What does the Failure to Reason with "Respectively" in Zero/Few-Shot Settings Tell Us about Language Models?

Ruixiang Cui[1], **Seolhwa Lee**[2], **Daniel Hershcovich**[1], **Anders Søgaard**[1]
[1]Department of Computer Science, University of Copenhagen
[2]Department of Computer Science, Technical University of Darmstadt
{rc,dh,soegaard}@di.ku.dk, seolhwa.lee@tu-darmstadt.de

Abstract

Humans can effortlessly understand the coordinate structure of sentences such as "Niels Bohr and Kurt Cobain were born in Copenhagen and Seattle, *respectively*". In the context of natural language inference (NLI), we examine how language models (LMs) reason with respective readings (Gawron and Kehler, 2004) from two perspectives: syntactic-semantic and commonsense-world knowledge. We propose a controlled synthetic dataset WikiResNLI and a naturally occurring dataset NatResNLI to encompass various explicit and implicit realizations of "respectively". We show that fine-tuned NLI models struggle with understanding such readings without explicit supervision. While few-shot learning is easy in the presence of explicit cues, longer training is required when the reading is evoked implicitly, leaving models to rely on common sense inferences. Furthermore, our fine-grained analysis indicates models fail to generalize across different constructions. To conclude, we demonstrate that LMs still lag behind humans in generalizing to the long tail of linguistic constructions.

1 Introduction

Transformer-based language models (LMs) (Devlin et al., 2019; Raffel et al., 2019; Brown et al., 2020) induce useful representations for a wide range of natural language understanding (NLU) tasks, including natural language inference (NLI; Wang et al., 2018; Hu et al., 2020), especially in in zero-shot or few-shot settings. To what extent this usefulness results from memorization, generalization or the ability of LMs to draw common sense inferences remains an open question.

To approach it, the linguistic phenomenon of respective readings (Gawron and Kehler, 2004) serves as an excellent probe. This phenomenon has so far been underexplored in NLP, even though it has been studied extensively in linguistic semantics (McCawley, 1968; Pullum and Gazdar, 1982; Dalrymple and Kehler, 1995; Eggert, 2000). In English, "respectively" is a rare word[1] used to establish a one-to-one mapping between two sets of participants and to distribute predicates over sets (Okada, 1999). For example, in Figure 1, the first conjunct in the subject corresponds to the first conjunct in the object and the second conjunct in the subject corresponds to the second conjunct in the object. The respective relation is bijective and respects the relative order of the elements of two different coordinate expressions; it is, in other words, cross-serial. "Respectively" can have different syntactic or semantic properties depending on the context, e.g., as a conjunction or adverb.

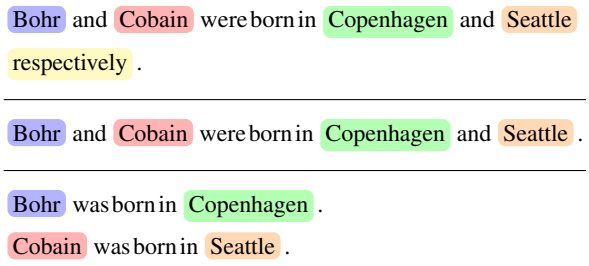

Figure 1: An example of explicit (top, evoked by "respectively") and implicit (middle, with no overt marker) respecitve readings. Humans can infer that both sentences have the same "cross-serial" meaning (bottom) by relying on commonsense knowledge (that a person is only born in one location) and world knowledge (that Copenhagen and Seattle are mutually exclusive).

In this paper, we investigate how LMs reason with respective readings. We propose two datasets, WikiResNLI (a controlled synthetic dataset) and NatResNLI (a naturally occurring dataset) to cover various explicit and implicit realizations of "respectively". Our research questions are:

1. Can NLI models reason with "respectively" constructions in zero-shot settings?

[1]In terms of frequency, in the British National Corpus, "respectively" is ranked 13,606th among 18,089 words, and 233rd among 429 adverbs (Leech et al., 2014).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8786–8800

2. Can LMs generalize from explicit to implicit respective readings?

3. Can LMs generalize from synthetic to natural respective readings?

4. What cues do LMs leverage for prediction?

We experiment with state-of-the-art LMs and analyze the results to gain insights into the limitations of current models and potential directions for future research. We show that LMs are able to generalize effectively in a few-shot learning scenario when the word "respectively" is present. However, when the reading is evoked implicitly, a greater number of training instances are necessary. LMs require significantly more instances to generalize to naturally occurring datasets than humans. In conclusion, our study demonstrates that LMs continue to exhibit a deficit in generalizability to infrequent linguistic constructions with limited coverage in their training data.

2 Respective Readings

Respective readings are closely related to several types of readings instantiated by plurals and mass terms: distributive readings, collective readings and cumulative readings (Champollion, 2015).

Distributive readings. These usually refer to the application of a predicate to the subsets of a set or group. As for sentence 1(a), it is equivalent to "John smiled and Mary smiled". The reading is available because of the nature of the predicate is *atomic* (Winter, 2002), similar instances including "sing" and "sleep". Distributive reading can be enforced with overt distributive markers, i.e., "every" and "each" (Scha, 1984). In example 1(b), we enforce the reading by adding "each" at the end of the sentence so as to rule out the reading "John and Mary earn 200 dollars together".

1. (a) **Distributive reading:** John and Mary smiled.
   (b) **Distributive reading with an enforced marker:** John and Mary earn 200 dollars *each*.

Collective readings. These are the opposite of distributive readings in that the predicates apply to the whole plural entity instead of individuals. The quantifiers "all" and "most" instead of "every" and "each" are usually compatible with collective readings as in example 2(b) (Dowty et al., 1987).

2. (a) **Collective reading:** The men gathered.
   (b) **Collective reading with overt marker:** *All* of the men gathered.

Cumulative readings. These involves two entities but in a symmetric non-scopal relation as in the canonical example 3 (Scha, 1984). The sentence can be paraphrased into "There are three boys and two girls, each of the three boy saw at least one of the two girls, and each of the two girls was seen by at least one of the three boys.". It is discussed sometimes with weak reciprocity (Langendoen, 1978).

3. **Cumulative reading:** Three boys saw two girls.

Respective readings. These are thought to be a special case of cumulative readings in which a bijective relation holds between the two (or more) sets of entities that enter into the cumulative relation (Chaves, 2012). For example 4(a), the pair (Emiliano Zapata, Morelos) and the pair (Gerhart Münch, Michoacán) are grouped under the *died in* relation. Respective reading can also arise without the adverb *respectively*, and the absence is even sometimes preferred. As in example 4(b), the binomial expression "husband and wife" is so strong that the adverb "respectively" is unwarranted.

4. (a) **Respective reading with overt marker:** Emiliano Zapata and Gerhart Münch. died in Morelos and Michoacán, *respectively*.
   (b) **Respective reading without overt marker:** John and Mary are husband and wife.

3 An NLI Benchmark for "Respectively"

Understanding the coordinate structures in respective readings is effortless for humans, but it remains a question whether LMs, after being pre-trained on billions of tokens and fined-tuned on thousands of NLI instances, can reliably process them.

To probe LMs' behaviour in the presence of respective readings, we construct two English NLI datasets: WikiResNLI, a synthetic dataset based on an analogy corpus, and NatResNLI, a dataset sourced and created from natural occurrences. We release both datasets on Github[2] and describe the detailed creation steps below.

3.1 Synthetic Dataset: WikiResNLI

To generate a controlled synthetic challenge set for reasoning with respective readings, we exploit a useful relationship between coordination constructions and *analogies*. Analogy is concerned with similarities between observable properties and causal similarities.

[2] https://github.com/ruixiangcui/WikiResNLI_NatResNLI

	Denotation	Natural Language Example
Premise:	w_1 and w_3 p w_2 and w_4, respectively.	*Emiliano Zapata* *and* *Gerhart Münch* *died in* *Morelos* *and* *Michoacán*, *respectively*
Hypotheses:		
Entailment (1), 1S1O	w_1 p w_2.	*Emiliano Zapata* *died in* *Morelos*.
Entailment (2), 1S1O	w_3 p w_4.	*Gerhart Münch* *died in* *Michoacán*.
Contradiction (1), 1S1O	w_1 p w_4.	*Emiliano Zapata* *died in* *Michoacán*.
Contradiction (2), 1S1O	w_3 p w_2.	*Gerhart Münch* *died in* *Morelos*.
Contradiction (3), 1S2O	w_1 p w_2 and w_4.	*Emiliano Zapata* *died in* *Morelos* *and* *Michoacán*.
Contradiction (4), 1S2O	w_3 p w_2 and w_4.	*Gerhart Münch* *died in* *Morelos* *and* *Michoacán*.
Contradiction (5), 2S1O	w_1 and w_3 p w_2.	*Emiliano Zapata* *and* *Gerhart Münch* *died in* *Morelos*.
Contradiction (6), 2S1O	w_1 and w_3 p w_4.	*Emiliano Zapata* *and* *Gerhart Münch* *died in* *Michoacán*.

Table 1: Example analogy in the spirit of Garneau et al. (2021). Both entity pairs (w_1, w_2 ; w_3, w_4) share the p relation. Object entities are *unique* in that given an entity pair and a subject, the fourth is uniquely determined. We generate eight hypotheses for each premise: 1S1O refers to one subject and one object, 1S2O refers to one subject and two objects and 2S1O refers to two subjects and one object.

Analogy dataset. Garneau et al. (2021) proposed WiQueen, a multilingual analogy dataset consisting of 78,000 analogies extracted from Wikidata. A subset of 9,000 instances is annotated where all four entities are *unique*. These are the analogies in which all relations are informative (Newman-Griffis et al., 2017). See Table 1 for an example. Their experiment showed that pretrained LMs can predict 29% of analogous entities in a zero-shot setting and 41% after training. This indicates that analogical knowledge already exists in pretrained models and can be enhanced by training.

Generating premises with "respectively". Given four analogical entities $\langle w_1, w_2, w_3, w_4 \rangle$ and the predicate p, we form a natural language premise consisting of the analogical information in a respective reading setting of 5(a) after adapting p for phrasing and conjugation. Such a premise is unambiguous and equivalent to 5(b), where the predication is distributed over the two pairs of entities. 5(a) is marked by an explicit respective reading indicator. As an implicit respective reading case, 5(c) has the same meaning as 5(b) but there is no explicit respective operator. In such implicit cases, the predicate p is usually mutually exclusive in that each subject can have only one object. For example, in Sentence 6(a) a person can only die in one place but not two places. Non-mutually exclusive predicates are disqualified for an implicit respective reading since they causes ambiguity, as in Sentence 6(b).

5. (a) w_1 and w_3 p w_2 and w_4, respectively.
   (b) w_1 p w_2 and w_3 p w_4.
   (c) w_1 and w_3 p w_2 and w_4.

6. (a) Emiliano Zapata and Gerhart Münch died in Morelos and Michoacán.
   (b) John and Mary ate a falafel and a tortilla.

Generating hypotheses. We subsequently generate hypotheses and pair them with the generated explicit and implicit premises. In Table 1, we show the rules to write entailment or contradiction hypotheses given a premise created from the analogical entities and properties.

Statistics. The resulting dataset, which we call WikiResNLI$_{\text{EXPLICIT}}$, contains 2,317 premises with different analogical entities, each of which has two entailment hypotheses and six contradiction hypotheses, resulting in 18,536 premise-hypothesis pairs in total. The dataset has 139 different predicates derived from Wikidata properties. For the development set, we randomly sample 13 predicates from the 126 predicates left and trimmed them if the number of premises for each predicate exceeds 100. We have 1,312 premise-hypothesis pairs for the development set. The rest is used as the training set, with 1,577 premises and 12,616 premise-hypothesis pairs.

Generating premises with implicit "respectively". We aim to test whether LMs can reason with respective readings and generalize from explicit construction to instances without overt markers. For this purpose, we derive an implicit dataset from WikiResNLI$_{\text{EXPLICIT}}$ by simply removing the

Human Reference	Entailment	Neutral	Contradiction
Entailment	93.4	2.1	4.5
Contradiction	5.9	4.1	90

Table 2: NatResNLI human annotated label distribution in percentages for each assigned reference label. Humans mostly agree with the pre-assigned reference labels (demonstrated in Table 1), but not always.

word "respectively" from the premises. We call this dataset WikiResNLI$_{\text{IMPLICIT}}$. In this process, we need to pay special attention to the fact that ambiguity usually occurs in the 1S2O setting when the predicates allow conjunction of objects; given the sentence 6(b), it is ambiguous whether the hypothesis "John ate a falafel and a tortilla" is entailed. To form a high-quality test set for WikiResNLI$_{\text{IMPLICIT}}$, we first need to exclude the ambiguous contradiction hypotheses. Therefore, two of the authors manually annotate the 139 predicates for whether they allow a single subject predicating conjunction of two objects. In total, 13 predicates are annotated by both authors as unambiguous. Subsequently, we keep only the premises with these predicates from the complete WikiResNLI, and for each predicate, we cap it if the number of premises exceeds 100. Eventually, we are left with 451 premises for the 13 predicates. The 3,608 premise-hypotheses pairs are used as the test set.

3.2 Naturally-occurring Dataset: NatResNLI

While the synthetic dataset is well-controlled, it does not necessarily cover the natural usage of "respectively". To address this, we also collect a dataset of naturally-occuring usages.

Collecting premises. As data resources for "respectively" in publicly available naturally-occuring data, we leverage two online dictionaries[3] and a writing advice blog,[4] which provide English examples containing specific words in real-world examples. We curate the sentences that included "respectively" and further filter some of them to avoid context ambiguity. In total, 76 sentences remain as the premise set.

[3] https://sentence.yourdictionary.com/respectively and https://www.dictionary.com/browse/respectively

[4] https://crosstalk.cell.com/blog/how-to-use-respectively-respectfully

Generating hypotheses. Two of the authors manually write hypotheses based on the fine-grained categorization of Table 1 for each collected premise. Given that the labels are pre-assumed, and to determine whether these inference relations align with humans, we employ crowd workers to verify them. See the annotation details in Appendix A.

Statistics. The resulting dataset, which we call NatResNLI, consists of 76 premises and 608 hypotheses. The average sentence lengths of NatResNLI's premise and hypothesis are 20.1 and 10.1, respectively. Sentences have 2.32 conjnucts in average, with 4 as the maximum.

Variety. NatResNLI's sentences have more complicated linguistic constuctions than WikiResNLI, such as relative clauses, e.g., sentence 7(a), implicit coreferences in sentence 7(b), and inverted sentences in sentence 7(c).

7. (a) The annual value of the Hulse endowment is between £800 and £900, of which eight-tenths go to the professor of divinity and one-tenth to the prize and lectureship, respectively.
   (b) In 1910 the export of palm kernels was 6,141 tons, of palm oil 2,160 tons; in 1916 the figures were 22,391 tons and 3,852 tons respectively.
   (c) Above this, approached by a stair, are the Lesche and the theatre, occupying respectively the north-east and northwest corner of the precinct.

Inter-annotator Agreement. The inter-annotator agreement (Fleiss' kappa; Fleiss, 1971) of the workers for NatResNLI is 0.65, lower than ANLI's (0.67–0.74) and SNLI's (0.70). This can be attributed to that we have five annotators rather than the commonly chosen three annotators, as a larger number of annotators can sometimes lead to more diverse interpretations and disagreements, potentially lowering the inter-annotator agreement.

Verification of pre-assigned labels. In Table 2, we calculate the average agreement percentage of human annotation with reference labels, showing that humans do not always agree with them. Investigating the examples where the majority votes are distinct from the pre-assigned labels, we find nine instances distributed over four premises. For the sentence in 8(a), humans actually correct the label as the respective reading here does not cause a mutually exclusive effect. For sentence 8(b), humans show more caution towards sentence ambiguity caused by unknown world knowledge of Kilia and Dniester's locations, and hence the neutral label.

Model	Training data	entailment	contradiction			*overall*
			1S1O	1S2O	2S1O	
RoBERTa	MNLI	99.7	47.9	0.3	4.5	38.1
	S,M,F,ANLI	100	55.1	0.1	1	39.1
ALBERT	S,M,F,ANLI	99.8	31.6	3.1	3.9	34.6
DeBERTa-v3	MNLI	99.4	36.1	0.6	3.9	35
	M,F,ANLI	98.8	40.2	3.9	10.8	38.4
	M,F,Ling, WANLI	100	77.8	36.7	59.4	68.5

Table 3: Zero-shot performance on the $\text{WikiResNLI}_{\text{EXPLICIT}}$ test set.

Model	Training data	entailment	contradiction			*overall*
			1S1O	1S2O	2S1O	
RoBERTa	MNLI	97.1	26.4	0.4	8.6	33.1
	S,M,F,ANLI	99.9	23.5	0.3	3.4	31.8
ALBERT	S,M,F,ANLI	100	14	0.1	0.8	28.7
DeBERTa-v3	MNLI	99.3	25.5	1.4	5.2	32.9
	M,F,ANLI	96.9	26.6	5.7	16.9	36.5
	M,F,Ling, WANLI	100	59.3	2.4	13.2	43.7

Table 4: Zero-shot performance on the $\text{WikiResNLI}_{\text{IMPLICIT}}$ test set.

8. (a) **Premise**: The annual value of imports and exports exceeds seven and nine million sterling respectively. **Hypothesis**: The annual value of imports and exports exceeds seven million sterling. **Pre-assigned Label**: contradiction. **Majority Vote**: entailment

   (b) **Premise**: In that year a Turkish fleet captured the strongholds of Kilia and Akkerman, commanding respectively the mouths of the Danube and Dniester. **Hypothesis**: In that year a Turkish fleet captured the stronghold of Kilia, commanding the mouths of the Danube and Dniester. **Pre-asigned Label**: contradiction. **Majority Vote**: neutral

Considering human annotations as ground truth, we discard the pre-assigned labels and adopt the majority votes as the final labels for NatResNLI.

4 Experiments

We begin our experiments with the datasets by addressing our first research question:

Question 1 *Can NLI models reason with the coordinate structure in "respectively" construction in zero-shot settings?*

Given the popularity of NLI as a classification task to test LMs' ability of language understanding, many works have proposed new models achieving state-of-the-art results on datasets such as SNLI (Bowman et al., 2015), MultiNLI (MNLI; Williams et al., 2018) and ANLI (Nie et al., 2020). On the GLUE leaderboard,[5] the state-of-the-art models have surpassed 90% and 95% accuracy on MNLI and QNLI which are deemed as solved challenges. ANLI has been one of the most challenging tasks in recent years, and the latest models such as DeBERTa-v3-large (He et al., 2021; Moritz et al., 2022) and PaLM 540B (Huang et al., 2022) can achieve 64% and 67.9%, respectively. While many works use ANLI as a medium to exhibit the models' growing reasoning ability, few of them analyze in depth in which case it fails and at which stage it gets to learn certain linguistic abilities.

We report the zero-shot performances of three LMs fine-tuned with different combinations of NLI corpora. The models include RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) and DeBERTa with fine-tuning data of MNLI, SNLI, ANLI, FEVER-NLI (Nie et al., 2019), LingNLI (Parrish et al., 2021) and WANLI (Liu et al., 2022).

The experiment results on $\text{WikiResNLI}_{\text{EXPLICIT}}$ and $\text{WikiResNLI}_{\text{IMPLICIT}}$ are presented in Table 3 and Table 4, respectively.

As can be seen in Table 3, models cannot fully correctly reason with respective readings. The best model, DeBERTa, only achieves 35% accuracy if fine-tuned with MNLI, and will reach 68.5% if fine-tuned with almost all NLI training datasets mentioned above. It gains a large increase in the 1S1O setting by 41.7%. However, the accuracy on 1S2O is still at a chance level, and the 2S1O setting performance is only approaching around 60%, leaving room for improvement.

The performance on $\text{WikiResNLI}_{\text{IMPLICIT}}$ is even worse, as indicated in Table 4. Similarly, DeBERTa is again the best performance model on the dataset, with an accuracy of 43.7% if fine-tuned with all NLI corpora. The accuracy is just 10% above the chance level, and it completely fails in the 1S2O and 2S1O settings.

Results on both datasets show that when training with more data, models improve on respective readings. However, the question of what leads to improvement remains. We examine how many times explicit respective readings appear in the training and testing datasets of MNLI, SNLI Fever-NLI and ANLI. We find that the adverb "respectively" occurs 177 and 12 times in the MNLI training and dev sets, 15 and 0 times in the SNLI training and test sets, 1,064 and 64 times in the Fever-NLI training and test sets, and 216 and 5 times in the combined ANLI training and dev sets. We randomly sampled a subset of each dataset and manually check whether they tackle reasoning over coordination structure. We find that in most cases, "respectively" works simply as a context word and has little to do with the actual inference relations. Thus it is

[5] https://gluebenchmark.com/leaderboard

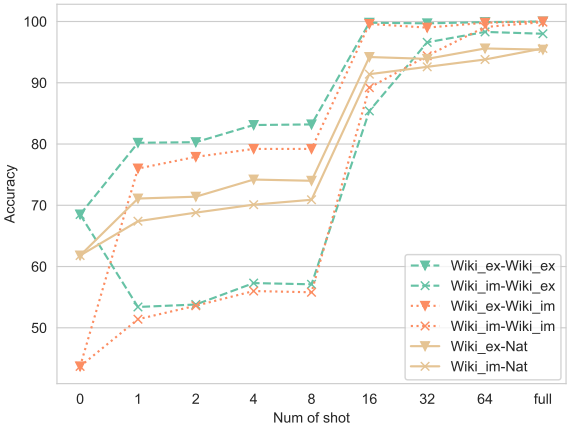

Figure 2: Overall performance of DeBERTa on WikiResNLI$_{\text{EXPLICIT}}$, WikiResNLI$_{\text{IMPLICIT}}$ and NatResNLI from zero-shot to fully supervised. *Wiki_ex-Wiki_ex* refers to training with WikiResNLI$_{\text{EXPLICIT}}$ instances and evaluating on WikiResNLI$_{\text{EXPLICIT}}$ test set. Similarly, *Wiki_im-Nat* refers to training with WikiResNLI$_{\text{IMPLICIT}}$ and testing on NatResNLI.

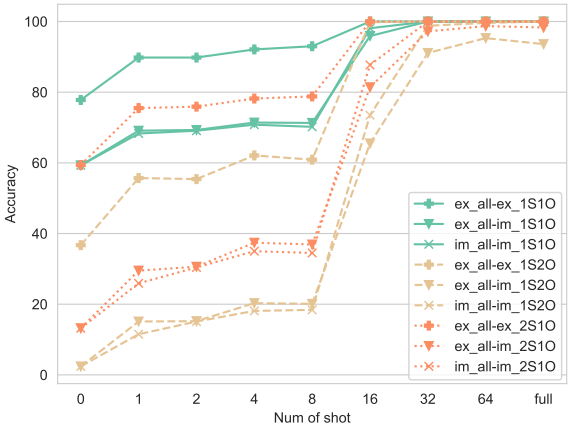

Figure 3: DeBERTa's Performances on WikiResNLI$_{\text{IMPLICIT}}$ after fine-tuning on WikiResNLI$_{\text{EXPLICIT}}$ or WikiResNLI$_{\text{IMPLICIT}}$. The result is broken down by contradiction fine-grained set.

# Shots / **Type**	1	2	4	8	16	32	64	Full
All	8	16	32	64	128	256	512	12,616
Basic	4	8	16	32	64	128	256	6,308

Table 5: Number of training instances for each number of shots. A "shot" contains multiple training instances since we always take a premise along with all of its generated hypotheses—8 in the general case and 4 in the basic case.

still not clear whether it is simply the exposure to the explicit cues (the word "respectively") or some instances with implicit coordinate structures that result in the performance improvement. We thus ask the following three research questions and experiment with few-shot learning.

Question 2 *Can LMs Generalize from Explicit to Implicit Respective Readings?*

Instances of WikiResNLI have the coordinate structures of an equal number of conjuncts, and linguists have argued that such semantic relations are reflected in the syntactic relations (Goodall, 1987; Moltmann, 1992). It is essentially semantic but also relies on pragmatically available information of the truth conditions. Respective readings in fact also commonly omit explicit lexical indicators but remain available and preferred as 2(a) (Gawron and Kehler, 2004). We are therefore interested in whether LMs can learn the semantic-pragmatic meaning of respective reading sentences rather than only making use of lexical and syntactic cues.

We fine-tune the DeBERTa model previously fine-tuned with M, F, Ling and WANLI with different numbers of WikiResNLI$_{\text{EXPLICIT}}$ examples without a dev set, since we do not want to bias the model towards our datasets hence hurting performance on the other NLI tasks. We fine-tune the model with WikiResNLI$_{\text{EXPLICIT}}$ and WikiResNLI$_{\text{IMPLICIT}}$ separately and report the overall accuracy on both dataset in Figure 2. Training with WikiResNLI$_{\text{EXPLICIT}}$ contributes to a steady performance increase on both WikiResNLI$_{\text{EXPLICIT}}$ and WikiResNLI$_{\text{IMPLICIT}}$. Especially, 1-shot learning enhances the performance clearly, with a 10% increase for in-domain evaluation, and a remarkable 30% increase for explicit to implicit generalization. The improvements are small from 1-shot to 8-shot. Only at 16-shot, both WikiResNLI$_{\text{EXPLICIT}}$ in-domain learning and transferring to WikiResNLI$_{\text{IMPLICIT}}$ reach 100% accuracy. This shows the possibility to learn respective readings, despite the need to see relevant instances 128 times (see Table 5).

Interestingly, in-domain few-shot learning of WikiResNIL$_{\text{IMPLICIT}}$ witnesses a relatively cold start. The accuracy does not increase above 60% until 16 shots. Generalization from implicit respective reading to explicit reading is surprisingly not reaching 100% accuracy even after full supervision. We are keen to investigate what types of instances are difficult to learn for explicit to implicit respective reading generalization. In Figure 3, we break down WikiResNLI$_{\text{IMPLICIT}}$ with contradiction labels

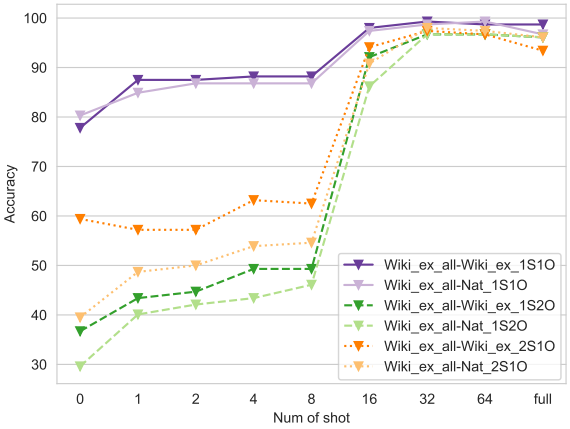

Figure 4: Performance of DeBERTa on NatResNLI after being fine-tuned on $\text{WikiResNLI}_{\text{EXPLICIT}}$. To facilitate comparison, we mark performances on $\text{WikiResNLI}_{\text{EXPLICIT}}$ in darker colours.

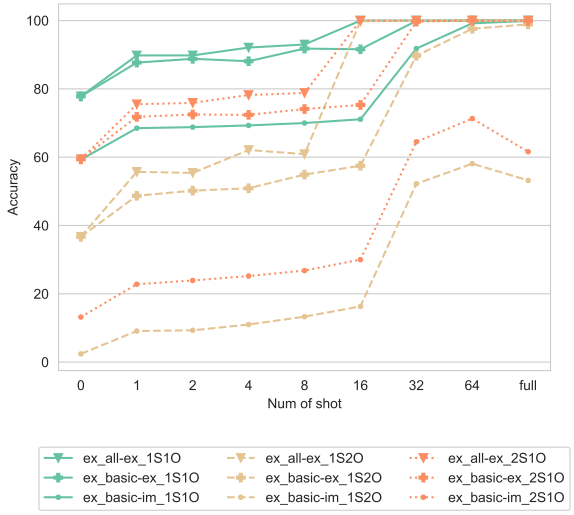

Figure 5: Performance of DeBERTa on $\text{WikiResNLI}_{\text{EXPLICIT}}$ and $\text{WikiResNLI}_{\text{IMPLICIT}}$ after being fine-tuned only with the basic types (entailment and 1S1O contradiction) of $\text{WikiResNLI}_{\text{EXPLICIT}}$.

by categories (1S1O, 1S2O and 2S1O) and plot the accuracy against number of shot.

As can be seen, the performance on explicit readings is always better than on implicit readings across all three contradiction types. Among them, 1S2O and 2S1O instances are the most difficult. Their accuracies are below 40% and 20%, respectively before 16 shots. And only until 32 shots do both types reach above 95% accuracy. Unlike in-domain learning, 1S2O never gets perfectly solved.

Question 3 *Can LMs Generalize from Synthetic to Natural Respective Readings?*

WikiResNLI is a synthetic dataset, and it remains unclear whether models can reason with respective readings in realistic settings if we generate enough synthetic data and feed it to models. With NatResNLI, we are able to investigate LM's respective reading reasoning generalizability from synthetic to natural data and its alignment with humans.

We evaluate the models fine-tuned with $\text{WikiResNLI}_{\text{EXPLICIT}}$ on NatResNLI and plot the performance in Figure 4. We can observe that scores on NatResNLI are almost always lower than on WikiResNLI due to domain drift. Particularly, 1S2O and 2S1O are 10% and 20% lower in zero-shot settings. 1S2O manage to reach on-par performance with WikiResNLI after 16 shots, while 2S1O after 32 shots.

Interestingly, the models are able to surpass 95% after 32 shots, while pre-assigned labels only have 90% match (see Table 2). Although we are comparing a rule-based method with 32-shot (256 examples) training, we can conclude that models are able to align with humans for respective reading reasoning. In addition, we notice that for 1S2O and 2S1O generalization, the complex linguistic structures discussed in Section 3.2 do have a high impact in the low-number few-shot learning, but the difficulty diminished as more training data are used.

Question 4 *What Cues do LMs Rely on?*

So far we have discussed LMs' ability to generalize on the syntactic-semantic level, from explicit to implicit and from synthetic to natural in respective readings. But it is yet to be determined whether the model is simply adopting the lexical-syntactic heuristics for prediction and whether it leverages common sense and world knowledge. If models can reason over basic hypothesis structures (1S1O entailment and 1S1O contradiction), it would be expected they are aware that the one-to-one relation correspondences should exclude 1S2O and 2S1O propositions due to common sense and world knowledge. Although there are cases such as 8(a) where one object entity includes the other in NatResNLI, all cases of the WikiResNLI test set disallow the situation due to the mutually exclusive properties.

Therefore, we fine-tuned the DeBERTa models with only $\text{WikiResNLI}_{\text{EXPLICIT}}$ instances of basic structures and evaluated their performances on both $\text{WikiResNLI}_{\text{EXPLICIT}}$ and $\text{WikiResNLI}_{\text{IMPLICIT}}$ 1S2O and 2S1O. The results can be seen in Figure 5.

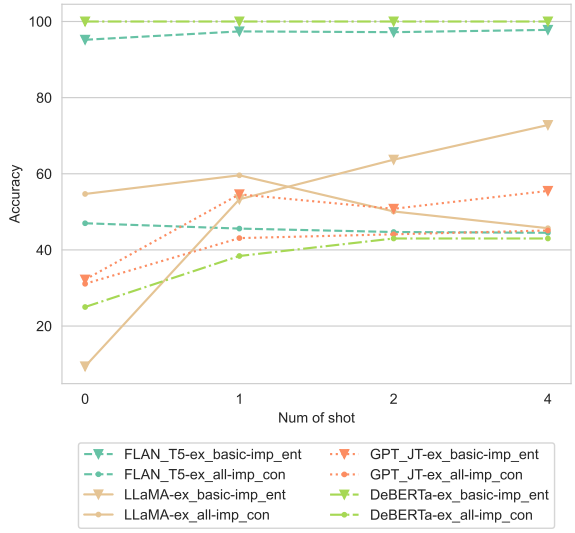

Figure 6: LLaMA, FLAN-T5, GPT-JT and DeBERTa's performances on $\text{WikiResNLI}_{\text{IMPLICIT}}$ after in-context learning of $\text{WikiResNLI}_{\text{EXPLICIT}}$. The last suffix *ent* of a legend means the performance on entailment pairs and *con* on contradiction pairs.

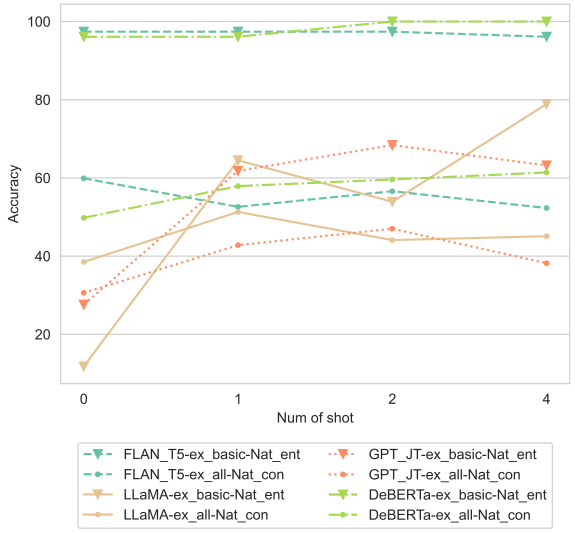

Figure 7: LLaMA, FLAN-T5, GPT-JT and DeBERTa's performances on NatResNLI after in-context learning of $\text{WikiResNLI}_{\text{EXPLICIT}}$.

We can observe that the generalization from basic structures to unseen structures is indeed difficult: while training with all structures and evaluating will all structures achieve perfect scores on 1S2O and 2S1O of $\text{WikiResNLI}_{\text{EXPLICIT}}$ at 16 shots, training with basic structures are only 58% and 75% accuracies. It is worth noting that all fine-tuning instances have either entailment or contradiction labels, and therefore a random-guessing baseline would be 50% instead of 33.3%.

The generalization performances from explicit respective readings with basic structures to implicit 1S2O and 2S1O are more disappointing. At 16 shots, the accuracies are only 18% and 30%, respectively, well below the chance level. Even full supervision can only achieve around 60% accuracy for both structures. The results indicate that the models do not effectively learn the abstract respective reading relations due to not understanding the commonsense and world knowledge.

We look into the intersection errors of 32-shot, 64-shot and fully-supervised models which are fine-tuned on $\text{WikiResNLI}_{\text{EXPLICIT}}$ and are evaluated on $\text{WikiResNIL}_{\text{IMPLICIT}}$. There are 358 1S2O and 248 2S1O instances that are consistently mistaken by the models. The top-5 frequent properties are: twinned administrative bodies, took place, are capitals of, buried in, and family names. Knowledge about relative location 9(a) and knowledge about humans 9(b) thus seem to play an important role in reasoning with implicit respective readings.

9. (a) **Premise**: Battle of Tours and Battle of Verdun took place in Poitiers and Verdun. **Hypothesis**: Battle of Tours took place in Poitiers and Verdun. **WikiResNLI Label**: contradiction. **Prediction**: entailment

   (b) **Premise**: Theresa of León and Maria Solé Cuñat died in Galicia and Catalonia. **Hypothesis**: Theresa of León and Maria Solé Cuñat died in Galicia. **WikiResNLI Label**: contradiction. **Prediction**: entailment

Impact on other NLI tasks. We evaluate all models fine-tuned with WikiResNLI above on other NLI tasks, i.e, MNLI-m and ANLI-R3, to check whether fine-tuning on such a label-imbalanced dataset hurts performance. Interestingly, full supervision with $\text{WikiResNIL}_{\text{IMPLICIT}}$ of basic structures results in new state-of-the-art performance for DeBERTa. On MNLI-m, the score improves from 90.8% to 91.4%; and on ANLI-R3, the performance raises from 63.6% to 64.1%.

Experiments on LLaMA, FLAN-T5 and GPT-JT Significant advancements in large generative LMs have been achieved in the realm of general natural language understanding. These improvements can be attributed to enhanced training strategies, such as incorporating code and human instructions into pretraining/fine-tuning data and RLHF (Christiano et al., 2017; OpenAI, 2023). We assess the zero-shot and in-context learning abilities of three open-source generative models, that is, LLaMA-7B (Touvron et al., 2023), FLAN-T5-XL (Chung

et al., 2022) and GPT-JT-6B (Wang and Komatsuzaki, 2021; Together, 2022). In this study, our focus is on two representative scenarios, namely generalizing from explicit to implicit readings and generalizing from synthetic to natural readings. We adopt the template *{premise} Question: Does this imply that {hypothesis}?* as it attains top-tier results for NLI tasks (Webson and Pavlick, 2022).

Figure 6 illustrates the explicit to implicit generalization results. Notably, FLAN-T5 achieved a near-perfect score on zero-shot entailment pairs, comparable to the fine-tuned DeBERTa. However, GPT-JT, despite being instruction-tuned on NLI datasets, performed at a mere chance level on entailment pairs, while LLaMA scored below 10% accuracy. In terms of contradiction instances, all three models scored below 60% accuracy, with in-context learning offering limited improvement at the 4-shot level. Specifically, FLAN-T5's performance decreased after in-context learning.

For the generalization from WikiResNLI to NatResNLI, in Figure 7, we observed similar trends as in the previous experiments. FLAN-T5 outperformed the other models on entailment instances, and LLaMA demonstrated significant improvement within a few shots. However, for contradiction pairs, all models experienced only a modest increase in accuracy from 1 to 4 shots, with the highest accuracy remaining below 60%.

To conclude, while large generative models have made significant strides in natural language understanding, they still face substantial challenges in reasoning with respective readings, highlighting the need for further research and development in the long tail of linguistic constructions.

5 Related Work

Logical relations between two sentences are a core aspect of language understanding (Frege, 1879; Heijenoort, 1967; Blackburn et al., 2006). To facilitate large-scale model evaluation, NLP researchers have developed manually labelled NLI corpora, typically for 2/3-way classification (Dagan et al., 2013; Bowman et al., 2015; Williams et al., 2018). In recent years, researchers start to analyze the characteristics of these datasets, such as annotation artefacts (Gururangan et al., 2018), syntactic heuristics (McCoy et al., 2019) and adversarial collection process (Williams et al., 2022).

In computational linguistics, distributive predication has been analyzed through means of distributivity operators (Massey, 1976; Link et al., 1983; Roberts, 1987; Lasersohn, 1998). And linguists have been working on extending first-order logical forms to include distributive and collective readings (Martin, 1981; Alshawi and van Eijck, 1989). Scha and Stallard (1988) present a recursive translation rule scheme to account for multi-level plurals. Aone (1991) proposed a reasoner consisting of domain-dependent constraints and domain-independent axioms for collective and distributive ambiguity. Shaw and McKeown (2000) described a simplified quantifier system to minimize distributive and collective ambiguities.

Respective readings have not yet been studied in modern NLP. Relevant works include plural understanding, which has been studied as a coreference resolution task (Jain et al., 2004; Zhou and Choi, 2018; Yu et al., 2020b). Manshadi et al. (2011) proposed quantifier scope annotation in which plurals are annotated with distributive and collective readings at the constraint level. Yu et al. (2020a) show that LMs are better at reflexive anaphora tasks with distributive than collective constructions.

6 Conclusions

The "respectively" construction is simple yet entails multiple levels of reasoning skills, including syntactic-semantic and commonsense-world knowledge. It is crucial that when an out-of-the-box model cannot reason over it, it should be able to learn with as few examples as possible. We proposed two datasets, WikiResNLI (a controlled synthetic dataset) and NatResNLI (a naturally occurring dataset) to probe their ability to do so in zero-shot and few-shot settings. We find that explicit reasoning is easier to learn than implicit reasoning, and LMs fail to generalize when common sense inference is needed. We confirm that diverse and complex training data are necessary to achieve human-level performance.

7 Limitation

Linguistic studies have shown that respective readings are not necessary to have two coordinate structures in the same sentence (Dalrymple and Kehler, 1995). Both WikiResNLI and NatResNLI have only one sentence in the premise and do not exhaust all possible and complicated realizations of respective readings. However, we are able to discuss and investigate LMs' generalizability with "respectively" with three constructions, i.e., 1S1O,

1S2O and 2S1O.

Our experiments are English-specific and are limited to LMs that can be run with an academic budget. However, our conclusions about the generalizability towards respective readings should be viewed as language-agnostic given there are linguistic constructions under-discussed in many other languages and it is worth researchers' attention to study them.

8 Acknowledgments

We would like to thank the members of the CoAStaL NLP group and the anonymous reviewers for their helpful suggestions.

References

Hiyan Alshawi and Jan van Eijck. 1989. Logical forms in the core language engine. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Chinatsu Aone. 1991. Resolution of collective-distributive ambiguity using model-based reasoning. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Berkeley, California, USA. Association for Computational Linguistics.

Patrick Blackburn, Johan van Benthem, and Frank Wolter. 2006. *Handbook of Modal Logic*. Elsevier.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Lucas Champollion. 2015. Ten Men and Women Got Married Today: Noun Coordination and the Intersective Theory of Conjunction. *Journal of Semantics*, 33(3):561–622.

Rui Pedro Chaves. 2012. Conjunction, cumulation and respectively readings. *Journal of Linguistics*, 48:297 – 344.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Mary Dalrymple and Andrew Kehler. 1995. On the constraints imposed by respectively. *Linguistic Inquiry*, pages 531–536.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Dowty et al. 1987. Collective predicates, distributive predicates, and all. In *Proceedings of the 3rd ESCOL*, pages 97–115. Ohio State University Ohio.

Randall Eggert. 2000. Grammaticality and context with respect to and... and or... respectively. In *Proceedings of the Thirty-Sixth Annual Meeting of the Chicago Linguistics Society (CLS-36)*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gottlob Frege. 1879. *Begriffsschrift: Eine der Arithmetischen Nachgebildete Formelsprache des Reinen Denkens*. Halle a.d.S.: Louis Nebert.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2021. Analogy training multilingual encoders. In *AAAI*.

Jean Mark Gawron and Andrew Kehler. 2004. The semantics of respective readings, conjunction, and filler-gap dependencies. *Linguistics and Philosophy*, 27:169–207.

Grant Goodall. 1987. *Parallel Structures in Syntax: Coordination, Causatives, and Restructuring*. Cambridge University Press.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jean van Heijenoort. 1967. Logic as calculus and logic as language. In *Proceedings of the Boston Colloquium for the Philosophy of Science 1964/1966*, pages 440–446. Springer.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *ArXiv*, abs/2210.11610.

Prateek Jain, Manav Ratan Mital, Sumit Kumar, Amitabha Mukerjee, and Achla M. Raina. 2004. Anaphora resolution in multi-person dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 47–50, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

D Terence Langendoen. 1978. The logic of reciprocity. *Linguistic inquiry*, 9(2):177–197.

Peter Lasersohn. 1998. Generalized distributivity operators. *Linguistics and philosophy*, pages 83–93.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Godehard Link et al. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. *Formal semantics: The essential readings*, 127:147.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mehdi Manshadi, James Allen, and Mary Swift. 2011. A corpus of scope-disambiguated English text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 141–146, Portland, Oregon, USA. Association for Computational Linguistics.

William A. Martin. 1981. Formal roles, co-descriptors, and the representation of quantified English expressions. *American Journal of Computational Linguistics*, 7(3):137–148.

Gerald J Massey. 1976. Tom, dick, and harry, and all the king's men. *American Philosophical Quarterly*, 13(2):89–107.

James D. McCawley. 1968. The role of semantics in a grammar. In Emmon Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 124–169. Holt, Rinehart, and Winston.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Friederike Moltmann. 1992. *Coordination and Comparatives*. Ph.D. thesis, MIT.

Laurer Moritz, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert - nli.

Denis Newman-Griffis, Albert Lai, and Eric Fosler-Lussier. 2017. Insights into analogy completion from the biomedical domain. In *BioNLP 2017*, pages 19–28, Vancouver, Canada,. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Sadayuki Okada. 1999. On the function and distribution of the modifiers respective and respectively. *Linguistics*, 37(5):871–903.

OpenAI. 2023. Gpt-4 technical report.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R.

Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geoffrey K Pullum and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):471–504.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Craige Roberts. 1987. *Modal subordination, anaphora, and distributivity*. University of Massachusetts Amherst.

Remko Scha. 1984. Distributive, collective and cumulative quantification. *Truth, Interpretation, and Information, GRASS*, 2:131–158.

Remko Scha and David Stallard. 1988. Multi-level plurals and distributivity. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Buffalo, New York, USA. Association for Computational Linguistics.

James Shaw and Kathleen McKeown. 2000. Generating referring quantified expressions. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 100–107, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Together. 2022. Releasing GPT-JT powered by open-source AI. https://www.together.xyz/blog/releasing-v1-of-gpt-jt-powered-by-open-source-ai.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Yoad Winter. 2002. *Flexibility principles in Boolean semantics: The interpretation of coordination, plurality, and scope in natural language*. MIT press.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020a. Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020b. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Appendices

A Annotation Details

We employ Amazon Mechanical Turk workers. A qualified worker is one who has completed more than 10,000 HITs and has an approval rate greater than 99%. We set the location to the United States as there was no option to choose language proficiency. They are shown only three examples with entailment, neutral and contradiction labels before annotation. For each premise-hypothesis pair, five workers were asked to annotate the entailment relation (entailment, neutral or contradiction) following the guidelines of Nie et al. (2020). The worker gains a reward of 12 cents. Based on the workers' feedback, our hourly rate ranges between 16 to 27 US dollars, which is above the federal or Californian hourly wage. In total, 170 annotators participated in the step of label annotation of the hypotheses written by the authors. The number of HITs (annotation) per worker ranges from 5 to 200 based on their wishes. We assure to have 5 annotations per each premise-hypothesis pair.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☒ A2. Did you discuss any potential risks of your work?
We provide a NLI dataset and we do not see a risk for that

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
section 3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
section 3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
section 3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 3

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendices

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendices

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.