Syntax and Semantics Meet in the "Middle": Probing the Syntax-Semantics Interface of LMs Through Agentivity

Lindia Tjuatja, Emmy Liu, Lori Levin, Graham Neubig
Language Technologies Institute
Carnegie Mellon University
{ltjuatja, mengyan3, lsl, gneubig}@cs.cmu.edu

Abstract

Recent advances in large language models have prompted researchers to examine their abilities across a variety of linguistic tasks, but little has been done to investigate how models handle the interactions in meaning across words and larger syntactic forms—i.e. phenomena at the intersection of syntax and semantics. We present the semantic notion of *agentivity* as a case study for probing such interactions. We created a novel evaluation dataset by utilizing the unique linguistic properties of a subset of optionally transitive English verbs. This dataset was used to prompt varying sizes of three model classes to see if they are sensitive to agentivity at the lexical level, and if they can appropriately employ these word-level priors given a specific syntactic context. Overall, GPT-3 `text-davinci-003` performs extremely well across all experiments, outperforming all other models tested by far. In fact, the results are even better correlated with human judgements than both syntactic and semantic corpus statistics. This suggests that LMs may potentially serve as more useful tools for linguistic annotation, theory testing, and discovery than select corpora for certain tasks.

1 Introduction

Consider the English sentences in (1) below:

(1) a. This author writes easily.
    b. This passage writes easily.

These sentences display an interesting property of certain optionally transitive verbs in English. Although they share an identical surface syntactic structure—a noun phrase in subject position followed by the intransitive form of the verb and an adverb phrase modifying the verb—they entail very different things about the roles of their subjects.

The subject of (1a) is someone that does the action of writing; in other words, *this author* is an **agent** in the writing event. On the other hand, the subject of (1b), *this passage*, doesn't do any writing—it is what is created in the event of writing. In contrast to *this author*, *this passage* is a **patient**. The agent and patient roles are not discrete categories, but rather prototypes on opposite ends of a continuum. These "protoroles" have a number of contributing properties such as causing an event for agents and undergoing change of state for patients (Dowty, 1991).

The contrast between the minimal pair in (1) suggests that there are lexical semantic properties of the subjects that give rise to these two distinct readings: one that describes how the subject generally *does* an action as in (1a), and another that describes how an event generally unfolds when the subject *undergoes* an action as in (1b). Intuitively, a speaker may know from the meaning of *author* that authors are animate, have some degree of volition, and typically write things, whereas passages (of text) are inanimate, have no volition, and are typically written. The knowledge of these aspects of meaning must somehow interact with the syntactic form of the sentences in (1) to disambiguate between the two possible readings, and an agent or patient role for the subject follows from the meaning of the statement as a whole.

Now consider the (somewhat unusual) sentences in (2) which use the transitive form of *write*:

(2) a. Something writes this author easily.
    b. This passage writes something easily.

At first glance, the above sentences (with the same sense of *write* as in 1) are infelicitous unless we imagine some obscure context where *this author* is something like a character in a text and *this passage* is somehow anthropomorphized and capable of writing; these contexts go against our natural intuitions of the semantics of "passage" and "author".[1] Unlike the syntactic form of the sentences

[1] There is another reading of (2a) that uses a different sense of *write*, where *this author* is a recipient (*Something writes*

in (1), the explicit inclusion of both arguments (subject and direct object) now forces whatever is in subject position to be the agent and whatever is in object position to be more like a patient, regardless of the typical semantic properties of the arguments.

Taken together, the examples in (1) and (2) illustrate a compelling interaction at the *syntax-semantics interface*. More specifically, we see a two-way interaction: first, near-identical surface forms acquire completely different entailments about their subjects *solely* depending on the choice of subject, while conversely certain syntactic forms can influence the semantic role of an argument *regardless* of the usual behavior of said argument. We aim to investigate the linguistic capabilities of language models with regards to this interaction.

Prior work in studying LMs as psycholinguistic subjects has largely focused on syntax and grammatical well-formedness (Futrell et al. 2019; Linzen and Baroni 2021, inter alia). However, as illustrated in the above examples, there are instances of near-identical syntactic structures that can give rise to different meanings depending on the individual lexical items as well as surrounding context. Thus evaluating LMs on syntax, while a necessary starting point, does not give us a sufficient measure of LM linguistic capabilities. While other work such as Ettinger (2020), Kim and Linzen (2020), and Misra et al. (2022) (among others) evaluate LMs on a variety of tests involving semantics and pragmatics, they do not investigate the interaction between the meanings associated with syntactic forms and those of individual lexical items.

Thus, we not only need to evaluate syntax and utilization of semantic knowledge, but we also need to understand how interactions of meaning at different linguistic levels—i.e. morphological, lexical, phrasal—may alter model behavior. Exploring phenomena within the syntax-semantics interface is a compelling approach as it gives us access to specific aspects of semantics while allowing precise control over syntactic form between levels.

In this work, we probe the syntax-semantics interface of several language models, focusing on the semantic notion of agentivity. We do this by prompting models to label nouns in isolation or in context as either agents or patients from a curated test set of noun-verb-adverb combinations that display the alternation shown in example (1). We then compare the performance of LMs to both human judgements and corpus statistics.

Probing for LMs for their knowledge of agentivity in syntactic constructions as in (1) and (2) is a particularly insightful case study as it allows us to explore three interconnected questions in a highly controlled syntactic setting:

I. Do models display sensitivity to aspects of word-level semantics independent of syntactic context, and is such sensitivity aligned with human judgements? (§3.1)

II. Can models employ lexical semantics to determine the appropriate semantics of a sentence where the syntax is ambiguous between readings (as in 1)? (§3.2)

III. Can models determine the semantics of a sentence from syntax, disregarding lexical semantics when necessary (as in 2)? (§3.3)

Additionally, the relatively infrequent pairings of semantic function and syntactic form of sentences such as (1b) are also interesting from a learnability and acquisition perspective for both LMs and humans. How both come to process and acquire exceptions to a general "rule" has been a topic of debate since early connectionist models (Rumelhart and McClelland, 1986). Hence, knowledge of LM capabilities in acquiring and processing these linguistic anomalies may serve as valuable insight to linguists, cognitive scientists, and NLP practitioners alike.

2 Methodology

We constructed three experiments, each targeting one of the above questions through the lens of agentivity. We will first give a broad overview of each, and then go into detail about the general approach.

Experiment 1 (§3.1) tests whether language models are sensitive to the word-level semantics of nouns with regards to agentivity, such as whether nouns like *author* and *passage* are more likely to be agents or patients without any surrounding context. This is analogous to the idea that speakers have intuition for how entities prototypically act in events, e.g. that *authors* write and *passages* are written, and that this extends to how we categorize their roles in events (Rissman and Majid, 2019).

Experiment 2 (§3.2) tests whether language models can disambiguate between the possible readings of sentences of the form in (1)—i.e. if

(to) this author easily). Regardless, given that the agent and patient roles as defined by Dowty (1991) are prototypes on a scale, *this author* in the recipient reading is closer to the patient role.

Exp 1: noun (lexical level)	Exp 2: intransitive (ambiguous mapping)	Exp 3: transitive (deterministic mapping)
noun: John agent/patient: agent	Sentence: John walks quickly. Is John an agent or a patient?: agent	Sentence: Jack throws something easily. Is Jack an agent or a patient?: agent
noun: vase agent/patient: patient	Sentence: This vase breaks easily. Is vase an agent or a patient?: patient	Sentence: Something hires the nurse swiftly. Is nurse an agent or a patient?: patient
noun: nurse agent/patient: agent	Sentence: This nurse works swiftly. Is nurse an agent or a patient?: agent	Sentence: The hammer breaks something quickly. Is hammer an agent or a patient?: agent
noun: mango agent/patient: patient	Sentence: This mango blends well. Is mango an agent or a patient?: patient	Sentence: Something blends the mango well. Is mango an agent or a patient?: patient
noun: <noun> agent/patient:	Sentence: <intr-agent/intr-patient> Is <noun> an agent or a patient?:	Sentence: <trans-agent>/<trans-patient> Is <noun> an agent or a patient?:

Figure 1: Prompt setup for each experiment. Note that the examples given for Exp 1 are not meant to be hard labels, rather they are "tendencies" for these nouns. In Exp 2, the noun itself determines whether the sentence is considered **intr-agent** or **intr-patient**; in Exp 3, we force the noun to take the agent or patient role by placing it in subject (**trans-agent**) or object (**trans-patient**) position.

they can identify whether the syntactic subject is an agent or a patient when the verb can allow for either. Sentences with the intransitive form of the verb that describe how the subject (an agent) does an action demonstrate *object drop* (as the direct object of the normally transitive verb is "dropped"), while sentences that describe how an event unfolds when the subject (a patient) undergoes an action are called *middles*, short for the linguistic term *dispositional middle* (van Oosten 1977; Jaeggli 1986; Condoravdi 1989; Fagan 1992, inter alia).[2] In our experimental setup, we will refer to these as **intr-agent** and **intr-patient**, respectively. If a model can do this task successfully by employing semantic information about the noun, we would expect not only to see that nouns in subject position are classified correctly as agents or patients, but also that these predictions for the most part correlate to the predictions in the first experiment.

Finally, **Experiment 3** (§3.3) tests whether language models can disregard word-specific priors to identify whether the noun of interest in a sentence with a transitive verb (such as those in 2) is an agent or patient. Since the semantic role of the noun maps directly to its syntactic position in these sentences, all subjects should be agents and all objects should be patients. For our test set, we create sentences where the position of the noun is the subject (**trans-agent**) and sentences where it is the object (**trans-patient**) for every noun.

2.1 General approach and data curation

In all of these experiments, we rely on the prompting paradigm to elicit LM probabilities of an "agent" or "patient" label for a given noun in isolation or within a sentence. Our prompting method consists of four examples with gold labels, followed by the unlabeled test example in the same format, as shown in Figure 1. As this task has not been explored in prior literature, we had to construct our own examples to test on.

The highly controlled syntactic setting that allows us to explore the alternation in agentivity as displayed in (1) and (2) is a double-edged sword—while this setting provides us with a minimal pair, it also restricts the types of verbs that work in this experimental setup. The second (**intr-agent** vs. **intr-patient**) and third (**trans-agent** and **trans-patient**) experiments require verbs that are optionally transitive and have no preference for whether an agent or a patient is the subject of the intransitive form, as in (1). These requirements together highly constrain the class of verbs that work in this experimental setup, and as far we can tell there exists no definitive list in the linguistics literature of English verbs that display both properties.

As a starting point to curate a list of verbs, we consulted literature on verbs that display object drop (Gillon 2012; Fillmore 1986, as well as Levin 1993 for an overview of English verb classes). We compiled a list of 23 verbs (see Appendix A), though this list is certainly non-exhaustive. For each verb, we list nouns and adverbs that can work in combination with each other in all of the templates in Table 1. Criteria for adding nouns and adverbs are listed in the Appendix B.

[2] Note that in English, dispositional middles also allow for what are considered non-patient promoted objects (such as paths, e.g. *The desert crosses easily*) (Tenny 1994, 1992), but for convenience we will treat them as being in the same category as patients.

In total, we have 233 unique nouns and a total of 820 noun-verb-adverb combinations. Out of these combinations, 343 form **intr-agent** sentences and 477 form **intr-patient** sentences. Since we can put any noun into syntactic subject or object position for the transitive sentences, we have 820 sentences each for **trans-agent** and **trans-patient**.

Sentence	**Template**
	This **<noun> <verb> <adverb>**.
intr-agent	This author writes easily.
intr-patient	This paper writes easily.
	This **<noun> <verb>** something **<adv>**.
trans-agent	This author writes something easily. This paper writes something easily.
	Something **<verb>** this **<noun> <adv>**.
trans-patient	Something writes this author easily. Something writes this paper easily.

Table 1: Templates for experiments 2 and 3. Sentences highlighted in pink contain a **<noun>** with an "agent" label, while those in blue with "patient".

2.2 Approximating "ground truth" agentivity labels for nouns out of context

Getting a gold "agent" or "patient" label is straightforward in the experiments with nouns in context: for sentences with the intransitive this was done ad hoc during data curation, and for sentences with the transitive this is a one-to-one mapping to syntax. However, using a hard label for nouns in isolation is problematic as a semantic role label is meaningless without context of the event; in principle, given an appropriate context, anything can act upon something else or have something done to it (literally or figuratively).

To get around this, we have two methods for finding an approximate label for the "typical" agentivity of a noun. The first was to collect human judgements. 19 annotators (native/fluent bilingual English proficiency) were given nouns without any context and were tasked to judge how likely each noun is to be an agent in any arbitrary event where both an agent and patient are involved. Their judgements were collected via ratings on a scale from 1 (very unlikely to be an agent) to 5 (very likely to be an agent). For nouns that have multiple common word senses (e.g. "model" can refer to both a fashion model or machine learning model, among other things) we include a disambiguating description. This description does not contain any verbs or other explicit indications of what events the noun may occur in (e.g. for "model", we give human annotators "model (person)").[3] We then average the ratings across all annotators and normalize so that the values fall between 0 and 1. To calculate inter-annotator agreement, we randomly divide the annotators into two groups (of 9 and 10), average their ratings for each noun, and calculate the correlation between the two; doing this seven times yields an average inter-group correlation of 0.968.

The second method uses statistics from linguistically annotated corpora as a proxy for the "typical" agentivity of a noun. We do this by calculating the frequency of "agenthood" for a noun (**agent ratio**), i.e. dividing the number of times the noun appears as an agent by the number of times it is either an agent or patient. The ideal annotated corpus for this would be one with semantic role labels such as Propbank (Kingsbury and Palmer, 2002), where the "ARG0" label corresponds to agent and "ARG1" to patient. However, many of the nouns in our data appeared only a few times in Propbank or not at all—out of all 233 nouns, only 166 of them occurred within an ARG0 or ARG1 span.[4]

Thus, we also tried utilizing syntax as a proxy using Google Syntactic Ngrams biarcs (Goldberg and Orwant, 2013), as it is significantly larger. The biarcs portion of the corpus covers dependency relations between three connected content words, which includes transitive predicates. To calculate a similar ratio, we divide the number of times a noun occurs as a subject by the total number of subject and direct object occurrences (we call this the **subject ratio**). A value closer to 1 should correlate with a tendency to occur more often as an agent, as agents are generally coded as subjects of English transitive verbs and patients as direct objects. All but one of our nouns contained at least one instance of occurring with a "nsubj" or "dobj" label.

3 Experimental Results

We evaluate BLOOM (Scao et al., 2022), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020) models of varying sizes for all experiments. Since previous work has shown that models are highly sensitive to the ordering of examples (Lu et al., 2021), we run each experiment twice: once with the order shown in Figure 1 where an agent

[3] Additional details on collecting human ratings can be found in Appendix C.

[4] We used Propbank annotations for BOLT, EWT, and Ontonotes 5.0 from https://github.com/propbank/propbank-release.

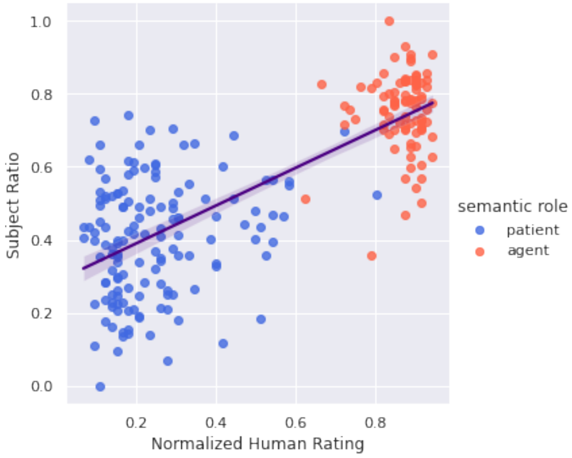


Figure 2: Correlation between subject ratio (from Google Syntactic Ngrams) and human ratings for each noun ($r = 0.762$). The semantic role label is the role the noun takes as the subject of the intransitive verb within our test set.

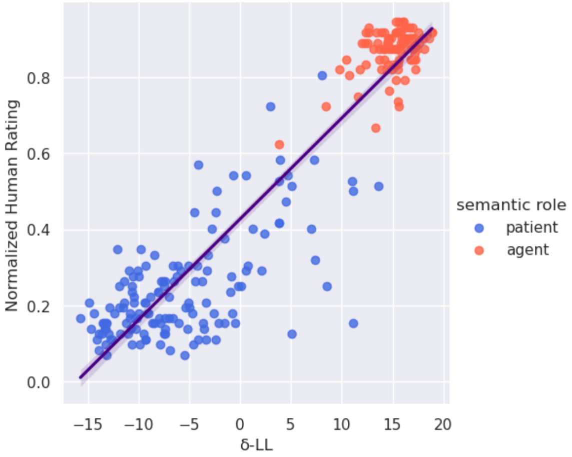


Figure 3: Correlation between δ-LL in Experiment 1 for GPT-3 `davinci-003` and the normalized human rating in the APAP experiment. Note that a negative δ-LL means the "patient" label is more likely.

is first (APAP ordering) and again with the first example moved to the bottom (PAPA ordering). We compare models based on their average performance across both orderings. Note, however, that some models are more sensitive to orderings than others; some models (like `text-davinci-003`) are largely invariant to example ordering. In Appendix D, we report results from both experiments.

3.1 Exp 1: Agentivity at the lexical level

In order to see if models are sensitive to the notion of how "typically" agentive a noun is, we compare the difference in log-likelihood between predicting "agent" or "patient" for that noun (δ-LL) with the normalized human ratings as well as corpus statistics from Google Syntactic Ngrams and Propbank.

Before we compare models with Ngrams and Propbank, we first ask how well-correlated both are with human ratings. We find that the subject ratio calculated from occurrence counts in Google Syntactic Ngrams is positively correlated with the average human rating with Pearson's r of 0.762, though the human rating has a stronger divide between agents and patients. This can be seen in Figure 2. When comparing with humans, using Syntactic Ngrams for this task actually turns out to be better than using Propbank: for the 166 nouns that occur with ARG0/1 labels, there is a correlation of 0.555 with human ratings (see Appendix E for details).

Overall, as seen in Table 2, we find that most models have a weak correlation with human ratings, with the exception of GPT-3 `text-davinci-003` (henceforth `davinci-003`), shown in Figure 3. We also see that `davinci-003` is not only both better correlated with human judgements than with corpus statistics, but surprisingly there is also a stronger correlation between its δ-LL and human ratings than between these proxies (syntactic and semantic) and human ratings. In fact, `davinci-003` is extremely close to the average inter-annotator group correlation, and furthermore this correlation is largely invariant to the ordering of prompts.

The observation that `davinci-003` is better correlated with human judgement than both syntactic (Ngrams) and semantic (Propbank) corpus statistics is intriguing as both types of corpora have been used in modeling prediction of *thematic fit*, or how well a noun fulfills a certain thematic role with a verb (Sayeed et al., 2016). Thus, we may naturally expect this to also work well with "general tendencies" or typicality judgements for nouns by themselves. However, it seems that such corpora may be too small or genre-biased to fully capture the nuances of human judgements, and such judgements may be better captured by LMs that have seen vast quantities of data across a wide variety of domains, even without explicit human annotation.

3.2 Exp 2: Disambiguating agentivity with the intransitive

In this experiment, we evaluate models along two metrics: how accurate the model is in predicting the

Model	Human	Ngrams	PB
BLOOM `560m`	0.549	0.519	0.377
BLOOM `1b1`	0.374	0.358	0.291
BLOOM `1b7`	0.340	0.288	0.278
BLOOM `3b`	0.305	0.348	0.231
BLOOM `7b1`	0.016	-0.129	0.011
GPT-2 `small`	0.650	0.569	0.463
GPT-2 `medium`	0.394	0.451	0.333
GPT-2 `large`	0.499	0.544	0.412
GPT-2 `xl`	0.358	0.349	0.227
GPT-3 `ada-001`	0.594	0.575	0.490
GPT-3 `babbage-001`	0.311	0.337	0.158
GPT-3 `curie-001`	0.107	0.181	0.128
GPT-3 `davinci-001`	0.467	0.461	0.330
GPT-3 `davinci-003`	**0.939**	**0.730**	**0.574**
Inter-annotator	0.968	–	–
Google Syntactic Ngrams	0.762	–	–
Propbank	0.555	–	–

Table 2: Correlation between the difference in log-likelihood of predicting "agent" or "patient" with human ratings, subject ratio calculated from Google Syntactic Ngrams (232/233 nouns), and agent ratio calculated from Propbank (166/233 nouns), averaged across APAP and PAPA experiments.

correct label in context and how strongly correlated the δ-LL in this experiment is with the δ-LL from Experiment 1.

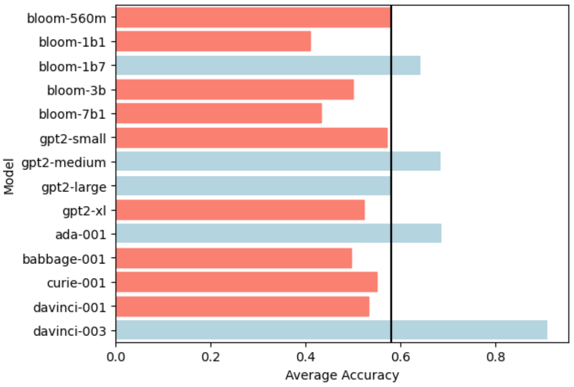


Figure 4: Average accuracy for predicting the label of nouns in **intr-agent/intr-patient** sentences. The black line indicates majority class performance; blue bars indicate above majority class performance.

Figure 4 shows the accuracy of each model in predicting (giving a higher probability to) the correct semantic label. Over half of the models do not achieve chance performance (predicting the majority class $\approx$ 0.582). Interestingly, we find that there is no monotonic increase in performance for this task with respect to model size (Kaplan et al., 2020)—for example, performance drops drastically between `text-ada-001` and `text-babbage-001`. This is also the case in Experiment 1.

We also evaluate how strongly correlated the δ-LL between predicting "agent" or "patient" for the noun in subject position of the intransitive is with the δ-LL of the noun in isolation. Since the role of the noun in the intransitive is heavily dependent on the meaning of the noun itself, if a model is using this information to disambiguate we would expect that the δ-LL in this experiment is correlated with δ-LL from Experiment 1. Furthermore, we would also want it to be strongly correlated with our approximate "ground truth" measures for agentivity, especially human ratings.

These correlations are shown in Table 3. As expected, `davinci-003` displays a strong relationship between the δ-LL from intransitive sentences with the δ-LL from Experiment 1, and furthermore also has a strong correlation with human ratings. Like in Experiment 1, `davinci-003`'s performance is invariant to changes in example orders.

Model	Noun δ-LL	Human	Ngrams	PB
BLOOM `560m`	0.605	0.217	0.147	0.100
BLOOM `1b1`	0.702	-0.0344	0.0200	0.0511
BLOOM `1b7`	0.540	0.706	0.562	0.441
BLOOM `3b`	0.258	0.280	0.190	0.0871
BLOOM `7b1`	0.385	0.161	0.124	0.0689
GPT-2 `small`	0.655	0.424	0.309	0.290
GPT-2 `medium`	0.611	0.523	0.516	0.505
GPT-2 `large`	0.551	0.609	0.489	0.447
GPT-2 `xl`	0.548	0.507	0.445	0.363
GPT-3 `ada-001`	0.541	0.496	0.358	0.307
GPT-3 `babbage-001`	0.127	-0.176	-0.170	-0.125
GPT-3 `curie-001`	0.130	0.156	0.189	0.0953
GPT-3 `davinci-001`	0.487	0.647	0.515	0.376
GPT-3 `davinci-003`	**0.914**	**0.919**	**0.715**	**0.567**

Table 3: Correlation between the δ-LL from **intr-agent/intr-patient** sentences with the δ-LL from the noun in isolation, human ratings, subject (Google Syntactic Ngrams), and agent ratios (Propbank).

3.3 Exp 3: Agentivity with the transitive

As previously discussed, the syntactic position of the noun in the transitive sentences (subject or object) directly map to their semantic roles (agent and patient, respectively). Figure 5 shows accuracy split by **trans-agent** and **trans-patient**.

As in the previous experiments, GPT-3 `davinci-003` outperforms all other models (0.994 for **trans-agent** and 0.991 for **trans-patient**—it is actually the *only* model which performs significantly above chance for both Experiments 2 and 3, and is also consistent across both example orderings.

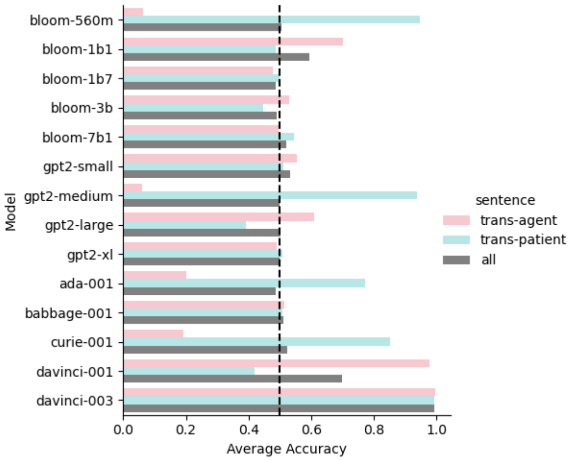


Figure 5: Average accuracy across **trans-agent**, **trans-patient**, and all transitive sentences. The dashed line indicates chance performance.

4 A Closer Look at davinci-003

Given that GPT-3 `davinci-003` does extremely well, a natural question to ask is whether `davinci-003` "fails" in similar ways to humans—i.e. we can see whether the nouns that are misclassified in the intransitive sentence setting (§3.2) are more ambiguous to humans as well.

In both APAP and PAPA orderings, all or nearly all of what `davinci-003` gets incorrect are patient subjects; all 78 incorrectly classified subjects of sentences in the APAP ordering are patients, and 69 of the 70 incorrect subjects in the PAPA ordering are patients. From this, one way to answer the above question is to compare this subset of nouns with the subset of nouns with a "patient" label (in the intransitive construction) that humans tend to rate as more agentive.

4.1 Animacy and thematic fit

Table 4 lists the latter subset of nouns, i.e. the most "agent-like" nouns with a "patient" label in the intransitive construction. Recall that human annotators were asked to rate each noun in isolation from a scale from 1 (very unlikely to be an agent) to 5 (very likely to be an agent) which is then normalized to a scale from 0 to 1, whereas the gold labels for nouns are determined by role it takes in the constructed (in this case, intransitive) sentences.

Animate nouns, such as "model (person)", "animal", and "fish" are unsurprisingly in this list, as many linguists have noted that the notion of agentivity is closely related to animacy (Silverstein 1976; Comrie 1989, inter alia). However, across both orderings, the only noun that was misclassified was "model" in the sentence *This model photographs beautifully/nicely*. Nevertheless, it could be argued that an agent interpretation in this context is plausible.

It appears that there are two interactions that are occurring in the above example. First, we must consider the *selectional restrictions* and of the verb, i.e. what arguments are allowable in the event described by the verb (Chomsky 1965; Katz and Fodor 1963). While selectional restrictions are traditionally viewed as binary features, a weaker, gradient version of this is *selectional preferences*, or the degree to which an argument fulfills the restrictions of the event (Resnik, 1996). A closely related notion to this is *thematic fit*, which is how much a word fulfills these preferences.

Secondly, the *Animacy Hierarchy*—of which humans are at the top—plays a role in such selectional restrictions and preferences, and thus in thematic fit (Trueswell et al., 1994). Since *photograph* requires a human-like entity as an agent, it could be argued that the interpretation of "model" being an agent in this sentence is not invalid (though likely a less salient interpretation by English speakers), as nothing in the "photographing" event rules out a subtype of a human "model" being the agent. This contrasts with the example with "animal" in our test set (*This animal photographs beautifully/nicely*), which would be far less acceptable with an animal agent interpretation, and falls below "human model" in the Animacy Hierarchy.

4.2 Verbs with vehicle objects

The other class of nouns present in Table 4, which also happen to be the remaining nouns, are vehicles. With regards to the relationship between animacy and agentivity, prior work such as Zaenen et al. (2004) has noted that "intelligent machinery" (such as computers and robots) and vehicles also often act as animates (below humans and above inanimates). Interestingly, nearly half of the examples that `davinci-003` gets wrong are sentences containing verbs with vehicle objects (*This car/vehicle/SUV/tractor/etc. drives nicely, This jet/plane/aircraft/etc. flies smoothly*). In fact, the examples that `davinci-003` gets the "most wrong" (higher $LL_{incorrect} - LL_{correct}$) are sentences with these verb-noun combinations.

Like the above examples with "model", some of

Noun	Human	Ngrams	Noun δ-LL
model (person)	0.806	0.523	8.06
animal	0.722	0.699	2.97
jet	0.583	0.562	7.27
aircraft	0.583	0.551	3.92
fish	0.569	0.467	-4.08
vehicle	0.542	0.468	4.66
bus	0.542	0.394	0.537
tank	0.542	0.564	-0.639
plane	0.528	0.565	11.1
car	0.528	0.565	3.83
motorcycle	0.514	0.184	5.11
truck	0.514	0.437	13.6
SUV	0.480	0.500	-2.27
tractor	0.401	0.500	11.2

Table 4: Nouns in **intr-patient** sentences with normalized human ratings ≥ 0.5, along with their subject ratio from Google Syntactic Ngrams and the average δ-LL from nouns in isolation (3.1). The average δ-LL for "patient" nouns ranges from -15.7 to 13.6. Note that *model* was presented to annotators with a disambiguating word sense (*person*).

these sentences have a possible alternative reading and are more ambiguous compared to sentences with verbs like *sell* (as in, *This car sells well.*). More specifically, they have a possible (though also less salient) unergative reading: e.g. in *This jet flies smoothly*, it could be a statement about how the jet flies on its own as opposed to about how the jet flies when someone flies it. Out of all the sentences in the test set, these are the only ones (along with some sentences with "turn") where the **intr-agent** has this possible unergative reading.

5 Related Works

There has been extensive work in the psycholinguistics literature investigating how humans make use of the relationship between events described by verbs and nouns that may participate in these events, which is especially relevant to the analysis described in §4.1. Works such as Tanenhaus et al. (1989) and Trueswell et al. (1994) have shown that humans utilize information about thematic fit to resolve ambiguity in sentence processing, mainly focusing on garden-path sentences.

Along this line of work, McRae et al. (1998) and Padó (2007) created human judgement datasets for thematic fit by asking humans to rate nouns associated with events (e.g. a crook arresting/being arrested by someone) on a scale from 1 (very uncommon/implausible) to 7 (very common/plausible). As stimuli, humans are given the noun, the verb describing the event, and the role of the noun. While this setup is similar to our dataset, they focus on the explicit relationship between the event and the noun, while our data is meant to focus on the relationship between the prototypical role of a noun (out of context) and its role in a controlled syntactic environment. Furthermore, as we would like the agent/patient distinction to be a minimal pair resulting changing the noun in an identical surface form, the sets of nouns and verbs between their studies and ours only partially overlap.

This study also follows a well-established line of work on LMs as psycholinguistic subjects (Futrell et al. 2019; Ettinger 2020; Linzen and Baroni 2021, inter alia). A large portion of this work focuses on probing LMs for sensitivity to the well-formedness of sentences containing various syntactic structures such as subject-verb agreement (Linzen et al., 2016), relative clauses (Gulordava et al. 2018; Ravfogel et al. 2021), and filler-gap dependencies (Wilcox et al., 2018), among others. A closely-related work by Papadimitriou et al. (2022) investigates how BERT classifies grammatical role of entities in non-prototypical syntactic positions, similar to our setup in Experiment 3.

There have also been works on evaluating and probing LMs for semantic/pragmatic knowledge. Ettinger (2020) created a suite of tests drawn from human language experiments to evaluate commonsense reasoning, event knowledge, and negation. The COGS challenge (Kim and Linzen, 2020), which contains related tests to ours with regards to argument alternation, tests for whether LMs can learn to generalize about passivization and unnacusative-transitive alternations in English. Misra et al. (2022) test LMs for their ability to attribute properties to concepts and further test property inheritance. With regards to lexical semantics, Vulić et al. (2020) investigate how type-level lexical information from words in context is stored in models across six typologically diverse languages.

However, our work is distinct from both previous syntax- and semantics-focused probing and evaluation in its focus on the interactions between the aspects of meaning in individual lexical items with larger syntactic structures or constructions. Nevertheless, methodologies from these research areas have informed the construction of our experiments. Our use of minimal pairs to form sentences with contrasting semantic roles is similar to the construction of the BLiMP dataset (Warstadt et al., 2020) and other test suites. Furthermore, we treat the "agent"/"patient" labelling task as classification based on the generation probabilities of the labels, following Linzen et al. (2016)'s method of using generation probabilities for grammaticality judgements.

Another relevant recent line of work within NLP is inspired by Construction Grammar (CxG), a branch of theories within cognitive linguistics that posits that *constructions*—defined as form-meaning pairings—are the basic building blocks of language (Goldberg 1995; Croft 2001, inter alia). Mahowald (2023) conducted a similar prompting experiment on the English Article-Adjective-Numeral-Noun construction, though this was focused on grammaticality judgements as opposed to aspects of semantics. Weissweiler et al. (2022) probe for both syntactic and semantic understanding of the English comparative correlative. Our study differs in that we analyze the impact of individual lexical items in what otherwise appears to be an identical syntactic construction, as opposed to analyzing competence of the construction as a whole. Finally, Li et al. (2022) find that sentences sharing the same argument structure constructions (ASCs) are closer in the embedding space than those sharing the main verb; in light of our results, an interesting direction would be to see if sentences of the same surface construction may cluster based on finer-grained semantic distinctions.

One consequence of our work—specifically with regards to davinci-003's extremely high correlation with human judgements—is the potential for LMs as a tool for discovery in theoretical linguistics. This also has been argued recently by Petersen and Potts (2022), who demonstrate this in the realm of lexical semantics through a case study of the English verb *break*.

6 Conclusion

In order to gain insight into the behavior of LMs with respect to the syntax-semantics interface, we created a suite of prompting experiments focusing on agentivity. We prompt varying sizes of BLOOM, GPT-2, and GPT-3 to see if they are sensitive to aspects of agentivity at the lexical level, and then to see if they can either utilize or discard these word-level priors given the appropriate syntactic context. GPT-3 davinci-003 performs exceptionally well in all three of our experiments—outperforming all other models tested by far—and is even better correlated with human judgements than some proxy corpus statistics. We find it surprising that davinci-003 is able to capture an abstract notion of agentivity extremely well, but this ability does not appear to come from the size of the model alone as performance does not increase monotonically across any of the model families tested. What aspects of model training/data contribute to davinci-003's (or other models') performance on linguistic tasks may be an interesting area for future work.

Furthermore, a qualitative analysis of what davinci-003 gets incorrect reveals examples involving a number of linguistic confounders that make them more ambiguous to humans as well. The model's ability to "pick out" these linguistically interesting examples, combined with the high correlation with human ratings in Experiment 1, showcases the potential of LMs as tools for linguistic discovery for new phenomena, such as finding new classes of words or syntactic constructions that behave in unexpected ways. We hope these results encourage a more lively discussion between NLP researchers and linguists to unlock the potential of LMs as tools for theoretical linguistics research.

7 Limitations

While the use of a particular subset of English transitive verbs allows us to have precise control over the surface forms we are evaluating LMs on, this restricts our scope to a specific alternation in one language as well as a relatively small evaluation set. Nevertheless, we hope the methodology presented in this work can be extended to other phenomena across languages.

Additionally, while we explored a variety of ways to prompt these models, it may be the case that the prompt is non-optimal and therefore does not elicit the best possible output with respect to the task. Furthermore, the "prompt" to elicit human judgements is not the same as the prompt given to models, nor are the output formats (humans are asked to respond on a discrete scale from 1-5, while models are evaluated by their label log likelihoods). Evaluating whether the methodology in this line of work is a fair comparison between models and humans may be an interesting direction for future work.

8 Acknowledgements

This work was supported by NSF CISE RI grant number 2211951. For helpful comments and suggestions, we thank Kyle Mahowald and John Beavers, as well as the three anonymous reviewers and meta-reviewer. Finally, we would like to thank Thomas Lu, Saujas Vaduguru, Leilani Zhang, Russell Emerine, Vijay Viswanathan, Jeremiah Milbauer, Qianli Ma, Alex Wilf, Amanda Bertsch, Sireesh Gururaja, Yingshan Chang, Clara Na, Zhiruo Wang, Cathy Jiao, Simran Khanuja, Atharva Kulkarni, Anubha Kabra, Leena Mathur, and Shuyan Zhou for their help in annotating nouns in our dataset.

References

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.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT press.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Cleo Condoravdi. 1989. The middle: Where semantics and morphology meet. In *MIT Working Papers in Linguistics 11*, pages 16–31. MIT Press.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Thomas Ernst. 2001. *The syntax of adjuncts*, volume 96. Cambridge University Press.

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

Sarah M. B. Fagan. 1992. *The Syntax and Semantics of Middle Constructions*. Cambridge University Press, Cambridge.

Charles J Fillmore. 1986. Pragmatically controlled zero anaphora. In *Annual Meeting of the Berkeley Linguistics Society*, volume 12, pages 95–107.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Brendan S Gillon. 2012. Implicit complements: a dilemma for model theoretic semantics. *Linguistics and Philosophy*, 35:313–359.

Adele E.. Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Ray S Jackendoff. 1972. Semantic interpretation in generative grammar.

Osvaldo A. Jaeggli. 1986. Passive. *Linguistic Inquiry*, 17:587–622.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jerrold J Katz and Jerry A Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *CoRR*, abs/2104.08786.

Kyle Mahowald. 2023. A discerning several thousand judgments: GPT-3 rates the Article + Adjective + Numeral + Noun construction. *arXiv preprint arXiv:2301.12564*.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. 2022. COMPS: Conceptual minimal pair sentences for testing property knowledge and inheritance in pre-trained language models. *arXiv preprint arXiv:2210.01963*.

Ulrike Padó. 2007. The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying arguments, BERT doesn't care about word order...except when it matters. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 203–205, online. Association for Computational Linguistics.

Erika Petersen and Christopher Potts. 2022. Lexical semantics with large language models: A case study of English *break*. Ms., Stanford University.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Lilia Rissman and Asifa Majid. 2019. Thematic roles: Core knowledge or linguistic construct? *Psychonomic bulletin & review*, 26(6):1850–1869.

David E Rumelhart and James L McClelland. 1986. On learning the past tenses of English verbs.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 99–105, Berlin, Germany. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Michael Silverstein. 1976. Shifters, linguistic categories, and cultural description. *Meaning in anthropology*.

Michael K. Tanenhaus, Greg Carlson, and John C. Trueswell. 1989. The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes*, 4(3-4):SI211–SI234.

Carol Tenny. 1992. The aspectual interface hypothesis. pages 490–508. CSLI Publications, Stanford.

Carol Tenny. 1994. *Aspectual Roles and the Syntax-Semantic Interface*. Kluwer, Dordrecht.

J.C. Trueswell, M.K. Tanenhaus, and S.M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33(3):285–318.

Jeanne van Oosten. 1977. Subjects and agenthood in English. In *CLS 13*, pages 451–471.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. *arXiv preprint arXiv:2210.13181*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M Catherine O'Connor, and Thomas Wasow. 2004. Animacy encoding in English: Why and how. In *Proceedings of the workshop on discourse annotation*, pages 118–125.

A Noun-Verb-Adverb Combinations

verb	***sells***
nouns	patients: *toy, book, novel, magazine, hat, lotion, album, car, SUV, product, make, item, CD, drug, snack* agents: *salesman, saleswoman, businessman, businesswoman, trader, peddler, telemarketer, dealer, shopkeeper*
adverbs	*easily, well, quickly*
verb	***drives***
nouns	patients: *car, SUV, truck, convertible, vehicle, tank, bus, tractor, van* agents: *driver, person, chauffeur*
adverbs	*nicely, smoothly, well*
verb	***flies***
nouns	patients: *plane, kite, jet, aircraft* agents: *pilot, person, aviator, captain*
adverbs	*nicely, smoothly, well*
verb	***cooks***
nouns	patients: *mushroom, pepper, fish, salmon, tuna, fillet, vegetable, herb, meat, ingredient, steak* agents: *chef, cook, baker, caterer*
adverbs	*nicely, well, terribly*
verb	***bakes***
nouns	patients: *pizza, potato, bread, cake, pastry, dough, pie, clay* agents: *patissier, chef, cook, baker, person, confectioner*
adverbs	*nicely, well, terribly*
verb	***reads***
nouns	patients: *passage, poem, verse, line, passage, script, abstract, essay, letter, report* agents: *student, orator, person, narrator, announcer, broadcaster, teacher*
adverbs	*nicely, well*
verb	***paints***
nouns	patients: *wall, fabric, glass, canvas, wood, surface, panel* agents: *painter, artist, person, illustrator, portraitist*
adverbs	*easily, terribly, well, beautifully*
verb	***writes***
nouns	patients: *section, passage, proposal, code, essay* agents: *student, person, notetaker, journalist, scribe, doctor, professor, essayist, blogger, poet, novelist, author*
adverbs	*quickly, easily*
verb	***performs***
nouns	patients: *routine, song, choreography, sonata, concerto, scene* agents: *musician, person, actor, comedian, dancer, singer, soloist*
adverbs	*easily*
verb	***photographs***
nouns	patients: *building, animal, landscape, lake, mountain, model, view* agents: *photographer, cameraman*
adverbs	*nicely, beautifully*
verb	***plays***
nouns	patients: *cello, piano, violin, instrument, flute, clarinet* agents: *musician, violinist, cellist, pianist, drummer, flutist, clarinetist*
adverbs	*nicely, beautifully*
verb	***cuts***
nouns	patients: *meat, cardboard, packaging, board, paper, fabric* agents: *hairdresser, barber, butcher, chef*
adverbs	*nicely, roughly, cleanly, effortlessly*
verb	***cleans***
nouns	patients: *jewelry, window, countertop, floor, surface, carpet, windshield, mirror, pot, silverware, bedding* agents: *janitor, maid, cleaner, housekeeper, busboy, waiter, waitress*
adverbs	*easily, quickly, effortlessly*
verb	***washes***
nouns	patients: *bottle, tub, shirt, car, windshield, dish, bedding, blanket, bowl* agents: *worker, maid, cleaner, busboy*
adverbs	*easily, quickly*
verb	***shaves***
nouns	patients: *beard, stubble, sideburn* agents: *barber, hairdresser*
adverbs	*neatly, nicely, smoothly*

verb	***packs***
nouns	patients: *crate, lunchbox, basket, container, coat, jacket, bag, duffle, food, suitcase, tent, backpack* agents: *mover, traveller, clerk, worker, backpacker, roadtripper, hiker, camper*
adverbs	*well, easily*
verb	***stitches***
nouns	patients: *silk, quilt, cotton, cut, cloth, fabric, wound* agents: *surgeon, tailor, machine, upholsterer, dressmaker*
adverbs	*easily, smoothly, nicely, poorly*
verb	***embroiders***
nouns	patients: *cushion, thread, cloth, fabric* agents: *tailor, seamster, seamstress*
adverbs	*well, nicely, beautifully, poorly*
verb	***knits***
nouns	patients: *yarn, wool, pattern* agents: *person, lady, man, woman*
adverbs	*well, nicely, beautifully, poorly, easily*
verb	***sews***
nouns	patients: *fabric, material* agents: *tailor, seamster, machine*
adverbs	*well, nicely, beautifully, poorly*
verb	***turns***
nouns	patients: *screw, knob, car, bike, motorcycle, valve, handle* agents: *driver, racer, motorist, pilot*
adverbs	*smoothly, easily, nicely, roughly*
verb	***carves***
nouns	patients: *pumpkin, wood, stone, gem, ice, steak, turkey* agents: *sculptor, person, jeweler, artisan, carver*
adverbs	*beautifully, nicely, cleanly, flawlessly*
verb	***sculpts***
nouns	patients: *wood, stone, marble, ice, clay* agents: *sculptor, person, potter, mason, carver*
adverbs	*beautifully, nicely, cleanly*

B Data Curation Criteria

After collecting a list of optionally transitive verbs that appear as intransitive via object drop (agent subject) or object promotion in the form of the middle construction (patient subject), we then had to curate adverbs and nouns that work in the templates as described in Table 1.

Adverbs must be manner adverbs, but they should not be *agent-oriented* adverbs (Jackendoff 1972; Ernst 2001) that express the mental state of the agent. Examples of such adverbs include *furiously, happily, angrily*, etc.

Then for each verb and a list of adverbs for each verb, we come up with a list of patient and agent nouns. All of the nouns must work in intransitive and transitive templates using the same sense of the verb. For nouns added as patients in the intransitive, the noun must not be an entity that causes the event described by the verb. Furthermore, it should not be necessarily oblique in the transitive form. In the example below, *needle* cannot be the direct object of the transitive and can only appear in the *with* prepositional phrase, so we do not include it in the list of nouns:

(4) a. This needle sews easily.

b. The tailor sews easily with this needle.

c. *The tailor sews this needle easily.

For nouns added as agents, in the intransitive it must be clear that the noun is the one doing the action. For human agents, we try to add agents that are most closely associated to the action described for the event, especially with those that tend to take human direct objects in the transitive form, such as *shave*.

C Human Annotation Details

We had 19 human annotators rate all 233 unique nouns on Google Forms. Each annotator saw a different random order of the nouns and were presented with 10 nouns on each page of the form, though they could go back to alter previous responses. All annotators are fluent in English. Annotators were also asked to self-identify as native or non-native speakers; 14 of 19 consider themselves native speakers.

For nouns that have multiple common and highly distinct word senses, we gave annotators a short disambiguating description. This description does not contain any verbs or any other indicator for what types of events the entity may occur in. A list of these nouns with their disambiguating description is given in Table 5.

Noun	Description
make	product of a particular company, such as of a car
plane	airplane
kite	a light frame covered with paper, cloth, or plastic, often with a stabilizing tail
jet	aircraft
line	of a text/a poem/etc.
passage	of a text/an essay/etc.
panel	of wood/a hard surface/etc.
model	person
routine	a part of an entertainment act
board	a long, thin, flat piece of wood or other hard material
letter	a sheet of paper with words on it in an envelope
proposal	a formal plan or suggestion
turkey	meat

Table 5: Nouns and disambiguating descriptions given to annotators.

C.1 Instructions provided to annotators

An **agent** is something that initiates an action, possibly with some degree of volition. In other words, nouns that tend to be agents have a tendency to do things.

A **patient** is something that undergoes an action and often experiences a change. In other words, nouns that tend to be patients have a tendency to have things done to it.

In this form, you are tasked to annotate how "agentive" you think a noun typically is—in other words, how likely it is to be an agent or a patient when an action involving both an agent and a patient occur.

Ex: The plant was watered by John.
The plant = patient
John = agent

Ex: The sun burns John.
The sun = agent
John = patient

A more formal definition is given by Dowty (1991), who outlines contributing properties of agents and patients:

(1) Contributing properties for the Agent Proto-Role:

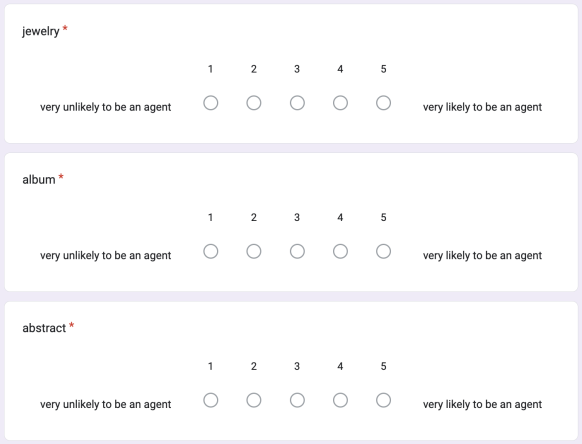
- volitional involvement in the event or state — sentience (and/or perception)
- causing an event or change of state in another participant
- movement (relative to the position of another participant)
- (exists independently of the event named by the verb)

(2) Contributing properties for the Patient Proto-Role:

- undergoes change of state
- incremental theme (something that changes incrementally over the course of an event)
- causally affected by another participant
- stationary relative to movement of another participant
- (does not exist independently of the event, or not at all)

For the sake of simplicity, disregard events described by reflexives (such as John shaved himself). For each of the following nouns, rate it on the following scale:

1 = very unlikely to be an agent
2 = somewhat unlikely to be an agent
3 = no preference between agent and patient
4 = somewhat likely to be an agent
5 = very likely to be an agent



jewelry *

	1	2	3	4	5	
very unlikely to be an agent	○	○	○	○	○	very likely to be an agent

album *

	1	2	3	4	5	
very unlikely to be an agent	○	○	○	○	○	very likely to be an agent

abstract *

	1	2	3	4	5	
very unlikely to be an agent	○	○	○	○	○	very likely to be an agent

Figure 6: Example of Google Form question format given to annotators.

Model	APAP	PAPA	δ
BLOOM `560m`	0.566	0.531	0.036
BLOOM `1b1`	0.384	0.365	0.019
BLOOM `1b7`	0.308	0.371	0.062
BLOOM `3b`	0.476	0.133	0.343
BLOOM `7b1`	-0.118	0.150	0.268
GPT-2 `small`	0.648	0.652	0.004
GPT-2 `medium`	0.420	0.367	0.053
GPT-2 `large`	0.501	0.496	0.005
GPT-2 `xl`	0.486	0.231	0.255
GPT-3 `ada-001`	0.589	0.598	0.009
GPT-3 `babbage-001`	0.394	0.228	0.166
GPT-3 `curie-001`	0.418	-0.204	0.622
GPT-3 `davinci-001`	0.579	0.356	0.223
GPT-3 `davinci-003`	**0.934**	**0.943**	0.010

Table 6: **Experiment 1**: Correlation between the difference in log-likelihood of predicting "agent" or "patient" with human ratings for nouns in isolation in both example orderings.

Model	APAP	PAPA	δ
BLOOM `560m`	0.214	0.219	0.005
BLOOM `1b1`	-0.096	0.027	0.124
BLOOM `1b7`	0.618	0.795	0.177
BLOOM `3b`	0.049	0.512	0.463
BLOOM `7b1`	0.050	0.272	0.223
GPT-2 `small`	0.658	0.190	0.468
GPT-2 `medium`	0.546	0.500	0.047
GPT-2 `large`	0.632	0.586	0.045
GPT-2 `xl`	0.484	0.531	0.047
GPT-3 `ada-001`	0.574	0.417	0.157
GPT-3 `babbage-001`	-0.030	-0.322	0.292
GPT-3 `curie-001`	0.045	0.266	0.221
GPT-3 `davinci-001`	0.673	0.622	0.051
GPT-3 `davinci-003`	**0.927**	**0.911**	0.017

Table 7: **Experiment 2**: Correlation between the difference in log-likelihood of predicting "agent" or "patient" with human ratings for nouns in intransitive sentences in both example orderings.

D Results by Example Order

Tables 6, 7, and 8 show performance in both APAP and PAPA orderings in Experiments 1 (nouns in isolation), 2 (nouns in intransitive sentences), and 3 (nouns in transitive sentences) respectively. For simplicity, we only report correlations with human judgements.

Both GPT-3 `davinci-001` and `davinci-003` are very robust to changes in example ordering for all three experiments, as are BLOOM `560m` and `1b1`. The three largest BLOOM models are remarkably sensitive to ordering, especially in Experiment 3, as are GPT-2 `xl` and GPT-3 `curie-001` and `babbage-001`.

E Propbank Statistics

When calculating model correlations with Propbank, we use all nouns with at least one occurrence of appearing within an ARG0/1 span in the parse tree to maximize the number of nouns we can compare with. However, we recognize that this may mess with correlation values since nouns with only one occurrence will have values at either 0 or 1. Furthermore, depending on the role the noun has in that particular sentence, it may push its agent rating to the opposite end of the spectrum compared to its "typical" behavior. Thus, we also tried calculating the correlation only for nouns that occur some greater number of times (within an ARG0/1 span) in Propbank. We call the minimum number of times the noun must appear the **count threshold**.

Figure 7 plots the Propbank agent ratio correlation with human ratings against the count threshold (in green). We also plot the number of nouns that meet this count threshold (in blue). The minimum count threshold to have a greater correlation than Google Syntactic Ngrams (pink line) is 27, however only 33 nouns meet this threshold. To meet meet the average human inter-annotator group correlation, the threshold is 268; only two nouns meet this.

F Adjusting Threshold for Exp 2

We also considered the possibility that the models may have a bias towards either the "agent" or "patient" label and may actually be correctly classifying nouns given an appropriate non-zero threshold for δ-LL. To account for this, we recalculate accuracies with thresholds that provide the best performance for each model as an "upper bound" for performance, as seen in Figure 8. After this adjustment, all models do at least as well as predicting the majority class, with GPT-2 `xl` experiencing the largest gain in accuracy. Nevertheless, GPT-3 `davinci-003` still outperforms all other models by far.

	trans-agent			trans-patient		
Model	**APAP**	**PAPA**	δ	**APAP**	**PAPA**	δ
BLOOM 560m	0.034	0.090	0.056	0.962	0.932	0.031
BLOOM 1b1	0.620	0.781	0.161	0.516	0.457	0.059
BLOOM 1b7	0.940	0.013	0.927	0.007	0.989	0.982
BLOOM 3b	1.000	0.059	0.941	0.000	0.895	0.895
BLOOM 7b1	0.974	0.017	0.957	0.088	1.000	0.912
GPT-2 small	0.313	0.796	0.483	0.811	0.210	0.600
GPT-2 medium	0.121	0.000	0.121	0.877	1.000	0.123
GPT-2 large	0.829	0.389	0.440	0.163	0.623	0.461
GPT-2 xl	0.978	0.001	0.977	0.018	1.000	0.982
GPT-3 ada-001	0.313	0.089	0.224	0.611	0.933	0.322
GPT-3 babbage-001	0.987	0.044	0.943	0.023	0.994	0.971
GPT-3 curie-001	0.353	0.034	0.319	0.740	0.963	0.224
GPT-3 davinci-001	0.987	0.968	0.018	0.413	0.427	0.013
GPT-3 davinci-003	0.996	0.993	0.004	0.999	0.984	0.015

Table 8: **Experiment 3**: Accuracy in both example orderings for predicting the role of the noun in transitive sentences, where **trans-agent** corresponds to the noun in subject position and **trans-patient** to object position.

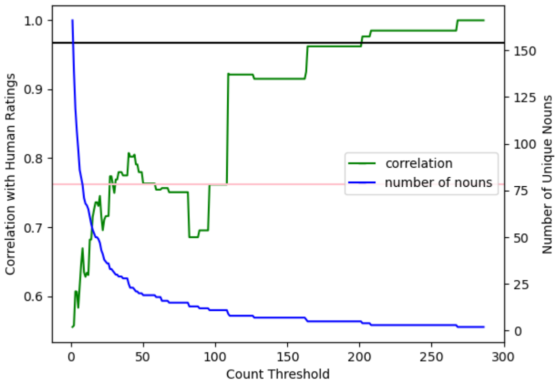


Figure 7: Count threshold versus the correlation between noun agent ratios and human ratings and the number of unique nouns that surpass the threshold. The pink horizontal line shows the correlation of Google Syntactic Ngrams with human ratings; the black line shows the average inter-annotator group correlation.

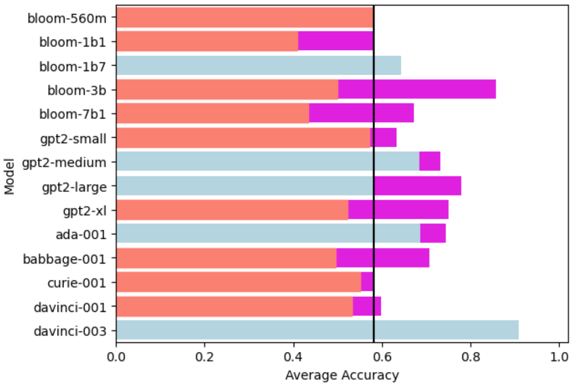


Figure 8: Average accuracy for predicting the label in **intr-agent/intr-patient** sentences with adjusted thresholds. After this adjustment, all models are at or above majority class accuracy. Magenta segments show increase in performance.