NC or not. The data was created by extracting noun-noun bigrams from the Google Ngram corpus (Brants, 2006). To simulate novel NCs, the models were trained on bigrams that only appeared in the corpus until the year 2000 and evaluated on bigrams that only appeared after 2000. Since GPT-3 was trained on recent data, we had to make sure that we only include the most infrequent NCs.

Test Set	NCI	NCC
Human Performance	-	73.33
GPT-3	95.64	83.81

Table 4: Human evaluation performance (percent of correct paraphrases) of paraphrases proposed by people or generated by GPT-3 for the NCI and NCC test sets.

We thus further refined the data from Dhar and van der Plas (2019) by including only the 500 most infrequent NCs based on their frequency in a large-scale text corpus, C4 (Raffel et al., 2020). We then semi-automatically filtered out named entities, compounds that were part of larger expressions, and NCs with spelling errors. Finally, we manually chose only the NCs for which we could come up with a plausible interpretation, leaving us with 83 NCs in total.

We added 22 more NCs that we extracted in a similar manner from the Twitter sentiment 140 dataset (Go et al., 2009). We expected to find more "ad-hoc" NCs in tweets than in more formal texts such as news. Due to the age and size of this dataset, we filtered the NCs based on frequency in C4, setting the threshold to 250 occurances. Overall, our NCC test set contains a total of 105 NCs.

Paraphrases. We collected reference paraphrases for the NCC test set using MTurk. We showed workers the target NC and asked them to paraphrase the NC or give their best estimate if they are unfamiliar with the NC. We used the same qualifications as in Sec 3.3, and paid $0.12 per HIT.

4.2 Evaluation

We focus on GPT-3 due to its almost perfect performance on NCI. We evaluated GPT-3 on the NCC test set using the few-shot setup described in Sec 3.2. We selected the few-shot examples from the NCI training set.

We focus on human evaluation (as described in Sec 3.3), which is more reliable than automatic metrics. We asked workers to judge the validity of both human-written and GPT-3 generated paraphrases.

Table 4 shows that GPT-3 performs significantly better than humans at this task. GPT-3 benefits from access to huge amounts of data. We conjecture that even though the target NCs are rare in its training data, it likely observed similar NCs, and is able to generalize and make sense of new concepts. At the same time, while humans are in general ca-

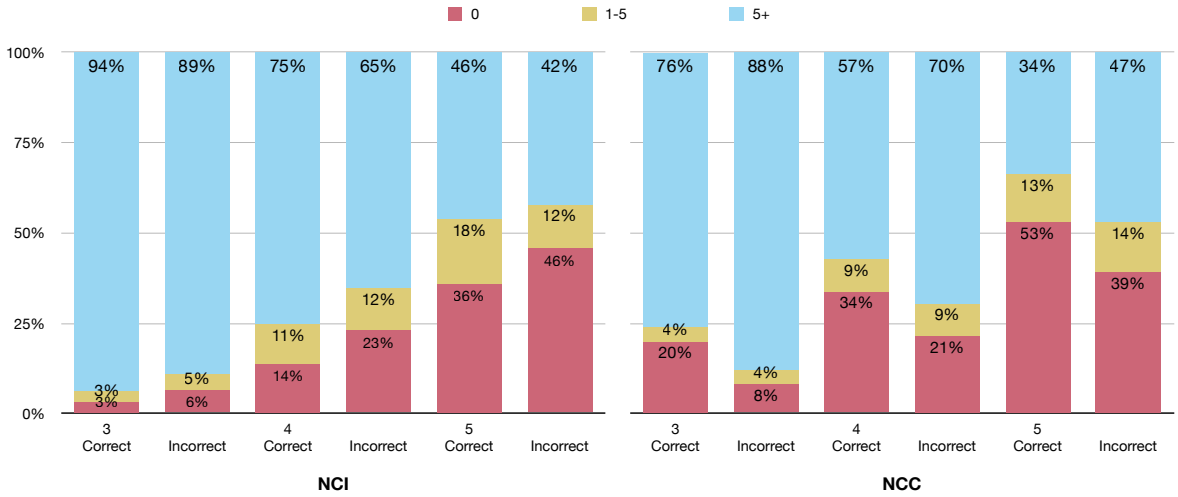

Figure 2: The percent of n-grams among the generated paraphrases (for $n = \{3, 4, 5\}$) that occur in the C4 corpus 0, 1-5, or 5+ times, for each of the NCI and NCC test sets, grouped by correct vs. incorrect generated paraphrases.

pable of coming up with a plausible interpretation for an unfamiliar concept, it is an effortful and cognitively taxing task. We hypothesize that in a setup other than crowdsourcing, i.e. given more time or incentive, human performance may increase.

Compared to its performance on NCI, GPT-3's performance on NCC shows a significant drop. This may suggest that GPT-3 struggles to reason about certain rare NCs, which we investigate in the next section.

5 Does GPT-3 Parrot its Training Data?

While GPT-3 performs fairly well on NCC, looking at failure cases brings up interesting observations. For example, one of its responses for *chocolate crocodile* was "A large, aggressive freshwater reptile native to Africa". This response seems to have ignored the *chocolate* part of the NC entirely, and opted to provide an answer to "What is a crocodile?". Much like a student who doesn't know the answer to a question so instead regurgitates everything they memorized about the topic in hopes that it will include the correct answer.[3]

To quantify the extent to which GPT-3 may be parroting its training corpus, we look at n-gram overlap between GPT-3's generated paraphrases and the large-scale web-based corpus C4 (Raffel et al., 2020).[4]

Figure 2 displays the percents of n-grams among the generated paraphrases (for $n = \{3, 4, 5\}$) that occur in the C4 corpus 0, 1-5, or 5+ times, for each of the NCI and NCC test sets. The results are presented separately for paraphrases deemed correct and incorrect by human evaluators.

We learn several things from Figure 2. First, the generated paraphrases often had significant overlap with the corpus (34-94%). As expected, trigrams are copied more than 4-grams, which are copied more than 5-grams, as those tend to be rarer.

Second, for the NCI test set, for each n, we see that n-grams from the correct paraphrases are copied from the web more often than n-grams from the incorrect paraphrases. The trend is reversed for NCC, where incorrect paraphrases are copied from the web more often than correct ones. Naturally, the copying strategy is less useful for NCC, which requires reasoning about new concepts. When GPT-3 generates correct paraphrases for NCC, their n-grams tend to not appear in the web at all.

We reach a similar conclusion by looking at the percent of n-grams in correct vs. incorrect paraphrases that are copied from the web. The vast majority of n-grams copied from the web (97%) for the NCI test set were correct, as opposed to only 80% for NCC.

6 Conclusion

We evaluated PLMs on their ability to paraphrase existing and novel noun compounds. For interpre-

[3] A similar phenomenon was also demonstrated in concurrent work from Li et al. (2022). They showed that for instance, GPT-3 defines a *banana table* as a *banana* rather than a *table*, differently from humans.

[4] We don't have access to the GPT-3 training corpus, but it included Common Crawl, web texts, books, and Wikipedia. C4 (Raffel et al., 2020) is a colossal, cleaned version of Common Crawl, thus it is the closest to GPT-3's training corpus.

tation of existing NCs (NCI), we released a cleaned version of the SemEval 2013 dataset, with manual correction and automatic augmentation of paraphrases, and proposed additional evaluation metrics to overcome limitations described in prior work. GPT-3 achieved near perfect performance on this new test set. We then investigated the task of noun compound conceptualization (NCC). NCC evaluates the capacity of PLMs to interpret the meaning of new NCs. We showed that GPT-3 still performs reasonably well, but its success can largely be attributed to copying definitions or parts of definitions from its training corpus.

7 Limitations

Human performance on NCC. The human accuracy on NCC was 73%, compared to 83% for GPT-3. We know from cognitive science research that humans are capable of forming new concepts based on existing ones (Connell and Lynott, 2012). Moreover, we manually selected NCs in the NCC test set that we could come up with a plausible interpretation for. The fact that 27% of the paraphrases proposed by MTurk workers were judged as incorrect could be explained by one of the following. The first explanation has to do with the limitations of crowdsourcing. To earn enough money, workers need to perform tasks quickly, and conceptualization is a slow cognitive process. On top of that, a worker that has already spent considerable amount of time trying to come up with a plausible interpretation for a new NC, is incentivized to submit any answer they managed to come up with, regardless of its quality. Skipping a HIT means lost wages. In a different setup, we hypothesize that human performance may increase for this task.

The second explanation has to do with the evaluation setup. We asked people to judge paraphrases as correct or incorrect. Upon manual examination of a sample of the human-written paraphrases, we observed a non-negligible number of reasonable (but not optimal) paraphrases that were annotated as incorrect. For future work, we recommend doing a more nuanced human evaluation that will facilitate comparing the outputs of humans and models along various criteria.

The work focuses only on English. Our setup and data construction methods are fairly generic and we expect it to be straightforward to adapt them to other languages that use noun compounds. With that said, languages such as German, Norwegian, Swedish, Danish, and Dutch write noun compounds as a single word. Our methods will not work on these languages without an additional step of separating the NC into its constituent nouns, similar to unblending blends (Pinter et al., 2020b). In the future, we would like to investigate how well PLMs for other languages perform on NCI and NCC, especially for low-resource languages.

Limitations of automatic metrics for generative tasks. Automatic metrics based on n-gram overlap are known to have low correlation with human judgements on various NLP tasks (Novikova et al., 2017). In particular, they penalize models for lexical variability. To mitigate this issue, we semi-automatically expanded the set of reference paraphrases using WordNet synonyms. Yet, we still saw inconsistencies with respect to the automatic metrics and human evaluation on NCI. The automatic metrics showed a clear preference to T5, which thanks to the supervision, learned to generate paraphrases that more closely resembled the style and syntax of the references. GPT-3's paraphrases, which were almost all judged as correct by human annotators, were penalized by the automatic metrics for their free form (e.g., they didn't always include the constituent nouns). For this reason, we focused only on human evaluation for NCC.

8 Ethical Considerations

Data Sources. All the datasets and corpora used in this work are publicly available. The cleaned version of the NCI dataset is based on the existing SemEval 2013 dataset (Hendrickx et al., 2013). The NCs for the new NCC test set were taken from another publicly-available dataset (Dhar and van der Plas, 2019), based on frequencies in the Google Ngram corpus (Brants, 2006). To quantify Ngram overlap, we used the Allen AI version of the C4 corpus (Raffel et al., 2020; Dodge et al., 2021) made available by the HuggingFace Datasets package.[5]

Data Collection. We performed human evaluation using Amazon Mechanical Turk. We made sure annotators were fairly compensated by computing an average hourly wage of $15, which is well above the US minimum wage. We did not collect any personal information from annotators.

Models. The models presented in this paper are for a low-level NLP task rather than for an appli-

[5] https://huggingface.co/datasets/c4

cation with which users are expected to interact directly. The generative models are based on PLMs, which may generate offensive content if prompted with certain inputs.

Acknowledgements

This work was funded, in part, by an NSERC USRA award, the Vector Institute for AI, Canada CIFAR AI Chairs program, an NSERC discovery grant, and a research gift from AI2.

References

Thorsten Brants. 2006. Web 1t 5-gram version 1. *http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp? catalogId= LDC2006T13*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Louise Connell and Dermot Lynott. 2012. Flexible shortcuts: Linguistic distributional information affects both shallow and deep conceptual processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Fintan J. Costello and Mark T. Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349.

Aliya Deri and Kevin Knight. 2015. How to make a frenemy: Multitape FSTs for portmanteau generation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–210, Denver, Colorado. Association for Computational Linguistics.

Prajit Dhar and Lonneke van der Plas. 2019. Learning to predict novel noun-noun compounds. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 30–39, Florence, Italy. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford 1.12*.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Su Nam Kim and Timothy Baldwin. 2007. Interpreting noun compounds using bootstrapping and sense collocation. *Proc. of the Pacific Association for Computational Linguistics (PACLING)*.

Su Nam Kim and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 648–658, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mirella Lapata and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Mark Lauer. 1995. Designing statistical language learners: Experiments on noun compounds. *Ph. D. Thesis, Department of Computing Macquarie University*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Siyan Li, Riley Carlson, and Christopher Potts. 2022. Systematicity in GPT-3's interpretation of novel English noun compounds. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 717–728, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

Preslav Nakov and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *Proceedings of the 12th international conference on Artificial Intelligence: methodology, Systems, and Applications*, pages 233–244.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Marius Paşca. 2015. Interpreting compound noun phrases using web search queries. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 335–344, Denver, Colorado. Association for Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Max Bittker. 2020a. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6509–6515, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020b. Will it unblend? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.

Girishkumar Ponkiya, Rudra Murthy, Pushpak Bhattacharyya, and Girish Palshikar. 2020. Looking inside noun compounds: Unsupervised prepositional and free paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4313–4323, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Vered Shwartz. 2021. A long hard look at MWEs in the age of language models. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, page 1, Online. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211, Melbourne, Australia. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden. Association for Computational Linguistics.

Tim Van de Cruys, Stergos Afantenos, and Philippe Muller. 2013. MELODI: A supervised distributional approach for free paraphrasing of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 144–147, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Edward J Wisniewski. 1997. When concepts combine. *Psychon Bull Rev.*, page 167–183.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2, 3

☑ B1. Did you cite the creators of artifacts you used?
2, 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
8

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
8

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Irrelevant for this type of data (as discussed in Section 8)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2

C ☑ Did you run computational experiments?

2, 3, 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Most of this is irrelevant to our experiments. We mentioned the exact models we used (sections 2, 3). We will add the GPU hours for the camera-ready version.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2, 3

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
2, 3, 8

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

2, 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
2, 3 (in the text - not a full HIT template with all the examples etc. We can include this as an appendix to the camera-ready version if needed)

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
2, 3, 8

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Irrelevant (we didn't collect private information as we discuss in Section 8)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Irrelevant (we didn't collect private information as we discuss in Section 8)

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
2, 3