Do LLMs exhibit Nominal Compound Understanding, or just Nominal Understanding?

Stanford CS224N Custom Project

Nathan Chi
Department of Computer Science
Stanford University
nchi1@stanford.edu

Elijah Song
Department of Computer Science
Stanford University
elijahs2@stanford.edu

Abstract

Nominal compounds (compounds consisting of two juxtaposed nouns), are particularly difficult for LLMs to understand. What makes them challenging is that across the spectrum of nominal compounds, there exist such different relationships between the two constituent nouns. Disentangling the relationship between the two constituents is not trivially easy, but certainly can be reasoned through by a native speaker. To this end, we are curious if LLMs are able to achieve the same degree of logical semantic reasoning about these compounds as humans can. We probe the ability of a suite of language models to understand the relationship between constituent nouns in various nominal compounds, finding that **fine-tuned LLMs are able to achieve remarkable success.** Furthermore, we find that the addition of multimodality does not help with this reasoning task, suggesting that visual cues, at least in the current LLM regime, do not aid logico-linguistic reasoning.

1 Key Information to include

- External Mentors: Ryan A. Chi, Ethan A. Chi
- External Collaborators (if you have any): N/A
- Sharing project: No
- Team contributions: The work for this project was fairly split and divided evenly. Nathan Chi played a major role in conducting fine-tuning experiments on BERT and BERT infilling, while Elijah Song focused on fine-tuning GPT models, working on CLIP, and calibrating baseline approaches. Both partners contributed equally to this report.

2 Introduction

Large language models (LLMs) are impressively fluent in a variety of textual domains. However, it is not always clear whether said fluency is simply a *mirage* of understanding — so-called "stochastic parroting" (Bender et al., 2021) — or actual semantic comprehension. For this reason, there is strong collective interest in the NLP research community toward developing benchmarks (Srivastava et al., 2023; Liang et al., 2023) to gauge LLMs' aptitude for *reasoning* tasks, the successful completion of which involves more than merely rote memorization.

It is known in the linguistic literature that identifying the relation between the two nouns in a compound requires more than simply first-order semantic comprehension. Consider the phrase *hymnbook*. It is probably straightforward to a native speaker that *hymnbook* $\rightarrow$ *book THAT HAS hymn*, but it would be amiss to generalize this to a broadly applicable rule. For instance, *library book* $\rightarrow$ *book IN library* (not *book THAT HAS library*).

This variation in meaning is belied by the uniformity of the nominal compound's *surface* structure (Chomsky 1957). On this note, Bauer and Tarasova (2013) describe nominal compounds as exhibiting

"superficial neutralization of semantic relationships." That is, the relationship between two nouns in a nominal compound can be wildly different, even though most nominal compounds basically look the same. In total, the Tratz and Hovy (2011) classification scheme groups relationships between the two nouns in nominal compound bigrams (N_1 and N_2) into 37 mutually exclusive categories and 12 over-arching nominal groups, which are included in 4.

We posit that much like any human speaker would need more than a *passing* understanding of language to semantically unpack these compounds, any computational linguistic system (including an LLM) will need more than *statistical rules* to successfully complete this task. To this end, we implement a set of neural techniques to determine with LLMs are able to achieve human-like levels of understanding. We probe the ability of a broad suite of language models to understand the relationship between component nouns, finding that fine-tuned LLMs are able to achieve reasonable success. Furthermore, we find that the addition of multimodality does not significantly impact model performance — and is no more effective than off-the-shelf infilling.

3 Related Work

In the past, work primarily focused on *compositionality analysis* of noun compounds: or, in other words, to what extent noun compounds can be expressed as a function of their constituent nouns (Reddy et al., 2011; Biemann and Giesbrecht, 2011). This body of early work was supplemented by Jana et al. (2019), which leveraged Poincaré embeddings to represent noun phrases. Yet other work focused around the related task of literality prediction — how closely does the meaning of a given compound match the literal composite of its component nouns? Furthermore, past work noted that infilling templates with pre-trained masked language models performed well for generating paraphrased versions of a nominal compound (Ponkiya et al., 2020). In particular, Ponkiya et al. (2020) find that reframing the task of identifying relationships between nominal compounds as both structured infilling and semi-structured free paraphrasing achieves effective results.

In terms of classifying challenging compound nominalizations, Lee et al. (2022) serves as a major step forward. Building upon their past work in Lee et al. (2021), they present an expanded dataset of annotated noun-modifier compound nominalizations — and their corresponding relationships, which they characterize as three basic categories (`NOUN`, `ADVERB`, and `NIL`). They also propose a preliminary, unsupervised approach to apply graph-based features to (1) classify relationships between nouns and modifiers and (2) select the most accurate paraphrases. Lee et al. (2022) address the twin tasks of paraphrasability prediction (i.e., whether or not a compound normalization is paraphrasable) and paraphrase generation. Their primary approach hinges on graph-based features: they apply the Abstract Meaning Representation (AMR) approach to represent sequences of text as a graph, where nodes represent tokens and edges represent inter-token relationships.

In terms of datasets, there is a marked lack of meaningful datasets of nominal compounds, the largest of which is Tratz and Hovy (2011). Despite being the only large dataset for the subject, Tratz and Hovy (2011) presents several difficulties chiefly owing to its extremely complex and idiosyncratic typology, which does not clearly match subjective opinions (as described in section 5.1). Other datasets include Ponkiya et al. (2018), a dataset of 2,600 examples of nominal compounds classified by Levi (1978) taxonomy. However, this dataset was inaccessible, even after contacting the authors to clarify. Additional datasets in non-English languages also include Wilkens et al. (2017), which is a Portuguese-language nominal compound compilation.

4 Approach and Experimental Details

4.1 Task Definition

We define the **nominal compound disambiguation task** as follows:

Given a noun compound, what category is the relationship between the two nouns?

where $n_{\text{category}} = 10$. The distribution over categories is as follows:

```
purpose          0.238050
objective        0.195595
topical          0.090336
```

```
attribute         0.081797
causal            0.078366
containment       0.078046
loc_part_whole    0.076530
complement        0.067353
owner_emp_use     0.062485
time              0.031442
```

demonstrating that the plurality-class baseline achieves an accuracy of **23.8%**.

Ambiguity In practice, many categories are highly ambiguous, and some entities could fall into multiple categories. (For example, "gold mine" could conceivably fall into *containment* (gold is contained in the mine) or *complement* (gold is mined)). However, because our dataset is only labeled for one category at a time, we frame the problem as a multi-way classification task.

Brief category distributions:

```
‘attribute‘: N1 describes N2, or N2 describes N1.
‘causal‘: N1 caused N2, or N1 performed N2, or N1 is used to perform N2.
‘complement‘: The phrase can be rephrased as N2 of N1.
‘containment‘: N2 contains N1
‘loc_part_whole‘: N1 is the location where N2 is from or occurs.
‘objective‘: N1 is the grammatical object of N2.
‘owner_emp_use‘: N1 is the owner, experiencer, or employer of N2.
‘purpose‘: N1 is the purpose of N2 (N2 performs N1; N2 is used to visit/use/propel/conserve/modif
‘time‘: N1 is the time when N2 occurs.
‘topical‘: N1 is the topic of N2 (N2 is interested in N1; N2 observes N1; N2 depicts N1).
```

4.2 Method: Infilling

A masked language model is inherently a classifier. We exploit the fact that the problem has inherent *structure* – i.e. the category is 'causal' if N_1causesN_2 – and compute score(category) as the masked language modelling score Salazar et al. (2020): $P(\text{word}_{\text{category}}|N_1\,\texttt{[MASK]}\,N_2)$. In particular:

1. **Input**: $n_2\ n_1$
2. **Template**: *In the phrase* $n_2\ n_1$*,* n_2 *[MASK]* n_1. (and reverse, with n_1 [MASK] n_2).
3. We extract normalized softmax probabilities for each of the 10 classifications described in (Tratz and Hovy, 2011).
4. We normalize probabilities as z-scores, then choose the classification and associated order with the highest resulting score.

Note that we use `bert-base-cased` for all experiments.

4.3 Method: prompting

We conducted a zero-shot baseline for both GPT-3.5 (`gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-0125-preview`), using temperature $= 0$ (as we are only interested in the MLE response). We prompt GPT (Brown et al., 2020; Achiam et al., 2023) under a zero-shot environment to generate the most probable baseline predicted relation between n_1 and n_2.

4.3.1 Fine-tuning BERT

We fine-tuned BERT (Devlin et al., 2018) by adding a linear classification head. We trained on the entire test set for 10 epochs with early stopping on an associated validation split provided by the dataset authors.

Model classes We examined tested varying model sizes of BERT and RoBERTa (an optimized BERT model which adopts a larger training dataset and modifies the training process and objective to

improve performance). Upon experimentation, we found that one linear layer was adequate enough for a sufficiently discriminative classification head.

We implemented BERT finetuning in PyTorch, and the model was trained on a single Tesla T4 for 45 minutes per model. Our best-performing BERT model was trained after 6 epochs. We utilize a learning rate of $5e^{-5}$, $\epsilon = 1e^{-8}$, and a batch size of 32 for more efficient training.

4.3.2 Fine-tuning GPT

We fine-tuned `gpt-3.5-turbo-0125` for one epoch on the entire train set, and fine-tuned with the validation set. We implemented fine-tuning GPT with OpenAI's API.

4.3.3 CLIP

We initially hypothesize that evaluating with an additional modality is likely correlated with an increase performance: to this end, we evaluate a zero-shot CLIP model on the `Tratz-coarse` dataset with images scraped on Google Search (as described). The zero-shot CLIP model was conducted using the scraped images from the web and the labels ranked by CLIP consisted of a description of each category as applied to each example in the testing set. The CLIP model returns a image-text similarity score for each of the labels, where a larger score indicates that the image and the text are more similar to each other. Softmax normalization was applied to this output of the CLIP model and the category with the largest softmax probability was selected as the CLIP model's answer.

5 Experiments

5.1 Data

Dataset	Train	Validation	Test
Modified `Tratz-coarse`	12,531	842	3,335
Scraped images	N/A	N/A	3,335

Table 1: Dataset details

We evaluate on a modified version of the Tratz and Hovy (2011) dataset, which contains 19,158 nominal compounds classified into two levels of specificity: 12 distinct coarse-grained relations (`Tratz-coarse`) and 37 distinct fine-grained relations (`Tratz-fine`). Notably, the Tratz and Hovy (2011) is highly subjective — the task itself of classifying noun compounds into a fixed number of semantic relations is quite noisy (Shwartz, 2019; Shwartz and Waterson, 2018). Past analytic work on Tratz and Hovy (2011) suggested that a number of noun compounds fit into more than just one category, with multiple relations in `Tratz-fine` overlapping in meaning. This lack of quality is mainly attributable to the use of crowd-sourced data. Due to the lack of other comparable data, we train and evaluate on Tratz and Hovy (2011). However, to account for this irregularity, we only evaluate on the `Tratz-coarse` dataset, under the assumption that grouping such arbitrarily defined data groups together would result in more realistic performance.

Furthermore, we remove the "cause" and "other" categories due to their relatively small numbers in the train set and the difficulty in defining the relations, reducing the size of our modified `Tratz-course` dataset to 16,708. We follow the data splits from (Shwartz and Waterson, 2018), where the data is split in a 75:20:5 train-test-validation ratio. The dataset was retrieved from the repository for (Shwartz, 2019), available here[1]. The categories we used can be found in the Appendix.

For our multimodal CLIP approach, we utilized Google's Custom Search JSON API (Google) to scrape images corresponding to each nominal compound in our test set. We performed manual review of the images to ensure that the scraped images were relevant to the task at hand, and stored a list of the relevant image URLs with each example in the modified `Tratz-coarse` to form our Multimodal Image Dataset.

[1] `https://github.com/vered1986/NC_embeddings`

5.2 Evaluation method

Given that this is a classification task, we evaluate using accuracy alone. We considered manually annotating the dataset for multiclass classification (as nominal compounds can fall into multiple categories as described above). However, we ultimately abandoned this idea, due to the extensive human labor in re-classifying 16,708 examples.

5.3 Results

Method	Test Accuracy	Dataset
BERT Infilling	0.1363	Modified `Tratz-coarse`
Plurality Baseline	0.2381	Modified `Tratz-coarse`
GPT 3.5 Baseline	0.2504	Modified `Tratz-coarse`
GPT 4 Few-Shot Completions Baseline	0.3748	Modified `Tratz-coarse`
GPT 3.5-Turbo Few-Shot Completions	0.2834	Modified `Tratz-coarse`
GPT 3.5-Turbo Fine-Tuned	0.7010	Modified `Tratz-coarse`
BERT Fine-Tuned	0.8824	Modified `Tratz-coarse`
BERT Fine-Tuned	0.8824	Modified `Tratz-coarse`
CLIP	0.1364	Modified `Tratz-coarse` + scraped images

Table 2: All accuracies reported test set.

We find that fine-tuning LLMs achieve significantly better performance than our baseline approaches on nominal compound classification. Fine-tuning BERT was particularly effective at this task, achieving a maximum performance of **88.24%** on the test set (an absolute increase over the GPT-4 baseline of approximately 51%), which was significantly better than we expected. The BERT model's strong performance is likely due to its bidirectional encoding nature, which enables a better understanding of context — especially in this task, when both nouns are arguably equally important.

Notably, fine-tuning BERT performed better than fining-tuning GPT 3.5 on all labels in the dataset – often by large amounts. For example, fine-tuned BERT predicted nominal compounds labeled "Attribute" correctly 82.6% of the time, while fine-tuned GPT 3.5 only classified 33.7% of these compounds correctly. Importantly, our BERT architecture also relied directly on training a classifier on top of BERT's hidden state representations — this direct use of a classification head may have meant that the model was more adapted to the downstream task than GPT. Furthermore, we postulate that classifying on numerical labels (class labels) as in BERT effectively blinded the categories, which may have reduced any unintentional bias associated with the class labels as in our GPT approach.

As expected, few-shot learning for GPT 3.5 only performed marginally better than the baseline zero-shot approach, with test accuracies of 28.34% and 25.04%, respectively. This is likely due to the fact that providing GPT with a few examples of nominal compound classification is unlikely to enable the model to reason more strongly about these complex topics without an excessively large number of examples, which would render the cost of experimentation prohbitively expensive.

We find that BERT infilling performs surprisingly poorly, despite prior work (Ponkiya et al., 2020) suggesting otherwise. To this end, we hypothesize that the classification schema of the Tratz taxonomy is not conducive toward multimodal classification — that is, rather than expressing nominal compounds ("celebrity chef") as simple predicate-noun combinations ("chef IS celebrity"), the Tratz taxonomy would classify it under a subjective and loosely defined broad category (attributive).

Ultimately, CLIP performed extremely poorly with a test accuracy of 13.64%, well below expectations and below even the plurality baseline of 23.81%.

We hypothesize that the lack of improved results is likely due to the relative difficulty of expressing complex concepts (particularly abstract ones) in images, making the calculating of image-text similarity scores quite difficult. Images tend to represent simple, concrete nouns (take for example, "apple") but may not be able to effectively represent the ambiguous meaning of many nominal compounds. The lack of ability for multimodal and image-based methods to analyze nominal compounds is not surprising, however, as the human analysis of nominal compounds is one expressed

Accuracy per label, BERT and GPT

label	accuracy_bert	accuracy_gpt	count
causal	0.775	0.502	249
loc_part_whole	0.806	0.565	283
attribute	0.826	0.337	264
owner_emp_use	0.829	0.649	222
containment	0.885	0.682	217
time	0.888	0.854	89
complement	0.899	0.518	228
purpose	0.902	0.843	794
topical	0.932	0.767	322
objective	0.940	0.843	667

Figure 1: Accuracy per Label

BERT confusion matrix

true_label \ predicted_label	attribute	causal	complement	containment	loc_part_whole	objective	owner_emp_use	purpose	time	topical
attribute	218	6	6	5	6	5	0	13	3	2
causal	4	193	2	6	4	23	2	8	0	7
complement	2	2	205	2	3	1	5	4	2	2
containment	2	0	3	192	2	0	3	13	0	2
loc_part_whole	7	8	3	7	228	10	11	5	2	2
objective	4	13	1	1	2	627	2	10	1	6
owner_emp_use	1	16	1	2	5	4	184	7	1	1
purpose	10	6	6	12	16	10	7	716	3	8
time	2	0	0	0	0	3	0	1	79	4
topical	4	3	2	2	3	2	2	4	0	300

Figure 2: A confusion matrix from our fine-tuned BERT experiment, demonstrating that the model often misclassifies a Causal relationship for Ownership, Employement, & Use and Objective relationships.

mainly in text, not image. Perhaps a series of multiple images could be better represent a nominal compound in a multimodal model, capturing greater semantic meaning.

Interestingly, CLIP performs at almost exactly the same accuracy level as BERT infilling, further suggesting that including the image modality does not aid linguistic reasoning under the Tratz taxonomy.

6 Analysis

In 2, we provide a confusion matrix of BERT nominal compound classifications. We see that BERT, our best-performing model, tends to misclassify a Causal relation as a `Objective` or `Ownership, Employment or Use and Objective` one. We posit this is due to the loosely defined nature of the Causal relationship —- and Tratz's taxonomic classifications, in general. There are, for example, many objects that are likely to be both caused and possessed. This, paired with the fact that Tratz only includes one category per nominal compound (despite multiple possibilities) which would result in multiple "mis-classifications," at least quantitatively. Qualitatively, we find that a number of these mixups are quite natural to a reader: for instance, "makeup artist" is classified as "causal" according to Tratz, rather than "Objective" or "Employment." To a human reader, this seems objectively wrong — and indeed, the BERT model falls to the same failure case. This failure can perhaps be seen as a weakness of the Tratz dataset — more so than our architecture.

Category	Image-Text Similarity Score
Objective	30.5502
Causal	32.2299
Purpose	31.2561
Ownership, Employment, or Use	31.4049
Time	32.7364
Location and Whole+Part	31.7102
Composition and Containment	32.3126
Topical	31.7183
Complement	31.7183
Attributive and Equative	31.5940

Table 3: CLIP cosine similarity scores by category, for “living standard", bounded by 1 - 100.

Figure 3: An image that we scraped to compile our dataset, representing the abstract nominal compound “living standard.”

The CLIP model tends to perform poorly with nominal compounds that represent abstract concepts. For example, the abstract nominal compound “living standard” is classified by CLIP as Causal (implying that “living” causes the “standard” in “living standard”) instead of as Complement, wherein “living” describes the particular nature of the “standard” in “living standard.” This is likely due to the fact that “living standard” is difficult to visualize as an image – indeed, the image-text similarity score output of CLIP on this image is nearly identical for all of the labels, and the model’s selection of “Causal” can be attributed to noise instead of reasoning.

In contrast, the CLIP model is able (at least to some extent) reason about concrete nominal compounds. For example, the concrete nominal compound “police bus” is correctly classified by CLIP as Ownership, Employment, and Use, where in the “police” are the users of the “bus.” The image-based CLIP model works well on concrete inputs as opposed to abstract inputs, as expected, testifying to a limitation of a multimodal approach when dealing with complex reasoning.

Figure 4: An image that we scraped to compile our dataset, representing the concrete nominal compound “police bus.”

7 Conclusion

We find that fine-tuned BERT achieves near-human results, with 88.24% accuracy, and conclude that multimodal modeling does not present an appreciable advantage over pure language model. This is likely due their struggles with reasoning about relations between two abstract nouns, which is far less compositional (and thus straightforward). Some limitations with this work include the use of the Tratz dataset, which is inherently limited due to its poor quality.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Laurie Bauer and Elizaveta Tarasova. 2013. The meaning link in nominal compounds. *SKASE Journal of Theoretical Linguistics*, 10(3).

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality*, pages 21–28.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Google. Using rest to invoke the api | programmable search engine. `https://developers.google.com/custom-search/v1/using_rest`. [Accessed 02-03-2024].

Abhik Jana, Dmitry Puzyrev, Alexander Panchenko, Pawan Goyal, Chris Biemann, and Animesh Mukherjee. 2019. On the compositionality prediction of noun phrases using poincar\'e embeddings. *arXiv preprint arXiv:1906.03007*.

John Sie Yuen Lee, Ho Hung Lim, and Carol Webster. 2022. Unsupervised paraphrasability prediction for compound nominalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3254–3263.

John SY Lee, Ho Hung Lim, and Carol Webster. 2021. Paraphrasing compound nominalizations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8023–8028.

Judith N Levi. 1978. The syntax and semantics of complex nominals. *(No Title)*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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.

Girishkumar Ponkiya, Kevin Patel, Pushpak Bhattacharyya, and Girish Palshikar. 2018. Towards a standardized dataset for noun compound interpretation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th international joint conference on natural language processing*, pages 210–218.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Vered Shwartz. 2019. A systematic comparison of english noun compound representations.

Vered Shwartz and Chris Waterson. 2018. Olive oil is made *of* olives, baby oil is made *for* babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 218–224, New Orleans, Louisiana. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari,

Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Stephen Tratz and Eduard Hovy. 2011. A fast, accurate, non-projective, semantically-enriched parser. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1268.

Rodrigo Wilkens, Leonardo Zilio, Silvio Cordeiro, Felipe SF Paula, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2017. Lexsubnc: A dataset of lexical substitution for nominal compounds. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)-Short papers*.

A Appendix

Category	Meaning	Example
Objective	N_1 is the logical grammatical object of N_2	biotechnology research
Causal	N_1 engaged in, provided, caused, justified, is a means in the process of, or is used by N_2	government figure
Purpose	N_2 performs, engages in, creates, obtains, mitigates, opposes, organizes, or has a purpose related to N_1	labor market
Ownership, Employment, or Use	N_1 owns, experiences, uses, or receives from N_2 *or* N_2 works for N_1	government technocrat
Time	N_2 exists, occurs during, or is created during N_1	winter holiday
Location and Whole+Part	N_1 is the location where N_2 is at *or* N_2 is a part, piece, or member of N_1	water spider
Composition and Containment	N_1 composes, constitutes, or is contained in N_2 *or* N_2 specifies the amount of N_1	stock portfolio
Topical	N_2 discusses, depicts, teaches, or contains info related to N_1 *or* N_1 is the topic of N_2	property deal
Complement	N_1 describes the nature or quality of N_2	earth tone
Attributive and Equative	N_1 is or is an instance of N_2, *or* N_1 is an adjective-like noun	core tradition

Table 4: Description of each category in our dataset according to Tratz and Hovy (2011), with an example of each.