Systematicity in GPT-3's Interpretation of Novel English Noun Compounds*

Siyan Li
Stanford University
siyanli@stanford.edu

Riley Carlson
Stanford University
rileydc@stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

Abstract

Levin et al. (2019) show experimentally that the interpretations of novel English noun compounds (e.g., *stew skillet*), while not fully compositional, are highly predictable based on whether the modifier and head refer to artifacts or natural kinds. Is the large language model GPT-3 governed by the same interpretive principles? To address this question, we first compare Levin et al.'s experimental data with GPT-3 generations, finding a high degree of similarity. However, this evidence is consistent with GPT-3 reasoning only about specific lexical items rather than the more abstract conceptual categories of Levin et al.'s theory. To probe more deeply, we construct prompts that require the relevant kind of conceptual reasoning. Here, we fail to find convincing evidence that GPT-3 is reasoning about more than just individual lexical items. These results highlight the importance of controlling for low-level distributional regularities when assessing whether a large language model latently encodes a deeper theory.

1 Introduction

English noun compounds (e.g., *birthday cake*, *toy store*) have long been central to debates about the limits of compositionality in natural language semantics. For Partee (1995), they are essentially memorized lexical items whose meanings cannot be predicted from their parts. In contrast, Downing (1977), Wisniewski and Love (1998), and Levi (1978) identify a range of systematic constraints on compound interpretation, suggesting underlying cognitive and linguistic properties that are more complex and productive than memorization would suggest. More recently, Levin et al. (2019) show that, for large classes of noun compounds, the meanings are highly predictable. Their evidence comes from production and comprehension experiments with novel noun compounds (e.g., *stew skillet*, *swamp squash*), for which participant interpretations are highly consistent. These results suggest that, while noun compound meanings are not fully compositional, they are *systematic* in the sense of Fodor and Pylyshyn 1988.

In this paper, we ask whether the large language model (LLM) GPT-3 (Brown et al., 2020) is similarly systematic in its handling of novel compounds. Our comparisons focus on the free-response comprehension experiments of Levin et al. (2019). Levin et al.'s guiding hypothesis is that the observed systematicity traces to two over-arching conceptual hypotheses. The *Event-Related Hypothesis* says that, for compounds referring to artifacts (e.g., *stew skillet*), the modifier will tend to convey information about the creation or use of the compound's referent. The *Essence-Related Hypothesis* says that, for compounds referring to natural kinds (e.g., *swamp squash*), the modifier will tend to identify essential properties of the compound's referent. In other words, deep conceptual properties of the component words are constraining how people interpret novel compounds. Is GPT-3 governed by the same interpretive principles? We report on three experiments seeking to address this question.

In Experiment 1, we find that, when prompted with novel compounds, GPT-3's behavior is strikingly aligned with the interpretations human participants gave in Levin et al.'s experiments. In Experiment 2, we employ the same protocol but with new noun compounds involving a wider range of modifier–head relationships, and we see essentially the same behavior from GPT-3. These results are consistent with the claim that GPT-3 is governed by the Event- and Essence-Related Hypotheses, but it is far from conclusive, since the model may simply be relying on regularities in the interpretations of individual words, rather than reasoning about artifacts or natural kinds as abstract conceptual categories. In Experiment 3, we seek to decouple these two explanations by constructing prompts in

*Experimental materials available at https://github.com/siyan-sylvia-li/systematicity_gpt3/

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 717–728
December 7-11, 2022

which random strings are defined as natural kinds or artifacts and then used in novel compounds. In this setting, GPT-3 is much less successful, suggesting that Experiments 1 and 2 may have confounded statistical distributions of tokens with deeper conceptual understanding. Overall, this highlights the importance of controlling for low-level distributional regularities when assessing whether an LLM latently encodes a deeper conceptual or linguistic theory.

2 Background

2.1 English Noun Compounds

We follow Levin et al. (2019) in focusing on English *endocentric* noun compounds like *soup spoon* and *string bean*. For our purposes, the crucial feature of endocentric compounds is that they entail their head noun: a soup spoon is a spoon, a string bean is a bean, and so forth. By contrast, *exocentric* compounds like *ladyfinger* and *paperback* do not entail either of their component parts (e.g., a ladyfinger is neither a lady nor a finger, but rather a sweet treat). Exocentric compounds may simply be memorized lexical items.

Downing 1977 is a groundbreaking study that uses the production and comprehension of novel compounds to explore the systematicity of noun compounds. Downing proposes that compounds are devices for communicating about objects by identifying their salient features. Inspired by Downing (1977)'s suggestions, Levi (1978) studies recovery of deleted information in dialogue. Specifically, when compounds are created to describe objects, the relationship between the compound head and modifier is often omitted. Levi formalizes nine semantic categories for relationships between compound heads and modifiers. Wisniewski and Love (1998) examine noun–noun compounds describing office supplies versus wildlife, and discover systematic differences in the relationships between compound heads and modifiers for these two types of entities.

2.2 LLMs and Linguistic Creativity

LLMs have been assessed in a range of tasks involving constrained creativity with language.

Malkin et al. (2021) show that GPT-3 can define novel nonsense words in ways that seem plausible to human evaluators. Similarly, Pinter et al. (2020) study blends like *thruple* ('three-person couple'), alongside more transparent cases (e.g., *quiz-maker*) and more opaque ones (*deathbox*, 'dangerous car').

Chakrabarty et al. (2021) find that enhancing GPT-2 with contextual or literal knowledge outperforms few-shot GPT-3 when continuing figurative narratives containing idioms and similes.

Yu et al. (2020) generate homophonic puns using a constraint selection process that rewrites sentences into puns in a semantically naturalistic manner, while Mittal et al. (2022) first generate context words related to different pun word senses through GPT-3, then combine context words with separate word senses to create homographic puns.

Idioms provide an interesting comparison with noun compounds, since we do not expect a high degree of predictability for them (though idioms do show some aspects of compositonality; Nunberg et al. 1994). Socolof et al. (2022) use BERT and XLNet to develop contextual metrics for idiom classification, and Tan and Jiang (2021) probe BERT's and ERNIE2's capacities to distinguish literal and idiomatic uses of a potentially idiomatic expression, and to identify the proper paraphrases of idiomatic expressions in a given context.

Novel noun compounds have the potential to offer important new insights in this area. First, unlike novel blends and idioms, novel compounds are easy to create, and Levin et al.'s findings indicate that we can expect high predictability for them. Second, Levin et al.'s hypotheses engage directly with deeper cognitive notions, rather than being purely about linguistic forms.

3 Experiment 1: Free-Response with Levin et al.'s Novel Compounds

In Levin et al.'s novel compound comprehension study,[1] participants gave free-form textual responses to prompts like "Imagine that you encounter the compound `X`. What would you think this refers to?" Expert labelers then annotated each response for its metarelation (relationship between the compound head and its modifier) and metarelation subtype, using definitions summarized in Table 1. The final experimental dataset consists of 798 participant-created explanations for 38 different novel compounds. 141 explanations were excluded in accordance with the coders' manual.

3.1 Methods

Our experiment essentially treats GPT-3 (specifically, the Instruct-GPT Davinci model) as a new

[1]Data available at `https://osf.io/t43kd/`

Meta-relation	Subtypes
Event	made of, method, purpose, time, used by, object-nom
Essence	borrowed, color, dimension, distinctive part, taste/smell, location, social/political
Other	named after, value, whole-part, other

Table 1: Meta-relations and their corresponding subtypes present in this study. Complete definitions are given in Levin et al. 2019, Appendix A.

participant in this experiment. For each compound, we create three separate prompts:

Natural: `Imagine that you encounter the compound X. What would you think this refers to?`

Structured: `Compound: X.\n\nExplanation:`

Few-shot

`Compound: X1.\n\nExplanation: E1\n\n`
`Compound: X2.\n\nExplanation: E2\n\n`
`Compound: X3.\n\nExplanation: E3\n\n`
`Compound: X.\n\nExplanation:`

In the few-shot prompt, the three examples are randomly chosen from the novel compounds in Levin et al. 2019. Given a randomly selected compound X_i, we then sample, from Levin et al.'s data, a human-generated explanation E_i from the majority metarelation for X_i.

We obtain top-1 samples from the model with temperatures 0.2, 0.4, 0.7 and 0.9, to simulate different response behaviors, from almost deterministic to highly variable. Combining the temperatures and prompts results in 12 conditions for each of the 38 compounds, for a total of 456 generations. Each generation was annotated by each of us (the authors) using the coding framework of Levin et al.. We saw only the compound and the generated text; the prompt and temperature value were hidden. The Fleiss' kappa for these annotations is 0.844 at the level of meta-relations and 0.767 at the level of relation sub-types.

3.2 Results

Figure 1 provides the correlations with Levin et al. (2019)'s human response data for our three prompt types and four temperature settings. The correlations are between metarelation distributions from the human comprehension study and coded GPT-3 generations. We provide more detail on generating the figure in Appendix A. On a high level, Levin et al. (2019) computed the percentages of coded metarelations for different noun compound types

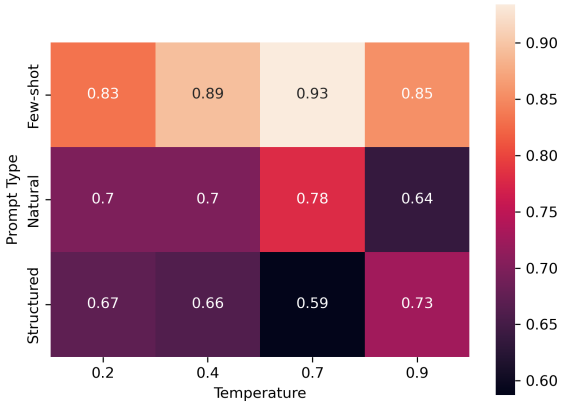

Figure 1: Experiment 1 correlations between model generations' metarelation distributions and Levin et al. (2019) metarelation distributions. Comparatively, temperature 0.7 with few-shot prompting appears optimal.

based on the semantic property of the heads and the modifiers. Specifically, the novel compounds are divided into four categories based on whether a compound component is a natural kind or is artificial. We compute the percentages of metarelations exhibited by GPT-3 generations, and average the Pearson's correlations of each compound category between our percentages and those of Levin et al. (2019). All of the correlations are very high, with the few-shot prompt and 0.7 temperature parameter giving the highest correlation, at 0.93. Overall, these results indicate that GPT-3 is returning definitions whose interpretations align with those given by humans.

Another important dimension is the rate at which definitions need to be excluded because they do not conform to the constraints of endocentric compounds or are otherwise unintelligible. Figure 2 provides the exclusion rates for all our model variants and Levin et al. (2019)'s human response data. GPT-3's exclusion rates are actually much lower than the human rate, suggesting that, in some sense, it may be *more* systematic than humans when confronted with novel compounds.

	Temperature			
	0.2	**0.4**	**0.7**	**0.9**
Natural	13.15	10.52	21.05	28.94
Structured	28.94	36.84	28.94	34.21
Few-shot	0.0	0.0	5.26	2.63
Human	17.67			

Figure 2: Experiment 1 exclusion percentages per prompt/temperature combination. We also report the percentage of human-generated responses that were excluded in Levin et al. (2019).

	Temperature			
	0.2	**0.4**	**0.7**	**0.9**
Natural	5.56	11.43	5.56	5.88
Structured	14.29	17.14	19.44	44.12
Few-shot	5.71	5.71	11.43	8.57

Figure 3: Experiment 2 exclusion percentages per prompt/temperature combination.

4 Experiment 2: Free-Response with New Novel Compounds

We are unable to determine whether Levin et al. (2019)'s experimental data might have been part of GPT-3's training or fine-tuning data. In addition, their set of compounds covers a relatively small part of the full space defined by the coding manual, which might make the problem artificially easy for the model. To address both of these concerns, we created a new set of compounds that keyed into different parts of Table 1, and we repeated our experiment with GPT-3 using this new set.

The additional compounds do not have any lexical overlap with those of Levin et al. 2019. Furthermore, we observed in the previous experiment that the generations for Levin et al. (2019)'s novel compounds are categorized primarily as the "purpose" metarelation subtype. Therefore, we also include compounds that, according to our intuitions, should have more varied metarelation subtypes. We use the same temperatures and prompts combinations.

Table 3 provides the exclusion percentages for this experiment. The rates are overall slightly higher than for Experiment 1, but they are still strikingly low, especially for the few-shot prompt. The Fleiss kappas for the annotations are also very similar to Experiment 1: 0.786 for meta-relations and 0.714 for relation sub-types. Appendix B provides additional details on this experiment.

5 Experiment 3: In-Context Conceptual Reasoning

The results of Experiments 1 and 2 are consistent with GPT-3's behavior being governed by the Event- and Essence-Related Hypotheses at some level, but many other explanations are available. In particular, it could very well be that GPT-3 is reasoning about individual lexical items without truly being able to identify them as artifacts or natural kinds or relate such classifications to its overall generations. To address this confound, we conducted a third experiment.

5.1 Methods

For this experiment, the target compound is created from the nonsense strings `gmtomflxri` and `putrlv` using a prompt that requires implicit inferences about the artifact vs. natural kind status of these terms based on their in-context definitions:

```
A modifier is definition1.
A head is definition2.
A compound is definition3.\n\n
A gmtomflxri is definition4.
A putrlv is definition5.
A gmtomflxri putrlv is
```

We used pure nonsense strings to avoid inadvertently using nonce words that the model might have seen in other contexts. We use four noun compounds for each combination of artifact and natural kind for head and modifier categories. The basis compounds are *kitchen knife*, *tree frog*, *strawberry cookie*, and *coffee bean*. This results in a total of 16 new compounds per basis compound.

For each basis compound, we generate new compounds from the components with specific relationships to the original components. We define these relationships along two axes: (1) Match (M) or Different (D) in noun category, and (2) Close (c) or Far (f) in semantic space. For example, *bedroom* is an Mc instance of *kitchen* because they are both artifacts and similar in concept, and *fur* is a Df instance of *kitchen* because *fur* is a natural kind and is not similar in concept. To decide semantic proximity, we utilize both WordNet (Miller, 1995) for Match instances and GloVe embeddings (Pennington et al., 2014) for Different instances. (See Appendix C.1 for details.)

For a concrete example of our prompting process, assume that we select the basis compound

strawberry cookie and the relationship Mc/Mf. Using the pre-defined list of basis compounds and their variations (Table 6 in Appendix C.3), we select the Mc word of *strawberry*, *banana*, and the Mf word of *cookie*, *table*. Referencing the list of definitions corresponding to each word defined in the appendix, we have the following prompt:

```
A strawberry is sweet fleshy red
fruit.
A cookie is any of various small
flat sweet cakes.
A strawberry cookie is a cookie
made with strawberries.\n\n
A gmtomflxri is a tropical
yellow fruit.
A putrlv is a piece of
furniture.
A gmtomflxri putrlv is
```

The model generations were annotated by us. The Fleiss' kappas are very similar to those of Experiments 1 and 2: 0.821 for the meta-relations and 0.803 for the subtypes.

5.2 Results

The first thing that stands out about these new results is that the exclusion rate is much higher. Table 2 shows the results for the temperature parameter of 0.2, which had the lowest exclusion rates overall. Approximately 30% of the exclusions are inversions of heads and modifiers (e.g., defining *banana table* as a banana instead of as a table). This suggests inconsistent knowledge of this core structural distinction when presented with nonce words. It is unclear if such core structural knowledge is missing in Experiment 1 and 2, since the results may have been masked by lexical regularities. Other instances of exclusion include vague definitions, directly concatenating head and modifier definitions together without proper explanations, and copying the prompt (e.g., *cider nut* is often defined as *coffee bean* because *coffee bean* is a part of the prompt).

For the explanations that were not excluded, we calculate the distributions of metarelation subtype coding per compound type, similar to Levin et al. (2019). The percentages of meta-relations per compound type are presented in Table 3. While there is limited correlation between compound type with the percentage of meta-relations coded as "Events", there does not appear to be as clear a correlation with the original study as in Experiment 1; the Pearson's correlation to the original distribution is only 0.105 (compared to 0.93 in Experiment 1).

Head / *Modifier*	Mc	Mf	Dc	Df	Avg.
Mc	25.0	50.0	0.0	75.0	37.5
Mf	0.0	0.0	0.0	0.0	0.0
Dc	50.0	25.0	75.0	33.3	45.8
Df	75.0	25.0	50.0	50.0	50.0
Avg.	37.5	25	31.25	39.5	

Table 2: Exclusion percentage for every head-modifier combination in Experiment 2, with temperature 0.2.

Type	Event	Percep.	Env.	Other	N
art-art	35.1	8.1	56.8	0	37
art-nk	46.9	0	53.1	0	32
nk-art	73.9	17.4	6.5	2.2	46
nk-nk	33.3	18.0	20.5	28.2	39

Table 3: Coded metarelation distribution per compound type for "Alien" compound experiments. "Art" means "artifact" and "nk" means "natural kind". "Art-nk" indicates the type of compounds with an artifact head and a natural kind modifier.

6 Discussion and Conclusion

This paper explored GPT-3's handling of novel English endocentric noun compounds, building on psycholinguistic research by Levin et al. (2019). Experiments 1 and 2 suggested that GPT-3 might be governed by the same over-arching conceptual and linguistic principles that shape human interpretations of these forms. However, Experiment 3 probed that conclusion more deeply and failed to find evidence for this stronger claim: GPT-3 may instead be memorizing token distributions rather than reasoning about the underlying conceptual categories on which Levin et al. (2019)'s theory depends.

7 Limitations

In Experiments 1 and 2, we assumed that the model already knows the distinction between compound heads and modifiers, yet in Experiment 3 we discovered that head–modifier reversal is a major factor in exclusions. Therefore, the assumption that GPT-3 inherently understands compound head versus modifier may be incorrect. This warrants a

more thorough investigation, to examine GPT-3 generations of flipped versions of compounds (e.g., evaluate both *duck potato* and *potato duck*). Although the switching of the head and modifier was not random enough to suggest that the model has no understanding of noun compounds, this investigation could help determine whether the model inherently understands the relationship between the head and modifier.

Additionally, the statistical distribution of compound components in the training data of GPT-3 may be a confounding variable in our experiment results. However, we were unable to test this theory using statistical analyses due to the model's training data not being publicly available. If the statistical distribution of the compound components in the training data were a confounding variable, then it may be the case that the compounds used did not thoroughly probe the models understanding of noun compounds and that other compounds should be used. Experiment 2 reassures us somewhat, but it is not a substitute for a thorough audit of the training data.

The GPT-3 model annotations were done by the three authors of the paper. While it might seem better to train separate annotators, this is not a simple matter, as its requires deep linguistic expertise. Levin et al. (2019) did much of their coding themselves, presumably for this reason. Importantly, when we did the coding, we were not able to see which model variant produced the generation being evaluated, which allowed for an unbiased comparison across model types even factoring in any annotator biases. In addition, we note that one of our experiments led to a positive result for GPT-3 and the other negative, consistent with us having no particular preferred outcome for the paper's findings. The high Fleiss kappa scores further support the claim that our work simply implemented the coding manual from Levin et al. 2019. All our annotations are included in our public repository: `https://github.com/siyan-sylvia-li/systematicity_gpt3/`.

References

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.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2021. It's not rocket science : Interpreting figurative language in narratives. *CoRR*, abs/2109.00087.

Pamela Downing. 1977. On the creation and use of english compound nouns. *Language*, pages 810–842.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Judith N Levi. 1978. *The syntax and semantics of complex nominals*. Academic Press.

Beth Levin, Lelia Glass, and Dan Jurafsky. 2019. Systematicity in the semantics of noun compounds: The role of artifacts vs. natural kinds. *Linguistics*, 57(3):429–471.

Nikolay Malkin, Sameera Lanka, Pranav Goel, Sudha Rao, and Nebojsa Jojic. 2021. GPT perdetry test: Generating new meanings for new words. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5542–5553, Online. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. Ambipun: Generating humorous puns with ambiguous context.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Barbara H Partee. 1995. Lexical semantics and compositionality. In Lila R. Gleitman and Mark Liberman, editors, *Invitation to Cognitive Science*, volume 1, pages 311–360. MIT Press, Cambridge, MA.

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.

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

Michaela Socolof, Jackie Cheung, Michael Wagner, and Timothy O'Donnell. 2022. Characterizing idioms: Conventionality and contingency. In *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 4024–4037, Dublin, Ireland. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Edward J Wisniewski and Bradley C Love. 1998. Relations versus properties in conceptual combination. *Journal of memory and language*, 38(2):177–202.

Zhiwei Yu, Hongyu Zang, and Xiaojun Wan. 2020. Homophonic pun generation with lexically constrained rewriting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2870–2876, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *CoRR*, abs/1802.01886.

Appendix

A Experiment 1: Additional Details

A.1 Details for Generating the Correlation Heatmap

The correlation heatmap in Figure 1 is created by calculating the average correlation between coded metarelations (not including subtypes) from Levin et al. (2019) and GPT-3 generations. There are four noun compound types, and each of them has different percentages for event, location, and perceptual in both Levin et al. 2019 and our experiments. For a specific set of GPT-3 generations, we first compute the percentage of each metarelation for each of the four noun compound types. Then, for each noun compound type, we compute the Pearson's correlations between the metarelation percentage vector of the generations with the metarelation vector from Levin et al. (2019). Finally, we average the coefficients across the four compound types. This number is then reported as the correlation coefficient for that specific set, in this case a prompt/temperature combination. The distribution from Levin et al. (2019) is in Table 5.

A.2 Examples of Experiment 1 Generations

We present examples of both GPT-3 generations and human-created definitions from Levin et al. (2019):

Compound	GPT-3 Generation
water spatula	A water spatula is a kitchen utensil that is used to help transfer liquids from one container to another. ...
stew skillet	A skillet used to make stews.
duck screen	A duck screen is a type of netting that is used to keep ducks and other waterfowl out of areas where they are not wanted.

Compound	Human Definition
water spatula	A toy with a broad flat edge, used to splash people at the pool.
stew skillet	It is a skillet that you make stew on to make it taste really good.
duck screen	A piece of artificial camouflage used by duck hunters.

Overall, more exclusions happen in higher temperatures, and few-shot prompting drastically decreases the exclusion rate among generations. However, we also observe that few-shot-prompted generations have lower diversity in generated content, even with high temperatures.

A.3 Exclusion Patterns

Examples of excluded model generations are presented in Table 4. Additional analyses of the excluded model generations reveal that GPT-3 makes both human-like and more machine-like mistakes. While some excluded generations are similar to excluded human definitions, such as failing to define the head of the compound or being vague, some exclusion cases are distinctly non-human. For example, the model sometimes defines the head and the modifier separately instead of as one word.

Interestingly, the model seems to recognize the novelty of these compounds. The model often either prefaces the definition with equivalents of "the term `X` is not a common one" or states that "`X` is not a real compound".

One compound that the model often produces excluded generations for is *duck potato*. Specifically, the model attempt to implicitly define *duck* with water-related attributes, stating that duck potatoes are plants that grow in wet areas. Although the model sometimes explains its reasoning (e.g., "the first thing that comes to mind when I see the word

Compound	Excluded Explanation	Reason
panini leaf	A type of sandwich that is usually made with panini bread.	Defining only compound modifier
stick broccoli	The word "stick" is a verb meaning to pierce or to fasten. In this sentence, it is used as a verb meaning to fasten. The word "broccoli" is a noun meaning a type of vegetable.	Defining each part separately
stick broccoli	The term "stick broccoli" is not a real compound, so it is difficult to say what it would refer to.	Not defining the compound
stew chickpea	A stew made with chickpeas.	Defining wrong compound head

Table 4: Example generations that are excluded.

Compound type	Example	Event	Perceptual	Environmental	N
art mod, art head	e.g. stew skillet	93%	7%	0%	N = 195
nk mod, art head	e.g. stream wheel	88%	12%	0%	N = 157
art mod, nk head	e.g. stew chickpea	66%	34%	0%	N = 157
nk mod, nk head	e.g. stream vegetable	15%	34%	51%	N = 148

Table 5: Distribution of metarelations per type of compounds from Levin et al. (2019).

'duck' is the waterfowl. Therefore, I would think that a duck potato would be a potato that is found in the water, perhaps growing on the stem of a water plant."), we exclude these explanations because they do not define a clear enough modifier–head relationship.

A.4 Generation Stability

We are interested in quantifying how different prompt types affect generation diversity. Therefore, we compute the Self-BLEU scores (Zhu et al., 2018) of different generations per prompt type. For prompt type 1, the self-BLEU score is 58.24; for prompt type 2, the score is 52.49; for prompt type 3, however, the score is 77.39. Given that prompt types 2 and 3 differ only in the additional few-shot examples in 3, this suggests that few-shot prompting provides more information, hence higher stability, for this task.

B Experiment 2: Additional Details

Although the novel compounds from Experiment 1 are unlikely to be present in GPT-3's training data, we generated additional novel compounds to try to factor out any possibility that GPT-3's generations are shaped by it having seen Levin et al. (2019)'s materials during training. The additional compounds do not have any lexical overlap with Levin et al. (2019).

Furthermore, we observed in the previous experiment that the generations for Levin et al. (2019)'s novel compounds are categorized primarily as the "purpose" metarelation subtype. Therefore, we also include compounds that, according to our intuitions, should have more varied metarelation subtypes.

Here is a subset of our additional compounds with their intended metarelation subtypes:

Compound	Intended Subtype
Bundt chair / tiger	Borrowed
Biden burrito / flower	Named after
Wonder grass / popsicle	Value
Elephant necktie / eggplant	Used-by
Leaf necklace / stick	Distinctive part
Monster bag / squid	Dimension

We use the same temperature and prompt combinations as Experiment 1.

B.1 Exclusion Rate

Given that no large collection of human-created explanations are available for our additional novel

compounds, we primarily use exclusion rate as the metric. Here, we modify the exclusion criteria from the previous experiment; we no longer exclude examples classified as "other property" by the majority of the raters, to explore the model's behavior in this category of noun compounds. The corresponding exclusion rates for each temperature–prompt combination are presented in Table 3 in the main text.

B.2 Generation Stability

We observe some variation in generations overall, as there is more diversity in compound metarelations. For instance, for the compound *Biden burrito*, the following explanations are produced:

1. The Biden burrito would most likely refer to a burrito that is made in the style of Vice President Joe Biden. This could mean that the burrito is filled with Biden's favorite foods, or that it is made to look like Biden himself.
2. The most likely interpretation of "Biden burrito" would be a burrito named after United States Vice President Joe Biden.
3. Burrito that is eaten by Biden.
4. The Biden burrito is a political term used to describe a 2020 United States presidential election campaign event in which then-Presidential candidate Joe Biden ate a burrito on a live stream.

We can see that the generated explanations span across multiple metarelation subtypes and are relatively diverse for this specific instance due to its inherent ambiguity.

Again, we compute the Self-BLEU scores of generations from each prompt type. Similar to what we observed in exclusion rate, the difference seems to diminish between different prompt types. The scores for Natural, Structured, and Few-shot prompts are 64.28, 54.06, and 65.68 respectively.

C Experiment 3: Additional Details

C.1 Proximity

When using WordNet for Mc, we identify nouns under a sufficiently specific cluster that is in the WordNet tree path to ensure path similarity, and locate the most common noun using unigram frequency.[2] We perform a similar operation for finding Mf instances, except using higher-level clusters to obtain low path-similarity nouns. For GloVe, we identify the most similar words and identify the first ones that differ in noun category as Dc instances, and use the least similar among the top 15,000 most similar words for Df instances. We use two different tools for this process because path similarity is only more informative when the two words have higher overlap in WordNet tree paths, and GloVe embeddings primarily capture semantic proximity without prioritizing noun categories.

[2] https://www.kaggle.com/datasets/rtatman/english-word-frequency

C.2 Exclusion Rates

Similar to prior experiments, we prompt GPT-3 under 4 different temperatures: 0.2, 0.4, 0.7, and 0.9. We report the results from temperature 0.2 in the body of the paper because it is the best-performing. Here we include the exclusion rates of different head-modifier relation combinations from the other temperatures.

Temperature: 0.4				
Head	**Mc**	**Mf**	**Dc**	**Df**
Modifier				
Mc	25.0	50.0	25.0	50.0
Mf	0.0	25.0	25.0	0.0
Dc	50.0	75.0	25.0	33.3
Df	100.0	25.0	75.0	50.0

Temperature: 0.7				
Head	**Mc**	**Mf**	**Dc**	**Df**
Modifier				
Mc	25.0	50.0	50.0	75.0
Mf	0.0	50.0	25.0	50.0
Dc	25.0	50.0	50.0	50.0
Df	50.0	25.0	100.0	50.0

Temperature: 0.9				
Head	**Mc**	**Mf**	**Dc**	**Df**
Modifier				
Mc	25.0	50.0	50.0	75.0
Mf	0.0	25.0	25.0	25.0
Dc	50.0	50.0	50.0	25.0
Df	75.0	25.0	25.0	50.0

C.3 Full Materials

C.3.1 Basis Compound Definitions

We report the list of definitions we use as a part of the basis compound definitions.

Baseline Word	Mc Word	Mf Word	Dc Word	Df Word
kitchen	bedroom	trunk	vegetable	fur
knife	scissors	magazine	metal	shrubbery
tree	bush	tomato	garden	blender
frog	reptile	doe	slipper	satin
strawberry	banana	bamboo	shortcake	overcoat
cookie	pancake	table	egg	oak
coffee	cider	medicine	fruit	kangaroo
bean	nut	flower	chili	gear

Table 6: Each baseline component of the compounds and their corresponding generated words based on matching or different noun category (M/D) and close or far semantic similarity (c/f)

Compound	Definition
kitchen knife	A "kitchen" is a room equipped for preparing meals. A "knife" is edge tool used as a cutting instrument; has a pointed blade with a sharp edge and a handle. A "kitchen knife" is a knife used in cooking .
tree frog	A "tree" is a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown. A "frog" is any of various tailless stout-bodied amphibians with long hind limbs for leaping; semiaquatic and terrestrial species. A "tree frog" is a frog that lives in trees.
strawberry cookie	A "strawberry" is sweet fleshy red fruit. A "cookie" is any of various small flat sweet cakes. A "strawberry cookie" is a cookie made with strawberries.
coffee bean	A "coffee" is a beverage consisting of an infusion of ground coffee beans. A "bean" is any of various edible seeds of plants of the family Leguminosae used for food. A "coffee bean" is a bean used to make coffee.

C.3.2 Definitions of Derived Compounds

The full list of derived compounds from the four basis compounds are in Table 6. To reiterate, these new compounds created only inform us of what definitions should be a part of the prompt, and the specific words are not shown to the model in the prompt.

For these derived compounds, we first identify the corresponding WordNet definition, then truncate and modify these definitions to prevent too much leakage of information so that the model should not be able to pinpoint the exact word from the prompt. The complete list of compound components used in the experiments, including both the original and the modified definitions, is presented below.

Basis word: kitchen

bedroom:

- Original: a room used primarily for sleeping
- Modified: a room with furniture

trunk:

- Original: compartment in an automobile that carries luggage or shopping or tools
- Modified: compartment in an automobile

vegetable:

- Original: edible seeds or roots or stems or leaves or bulbs or tubers or nonsweet fruits of any of numerous herbaceous plant
- Modified: edible plant

fur:

- Original: dense coat of fine silky hairs on mammals
- Modified: part of a mammals body

Basis word: knife

scissors:

- Original: an edge tool having two crossed pivoting blades
- Modified: a tool with blades

magazine:

- Original: a periodic publication containing pictures and stories and articles of interest to those who purchase it or subscribe to it
- Modified: a publication with pictures

metal:

- Original: any of several chemical elements that are usually shiny solids that conduct heat or electricity and can be formed into sheets etc.
- Modified: any shiny solid

shrubbery:

- Original: a collection of shrubs growing together
- Modified: a collection of plants

Basis word: tree

bush:

- Original: a low woody perennial plant usually having several major stems
- Modified: a low woody plant

tomato:

- Original: mildly acid red or yellow pulpy fruit eaten as a vegetable
- Modified: pulpy fruit

garden:

- Original: a yard or lawn adjoining a house
- Modified: an outside area

blender:

- Original: an electrically powered mixer with whirling blades that mix or chop or liquefy foods
- Modified: an appliance

Basis word: frog

reptile:

- Original: any cold-blooded vertebrate of the class Reptilia including tortoises, turtles, snakes, lizards, alligators, crocodiles, and extinct forms
- Modified: cold-blooded vertebrate

doe:

- Original: mature female of mammals of which the male is called 'buck'
- Modified: mature mammals in the forest

slipper:

- Original: low footwear that can be slipped on and off easily
- Modified: low footwear

satin:

- Original: a smooth fabric of silk or rayon
- Modified: a fabric

Basis word: strawberry

banana:

- Original: any of several tropical and subtropical treelike herbs of the genus Musa having a terminal crown of large entire leaves and usually bearing hanging clusters of elongated fruits
- Modified: a tropical yellow fruit

bamboo:

- Original: woody tropical grass having hollow woody stems; mature canes used for construction and furniture
- Modified: woody tropical plant

shortcake:

- Original: very short biscuit dough baked as individual biscuits or a round loaf
- Modified: short biscuit

overcoat:

- Original: a heavy coat worn over clothes in winter
- Modified: a piece of clothing

Basis word: cookie

pancake:

- Original: a flat cake of thin batter fried on both sides on a griddle
- Modified: a flat cake

table:

- Original: a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs
- Modified: a piece of furniture

egg:

- Original: oval reproductive body of a fowl (especially a hen) used as food
- Modified: food from a farm animal

oak:

- Original: a deciduous tree of the genus Quercus
- Modified: a tree

Basis word: coffee

cider:

- Original: a beverage made from juice pressed from apples
- Modified: a beverage made from juice

medicine:

- Original: something that treats or prevents or alleviates the symptoms of disease
- Modified: something that treats sickness

fruit:

- Original: the ripened reproductive body of a seed plant
- Modified: a part of a plant

kangaroo:

- Original: any of several herbivorous leaping marsupials of Australia and New Guinea having large powerful hind legs and a long thick tail
- Modified: a mammal

Basis word: bean

nut:

- Original: usually large hard-shelled seed
- Modified: hard-shelled seed

flower:

- Original: a plant cultivated for its blooms or blossoms
- Modified: a colorful plant

chili:

- Original: ground beef and chili peppers or chili powder often with tomatoes and kidney beans
- Modified: a southern soup

gear:

- Original: a toothed wheel that engages another toothed mechanism in order to change the speed or direction of transmitted motion
- Modified: a wheel