Adjudicating LLMs as PropBank Annotators

Julia Bonn*[1], Harish Tayyar Madabushi*[2], Jena D. Hwang[3], Claire Bonial[4]
[1]University of Colorado, Boulder,[2] University of Bath, [3]Allen Institute for AI,
[4]Army Research Lab
julia.bonn@colorado.edu

Abstract

We evaluate the ability of large language models (LLMs) to provide PropBank semantic role label annotations across different realizations of the same verbs in transitive, intransitive, and middle voice constructions. In order to assess the meta-linguistic capabilities of LLMs as well as their ability to glean such capabilities through in-context learning, we evaluate the models in a zero-shot setting, in a setting where it is given three examples of another verb used in transitive, intransitive, and middle voice constructions, and finally in a setting where it is given the examples as well as the correct sense and roleset information. We find that zero-shot knowledge of PropBank annotation is almost nonexistent. The largest model evaluated, GPT-4, achieves the best performance in the setting where it is given both examples and the correct roleset in the prompt, demonstrating that larger models can ascertain some meta-linguistic capabilities through in-context learning. However, even in this setting, which is simpler than the task of a human in PropBank annotation, the model achieves only 48% accuracy in marking numbered arguments correctly. To ensure transparency and reproducibility, we publicly release our dataset and model responses.

Keywords: PropBank, Semantic Role Labeling, LLM Evaluation

1. Introduction

The increasing generative power of LLMs presents ample opportunity for NLP resource practitioners to employ it for large-scale annotation efforts, which have traditionally been costly and labor intensive. Various studies have touted the promises of these large scale language models' capabilities for syntactic and semantic analyses (c.f. Tan et al. 2024; Savelka and Ashley 2023; Shin and Van Durme 2022). Other works suggest that they are still yet to achieve the type of capabilities that are needed to make them truly useful in language resource building capacities (Lu et al., 2023; Ettinger et al., 2023; Bonial and Tayyar Madabushi, 2024). In this work,[1] we empirically test the feasibility for using state-of-the-art LLMs for conducting large scale linguistic annotation, using PropBank as a test bed. More concretely, we ask, do GPT-3.5 and GPT-4, which excel in language generative capabilities, possess the ability to produce viable PropBank annotation?

Our choice of PropBank annotation as the testbed is motivated by the ways in which PropBank annotation is rooted in both syntax and semantics. Although the task of PropBank is primarily semantic role labeling, the semantic roles assigned depend upon the choice of a given relation's coarse-grained sense. Sense distinctions in PropBank were made based upon differences in semantic roles as well as syntactic behaviors—namely the subcategorization frame of a relation or the ways in which the semantic arguments are realized syntactically (e.g., as subjects, direct objects, or obliques). Thus, PropBank senses or "rolesets" reflect a set of semantic roles that are realized in a syntactically distinct way. As a result, PropBank is a powerful resource that provides explicit mappings between particular syntactic patterns of argument expression and the semantic roles of those arguments, enabling a shallow semantic analysis facilitated by clearly recognizable syntactic patterns. Given that LLMs have been touted for their abilities with respect to both syntax and semantics, we seek to test whether the mapping of syntactic constituents to particular semantic roles can be accomplished by LLMs.

The primary contribution of this research is an initial assessment of the meta-linguistic capabilities of LLMs, where we design three prompts meant to dissect LLMs' abilities with respect to the PropBank tasks of both argument annotation and sense or roleset annotation. We test LLMs' ability to accomplish roughly the equivalent task as human PropBank annotation via few-shot in-context prompting. We also test two additional settings: a more difficult setting testing LLMs' zero-shot knowledge of PropBank (no in-context examples); and an easier setting, where the LLM is provided with not only the few-shot examples but also the correct roleset with expected roles.

Our findings show that even GPT-4 (best model)

*Equal contribution

[1]Dataset and model responses available at https://github.com/H-TayyarMadabushi/Adjudicating-LLMs-as-PropBank-Annotators

5th Workshop on Designing Meaning Representations (DMR 2024) @LREC-COLING-2024, pages 112–123

Model	Setting	Match Types Exact	Core-Arg	Num-Arg
GPT-3.5	0-shot	8.6%	8.6%	17.1%
	3-shot	11.4%	17.1%	37.1%
	3-shot+rs	2.9%	2.9%	20.0%
GPT-4	0-shot	8.6%	17.1%	34.3%
	3-shot	14.3%	20.0%	42.9%
	3-shot+rs	**22.9%**	**22.9%**	**48.6%**

Table 1: We report on positive matches for GPT-3.5 and GPT-4 over three prompt settings: 0-shot, 3-shot, and 3-shot with roleset (3-shot+rs).

Construction	N	Match Types Exact	Core-Arg	Num-Arg
Transitive	14	50.0%	50.0%	85.7%
Intransitive	13	7.7%	7.7%	23.1%
Middle	8	12.5%	12.5%	25.0%

Table 2: We report the percentage of positive matches for our **best-performing prompt and model** combination: GPT-4 with the few-shot prompt that includes the correct roleset. N refers to the number of instances available for each construction.

generally struggles to assign correct semantic roles to the arguments across syntactic realizations, achieving 42.9% accuracy in the few-shot setting (Table 1). When the roleset is predefined alongside examples, the model performance does improve to 48.6%; however, that is abysmally low in comparison to the reported PropBank human average of 88.3%. As expected, the zero-shot setting is the most difficult for the LLMs (34.3%).

Furthermore, we show that GPT-4's relatively poor performance stems from its apparent inability to generalize semantics across the various syntactic realizations. The highest successes are attributed to the transitive construction (best 85.7%) where syntax maps canonically to PropBank's argument numbering (i.e. Arg0-5). For intransitive and middle voice constructions, performance drops considerably (best 25.0%).

2. Background & Motivation

2.1. PropBank Annotation

Born in the early 2000s, The Proposition Bank (PropBank) changed the world of lexical semantics in NLP by using syntactic parses as a scaffolding for the much more difficult problem of parsing meaning. The underlying idea was that English verbs exhibit patterns in the way they structure their participants both syntactically and semantically, and so by tagging syntactic arguments of a verb with semantic role labels, a system could be trained to understand fundamental propositional semantics (i.e. who did what to whom, when and how?) using syntactic cues (Palmer et al., 2005).

PropBank's main innovation was in creating a large scale inventory of rolesets (sense disambiguated predicate argument structures) for English verbs, and then having expert human annotators apply them to syntactic parse trees from the Penn TreeBank (Taylor et al., 2003). The PropBank roleset lexicon consists of verb lemmas organized into frame files. Each frame file contains one or more rolesets representing the different semantic senses associated with the verb, with each roleset providing a predicate label, a written sense definition, and a list of roles corresponding to the semantically-essential participants of the event. PropBank roles are numbered and given short written descriptions rather than more traditional thematic role labels as a way of splitting the difference between semantic and syntactic primacy of the argument. For example, Arg0s correspond to proto-agents (Dowty, 1991), which also tend to occur as syntactic subjects on verbs, and Arg1s generally correspond to proto-patients, which often occur as syntactic objects. Consider, for example, the following rolesets for the verb *deal*:[2]

Verb: deal
Roleset: deal.01 (*handle, deal with, transaction*)
ARG0: dealer (or all dealers)
ARG1: co-dealer
ARG2: subject/type of transaction
ARG3: value of transaction

Roleset: deal.02 (*play cards, distribute something*)
ARG0: distributor
ARG1: cards, thing distributed
ARG2: other player(s), distributed to

The annotation schema itself was relatively simple. For every instance of a verbal relation in a corpus sentence, annotators would first select a roleset and then tag the nodes in the parse tree governed by the verb with either a) a numbered argument from the roleset, or b) one of a small inventory of general semantic modifier args (ArgMs, e.g., ArgM-LOC (location), ArgM-DIS (discourse markers), ArgM-MNR (manners and instruments)) (Bonial et al., 2010). Annotators were presented all of the instances of a given verb lemma from the corpus as a single task, and were able to see all of the rolesets associated with that lemma in a dropdown menu (Choi et al., 2010). For each roleset, they were able to see the definition, the roles with their descriptions, and they were able to

[2]All rolesets provided in this paper are copied directly without changes from https://propbank.github.io/v3.4.0/frames/.

open a window that showed a variety of annotated example sentences.

One of PropBank's greatest successes was that, across a wide range of corpora and domains, human annotators were able to make these judgments easily and consistently. Inter-annotator agreement (IAA) was consistently high for PropBank—Bonial et al. (2017) report "exact match" (all constituents and arguments match precisely) IAA for English verbal relations at 84.8%, and "core-arg match" (numbered arguments match and ArgMs match, but the specific ArgM, such as Temporal or Locative, need not match) of 88.3%.

2.2. Related Works & Motivation

The benefits of being able to produce annotations with little training data has become an alluring prospect for resource practitioners in NLP. In the recent years, LLMs have been used to collect large-scale datasets (c.f. Liu et al. 2022; Shin et al. 2020) or to distill data to enable smaller models (c.f. Bhagavatula et al. 2022; West et al. 2021) as a means of reducing the cost burden that large-scale annotation efforts may incur. These achievements have been made possible by LLMs' capability to produce impressive generations, which have been attributed to an emergent capability to do semantic reasoning (Srivastava et al., 2023; Wei et al., 2022).

Recently, however, several works have cast scrutiny over the LLM capabilities for grasping semantic components of language and for targeted semantic analysis. Lu et al. (2023) have suggested that ability to tackle complex tasks is not necessarily emergent. Rather, models are adept at leveraging in-context learning to tackle complex tasks.[3] To refine our understanding and better delineate the parameters necessary to prompt LLMs to exhibit complex analytical abilities, we undertake experiments employing prompts both with and without illustrative examples. These experiments aim to establish the optimal prompt format conducive to eliciting LLM abilities that enable us to solve metalinguistic tasks such as this, while also serving as a method for exploring the capabilities of and limitations of LLMs.

In terms of the level of semantic analysis LLMs are able to accomplish, some research shows that larger LLMs are able to sort sentences by semantic similarity based on constructional semantics (e.g., grouping together *She blinked the tears off of her eyelashes* and *She wiped the flour off of the table*), while smaller LLMs are only able to sort sentences by lexical semantics (e.g., grouping together *blink, cough, breathe* regardless of their broader constructional setting) (Li et al., 2022). However, recent research suggests that even the largest models (GPT-4) are unable to recognize the semantic similarity of events expressed in argument structure constructions (Goldberg, 2003), such as the resultative, when non-canonical verbs are found in these constructions (e.g., *He yelled himself hoarse* as opposed to *He made himself hoarse by yelling*) (Bonial and Tayyar Madabushi, 2024). Even if LLMs are able to group some sentences by semantic and constructional similarity, there is evidence suggesting that the models are not able to infer the appropriate semantics from constructions such as *The more I study it, the less I understand it* (Weissweiler et al., 2022).

Wilson et al. (2023) evaluate the extent to which models in the BERT family are able to generalize different types of linguistic knowledge, including what they call "Type 2 knowledge," which allows speakers to predict word occurrences in new, structurally related contexts they have not explicitly encountered before, based on their understanding of how thematic roles are typically assigned across different grammatical structures. The authors use fine-tuning and introduce novel tokens in a fixed structural context to evaluate the extent to which pre-trained language models generalize to Type 2 knowledge. The authors find that PLMs can generalize to Type 2 knowledge only to a very small extent, and do not generalize across active and passive sentences. While these results are certainly relevant to our own research question, we emphasize that we are testing much larger models, where research has suggested distinct potential for in-context learning (Wei et al., 2023).

Moreover, Ettinger et al. (2023) have shown that LLMs can readily achieve surface level semantic analysis such as locating the main predicate and its core arguments (i.e. retrieving the "who-did-what-to-whom"). However, when tasked to capture a more complex semantic analysis as required by the structured AMR framework, the models fail miserably even when presented with a diverse set of in-context examples. Thus, in this work we turn to PropBank, which provides a relatively simple semantic annotation framework revolving around identifying the who-did-what-to-whom information of a verb, which may be a more reasonable level of semantic decomposition for LLMs to grasp.

However, despite the simplicity of the PropBank framework, we also recognize the annotation demands a level of comprehension beyond that of mere pattern recognition. It requires the comprehension of the elements of the sentence and their associated forms. Thus, we hypothesize that the effectiveness of LLMs on this task is likely to be limited, especially due to the complexity of this task which requires a certain "understanding" of the meaning of sentences. This is especially likely

[3] In-context learning refers to the capability of LLMs to perform tasks based on minimal examples.

given the propensity of LLMs to generate linguistically fluent, but factually or logically inconsistent sentences (Rawte et al., 2023).

In light of the evolving discourse surrounding LLMs and their capabilities, this work aims to explore the utility of LLMs in generating PropBank annotations. Specifically, we aim to answer the following research questions: a) How effective are LLMs at generating PropBank annotations, and b) What is the most effective way of prompting LLMs for the purpose of PropBank annotation?

3. Evaluation Framework

3.1. Verb and Construction Targets

To capture a wide variety of semantic and syntactic realizations, we select 7 verbs from 7 distinct VerbNet Classes (Schuler, 2005) for the evaluation and analysis of LLM capability for PropBank annotation. The verbs are listed in Table 3. While PropBank annotations are inclusive of both verb and non-verbal relations (e.g., `pitch.04` serves both *the White House **pitch*** and *the proposal **pitched** by the White House*), for the purposes of this work, we specifically focus only on verbal relations.

These verbs are selected on the basis of their ability to participate in three syntactic realizations (henceforth, constructions): transitive, intransitive, and middle voice. These constructions map semantic arguments to their syntactic element quite distinctly. As such, they allow us to evaluate if LLMs can appropriately assign what are generally Arg0 prototypical agents and Arg1 prototypical patients to the correct arguments, despite these fundamental constructional differences. For example, in intransitive realizations, the subjects may be animate Agents or Causes (e.g., *John writes well*), but we may also see inanimate Patients undergoing a change of state (e.g., *The chair broke*). In the middle voice, it is the Theme or Patient that sits in the subject position with the Agent unmentioned (e.g., *the cards deal smoothly*). Further details on the data collection and distinction between intransitive and middle voice can be found in Appendix A. Thus, each evaluation instance requires the model to cue on both the syntactic and lexical semantic information to determine whether it is Arg0 or Arg1 that likely sits in the subject position. From PropBank IAA, we know that human annotators can easily track these alternations. In this work, we investigate whether the models can do so as well.

3.2. Evaluation Set and Data Source

For compiling our exploratory corpus for evaluating LLMs, we leverage the Corpus of Contemporary American English (COCA) (Davies, 2008), which enables targeted search for particular verb in the syntactic realizations of our interest. As COCA does not furnish PropBank annotations, the extracted sentences are annotated for verbal relation targets by three of the authors previously trained extensively in PropBank annotation standards.

From COCA, we extract sentences for each of the 7 verbs with 5 usages per verb (aiming for 2 transitive, 2 intransitive, and 1 middle voice construction) resulting in a total of 35 sentences in the evaluation set. Additionally we extract 3 instances corresponding to the three constructions for in-context examples used in our few-shot setting. Further details are included in Appendix A.

The purpose of this annotated dataset is an initial exploration of LLM capabilities; it is not large enough to serve as a full diagnostic evaluation set. Although we considered leveraging some of the existing PropBank corpus annotations, we opted to annotate new sentences not included in any past PropBank release to avoid the possibility that the model's training data included the existing annotated corpora.

3.3. Models & Prompting Strategies

The capabilities of LLMs are inherently determined by the extent of their training and the scale of their parameters. As such, in assessing the proficiency of LLMs as effective PropBank annotators, our analysis centers on two prominent and powerful language models, GPT-3.5-turbo-0301 and GPT-4-0613. The experiments are conducted via the OpenAI API using a temperature setting of 0. A temperature of 0 is chosen to enforce deterministic output generation, wherein the models select the most probable next token thus ensuring reproducibility. Due to the deterministic nature of our experiments, we run each of them once.

The choice of the specific prompts employed when interfacing with LLMs has been identified as a critical factor influencing their performance. We employ the following three prompting formats:

- **0-shot setting**: The model is instructed to annotate the provided sentence using PropBank annotations. This is a setting we expect to be harder than human PropBank annotation as no examples nor rolesets are made available.
- **3-shot setting**: The model is provided with 3 examples in a setting that is roughly equivalent to a human PropBank annotation set up—examples are given and, in addition to completing annotation, the annotator must decide on the roleset.
- **3-shot roleset setting** (3-shot+rs): Along with the examples, the model is provided with the roleset associated with the input sentence. This setting is easier than human annotation—examples and the rolesets with expected roles are provided.

Verb	VerbNet Class	Corpus Example (corresponding construction in parenthesis)
break	Break-45.1	I think you were badly cut when the chair broke under you. (intransitive)
pour	Pour-9.5	The beer pours a hazy yellow color with a huge white head. (middle)
write	Say-37.7-1	It was due to illness and the doctor wrote a letter saying I couldn't fly. (transitive)
deal	Give-13.1	I saw that dude dealing drugs. (transitive)
smell	See-30.1	Butterflies smell with their feet. (intransitive)
parse	No VN Entry	Spivak is the most gender-free pronoun that parses well in English... (middle)
rain	Weather-57	In spring and fall it rains occasionally. (intransitive)
hike	Run-51.3.2	This trail hikes through a portion of the historic area...(middle)

Table 3: We focus on 7 verbal relations for evaluation set with 5 usages for each verb for a total of 35 sentences. We use the 8th verbal relation ("hike") for in-context examples for prompting.

In addition to the specific prompt format, the exact wording of the prompt itself has been found to have an effect on the output generated by LLMs. Given the inexact nature of prompt engineering, we conduct preliminary tests focused on subjective assessments of output variations on a small number of test samples. While there could always be a more effective prompt, identifying such an optimal prompt is not straightforward. Additionally, our aim is to assess how annotators typically would interact with LLMs.

In human annotation, examples provided during annotation do not necessarily use the same verb or voice as the sentence being annotated. Thus, in selecting examples, we always use the same (static) set of examples, involving the verb *hike*, which differ from the evaluation dataset.[4] We conducted experiments using a range of prompts aimed at identifying the most effective wording and format, using a small subset of our data. The final prompt we use is shown Appendix B.

We also note that providing in-context examples allows us to evaluate models that may not have been explicitly trained with PropBank annotations. By incorporating in-context examples, we circumvent the need for models to undergo specific fine-tuning (also called instructional fine-tuning) for understanding instructions pertaining to PropBank.

3.4. Metrics for Evaluation

We use three evaluation metrics that mirror evaluation metrics used to report human IAA for PropBank annotation (see Albright et al. (2013) for a summary of metrics). Exact match represents the strictest match, while the rest are more relaxed measures.

- **Exact Match:** LLM annotation matches the manually produced, gold standard annotation with respect to constituent boundaries as well as the same role number or the same ArgM type identified for each phrases.
- **Core-Arg Match:** LLM's constituent boundaries match and have the same numbered roles labeled as the human annotation. ArgMs also match in terms of *being* argMs, although the distinctions between the individual ArgM labels is ignored. This relaxed measure allows for ArgM type differences as observed in human annotation. For example, *The paper presented **at a 2020 ACL*** could plausibly be marked as either ArgM-TMP or ArgM-LOC.
- **Number-Arg Match:** LLM and human annotations are matched with respect to the heads of argument phrases (correct participant is identified, ignoring precise constituent boundaries), and with respect to numbered arguments only, ignoring ArgM annotations. Here, we are are primarily interested in the correct assignment of Arg0 and Arg1 despite syntactic differences in their realization or their omission.[5]

4. Results

Here we report results for both GPT-3.5 performance and GPT-4 performance across our three different prompt settings: zero-shot, 3-shot without the roleset information given, and 3-shot with the correct roleset given in the prompt. In Table 1, we report the percentage of positive matches across each of the match types from strictest to loosest: Exact, Core-Arg, and Numbered-Arg match. In the sections to follow, we discuss and provide match

[4]While the *hike* examples are provided in transitive, intransitive, and middle constructions, we acknowledge that there may be an effect of using a single verb across the few-shot examples. We provide a follow-on experiment in Section 5 that examines results where the prompt verb and voice match the test usage.

[5]The roleset specification in the 3-shot+rs setting includes the description of the core arguments only, without reference to the various ArgMs that PropBank allows. Thus, outside of the ArgMs included in examples in the few-shot setting, the model is given no guidance on ArgM annotation, whereas human PropBank annotators would be trained to identify ArgM types. Number-Arg Match is designed to assess the generation without unfairly penalizing the model for mistakes in ArgM.

and error examples for each prompt setting and finally for the different sentence types (transitive, intransitive, middle voice).

4.1. Zero-Shot Setting

In the zero-shot setting, we prompt the model, "Given the following verb and sentence, produce a PropBank annotation of the verb sense and its arguments." In this setting, we provide only the target verb and sentence, we do not provide potential rolesets or the correct roleset. With this prompt, GPT-3.5 is only able to provide exact and core-arg matches for two relatively straightforward transitive sentences:

1. (This reporter)-ARG0 smells-REL (another Emmy)-ARG1[6]
2. (A fringe of activists)-ARG0 broke-REL (some doors and windows of the halls)-ARG1 and committed two minor assaults.

The numbered-arg matches that GPT-3.5 is able to obtain are also largely (5 of 7 matches) of the transitive type.

GPT-4 performs much better than GPT-3.5 in the zero shot setting. GPT-4 matches on the same transitive sentences that GPT-3.5 was able to match in this setting, and it is also able to provide core-arg matches for 2 intransitives and 2 middle voice usages, including, for example:

3. GPT-4 Annotation: Use the heavy floss because (the fine floss)-ARG1 breaks-REL (easily)-ARGM-ADVERBIAL
4. GPT-4 Annotation: (This fellow)-ARG0 writes-REL (abominably)-ARGM-ADVERBIAL

Note that the above sentences are only core-arg matches, as opposed to exact matches, due to differences in the specific ArgMs marked. The gold standard marks what was annotated by GPT-4 as ArgM-Adverbial instead as ArgM-Manner. Interestingly, as we describe in the next section, GPT-4 is not able to correctly annotate the above sentences in the few-shot setting.

4.2. Few-Shot Setting

In the few-shot setting, we prompt the model in the same way, but we also provide three example annotations that all use the verb *hike*, exemplified in transitive, intransitive, and middle voice constructions. We then provide the target verb and sentence. We do not provide any information regarding the relevant roleset. Thus, this setting is very similar to what a human annotator would face, as they would not have seen the particular target annotation instance before, though they may have seen variety of other PropBank annotation examples during their training. Note that some generalization is required in moving from the examples of a different verb and the alternations in Arg0 and Arg1 seen for that verb, and the parallel syntactic alterations for the target verb.

While both GPT-3.5 and GPT-4 show improvement in this setting, the improvement is not as straightforward as one might expect. Specifically, the gains are made primarily with respect to superior annotation of transitive usages. For example, GPT-4 in particular fails to match on the middle and intransitive sentences (3) and (4) above by shifting the Arg1, *the fine floss*, to an Arg0, while also shifting the manner adjunct, *abominably* to an Arg1. We hypothesize therefore that adding the examples for comparison causes the model to overgeneralize where numbered arguments should be used, and specifically where Arg0 should be used, perhaps given that most of the *hike* examples involve an Arg0 subject.

4.3. Simplified Annotation Task in Few-Shot Setting

In the final prompt setting we provide the most information, simplifying the annotation task by including the correct roleset in the prompt. Thus, in addition to examples of how argument numbers are applied across the three constructions from the verb *hike*, we also describe explicitly how the argument numbers map to the semantic roles, expressed in natural language (as opposed to traditional thematic role labels), for the target verb.

We find that GPT-3.5 performs worse in this setting, with the numbered-arg matches falling from 37.1% in the few-shot setting to 20.0% when we now provide the roleset. When we examine where new errors were introduced in this setting, we find that example (2), which was consistently annotated correctly in the zero-shot and few-shot (without the roleset) settings, is no longer annotated correctly. Instead, the model over-extends the application of the numbered arguments specified in the roleset (see Figure 1), which was provided to the model .

5. (A fringe of activists)-ARG0 broke-REL (some doors and windows of the halls)-ARG1 (away from the halls)-ARG4

Note that the phrase GPT-3.5 assigns as the Arg4 (thing broken away from) is not present in the original sentence. The model adds this to the annotation despite explicit prompting to only use the words found in the sentence. Similarly, the inclusion of the roleset seems to have entirely derailed GPT-3.5's annotation, resulting in particularly

[6]We use this notation to express the gold standard annotation, we did not expect or require the LLMs to output in this format.

Verb: break
Roleset: break.01 (break, cause to not be whole)
ARG0: breaker
ARG1: thing broken
ARG2: instrument
ARG3: pieces
ARG4: arg1 broken away from what?

Verb: smell
Roleset: smell.02 (emit an odor)
ARG1: stinky thing
ARG2: attribute of arg1

Figure 1: PropBank rolesets *break.01* and *smell.02*

widespread (and incorrect) application of the numbered arguments:

6. GOLD annotation: I think you were badly cut when (the chair)-ARG1 broke-REL (under you)-ARGM-LOCATION

7. GPT-3.5 annotation: I think you were (badly)-ARG3 cut when (the chair)-ARG0 (broke)-REL under (you)-ARG1

GPT-4, in contrast, achieves the best performance in this setting for all match types, with a best result of 48.6% numbered-arg matches overall. Notably, most of this improvement comes in adding matches for the intransitive and middle voice usages, for example, achieving an exact match (whereas no other settings produced any type of match) on this usage of *smell* (see Figure 1).

8. (Our guy)-ARG1 smells-REL (incredible)-ARG2

Thus, we hypothesize that when the mapping from the roleset to the usage in question is particularly simple and clear, the model is able to precisely apply the roleset information. However, we acknowledge that it cannot handle cases beyond the simple with much success.

4.4. Constituent Matching

A key difference between our prompt setup and the information presented to human annotators is that humans are asked to place the PropBank argument labels on top of Penn TreeBank constituency parses (Marcus et al., 1994). The annotators are instructed place labels only on constituents that are sisters to the verb phrase (i.e. the subject) and sisters of the verb (i.e. the direct object) (Bonial et al., 2010), which is enforced by the PropBank annotation tool (Choi et al., 2010).[7] Pradhan et al. (2022) attribute some of the high IAA to the fact that the placement of annotations is clearly constrained by the syntactic tree.

[7]This training follows generative assumptions that the verbal relation assigns theta roles to its arguments, and that its arguments appear in these positions and only these positions.

Verb: rain
Roleset: rain.01 (rain)
ARG0: metaphorical agent
ARG1: metaphorical rain
ARG2: rained upon

Figure 2: PropBank rolesets *rain.01*

In our prompting experiments, we do not provide the syntactic tree corresponding to the sentence. Thus, in this section we explore the extent to which our best-performing model, GPT-4, is able to provide constituent matches with the gold standard. A constituent match is based solely on what phrases are treated as annotated arguments, where the argument labels themselves are entirely ignored. We find that in the zero-shot setting, GPT-4 obtains positive constituent matches in 42.9% of the annotations. In the 3-shot setting where no roleset is given, constituent matches are made for 51.4% of the sentences. Finally, in the 3-shot setting where the roleset is given, constituent matches drop slightly to 48.6%. The fact that constituent matches are hovering around 50% is a trend that suggests that constituent matching is likely a large source of annotation error.

To gain a sense of what the constituent mismatches look like, consider the following example, given the following roleset for *rain* (Figure 2):

9. GOLD annotation: On days (when)-ARGM-TEMPORAL (it)-ARG0 rains-REL (nonstop)-ARGM-TEMPORAL, they throw sheets of plastic over their hung wash.

10. GPT-4 annotation: (On days when)-ARGM-TEMPORAL (it)-ARG0 rains-REL (nonstop)-ARGM-ADVERBIAL, they throw sheets of plastic over (their hung wash)-ARG2

Note that *their hung wash* is what might have been rained upon, had they not thrown sheets of plastic over it. Thus, while there may be some plausible justification for calling this Arg2, it is not in a syntactic position to be considered a PropBank argument for *rain*.

The numbered-arg match type does not require constituent matches, but instead asks if the numbered arguments are assigned correctly to phrases with the same head. Thus, there are instances in our data where the constituents annotated do not match, but the annotation is assigned a numbered-arg match. Generally, these are cases where the model fails to annotate an adjunct argument altogether, or when constituent boundaries are slightly off; for example, consider the following case

of numbered-arg match that is not a constituent boundary match:

11. GOLD annotation: (The Palestinians)-ARG0 rained-REL (stones)-ARG1 (down)-ARGM-DIRECTION (onto Jews praying at the Western Wall below)-ARG2, (injuring 11)-ARGM-ADVERBIAL

12. GPT-4 annotation: (The Palestinians)-ARG0 rained-REL (stones)-ARG1 down (onto Jews praying at the Western Wall below, injuring 11)-ARG2

4.5. Trends Across Transitive, Intransitive, Middle

A key research question in this evaluation is whether or not LLMs can act as PropBank annotators, where the most critical aspect of the annotation is correctly assigning argument numbers across different syntactic realizations of the same relation. Thus, in this section, we focus on performance across transitive, intransitive, and middle voice constructions. Note that our evaluation includes the same 7 verbs exhibited in each of these construction types, and the few-shot examples are also one of each construction. For this analysis, we focus on our best-performing model and prompt combination—GPT-4 in the 3-shot setting with the correct roleset provided.

As we can observe in Table 2, the model achieves by far the most matches (85.7% numbered-arg matches) for transitive usages. The model can only achieve the most relaxed measure, numbered-arg match, about 25% of the time across intransitive usages (23.1 %) and middle voice usages (25.0%). Again, our dataset is small, but from this trend, we conclude that even at its best performance, GPT-4 cannot identify the same semantic roles arising in distinct syntactic realizations. Overall, we see that even for transitives, the best performing model and prompt combination achieves a core-arg match of only 50.0%. We contrast this with the human IAA reported in Bonial et al. (2017), where people achieve an exact match IAA of 84.8% for verbal relations and a core-arg match IAA of 88.3%—and those agreement rates are for verbal relations realized in a wide variety of syntactic realizations.

5. Discussion & Follow-On Experimentation

We started out our study by asking two questions regarding LLM capabilitily with respect to its (in-) ability for to perform PropBank annotation: (a) How effective are LLMs at this task, and (b) What is the most effective way of prompting LLMs for this task. Based on our results, we observe that there is little evidence of any zero-shot meta-linguistic knowledge enabling PropBank annotation. There is some evidence that the larger model can do better with more information—in-context learning is certainly required for the ability to do PropBank annotation. Specifically, we conclude that LLMs are *not* a good replacement for expert linguistic annotators in generating PropBank annotations, and the use of in-context examples is helpful in better guiding LLMs towards the kind of annotations that are more accurate.

To further validate this conclusion, we conducted additional in-context experiments: Concretely, we assessed the models' ability to correctly perform PropBank annotation when in-context examples have the same verb and voice as the target usage to be annotated. This enabled us to gauge the model's capability in a scenario with minimal variation between the in-context example and the model's requirements. Our findings consistently demonstrate that both GPT-3.5 and GPT-4 perform better on this version of the task than on the original one. In fact, we observed that providing explicit information related to the roleset helps models correctly complete the task in instances where they previously do not. Overall, these results indicate that models seem to be effective in following explicit instructions in the form of templated in-context examples, as opposed to being able to generalize from generic instructions akin to those presented to humans.

Importantly, this indicates that resources such as PropBank continue to be useful and, indeed, essential despite the effectiveness of LLMs, regardless of their size. Not only are these datasets helpful in probing the capabilities and limitations of LLMs, they are also likely to be useful in augmenting LLMs with additional capabilities including, for example, a sample-efficient and nuanced interpretation of input sentences.

6. Conclusion and Future Work

Our research indicates that while LLMs may excel at producing natural language text, they also show astonishingly poor capabilities to generalize semantically, especially when it comes to the capacity to produce meta-linguistic annotations that adhere to the annotation standards of the PropBank framework. However, we also show the utility of in-context examples and positive effect of carefully designed prompts in producing better LLM meta-linguistic generations. As PropBank and other linguistic resources remain valuable for semantic analysis, our work suggests that continued research and investment is needed in exploring how to best support in-context learning of meta-linguistic knowledge.

It's worth underscoring that the goal of this study was to assess current model capabilities to do basic meta-linguistic annotation, rather than developing methods by which we can empower models to do PropBank annotation. Our finding that models fail to perform even for manually-selected prototypical constructions with sufficiently clear prompts indicates a failure in meta-linguistic capabilities. Future works to expand on evaluation dataset size will be required to reveal the prevalence of this problem and further explorations with prompt engineering would be necessary to assess the depth of brittleness of model capabilities.

Thus, an immediate future work includes the expansion of the evaluation dataset. While the small size of this dataset is appropriate for the present work that is aimed at an initial exploration of LLM capabitilies, the expansion of this dataset would be necessary to scale up to a full diagnostic set for evaluating models. We expect that a larger evaluation set will be helpful to discover further insights, giving us the capability to make more robust generalizations with regard to model capabilities. Also, the present work was limited to GPT-3.5 and GPT-4. Future directions include expanding the evaluation over other models of varying scale and attested capabilities.

In this work, we have specifically focused on the inclusion of 3-shot and roleset information for prompting experiments. Future studies include an expansion on the prompting types and varieties to better assess and categorize the errors observed in models with the goal of providing more insightful recommendations for meta-linguistic prompting for PropBank annotation.

A broader application of this work is the possibility of leveraging LLMs for building up semantic resources for lower-resource languages with limited capacity for mass-annotation efforts like crowdsourcing. It is yet unclear what the extent of **multilingual** meta-linguistic capabilities of LLMs are. However, a wider net of experiments that include verb-argument behavior different from that of English is a compelling future direction of this research.

7. Ethical Considerations and Limitations

Dataset Size. The goal of the work was to take pulse of LLM capabilities regarding PropBank annotation for the purpose of a close-up manual analysis of the successes and mistakes the LLMs make in the annotation process. For this purpose, the size of the dataset was suitable. However, because the dataset used in this work is indeed very small, we do not recommend the set to be used as a full diagnostic evaluation set.

English Centricity. PropBank is available not only for English, but a wide number of languages and domains. PropBank lexicons and/or corpora now exist for for Chinese (Xue, 2006), Korean (Palmer et al., 2006), Arabic (Zaghouani et al., 2010), Hindi (Vaidya et al., 2013), Portuguese (Duran and Aluísio, 2012), Finnish (Haverinen et al., 2014), and Turkish (Şahin and Adalı, 2018), just to mention those we know well. This work, however, focuses on the aspects of PropBank annotation that is relevant to English only. The findings we offer may not hold for other languages.

8. Bibliographical References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Chandra Bhagavatula, Jena D Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2022. I2d2: Inductive knowledge distillation with neurologic and self-imitation. *arXiv preprint arXiv:2212.09246*.

Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*.

Claire Bonial, Kathryn Conger, Jena D Hwang, Aous Mansouri, Yahya Aseri, Julia Bonn, Timothy O'Gorman, and Martha Palmer. 2017. Current directions in english and arabic propbank. *Handbook of linguistic annotation*, pages 737–769.

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Jinho D Choi, Claire Bonial, and Martha Palmer. 2010. Propbank instance annotation guidelines using a dedicated editor, jubilee. In *LREC*. Citeseer.

Mark Davies. 2008. The corpus of contemporary american english (coca): 560 million words, 1990-present.

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

Magali Sanches Duran and Sandra Maria Aluísio. 2012. Propbank-br: a brazilian treebank annotated with semantic role labels. In *LREC*, pages 1862–1867.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "you are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48:493–531.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. *arXiv preprint arXiv:2202.12246*.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning?

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. Korean propbank. *LDC Catalog No.: LDC2006T03 ISBN*, pages 1–58563.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-Bettner, and Martha Palmer. 2022. Propbank comes of age—larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

Gözde Gül Şahin and Eşref Adalı. 2018. Annotation of semantic roles for the turkish proposition bank. *Language Resources and Evaluation*, 52:673–706.

Jaromir Savelka and Kevin D Ashley. 2023. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5417–5425, Seattle, United States. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks: Building and using parsed corpora*, pages 5–22.

Ashwini Vaidya, Martha Palmer, and Bhuvana Narasimhan. 2013. Semantic roles for nominal predicates: Building a lexical resource. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 126–131.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

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*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Nianwen Xue. 2006. A chinese semantic lexicon of senses and roles. *Language resources and evaluation*, 40:395–403.

Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised arabic propbank. In *Proceedings of the fourth linguistic annotation workshop*, pages 222–226.

9. Language Resource References

A. Data Collection Details

For each verb, we leveraged COCA search to find instances of the verbs in transitive, intransitive, and middle voice usages. This allowed us to specify, for example, expected noun phrases in both the preverbal and postverbal positions for the transitive voice, the expected subject noun phrase and generally a prepositional phrase for intransitives, and finally the expected subject noun phrase and generally a postverbal adverbial phrase for the middle voice. From the search results, we attempted to select relatively simple sentences where the target verb was the matrix verb. We selected 2 instances of both transitive and intransitive usages, where one usage was relatively concrete (e.g., *...a fringe of activists broke some doors and windows of the halls and committed two minor assaults* and one usage was more abstract (e.g., *...this number broke all records for a single registration day.* Note that because PropBank senses are relatively coarse-grained, such usages are generally classed as the same sense as their semantics are similar as is the argument structure (Bonial et al., 2010).

Finding middle voice usages was more challenging, as these are less frequent and often isolated to advertising language. If we were unable to find the target verbs in middle voice usages in COCA, we completed secondary web searches and were able to find such usages in product reviews. Given that these usages are less frequent, we included and annotated only one middle voice usage for each verb, with the exception of the verb *parse*, for which we could find only one intransitive usage but many middle voice usages. Thus, we included one intransitive and two middle voice usages in addition to two transitive usages for it.

We acknowledge that the defining criteria of both intransitive and middle voice can be challenging. Although our defining criteria may be debatable, we note that we do not necessarily believe that a mis-classification would significantly alter the findings of our primary research question here, as we were primarily searching for distinct syntactic realizations of the same verb to determine if LLMs could track the semantic roles across those distinct realizations. Middle constructions are both particularly challenging and particularly interesting as they can be syntactically identical to intransitives (e.g., *This cake cuts beautifully*), but are semantically distinct as the *cake* is not doing the *cutting*.

B. Full Prompts

We present our full prompts, where curly brackets are placeholders for instances from our 35-sentence evaluation set.

Version 0 - "zero-shot-NoRoleset" (less info than given to an annotator)
Given the following verb and sentence, produce PropBank annotations of the verb sense and its

arguments. Limit your annotation to the words in the sentence provided.

Annotate this:
Sentence:
Verb:

Version 1 - "3-examples-NoRoleset" (same info given to an annotator)
Given the following verb and sentence, produce PropBank annotations of the verb sense and its arguments. Limit your annotation to the words in the sentence provided.

Example 1:
Sentence: They went to India and Nepal, stayed in hostels and hiked mountains.
Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Arguments:
Arg0: They
Rel: hiked
Arg1: mountains

Example 2:
Sentence: Connor Kobal hikes regularly in Boulder Mountain Park.
Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Arguments:
Arg0: Connor Kobal
Rel: hikes
ArgM-TMP: regularly
ArgM-LOC: in Boulder Mountain Park.

Example 3
Sentence: This trail hikes through a portion of the historic area and then up to a ridge overlooking Stone Valley.
Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Arguments:
Arg1: This trail
Rel: hikes
ArgM-DIR: through a portion of the historic area and then up to a ridge overlooking Stone Valley.

Annotate this:
Sentence:
Verb:

Version 2 - 3-examples-Roleset (more info than given to an annotator)
Given the following verb and sentence, produce PropBank annotations of the verb sense and its arguments. Use the roleset information provided to produce the annotation. Limit your annotation to the words in the sentence provided.

Example 1:
Sentence: They went to India and Nepal, stayed in hostels and hiked mountains.
Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Roleset:
ARG0: causer of motion
ARG1: path of motion; location
Arguments:
Arg0: They
Rel: hiked
Arg1: mountains

Example 2:
Sentence: Connor Kobal hikes regularly in Boulder Mountain Park.
Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Roleset:
ARG0: causer of motion
ARG1: path of motion; location
Arguments:
Arg0: Connor Kobal
Rel: hikes
ArgM-TMP: regularly
ArgM-LOC: in Boulder Mountain Park.

Example 3
Sentence: This trail hikes through a portion of the historic area and then up to a ridge overlooking Stone Valley.
Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Roleset:
ARG0: causer of motion
ARG1: path of motion; location
Arguments:
Arg1: This trail
Rel: hikes
ArgM-DIR: through a portion of the historic area and then up to a ridge overlooking Stone Valley.

Annotate this:
Sentence:

Verb:
Sense:
Roleset: