# Can Large Language Models Interpret Noun-Noun Compounds? A Linguistically-Motivated Study on Lexicalized and Novel Compounds

**Anonymous ACL submission**


## Abstract

Noun-noun compounds interpretation is the task where a model is given one of such constructions, and it is asked to provide a paraphrase, making the semantic relation between the nouns explicit, as in *carrot cake* is "a cake *made of* carrots." Such a task requires the ability to understand the implicit structured representation of the compound meaning.

In this paper, we test to what extent the recent Large Language Models can interpret the semantic relation between the constituents of lexicalized English compounds and whether they can abstract from such semantic knowledge to predict the semantic relation between the constituents of similar but novel compounds (e.g., *carrot dessert*). We test both Surprisal metrics and prompt-based methods to see whether i.) they can correctly predict the relation between constituents, and ii.) the semantic representation of the relation is robust to paraphrasing.

Using a dataset of lexicalized and annotated noun-noun compounds, we find that LLMs can infer some semantic relations better than others (with a preference for compounds involving concrete concepts). When challenged to perform abstractions and transfer their interpretations to semantically similar but novel compounds, LLMs show serious limitations.


## 1 Introduction

Noun-noun compounds represent an important challenge for all the applications related to Natural Language Understanding, given the implicit semantic relation assumed between the two components, namely: head and modifier (Nakov, 2008b). Their correct interpretation is an essential step for several Natural Language Processing applications such as question answering, machine translation, and information extraction. For example, if a question answering system is asked something about *birthday cake*, it must understand that the user is talking about a cake made *for birthdays;* while if it is asked about a *carrot cake,* it must understand that the query refers to a cake made *with carrots* (not *for* carrots). The capacity to grasp the semantic connection underlying the pairing of two terms in a compound represents a form of abstraction inherent to human cognition, applicable to concrete and abstract concepts alike (concrete such as *carrot cakes* and abstract such as *bank loans*). This skill is often wielded even for never-encountered-before compounds (Van Jaarsveld and Rattink, 1988).

Previous research stressed the role of structured world knowledge in the interpretation of compounds (Wisniewski and Love, 1998; Ó Séaghdha, 2008), which includes the knowledge of the constituent entities and their potential relations. Moreover, people are able to interpret novel compounds by abstracting from knowledge based on past experiences with similar conceptual combinations (Gagné and Spalding, 2006b; Gagné and Shoben, 1997, 2002, among others) and to extend them by relying on analogical comparisons (Krott, 2009). Can the modern Large Language Models (LLMs) do the same?

The main goal of our study is to propose a more refined methodology to understand when and how LLMs are capable of performing abstractions that humans routinely do, namely: understanding the semantic relation existing between the two components of a lexicalized compound and then extending such relation to novel compounds that are constructed in such a way to maintain the semantics of the original components. To do so, we manually manipulated existing compounds by replacing one of the two terms (head or modifier) with their hypernym, namely a word denoting a superordinate concept (Cruse, 1986). This allowed us to generate novel compounds such as *birthday dessert* and *event cake*, based on the lexicalized *birthday cake*. To test the LLM's ability to understand the seman-

| compound | coarse-grained (Tratz, 2011) | fine-grained (Tratz, 2011) | Hatcher-Bourque (Pepper, 2022) | paraphrase (Pepper, 2021) |
|---|---|---|---|---|
| *plastic bag* | containment | SUBSTANCE -MATERIAL-INGREDIENT | COMPOSITION-R | a bag that is composed of plastic |
| *trash bag* | containment | CONTAIN | CONTAINMENT-R | a bag that contains trash |
| *supermarket shelf* | loc_part_whole | LOCATION | LOCATION | a shelf that is located in a supermarket |
| *car door* | loc_part_whole | WHOLE+ PART_OR _MEMBER_OF | PARTONOMY | a door that is part of a car |
| *food company* | purpose | CREATE-PROVIDE-GENERATE-SELL | PRODUCTION | a company that produces food |
| *bank loan* | causal | CREATOR-PROVIDER-CAUSE_OF | PRODUCTION-R | a loan that a bank produces |
| *research group* | purpose | PERFORM& ENGAGE_IN | PURPOSE | a group intended for research |
| *art class* | topical | TOPIC | TOPIC-R | a class that is about art |

Table 1: Semantic relations of Tratz (2011) and their mapping onto the Hatcher-Borque classification.

tics of lexicalized and novel compounds, we assess whether Surprisal, a metric directly based on the log probabilities of the LLMs, is able to differentiate between the possible interpretations of a compound. We hypothesize that LLMs may be accurate in recognizing the correct semantic relation holding between the two components of a lexicalized compound. Moreover, if we were to observe any differences in the performance across different types of compounds, we would argue that such differences may be (at least partially) explained by the concreteness of the compound, in line with previous psychological findings showing that concrete concepts are processed more easily than abstract ones (Jessen et al., 2000). As a complement to Surprisal analyses, we performed a metalinguistic prompt asking to identify the correct interpretation of a compound from a list of options. We relied on LLMs trained with Instruction tuning, a method that has recently been proposed to enhance the generalization capability of LLMs, and assessed the performance of some of the most popular architectures on this task. Our contributions can be summarized as follows:

1. To the best of our knowledge, we are the first to investigate compound interpretation with the most recent LLMs, including instruction-tuned variants;

2. We introduce a dataset designed to manipulate compounds at several levels of linguistic information and present a methodology to generate novel compounds that could be helpful for future investigations.

## 2 Related Work

The problem of the interpretation of compounds has generally been addressed via two different tasks: the first one is the classification in a limited inventory of ontological/semantic relations holding between the two nouns (Nastase and Szpakowicz, 2003), and the second one is the free generation of a paraphrase describing the same relations (Hendrickx et al., 2013; Shwartz and Waterson, 2018; Shwartz and Dagan, 2019). With the introduction of Transformer-based language models, several studies have proposed to investigate their internal representations to understand how the constituent meanings are composed (e.g. Ormerod et al. (2023); Miletić and Schulte im Walde (2023); Buijtelaar and Pezzelle (2023)), and if and to what extent they are able to generalize to interpret unseen compounds (Li et al., 2022).

Coil and Shwartz (2023) proposed a few-shot model based on GPT-3 to tackle interpretation, and they were able to achieve almost perfect performance on a SemEval noun compounds benchmark by Hendrickx et al. (2013). However, by measuring the ngram overlap between the generated paraphrases and the C4 corpus, they found that GPT-3 might just be parroting word sequences seen in the training data, and the strategy turned out to be less effective with rare or novel compounds.

Is the knowledge encoded in recent LLMs, including instruction-tuned ones, sufficient to interpret the relation between constituent nouns and to generalize the interpretations to novel compounds? Language models retain a non-trivial amount of knowledge about the world, and this is reflected in the log probability scores that they assign to real world situations and events described by natural language sentences (Pedinotti et al., 2021; Kauf et al., 2023); moreover, the recent progress on instruction tuning have led to even better alignment with conceptual representations in the human brain (Aw et al., 2023). Therefore, our investigation will focus on three of the most popular LLMs (Llama-2, Falcon and Mistral), both in their Base and in their Instruct version, to see if instruction tuning leads to performance improvements also in the interpretation of compounds.

## 3 Do LLMs Grasp Semantic Relations in Lexicalized Noun Compounds?

### 3.1 Data

For our experiment, we selected compounds from two previously released datasets. Tratz (2011) gathered in his dataset around 19K compositional noun compounds human-annotated with a semantic relation (37 fine-grained relations, 12 coarse-grained relations). Conversely, Muraki et al. (2023) collected concreteness ratings for over 60K multiword expressions from 2,825 online participants. Expressions were rated from 1 to 5, where 1 indicates that the expression was very abstract and 5 that the expression was very concrete. In order to use the concreteness ratings collected by Muraki as a predictor for the accuracy of the LLMs in identifying the correct semantic relation existing between head and modifier, we retained only the compounds from Tratz associated with concreteness ratings in Muraki. The intersection of the two datasets resulted in 2,268 noun-noun compounds annotated with word and bigram frequency (extracted from enTenTen20 corpus), concreteness score, semantic relation class, and the semantic type of the compound (provided by three annotators who followed the coding scheme of Villani et al., 2024). We believe that the more linguistic features are added to a compound, the more we can glimpse which factors influence LLMs' plausibility of noun compounds.

Additionally, we associated a paraphrase created for each compound for the following reason. Using abstract semantic categories to describe compounds is considered problematic because i.) it is unclear which relation inventory is the best one, ii) such relations capture only part of the semantics (e.g., classifying *malaria mosquito* as CAUSE obscures the fact that mosquitos do not directly cause malaria, but just transmit it), and iii.) multiple relations are possible (Nakov, 2008a). Therefore, common compound datasets used in NLP typically provide linguistic paraphrases of compounds produced by human annotators. However, if multiple paraphrases are reported for each compound, this causes an exponential generation of similar paraphrases in the data; for instance, *golf course* can be "course for golf," "course for playing golf," "course for the game of golf," etc. (from Hendrickx et al. (2013)).

We decided to follow a different approach to reduce the variability of paraphrases. We converted Tratz's relations into the Hatcher-Bourque classification (Pepper, 2022), a classification of semantic relations suitable for typologically different languages. The classification comprises 17 low-level relations, and some of them can be reversible (the first word of the compound, usually the modifier, is the semantic head). These relations are grouped according to the three high-level relations (similarity, containment, and direction). We chose this classification not just because it was conceived to be cross-linguistically consistent but also because Pepper (2021) proposed an Excel-based tool for the computer-assisted analysis of semantic relations called the "Bourquifier". For instance, the relation USAGE, which expresses the relation between something that is "used" and the entity ("user") that uses it, can be translated as "an H that an M uses" (e.g., a *lamp oil* is "(an) oil <u>that</u> a lamp <u>uses</u>"). Conversely, *animal doctor* is annotated with the semantic class PURPOSE and expresses the relation between an entity and its purpose, and it is paraphrased as "a doctor <u>intended for</u> animals." We used the Bourquifier as a template to create compound paraphrases; as a result, compounds classified under the same semantic relation have a similar paraphrase.

For the present study, we selected only compounds with a clear map between Tratz and Hatcher-Bourque classifications from the overall dataset, disregarding ambiguous compounds or odd paraphrases. The final subset consists of 668 lexicalized (and compositional) noun-noun compounds (henceforth, **LNC**) and contains compounds for nine semantic relations. Table 1 illustrates the final relations together with the associated paraphrase.

In addition, we used the dataset of Nakov (2008b), which contains 250 compounds anno-

| Relation | Count | Mean Conc |
|---|---|---|
| COMP-R | 85 | 4.47 |
| CONT-R | 54 | 4.49 |
| LOCATION | 107 | 4.15 |
| PARTONOMY | 16 | 4.58 |
| PROD-R | 13 | 3.18 |
| PRODUCTION | 47 | 4.34 |
| PURPOSE | 270 | 4.01 |
| TOPIC-R | 66 | 3.30 |
| USG-R | 10 | 4.24 |

Table 2: Statistics of frequency and mean concreteness ratings for the LNC dataset.

| Relation | Verb | Count |
|---|---|---|
| ABOUT | *involve* | 18 |
| BE | *be* | 42 |
| CAUSE1 | *cause* | 8 |
| CAUSE2 | *be caused by* | 17 |
| FOR | *contain* | 16 |
| FROM | *come from* | 22 |
| HAVE1 | *contain* | 14 |
| HAVE2 | *come from* | 14 |
| IN | *occur in* | 22 |
| MAKE1 | *make* | 4 |
| MAKE2 | *be made of* | 20 |
| NOMIN:ACT | *be made by* | 15 |
| NOMIN:AGENT | *give* | 6 |
| NOMIN:PATIENT | *work for* | 5 |
| NOMIN:PRODUCT | *be made by* | 11 |
| USE | *use* | 16 |

Table 3: Descriptive statistics for Nakov dataset. We report the most frequent verbal expression associated with each of the 16 semantic relations.

tated with 16 semantic classes (coming from the classification by Levi (1978)) and human-proposed paraphrasing verbs (see Table 3). For our purposes, we selected the most frequently produced verb expressing the correct underlying relation for each compound and created a short sentence. For example, *beacon grease* becomes "(a) grease that comes from (a) bacon." This dataset serves as a diagnostic test for the evaluation of our dataset. Specifically, we assess whether LLMs show higher performance when asked to recognize paraphrases that are generated spontaneously by humans instead of those generated from the Bourquifier templates.

### 3.2 Methods

**Models** We evaluated three open-source LLMs and their instruction-tuned variant: Llama-2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), and Mistral (Jiang et al., 2023). All models are open-source, pre-trained generative text models with 7 billion parameters. As a baseline, we selected BERT-large-uncased (Devlin et al., 2019), a bi-directional masked language model, and GPT2-xl (Radford et al., 2019).[1]

**Tasks** The aim of this study is to evaluate whether LLMs are able to correctly identify the semantic relation underlying noun-noun compounds. We propose not to make the model generate the correct paraphrase but to pick the correct one from a list of possible paraphrases. From the LNC dataset, we used the Bourquifier templates to make implausible paraphrases of the compound. For the Nakov dataset, we selected the most frequent verbal phrase associated with each relation (Table 3) and used it to create the distractors.

We designed two complementary tasks to evaluate the ability to interpret compounds: i.) direct probability measures and ii.) metalinguistic prompting. In the first task, we compute the *Surprisal* at the sentence level. The Surprisal $S_t$ of the single token $t_i$ is defined as the negative of the log probability of $t_i$, conditioned on the preceding sentence tokens $w_{<i}$. The Surprisal of the overall sentence ($S_s$) is then defined as the sum of the Surprisals of each token ($S_t$), normalized by the length of the sentence:

$$S_s = \frac{\sum_{t \in S}^{T} S_t}{count(t)} \tag{1}$$

For BERT, a bidirectional masked language model, the Surprisal of sentences was computed using a modified version of the metric by Kauf and Ivanova (2023). In short, each sentence token is successively masked, the Surprisal score is retrieved by using the sentence context in a masked language modeling setting, and then the partial scores finally get summed; additionally, for out-of-vocabulary words, all the tokens within the word also get masked, and not just the target one (this helps to avoid the probability overestimation of rare words). The Surprisal scores were extracted using the `minicons` library v. 0.2.33 (Misra, 2022).

Our assumption is that the correct paraphrase of a given compound ($good_{NC}$) should have a lower Surprisal score than the scores of all incorrect al-

[1] We only focus on open LLMs i.) for reproducibility reasons, and ii.) because we are interested in comparing the Base and the Instruct version of the very same models.

| | | baselines | | LLMs (Base) | | | LLMs (Instruct) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BERT-large | GPT2-xl | Llama-2 | Falcon | Mistral | Llama-2 | Falcon | Mistral |
| LNC | Acc | 0.262 | 0.338 | 0.401 | 0.433 | 0.403 | **0.448** | 0.38 | 0.428 |
| | MRR | 0.509 | 0.542 | 0.583 | 0.595 | 0.569 | **0.599** | 0.557 | 0.592 |
| Nakov | Acc | 0.484 | 0.548 | 0.592 | 0.568 | 0.6 | 0.632 | 0.56 | **0.648** |
| | MRR | 0.641 | 0.682 | 0.722 | 0.707 | 0.73 | 0.746 | 0.698 | **0.756** |

Table 4: Surprisal results on the LNC and Nakov datasets.

ternatives ($bad_{NC}$).

$$\forall s \in bad_{NC}, S(good_{NC}) < S(s) \qquad (2)$$

As a more natural way of evaluating the performance of instruct-tuned models, we decided to prompt them to select the best paraphrases for a given compound. Specifically, the model is asked to choose a correct paraphrase from a list of expressions (full example in Appendix A):

```
Which is the most likely description
of "olive oil"?
1. an oil that uses olives;
...
9. an oil that is composed of olives
```

We ran three different versions of prompting strategies: zero-shot (no examples of the task are provided), one-shot, and three-shot learning (one and three examples are provided, respectively). Since we observed inconsistent output from the zero-shot prompting, we just reported results for the other two settings. For this task, we selected only the instruction-tuned variants of Llama-2, Falcon, and Mistral and used the same hyperparameters for all models[2]. All experiments were run on Colab TPU and A100.

### 3.3 Results

**Surprisals** Table 4 reports LLMs' performance over the two datasets. We computed two different performance metrics: i.) *Accuracy*, the proportion of compounds where the model assigns the lowest Surprisal to the correct paraphrase, and ii.) *Mean Reciprocal Rank* (MRR). For this metric, we ranked the paraphrases in terms of their Surprisal (from the smallest values to the largest ones) and computed the multiplicative inverse of the rank of the correct answer (1 if it is in the first place, 0.5 for the second, and so on). The overall Accuracy of recent LLMs is higher than the two baselines (BERT: 26,2%; GPT2: 33,8%), with BERT performing poorly. The MRR scores align with Accuracy. Instruction-tuned variants are not consistently better than their pre-trained variants: Llama-2 Instruct reaches a statistical significance of the improvement over the Base model, but the opposite trend is observed for Falcon, whose instruction-tuned version performs statistically worse than its Base counterpart. Finally, Mistral's improvement of the Instruct model over the Base one does not reach statistical significance. Considering the instruction-tuned models, Llama-2 gains the highest performance (44,8%), but there is no statistical difference with Mistral (42,2%), while both models are statistically better than Falcon (38%)[3]. Overall, 200 compounds are always correctly categorized by Llama-2, Falcon, and Mistral.

To further gather an idea of which semantic relations are commonly mistaken by all models and to identify similar patterns across their Surprisal distributions, we also computed each class's accuracy. In this case, per-class Accuracy is considered as the proportion of compounds where the model assigns the lowest Surprisal to the paraphrase of the correct class over the total compounds annotated with that semantic relation/class in the gold standard. By looking at Figure 1, the analysis by category reveals an interesting trend across LLMs: some semantic relations have higher Accuracy (COMP-R and PRODUCTION are almost perfect), whilst others are commonly mistaken (PURPOSE, PROD-R, and TOPIC-R). It is worth noticing that the semantic relations that are less understood are also the ones referring to less concrete referents (the average of concreteness ratings is 3.18 for PROD-R and 3.30 for TOPIC-R, cf. Table 2). A binomial generalized linear mixed model demonstrates that there is a positive, significant ef-

[2] `Temperature:0, do_sample:False, top-k:10, top-p:5, max-tokens:50, frequency and presence penalty:0.`

[3] We determine the significance of differences between model accuracies with McNemar's Chi-Square Test, applied to a 2x2 contingency matrix containing the number of correct and incorrect answers. Statistical significance is reached when $p$-value $< 0.01$.

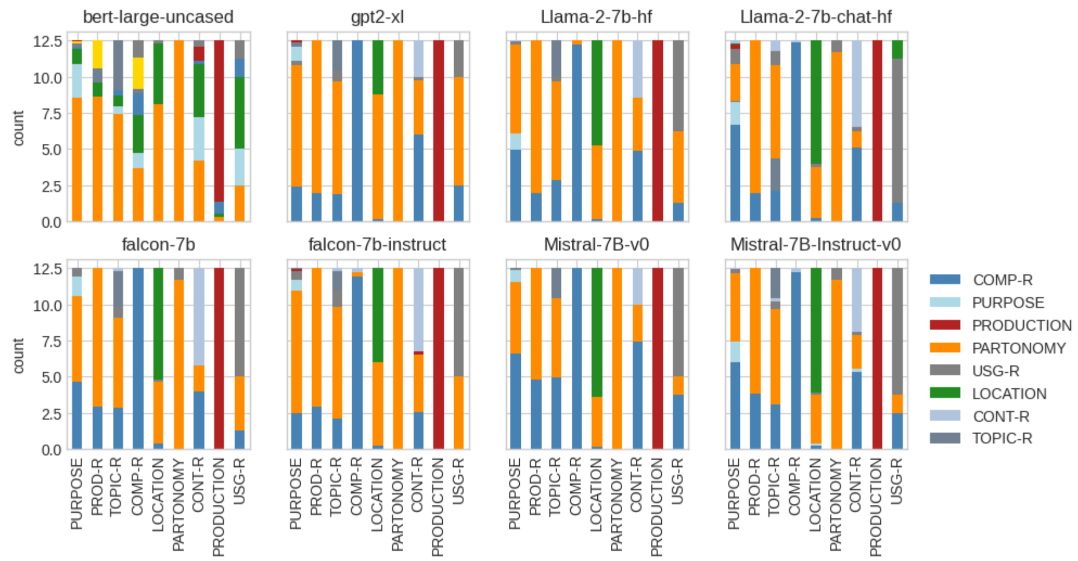


Figure 1: LNC dataset: Distribution of semantic relations with the lowest Surprisal scores for each relation.

fect between Accuracy (dependent variable) and concreteness (independent variable) (coefficient= 0.703, SE=0.133, $p<0.001$), showing that accuracy increases with concreteness (AIC: 894.7 BIC: 908.2).

The evaluation of the Nakov dataset gives similar results (cf. Appendix B): Instruction-tuned LLMs have higher performance (Llama-2: 63,2%, Mistral: 64,8%). In this case, the compounds that are accurately recognized are from the semantic classes of FROM (between 86-94% of accuracy), CAUSE2 (between 84-94% of accuracy), MAKE2 (between 80-92% of accuracy), and NOMINALIZATION_PATIENT (but this group consists of 5 compounds only).While accuracy scores are higher than those computed for our LNC dataset, this outcome does not demonstrate that a more naturalistic input changes the Surprisal distributions.

**Prompting** The results of the prompting experiment are in line with Surprisal scores. As reported in Table 5, Mistral obtains the highest values, reaching 59% of Accuracy in the 1-shot setting. It is interesting to notice that adding examples to the prompt negatively affects the models' answers. Considering the best variant, PRODUCTION is almost always identified correctly (96%), but its counterpart PROD-R is hardly chosen (15%).

The evaluation of Nakov compounds (Table 6) is in line with the LNC dataset, and Mistral performs very well in both settings (one-shot:80%, three-shot:75%). Overall, the best model is more confused with the ABOUT relation (just 61% of accuracy). Finally, the models sometimes tend to justify their choice, giving us an idea of what their interpretation is. Interestingly, they do not hallucinate but answer coherently even when they fail to select the preferred option. This qualitative analysis of the answers further confirms that instruction-tuned LLMs can provide definitions similar to human ones but do not always process the underlying relation encoded into the semantics of compounds.

| **model** | **1-shot** | **3-shot** |
|---|---|---|
| Llama-2-7B-chat-hf | .41 | .18 |
| Mistral-7B-Instruct | .59 | .56 |
| Falcon-7B-Instruct | .15 | .14 |

Table 5: Prompt Accuracy over the LNC dataset.

| **model** | **1-shot** | **3-shot** |
|---|---|---|
| Llama-2-7B-chat-hf | .42 | .33 |
| Mistral-7B-Instruct | .80 | .75 |
| Falcon-7B-Instruct | .15 | .21 |

Table 6: Prompt Accuracy over the Nakov dataset.

## 4 Are LLMs Generalizing Semantic Relations over Novel Compounds?

Interpreting a novel compound (e.g, *birthday dessert*) involves both the conceptual and lexical systems; one must: i.) access the concepts denoted by the words and ii.) select a relation (e.g., a dessert *intended for* a birthday) to form a unified conceptual representation (Gagné and Spalding, 2006b). Coil and Shwartz (2023) observed that even for rare compounds, GPT-3 is able to generalize and make sense of new concepts, but the model tends to parrot incorrect paraphrases from the training set more often than correct ones.

We hereby designed and explored a diagnostic dataset to investigate how LLMs deal with novel compound interpretation. Instead of relying on randomly generated infrequent combinations, we manipulated our original dataset of lexicalized compounds by replacing the head or the modifier with one of its hypernyms in order to answer the following questions: i.) Can LLMs generalize (i.e., can they abstract) an implicit semantic relation that ties the two constituents of a conventional compound and transfer it to a semantically similar but novel compound? ii.) Does the LLMs performance change, as a function of the type of component (head or modifier) replaced for the construction of the novel compound?

### 4.1 Data and Methods

From the original dataset, we extracted the hypernyms of the head and modifier using WordNet 3.0[4]. Only hypernyms occurring more than 1000 times in the enTenTen20 corpus were selected. The frequency of the new bigram (the novel compound) was then calculated, and only meaningful expressions with a frequency of occurrence lower than 30 were retained as novel compounds. For instance, given the compound *apple orchard* ("an orchard that produces apples"), we created the compounds *pome orchard* ("an orchard that produces pomes") as a novel compound with the same head (*sameHead*) but replaced modifier, and *apple parcel* ("a parcel that produces apples"), as a same modifier (*sameMod*) but replaced head novel compound. This diagnostic dataset consists of 64 novel compounds covering four semantic relations: CONTAINMENT-R, LOCATION, PRODUCTION, and PURPOSE.

### 4.2 Results

**Surprisals** We computed the Surprisal scores on the novel compounds' paraphrases containing the original semantic relation (e.g., "pome orchard is an orchard that produces pomes") and compared them with the Surprisals of the corresponding distractor paraphrases (e.g., "pome orchard is an orchard that is located in pomes"), following the same methodology presented in the previous experiment. As expected, the results are lower than the previous experiment. An aspect to notice is that the models tend to assign the lower score to the paraphrase of the original semantic relation more often when the head is fixed (blue bar) than when the modifier is fixed (orange bar, Figure 2). This is valid for Llama-2 (Base and Instruct), Falcon Base, and Mistral Instruct. It is worth noticing that BERT performs better than some of the larger models but shows the opposite trend.

[4]We queried WordNet by relying on NLTK package, version 3.8.1. (Bird et al., 2009).

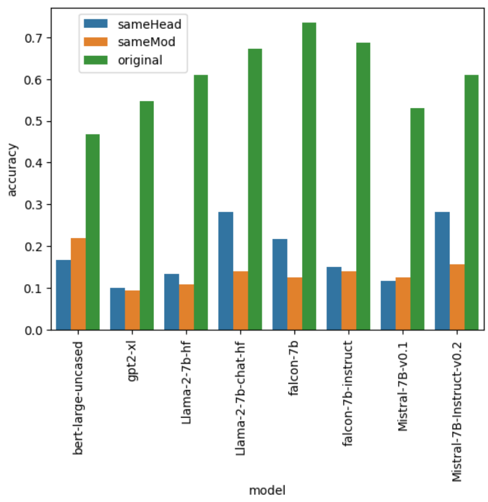


Figure 2: Surprisal Accuracy over the NNC dataset.

**Prompting** We observe that Llama-2 and Falcon perform poorly on this task, while Mistral achieves good performance, obtaining accuracy scores of .578 (1-shot) and .531 (3-shot) for the sameHead part of the NNC dataset and .469 (1-shot) and .30 (3-shot) for the sameDep. Considering just the results for this model, we observe that changing the head or the modifier affects the ability of the model to recognize good paraphrases differently. For example, the CONTAINMENT-R relation (CONT-R) is not particularly problematic when the novel word is the modifier (accuracy 1-shot: .474, 3-shot: .737). For instance, the novel compound *equipment box* (from *glove box*) is correctly paraphrased as "a box that contains equipments" by the two versions of Mistral. However, performance drops when changing the head (accuracy 1-shot: .37, 3-shot: .05). Given the previous example, Mistral (3-shot setting) associated to the compound *glove container* (from *glove box*) the paraphrase "a container intended for gloves" instead of "a container that contains gloves," which

| model | sameHead | | sameMod | |
|---|---|---|---|---|
| | 1 shot | 3 shot | 1 shot | 3 shot |
| Llama-2-7B-chat-hf | .156 | .172 | .141 | .219 |
| Mistral-7B-Instruct | **.578** | .531 | **.469** | .30 |
| Falcon-7B-Instruct | .047 | .063 | .079 | .047 |

Table 7: Prompt accuracy over the NNC dataset.

should be expected if the model retains the same semantic relation of the original compound. From a qualitative analysis, we observed a tendency for the model to answer with the PURPOSE category instead of the appropriate one; indeed, this category gets the highest number of correct answers among the four classes (1-shot: .82; 3-shot: .53).

## 5 General Discussion

This paper evaluated recent LLMs on their ability to interpret Noun-Noun compounds and, specifically, to correctly identify the semantic relation underlying existing and novel compounds.

For the interpretation of existing compounds (LNCs), we released a dataset that assembles several linguistic and conceptual features associated with each compound, extracted from previous resources or added by the authors (concreteness, semantic type, semantic relation from different classifications) together with a limited set of paraphrases generated from Pepper (2022)'s classification. LLMs accuracy was tested on both the Surprisal scores and metalinguistic knowledge extracted by prompting strategies. In both settings, the models showed different performance levels in the identification of different semantic relations. Some relations like PRODUCTION are easy to recognize; that is, its paraphrase is the most expected (considering Surprisal scores) and more frequently identified in a metalinguistic prompt task. Moreover, compounds characterized by higher concreteness were interpreted more accurately overall, as hypothesized. This effect may be explained by the so-called *concreteness effect* (Jessen et al., 2000), which suggests that concrete concepts are processed faster and more easily than abstract ones.

Previous studies reported that LLMs generate compound definitions that highly resemble human-generated paraphrases, reaching an almost perfect performance. However, the analyses presented here reveal that they are not as perfect when asked to identify the correct paraphrase, given alternatives. Our outcomes confirm what was observed by Coil and Shwartz (2023): LLMs' performance can largely be attributed to parroting definitions or parts of definitions extracted from the training corpora. However, it is unclear to what extent LLMs extract the relational linguistic patterns they learn from corpora and use them to hypothesize about the most likely relationship underpinning a noun compound. In other words, while the models can somehow interpret the semantic relation underlying compounding, there is still a question far from being completely answered: what linguistic properties make compounds more or less difficult to interpret by LLMs? For this reason, we believe that more effort should be made in designing a comprehensive dataset of noun-noun compounds annotated with different factors influencing the plausibility of the noun compounds.

The second experiment represents the first attempt to model novel compounds to understand LLMs' abilities to abstract and transfer knowledge. According to previous studies, the interpretation of a novel combination relies on previous language experience (Gagné and Shoben, 1997, 2002; Gagné and Spalding, 2006a, among others). That is, people are able to interpret novel compounds by abstracting from their knowledge of past experiences with similar conceptual combinations, which provide an analogical basis for the production and interpretation of novel compounds (Krott, 2009) [5]. We manipulated a subset of lexicalized compounds by replacing the modifier or the head word with a hypernym and observed how harder it is for the LLMs to interpret the generated compounds. As expected, language models are challenged by this task, but we observe that they still look for a suboptimal solution. For instance, they choose the PURPOSE relation, which has a more general paraphrase (*indended for*) than other relations (such as LOCATION or CONTAINER). We believe that this task could provide a window into a specific aspect of the creative abilities of LLMs.

In conclusion, the present study illustrates that there are still questions unanswered regarding how LLMs interpret compounding. Future works will focus on expanding both the LNC and the NNC datasets, including more linguistic features and evaluating the acceptability of selected paraphrases with human judgments.

[5] On the possible role of analogy compositionality process, see also the vector composition approach presented in Rambelli et al. (2022).

## Limitations

**The work focuses only on English** The present dataset and work are focused only on English. Expanding the dataset to other languages would be beneficial, but we currently lack the same amount of resources for other languages annotated with the same amount of linguistic information, such as concreteness ratings and semantic relations. However, we chose Pepper (2022) classification precisely because it has been implemented to be suitable across languages, and the Bourquifier templates could be easily converted into other languages. Additionally, the methodology presented to generate novel compounds could be replicated for other languages by relying on language-specific WordNet versions released within the OpenMultiWordNet project (Bond and Paik, 2012; Bond et al., 2016), accessible through the NLTK package.

**Prompting strategies are conservative** For the present study, we evaluated models in a conservative setting by using a low temperature. Further studies could investigate how the same models with higher temperatures answer, that is, how augmenting the linguistic creativity of LLMs affects models' performance on compound interpretations. An additional limitation concerns the prompt used. We evaluated all LLMs on the question "Which is the most likely description of COMPOUND?" followed by a list of possible paraphrases. However, we did not test whether other questions could improve the models' accuracy, nor did we explore whether different examples within the prompt could yield varied outcomes.

**Comparing LLMs' performance over humans' judgments** A limitation of this dataset comes from the annotations of Tratz (2011). We used an aggregated version of this dataset, so it is impossible to determine the degree of agreement across annotators for each compound. However, literature reports that some expressions show greater entropy of conceptual relations, i.e., greater competition between possible underlying semantic relations (Benjamin and Schmidtke, 2023). This information could be useful for a more fine-grained evaluation of LLMs' performance. A related consideration is that, when collecting paraphrases for compounds, there can be various relationships with different degrees of acceptability (Spalding and Gagné, 2014; Benjamin and Schmidtke, 2023), while we simplify by assuming there is only one correct relationship. While it was out of the scope of the present paper, we would further investigate these hypotheses and collect the acceptability of paraphrases for both lexicalized and novel compounds.

## Ethics Statement

**Data** The datasets used to build our LNC dataset are publicly available online. Concreteness ratings of Muraki et al. (2023) can be downloaded from the authors' OSF project: `https://osf.io/ksypa/`. For the Tratz (2011) dataset, we used the data released by Shwartz and Dagan (2018) at `https://github.com/vered1986/panic/tree/master/classification/data`. (Nakov, 2008b) dataset is available from the SIGLEX-MWE archive (`https://multiword.sourceforge.net/PHITE.php%3Fsitesig%3DFILES%26page%3DFILES_20_Data_Sets`) under Creative Commons Attribution 3.0 Unported License. We will release all additional data and code used in the present experiment.

**Models** For reasons of replicability, we used only open-access models available from huggingface. Given a limited GPU, we relied on 7 billion parameter models and used quantization techniques to reduce memory and computational costs (we used the `bitsandbytes` library).

There are well-known ethical concerns about LLMs, which have been shown to produce factually incorrect output, which may generate offensive content if prompted with certain inputs. Instruction-tuned LLMs have been trained to reduce the harm of model responses, as we also observed in our analyses. For instance, when asked to choose the correct paraphrase, the Llama-2 answered: "*It is important to clarify that child pornography is a criminal and morally reprehensible activity. Therefore, none of the descriptions provided accurately describe child pornography. Instead, it is essential to understand that child pornography involves the production..*". However, some responses may still contain offensive content. Finally, any demonstrations of LLMs' linguistic generalizations should not imply that they are safe to use or can be expected to behave in a way that is aligned with human preferences and values.

## References


Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon Series of Open Language Models. *arXiv preprint arXiv:2311.16867*.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning Aligns LLMs to the Human Brain. *arXiv preprint arXiv:2312.00575*.

Shaina Benjamin and Daniel Schmidtke. 2023. Conceptual Combination during Novel and Existing Compound Word Reading in Context: A Self-paced Reading Study. *Memory & Cognition*, pages 1–28.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Francis Bond and Kyonghee Paik. 2012. A Survey of Wordnets and their Licenses. In *Proceedings of the Global WordNet Conference*.

Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. Cili: The Collaborative Interlingual Index. In *Proceedings of the 8th Global WordNet Conference*.

Lars Buijtelaar and Sandro Pezzelle. 2023. A Psycholinguistic Analysis of BERT's Representations of Compounds. In *Proceedings of EACL*.

Jordan Coil and Vered Shwartz. 2023. From Chocolate Bunny to Chocolate Crocodile: Do Language Models Understand Noun Compounds? In *Findings of ACL*.

D Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Christina L Gagné and Edward J Shoben. 1997. Influence of Thematic Relations on the Comprehension of Modifier–noun Combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):71.

Christina L Gagné and Edward J Shoben. 2002. Priming Relations in Ambiguous Noun-noun Combinations. *Memory & Cognition*, 30(4):637–646.

Christina L Gagné and Thomas L Spalding. 2006a. Conceptual Combination: Implications for the Mental Lexicon. *The Representation and Processing of Compound Words*, pages 145–168.

Christina L Gagné and Thomas L Spalding. 2006b. Using Conceptual Combination Research to Better Understand Novel Compound Words. *SKASE Journal of Theoretical Linguistics*, 3(2):9–16.

Iris Hendrickx, Preslav Nakov, Stan Szpakowicz, Zornitsa Kozareva, Diarmuid O Séaghdha, and Tony Veale. 2013. SemEval-2013 Task 4: Free Paraphrases of Noun Compounds. *Proceedings of SemEval*.

Frank Jessen, Reinhard Heun, Michael Erb, D-O Granath, Uwe Klose, Andreas Papassotiropoulos, and Wolfgang Grodd. 2000. The Concreteness Effect: Evidence for Dual Coding and Context Availability. *Brain and Language*, 74(1):103–112.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of ACL*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely. *Cognitive Science*, 47(11):e13386.

Andrea Krott. 2009. The Role of Analogy for Compound Words. In *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

Judith N Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Siyan Li, Riley Carlson, and Christopher Potts. 2022. Systematicity in GPT-3's Interpretation of Novel English Noun Compounds. In *Findings of EMNLP*.

Filip Miletić and Sabine Schulte im Walde. 2023. A Systematic Search for Compound Semantics in Pretrained BERT Architectures. In *Proceedings of EACL*.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Emiko J Muraki, Summer Abdalla, Marc Brysbaert, and Penny M Pexman. 2023. Concreteness Ratings for 62,000 English Multiword Expressions. *Behavior Research Methods*, 55(5):2522–2531.

Preslav Nakov. 2008a. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In *Artificial Intelligence: Methodology, Systems, and Applications: 13th International Conference, AIMSA*

*2008, Varna, Bulgaria, September 4-6, 2008. Proceedings 13*, pages 103–117. Springer.

Preslav Nakov. 2008b. Paraphrasing Verbs for Noun Compound Interpretation. In *Proceedings of the LREC Workshop on Multiword Expressions*.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring Noun-modifier Semantic Relations. In *Proceedings of IWCS*.

Diarmuid Ó Séaghdha. 2008. Learning Compound Noun Semantics. Technical report, University of Cambridge.

Mark Ormerod, Barry Devereux, and Jesús Martínez del Rincón. 2023. How is a "Kitchen Chair" like a "Farm Horse"? Exploring the Representation of Noun-Noun Compound Semantics in Transformer-based Language Models. *Computational Linguistics*, pages 1–33.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge. In *Proceedings of *SEM*.

Steve Pepper. 2021. The Bourquifier: An Application for Applying the Hatcher-Bourque Classification (Version 3)[MS Excel]. https://www.academia.edu/83122396.

Steve Pepper. 2022. Hatcher-Bourque: Towards a Reusable Classification of Semantic Relations. In *Binominal Lexemes in Cross-Linguistic Perspective*, pages 303–354. De Gruyter Mouton.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2022. Compositionality as an Analogical Process: Introducing ANNE. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Vered Shwartz and Ido Dagan. 2018. Paraphrase to Explicate: Revealing Implicit Noun-compound Relations. In *Proceedings of ACL*.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

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 NAACL*.

Thomas L Spalding and Christina L Gagné. 2014. Relational diversity affects ease of processing even for opaque english compounds. *The Mental Lexicon*, 9(1):48–66.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Stephen Tratz. 2011. *Semantically-enriched Parsing for Natural Language Understanding*. Ph.D. thesis, University of Southern California.

Henk J Van Jaarsveld and Gilbert E Rattink. 1988. Frequency Effects in the Processing of Lexicalized and Novel Nominal Compounds. *Journal of Psycholinguistic Research*, 17:447–473.

Caterina Villani, Adele Loia, and Marianna M. Bolognesi. 2024. The Semantic Content of Concrete, Abstract, Specific, and Generic Concepts. *Language and Cognition*, page 1–28.

Edward J Wisniewski and Bradley C Love. 1998. Relations versus Properties in Conceptual Combination. *Journal of Memory and Language*, 38(2):177–202.

# A Experiment 1-Prompt Example

We report below an example of the prompt used as an example (1-shot setting) for the LNC dataset.

```
Which is the most likely description of "olive oil"?

    1. an oil that uses olives;
    2. an oil that is part of olives;
    3. an oil that olives produce;
    4. an oil that produces olives;
    5. an oil that contains olives;
    6. an oil that is about olives;
    7. an oil that is composed of olives;
    8. an oil that is located in olives;
    9. an oil intended for olives
```

We report below an example of the prompt used as an example (1-shot setting) for the Nakov dataset.

```
Which is the most likely description of “pumpkin pie”?

    1. a pie that uses a pumpkin;
    2. a pie that is caused by a pumpkin;
    3. a pie that is made from a pumpkin;
    4. a pie that gives a pumpkin;
    5. a pie that comes from a pumpkin;
    6. a pie that is made by a pumpkin;
    7. a pie that causes a pumpkin;
    8. a pie that is a pumpkin;
    9. a pie that involves a pumpkin
```

# B Experiment 1-Additional Analyses

As for the LNC dataset, we plot the distribution of semantic relations with the lowest Surprisal scores inside each class for the Nakov dataset. Figure 3 allows us to grasp common errors across LLMs.

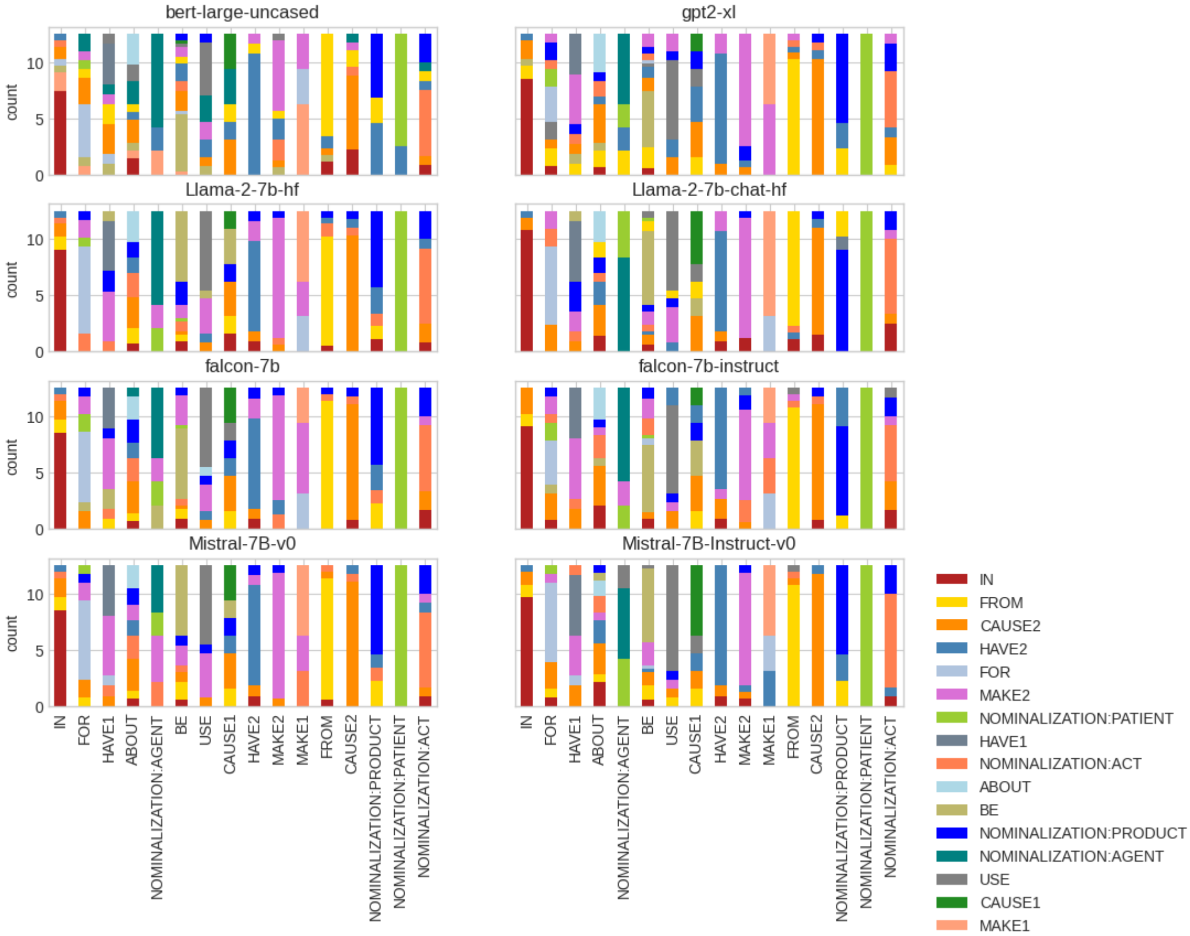


Figure 3: Navok dataset: Distribution of semantic relations with the lowest Surprisal scores for each relation.

# C Experiment 2 - Additional results

| | NNC | |
|---|---|---|
| | **sameHead** | **sameMod** |
| BERT-large | .219 | **.167** |
| GPT2-xl | .100 | .094 |
| Llama-2 (Base) | .133 | .109 |
| Falcon (Base) | .217 | .125 |
| Mistral (Base) | .117 | .125 |
| Llama-2 (Instruct) | **.283** | .141 |
| Falcon (Instruct) | .150 | .141 |
| Mistral (Instruct) | **.283** | **.156** |

Table 8: Surprisal accuracy of instruction-based models on the NNC dataset, distinguishing when we substitute the first word (*sameHead*) or the second word (*sameMod*) of a compound with a hypernym.