An Empirical Study of Focus Identification in Large Language Models with Modal Contexts

Anonymous authors
Paper under double-blind review

Abstract

We examine whether pre-trained large language models GPT-2 and Flan-T5 use prior context to accurately identify the focus of negated sentences. We present to the models procedurally-generated negated sentences and procedurally-generated prior contexts to prime the models to generate text coherent with a particular focus in the negated sentences. We examine model accuracy with and without explicit modal operators for necessity and possibility and formalize our examples in modal logic. We find that larger models of the same class are much more likely to generate sentence continuations that cohere with the provided context, but even the larger models often struggle to outperform a random baseline with respect to accuracy. Certain necessity modals in the prior context greatly increase the likelihood of a continuation that is both relevant to the context and consistent with the primed negation focus.

1 Introduction

Despite the state-of-the-art results on many natural language processing (NLP) datasets (Devlin et al., 2019; Radford et al.; Brown et al., 2020), deep neural network (DNN)-based NLP models still struggle to fully capture negation and semantic compositionality (Hossain et al., 2020; Tang et al., 2021)—two linguistic phenomena that are necessary for humans to fully represent the meaning of a sentence (Hossain et al., 2020; Maruyama, 2021)—and the degree to which LLMs can take advantage of implicit and pragmatic information in text is not well understood.

Negation changes the meaning of a sentence by reversing the polarity of the particular constituent to which it applies. It has posed a challenge for NLP for the duration of the field's existence. The pragmatics of language, especially as represented in text, are difficult to both model and evaluate at scale, requiring carefully targeted examples that do not lead to false analyses due to confounding statistical patterns in the data.

Humans handle with great alacrity the ambiguity in both spoken and written language. In spoken English, to change the focus of a word, one can change intonation or stress, which is often lost in plain text. For example, the sentence, "Alex likes *baseball*" has different implications from "*Alex* likes baseball", in which the italicized text is stressed. But even in text, given certain prior contexts, a human reader effortlessly interprets where the focus is, even adding complex truth conditions to the semantic representation. We take advantage of the fact that the focus of the negation can be facilitated by manipulating the prior context, thereby restricting the number of plausible and felicitous immediate completions of a given sentence.

We examine whether pre-trained LLMs manage to do what humans do effortlessly: correctly identify the focus of negated sentences *based only on prior context*. We do this for contexts both with and without explicit modal operators by designing prompts for the LLMs. While modality has been neglected in much of computational linguistics (Morante & Sporleder, 2012), we formalize the truth conditions of our prompts in modal logic, which is sufficiently powerful to specify each possible interpretation. This formalization clarifies the truth conditions of the prompts and the acceptable completions and encourages researchers to think carefully about the encoded meaning.

This paper proceeds as follows: In Section 2, we survey some of the literature on the abilities of neural language models and machine translation models to represent negation and incorporate implicit information in natural language. In Section 3, we describe our experimental setup, with examples of the prompts we use and their logical forms, before discussing results in Section 4. Finally, in Section 5 we summarize some key points and observations, conclude, and suggest some avenues for future work.

2 Background and Related Work

We now review related background on modality, focus, and negation. As we examine the focus of negated sentences as determined by prior context, our work lies at the intersection of NLP and several fields of linguistics.

2.1 First-order logic and Higher-order Logic

First-order logic allows specification and inference by use of conjunction ($P \wedge Q$) disjunction ($P \vee Q$), implication, ($P \implies Q$), and negation ($\neg P$). The truth conditions of natural language sentences can be specified precisely by conversion to logical form, a tradition which goes back to at least Aristotle but was formalized by nineteenth and twentieth century logicians as propositional logic. A sentence such as "John likes Mary" can be represented as a proposition, $\textsc{likes}(\text{John}, \text{Mary})$, where the predicate LIKES takes two arguments. First-order logic (FOL) adds existential and universal quantifiers, allowing for the formal specification of words such as *some* and *all*. When applied to a variable, the existential quantifier $\exists x$ requires that there exists at least one x to which the predicates taking x as an argument apply, and the universal quantifier $\forall$ requires that the corresponding predicates apply to all x. For example, "Mary likes someone" is $\exists m \exists p[\textsc{mary}(m) \wedge \textsc{likes}(m, p) \wedge \textsc{person}(p)]$, which can be read "There exist m and p such that m is Mary, m likes, p and p is a person." From this, logical inferences can be made that must be consistent with what is explicit in the sentence.

Higher-order logics (HOL) add more quantifiers than $\exists$ and $\forall$. Second-order logic (SOL) adds quantification over predicates, which we use. In FOL, we cannot write, $\exists P$, where P is some predicate such as EATS, but SOL allows this. In SOL, we can represent "Alex does something" as $\exists a \exists P[\textsc{alex}(a) \wedge P(a)]$. Modal logic, described in Section 2.2, is another HOL.

2.2 Modal Logic

To formalize the truth conditions of the prompts in our experiments, we use modal logic with second-order quantification over predicates, which is sufficiently expressive for our immediate purposes. Modal logic has a long history of use as a semantic formalism. It extends first-order logic with two operators, *necessity* ($\Box$) and *possibility* ($\Diamond$), where $\Box P$ indicates "It is necessary that P" and $\Diamond P$ means "It is possible that P". $\Box P$ indicates that P is true in all possible worlds, while $\Diamond P$ indicates that P is true in at least one. Modals can be further classified into more precise interpretations, such as deontic and epistemic (Palmer, 2001), though other distinctions are possible (Fintel, 2006). Modality can be introduced by auxiliaries (e.g., *can*, *must*), verbs (e.g., *need*), adverbs (e.g., probably, definitely), etc. Furthermore, *covert* modality (Bhatt, 2006) is introduced implicitly, without explicit modal operators.[1]

Portner (2009) describes the utility of modal logic for formalizing modality: modal logic has been useful for clarifying the many kinds of modality, how they behave, and how they are triggered. Modality also explicitly accounts for hypothetical and counterfactual worlds, which may confound next-word predicting language models. Portner (2009) notes that "a semantic theory which does not attend to modality will be radically simpler than one which does, and so will provide a much less accurate overall picture of the nature of meaning in human language." Given its first-class place in semantics, it is a useful framework for studying the nature and behavior of the language used and produced by LLMs.

[1] See Portner (2009) or Fintel (2006) for a thorough introduction to the many types of modality.

2.3 Focus and Negation

The same can be said of focus and negation. In English, focus can be locally coerced or via intonation and stress, but also encouraged via prior context alone (Section 2.5). We are interested in whether the truth conditional meaning of the generated text is consistent with the restrictions facilitated by the prior context.

The interaction of focus and negation has received little attention in NLP. There is work on modality/negation annotation and tagging for machine translation (Baker et al., 2012) and using modality and polarity in factuality detection (Saurí & Pustejovsky, 2012) and scope resolution for speculation detection (Velldal et al., 2012; Khandelwal & Britto, 2020). Closer to our work, Hossain et al. (2020) predict the focus of negated sentences with a BiLSTM+CRF using ELMO embeddings, finding scope prediction to be very helpful; and earlier work identifies the focus of negation based on prior context (Rosenberg & Bergler, 2012; Matsuyoshi et al., 2014; Zou et al., 2014; 2015; Shen et al., 2019) using data from the *SEM 2012 shared task (Morante & Blanco, 2012). We also study focus detection, but do so implicitly in a zero-shot LLM context in a more targeted manner in which we control the context and check for consistency.

2.4 Logical Reasoning in LLMs and Linguistic Evaluation

Related to but different from our work, there has been some work on examining the ability of LLMs and LLM-derived models to generalize logical reasoning (Huang & Chang, 2022). Creswell et al. (2022) examined the ability of LLMs to perform few-shot, multi-step logical reasoning on ten tasks and engineer prompts and use fine-tuning to improve LLMs' performance on these tasks. Han et al. (2022) claims that GPT-3 only achieves slightly better than a random baseline on its first-order logic annotated dataset, while Saparov et al. (2023) examine out-of-distribution deductive reasoning and argue that LLMs can learn to generalize multi-step proofs if shown explicit examples. Chain-of-thought prompting approaches improve LLMs' ability to output text in a logically consistent way in several contexts (Wei et al., 2022; Kojima et al.). McCoy & Linzen (2018) introduce a challenge dataset for non-entailed subsequences, and Kassner & Schütze (2019) look specifically at negated sentences and linguistic priming of LLMs, in line with other work using LLMs as psycholinguistic subjects (**?**)

Though this work is relevant since we are in part concerned with the truth conditions of the LLM prompts and outputs, We are not evaluating the ability of LLMs to perform logical reasoning per se; we are using logic to formalize, evaluate, and reason about the ability of LLMs to perform a pragmatic task—to play a pragmatic game by using the prior context that ideally will prime its output—that reveals whether the base LLM correctly infers the focus of the negation and, secondarily, whether the LLM will play the game at all.

2.5 Focus-mediated Truth Conditions

In this paper, we examine how LLMs identify the focus of negated sentences by examining how they generate sentences given certain prior contexts. We now explain with an example what we are examining and how it can be formalized.

Consider the sentence *John does not like Mary* (Lee, 1985). The focus of the negation can apply to *John* (the subject), *not* (the negation itself), *like* (the verb), or *Mary* (the object).[2] The focus changes the truth conditions in complex ways, which the logical representation makes clear.

(1) John does not like Mary.
$\exists m \exists j [\textsc{mary}(m) \wedge \textsc{john}(j) \wedge \neg \textsc{likes}(j, m)]$

   a. **John** does not like Mary.
   $\exists m \exists j \exists p [\textsc{mary}(m) \wedge \textsc{john}(j) \wedge \neg \textsc{likes}(j, m) \wedge \Diamond \textsc{likes}(p, m) \wedge p \neq j)]$

   b. John **does not** like Mary.
   $\exists m \exists j [\textsc{mary}(m) \wedge \textsc{john}(j) \wedge \neg \textsc{likes}(j, m)]$

[2]This sentence has do-support, and while spoken emphasis could be placed on *does*, the resulting meaning would be the same as applying it to *not*.

c. John does not **like** Mary.
$\exists m \exists j \exists P[\text{MARY}(m) \wedge \text{JOHN}(j) \wedge \neg \text{LIKES}(j, m) \wedge \Diamond P(j, m)]$

d. John does not like **Mary**
$\exists m \exists j \exists p[\text{MARY}(m) \wedge \text{JOHN}(j) \wedge \neg \text{LIKES}(j, m) \wedge \Diamond \text{LIKES}(j, p)]$

In each variation of Example 1, the truth conditions of negation-focused sentence remain intact, but the possibility of additional truth conditions is introduced. Example 1a indicates the possible existence of someone other person p who *does* like Mary (reversing the negation of the predicate *like* when applied p and Mary). 1b is equivalent to an unmarked utterance. 1c requires the possible existence of an unknown *predicate* (transitive verb) P that applies to John and Mary, such as *love*, which may or may not subsume or contradict the original predicate. Finally, in 1d, there may exist some other person p whom John does like, canceling the negation for j John and p. Call the additional truth conditions imposed by the marked utterance proposition F. One key insight is this: while one cannot know the precise truth conditions Q—which may remove the need for the $\Diamond$ operator in F—without more context (e.g., whether *Jack* loves Mary in 1a), the felicitous Q must be consistent with F, only potentially more specific. E.g., if *Jack* loves Mary, then P is redundant since the set of worlds wherein Q is true is a proper subset of the set of worlds in which P is true.

We furthermore use the same prompts in an additional set of experiments with modal operators that can either weaken or strengthen the prior context. These are described in Section 3.1.3

3 Experiments

We now discuss experiments to test how well GPT-2 (Radford et al.) and Flan-T5 (Chung et al., 2022) model the pragmatic nuances of negation: those sentences wherein a negated sentence may apply the negation to any one of its multiple constituents.

We prompt each model with negated sentences, preceded by **prior context** and followed by one of three conjunctions: *but*, *although*, as well as the *if-then* construction. For the if-then construction, the negated sentence shown to the model begins with the word *if*, and the model must complete the sentence after the word *then*. The use of a conjunction at the end of each prompt primes the model to output a continuation of the prompt. We choose conjunctions specifically to indicate any logically coherent continuation of the input must offer an *alternative* to the negated sentence in the prompt. Thus, how the model completes the sentence after the conjunction indicates on which word the focus of the prompt must be.

Unlike Hosseini et al. (2021) and Kalouli et al. (2022), the negation we examine is not the kind applied to common knowledge or universally true or false statements. Instead, we use only sentences whose truth value depends on previously established knowledge, the prior context we provide. Presupposed, given, or implied knowledge can all support multiple interpretations of the same utterance—for example, factive verbs, whose meanings presuppose the truth value of an utterance (Erdmann, 1974; Kripke, 2009; Grissom II & Miyao, 2012).

Consider the Example 2 prompt.

(2) Sam plays something. Sam doesn't play **baseball**, but
$\exists s \exists a \exists b[\text{SAM}(s) \wedge \text{BASEBALL}(b) \wedge \Diamond \text{PLAYS}(s, a) \wedge \neg \text{PLAYS}(s, b)]$

This sentence is divided into three parts: the prior context (*Sam plays something*), the negated sentence (*Sam doesn't play baseball*), and the conjunction (*but*). The prior context introduces Sam and the fact that Sam plays something, $\exists s \exists a[\text{SAM}(s) \wedge \text{PLAY}(s, a)]$. The bolded **baseball** is the focus of negation in the negated sentence. Since *baseball* is negated, a reasonable continuation of the sentence would be something like *Sam plays* ***volleyball***, because we are working with the knowledge that a) Sam plays something (which must obviously be playable) and b) that something is not baseball.

3.1 Methodology

We use small and large versions of pretrained GPT-2 with 124 million and 1.5 billion parameters, respectively (Radford et al.), and pretrained Flan-T5 Small and XXL with 80 million and 11 billion parameters, respectively (Chung et al., 2022), to examine the effect of model size within the same model class. All prompts given to the models are generated procedurally. For each of ten intransitive verbs and ten transitive verbs , we generate an input of the form `[prior context] + [negated sentence] + [conjunction]`.

3.1.1 Generating Negated Sentences

Each negated sentence has the form `[subject] + [negation] + [optional object]`. All subjects are the names of people and all are singular.

One negated sentence is generated for each verb. To generate a negated sentence, a subject (here, a person's name) is randomly chosen from a list of fifteen names, twelve of which were generated using a name-generating website.[3] The remaining three names are manually added and are gender-neutral. After selecting a subject, we concatenate a negation construction to the sentence, randomly chosen from a list of three negation constructions (*does not*, *doesn't*, and *didn't*) to represent different past and present tense negations. Next, the verb is added to the sentence. If the verb is intransitive, the sentence generation process stops here.

If the verb is transitive, an object is added to the negated sentence. Every verb has annotation for the types of objects it can logically take. These object types fall into five categories: `sports and games`, `instruments`, `school subjects`, `food`, and `people`. Note that verbs cannot necessarily take objects of all five types; for example, the verb *eat* cannot take objects from the `sports and games` category. To select the verb's object, we randomly select one of the object categories that a verb can take and then randomly choose an object from a list of objects of that particular type. E.g., if the verb is the transitive verb *like*, it can take four types of objects: `sports and games`, `school subjects`, `people`, and `food`. To choose a specific object for the verb, an object category is randomly chosen (say, `food`), and then the program randomly chooses the object from a list of food words (e.g., vegetables).

All objects are one word to avoid potential ambiguity over which of the words in a multi-word object constituent is the focus of negation. E.g., negating the phrase *warm bread* could lead to two different interpretations that negate only one of the words: *not* ***warm*** *bread* ($\exists x[\neg\text{WARM}(x) \wedge \text{BREAD}(x)$) or *not warm* ***bread*** ($\exists x[\text{WARM}(x) \wedge \neg\text{BREAD}(x)]$). The only exception to the one-word object requirement is in the `instruments` category, where each instrument is preceded by the determiner *the* (i.e, *the* cello, *the* flute) to increase naturalness.

3.1.2 Generating Prior Context

After a negated sentence is constructed, a prior context sentence is generated for that sentence. This context establishes which word in the negated sentence is the focus of negation. Different types of prior context are generated for intransitive verbs and transitive verbs. Given a negated sentence with an intransitive verb of the form `subject + negation + intransitive verb`, two types of prior context are generated in the form shown in Example 3a and 3b. This is followed by an incomplete sentence in the form of 3c.

(3) a. Someone `[verb]`.
$\exists s[\text{PERSON}(s) \wedge \text{VERB}(p)]$

b. `[subject]` does/did something.
$\exists s \exists P[\text{PERSON}(s) \wedge P(s)]$

c. `[subject]` does not `[verb]` + `[conj]`
$\exists s[\text{PERSON}(s) \wedge \neg\text{VERB}(s)]$

As an example, given a negated sentence *Una doesn't read*, the generated prior context is either sentences in 4a or 5a.

[3] https://diversenamesgenerator.com/

(4) a. **Someone** reads.
$\exists p[\text{PERSON}(p) \wedge \text{READS}(p)]$
b. Una doesn't read.
$\exists u[\text{UNA}(u) \wedge \neg\text{READS}(u)]$
c. $\implies$ Someone other than Una reads.
$\exists u \exists p[\text{UNA}(u) \wedge \neg\text{READS}(u) \wedge \text{PERSON}(p) \wedge \text{READS}(p) \wedge u \neq p]$

(5) a. Una **does something**.
$\exists u \exists P[\text{UNA}(u) \wedge P(u)]$
b. Una doesn't read.
$\exists u[\text{UNA}(u) \wedge \neg\text{READS}(u)]$
c. $\implies$ Una does something other than reading.
$\exists u \exists P[\text{UNA}(u) \wedge \neg\text{READS}(u) \wedge P(u)]$

Taken together, the sentences in Example 4 indicate that there exists someone who reads (4a) who is not Una (4b).[4] The introduction of the named entity Una with the same predicate in 4b is felicitous when the focus in 4a was *someone*. Thus, there exists someone who is not Una who reads (4c). When paired with the negated sentence *Una doesn't read*, the focus of the unmarked sentence 4a is *Una*, while in sentence 5a it is the verb. Note that the prior context does not include any negation.

For negated sentences with transitive verbs, the subject and the object, but not the verb, can be negated. This is because it is harder to substitute a diverse set of transitive verbs with a generic verb phrase such as *do something* when the verb is enacting a unique action on the object. Given a negated sentence with a transitive verb of the form `[subject]` + `[negation]` + `[verb]` + `[object]`, the two types of prior context for the sentence are generated from the template in Example 6.

(6) a. **Someone** + `[verb]` + `[object]`.
b. `[subject]` + `[verb]`
+ **someone/something** [5]

Thus, given a negated sentence *Una doesn't like vegetables*, the prior context generated is shown in Example 7:

(7) Negated sentence: Una doesn't like vegetables.
Generated contexts:
a. **Someone** likes vegetables.
$\exists p \exists v[\text{PERSON}(p) \wedge \text{VEGETABLE}(v) \wedge \text{LIKES}(p, v)]$
b. Una likes **something**.
$\exists u \exists x[\text{UNA}(u) \wedge \text{LIKES}(u, x)]$

3.1.3 Explicitly Modal Contexts

We also generate modal variations of the prior contexts detailed above which explicitly add possibility and necessity modals to the prior context. This is motivated by the fact that adding necessity or possibility modals changes the strength of the coercion of the focus. For instance, saying "Mary might eat something" (possibility modal) has weaker priming than "Mary eats something". The prompt *It is necessary that someone reads. Una doesn't read, but* establishes not only that Una doesn't read, but that someone logically *must* read, more strongly insisting that the model infer who that person is.

Given any prior context of the from `[subject/someone]` + `[verb]`/(did/does) something + `[optional object]`, the following forms are generated:

(8) a. **It is necessary that** (subject/someone) + `[verb]`/(did/does) something + `[optional object]`
$\Box P$

[4] It's possible to coerce these sentence to shift their focus, but this is not the typical reading.
[5] The word "someone" is used to stand in for animate objects while "something" is used to stand in for inanimate objects.

b. **It is possible that** (subject/someone) + `[verb]`/(did/does) something + `[optional object]`
$\Diamond P$

c. (subject/someone) + **must/must have** `[verb]`/(do/done) something + `[optional object]`
$\Box P$

d. (subject/someone) + **should/should have** `[verb]`/(do/done) something + `[optional object]`
$\Box P$ or $\neg P$[6]

e. (subject/someone) + **has to/had to** `[verb]`/(do) something + `[optional object]`
$\Box P$

For each negated sentence generated, six types of prior contexts are created: the original prior context plus the five types of prior context incorporating possibility and necessity listed in Example 8. Each type of prior context is then concatenated to its own copy of the negated sentence, forming an almost-complete input prompt.

To form complete input prompts, for each `prior context + negated sentence` construction, three versions of it are created: one with the word *but* following the negated sentence, another with *although* following it, and a final one with *if* preceding it and *then* following it. Thus, for an incomplete input prompt of the form `prior context + negated sentence`, the following complete prompts are generated:

- `[prior context]` + `[negated sentence]` + **but**
- `[prior context]` + [negated sentence] + **although**
- `[prior context]` + **If** + `[negated sentence]` + **then**

We use *but*, *although*, and *if-then* constructions for each `prior context + negated sentence` input to prime the LLM to output a logical alternative to the negated sentence in the prior context. I.e., if the prompt to the model is *It is necessary that Una likes something. Una doesn't like vegetables, but*, the conjunction *but* signals the continuation of the prompt with an alternative to the negated statement. Perhaps Una doesn't like **vegetables**, but Una likes **fruit**. Since this scenario focuses on what Una likes—we are given that Una likes *something* but it has not yet been specified—it is felicitous and logical to follow *but* from the prompt with something specific that Una likes. It would be infelicitous, however, to follow *but* with something like *Osamu likes vegetables*, because this focuses the person doing the liking as opposed to the particular food being liked.

4 Results

In this section, we describe the results of our experiments, beginning with our evaluation method and baseline calculation, and moving to our experimental results on four large language models: two GPT-2 models and two Flan-T5 models.

4.1 Evaluation

To evaluate each language model, we manually inspect each of the 720 LLM sentence completions.[7] If the completion is consistent with the focus of the negated sentence being on the word suggested by the context, we consider the example correct; otherwise, it is marked incorrect. We calculate accuracy for each of the models.[8]

[6]The deontic sense of *should have* may suggest $\neg P$ (in the present) when followed by *but*. Deontic logics explicitly account for "ought to" cases. Some modal logics, such as the Kripke modal logic **K**, do not include the reflexivity axiom $\Box P \rightarrow P$; thus we could define this sentence as $\Box P \vee \neg P$ in these without violating logical consistency, but is a desirable property for epistemic modals. Traditionally, *should* has been analyzed as a necessity modal. A better solution would be to incorporate deontic operators to account for these.

[7]The annotator is one of the authors, a native English speaker.

[8]In an unconstrained linguistic generation setting such as this one, many "creative" prompts are possible, and determining that a generated continuation is "wrong" can sometimes be subjective. We operate under the assumption that the context we

4.2 Calculating a Baseline

To best contextualize the results, we calculate a baseline accuracy of a model whose output randomly guesses the focus of negation in the input prompt. We can calculate a baseline by simply counting the number of words on which each case can focus and calculating the probability of such a random guess. We have two cases: intransitive and transitive verbs. Overall random accuracy is

$$\begin{aligned} P(\text{correct}) &= P(\text{correct}, \text{intrans}) + P(\text{correct}, \text{trans}) \\ &= P(\text{intrans})P(\text{correct}|\text{intrans}) \\ &+ P(\text{trans})P(\text{correct}|\text{trans}), \end{aligned}$$

yielding a random baseline accuracy of 0.292. where $P(\text{correct}|\text{intransitive}) = \frac{1}{3}$ since there are three possible words of focus (the subject, negation, or verb phrase) and $P(\text{correct}|\text{transitive}) = \frac{1}{4}$ because there are four (subject, negation, verb phrase, or object). There is an equal number of intransitive and transitive prompts, so $P(\text{intransitive}) = P(\text{transitive}) = \frac{1}{2}$. Substituting these values, we have

$$P(\text{correct}) = \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{4} = 0.292. \tag{1}$$

Sentences with intransitive verbs can negate the subject, the negation cue itself, or the verb.

4.3 Language Model Results

GPT-2 XL inferred the focus of negation in line with our correctness judgments in 164 out of 720 prompts, resulting in an accuracy of 0.228. This falls below the random baseline. GPT-2 correctly inferred the focus of negation in 94 out of 720 prompts, resulting in an even lower accuracy of 0.131. Flan-T5 XXL correctly inferred the focus of negation in 257 out of 720 prompts, resulting in an accuracy of 0.357. This is slightly above the random baseline. Flan-T5 Small correctly inferred the focus of negation in 38 out of 720 prompts, resulting in an accuracy of 0.053. Outputs ranged from very natural to logically contradicting the prompt. The following output segments were both highly natural and, thus, labeled as correctly inferring the focus of negation (the model output is in bold):

(9) Boram had to play something. Boram didn't play the saxophone, but **he did play a trumpet, an oboe, and more...** (GPT-2 XL)
Context:
$\exists b \exists x[\textsc{boram}(b) \wedge \Box\textsc{play}(b, x) \wedge \neg\textsc{sax}(x)] \wedge Q$

After Completion:
$\exists t \exists o \exists y \exists b \exists x[\textsc{boram}(b) \wedge \Box\textsc{play}(b, x) \wedge \neg\textsc{sax}(x) \wedge \textsc{play}(b, t) \wedge \textsc{play}(b, o) \wedge \textsc{oboe}(o) \wedge \textsc{trumpet}(t) \wedge \textsc{play}(b, y)]$

(10) Someone must have worked. Osamu didn't work, but **there were a couple of other people who did.** (GPT-2 XL)
Context:
$\Box\exists p[\textsc{worked}(p)]$
$\wedge \exists o[\textsc{osamu}(o) \wedge \neg\textsc{worked}(o)] \wedge Q$
After Completion:
$\Box\exists p \exists o \exists q \exists r[\textsc{worked}(p) \wedge \textsc{osamu}(o) \wedge \neg\textsc{worked}(o) \wedge \textsc{worked}(q) \wedge \textsc{worked}(r)]$

(11) It is necessary that someone reads. Boram doesn't read, but **Boram's son reads.** (Flan-T5 XXL)
Context:

have given should lead to the kinds of immediate outputs we have laid out, though this may be a contestable premise in some circumstances. Despite this, our results objectively show how well each model generated outputs in line with our (subjective) expectations.

$\Box\exists p[\text{READS}(p)] \wedge \exists b[\text{BORAM}(b) \wedge \neg\text{READS}(b)] \wedge Q$
After Completion:
$\overline{\Box\exists p[\text{READS}(p)] \wedge \exists b[\exists s\text{BORAM}(b) \wedge \neg\text{READS}(b) \wedge \text{SON}(b, s) \wedge \text{READS}(s)]}$

(12) Jordan must love something. If Jordan does not love chess, then **Jordan must love something else.** (Flan-T5 XXL).
Context:
$\exists j\exists c\Box\exists x[\text{JORDAN}(j)$
$\wedge \text{LOVES}(j, x) \wedge \text{CHESS}(c) \wedge (\neg\text{LOVES}(j, c) \rightarrow Q)]$
After Completion:
$\overline{\exists j\exists c\Box\exists x[\text{JORDAN}(j)}$
$\wedge \text{LOVES}(j, x) \wedge \text{CHESS}(c)$
$\wedge (\neg\text{LOVES}(j, c) \rightarrow x \neq c)]$

(13) Jordan loves something. Jordan does not love chess, but **he loves something else**. (GPT-2)
Context:
$\exists j\exists c\exists x[\text{JORDAN}(j)$
$\wedge \text{LOVES}(j, x)] \wedge \text{CHESS}(c) \wedge \neg\text{LOVES}(j, c)] \wedge Q$
After Completion:
$\overline{\exists j\exists c\exists x[\text{JORDAN}(j) \wedge \text{LOVES}(j, x)]} \wedge \text{CHESS}(c)$
$\wedge \neg\text{LOVES}(j, c) \wedge c \neq x]$

(14) It is possible that Abdulaziz knows someone. Abdulaziz does not know Hani, but **does know Hani's son...** (GPT-2)
Context:
$\exists a\exists h\Diamond\exists p[\text{ABDULAZIZ}(a) \wedge \text{PERSON}(p) \wedge \text{HANI}(h) \wedge \neg\text{KNOWS}(a, h)] \wedge Q$
After Completion:
$\overline{\exists a\exists h\Diamond\exists p[\text{ABDULAZIZ}(a) \wedge \text{PERSON}(p) \wedge \text{HANI}(h) \wedge \neg\text{KNOWS}(a, h) \wedge \text{SON-OF}(p, h) \wedge \text{KNOWS}(a, p)]}$

(15) It is possible that someone sings. Taguhi doesn't sing, but **no one sings**. (Flan-T5 Small)
Context:
$\Diamond\exists p\exists t[\text{PERSON}(p) \wedge \text{SINGS}(p) \wedge \text{TAGUHI}(t) \wedge \neg\text{SINGS}(t)] \wedge Q$
After Completion:
$\overline{\Diamond\exists p\exists t[\text{SINGS}(p) \wedge \text{TAGUHI}(t) \wedge \neg\text{SINGS}(t)]} \wedge \neg\exists p[\text{PERSON}(p) \wedge \text{SINGS}(p)]$

(16) Boram has to do something. Boram doesn't read, although **Boram is a writer**. (Flan-T5 Small)
Context:
$\exists b\Box\exists P[\text{BORAM}(b) \wedge P(b) \wedge \neg\text{READS}(b)] \wedge Q$
After Completion:
$\overline{\exists b\Box\exists P[\text{BORAM}(b)} \wedge \neg\text{READS}(b) \wedge P(b) \wedge \text{WRITER}(b)]$

The following output segments logically contradicted the input prompt and were thus labeled as not correctly inferring the focus of negation (the model output is bolded):

(17) Hani must do something. Hani does not write, but **Hani writes.** (Flan-T5 XXL)
Contradiction: $\text{WRITES}(h) \wedge \neg\text{WRITES}(h)$

(18) It is possible that Taguhi does something. Taguhi doesn't sing, although **Taguhi does sing.** (Flan-T5 XXL)
Contradiction: $\text{SINGS}(t) \wedge \neg\text{SINGS}(t)$

(19) It is possible that Una does something. Una does not dance, but **dances** (Flan-T5 Small)
Contradiction:
$\text{DANCES}(u) \wedge \neg\text{DANCES}(u)$

The following examples have interpretations which contradict the prompt and others which do not, depending on the properties of the logic used to represent them. In Example 20, if we assume that $\forall p[\text{WRITER}(p) \rightarrow$

WRITES(p)] (where WRITER and WRITES can be any such noun-verb pair, as in Examples 22 and 24), which implies $\forall p[\neg\text{WRITES}(p) \rightarrow \neg\text{WRITER}(p)]$, we have a contradiction. Likewise, $\forall p[\text{WRITER}(p) \leftrightarrow \text{WRITES}(p)]$ would lead to even more contradictions in similar cases. For some examples, the pragmatics *may* facilitate a non-contradictory reading. In Example 21, the first instance of *sleep* could pragmatically mean *sleep now* or *sleep soon*, while the second instance may mean *sleep later*.

(20) It is necessary that someone writes. Hani does not write, but **she writes carefully when writing...** (GPT-2)

(21) Takuto must do something. If Takuto does not sleep, then **he will sleep till the very next day...** (GPT-2 XL)

(22) Someone must dance. Una does not dance, although **she is a very graceful dancer, and not a weak one either...** (GPT-2 XL)

(23) Someone ate dessert. Taguhi didn't eat dessert, but **she was eating her dessert...** (GPT-2)

(24) Boram played something. Boram didn't play the saxophone, although **Boram was a very good saxophonist**. (Flan-T5 Small)

Most "incorrect" outputs, however, did not follow the prompt but did not logically contradict it; they were nonsequiturs with respect to the prior context (not necessarily with respect to the negated sentence), pragmatically ignoring the prior context and leaving the logical unknowns unresolved. Some examples are:

(25) Osamu should have done something. Osamu didn't work, although **nobody in the rest of the division stood out too much.** (GPT-2 XL)

(26) Jordan loves something. Jordan does not love chess, but **you all can learn something from his games.** (GPT-2 XL)

(27) It was necessary that someone cooked. If Jenessa didn't cook, then **Jenessa would have to eat out.** (Flan-T5 XXL)

(28) Someone should swim. Taguhi doesn't swim, although **Taguhi is a mermaid.** (Flan-T5 XXL)

(29) Una has to do something. If Una does not dance, then **the others won't do anything...** (GPT-2)

(30) Someone writes. Hani does not write, but **there are multiple versions of his story...** (GPT-2)

(31) Someone had to like basketball. Boram didn't like basketball, although **Boram was a very talented basketball player.** (Flan-T5 Small)

(32) Someone loves chess. If Jordan does not love chess, then **Jordan will have to play chess.** (Flan-T5 Small)

In addition to calculating the overall accuracy of each model, we compare the accuracy of each model with respect to the transitivity of the verb, the type of prior context (if any), and the conjunction used in each prompt.

4.3.1 Transitivity

All four models performed better on prompts with intransitive verbs than on prompts with transitive verbs. On both intransitive and transitive verbs Flan-T5 XXL outperformed GPT-2 XL, Flan-T5 XXL outperformed Flan-T5 Small, GPT 2 outperformed Flan-T5 Small, and GPT-2 XL outperformed GPT-2.

Recall that random baseline accuracy for prompts with intransitive verbs $P(\text{correct}|\text{intrans}) = \frac{1}{3}$ and $P(\text{correct}|\text{transitive}) = \frac{1}{4}$. GPT-2 XL, GPT-2, and Flan-T5 Small performed worse than random on both intransitive and transitive verbs while Flan-T5 XXL performed slightly better (Figure 1).

4.3.2 Prior Context

We also compare each model's accuracy over the different types of prior context with explicit modality (*original*, *possible*, *necessary*, *must*, *should*, and *has to*). The accuracy for each type of prior context is determined by dividing the number of correct outputs whose prompt contained that prior context by the total number of outputs whose prompts contained that prior context (Figure 2).

Flan-T5 XXL consistently outperforms GPT-2 XL, which struggles to even meet the random baseline, suggesting that it has a bias that is not well attenuated by the prior context. On Flan-T5 XXL, corresponding with our intuition, sentences of the from $\Box P$ by far perform the best, with the exception of *necessary*, which achieves approximately baseline accuracy. Sentences of the form $\Diamond P$ underperform the original sentences with our explicit modal operators on Flan-T5 XXL. This also accords without intuition, since possibility operators may convey more uncertainty than no modal operator at all. Interestingly, however, GPT-2 outperforms Flan-T5 Small on all types of prior context except $\Diamond P$ (*possible*).

Flan-T5 XXL outperforms Flan-T5 Small on all types of prior context, and GPT-2 XL outperforms GPT-2 on all types of prior contexts, revealing a clear advantage for larger models of the same class.

4.3.3 Conjunctions

Finally, Figure 3 shows the accuracy of each model with respect to the three types of conjunctions—*but, although*, and *if-then*—found at the end of each prompt. These accuracies are computed by dividing the number of outputs labeled as correct for a particular conjunction by the total number of prompts that include the conjunction. Similarly to the models' accuracies with respect to transitivity, Flan-T5 XXL achieved the highest accuracy on type of conjunction.

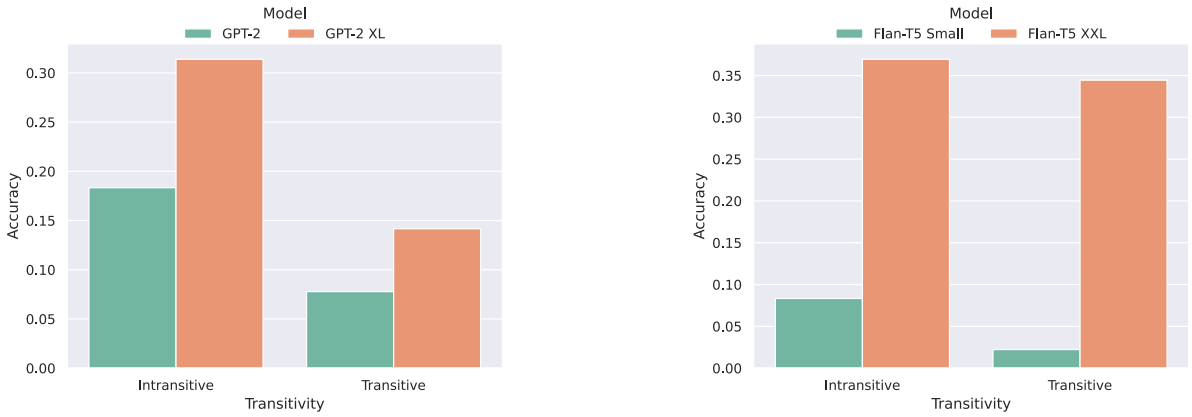

Figure 1: Accuracy for transitive and intransitive verbs on smaller and larger language models. Random baseline accuracy is 0.292. The smaller models perform worse than random (due to nonsequitur outputs) and the larger ones do slightly better than random. Larger models clearly vastly outperform smaller ones.

5 Discussion and Conclusion

We now make some general observations about our results, before concluding and suggesting some avenues for future work.

At the highest level, the fact that the aggregate accuracy for the transitive and intransitive cases did not reach 0.4 with any model on this ostensibly simple task—and in many cases failed to meet a random baseline—suggests that the base LLMs probability calculations do not model this phenomenon well, in ways that might not be obvious by, for example, calculating perplexity. We demonstrate three non-exclusive failure modes: wrong focus, logically contradictory outputs, and nonsequiturs with respect to the prompt. In the first two, the model tries to play the game and fails. In the last case, the model does not play the game at all and

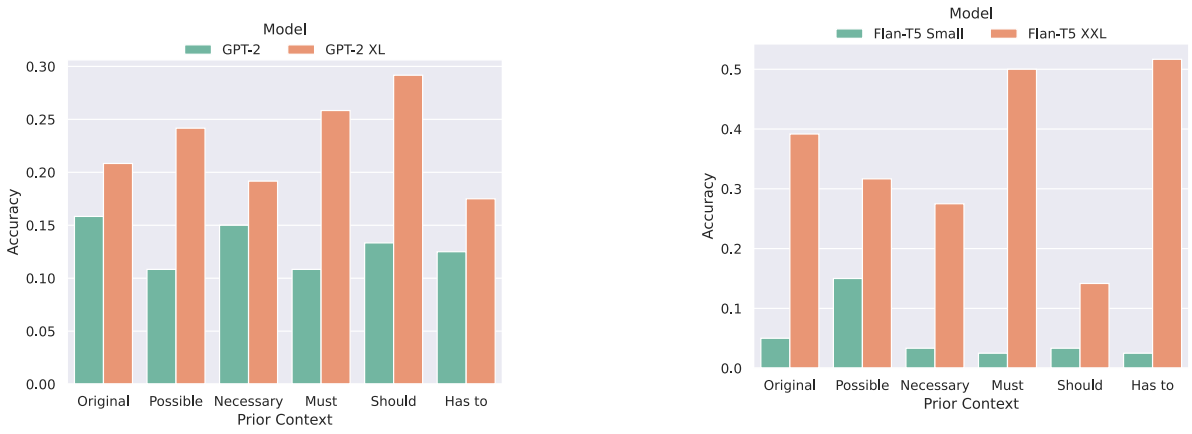

Figure 2: Accuracy with respect to the modality-inducing word in the prior context. Introducing modal terms greatly modulates accuracy. Necessity modals *must* and *has to* generate the expected outputs approximately half the time on Flan T5 XXL. On GPT2 XL, *must*, *should*, and *possible* lead to more consistent generations than the original, but still underperform a random baseline overall.

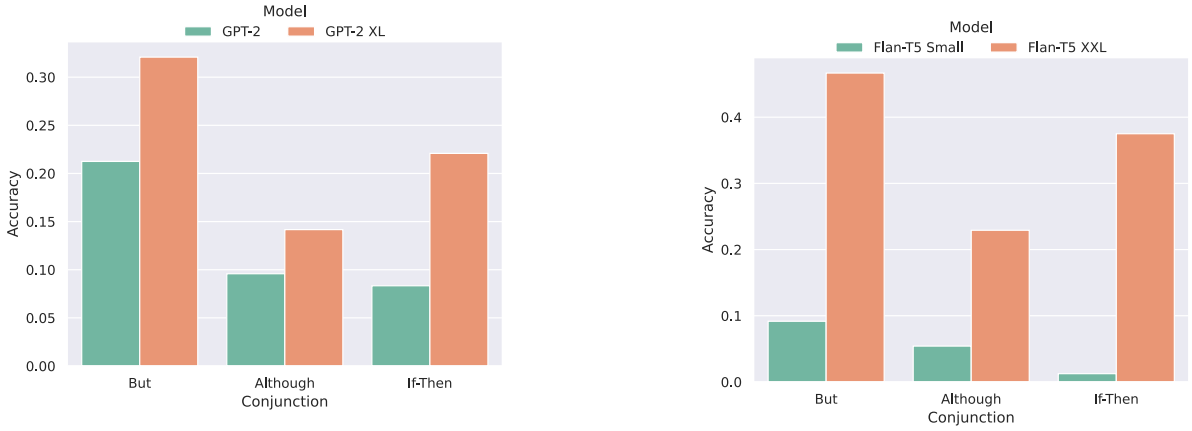

Figure 3: Accuracy with with respect to each of the three conjunctions used in the input prompts. *But* is the only conjunction that ever exceeds the baseline for the larger models.

completes the prompt in a way that *may* be felicitous with respect to the *local* (intrasentential) context but which is infelicitous with respect to the whole (intersentential) context, however brief. The last case is more ambiguous with regard to "correctness", but despite this, our results show how often the models played the game and did so correctly.

Variation in Prior Context Influence It is surprising that each model's accuracy with respect to different prior contexts in the prompt vary as much as they do. Each type of prior context was applied to each input prompt, so the frequency of different prior contexts is not a factor in the models' completions. The consistent, vast differences between the large and small models and between when adding certain necessity modals suggests that both model size and modal contexts modulate results.

Transitivity Underperformance The fact that all four models were better able to correctly infer the focus of negation in prompts with intransitive verbs could suggest that it was challenging for them to keep track of the relationship between the verb and its object.

Conjunctions The models better inferred the focus of negation from prompts ending with the word *but*, which may suggest that *but* could have appeared more often and in more varied environments in the training

data for each model than, say, *although*, which seems plausible. But it could also be the case that *although* simply offers a weaker or more vague contrast signal with respect to the completion.

Implications Our work attempts to evaluate an open-ended, zero-shot completion task with LLMs of various sizes, providing insight into how LLM reasoning is attenuated by size on a specific linguistic task that requires pragmatic rasoning. While we do not claim that the models explicitly do semantic reasoning, we show that there is wide variation in how LLMs perform the task in an open-ended setting and that the LLM completions are attenuated along the axes of LLM size and sentence modality. While individuals may contest the acceptability judgments of various prompts, the fact that the LLMs produce different outputs along these axes remains.

Limitations and Future Directions There are several limitations of the study that could have influenced these results. We designed the prompts, including the contexts, to prime the LLMs to complete the prompt in a way coherent with the both the incomplete negated sentence and the prior context in another sentence, but some of the LLM-generated continuations simply ignored the prior context from a pragmatic standpoint, as shown by the fact that they did not play the game we set up. Sometimes the LLM implicitly focused on a constituent other than the constituent specified in the prompt (e.g., focusing the verb instead of the subject). Additionally, in some outputs that contained two separate sentences, the first sentence contradicted the prompt while the second sentence correctly identified the focus of negation. Follow-up work could query human evaluators to complete the input prompts (without revealing the nature of the task) and compare their results to those of the models to get a better sense of how well the models are expected to do in relation to humans under similar constraints. Our work assumes—justifiably, we argue—that humans will not have trouble with this task.

We did not use plural subjects or objects in our prompts; nor did we consider tense. Future work can isolate these features in addition to those studied here to get a better sense of what these large language models are able or unable to represent about natural language.

Additionally, the sentences generated for these experiments are not representative of all varieties of English, so these results do not reflect a model's use of utterances across varieties of English. Future work can consider a more diverse set of inputs with respect to the structure, content, and variety of English to better understand how well neural models are able to interpret negation in a larger set of contexts. Future work can also expand this study to models of languages other than English, as the structure of different languages may pose different and unique challenges to pragmatic inference.

References

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. Modality and Negation in SIMT Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*, 38(2):411–438, 06 2012. ISSN 0891-2017. doi: 10.1162/COLI_a_00099. URL https://doi.org/10.1162/COLI_a_00099.

Rajesh Bhatt. *Covert Modality in Non-finite Contexts*. De Gruyter Mouton, Berlin, New York, 2006. ISBN 9783110197341. doi: doi:10.1515/9783110197341. URL https://doi.org/10.1515/9783110197341.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha

Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *Proceedings of the International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3Pf3Wg6o-A4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Peter Erdmann. Factive, implicative verbs and the order of operators. *Studia Linguistica*, 28(1):51–63, 1974. URL https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9582.1974.tb00604.x?casa_token=D1iDbL5R1rIAAAAA:4YDs8cNrmdVKlhH1zQ_kX_NT_Mdexz5XKVr6vthnXCLXjjjFWvRxUPhVo_TqPnLe7vxa2QcWaOxWhYHc.

Kai Von Fintel. Modality and language. In D. Borchert (ed.), *Encyclopedia of Philosophy*, pp. 20–27. Macmillan Reference, 2006. URL https://web.mit.edu/fintel/fintel-2006-modality.pdf.

Alvin Grissom II and Yusuke Miyao. Annotating factive verbs. In *International Conference on Language Resources and Evaluation*, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/757_Paper.pdf.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022. URL https://arxiv.org/pdf/2209.00840.pdf.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3869–3885, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.345. URL https://aclanthology.org/2020.findings-emnlp.345.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. Understanding by understanding not: Modeling negation in language models. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1301–1312, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.102. URL https://aclanthology.org/2021.naacl-main.102.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. Negation, coordination, and quantifiers in contextualized language models. In *Proceedings of International Conference on Computational Linguistics*, pp. 3074–3085, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.272.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*, 2019. URL https://arxiv.org/abs/1911.03343.

Aditya Khandelwal and Benita Kathleen Britto. Multitask learning of negation and speculation using transformers. In *Proceedings of the International Workshop on Health Text Mining and Information Analysis*, pp. 79–87. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.louhi-1.9.pdf.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Proceedings of Advances in Neural Information Processing Systems*, 35:22199–22213. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Saul A Kripke. Presupposition and anaphora: Remarks on the formulation of the projection problem. *Linguistic Inquiry*, 40(3):367–386, 2009. URL https://direct.mit.edu/ling/article-pdf/40/3/367/724537/ling.2009.40.3.367.pdf?casa_token=gTk3ey8mL-oAAAAA:xf-wCVrx49eTjsKmY2osCAJ3aVG0lQLYJfBJuNvbOpsOT3P2Zv8bqu4sWaMHASNb2j8tHlaD.

Hwan-Mook Lee. Negation and compositionality. In *Proceedings of the Korean Society for Language and Information Conference*, pp. 95–105. Korean Society for Language and Information, 1985. URL https://waseda.repo.nii.ac.jp/?action=pages_view_main&active_action=repository_action_common_download&item_id=28362&item_no=1&attribute_id=101&file_no=1&page_id=13&block_id=21.

Yoshihiro Maruyama. Learning, development, and emergence of compositionality in natural language processing. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–7, 2021. doi: 10.1109/ICDL49984.2021.9515636.

Suguru Matsuyoshi, Ryo Otsuki, and Fumiyo Fukumoto. Annotating the focus of negation in Japanese text. In *International Conference on Language Resources and Evaluation*, pp. 1743–1750, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/777_Paper.pdf.

R Thomas McCoy and Tal Linzen. Non-entailed subsequences as a challenge for natural language inference. *arXiv preprint arXiv:1811.12112*, 2018. URL https://arxiv.org/pdf/1811.12112.pdf.

Roser Morante and Eduardo Blanco. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Joint Conference on Lexical and Computational Semantics*, pp. 265–274, 2012. URL https://aclanthology.org/S12-1035.pdf.

Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012. URL https://direct.mit.edu/coli/article-pdf/38/2/223/1801323/coli_a_00095.pdf.

Frank Robert Palmer. *Mood and modality*. Cambridge University Press, 2001. URL https://academic.oup.com/edited-volume/34871/chapter-abstract/298322377?redirectedFrom=fulltext.

Paul Portner. *Modality*. Oxford University Press, 2009. URL https://global.oup.com/academic/product/modality-9780199292431?cc=us&lang=en&.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. URL http://static.cs.brown.edu/courses/cs146/assets/papers/language_models_are_unsupervised_multitask_learners.pdf.

Sabine Rosenberg and Sabine Bergler. Uconcordia: Clac negation focus detection at* sem 2012. In *Joint Conference on Lexical and Computational Semantics*, pp. 294–300, 2012. URL https://aclanthology.org/S12-1039.pdf.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples. *arXiv preprint arXiv:2305.15269*, 2023. URL https://arxiv.org/abs/2305.15269.

Roser Saurí and James Pustejovsky. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*, 38(2):261–299, 06 2012. ISSN 0891-2017. doi: 10.1162/COLI_a_00096. URL https://doi.org/10.1162/COLI_a_00096.

Longxiang Shen, Bowei Zou, Yu Hong, Guodong Zhou, Qiaoming Zhu, and AiTi Aw. Negative focus detection via contextual attention mechanism. In *Proceedings of Empirical Methods in Natural Language Processing*, pp. 2251–2261, 2019. URL https://aclanthology.org/D19-1230.pdf.

Gongbo Tang, Philipp Rönchen, Rico Sennrich, and Joakim Nivre. Revisiting negation in neural machine translation. *Transactions of the Association for Computational Linguistics*, 9:740–755, 2021. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00395/106793.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. Speculation and Negation: Rules, Rankers, and the Role of Syntax. *Computational Linguistics*, 38(2):369–410, 06 2012. ISSN 0891-2017. doi: 10.1162/COLI_a_00126. URL https://doi.org/10.1162/COLI_a_00126.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. Negation focus identification with contextual discourse information. In *Proceedings of the Association for Computational Linguistics*, pp. 522–530, 2014. URL https://aclanthology.org/P14-1049.pdf.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. Unsupervised negation focus identification with word-topic graph model. In *Proceedings of Empirical Methods in Natural Language Processing*, pp. 1632–1636, 2015. URL https://aclanthology.org/D15-1187.pdf.