Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks

Ruixiang Cui, Daniel Hershcovich, and **Anders Søgaard**
University of Copenhagen
{rc, dh, soegaard}@di.ku.dk

Abstract

Logical approaches to representing language have developed and evaluated computational models of quantifier words since the 19th century, but today's NLU models still struggle to capture their semantics. We rely on Generalized Quantifier Theory for language-independent representations of the semantics of quantifier words, to quantify their contribution to the errors of NLU models. We find that quantifiers are pervasive in NLU benchmarks, and their occurrence at test time is associated with performance drops. Multilingual models also exhibit unsatisfying quantifier reasoning abilities, but not necessarily worse for non-English languages. To facilitate directly-targeted probing, we present an adversarial generalized quantifier NLI task (GQNLI) and show that pre-trained language models have a clear lack of robustness in generalized quantifier reasoning.

1 Introduction

Quantifier words—such as *each* or *most* or *more than three*—have been extensively studied, both in logic and in linguistics (Westerståhl, 1989; Peters and Westerståhl, 2006), going all the way back to Frege (1879). In this paper, we examine the extent to which they present a challenge to modern NLU systems. Our analysis is motivated by three observations:

Quantifier words are abstract Unlike nouns, verbs and adjectives, quantifier words do not have referents out in the world. Rather, quantifier words specify relationships between sets of entities, events and properties. To provide intuitions about the semantics of quantifier words, and to be able to refer to quantifiers in a language-independent way, we rely on the notion of generalized quantifiers (Mostowski, 1957), as described in §2.

Quantifier words vary across languages Quantifier word inventories differ across languages.

QA_English	CONTEXT: *A piece of paper was later found on which he had written his last statements in **two** languages, Latin and German. Only **one** statement was in Latin and the rest in German.* QUESTION: *In what language were **most** statements written?* ANSWER: `German` PREDICTED ANSWER: `Latin and German`
NLI_Spanish	PREMISE: ***Más de tres** personas resultaron heridas en un accidente de **dos** vehículos en la Ruta **30** el lunes por la noche. (translation: **More than three** people were injured in a **two**-vehicle crash on Route **30** Monday evening.)* HYPOTHESIS: *Había **4** personas involucradas. (translation: There were **4** people involved.* LABEL: `Neutral` PREDICTED LABEL: `Entailment`

Table 1: Examples of quantifiers (marked in bold texts) in NLP tasks, with RoBERTa's prediction for QA and XLM-R's prediction for NLI after fine-tuning.

Often what is considered rough translation equivalents also differ in syntax, fine-grained semantics or pragmatics. Stateva et al. (2019) show, e.g., that perceptions of the numerical bounds of existential quantifiers differ across speakers of English, French, Slovenian, and German. Other papers showing discrepancies between quantifier systems include comparisons of Salish to English (Matthewson, 2001), Adyghe to English (Nikolaeva, 2012), or of Dutch, Hebrew and Bengali (Gil, 1982). The cross-linguistic differences in how generalized quantifiers are expressed motivates a cross-lingual error analysis, since quantifiers may contribute more to error when processing some languages rather than others.

Quantifier words are important Quantifier words are extremely important for tasks that require inference, including natural language inference, question answering, fact-checking, etc. Datasets have, for example, been developed for numerical reasoning in English (Dua et al., 2019). Several researchers have identified quantifier words as important sources of errors for natural language processing systems (Joshi et al., 2020); see Table 1

Generalized Quantifiers	Logical Denotation	Example
some(A)(B) = 1	$A \cap B \neq \varnothing$	This process is known to increase security in several ways.
all(A)(B) = 1	$A \subseteq B$	Everyone agreed the food was terrible.
more than k the(A)(B) = 1	$\lvert A \cap B \rvert > k$	They do let them go more than twice a week.
less than k the(A)(B) = 1	$\lvert A \cap B \rvert < k$	San Augustin Acolman has less than 1,000 residents.
k (A)(B) = 1	$\lvert A \cap B \rvert = k$	Please donate 100 million to the School of Nursing.
between p and k the(A)(B) = 1	$p < \lvert A \cap B \rvert < k$	The USA added ten states to its nation between 1800 and 1850.
the **p/k** (A)(B) = 1	$\lvert A \cap B \rvert = p \cdot (\lvert A \rvert / k)$	Captain Blood has 20/20 vision.
the **k%** (A)(B) = 1	$\lvert A \cap B \rvert = k \cdot (\lvert A \rvert / 100)$	The lending fund is always guaranteed 9% interest.
most (A)(B) = 1	$\lvert A \cap B \rvert > \lvert A \backslash B \rvert$	Most ZIP Codes cover roughly ten thousand addresses.
few (A)(B) = 1	$\lvert A \cap B \rvert < \lvert A \backslash B \rvert$	Only a few teenagers were still listening to Rock 'n' Roll.
each other (A)(B) = 1	$\forall a \in (A \cap B) \exists b \in (A \cap B)(a \neq b)$	All of these trails are located within the a one hour drive of each other.

Table 2: The categorization set of quantifiers for task analysis. The first six are Aristotelian/counting quantifiers and the following four are proportional quantifiers. The last one is a Ramsey quantifier (Schmerl and Simpson, 1982). For each quantifier, its logical denotation is listed in the second column. The third conlumn contains English examples with quantifiers taken from XNLI.

for examples of such errors. Unfortunately, most efforts have concentrated on subsets of quantifier words and on English.

Contributions We analyze how quantifiers are represented in NLU benchmarks, and how their occurrence at test time contributes to errors by neural language models (LMs). We derive a linguistically motivated 11-way categorization set for generalized quantifiers and look into their distribution in three steps: (a) monolingual NLI; (b) cross-lingual NLI; (c) cross-lingual question answering. We also propose GQNLI[1], an adversarial generalized quantifier NLI challenge dataset. Our work shows that (i) generalized quantifiers are pervasive and cause overall performance drops in NLU benchmarks; (ii) the contribution of quantifier words to system error varies across languages; and (iii) generalized quantifiers are particularly difficult for LMs in interaction with negation and subsumption.

2 Background

Generalized quantifiers (GQs) are developed upon first-order predicate logic, denoting relations between sets (Mostowski, 1957). Given a universe E, a quantifier Q would be treated as a mapping Q_E from the Cartesian product of powersets $\mathcal{P}(E) \times \mathcal{P}(E)$ to the set {*false,true*} or, as a binary relation on subsets of E (Dvořák and Holčapek, 2015). GQs are generalizations of the $\forall$,$\exists$ quantifiers from first-order predicate logic (Mostowski, 1957; Lindström, 1966; Montague, 1973; Bach et al., 1995; Keenan and Paperno, 2012). A generalized quantifier is, abstractly, a relation between sets. Generalized quantifier theory, while developed by logicians, is used by formal linguists to analyze the meaning of quantifier words in combination with referential expressions (Barwise and Cooper, 1981; Higginbotham and May, 1981).

Most human languages contain ways of expressing generalized quantifiers, and their semantics exhibit striking similarities across languages (Matthewson, 2004; Fintel and Matthewson, 2008; Steinert-Threlkeld, 2019). At the same time, generalized quantifiers can be instantiated very differently across languages due to pragmatic considerations (Grice, 1989) or cognitive economy and cost-benefit optimisation in the exchange of information (Levinson et al., 2000; Steinert-Threlkeld, 2021; Uegaki, 2022). Quantifier words also exhibit syntactic differences, e.g., with some languages having specialized words to express quantity, while others rely on metaphorical usage of common nouns (Katsos et al., 2012). In English, *most* is a determiner, but Spanish and French express the same concept through common nouns, *la mayoría* and *la majorité*. The relative stability of the core semantics of quantifiers makes a cross-linguistic comparison possible, but the syntactic and pragmatic variation associated with the expression of generalized quantifiers poses a challenge for multilingual NLU. We consult quantifier taxonomy studies (Keenan and Westerståhl, 1997; Peters and Westerståhl, 2006; Szymanik and Thorne, 2015; Szymanik, 2016) and derive a categorization set for quantifier analysis in NLU benchmarks. In Table 2, we list the 11-way quantifier categorization set and their logical denotation based on set theory.

While other foci of formal linguistics have attracted the attention of NLP researchers—including

[1] https://github.com/ruixiangcui/GQNLI

Quantifier	English						Cross-lingual
	MNLI_m	MNLI_mm	SNLI	ANLI_R1	ANLI_R2	ANLI_R3	XNLI
some	171	132	191	5	1	17	115
all	255	239	65	15	8	29	166
$> k$	14	23	8	10	16	14	16
$< k$	3	3	0	6	7	5	1
k	266	269	988	55	62	48	159
between	2	3	0	3	2	0	1
p/k	1	5	1	1	1	0	2
$k\%$	10	7	0	0	0	1	5
most	35	39	1	0	2	1	9
few	14	15	11	0	0	6	11
each other	4	3	35	0	0	2	5
Total	775	738	1300	95	99	124	499
Frequency	7.9%	7.5%	13.2%	9.5%	9.9%	12.4%	10.0%

Table 3: Quantifier distribution in four NLI tasks, among which three are monolingual English and one is cross-lingual. The table show statistics of the test set, if not available, dev set, of the target task. All but the last rows show the occurrence time of the type of quantifier in the first column. The last row represents the distribution rate of any quantifier in the dataset.

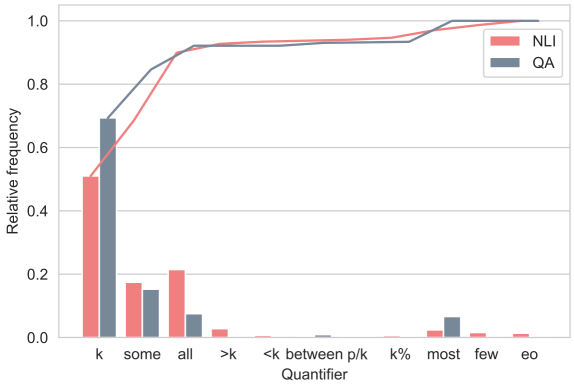

Figure 1: Relative distribution of quantifiers in NLI and QA tasks ranked by semantic complexity. The bars show the relative frequency of such quantifier and the lines indicate the cumulative frequency for a task.

coreference (Ogrodniczuk et al., 2019, 2020), negation (Hossain et al., 2020; Hartmann et al., 2021), and consistency (Li et al., 2019; Ribeiro et al., 2019; Asai and Hajishirzi, 2020; Geva et al., 2022)—there has been little work on generalized quantifiers as a source of error in NLU, let alone in multilingual NLU. It remains an open problem whether LMs represent the semantics of quantifiers words adequately, or if they provide a basis for resolving scopal ambiguities.[2]

3 NLU Benchmarks

We conduct an error analysis focusing on the role of generalized quantifiers in two NLU tasks, Natural Language Inference (NLI) and Question Answering (QA), which generally require understanding of quantifiers. For each type of task, both monolingual and cross-lingual evaluation are conducted. We focus on generalized quantifiers in the *hypotheses* in NLI examples—and on generalized quantifiers in the *question* fields in question answering. To this end, we identify quantifiers by the lemma and the universal dependency relation (Nivre et al., 2020) of a quantifier after preprocessing the sentences using *Stanza* (Qi et al., 2020). Take the sentence "The Yiddish culture has survived for more than a thousand years.", we annotate it as "The/*det* Yiddish/*amod* culture/*nsubj* have/*aux* survive/*root* for/*case* more/*advmod* than/*fixed* a/*det* thousand/*nummod* year/*obl* ./*punct*". By matching the regex pattern of the quantifier "more than k", in this case *"((more|great)∨advmod than∨(fixed|case)|at∨case least∨nmod) .+∨nummod .+∨(nsubj|obj|obl)"*, we approximate the surface form of the type "more than k".Through matching quantifier patterns, we are able to find entries in which quantifiers are instantiated. See Appendix A for the list of regex patterns we write to identify GQs. In Table 3 and Table 6, we present the statistics of the quantifier distributions in NLI and QA tasks, respectively. As can be seen, quantifiers are indeed widespread in NLU tasks, accounting for roughly 10% in NLI tasks and 5% in QA tasks. We will further discuss the statistics and experiments in the following section.

4 Quantifiers in English NLI Benchmarks

NLI is commonly framed as a three-way classification task with labels *entailment*, *contradiction* and *neutral* (Bowman et al., 2015a). While SOTA models exhibit low error rates on NLI benchmarks, it is unclear when they succeed or fail in their underlying reasoning. We are interested in whether generalized quantifers challenge modern NLI models. In our error analysis, we initially focus on three English NLI datasets, MultiNLI (MNLI; Williams et al., 2018), SNLI (Bowman et al., 2015a) and ANLI (Nie et al., 2020) as testbeds.

Table 3 presents statistics of quantifier distri-

[2]Note that generalized quantifiers are not always *explicit* in discourse. The sentence *inadequate sleep causes obesity* should be interpreted as *Most of those who do not sleep adequately, gain weight* (Zadeh, 1983). Such implicit quantifiers related to pragmatic variation are important for language understanding, but will be ignored in this work.

Quantifier	BERT							RoBERTa						
	M_m	M_mm	SNLI	A_R1	A_R2	A_R3	*weig.*	M_m	M_mm	SNLI	A_R1	A_R2	A_R3	*weig.*
some	**82.5**	**84.1**	**86.9**	100	**0**	47.1	**83.4**	**83**	**84.8**	**86.9**	100	100	**41.1**	**83.7**
all	**85.9**	88.3	**89.2**	**46.7**	**37.5**	**34.5**	**83.2**	**85.9**	92.1	**90.8**	**66.7**	**37.5**	**34.5**	**85.3**
$> k$	**85.7**	100	**87.5**	70	**43.8**	**42.9**	**73**	**85.7**	91.3	**87.5**	80	**37.5**	**28.5**	**68.2**
$< k$	100	100		**33.3**	57.1	80	**66.7**	100	100		83.3	85.7	100	91.7
k	87.2	**81.8**	92.4	**43.6**	**43.5**	**33.3**	84.8	88.3	**88.8**	92.9	**56.3**	61.3	**43.8**	87.8
between	100	100		66.7	50		**80**	100	**66.7**		**66.7**	50		**70**
p/k	100	**60**	100	100	100		**77.8**	100	**80**	100	100	**0**		**77.8**
$k\%$	90	100				100	94.4	**70**	**85.7**				**0**	**72.2**
most	**74.3**	**79.5**	**0**		50	**0**	**74.4**	**77**	**87.2**	100		59	**0**	**80.9**
few	**78.6**	**73.3**	**90.9**			**33.3**	**73.9**	**85.7**	**80**	**90.9**			**33.3**	**78.3**
each other	**75**	100	**85.7**			50	**84.1**	**50**	100	**88.6**			50	**84.1**
all GQs	**85**	**84.8**	**91.2**	**50.5**	**44.4**	**39**	**83.3**	**85.4**	**88.8**	**91.7**	**65.3**	56.5	**40.3**	**85.5**
full	86.5	86.1	91.3	58.6	48	43.2	84.4	89.5	89.4	92.3	71.7	49.6	49	87.3

Table 4: BERT and RoBERTa performance on NLI tasks. The *weig.* column represents the percentage of all true predictions in six subtasks over total instances. The penultimate row stands for the overall performance when quantifiers exist in a dataset. The last row reports the overall performance in a dataset. Number marked in bold signifies a lower score than the overall performance.

bution in these datasets, where we observe that, across, about 10% of all hypotheses contain quantifier words, indicating the pervasiveness of quantification. We also plot the frequency of quantifiers in NLI in Figure 1 and find the quantifier word distribution follows Zipf's law (Zipf, 1949). Note the top three most common quantifiers account for more than 90% of all.

Experiments and Results In order to investigate whether NLU systems can solve quantifiers in NLI, we experiment with two pretrained LMs: BERT[3] (Devlin et al., 2019) and RoBERTa[4] (Liu et al., 2019). We use the codebase by Nie et al. (2020). The training data combines SNLI, MNLI, FEVER-NLI (Nie et al., 2019) and ANLI.

In Table 4, we report the test set performance on SNLI and ANLI, and the dev set performance on MLNI *matched* and *mismatched* sections. We can observe that SOTA models suffer from performance drops across almost all quantification phenomena in every task. When it comes to performance over all quantifiers, the improvement from RoBERTa to BERT (2.2%) is less prominent than that over full datasets (2.9%), suggesting RoBERTa is particularly challenged.

Taking a closer look at error by category, proportional quantifiers seem harder to solve than Aristotelian/counting quantifiers. Except for *k%*, all proportional quantifiers—*p/k*, *most*, and *few*—are about 10% lower than the five counting quantifiers (except *less than k*) with BERT; and about 5% lower with RoBERTa. RoBERTa is not generally superior to BERT; e.g., for *k%*, BERT outperforms it by 22%. We show a pairwise analysis of how GQs affect performance when they appear in both the premises and hypotheses in the Appendix B. Generally, our results attest to the difficulty of resolving GQs in NLI benchmarks.

5 Quantifiers in Cross-lingual NLU Benchmarks

Quantifiers are acquired in similar orders across languages (Katsos et al., 2016), although languages express quantifiers in different ways. For example, there are eight different universal quantifiers with different level of distributivity in Malagasy (Matthewson, 2008). This poses challenges to training multilingual LMs and transfer learning. We are interested in whether quantifiers are universally and evenly challenging for all languages.

Quantifiers in Cross-lingual NLI We choose XNLI (Conneau et al., 2018), a manual translation of the development and test set of MNLI into 15 languages, for this multilingual error analysis. We should clarify that for XNLI, the authors annotate entailment labels for the English data only and apply them to the other languages. We do not assume label changes due to translation in this study, but it is worth investigate in the future. We choose five languages belonging to different language families, namely Arabic, Chinese, German, Spanish and Vietnamese as targets. The last column in Table 3 shows the numbers of quantifiers in XNLI. The distribution rate is 10%. Note that the universal

[3] `wwm_cased_L-24_H-1024_A-16`
[4] `roberta-large`

Quantifier	mBERT							XLM						
	en	zh	es	ar	vi	de	*weig.*	en	zh	es	ar	vi	de	*weig.*
some	85.2	**69.6**	80	**63.5**	**67.8**	74.8	73.4	85.2	70.3	79.1	71.3	73.9	**69.6**	**69.6**
all	**80.1**	**65.7**	**72.8**	69.3	**63.9**	74.1	**70.9**	**82.5**	**62.7**	74.1	67.5	71.7	73.5	**72**
$> k$	87.5	**50**	**68.8**	**43.8**	**56.2**	**62.5**	**61.6**	**81.2**	**62.5**	**56.2**	**62.5**	**50**	75	75
$< k$	100	100	100	100	100	100	100	100	100	100	100	100	100	100
k	86.2	**69.1**	80.5	71.7	76.7	82.4	77.7	83	66.7	78.6	71.7	74.2	81.1	75.8
between	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p/k	100	**50**	100	100	100	100	91.7	100	**0**	100	100	**50**	**50**	**66.7**
$k\%$	100	100	80	100	100	100	96.7	**80**	80	80	100	100	80	86.7
most	**55.6**	**55.6**	**66.7**	66.7	**33.3**	66.7	**57.4**	**55.6**	**33.3**	**66.7**	**55.6**	**44.4**	77.8	**55.6**
few	**72.7**	**54.5**	**72.7**	**63.6**	**45.5**	72.7	**63.6**	**63.6**	**36.4**	**54.5**	**63.6**	**54.5**	72.7	**57.5**
each other	**60**	**60**	**60**	**60**	80	80	**66.7**	**80**	**20**	**60**	**20**	**40**	**60**	**46.7**
all GQs	83	**67.1**	76.7	68.1	**68.3**	76.9	73.3	**82.4**	**64.2**	**75.7**	69.3	**71.4**	74.8	73
comp.	82.6	88.9	74.7	65.6	70.7	71.4	72.4	83.1	64.8	76.3	66.9	71.6	71.3	72.3

Table 5: Results of mBERT and XLM performance on XNLI tasks decomposed by quantifier categories.

quantifier is the most common quantifier in XNLI.

We fine-tune mBERT[5] (Devlin et al., 2019) and XLM[6] (Lample and Conneau, 2019) on the MNLI training set and evaluate them on XNLI. We report the results in Table 5. We find that performance varies across languages. For Chinese and Vietnamese, we see significant drops in performance for examples with GQs, whereas for Arabic and German, we see improvements. The results *per* quantifier are more homogeneous, however.

Similar to our results for English, we can see that the lowest accuracies in XNLI are with proportional quantifiers, such as *most* and *few*. But the gap in non-English languages is wider for these two categories, especially for Chinese, the difference reaches 30%. Other hard quantifiers include *all*, $> k$, $< k$, and *each other*.

Quantifiers in Cross-lingual QA Cross-lingual question answering (XQA) is another important NLU task that evaluates the cross-lingual transferability of LMs. We evaluate the effect of quantifiers on system errors across two XQA datasets, namely XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020). As demonstrated in Figure 1, quantifier word distributions in XQA tasks also follow Zipf's law, as in NLI tasks, but *k* is more frequent (perhaps because of a traditional emphasis on numerical reasoning), and we see less variance across languages. This is probably because question answering is targeting quantification less directly. To evaluate cross-lingual QA performance on GQs, we fine-tune mBERT and XLM-R[7] (Conneau et al., 2020) using Hu et al. (2020)'s architecture. We present results for mBERT in Table 7; for XLM-R results, please refer to Appendix D.

Quantifier	MLQA						XQuAD
	en	zh	es	ar	vi	de	...
some	66	39	41	44	37	33	12
all	31	14	26	21	19	16	7
$< k$	1	0	0	0	1	0	0
k	322	168	166	195	204	149	32
between	4	2	2	2	3	0	3
p/k	1	1	1	0	0	0	0
$k\%$	1	1	0	1	0	0	0
most	27	19	11	30	17	9	5
Total	453	244	247	293	281	207	59
Frequency	3.9%	4.7%	4.7%	5.4%	5.1%	4.5%	5.0%

Table 6: Quantifier distribution in two multilingual QA tasks, MLQA and XQuAD. We choose six common languages apprearing in both tasks to facilitate comparisons. XQuAD is strictly parellel while MLQA is not, hence only the latter has statistics by languges. Categories that no entry exists are omitted.

Just as with XNLI, LMs suffer from performance drops across all languages for almost all GQ phenomena with significant, cross-lingual variation. The most distinguished is that Exact Match (EM) suffers from a greater deterioration than F1 scores for all languages. For example, the weighted EM difference for mBERT on MLQA is 2.9% while the weighted F1 is 1%. As one example in Table 1, we observe that the plausible answers selected by models, while being incorrect, result in a sharper decrease of EMs comparing to F1s. Questions containing GQs also tend to have less verbal answers comparing to those without GQs, and therefore require higher precision.

Regarding cross-lingual comparisons, Chinese and Arabic are the two languages that do not have

[5] `multi_cased_L-12_H-768_A-12`
[6] `xlm-mlm-100-1280`
[7] `xlm-roberta-large`

Quantifier	XQuAD														MLQA													
	en		zh		es		ar		vi		de		*weighted*		en		zh		es		ar		vi		de		*weighted*	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
some	75	84.2	50	**55.5**	58.3	76.1	50	**50**	16.6	42.4	**33.3**	**43.8**	47.2	58.7	**59**	80	**28.2**	**52.1**	**34.1**	**59.2**	36.3	54.9	**5.4**	**24**	**33.3**	58.4	**32.7**	**54.8**
all	**28.5**	**62.2**	**14.2**	**35.2**	**28.5**	82	42.8	**52.3**	**14.2**	**29.4**	**28.5**	**56**	**26.1**	**52.9**	67.7	**79.8**	14.2	**46.4**	**38.4**	**62.8**	33.3	57.9	**10.5**	30.1	**31.2**	**51.6**	**32.6**	**54.8**
$< k$															**0**	**0**							0	**13.3**			**0**	6.7
k	78.1	90.1	68.7	80.4	56.2	**72.1**	**40.6**	64.3	**12.5**	35.7	56.2	77.1	52.1	70	74.9	**79.4**	47	63.4	**41.5**	**65.9**	**27.6**	50.3	**6.3**	**23.7**	**38.2**	**53**	39.3	**56**
between	100	100	**33.3**	**72.2**	66.6	93.3	100	100	**0**	**19**	**0**	**56.5**	50	73.5	**50**	88.5	50	83.3	**0**	**26.6**	**0**	68.7	**0**	**26.6**			**20**	58.7
p/k															100	100	**0**	**0**	**0**	**0**							**33.3**	**33.3**
$k\%$															100	100	**0**	**26.6**			**0**	**23.7**					**33.3**	**50.1**
most	**40**	**53.3**	**40**	**40**	**0**	**10**	**0**	**26.6**	**0**	**0**	**20**	**49.3**	**16.7**	**29.9**	**55.5**	**76**	47.3	62.1	**45.4**	**61.7**	30	**46.8**	**5.8**	**15.7**	**33.3**	**40.7**	**36.2**	**50.3**
all GQs	**70**	**83.2**	55	66.7	**50**	**70.3**	41.6	58.2	**11.6**	32.5	**43.3**	**65**	**45.3**	62.7	**63.5**	**79.2**	41.8	60.3	**39.6**	**63.7**	**29.3**	51.3	**6.4**	**23.6**	**36.1**	**53.2**	**36.1**	**55.2**
comp.	71.8	83.7	48	59.1	56	74.5	40.8	57.9	13.9	32.4	50.7	67.2	46.9	62.5	67.2	80.6	37.5	57.9	47.3	66	30	48.4	11.2	28	40.8	56	39	56.2

Table 7: Results of mBERT performance on XQA tasks decomposed by quantifier categories.

lower performance over GQs compared to the performance over the complete dataset. Despite the overall trends, subtle differences from XNLI performance still exist. For example, XLM-R is worse than mBERT on quantifier reasoning on XQuAD Chinese, especially at proportional quantifiers, but this is not the case on MLQA Chinese.

6 GQNLI

We have seen how quantifiers present challenges to NLI and QA models. Using an approach similar to ANLI (Nie et al., 2020) and DynaBench (Kiela et al., 2021), we use model difficulty (RoBERTa's) as a heuristic to select hard examples for a challenge dataset that can hopefully be used to evaluate any future progress on this. We propose GQNLI, a generalized quantifier NLI challenge dataset, consisting of 30 premises and 300 hypotheses. The average sentence lengths of hypothesis and premises are 15.97 and 7.35, respectively. Both numbers are comparable to those of MNLI, but lower than ANLI's (Williams et al., 2020). It should be noted that GQNLI is designed for evaluating future models; obviously not for benchmarking RoBERTa.

Dataset Creation Firstly, we manually create 100 premise-hypothesis pairs, in which various types of GQs appear. For each premise and hypothesis, the number of GQs varies from one to three. To choose the premises, we randomly sampled 100 premises with GQs from SNLI and ANLI test sets, respectively, and selected 10 premises in total, that we consider are semantically adequate for adding GQs and making simple hypotheses.

To construct the hypotheses, we rely on RoBERTa fine-tuned on MNLI and manually select examples about which the model is unsure or incorrect. To focus on GQs, we keep the challenge examples otherwise simple (Ribeiro et al., 2020), and avoid lexical variations in the hypotheses. Hard examples were found to be characterized by (i) mixing generalized quantifiers with other logical operators, such as subsumption or negation, and (ii) combining multiple different generalized quantifiers. We discuss these observations in Section 7.

Two of the authors annotated the examples. The inter-annotator agreement (Fleiss' kappa) was 0.895, substantially higher than ANLI's (0.672–0.740). It is worth noting that the level of semantic or pragmatic interpretation difference of GQs is reflected in the measurement.

We augmented the examples by substituting non-quantifier words (e.g., replacing "dogs" with "cats") while keeping the labels, to exclude the effect of specific lexical items. The resulting labels are uniformly distributed. Table 8 presents GQNLI statistics. Since the dataset is curated to probe the ability to reason with quantifiers, the distribution of generalized quantifiers does not follow Zipf's law; see §4. A list of GQNLI examples per category is shown in Appendix E.

Experiments and Results We evaluate seven types of models on GQNLI, fine-tuned with different combinations of NLI datasets. As data creation only relied on RoBERTa and MNLI, nothing prevents that models with different architectures and training data will perform well. They do not, however. The results are shown in Table 8.

We see that all models have great difficulty with GQNLI. With more training data, models improve, but the best performance is 42.7%, less than 10 points above chance level. In general, the counting quantifiers, especially the existential and universal quantifiers, are easier than proportional quantifiers. Particularly, most models struggle with *less than k* and *few*. This is in some contrast with the NLU tasks studied above, where these quantifiers were among the easiest.

We also observe unstable GQ reasoning ability in simple word substitution cases. For instance, it happens for DeBERTa fine-tuned with M, F, Ling, DocNLI that it predicted correctly the contradiction

Quantifier # Occurrence		some 27	all 51	$> k$ 51	$< k$ 33	k 170	between 21	p/k 24	$k\%$ 45	most 18	few 9	each other 36	*Overall* 485
Model	**Training Data**							**% Performance**					
ELECTRA	S,M,F,ANLI	25.9	15.7	19.6	27.3	24.1	71.4	0	26.7	11.1	11.1	22.2	26.7
BERT	S,M,F,ANLI	40.7	41.2	33.3	30.3	30.6	14.3	37.5	22.2	**61.1**	22.2	**41.7**	30
SBERT	S,M,F,ANLI	7.4	29.4	27.5	21.2	34.7	57.1	62.5	40	22.2	11.1	33.3	34.3
RoBERTa	MNLI	55.6	25.5	17.6	27.3	24.7	23.8	45.8	17.8	33.3	33.3	11.1	28.2
	S,M,F,ANLI	40.7	23.5	41.2	**45.5**	40	57.1	25	33.3	27.8	0	22.2	37
BART	MNLI	51.9	27.5	25.5	36.4	31.2	23.8	50.0	13.3	44.4	**66.7**	30.6	31.3
	S,M,F,ANLI	37.0	39.2	**58.8**	33.3	**47.6**	**76.2**	25.0	**53.3**	16.7	33.3	30.6	41.7
ALBERT	S,M,F,ANLI	44.4	37.3	49	27.3	44.1	57.1	50.0	46.7	16.7	0	33.3	42
DeBERTa-v3	MNLI	48.1	37.3	41.2	12.1	40	33.3	**66.7**	20	50	33.3	**41.7**	40.7
	M,F,ANLI	51.9	41.2	23.5	33.3	37.1	57.1	50.0	13.3	**61.1**	33.3	33.3	41.3
	M,F,Ling,DocNLI	**66.7**	**52.9**	35.3	**45.5**	44.1	66.7	50	31.1	50	33.3	**41.7**	**42.7**

Table 8: GQNLI statstics and seven types of models' performance with different combinations of training data. The second row shows the occurrence time of the type of GQ in GQNLI. The following rows show models' performance on the dataset. We tested most competitive models fine-tuned for NLI available on Hugging Face. All but ALBERT (`xxlarge`) and DeBERTa-v3 (`base`) are size `large`. S, M, F, Ling, A, DocNLI refer to SNLI, MNLI, Fever-NLI, LingNLI (Parrish et al., 2021), ANLI and DocNLI (Yin et al., 2021), respectively. Numbers in bold represent the highest accuracy in one category. Due to space limitation we provide the link to each model in the Appendix H.

relation between "There are six children standing on top of a yellow mountain. Two thirds wear red tops and one third wear green." and "Between 80% and 90% children do not wear red tops.", but incorrectly when "red" is substituted with "beige" and "green" with "cyan". We are yet to study what kind of cues lead to the instability. Our experiments suggest a lack of testing proportionality reasoning and robustness in existing benchmarks.

7 Discussion

Negation The interaction between negation words and quantifiers increases semantic complexity (Partee, 1970; Horn, 2010). We investigate whether this holds for NLI tasks, using negation cue detection to find all cases where a negation word and a quantifier appear in the hypotheses.

We break down the performances on negation of the seven models in Appendix F. As indicated, LMs overall have worse results for negation cases comparing to the entire dataset. We can see a majority of the models even predicted opposite labels for some GQ categories, with 0% accuracy. DeBERTa-v3 is no longer the best model, replaced by BERT and BART. The improvement by training with more data becomes unstable for reasoning over GQs with negation.

For a cross-lingual investigation of the interaction of GQs and negation, we find that in XNLI, the number of cases combining both phenomena is insufficient: we identified four such cases, involving only the quantifiers "all" and "more than." For English, mBERT predicted two cases successfully. For Chinese, German, Vietnamese and Arabic, one is correct. For Spanish, all are wrongly predicted.

It is evident that NLU models suffer from reasoning difficulties in certain cases when negation interacts with GQs, especially in cross-lingual evaluation. In future work, we are interested in expanding GQNLI to more instances and more languages to facilitate qualitative investigations.

Subsumption In generalized term subsumption languages (TSLs; Yen, 1991; Ali and Shapiro, 1993), a term a subsumes another term b if and only if the extension of a is a superset of the extension of b . Rather than surface number comparison, subsumption reasoning requires knowledge of the relations between supersets and subsets. For example, to decide whether "There are six dogs. Three brown dogs, a black dog and a white dog run along the green grass" entails "One dog sits", LMs should be aware that "six dogs" is a superset of the extension of the "brown dogs", "black dog" and "white dog". Another example in GQNLI is to infer whether "There are twelve singers on a stage, less than half from Argentina and one from Cape Verde" entails "Several singers do not come from Chile".

We annotate 63 cases out of the first 100 in GQNLI requiring subsumption reasoning. We show the statistics and results regarding subsumption in Appendix G. It can be seen that more training data leads to higher accuracies. Especially, DeBERTa fine-tuned with DocNLI, which unifies

the two classes "neutral" and "contradict" into a new class "not entail", has a significant improvement on subsumption cases with neutral label. The training bias give an advantage to the model on the subsumption subset, half cases of which are labelled neutral. But such bias has a negative effect on non-subsumption cases; the accuracy drops by 16.2% comparing to the model without training with DocNLI. It is worth investigating whether DocNLI is truly helping subsumption reasoning in future work. Subsumption is a key concept in the study of knowledge representation (Woods, 1991), but is neglected in current NLP research. The fact that LMs struggle to perform subsumption reasoning asserts the necessity to explicit tackle the problem.

8 Related Work

We examine the sensitivity of NLU models to generalized quantifiers. These models are designed to induce correlations from large volumes of data, not to reason symbolically with logical quantifiers. Such models have, nevertheless, been probed for logical knowledge.

Mul and Zuidema (2019), for example, show neural networks encode fragments of first-order logic and exhibit zero-shot generalization ability. Evans et al. (2018) present a neural architecture that improves performance on propositional logical inference. Bowman et al. (2015b) also suggest neural networks learn semantic representations for logical inference in natural languages. However, on the same task, Veldhoen and Zuidema (2017) find neural networks fail to do so on a more stringent test. Geiger et al. (2019) also show that neural networks fail to exhibit robust logical inference. Srivastava et al. (2018) use semantic parsers to encode quantifiers and improve zero-shot learning in classification tasks. Haruta et al. (2020) present a system that computes logical inference over GQs and see improvements on two specialized datasets, FraCaS (Cooper et al., 1994) and MED (Yanaka et al., 2019). None of these papers explicitly discussed generalized quantifiers, and all were limited to studying the ability of neural networks to capture the logical semantics of English.

Many studies have instead focused on LMs' ability to capture negation (Gururangan et al., 2018; Naik et al., 2018; Hossain et al., 2020; Ettinger, 2020; Hartmann et al., 2021) or coreference (Ye et al., 2020; Varkel and Globerson, 2020; Abdou et al., 2020). Others have focused on LMs' ability to reason with numbers (Johnson et al., 2020). DROP (Dua et al., 2019), for example, is a question answering dataset designed specifically to probe LMs' ability to count, add and subtract for answering factoid questions. Models have also been tailored for numerical reasoning (Geva et al., 2020; Zhang et al., 2020). Cobbe et al. (2021) proposes to use a verification task during pretraining of LMs to improve their ability to solve math word problems. Others have studied monotonicity inference (Hu et al., 2019; Yanaka et al., 2019, 2020), and Fang and Lou (2021) recently focused on the two quantifier words *part* and *whole* in an error analysis for named entity recognition.

Many NLU benchmarks contain quantifier words, but their influence on performance has not been studied systematically. One exception to this is that generalized quantifiers have been used to generate adversarial examples in the context of numerical reasoning (Naik et al., 2018; Nie et al., 2020). TaxiNLI (Joshi et al., 2020), which categorizes 15 types of reasoning abilities, is a dataset drawn from MNLI. In their taxonomy, the Quantifier category only refers to universal and existential quantifiers, *not* to generalized quantifiers, and ditto for Kim et al. (2019). All of the above focused on English, but in an extension to TaxiNLI, K et al. (2021) incorporated quantifiers into the Logic class and found a large cross-lingual transfer gap on LMs.

9 Conclusion

Quantifiers lie in the intersection of logic, linguistics and NLP research. It is essential for NLU systems to learn quantifier reasoning. We examined generalized quantifiers in multilingual NLU tasks with regards to their expressiveness and logical reasoning requirement. Our survey and experiments indicate quantifiers are neglected to a degree and cause significant performance drops for neural LMs. To better understand LMs' reasoning abilities, we release GQNLI, a novel generalized quantifier NLI challenge dataset. With the pervasiveness of generalized quantifiers, we stress that more efforts are necessary to investigate: (1) when and why models systematically fail when quantifiers interact with other operators; (2) how to improve cross-lingual transferability of quantifiers; (3) how we can exploit the theoretical results about generalized quantifiers from logic and linguistic studies,

so as to improve the logical inference ability of neural LMs.

Acknowledgements

We would like to thank Miryam de Lhoneux, Constanza Fierro, Desmond Elliott and the anonymous reviewers for their valuable feedback.

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online. Association for Computational Linguistics.

Syed S Ali and Stuart C Shapiro. 1993. Natural language processing using a propositional semantic network with structured variables. *Minds and machines*, 3(4):421–451.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Emmon W. Bach, Eloise Jelinek, Angelika Kratzer, and Barbara H. Partee. 1995. Quantification in natural languages.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence*, pages 241–301. Springer.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. 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.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015b. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. Fracas: A framework for computational semantics. *Deliverable D6*.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Antonín Dvořák and Michal Holčapek. 2015. Type 1 , 1 fuzzy quantifiers determined by fuzzy measures on residuated lattices. part iii. extension, conservativity and extensionality. *Fuzzy Sets Syst.*, 271(C):133–155.

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

Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. 2018. Can neural networks understand logical entailment? *arXiv preprint arXiv:1802.08535*.

Lei Fang and Jian-Guang Lou. 2021. Part & whole extraction: Towards a deep understanding of quantitative facts for percentages in text. *ArXiv*, abs/2110.13505.

Kai Fintel and Lisa Matthewson. 2008. Universals in semantics.

Gottlob Frege. 1879. Begriffsschrift. eine der arithmetischen nachgebildete formelsprache des reinen denkens.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4485–4495, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Mor Geva, Tomer Wolfson, and Jonathan Berant. 2022. Break, Perturb, Build: Automatic Perturbation of Reasoning Paths Through Question Decomposition. *Transactions of the Association for Computational Linguistics*, 10:111–126.

David Gil. 1982. Quantifier scope, linguistic variation, and natural language semantics. *Linguistics and Philosophy*, 5(4):421–472.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard 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.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 263–270, Online. Association for Computational Linguistics.

James Higginbotham and Robert May. 1981. Questions, quantifiers and crossing.

Laurence R. Horn. 2010. The expression of negation.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Hai Hu, Qi Chen, and Larry Moss. 2019. Natural language inference with monotonicity. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 8–15, Gothenburg, Sweden. Association for Computational Linguistics.

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 generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Devin Johnson, Denise Mak, Andrew Barker, and Lexi Loessberg-Zahl. 2020. Probing for multilingual numerical understanding in transformer-based language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 184–192, Online. Association for Computational Linguistics.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Napoleon Katsos, Chris Cummins, Maria-José Ezeizabarrena, Anna Gavarró, Jelena Kuvač Kraljević, Gordana Hrzica, Kleanthes K Grohmann, Athina Skordi, Kristine Jensen De López, Lone Sundahl, et al. 2016. Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, 113:9244 – 9249.

Napoleon Katsos, Maria-Jose Ezeizabarrena, Anna Gavarro, Jelena Kuvac Kraljevic, Gordana Hrzica, Kleanthes K Grohmann, Athina Skordi, Kristine Jensen de Lopez, Lone Sundahl, Angeliek Van Hout,

et al. 2012. The acquisition of quantification across languages: some predictions.

Edward L. Keenan and Denis Paperno. 2012. Handbook of quantifiers in natural language.

Edward L. Keenan and Dag Westerståhl. 1997. Generalized quantifiers in linguistics and logic. In *Handbook of Logic and Language*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Stephen C Levinson, C Stephen, and Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Per Lindström. 1966. First order predicate logic with generalized quantifiers. *Theoria*, 32(3):186–195.

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*, abs/1907.11692.

Lisa Matthewson. 2001. Quantification and the nature of crosslinguistic variation. *Natural Language Semantics*, 9(2):145–189.

Lisa Matthewson. 2004. On the methodology of semantic fieldwork. *International journal of American linguistics*, 70(4):369–415.

Lisa Matthewson. 2008. Quantification: A cross-linguistic perspective.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

Andrzej Mostowski. 1957. On a generalization of quantifiers. *Fundamenta Mathematicae*, 44(1):12–36.

Mathijs Mul and Willem Zuidema. 2019. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. 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.

Liudmila Nikolaeva. 2012. *Quantifiers in Adyghe*, pages 21–82. Springer Netherlands, Dordrecht.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Maciej Ogrodniczuk, Vincent Ng, Yulia Grishina, and Sameer Pradhan, editors. 2020. *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Barcelona, Spain (online).

Maciej Ogrodniczuk, Sameer Pradhan, Yulia Grishina, and Vincent Ng, editors. 2019. *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Minneapolis, USA.

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.

Barbara Hall Partee. 1970. Negation, conjunction, and quantifiers: Syntax vs. semantics. *Foundations of Language*, 6(2):153–165.

Stanley Peters and Dag Westerståhl. 2006. *Quantifiers in language and logic*. Oxford University Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

James H. Schmerl and Stephen G. Simpson. 1982. On the role of ramsey quantifiers in first order arithmetic. *J. Symb. Log.*, 47:423–435.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Penka Stateva, Arthur Stepanov, Viviane Déprez, Ludivine Emma Dupuy, and Anne Colette Reboul. 2019. Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology*, 10:957.

Shane Steinert-Threlkeld. 2019. Learnability and semantic universals. *Semantics and Pragmatics*, 12:4.

Shane Steinert-Threlkeld. 2021. Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, 23(10).

Jakub Szymanik. 2016. *Cognitive Processing of Quantifiers*, pages 51–83. Springer International Publishing, Cham.

Jakub Szymanik and Camilo Thorne. 2015. Semantic complexity of quantifiers and their distribution in corpora. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 64–69, London, UK. Association for Computational Linguistics.

Wataru Uegaki. 2022. The Informativeness/Complexity Trade-Off in the Domain of Boolean Connectives. *Linguistic Inquiry*, pages 1–39.

Yuval Varkel and Amir Globerson. 2020. Pre-training mention representations in coreference models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8534–8540, Online. Association for Computational Linguistics.

Sara Veldhoen and Willem Zuidema. 2017. Can neural networks learn logical reasoning? *CLASP Papers in Computational Linguistics*, page 34.

Dag Westerståhl. 1989. Quantifiers in formal and natural languages. In *Handbook of philosophical logic*, pages 1–131. Springer.

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. 2020. Anlizing the adversarial natural language inference dataset.

William A. Woods. 1991. Understanding subsumption and taxonomy: A framework for progress. In *Principles of Semantic Networks*.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand

monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

John Yen. 1991. Generalizing term subsumption languages to fuzzy logic. In *IJCAI*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Lotfi A Zadeh. 1983. A computational approach to fuzzy quantifiers in natural languages. In *Computational linguistics*, pages 149–184. Elsevier.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort.

Generalized Quantifiers	Regular Expressions
some(A)(B) = 1	(some|several|much|many)\/det .*\/(nsubj|obj|obl)|(some|several|much|many)\/nsubj|(some|several|much|many)\/amod \w+\/nsubj:pass
all(A)(B) = 1	(every|all|each)\/det .*\/(nsubj|obj|obl)|all\/det:predet .*\/(nsubj|obj|obl) | everything|everyone|everybody
more than k the(A)(B) = 1	((more|great)\/advmod than\/(fixed|case)|at\/case least\/nmod) .+\/nummod .+\/(nsubj|obj|obl)
less than k the(A)(B) = 1	((few|less)\/advmod than\/(fixed|case)|at\/case most\/amod) .+\/nummod .+\/(nsubj|obj|obl)
k (A)(B) = 1	\w+\/nummod .+\/(nsubj|obj|obl)
between p and k the(A)(B) = 1	between\/case \w+\/(nummod|nsubj|obj|obl) and\/cc \w+\/conj|between\/case .+\/(nummod|nsubj|obj|obl) %\/obl
the **p/k** (A)(B) = 1	\d+\/\d+\/(nummod|nsubj|obj|obl)|half\/nummod|third\/(nsubj|obj|obl)|fourth\/(nsubj|obj|obl)|fifth\/(nsubj|obj|obl)
the **k%** (A)(B = 1)	\d+\/nummod %\/(nsubj|obj|obl)
most (A)(B) = 1	most\/amod \w+\/(nsubj|obj|obl)|most\/nsubj:pass of\/case .+\/nmod
few (A)(B) = 1	few\/amod \w+\/(nsubj|obj|obl)|few\/nsubj:pass of\/case .+\/nmod
each other (A)(B) = 1	each\/det other\/(nsubj|obj|obl)

Table 9: Regular Expressions for generalized quantifiers.

Appendices

A Regular Expressions for Generalized Quantifiers

Table 9 lists the regex we use to parse generalized quntifiers in sentences augmented with universal dependency tags. The approach does not find all the generalized quantifiers exhuastively but rather approximates the common distributions.

B Pairwise Observation

While the analysis in Section 4 is based on quantifiers in hypotheses, next we consider the interaction of quantifiers in hypotheses and quantifiers in premises. To this end, we calculate the difference between overall performance and performance for premise-hypothesis pairs of GQs. In Figure 2, we visualize the results as heatmaps (see Table 10 for exact numbers of occurences and accuracies). Surprisingly, whenever quantifiers appear in both the premise and the hypothesis, LMs largely fail to predict the entailment. Percentage quantifiers, supposed to be semantically more complex than counting quantifiers, are not *de facto* harder in NLI. We studied all 27 cases of percentage quantifiers in the English NLI datasets, and found that in most cases, percentage quantifiers occurrences are *identical* across premises and hypotheses, i.e., triggering little or no inference. The other two proportional quantifiers, *most* and *few*, are hard for

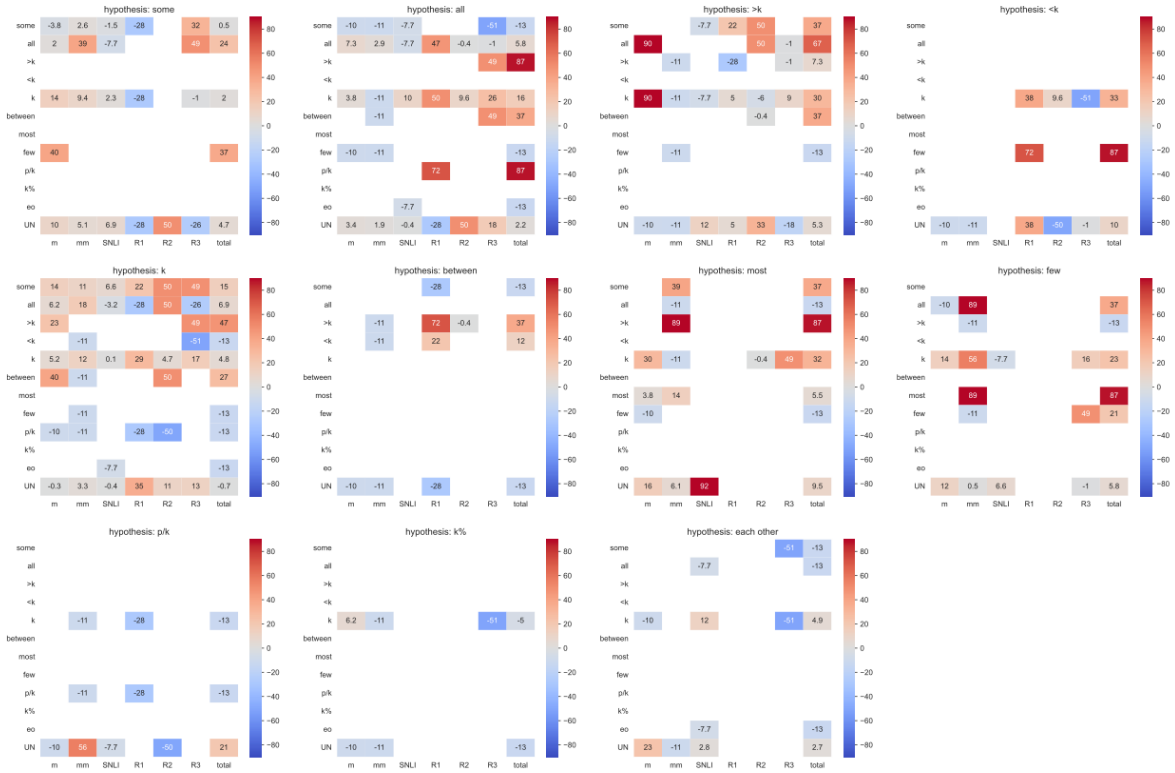

Figure 2: Fine-grained analysis of RoBERTa performance on 6 English NLI subtasks. Each heatmap represents hypotheses with a type of quantifier. The rows stand for premises with the quantifier of that label. The numbers are calculated as the accuracy over the whole dataset minus the fine-grained accuracy given a specific premise and hypothesis (the higher the number, the worse the performance). For each heatmap, the last column represents the accuracy gap weighted by all 6 tasks. "UN" stands for an entry where no explicit quantifier is identified.

LMs to resolve, e.g., in some quantifier pairs, models yield 0% accuracy. Although *each other* is supposed to be hardest to resolve due to the complex semantics of reciprocals (Szymanik and Thorne, 2015), it is not reflected in NLI tasks as such. The reason is similar to percentage quantifiers, while annotators intend to alter counting quantifiers when writing hypotheses, reciprocality is seldomly considered a linguistic ability that needs testing for NLU systems. And the annotation for Ramsey quantifier is simply a knockoff, making reciprocal relation identification unwarranted through shallow correlations.

C Fine-grained NLI Analysis

D XQA Result: mBERT and XLM-R

Table 11 compares the results of mBERT and XLM-R on two XQA tasks, XQuAD and MLQA.

E GQNLI Examples

Table 12 list one example per category in GQNLI.

F GQNLI Negation Cases

We present the results of seven models' performance on cases with negation cues in GQNLI in Table 13.

G GQNLI Subsumption Cases

See Table 14 for models 'performance on cases requiring subsumption reasoning in GQNLI. We also break down subsumption results by entailment labels into two categories: neutral and non-neutral.

H GQNLI Experiment Details

We reused the fine-tuned BERT and RobERTa in Section 4. The other fine-tuned LMs are from Hugging Face. We list the models and thier links in Table 15.

Hypothesis	Premise	MNLI_m_dev		MNLI_mm_dev		SNLI_test		ANLI_R1_test		ANLI_R2_test		ANLI_R3_test		Total		
		#occurrence	%Acc	#occurrence	%Acc	#occurrence	%Acc	#occurrence	%Acc	#occurrence	%Acc	#occurrence	%Acc	#occurrence	%Acc	#correctpred
some	some	45	93.3	38	86.8	16	93.8	1	100			6	16.7	106	86.8	92
	all	8	87.5	8	50	3	100					3	0	22	63.6	14
	>k															
	<k															
	k	12	75	10	80	40	90	4	100	0		2	50	68	85.3	58
	between															
	most															
	few	2	50											2	50	1
	p/k															
	k%															
	eachother															
	"unmatched"	110	79.1	83	84.3	137	85.4	1	100	1	0	8	75	340	82.6	281
all	some	11	100	12	100	2	100					1	100	26	100	26
	all	73	82.2	74	86.5	3	100	4	25	2	50	6	50	162	81.5	132
	>k											1	0	1	0	0
	<k															
	k	28	85.7	19	100	22	81.8	9	22.2	5	40	13	23.1	96	70.8	68
	between			1	100							1	0	2	50	1
	most															
	few	4	100	2	100									6	100	6
	p/k							1	0					1	0	
	k%															
	eachother	0		0		1	100	0		0		0		1	100	1
	"unmatched"	151	86.1	144	87.5	41	92.7	5	100	2	0	13	30.8	356	85.1	303
>k	some					1	100	2	50	1	0			4	50	2
	all	1	0							2	0	2	50	5	20	1
	>k			2	100			1	100			2	50	5	80	4
	<k															
	k	1	0	3	100	2	100	3	66.7	9	55.6	10	40	28	57.1	16
	between									2	50			2	50	1
	most															
	few			1	100									1	100	1
	p/k															
	k%															
	eachother															
	"unmatched"	12	100	18	100	5	80	6	66.7	6	16.7	3	66.7	50	82	41
<k	some															
	all															
	>k															
	<k															
	k							3	33.3	5	40	3	100	11	54.5	6
	between															
	most															
	few							1	0					1	0	0
	p/k															
	k%															
	eachother															
	"unmatched"	3	100	3	100			3	33.3	2	100	2	50	13	76.9	10
k	some	8	75	14	78.6	28	85.7	2	50	2	0	4	0	58	72.4	42
	all	12	83.3	14	71.4	22	95.5	1	100	3	0	4	75	56	80.4	45
	>k	3	66.7									2	0	5	40	2
	<k			2	100							1	100	3	100	3
	k	140	84.3	121	76.9	593	92.2	42	42.9	49	44.9	37	32.4	982	82.5	810
	between	2	50	2	100					1	0			5	60	3
	most															
	few			1	100									1	100	1
	p/k	1	100	1	100			1	100	1	100			4	100	4
	k%															
	eachother					7	100							7	100	7
	"unmatched"	118	89.8	137	86.1	383	92.7	11	36.4	13	38.5	11	36.4	673	88	592
between	some							1	100					1	100	1
	all															
	k			1	100			1	0	2	50			4	50	2
	between			2	100			2	50					4	75	3
	most															
	few															
	>p/k:more/greaterthanp/k															
	<p/k:fewer/lessthanp/k															
	p/k															
	k%															
	eachother															
	"unmatched"	2	100	1	100			1	100					4	100	4
most	some			2	50									2	50	1
	all			2	100									2	100	2
	>k			1	0									1	0	0
	<k															
	k	5	60	1	100					2	50	1	0	9	55.6	5
	between															
	most	7	85.7	4	75									11	81.8	9
	few	1	100											1	100	1
	p/k															
	k%															
	eachother															
	"unmatched"	23	73.9	30	83.3	1	0							54	77.8	42
few	some															
	all	1	100	1	0							0		2	50	1
	>k			1	100							0		1	100	1
	<k											0		0		0
	k	4	75	3	33.3	4	100					3	33.3	14	64.3	9
	between															
	most	0		1	0	0								1	0	0
	few	0		2	100	0						1	0	3	66.7	2
	p/k															
	k%															
	eachother															
	"unmatched"	9	77.8	9	88.9	7	85.7					2	50	27	81.5	22
p/k	some															
	all															
	>k															
	<k															
	k			2	100	0		1	100	0		0		3	100	3
	between															
	most															
	few															
	p/k			2	100			1	100					3	100	3
	k%															
	eachother															
	"unmatched"	1	100	3	33.3	1	100	0		1	100	0		6	66.7	4
k%	some															
	all															
	>k															
	<k															
	k	6	83.3	6	100							1	100	13	92.3	12
	between															
	most															
	few															
	p/k															
	k%															
	eachother															
	"unmatched"	4	100	1	100									5	100	5
eachother	some											1	100	1	100	1
	all					3	100							3	100	3
	>k															
	<k															
	k	1	100			15	80					1	100	17	82.4	14
	between															
	most															
	few															
	p/k															
	k%															
	eachother					1	100							1	100	1
	"unmatched"	3	66.7	3	100	19	89.5							26	84.6	22

Table 10: Statistics of pairwise analysis in Monolingual NLI Benchmarks

Quant.	mBERT en EM	mBERT en F1	mBERT zh EM	mBERT zh F1	mBERT es EM	mBERT es F1	mBERT ar EM	mBERT ar F1	mBERT vi EM	mBERT vi F1	mBERT de EM	mBERT de F1	mBERT *weighted* EM	mBERT *weighted* F1	XLM-R en EM	XLM-R en F1	XLM-R zh EM	XLM-R zh F1	XLM-R es EM	XLM-R es F1	XLM-R ar EM	XLM-R ar F1	XLM-R vi EM	XLM-R vi F1	XLM-R de EM	XLM-R de F1	XLM-R *weighted* EM	XLM-R *weighted* F1
XQuAD																												
some	75	84.2	50	55.5	58.3	76.1	50	**50**	16.6	42.4	**33.3**	**43.8**	47.2	58.7	**66.7**	**76.1**	**41.6**	**51.3**	**50**	**71.5**	66.7	73.6	66.7	**76.9**	66.7	80.6	59.7	71.7
all	**28.5**	**62.2**	**14.2**	**35.2**	**28.5**	82	42.8	**52.3**	**14.2**	**29.4**	**28.5**	**56**	**26.1**	**52.9**	**57.1**	91.8	**14.2**	**21.4**	**57.1**	**78.6**	**42.8**	**54.9**	85.7	85.7	**57.1**	79.3	**52.3**	**68.6**
$> k$																												
$< k$																												
k	78.1	90.1	68.7	80.4	56.2	**72.1**	**40.6**	64.3	**12.5**	35.7	56.2	77.1	52.1	70	75	87.4	53.1	58.8	**46.8**	**77.4**	65.6	86.3	62.5	85.4	62.5	86.9	60.9	80.4
between	100	100	**33.3**	**72.2**	66.6	93.3	100	100	**0**	**19**	**0**	**56.5**	50	73.5	100	100	66.7	66.7	**33.3**	**60**	100	100	100	100	**33.3**	**55.5**	72.2	80.4
p/k																												
$k\%$																												
most	**40**	**53.3**	**40**	**40**	**0**	**10**	**0**	**26.6**	**0**	**0**	**20**	**49.3**	**16.7**	**29.9**	**40**	**48**	**20**	**33.3**	**40**	**50**	**0**	**26.6**	**0**	**0**	**20**	**49.3**	**20**	**34.5**
few																												
each other																												
all GQs	**70**	**83.2**	55	66.7	**50**	**70.3**	41.6	58.2	**11.6**	32.5	**43.3**	**65**	**45.3**	62.7	**70**	**83.6**	**43.3**	**50.2**	**48.3**	**73.6**	60	76	68.3	83.6	**58.3**	80.3	**58**	74.6
comp.	71.8	83.7	48	59.1	56	74.5	40.8	57.9	13.9	32.4	50.7	67.2	46.9	62.5	74.5	86	43	52.8	61	80	53.3	71.7	58.1	78	61.1	77.1	58.5	74.3
MLQA																												
some	**59**	80	**28.2**	**52.1**	**34.1**	**59.2**	36.3	54.9	**5.4**	**24**	**33.3**	58.4	**32.7**	**54.8**	**69.6**	86.1	**33.3**	**60.6**	**41.4**	**70**	43.1	62.9	**43.2**	78	**45.4**	**61.1**	**46**	**69.8**
all	67.7	**79.8**	**14.2**	**46.4**	**38.4**	**62.8**	33.3	57.9	**10.5**	30.1	**31.2**	**51.6**	**32.6**	**54.8**	77.4	90.6	**35.7**	70	**42.3**	**66.4**	**38**	**60**	57.8	79.8	**37.5**	**51**	**48.1**	**69.6**
$> k$																												
$< k$	0	0							0	13.3			0	6.7	0	40							0	20			0	30
k	**74.9**	79.4	47	63.4	**41.5**	**65.9**	**27.6**	50.3	**6.3**	**23.7**	**38.2**	**53**	39.3	**56**	**69.2**	**82.1**	45.2	66.2	**48.7**	73.3	43	64.9	**48.5**	**71.9**	**46.3**	**62.1**	**50.2**	70.1
between	**50**	88.5	50	83.3	**0**	**26.6**	**0**	68.7	**0**	**26.6**			**20**	58.7	**50**	88.5	50	**50**	**50**	**65.3**	**0**	**54.6**	**0**	77.4			**30**	**67.2**
p/k	100	100	**0**	**0**	**0**	**0**							**33.3**	**33.3**	100	100	100	100	100	100							100	100
$k\%$	100	100	**0**	**26.6**			**0**	**23.7**					**33.3**	**50.1**	100	100	0	**26.6**			**0**	71.4					**33.3**	**66**
most	**55.5**	**7**	47.3	62.1	**45.4**	**61.7**	30	**46.8**	**5.8**	**15.7**	**33.3**	**40.7**	**36.2**	**50.3**	**59.2**	**76**	47.3	69.5	**45.4**	**59.5**	**40**	63.2	**47**	75.7	**22.2**	**31.7**	**43.5**	**62.6**
few																												
each other																												
all GQs	**63.5**	**79.2**	41.8	60.3	**39.6**	**63.7**	**29.3**	51.3	**6.4**	**23.6**	**36.1**	**53.2**	**36.1**	**55.2**	**69**	**83**	43	65.6	**46.9**	**71.5**	**41.9**	64.1	**47.6**	73.2	**44.4**	**59.8**	**48.8**	**69.5**
comp.	67.2	80.6	37.5	57.9	47.3	66	30	48.4	11.2	28	40.8	56	39	56.2	70.4	83.3	38.7	62.5	54.1	72.2	42.5	62.9	50.5	72.3	52.2	67.3	51.4	70.1

Table 11: Results of mBERT and XLM-R performance on XQA tasks decomposed by quantifier categories.

Quantifier	**Premise**	**Hypothesis**	**Label**
some	"There are six dogs. Three brown dogs, a black dog and a white dog run along the green grass."	"Some dogs sit."	Neutral
all	"In 2021, there are 490 million people in Africa living in extreme poverty, or 36% of the total population."	"Not all people in Africa live in extreme poverty."	Entailment
$> k$	"Two young men in blue stand over a stove and look at the camera while another young man in red stands behind them."	"At least two men wear red."	Contradiction
$< k$	"More than five guys chased two girls in the classroom."	"No less than four guys chased two girls in the classroom."	Entailment
k	"There are twelve singers on a stage, less than half from Argentina and one from Cape Verde."	"Two singers come from Argentina."	Neutral
between	"Only half out of six cleaners are sweeping up animal faeces from the street during a parade."	"Between four and five cleaners are sweeping up animal faeces."	Contradiction
p/k	"More than 50% but less than 65% of Americans worry about global warming."	"Two thirds of Americans worry about global warming."	Contradiction
$k\%$	"More than five guys chased two girls in the classroom."	"100% of the guys chased two girls in the classroom."	Neutral
most	"Two young men in blue stand over a stove and look at the camera while another young man in red stands behind them."	"Most men wear blue."	Entailment
few	"More than 50% but less than 65% of Americans worry about global warming."	"A few people from America do not worry about global warming."	Entailment
each other	"There are 100 villagers and 100 townsmen. Most villagers and most townsmen hate each other."	"All villagers and all townsmen hate each other."	Neutral

Table 12: GQNLI examples.

Quantifier		some	all	$> k$	$< k$	k	between	p/k	$k\%$	most	few	each other	*Overall*
# Occurrence with negation cues		9	6	6	9	18	3	6	6	6	9	3	81
Model	**Training Data**	**% Performance**											
ELECTRA	S,M,F,ANLI	33.3	16.7	0	66.7	50	100	0	100.0	16.7	11.1	33.3	27.5
BERT	S,M,F,ANLI	0	66.7	100	33.3	50	0	50	0	50	22.2	33.3	39.2
SBERT	S,M,F,ANLI	0	0	0	44.4	22.2	0	0	50	16.7	11.1	0	11.8
RoBERTa	MNLI	33.3	16.7	0	33.3	27.8	66.7	33.3	33.3	50	33.3	33.3	31.4
	S,M,F,ANLI	33.3	0	0	66.7	33.3	0	0	50	0	0	0	17.6
BART	MNLI	33.3	50.0	50.0	33.3	33.3	0	50	0	0	66.7	100	35.3
	S,M,F,ANLI	0.0	66.7	100.0	33.3	50.0	0.0	50.0	0.0	50.0	22.2	33.3	39.2
ALBERT	S,M,F,ANLI	44.4	0	33.3	44.4	33.3	0	50	50	0	0	0	25.5
DeBERTa-v3	MNLI	33.3	0	33.3	33.3	27.8	0	16.7	0	0	33.3	0	23.5
	M,F,ANLI	0	0	0	66.7	33.3	0	0	50	50	33.3	0	23.5
	M,F,Ling,DocNLI	33.3	0	0	0	11.1	66.7	50	33.3	50	33.3	0	27.5

Table 13: Models' performance on instances with negation cues in GQNLI.

Type # Occurrence		Subsumption (neutral) 90	Subsumption (non-neutral) 99	Subsumption (total) 189	Non-subsumption 111
Model	**Training Data**		**% Performance**		
ELECTRA	S,M,F,ANLI	3.3	37.4	21.2	36
BERT	S,M,F,ANLI	22.2	24.2	23.3	41.4
SBERT	S,M,F,ANLI	13.3	48.5	31.7	38.7
RoBERTa	MNLI	27.8	18.2	22.8	37.8
	S,M,F,ANLI	21.1	49.5	36.0	38.7
BART	MNLI	36.7	15.2	25.4	41.4
	S,M,F,ANLI	44.4	44.4	44.4	36.9
ALBERT	S,M,F,ANLI	33.3	49.5	41.8	42.3
DeBERTa-v3	MNLI	45.6	34.3	39.7	42.3
	M,F,ANLI	52.2	41.4	46.6	32.4
	M,F,Ling,DocNLI	86.7	32.3	58.2	16.2

Table 14: Models' performance on instances requiring subsumption reasoning.

Model	Training Data	Model's link
ELECTRA	S,M,F,ANLI	https://huggingface.co/ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli
BERT	S,M,F,ANLI	
SBERT	S,M,F,ANLI	https://huggingface.co/usc-isi/sbert-roberta-large-anli-mnli-snli
RoBERTa	MNLI	https://huggingface.co/roberta-large-mnli
	S,M,F,ANLI	
BART	MNLI	https://huggingface.co/facebook/bart-large-mnli
	S,M,F,ANLI	https://huggingface.co/ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli
ALBERT	S,M,F,ANLI	ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli
DeBERTa-v3	MNLI	https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli
	M,F,ANLI	https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli
	M,F,Ling,DocNLI	https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c

Table 15: Links to the models we use to test on GQNLI.