ANLIzing the Adversarial Natural Language Inference Dataset

Adina Williams, Tristan Thrush, Douwe Kiela
Facebook AI Research
{adinawilliams, tthrush, dkiela}@fb.com

Abstract

We perform an in-depth error analysis of the Adversarial NLI (ANLI) dataset, a recently introduced large-scale human-and-model-in-the-loop natural language inference dataset collected dynamically over multiple rounds. We propose a fine-grained annotation scheme for the different aspects of inference responsible for the gold classification labels, and use it to hand-code the ANLI development sets in their entirety. We use these annotations to answer a variety of important questions: which models have the highest performance on each inference type, which inference types are most common, and which types are the most challenging for state-of-the-art models? We hope our annotations will enable more fine-grained evaluation of NLI models, and provide a deeper understanding of where models fail (and succeed). Both insights can guide us in training stronger models going forward.

1 Introduction

Natural Language Inference (NLI) is one of the canonical benchmark tasks for research on Natural Language Understanding (NLU). NLI[1] has characteristics that make it desirable both from theoretical and practical standpoints. Theoretically, entailment is, in the words of Richard Montague, "the basic aim of semantics" (Montague, 1970, p. 223 fn.), and indeed meaning in formal semantics relies on necessary and sufficient truth conditions. Practically, NLI is easy to evaluate and intuitive even to non-linguists, enabling data to be collected at scale with crowdworker annotators. Moreover, many core NLP tasks can also easily be converted to NLI problems (White et al. 2017; Demszky et al. 2018; Poliak et al. 2018a i.a.) suggesting that NLI is generally seen as a good proxy for measuring models' overall NLU capabilities.

Benchmark datasets are essential for driving progress in NLP and machine learning (DataPerf Working Group, 2021). In recent years, large-scale NLI benchmarks like SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) have established a straightforward basis for comparison between trained models. However, with the advent of transformer language models, many benchmarks are now reaching saturation, leading some to wonder: have we solved NLI and, perhaps, NLU? However, the recent ANLI dataset (Nie et al., 2020a) illustrated that our models do not yet perform NLI in the robust and generalizable way that humans can. In this paper we ask: where do our models still fall short?

To improve towards general NLU, merely listing examples of failure cases is not by itself sufficient. We also need a quantifiable and finer-grained understanding of *which phenomena are responsible* for failures (or successes). Since the dynamic adversarial set up of ANLI encouraged human annotators to exercise their creative faculties to fool model adversaries, the data contains a wide range of possible inferences (as we will show). Because of this, ANLI is an ideal testbed for studying current model shortcomings, and for characterizing what future models will have to do in order to make progress on the NLI task.

Towards that end, we propose a genre-agnostic annotation scheme for NLI that classifies example pairs into 40 inference types. It is hierarchical, reaching a maximum of four levels deep, enabling analysis of model performance at a flexible level of granularity. We also contribute expert hand-annotations on the ANLI development sets (3200 sentence pairs) according to our scheme[2], thereby extending the usefulness of the ANLI dataset by making it possible to analyze future models. We

[1] Also known as recognizing textual entailment (RTE; Fyodorov et al. 2000; Dagan et al. 2006, i.a.).

[2] All annotations are publicly available at https://github.com/facebookresearch/anli/anlizinganli.

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 23-54.
Held on-line February 7-9, 2022

Context	Hypothesis	Rationale	Gold/Pred. (Valid.)	Tags
Eduard Schulte (4 January 1891 in Düsseldorf – 6 January 1966 in Zürich) was a prominent German industrialist. He was one of the first to warn the Allies and tell the world of the Holocaust and systematic exterminations of Jews in Nazi Germany occupied Europe.	Eduard Schulte is the only person to warn the Allies of the atrocities of the Nazis.	The context states that he is not the only person to warn the Allies about the atrocities committed by the Nazis.	C/N (CC)	Tricky, Prag., Numerical, Ordinal
Kota Ramakrishna Karanth (born May 1, 1894) was an Indian lawyer and politician who served as the Minister of Land Revenue for the Madras Presidency from March 1, 1946 to March 23, 1947. He was the elder brother of noted Kannada novelist K. Shivarama Karanth.	Kota Ramakrishna Karanth has a brother who was a novelist and a politician	Although Kota Ramakrishna Karanth's brother is a novelist, we do not know if the brother is also a politician	N/E (NEN)	Basic, Coord., Reasoning, Plaus., Likely, Tricky, Syntactic
Toolbox Murders is a 2004 horror film directed by Tobe Hooper, and written by Jace Anderson and Adam Gierasch. It is a remake of the 1978 film of the same name and was produced by the same people behind the original. The film centralizes on the occupants of an apartment who are stalked and murdered by a masked killer.	Toolbox Murders is both 41 years old and 15 years old.	Both films are named Toolbox Murders one was made in 1978, one in 2004. Since it is 2019 that would make the first 41 years old and the remake 15 years old.	E/C (EE)	Reasoning, Facts, Numerical Cardinal, Age, Basic, Coord., Tricky, Wordplay

Table 1: Examples from development set. 'corr.' is the original annotator's gold label, 'pred.' is the model prediction, 'valid.' is the validator label(s).

find that examples requiring models to resolve references, utilize external knowledge, and deploy syntactic abilities remain especially challenging. Our annotations are publicly available, and we hope they will be useful for benchmarking progress on particular inference types and exposing weaknesses of future NLI models.

2 Background

We proposes an inference type annotation scheme for the Adversarial NLI (ANLI) dataset, which was collected via a gamified, adversarial human-and-model-in-the-loop format using the Dynabench platform (Kiela et al., 2021; Ma et al., 2021). Human annotators are matched with a **target model** trained on existing NLI data, and tasked with finding examples that fooled it into predicting the wrong label. Dynamically collecting data has since been shown to have training-time benefits above statically collected data (Wallace et al., 2021). Other than being dynamic, ANLI was collected with a similar method to SNLI and MNLI: untrained crowdworkers are given a context—and one of three classification labels, i.e., Entailment, Neutral and Contradiction—and asked to write a hypothesis. Table 1 provides examples.

The ANLI dataset was collected in English over three rounds, with different target model adversaries each round. The first round adversary was a BERT-Large (Devlin et al., 2019) model trained on SNLI and MNLI. The second was a RoBERTa-Large (Liu et al., 2019) ensemble trained on SNLI and MNLI, as well as FEVER (Thorne et al., 2018) and the training data from the first round. The third round adversary was a RoBERTa-Large ensemble trained on all previous data, plus the training data from the second round, with the additional difference that the contexts were sourced from multiple domains (rather than just from Wikipedia, as in the preceding rounds). The ANLI dataset is split so that all development and test set data were human-validated as model-fooling.

The ANLI dataset creators encouraged crowdworkers to give free rein to their creativity (Nie et al., 2020a, p.8).[3] Annotators explored, then ultimately converged upon, inference types that challenged each round's target model adversary. For example, the target model in round 1 was often fooled by numbers (see §4), which means the development set from round 1 (i.e., A1) contains many NUMERICAL examples. Training a later rounds' adversary on A1 then should result in a model that does better on such examples. Ultimately, crowdworkers would be less successful at fooling later adversaries with numbers, and fewer NUMERICAL examples will end up in later development sets.[4] In this way, understanding how inference types dynamically shift across the ANLI development sets can illuminate the capabilities of the target models used to collect them.

3 Developing A Scheme for Annotating Types of Inferences in NLI

Categorizing sentential inference relations into types is by no means a new endeavor (see the Doctrine of Categories from Aristotle's *Organon*): ample research has aimed to understand model behavior and/or develop best annotation practices which ought to be incorporated. However, a scheme should be, at least to some extent, tai-

[3] Gamification generally results in wide coverage datasets (Joubert et al., 2018; Bernardy and Chatzikyriakidis, 2019).

[4] Assuming that models trained on later rounds don't suffer from catastrophic forgetting.

Top Level	Second Level	Description
Numeral	Cardinal	basic cardinal numerals (e.g., 56, 57, 0, 952, etc.).
	Ordinal	basic ordinal numerals (e.g., 1^{st}, 4^{th}, 72^{nd} etc.).
	Counting	counting references in the text, such as: *Besides A and B, C is one of the monasteries located at Mt. Olympus.* ⇒ *C is one of three monasteries on Mount Olympus.*
	Nominal	numbers as names, such as: *Player 37 scored the goal* ⇒ *a player was assigned jersey number 37.*
Basic	Comp.& Super.	degree expressions denoting relationships between things, such as: *if X is faster than Y* ⇒ *Y is slower than X*
	Implications	cause and effect, or logical conclusions that can be drawn from clear premises. Includes classical logic types such as Modus Ponens.
	Idioms	idioms or opaque multiword expressions, such as: *Team A was losing but managed to beat the other team* ⇒ *Team A rose to the occasion*
	Negation	inferences relying on negating content from the context, with "no", "not', "never", "un-" or other linguistic methods
	Coordinations	inferences relying on "and", "or", "but", or other coordinating conjunctions.
Ref.	Coref.	accurately establishing multiple references to the same entity, often across sentences, such as: *Sammy Gutierrez* is *Guty*
	Names	content about names in particular (e.g., *Ralph is a male name*, *Fido is a dog's name*, *companies go by acronyms*)
	Family	content that is about families or kinship relations (e.g., *if X is Y's aunt, then Y is X's nephew/niece and Y is X's parent's sibling*)
Tricky	Syntactic	argument structure alternations or changes in argument order (e.g., *Bill bit John* ⇒ *John got bitten.*, *Bill bit John* ⇏ *John bit Bill*)
	Pragmatic	presuppositions, implicatures, and other kinds of reasoning about others' mental states: *It says 'mostly positive' so it stands to reason some were negative.*
	Exhaustification	pragmatic reasoning where all options not made explicit are impossible, for example: *a field involves X, Y, and Z* ⇒ *X, Y and Z are the only aspects of the field*
	Translation	examples with text in a foreign language or using a foreign orthography.
	Wordplay	puns, anangrams, and other fun language tricks, such as *Margaret Astrid Lindholm Ogden's initials are MALO, which could be scrambled around to form the word 'loam'.*
Reasoning	Plausibility	the annotators subjective impression of how plausible a described event is (e.g. *Brofiscin Quarry is named so because a group of bros got together and had a kegger at it.* and *Fetuses can't make software* are unlikely)
	Facts	common facts the average human would know (like that the year is 2020), but that the model might not (e.g., *the land of koalas and kangaroos* ⇒ *Australia*), including statements that are clearly not facts (e.g., *In Ireland, there's only one job.*)
	Containment	references to mereological part-whole relationships, temporal containment between entities (e.g., *October is in Fall*), or physical containment between locations or entities (e.g., *Germany is in Europe*). Includes examples of bridging (e.g., *the car had a flat* ⇒ *The car's tire was broken*).
Imperfections	Error	examples for which the expert annotator disagreed with the gold label, such as the gold label of neutral for the pair *How to limbo. Grab a long pole. Traditionally, people played limbo with a broom, but any long rod will work* ⇒ *limbo is a type of dance*
	Ambig.	example pairs for which multiple labels seem to the expert to be appropriate. For example, with the context *Henry V is a 2012 British television film*, whether *Henry V is 7 years old this year* should get a contradiction or neutral label depends on what year it is currently as well as on which month Henry V began to be broadcast and when exactly the hypothesis was written.
	Spelling	examples with spelling errors.

Table 2: Summary of the Annotation Scheme. Toy examples are provided, ⇒ denotes entailment, ⇏ denotes contradiction. Only top and second level tags are provided, due to space considerations.

lored to the particular task at hand. Here, we balance these considerations and develop a novel NLI annotation scheme. We hope other large NLI datasets will be annotated according to our scheme to make even wider comparison possible.

Researchers have proposed many ways to 'crack open the black box' (Alishahi et al., 2019; Linzen et al., 2019), from uncovering lexical confounders or annotation "artifacts" (Gururangan et al., 2018; Geiger et al., 2018; Poliak et al., 2018b; Tsuchiya, 2018; Glockner et al., 2018; Geva et al., 2019) to evaluating generalization with diagnostic datasets (McCoy et al., 2019; Naik et al., 2018; Nie et al., 2019; Yanaka et al., 2019; Warstadt et al., 2019a; Geiger et al., 2020; Hossain et al., 2020; Jeretic et al., 2020; Warstadt et al., 2020; Schuster et al., 2020); see Zhou et al. (2020) for a critical overview. Specific to NLI, some have probed models to see what they learn (Richardson et al., 2019; Sinha et al., 2021b), honed data collection methods (Bowman et al., 2020; Vania et al., 2020; Parrish et al., 2021) and analyzed inherent disagreements between human annotators (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b; Nangia et al., 2021), all in the service of understanding and improving models (see Poliak (2020) for a recent survey). See Table 10 and §A.3 for comparisons between our annotation scheme and others.

To inventory possible inference types, three NLP researchers independently inspected data from ANLI A1. For consistency, we then discussed and merged codes, applying an inductive approach (Thomas, 2006; Blodgett et al., 2021). Our scheme—provided in abbreviated form in Table 2—has 40 tag types that can be combined to a depth of up to four (see the Appendix for more details in §A.1, and more examples in Table 14). The top level of the scheme was fixed by the original ANLI paper to five classes: NUMERICAL, BASIC, REFERENCE, TRICKY inferences, and REASONING.[5] We aimed to balance proliferating narrow tags (and potentially being overly expressive), and limiting tag count to enable generalization (potentially being not expressive enough). A hierarchical tagset achieves the best of both worlds—we can measure all our metrics at varying granularities while allowing for pairs to receive as many tags as are warranted (see Table 1).

Annotation. Annotating NLI data for inference types requires various kinds of expert knowledge,

[5]These top-level types were introduced for smaller subsets of the ANLI development set in § 5 of Nie et al. (2020a), which we drastically expand both in number and specificity of tag types, as well as in annotation scope.

Dataset	Subset	Numerical	Basic	Reference	Tricky	Reasoning	Error
A1	All	40.8	31.4	24.5	29.5	58.4	3.3
	C	18.6	8.2	7.8	13.7	11.9	0.7
	N	7.0	9.8	7.1	6.4	31.3	1.0
	E	15.2	13.4	9.6	9.4	15.2	1.6
A2	All	38.5	41.2	29.4	29.1	62.7	2.5
	C	15.6	11.8	10.2	13.6	15.5	0.3
	N	8.1	12.8	9.1	7.4	30.0	1.4
	E	14.8	16.6	10.1	8.1	17.2	0.8
A3	All	20.3	50.2	27.5	25.6	63.9	2.2
	C	8.7	17.2	8.6	12.7	14.9	0.3
	N	4.9	13.1	8.2	4.6	30.1	1.0
	E	6.7	19.9	10.7	8.3	18.9	0.8

Table 3: Percentages (of the total) of tags by gold label and subdataset. 'All' refers to the total percentage of examples in that round that were annotated with that tag. 'C', 'N', and 'E', refer to percentage of examples with that tag that receive each gold label.

i.e. with a range of complicated linguistic phenomena and the particularities of the NLI task. Our work is fairly unique in that examples are *only* tagged as belonging to a particular branch of the taxonomy when the annotator believed the tagged phenomenon is required for a human to arrive at the target label assignment. Mere presence of a phenomenon was insufficient, meaning that automation was impossible, and expert annotation was necessary.[6] A single annotator with a decade's training in linguistics and expertise in NLI both devised our scheme and applied it to the ANLI development set. Annotation was laborious, taking the expert several hundred hours.

Inter-annotator Agreement. Employing a single annotator may have downsides, if they inadvertently introduce personal idiosyncrasies into their annotations. NLI may be especially susceptible to this, as recent work uncovers much variation in human judgements for this task (Pavlick and Kwiatkowski, 2019; Min et al., 2020; Nie et al., 2020b). To understand whether our tags are individual to the main annotator, we employed a second expert (with 5 years of linguistic training) to re-annotate 300 random examples, 100 from each development set. Re-annotation took the second annotator approximately 35 hours (excluding training time). Further details on the scheme, guidelines, and process are in Appendix A.

We measure inter-annotator agreement for each tag independently. For each example, annotators agree on a tag if they both used that tag or both did not use that tag; otherwise they disagree. Average percent agreement between our annotators is 72% for top-level and 91% for low-level tags respectively (see Table 8 and §A.2 for further details). Recall that 50% would be chance (since we are measuring whether the tag was used or not between two annotators). Our inter-annotator agreement is comparable to a similar semantic annotation effort on top of the original RTE data (Toledo et al., 2012), suggesting we have reached an acceptable level of agreement for our setting, and that the main annotator is not very idiosyncratic.

4 Experiments

We investigate 8 models: the original ANLI target model adversaries[7], and five SOTA models[8]—a RoBERTa-Large (Liu et al., 2019), a BART-Large (Lewis et al., 2019), an XLNet-Large (Yang et al. 2019, an ELECTRA-Large (Clark et al., 2020), and an ALBERT-XXLarge (Lan et al., 2020)—finetuned on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), and ANLI rounds 1–3.

We report the tag distribution of the ANLI validation sets to establish an estimate of inference type frequency and explore what models may have learned as rounds progressed. To measure difficulty, we report models' correct label probability,

[6]Experts are well known to achieve higher performance than naïve crowdworkers when the task is linguistically complex (e.g., the CoLA subtask of the GLUE benchmark from Warstadt et al. (2019b), as well as Nangia and Bowman 2019, p. 4569, Basile et al. 2012; Bos et al. 2017, i.a.).

[7]For A2 and A3, which were ensembles, we randomly select a single RoBERTa-Large as the representative.

[8]https://github.com/facebookresearch/anli

and entropy on example pairs requiring each inference type (as accuracy on ANLI is still very low).

4.1 Tag Distribution

REASONING tags are the most common in the validation dataset, followed by NUMERICAL, TRICKY, BASIC, REFERENCE and then IMPERFECTIONS. The frequency of top-level tags are presented in Table 3, and for subtags in Table 15.

Walking through top-level types in turn, we find that NUMERICAL pairs are most common in A1. Since A1 contexts comprised the first few lines of Wikipedia entries—which often have numbers in them—this makes sense. A2, despite also using Wikipedia contexts, has a lower percentage of NUMERICAL examples, possibly because its target model—also trained on A1—improved on that category. In A3, the percentage of NUMERICAL pairs has dropped even lower. Between A1/A2 and A3, this drop in top level NUMERICAL tag frequency is due at least in part to a drop in the use of CARDINAL subtag, which results in a corresponding drop of third level DATES and AGES tags (in the Appendix). Overall, NUMERICAL pairs are more likely to have the gold label contradiction or entailment than neutral.

BASIC pairs are fairly common, with increasing frequency as rounds progress. Subtags LEXICAL and NEGATION rise sharply in frequency between A1 and A3; IMPLICATIONS and IDIOMS also rise—though they rise less sharply and are only present in $< 10\%$ of examples. COORDINATION and COMPARATIVES & SUPERLATIVES tag frequency stays roughly constant. Overall, BASIC examples tend to be gold labeled as entailment.

REFERENCE tags are rarest main tag type (present in 24.5% of A1 examples, rising slightly in A2 and A3). The most common subtag for REFERENCE is COREFERENCE with incidences ranging from roughly 16% in A1 to 26% in A3. Subtags NAMES and FAMILY maintain roughly constant low frequency across rounds, although there is a precipitous drop in NAMES tags for A3 (likely reflecting genre differences). Examples tagged as REFERENCE most commonly have entailment as their gold label for all rounds.

TRICKY inference types occur at relatively constant rates. A1 contains more examples with word reorderings than the others. PRAGMATIC examples are more prevalent in A1 and A3. A2 is unique in having slightly higher frequency of EXHAUSTIFICATION tags. WORDPLAY examples increase in A2 and A3 compared to A1. TRANSLATION pairs are rare (≈3%). On the whole, there are fewer neutral TRICKY pairs than contradictions or entailments, with contradiction being somewhat more common.

REASONING examples are very common across the rounds, with 50–60% of pairs receiving at least one. Subtagged FACTS pairs are also common, rising from 19% in A1 to roughly 25% of A2 and A3. CONTAINMENT shows the opposite pattern; it halves its frequency between A1 and A3. The frequency of third level LIKELY examples remains roughly constant whereas third level UNLIKELY and DEBATABLE examples become more common over the rounds. DEBATABLE tags rise to 3 times their rate in A1 by the third round, in part reflecting the contribution of different domains of text (see Table 7 for incidence on the procedural genre). On average, REASONING tags are more common for examples with a neutral gold label.

IMPERFECTION tags are rare across rounds ($\approx 14\%$ of example pairs receive that tag on average), and are slightly more common for neutral pairs. SPELLING imperfections are the most common second level tag type, at $\approx 5-6\%$ of examples. Examples marked as AMBIGUOUS and ERROR were rare at $\approx 3-5\%$.

4.2 Model Predictions by Tag

For each model-round-tag triple, we report (i) the average probability of the correct prediction and (ii) the entropy of model predictions (i.e., from the input to the softmax layer) in Table 4[9]. We report both because neither number is fully interpretable in itself. Measuring the probability mass the model assigned to the correct label gives a nuanced notion of accuracy, whereas entropy can be seen as a measure of difficulty, in the sense that it can tell us how (un)certain a model is in its predictions. If a particular model-round-tag triple has high entropy, then that tag was more difficult for that model to learn from that round's data. A given model-round-tag triple can have both high probability and high entropy, which would show that the round-tag pairing is difficult (given the entropy), but that the model succeeded, at least to some extent, in learning how to predict the correct label anyway (given the probability).

ALBERT-XXLarge performs best overall, with

[9]Metrics for the lower level tags in Table 16–Table 20.

Round	Model	Numerical	Basic	Ref. & Names	Tricky	Reasoning	Imperfections
	BERT (R1)	0.10 (0.57)	0.13 (0.60)	0.11 (0.56)	0.10 (0.56)	0.12 (0.59)	0.13 (0.57)
	RoBERTa Ensemble (R2)	0.68 (0.13)	0.67 (0.13)	0.69 (0.15)	0.60 (0.18)	0.66 (0.15)	0.61 (0.14)
	RoBERTa Ensemble (R3)	0.72 (**0.07**)	0.73 (**0.08**)	0.72 (0.08)	0.65 (0.09)	0.70 (**0.08**)	0.68 (**0.07**)
A1	RoBERTa-Large	0.73 (0.13)	0.75 (0.12)	**0.76** (0.10)	**0.70** (0.14)	0.75 (0.15)	0.68 (0.13)
	BART-Large	0.73 (0.10)	0.76 (**0.08**)	0.72 (**0.07**)	**0.70** (**0.08**)	0.70 (0.11)	**0.71** (0.08)
	XLNet-Large	0.73 (0.10)	0.74 (0.09)	0.75 (0.09)	**0.70** (0.10)	0.72 (0.09)	0.67 (0.08)
	ELECTRA-Large	0.71 (0.29)	0.66 (0.36)	0.68 (0.34)	0.62 (0.44)	0.63 (0.41)	0.63 (0.40)
	ALBERT-XXLarge	**0.74** (0.22)	**0.77** (0.18)	**0.76** (0.20)	0.65 (0.21)	**0.77** (0.18)	0.69 (0.22)
	BERT (R1)	0.29 (0.53)	0.30 (0.47)	0.29 (0.44)	0.25 (0.48)	0.31 (0.47)	0.33 (0.48)
	RoBERTa Ensemble (R2)	0.19 (0.28)	0.21 (0.26)	0.20 (0.25)	0.16 (0.23)	0.19 (0.24)	0.19 (0.27)
	RoBERTa Ensemble (R3)	0.50 (0.18)	0.43 (0.16)	0.41 (0.14)	0.44 (0.14)	0.45 (0.14)	0.33 (0.14)
A2	RoBERTa-Large	0.54 (0.22)	0.51 (0.21)	0.47 (0.17)	0.48 (0.22)	0.49 (0.20)	0.49 (0.19)
	BART-Large	0.55 (0.13)	0.52 (0.13)	0.48 (0.14)	0.48 (0.15)	0.50 (0.13)	0.42 (**0.10**)
	XLNet-Large	0.54 (**0.11**)	0.53 (**0.12**)	0.53 (**0.13**)	**0.52** (**0.12**)	0.50 (**0.10**)	0.44 (**0.10**)
	ELECTRA-Large	0.56 (0.36)	0.53 (0.40)	0.52 (0.40)	0.51 (0.45)	0.53 (0.38)	0.54 (0.39)
	ALBERT-XXLarge	**0.57** (0.28)	**0.57** (0.29)	**0.58** (0.28)	0.50 (0.26)	**0.56** (0.25)	**0.58** (0.32)
	BERT (R1)	0.34 (0.53)	0.34 (0.51)	0.32 (0.50)	0.29 (0.55)	0.32 (0.49)	0.31 (0.54)
	RoBERTa Ensemble (R2)	0.29 (0.47)	0.26 (0.54)	0.26 (0.57)	0.24 (0.58)	0.27 (0.55)	0.23 (0.58)
	RoBERTa Ensemble (R3)	0.20 (0.43)	0.23 (0.50)	0.24 (0.53)	0.25 (0.54)	0.25 (0.54)	0.23 (0.52)
A3	RoBERTa-Large	0.44 (0.32)	0.44 (0.26)	0.45 (0.25)	0.49 (0.25)	0.46 (0.27)	0.40 (0.23)
	BART-Large	0.51 (**0.14**)	0.50 (**0.14**)	0.49 (**0.14**)	**0.53** (0.18)	0.50 (**0.14**)	0.48 (0.17)
	XLNet-Large	0.52 (0.15)	0.49 (**0.14**)	0.49 (0.15)	0.51 (**0.14**)	0.52 (0.15)	0.43 (**0.14**)
	ELECTRA-Large	0.55 (0.46)	0.51 (0.45)	0.52 (0.44)	0.54 (0.44)	0.52 (0.48)	0.47 (0.49)
	ALBERT-XXLarge	**0.56** (0.39)	**0.57** (0.33)	**0.55** (0.36)	0.52 (0.32)	**0.54** (0.32)	**0.52** (0.33)
	BERT (R1)	0.22 (0.54)	0.26 (0.52)	0.26 (0.50)	0.21 (0.53)	0.26 (0.51)	0.27 (0.53)
	RoBERTa Ensemble (R2)	0.41 (0.26)	0.37 (0.33)	0.34 (0.37)	0.33 (0.34)	0.35 (0.33)	0.32 (0.37)
	RoBERTa Ensemble (R3)	0.52 (0.20)	0.44 (0.27)	0.41 (0.30)	0.45 (0.26)	0.45 (0.28)	0.39 (0.28)
ANLI	RoBERTa-Large	0.59 (0.21)	0.55 (0.20)	0.53 (0.19)	0.56 (0.20)	0.56 (0.21)	0.50 (0.19)
	BART-Large	0.61 (**0.12**)	0.58 (**0.12**)	0.54 (**0.13**)	**0.57** (0.14)	0.55 (0.13)	0.52 (0.12)
	XLNet-Large	0.61 (**0.12**)	0.58 (**0.12**)	0.56 (**0.13**)	**0.57** (**0.12**)	0.57 (**0.12**)	0.50 (**0.11**)
	ELECTRA-Large	0.62 (0.35)	0.56 (0.40)	0.56 (0.40)	0.56 (0.44)	0.55 (0.43)	0.54 (0.44)
	ALBERT-XXLarge	**0.64** (0.28)	**0.63** (0.27)	**0.61** (0.30)	0.56 (0.26)	**0.61** (0.25)	**0.59** (0.30)

Table 4: Mean correct label probability (highest bold) and mean entropy of label predictions (lowest bold) by model and top level tag. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58. See Appendix E: Table 16–Table 21 for full results on lower-level tags.

the highest label probability for the full ANLI development set for each top-level tag except for TRICKY, where it performs roughly as well as the others. BART-Large, XLNet-Large, and ELECTRA-Large are tied for second place, with RoBERTa-Large being a relatively close third. In general, the five SOTA models' probabilities of correct label differ by a few points, although BART-Large and XLNet-Large show markedly more certainty (i.e., lower entropy of predictions) than the others. It is clear that A1 is easier than A2 and A3, as measured by both higher correct label probability and lower entropy in general across models. A2 and A3 don't appreciably differ, although A3 generally has slightly lower correct label probabilities and higher entropies, meaning that A2 and A3 remain difficult for current models.

The ANLI model adversaries perform much worse that the SOTA models, having both lower mean probability of the correct label and often higher entropy: On A1 and A2, of three model adversaries, RoBERTa-Large (R3) also has the highest average label probability and lowest entropy (recall that RoBERTa-Large (R3) was one of the model adversaries in the ensemble, so its average prediction probability on A3 should be low).

Difficulty by Tag. Accuracy on ANLI is still fairly low (see Table 13), however it is still worth discussing which inference types confound our best current models. To understand our results, we have to be aware of how prevalent in the training corpus certain types are. We cannot necessarily expect a model to perform well on things it hasn't seen (although people often do, see Chomsky 1980). Because the ANLI training sets are not annotated, we will estimate the incidence of tags using the development sets (recall Table 3). To explore the relationship between phenomenon frequency and learnability by models, we split lower level tags into "common" tags are present in approximately 10% or more the ANLI development sets, while the rest are deemed "uncommon" (see Appendix E Table 16–Table 21 for more details).

Wikipedia	Fiction	News	Procedural	Legal	RTE
0.64 (0.24)	0.57 (0.29)	0.58 (**0.24**)	0.60 (0.28)	0.55 (0.39)	0.52 (0.52)

Table 5: Mean correct label probability (mean entropy of label predictions) for ALBERT-XXLarge by genre.

Tag	Wikipedia	Fiction	News	Procedural
Numerical	39.7%	3.5%	17.2%	10.5%
Basic	36.8%	41.0%	54.5%	48.5%
Reference	27.5%	21.0%	19.7%	14.5%
Tricky	29.0%	28.5%	25.3%	24.0%
Reasoning	61.7%	67.5%	59.6%	62.5%
Error	2.8%	3.5%	1.0%	2.5%

Table 6: Percentage of top-level tags in each genre.

Perhaps obviously, common inference types (e.g., REASONING-LEXICAL, NUMERICAL-DATES, REASONING-LIKELY) are easier for models to perform well on (according to higher correct label probability). More compellingly though, there were some common inference types that the models still behaved poorly on, namely REASONING-FACTS, REFERENCE-COREFERENCE, BASIC-NEGATION and TRICKY-SYNTACTIC. Since these tags are fairly frequent, it's reasonable to conclude that these types required more complex knowledge. For example, REASONING-FACTS, which includes knowing that "2020 is this year" or that "a software engineering tool can't enable people to fly".

Models can do fairly well on some uncommon tags, e.g., BASIC-COORDINATION and NUMERICAL-NOMINAL, REASONING-UNLIKELY, REFERENCE-NAMES, REASONING-CONTAINMENT, TRICKY-WORDPLAY. There are two potential explanations for this higher than expected performance: perhaps the SNLI, MNLI or FEVER training data has sufficient quantities these inference types or, alternatively, these types are somewhat easier to learn from fewer examples. Models do struggle with NUMERICAL-COUNTING, NUMERICAL-AGE, BASIC-IMPLICATIONS, REASONING-DEBATABLE, BASIC-IDIOM, TRICKY-PRAGMATICS, TRICKY-EXHAUSTIFICATION. Similarly, these failures can either be due to tag rarity or to their inherent difficulty. Future work could ask whether augmenting training data with more examples of these types boosts performance.

Overall, models struggle with examples requiring linguistic or external knowledge: the hardest top-level tag for all models is TRICKY, with REASONING and REFERENCE being next in line. Anywhere from one quarter to two thirds of data contains at least one of these tags, so models have been exposed to these inference types. NUMERICAL and BASIC examples are less difficult, but are by no means solved. On rounds A1–3, adversaries improve on NUMERICAL examples, suggesting that exposure to relevant NUMERICAL examples can enable modest improvement (see also Dua et al. 2019 for a related observation).

Tag	Wikipedia	Fiction	News	Procedural
Numerical	**0.65** (0.25)	**0.65** (0.27)	**0.67** (0.32)	0.66 (**0.26**)
Basic	0.64 (0.25)	0.55 (0.27)	0.56 (0.22)	0.61 (0.32)
Reference	**0.65 (0.22)**	0.50 (**0.23**)	0.52 (0.23)	**0.71** (0.29)
Tricky	0.56 (0.24)	0.52 (0.26)	0.64 (**0.19**)	0.57 (0.29)
Reasoning	0.66 (0.24)	0.61 (0.28)	0.55 (0.24)	0.56 (0.31)
Imperfection	0.63 (0.27)	0.62 (0.34)	0.60 (0.26)	0.53 (0.26)

Table 7: Mean correct label probability (mean entropy of label predictions) for ALBERT-XXLarge.

Summary. ALBERT-XXLarge performs slightly better than the others, but it is less certain in its predictions; XLNet-Large and BART-Large perform slightly worse, but have lower entropy. Top-level TRICKY[10], REASONING, and REFERENCE categories are still difficult for SOTA models, even though they are frequent. Of the lower level tags that appear in approximately 10% of the ANLI development sets, FACTS, COREFERENCE, NEGATION and SYNTACTIC example pairs remain difficult.

4.3 Overlap in Model Predictions

Generally, model outputs were somewhat correlated with ANLI gold labels represented as one-hot vectors (see Figure 1). ALBERT-XXLarge model outputs are the most positively correlated (Pearson's correlation) (≈ 0.5), RoBERTa-Large, BART-Large, XLNet-Large, and ELECTRA-Large have medium sized positive correlations, and the R2 and R3 RoBERTa-Large models have small positive correlations. BERT (R1) is slightly negatively correlated with gold labels. All differences were significant ($p < 0.01$).

However, different models made very similar predictions: RoBERTa-Large, BART-Large, XLNet-Large, and ALBERT-XXLarge correlated highly with each other (> 0.6), with ELECTRA-Large (> 0.5), and with A2 and A3 RoBERTa-

[10] TRICKY was the only inference type for which ALBERT-XXLarge wasn't the top performer; XLNet-Large performed somewhat better, largely due to stronger higher probability and lower entropy on linguistically sophisticated SYNTACTIC and PRAGMATIC examples.

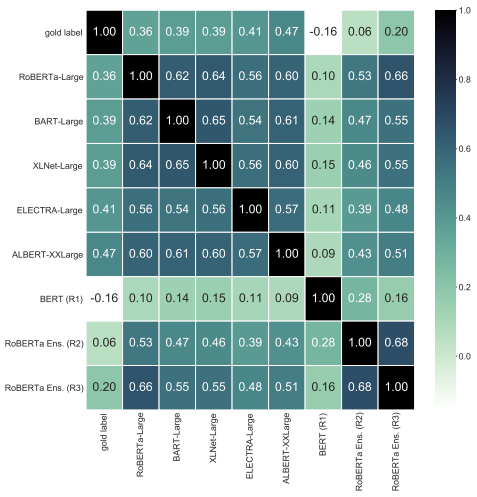

Figure 1: Correlation between gold labels and model outputs. All comparisons are significant $p < 0.01$.

Large models ($0.4 - 0.5$). RoBERTa-Large model predictions from A2 correlated with those from A3 (0.68). These results suggest that substantial improvement on ANLI may require radically new ideas, not just minor adjustments to the pretrain-finetune paradigm (c.f. Sinha et al. 2021a,b).

4.4 Analyzing Results by Genre

A3 was collected using contexts from a variety of text domains. Table 5 shows the performance of the highest performing model (ALBERT-XXLarge) across genres. Wikipedia is the least difficult genre (as well as the most frequent), Procedural is somewhat harder, then News (which is lower entropy), followed by Fiction, Legal, then RTE. Genres differ widely in how many of their examples have particular top-level tags (see Table 6). Across all genres, TRICKY and REASONING examples occur at roughly the same rates—with REASONING examples being very common across the board. Compared to the other genres, News text has more BASIC tags, and Wikipedia text has more NUMERICAL. Procedural text has the lowest rate of NUMERICAL and REFERENCE tags, but the highest rate of IMPERFECTION.

Table 7 breaks down of the performance of the ALBERT-XXLarge model by genre and tag (see Table 22 in the Appendix for the other models' performance). ALBERT-XXLarge performance on NUMERICAL examples is relatively stable across the genres, but for the other top level tags there is some variation that does not just reflect tag frequency. For example, the ALBERT-XXLarge model does better on BASIC and REASONING examples from Wikipedia, on REFERENCE examples from the Procedural genre, and on TRICKY examples from the News genre. This suggests that data from different genres could be differentially beneficial for training the skills needed for these top-level tags, suggesting that targeted upsampling could be beneficial in the future.

4.5 Other Analyses

Appendix B provides a detailed analysis of other dataset properties (word and sentence length, and most common words by round, gold label, and tag), where we show that ANLI and MNLI are relatively similar to each other but differ from SNLI. Crowdworker rationales from ANLI are explored in §B.1, Table 23–Table 24.

5 Conclusion

We release annotations of the ANLI development sets to determine which inference types are responsible for model success and failure, and how their frequencies change over dynamic data collection. Inferences relying on numerical or common sense reasoning are most prevalent, appearing in ≈40%–60% of examples. We finetuned a variety of transformer language models on NLI and compared their performance to the original target models used to adversarially collect ANLI. ALBERT-XXLarge performs the best of our 8 model sample, but there is still ample room for improvement in accuracy. Despite being frequent, examples requiring common sense reasoning, understanding of co-reference, negation and syntactic knowledge remain the most difficult. One could imagine explicit interventions to address this, perhaps incorporating insights from Sap et al. (2020), or using other modes of evaluation that explore model and data dynamics (Gardner et al., 2020; Swayamdipta et al., 2020; Rodriguez et al., 2021).

ANLI remains difficult: the huge GPT-3 model (Brown et al., 2020) barely made any progress, and even the recent DeBERTa model (He et al., 2021) cannot break 70% accuracy. We hope our annotations will inspire new innovations by enabling more fine-grained understanding of model strengths and weaknesses as ANLI matures.

Acknowledgements

Special thanks to Naman Goyal for converting the R2 and R3 RoBERTa models to something modern `fairseq` can load. Thanks as well to Yixin Nie, Mohit Bansal, Emily Dinan, Grusha Prasad, Nikita Nangia, and Alex Wang for comments and suggestions along the way. Thanks as well to several rounds of anonymous reviewers from the *ACL community and SCIL.

References

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, 25(4):543–557.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3196–3200, Istanbul, Turkey. European Languages Resources Association (ELRA).

Isaac I Bejar, Roger Chaffin, and Susan Embretson. 2012. *Cognitive and psychometric analysis of analogical problem solving*. Springer Science & Business Media.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *International Conference on Agents and Artificial Intelligence (ICAART2)*, pages 919–931.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. 2017. The Groningen Meaning Bank. In *Handbook of linguistic annotation*, pages 463–496. Springer.

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.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online. 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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Gennaro Chierchia et al. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, 3:39–103.

Noam Chomsky. 1980. *On cognitive structures and their development: A reply to Piaget*. Harvard University Press.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Peter Clark. 2018. What knowledge is needed to solve the RTE5 textual entailment challenge?

Robin Cooper, Crouch Dick, Jan van Eijck, Chris Fox, Joseph van Genabith, Han Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Brisco, Holger Maier, and Karsten Konrad. 1996. Using the framework. technical report lre 62-051r. The FraCaS consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment

challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

DataPerf Working Group. 2021. DataPerf: Benchmarking data for better ML. Technical report.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

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.

Yaroslav Fyodorov, Yoad Winter, and Nissim Francez. 2000. A natural logic inference system. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS-2)*. Citeseer.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences. *arXiv preprint arXiv:1810.13033*.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1161–1166. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

H Paul Grice. 1975. Logic and conversation. *1975*, pages 41–58.

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. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: decoding-enhanced BERT with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. 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.

Alain Joubert, Mathieu Lafourcade, and Nathalie Le Brun. 2018. The jeuxdemots project is 10 years old: What we have learned. *Games and Gamification for Natural Language Processing*, 22.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, Dieuwke Hupkes, and et al., editors. 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy.

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*.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*.

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.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

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.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of NLI models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. 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.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R Bowman. 2021. Does putting a linguist in the loop improve nlu data collection? *arXiv preprint arXiv:2104.07179*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2019. Probing natural language inference models through semantic fragments. *arXiv preprint arXiv:1909.07521*.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan. 2019. Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. "ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208, Uppsala, Sweden. Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Colorado.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked language modeling and the distributional

hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Assaf Toledo, Sophia Katrenko, Stavroula Alexandropoulou, Heidi Klockmann, Asher Stern, Ido Dagan, and Yoad Winter. 2012. Semantic annotation for textual entailment recognition. In *Mexican International Conference on Artificial Intelligence*, pages 12–25. Springer.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of LREC*.

Clara Vania, Ruijie Chen, and Samuel R. Bowman. 2020. Asking Crowdworkers to Write Entailment Examples: The Best of Bad options. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 672–686, Suzhou, China. Association for Computational Linguistics.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2021. Analyzing dynamic adversarial training data in the limit. *arXiv preprint arXiv:2110.08514*.

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.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019a. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. *International Conference on Learning Representations*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2020. The universal decompositional semantics dataset and decomp toolkit. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5698–5707, Marseille, France. European Language Resources Association.

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.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.

A Further Details on Annotation

A.1 Details of the Annotation Scheme

A full ontology, comprising all four levels, is provided together with examples in Table 14.

To give an idea of what sorts of information falls under each tag, we will go through them in turn. NUMERICAL classes refer to examples where numerical reasoning is crucial for determining the correct label, and break down into CARDINAL, ORDINAL—along the lines of Ravichander et al. (2019)—COUNTING and NOMINAL; the first two break down further into AGES and DATES if they contain information about either of these topics. BASIC consists of staple types of reasoning, such as lexical hyponymy and hypernymy (see also Glockner et al. 2018), conjunction (see also Toledo et al. 2012; Saha et al. 2020), and negation (see also Hossain et al. 2020). REFERENCE consists of pairs that require noun or event references to be resolved (either within or between context and hypothesis). TRICKY examples require either complex linguistic knowledge, say of pragmatics or syntactic verb argument structure, reorderings, word games (e.g., anagrams, acrostic jokes), and foreign language content (TRANSLATION).

REASONING examples require the application of reasoning outside of what is provided in the example alone; it is divided into three levels. The first is PLAUSIBILITY, which was loosely inspired by Bhagavatula et al. (2020); Chen et al. (2020), for which the annotator provided their subjective intuition on how likely the situation is to have genuinely occurred (for example 'when computer games come out they are often buggy' and 'lead actors get paid the most' are likely). PLAUSIBILITY also contains DEBATABLE examples, which depend on opinion or scalar adjectives like "big" (e.g. a big mouse is "big" for a mouse, but not big when compared to an elephant). The other two FACTS and CONTAINMENT refer to external facts about the world (e.g., 'what year is it now?') and relationships between things (e.g., 'Australia is in the southern hemisphere'), respectively, that were not clearly provided by the example pair itself.

There is also a catch-all class labeled IMPERFECTION that catches not only label "errors" (i.e., rare cases of labels for which the expert annotator(s) disagreed with the gold label from the crowdworker-annotator), but also spelling mistakes (SPELLING), event coreference examples (EVENTCOREF[11]), and pairs that could reasonably be given multiple correct labels (AMBIGUOUS). The latter are likely uniquely subject to human variation in entailment labels, *à la* Pavlick and Kwiatkowski (2019), Min et al. (2020), Nie et al. (2020b), since people might vary on which label they initially prefer, even though multiple labels might be possible.

Exhaustive List of Tags. In the actual dataset, tags at different levels are dash-separated, as in REASONING-PLAUSIBILITY-LIKELY. These include: BASIC CAUSEEFFECT, BASIC COMPARATIVESUPERLATIVE, BASIC COORDINATION, BASIC FACTS, BASIC IDIOMS, BASIC LEXICAL DISSIMILAR, BASIC LEXICAL SIMILAR, BASIC MODUS, BASIC NEGATION, EVENTCOREF, IMPERFECTION AMBIGUITY, IMPERFECTION ERROR, IMPERFECTION NONNATIVE, IMPERFECTION SPELLING, NUMERICAL CARDINAL, NUMERICAL CARDINAL AGE, NUMERICAL CARDINAL COUNTING, NUMERICAL CARDINAL DATES, NUMERICAL CARDINAL NOMINAL, NUMERICAL CARDINAL NOMINAL AGE, NUMERICAL CARDINAL NOMINAL DATES, NUMERICAL ORDINAL NUMERICAL ORDINAL AGE, NUMERICAL ORDINAL DATES, NUMERICAL ORDINAL NOMINAL, NUMERICAL ORDINAL NOMINAL DATES, REASONING CAUSEEFFECT, REASONING CONTAINMENT LOCATION, REASONING CONTAINMENT PARTS, REASONING CONTAINMENT TIMES, REASONING DEBATABLE, REASONING FACTS, REASONING-PLAUSIBILITY LIKELY, REASONING PLAUSIBILITY UNLIKELY, REFERENCE COREFERENCE, REFERENCE FAMILY, REFERENCE NAMES, TRICKY EXHAUSTIFICATION, TRICKY PRAGMATIC, TRICKY SYNTACTIC, TRICKY TRANSLATION, TRICKY WORDPLAY.

In addition to these tags, some top-level tags are associated with a -0 flag; these are very rare (less than 30 of these in the dataset). The zero-flag was associated with examples that didn't fall into any lower level categories. Finally, for the purposes of this paper, we collapsed two second-level tags BASIC CAUSEEFFECT and BASIC MODUS[12] into BASIC-IMPLICATIONS because these types were rare, we felt the two are related.

[11] SNLI and MNLI annotation guidelines required annotators to assume event coreference.

[12] MODUS labeled classical inference types such as Modus Ponens, Modus Tollens, etc.

Tag	Agreement (%)	A1 # of Tags	A2 # of Tags
REASONING	59.1%	176	226
BASIC	69.2%	122	128
REFERENCE	64.5%	88	136
NUMERICAL	88.6%	94	112
TRICKY	64.5%	89	105
IMPERFECTION	81.2%	44	56
EVENTCOREF	89.2%	11	29
REASONING-FACTS	54.8%	61	174
REFERENCE-COREFERENCE	66.2%	72	109
REASONING-PLAUSIBILITY	71.2%	104	70
BASIC-LEXICAL	73.9%	67	69
NUMERICAL-CARDINAL-DATES	**92.9%**	51	68
TRICKY-PRESUPPOSITION	74.9%	19	66
BASIC-NEGATION	**94.3%**	34	33
REFERENCE-NAMES	82.2%	22	45
NUMERICAL-CARDINAL	**92.9%**	23	38
BASIC-CONJUNCTION	87.9%	12	38
TRICKY-SYNTACTIC	88.2%	33	12
EVENTCOREF	89.2%	11	29
TRICKY-TRANSLATION	**92.6%**	15	23
TRICKY-EXHAUSTIFICATION	**94.6%**	22	14
IMPERFECTION-SPELLING	**93.3%**	15	15
REASONING-CONTAINMENT-LOCATION	**96.6%**	15	13
NUMERICAL-CARDINAL-AGE	**98.6%**	14	12
IMPERFECTION-NONNATIVE	**94.3%**	5	20
IMPERFECTION-LABEL	**93.6%**	8	17
IMPERFECTION-AMBIGUITY	**93.9%**	16	8
BASIC-COMPARATIVESUPERLATIVE	**95.3%**	17	5
REASONING-CONTAINMENT-TIME	**94.3%**	16	5
BASIC-CAUSEEFFECT	**95.9%**	8	12
NUMERICAL-ORDINAL	**98.6%**	9	9
NUMERICAL-CARDINAL-COUNTING	**99.3%**	7	9
NUMERICAL-CARDINAL-NOMINAL-DATES	**95.3%**	0	14
TRICKY-WORDPLAY	**96.9%**	8	5
NUMERICAL-CARDINAL-NOMINAL	**96.3%**	6	7
BASIC-IDIOM	**96.3%**	7	6
REFERENCE-FAMILY	**99.3%**	5	5
NUMERICAL-ORDINAL-DATES	**97.9%**	4	2
BASIC-0	**98.3%**	4	1
IMPERFECTION-0	**98.9%**	2	1
REASONING-CONTAINMENT-PARTS	**99.6%**	1	0
REASONING-0	**99.6%**	0	1
Aggregate	**91.1% (avg)**	713 (sum)	955 (sum)

Table 8: Interannotator agreement percentages (bold exceeded 90%) and tag counts for 300 randomly sampled examples. Tags are sorted by the number of usages of that tag by either annotator.

More Examples from the Annotation Guidelines. Some tags required sophisticated linguistic domain knowledge, so more the annotation guidelines included more examples (some will be provided here). For example, the TRICKY-EXHAUSTIFICATION is wholly novel, i.e., not adopted from, or similar to, any other semantic annotation scheme known to the authors. This tag marks examples where the original crowdworker-annotator assumed that only one predicate holds of the topic, and that other predicates don't. Often TRICKY-EXHAUSTIFICATION examples have the word "only" in the hypothesis, but that's only a tendency: observe the context, *Linguistics is the scientific study of language, and involves an analysis of language form, language meaning, and language in context* and the hypothesis *Form and meaning are the only aspects of language linguistics is concerned with*, which gets labeled as a contradiction.[13] For this example, the crowdworker-annotator wrote a hypothesis that excludes one of the core properties of linguistics provided in the context and claims that the remaining two they list are the only core linguistic properties.

To take another example, also a contradiction: For the context, *The Sound and the Fury is an American drama film directed by James Franco. It is the second film version of the novel of the same name by William Faulkner* and hypothesis *Two Chainz actually wrote The Sound and the Fury*, we have a TRICKY-EXHAUSTIFICATION tag. The Gricean Maxims of Relation and Quantity (Grice, 1975) require the writer of the original context to be maximally cooperative and informative, and thus, to list all the authors of *The Sound and Fury*. Since the context only listed Faulkner, we con-

[13]This example also receives BASIC-COORDINATION, and BASIC-LEXICAL-SIMILAR for "involves" and "aspects"/"concerned with".

clude that the book only had one author, Faulkner, and Two Chainz did *not* in fact (co-)author *The Sound and the Fury*.[14]

As we mentioned above, any one example sentence pair can receive multiple tags. An example with a hypothesis *George III comes after George II* would receive tags REFERENCE-NAMES (because we are comparing the names of two individuals), and NUMERICAL-ORDINAL (because we are comparing the roman numerals for first and second). A pair with the context *Sean Patrick Hannity...is an American talk show host, author, and conservative political commentator...* and the hypothesis *Hannity has dated a liberal* would receive the tags BASIC-LEXICAL (because of the relation between "conservative" and "liberal"), REFERENCE-COREFERENCE, (because of the coreference between "Sean Patrick Hannity" and "Hannity"), and REASONING-UNLIKELY (because it's unlikely given world knowledge that a liberal and a conservative commentator would date, although it's definitely possible).

The annotation guidelines also provided examples to aid in disentangling REFERENCE-NAMES from REFERENCE-COREFERENCE, as they often appear together. REFERENCE-COREFERENCE should be used when resolving reference between non-string matched noun phrases (i.e. DPs) is necessary to get the label: ***Mary Smith***$_i$ *was a prolific author.* ***She***$_i$ *had a lot of published works by 2010.*⇒***Smith***$_i$ *published many works of literature.* REFERENCE-NAMES is used when the label is predicated on either (i) a discussion of names, or (ii) resolving multiple names given to a person, but the reference in the hypothesis is an exact string match to one of the options: ***La Cygne***$_i$ *(pronounced* ***"luh SEEN"****) is a city in the south of France.*⇒***La Cygne***$_i$ *is in France.* Some examples require both REFERENCE-COREFERENCE and REFERENCE-NAMES tags: ***Mary Beauregard Smith, the fourth grand Princess of Winchester*** *was a prolific author.*⇒***Princess Mary*** *wrote a lot.*

A.2 Inter-Annotator Agreement

Annotation guidelines for each tag were discussed verbally between the two annotators during the training of the second expert. The main expert annotator trained the second by first walking through the annotation guidelines (i.e., Table 2), answering any questions, and providing additional examples taken from their experience as necessary. The second expert then annotated 20 randomly sampled examples from the R1 training set as practice.

	Average			Top Level Tags		
	Precision	Recall	F1	Precision	Recall	F1
A_1	0.55	0.42	0.44	0.59	0.73	0.61
A_2	0.42	0.55		0.73	0.59	

Table 9: Average Precision, Recall and F1 between our two annotators on 300 randomly selected development set examples. A_1 was taken with the original annotator as ground truth, A_2 with the second expert. Recall that X to Y precision is equivalent to Y to X recall.

The two annotators subsequently discussed their selections on these training examples when they differed. Of course, there is some subjectivity inherent in this annotation scheme, which crucially relies on expert opinions about what information in the premise or hypothesis could be used to determine the correct label. After satisfactorily coming to a conclusion (i.e., a consensus for all 20 examples), the second annotator was provided with another set of 20 randomly sampled examples, this time from the R3 training set (to account for genre differences across rounds), and again, discussion was repeated until consensus was reached. Several further discussions took place. Once both annotators were confident in the second expert annotator's understanding of the scheme, the secondary annotator was provided with 3 random selections of 100 examples (one from each development set) as the final set to calculate inter-annotator agreement from. The second annotator was also provided with the exhaustive tag list (above), which includes some splits that subcategorize the tags from Table 2 even further. The tags are visible in Table 8, along with percent agreement for each tag.

To provide additional NLI-internal context for our percent agreement results, we note that percent agreement on both top and lower level tags exceeds the percent agreement of non-experts on the task of NLI as reported in Bowman et al. (2015) and Williams et al. (2018). Recall that performing NLI is a subtask of our annotations (i.e., experts must check the NLI label to determine if there was an error and must also then tag contained phenomena that contribute to the label decision).

Since our annotation scheme incorporated some subjectivity—i.e., annotators tag phenomena they

[14]This pair also gets TRICKY-PRAGMATIC, and EVENT-COREF and BASIC-LEXICAL-SIMILAR tags.

Our Scheme's Tag	Other Scheme's Tag (Citation)
BASIC-NEGATION	Negation (Naik et al., 2018; Hossain et al., 2020; Geiger et al., 2020)
BASIC-LEXICAL-DISSIMILAR	Antonymy (Naik et al., 2018), Contrast (Bejar et al., 2012); Ch. 3[15]
BASIC-LEXICAL-SIMILAR	Overlap (Naik et al., 2018), Similar (Bejar et al., 2012); Ch. 3, hyponym/hypernym (Geiger et al., 2020), Lexical (Joshi et al., 2020)
BASIC-CAUSEEFFECT	Cause-Purpose (Bejar et al., 2012); Ch. 3, cause (Sammons et al., 2010), Cause and Effect (LoBue and Yates, 2011)
BASIC-COORDINATION	Conjoined Noun Phrases (Cooper et al., 1996), ConjNLI (Saha et al., 2020), "Connectives" (Joshi et al., 2020)
BASIC-COMPARATIVESUPERLATIVE	Comparatives (Cooper et al., 1996), , "Connectives" (Joshi et al., 2020)
NUMERICAL	numeric reasoning, numerical quantity (Sammons et al., 2010), Mathematical (Joshi et al., 2020)
NUMERICAL-CARDINAL	cardinal (Ravichander et al., 2019)
NUMERICAL-ORDINAL	ordinal (Ravichander et al., 2019)
REFERENCE-COREFERENCE	Anaphora (Inter-Sentential, Intra-Sentential) (Cooper et al., 1996), coreference (Sammons et al., 2010)
REFERENCE-COREFERENCE with REFERENCE-NAMES	Representation (Bejar et al., 2012); Ch. 3
REFERENCE-FAMILY	parent-sibling, kinship (Sammons et al., 2010)
REFERENCE-NAMES	name (Sammons et al., 2010)
REASONING-DEBATABLE	Cultural/Situational (LoBue and Yates, 2011), Defeasible Inferences (Rudinger et al., 2020)
REASONING-PLAUSIBILITY-LIKELY	Probabilistic Dependency (LoBue and Yates, 2011)
REASONING-CONTAINMENT-TIMES	Temporal Adverbials (Cooper et al., 1996), Space-Time (Bejar et al., 2012); Ch. 3, event chain, temporal (Sammons et al., 2010)
REASONING-CONTAINMENT-LOCATION	spatial reasoning (Sammons et al., 2010), Geometry (LoBue and Yates, 2011)
REASONING-CONTAINMENT-PARTS	Part-Whole, Class-Inclusions (Bejar et al., 2012); Ch. 3, has-parts (LoBue and Yates, 2011)
REASONING-FACTS	Real World Knowledge (Naik et al., 2018; Clark, 2018; Bernardy and Chatzikyriakidis, 2019)
TRICKY-SYNTACTIC	passive-active, missing argument, missing relation, simple rewrite, (Sammons et al., 2010)
IMPERFECTIONS-AMBIGUITY	Ambiguity (Naik et al., 2018)

Table 10: Comparisons between our tagset and tags from other annotation schemes.

believe a human would use to provide the NLI label for the example—annotators are likely to have different blindspots. Descriptively, annotators differed slightly in the number of tags they assign on average: the original annotator assigns fewer tags per example (Mean $= 2.25$, Std. $= 1.01$) than the second expert (Mean $= 3.02$, Std. $= 1.45$). The number of tags in the intersection of the two was predictably lower (Mean $= 1.20$, Std.$= 0.85$) than either annotator's average or the union (Mean $= 4.07$, Std. $= 1.55$).

In addition to agreement percentages that are reported in Table 8, we report average precision, recall, and F1 (a weighted average of the two) for our annotations in Table 9.[16] For percentages, we note that agreement was generally higher for rarer tags. The most frequent top-level tag, REASONING, had the lowest agreement, perhaps due to disagreements in REASONING-FACTS, where the subjectivity of decisions likely drove down agreement. Subjectivity might be expected for REASONING-PLAUSIBILITY examples as well, because it is hard to be sure whether a particular fact is necessary for the label (particular in the case of REASONING-PLAUSIBILITY-DEBATABLE. REASONING-PLAUSIBILITY also showed some disagreement, as people differ whether they feel compelled to note that the likelihood of a context is relevant for the label decision. Finally, we note that frequent lower level tags NUMERICAL-CARDINAL(-DATES) and BASIC-NEGATION had the highest agreement.

Although we report accuracy (i.e., percentage agreement), F1 is usually more useful than accuracy, especially if you have an uneven class distribution (as we do). For this reason, we additionally report F1, precision and recall between the two annotators (reporting statistics twice, once with each annotator taken to be ground truth). Precision, recall and F1 are all fairly high (recall that these three measures are upper bounded by 1), but are higher for top level tags than for the average of all tags. We believe this is an acceptable level of agreement, especially given the difficulty of the task and the fact that tags vary in how subjective their decisions are.

A.3 Direct Comparisons to other Annotation Schemes

Our scheme derives its inspiration from the wealth of prior work on types of sentential inference both within and from outside NLP—Cooper et al. (1996); Sammons et al. (2010); LoBue and Yates (2011); Jurgens et al. (2012); Jia and Liang (2017); White et al. (2017); Naik et al. (2018); Nie et al. (2019); Kim and Linzen (2020); Yu and Ettinger (2020); White et al. (2020), i.a. When one implements an annotation scheme, one must decide on the level of depth one wants to achieve. On the one hand, a small number of tags can allow for easy annotation (by non-experts or even automatically), whereas on the other, a more complicated and complete annotation scheme (like, e.g., Cooper et al. 1996; Bejar et al. 2012) can allow for a better understanding of the full range of possible phenomena that might be relevant. (Note: for contextualization, our tags are greater in tag number than Naik et al. (2018) but smaller and more manageable than Cooper et al. 1996 and Bejar et al. 2012). We wanted annotations that allow for an

[16]For all statistics that aggregate tag results, we did not include Imperfection tags, as imperfections can be difficult to spot and annotator differences for these tags typically only represent whether an annotator noticed a mistake when the other did not.

evaluation of model behavior on a phenomenon-by-phenomenon basis, in the spirit of Weston et al. (2016); Wang et al. (2018); Jeretic et al. (2020)—but unlike Jia and Liang (2017). We also wanted to be able to detect interactions between phenomena (Sammons et al., 2010). Thus, we implemented our hierarchical scheme (for flexible tag-set size) in a way that could provide all these desiderata.

Table 10 provides a by-tag comparison between our annotation scheme and several others. Only direct comparisons are listed in the table; in other cases, our scheme had two tags where another scheme had one, or vice versa. Some of these examples are listed below, by the particular inference types for each annotation scheme.

Several labels from Naik et al. (2018)'s annotation scheme concur with ours, but ours has much wider coverage. In fact, it is a near proper superset of their scheme. Both taxonomies have a NEGATION tag, an AMBIGUITY tag, a REAL WORLD KNOWLEDGE—which for us is REASONING-FACTS, and a ANONYMY tag—which for us is BASIC-LEXICAL-DISSIMILAR. Additionally, both annotation schemes have a tag for numerical reasoning. We didn't include "word overlap" as that is easily automatable and would thus be an inappropriate use of limited hand-annotation time. Instead, we include a more flexible/complex notion of overlap in our BASIC-LEXICAL-SIMILAR tag, which accounts not only for synonyms, but also for phrase level paraphrases.

Our scheme can handle nearly all of the inference types in Sammons et al. (2010). For example, their 'numerical reasoning' tag maps onto a combination of NUMERICAL tags and REASONING-FACTSfor us to account for external mathematical knowledge. A combination of their 'kinship' and 'parent-sibling' tags is present in our REFERENCE-FAMILY tag. One important difference between our approach and theirs is that we do not separate negative and positive occurrences of phenomena; both would appear under the same tag for us. One could imagine performing a further round of annotation on the ANLI data to separate positive from negative as Sammons et al. does.

Several of the intuitions of the LoBue and Yates (2011) taxonomy are present in our scheme. For example, their 'arithmetic' tag roughly corresponds to a combination of our NUMERICAL-CARDINAL and REASONING-FACTS (i.e., for mathematical reasoning). Examples labeled with their "preconditions" tag would receive our TRICKY-PRAGMATIC tag. Interestingly, our TRICKY-EXHAUSTIFICATION tag seems to be a combination of their 'mutual exclusivity', 'omniscience' and 'functionality' tags. Other relationships between our tags and theirs are in Table 10.

Many of our numerical reasoning types were inspired by Ravichander et al. (2019), which showed that many NLI systems perform very poorly on many types of numerical reasoning. In addition to including cardinal and ordinal tags, as they do, we take their ideas one step further and also tag numerical examples where the numbers are not merely playing canonical roles as degrees of measure (e.g., NUMERICAL-NOMINAL and NUMERICAL-COUNTING). We also expand on their basic numerical types by specifying whether a number refers to a date or an age. For any of their examples requiring numerical reasoning, we would assign NUMERICAL as a top level tag, as well as a REASONING-FACTS tag, as we described in the paragraph above. A similar set of tags would be present for their "lexical inference" examples where, e.g., it is necessary to know that 'm' refers to 'meters' when it follows a number; in this case, we would additionally include a TRICKY-WORDPLAY tag.

The annotation tagset of Poliak et al. (2018a) overlaps with ours in a few tags. For example, their 'pun' tag is a proper subset of our TRICKY-WORDPLAY tag. Their 'NER' and 'Gendered Anaphora' fall under our REFERENCE-COREFERENCE and REFERENCE-NAMES tags. Their recasting of the MegaVeridicality dataset (White and Rawlins, 2018) would have some overlap with our TRICKY-PRAGMATIC tag, for example, for the factive pair *Someone knew something happened.* $\Rightarrow$ *something happened.*. Similarly, their examples recast from Schuler (2005, VerbNet) would likely recieve our TRICKY-SYNTACTIC tag for argument structure alternation, in at least some cases.

Rozen et al. (2019)'s tagset also has some overlap with ours, although none directly. They present two automatically generated datasets: one targets comparative reasoning about numbers—i.e., corresponding to a combination of our NUMERICAL-CARDINAL and BASIC-COMPARATIVESUPERLATIVE tags—and the other targets dative-alternation—which, like (Poliak et al., 2018a)'s recasting of VerbNet, would

Dataset		Contexts		Statements	
		Word$_{Len.}$	Sent.$_{Len.}$	Word$_{Len.}$	Sent.$_{Len.}$
ANLI	All	4.98 (0.60)	55.6 (13.7)	4.78 (0.76)	10.3 (5.28)
	A1	5.09 (**0.69**)	54.1 (8.35)	4.91 (0.74)	11.0 (5.36)
	A2	5.09 (0.47)	54.2 (8.24)	4.80 (0.77)	10.1 (4.95)
	A3	**4.73** (0.50)	**59.2** (**21.5**)	4.59 (0.76)	9.5 (5.38)
	C	5.00 (**0.79**)	55.8 (13.8)	4.76 (0.73)	**11.4** (**6.51**)
	N	4.97 (0.47)	55.4 (13.8)	4.83 (0.78)	9.4 (4.49)
	E	5.00 (0.49)	55.7 (13.6)	4.75 (0.78)	10.3 (4.44)
MNLI	All	4.90 (0.97)	19.5 (13.6)	4.82 (0.90)	10.4 (4.43)
	M	**4.88** (**1.10**)	19.3 (14.2)	4.78 (0.92)	9.9 (4.28)
	MM	**4.93** (**0.87**)	19.7 (13.0)	4.86 (0.89)	10.8 (4.53)
	C	4.90 (0.97)	19.4 (13.6)	4.79 (0.90)	9.7 (3.99)
	N	4.90 (0.98)	19.4 (13.8)	4.79 (**0.85**)	10.9 (4.46)
	E	4.91 (0.96)	19.6 (13.5)	4.86 (**0.95**)	10.4 (4.71)
SNLI	All	4.31 (0.65)	14.0 (6.32)	4.23 (0.75)	7.5 (3.14)
	C	4.31 (0.64)	14.0 (6.35)	4.16 (0.71)	7.4 (2.90)
	N	4.31 (0.66)	13.8 (6.28)	4.26 (0.72)	8.3 (**3.36**)
	E	4.31 (0.64)	14.0 (6.31)	4.26 (**0.81**)	6.8 (2.90)

Table 11: Average length of words and sentences in contexts, statements, and reasons for ANLI, MultiNLI, SNLI. Average and (standard deviation).

Dataset	Word$_{Len.}$	Sent.$_{Len.}$	Count
All	4.54 (0.69)	21.05 (13.63)	3200
R1	**4.57** (0.65)	22.40 (13.80)	1000
R2	4.51 (**0.71**)	20.14 (12.96)	1000
R3	4.55 (0.70)	**20.81** (**14.11**)	1200
C	4.53 (0.70)	19.46 (12.64)	1062
N	4.52 (0.64)	**23.81** (**15.05**)	1066
E	**4.58** (**0.72**)	19.87 (12.66)	1070
Numerical	4.44 (0.65)	21.79 (13.21)	1036
Basic	**4.63** (0.69)	21.31 (13.92)	1327
Reference	4.53 (0.70)	20.04 (13.01)	868
Tricky	4.56 (**0.71**)	20.58 (13.22)	893
Reasoning	4.52 (0.66)	**21.82** (**14.08**)	1197
Imperfection	4.53 (**0.71**)	19.26 (13.06)	452

Table 12: Average length of words and sentences in rationales for ANLI. Average and (standard deviation).

probably correspond to our TRICKY-SYNTACTIC.

White et al. (2017) uses pre-existing semantic annotations to create an RTE/NLI formatted dataset. Their approach has several strong benefits, not the least of which is its use of minimal pairs to generate examples that can pinpoint exact failure points. For the first of our goals—understanding the contents of ANLI in particular—it would be interesting to have such annotations, and this could be a potentially fruitful future direction for research. But for the other—understanding current model performance on ANLI—it is not immediately clear to us that annotating ANLI for lexical semantic properties of predicates and their arguments (e.g., volition, awareness, and change of state) would help. In the end, it is an empirical question for future work.

From the above pairwise comparisons between existing annotation schemes (or data creation schemes), it should be clear there are many shared intuitions and many works are attempting to capture similar phenomena. We believe our tags thread the needle in a way that incorporates the best parts of the older annotation schemes while also innovating new phenomena and ways to view phenomena in relation to each other. In particular, very few of the schemes cited above arrange low level phenomena into a comprehensive multilevel hierarchy. This is one of the main benefits of our scheme. Our hierarchy allows us to compare models at multiple levels, and hopefully, as our models improve, it can allow us to explore transfer between different reasoning types.

B Dataset Properties

To further describe the ANLI dataset, we measure the length of words and sentences across all rounds and across all gold labels. We compare ANLI to SNLI and MNLI in Table 11. We also report length of rationales in Table 12. As the tables show, the statistics across classification labels are roughly the same within each dataset. It is easy to see that ANLI contains much longer contexts than both MNLI and SNLI. Overall, ANLI and MNLI appear more similar in statistics to each other than to SNLI, having have longer statements and longer words.

We analyzed the top 25 most frequent words (with stopwords removed based on the NLTK[17] stopword list) in development set contexts, statements, and rationales. We investigate frequent words for the entire dataset, by round, and by gold label (see Table 23), and by top-level annotation tag (see Table 24). The most frequent words in contexts reflect the domains of the original text. Since Wikipedia contexts were the most frequent in ANLI, words from Wikipedia including, for example 'film', 'album', 'directed', 'football', 'band', 'television' predictably figure prominently. References to nations, such as 'american', 'state', and 'national' are also common—perhaps reflecting a North American bias in the dataset.

Statements written by crowdworkers show a preference instead for terms like 'born', 'died', and 'people', suggesting again, that Wikipedia contexts, consisting largely of biographies, have a specific genre effect on constructed statements.

[17] https://www.nltk.org/

Model	A1	A2	A3	ANLI	hyperparameters
BERT (R1)	0	28	32	21	24-layer, 1024-hidden, 16-heads, 335M param.
RoBERTa (R2) Ens.	67	18	22	35	24-layer, 1024-hidden, 16-heads, 355M param.
RoBERTa (R3) Ens.	72	45	20	44	24-layer, 1024-hidden, 16-heads, 355M param.
RoBERTa-Large	74	51	46	56	24-layer, 1024-hidden, 16-heads, 355M param.
BART-Large	74	52	50	58	24-layer, 1024-hidden, 16-heads, 406M param.
XLNet-Large	74	52	51	58	24-layer, 1024-hidden, 16-heads, 340M param.
ELECTRA-Large	67	54	55	58	24-layer, 1024-hidden, 16-heads, 335M param.
ALBERT-XXLarge	**76**	**57**	**57**	**63**	12 repeating layer, 4096-hidden, 64-heads, 223M param.

Table 13: Accuracy for each model on the ANLI Development Sets (highest accuracy is bolded). Hyperparameters also provided. 'Ens.' refers to one model randomly selected from an ensemble of different seeds

Several examples appear in the top 25 most frequent words for both statements and contexts, including 'film', 'american', 'one', 'two', 'not', 'first', 'new', 'played', 'album', and 'city'. In particular, words such as 'one', 'first', 'new', and 'best' in contexts appear to be opposed by (near) antonyms such as 'two', 'last', 'old', 'least', and 'less' in statements. This suggests the words present in a context might affect how crowdworkers construct statements, potentially suggesting some lexical confounds in ANLI. Finally, we observe that the top 25 most frequent words in contexts are used roughly 3 times as often as the top 25 most frequent words in statements. This suggests that statements have wider and more varied vocabulary than contexts do.

B.1 Analyzing Annotator Rationales

We observe that the most frequent words in rationales differ from those in contexts and statements. The original annotators often use 'statement' and 'context' in their rationales to refer to example pairs, as well as 'system' to refer to the model; this last term is likely due to the fact that the name of the Mechanical Turk task used to employ crowdworkers in the original data collection was called "Beat the System" (Nie et al., 2020a, App. E). The set of most frequent words in rationales also contains, predictably, references to the model performance (e.g., 'correct', 'incorrect'), and to speech act verbs (e.g., 'says', 'states').

Interestingly, there is a higher number of verbs in the rationales denoting mental states (e.g., 'think', 'know', 'confused'), which suggests that the annotators could be ascribing theory of mind to the system, or at least using mental-state terms metaphorically—which could be due to the Nie et al. (2020a) data collection procedure that encourages crowdworkers to think of the model as an adversary. Rationales also contain more modals (e.g., 'probably', 'may', 'could'), which are often used to mark uncertainty, suggesting that the annotators are aware of the fact that their rationales might be biased by their human expectations. Finally, we note that the top 25 most frequent words used in rationales are much more common than are the top 25 most frequent words in contexts (by roughly two times) or in statements (by roughly 5-6 times). This suggests that vocabulary used for writing rationales is smaller than that in the contexts (from domains such as Wikipedia), and crowdworker annotated statements.

C Tag Breakdowns

Table 15 shows a breakdown of second-level tag incidence by top-level tag.

D Development Set Accuracies for 8 Transformer Models

Table 13 shows development set accuracies for all transformer models, by round. ANLI is still quite challenging, with even SOTA models barely exceeding 50% accuracy (although remember that the development set is approximately balanced 3-way classification, so we are beating random baseline). The ALBERT-XXLarge model achieves the highest accuracy on the full development set, reaching approximately 63% correct. On A1, the accuracy between the ALBERT-XXLarge and the other SOTA models hovers around two points, extending to 5–6 percentage points on A2, and 6–11 points on A3; the gap between ALBERT-XXLarge and the other SOTA models on the full ANLI development set hovers between 5 and 7 points.

E Model Predictions Breakdown by Tag

Model predictions by specific tags are in Table 16 (BASIC), Table 17 (NUMERICAL), Table 18 (REASONING), Table 19 (REFERENCE), Table 20 (TRICKY), Table 21 (IMPERFECTIONS).

For NUMERICAL, COUNTING is the hardest, which makes sense given that COUNTING examples are relatively rare, and require that one actually counts phrases in the text, which is a metalinguistic skill. ORDINAL is the next most difficult category, perhaps because, like COUNTING examples, ORDINAL examples are relatively rare.[18] For

[18] Additionally, it seems difficult for models to bootstrap their CARDINAL number knowledge for ORDINAL numbers. One might hope that a model could bootstrap its knowledge of the order of cardinal numbers (e.g., that *one* comes before *two* and *three*) to perform well on their corresponding ordinals, However, numerical order information doesn't seem to be generally applied in these models. Perhaps this is because

BASIC, IMPLICATION, IDIOM and NEGATION were more difficult than LEXICAL, COMPARATIVE & SUPERLATIVE and COORDINATION. For REFERENCE, there is a lot of variation in the behavior of different models, particularly for the NAMES examples, although also for COREFERENCE examples, making it difficult to determine which is more difficult. Finally, for TRICKY examples, WORDPLAY examples the most difficult, again because these require complex metalinguistic abilities (i.e., word games, puns, and anagrams), but they are followed closely by EXHAUSTIFICATION examples, which require a complex type of pragmatic reasoning.[19]

F Model Predictions Breakdown by Domain

Table 22 shows the breakdown by genre. Wikipedia results correspond with the overall dataset: ALBERT-XXLarge performs the best on everything except TRICKY (where XLNET-LARGE performs best. ALBERT-XXLarge performs best nearly across the board on procedural text (being narrowly edged out by ELECTRA-Large on REASONING) and fiction (where ELECTRA-Large performs best on REFERENCE, and where BART-Large and ELECTRA-Large jointly take top slot for TRICKY). Finally, the news genre has the most variation: ALBERT-XXLarge still performs well on BASIC, TRICKY, REASONING tags, although ELECTRA-Large narrowly beats it on NUMERICAL; XLNet-Large beats out all others on REFERENCE in the news genre by 3+ points.

We aim to characterize relative performance between the models and note variation between model performances on different genres. For example, BART and RoBERTa struggle with fiction (except for on the TRICKY tag). For example, ELECTRA-Large performs quite well on NUMERICAL examples from the Wikipedia, news, and procedural datasets, but poorly on NUMERICAL examples from Fiction. Similary BART-LARGE performs well on TRICKY examples from Wikipedia, fiction, and news, but struggles with TRICKY examples in procedural text. To give a final example, RoBERTa-Large and XLNet-Large do well on REFERENCE examples in procedural text and Wikipediate to some extent (and, for XLNet-Large, also news text), but they struggle with fiction (and, for RoBERTa-Large, also news). Since models do not perform similarly on particular tags across genres, we suggest they have not learned fully generalizable knowledge corresponding to these tag types.[20]

many common ordinal numbers in English are not morphologically composed of their cardinal counterparts (e.g., *one* and *first*, *two* and *second*.

[19] See Chierchia et al. (2004) for a summary of the linguistic theory on exhaustification, although we adopt a wider definition of the phenomenon for the tag here as in Table 14.

[20] Although we analyze examples in the aggregate to abstract away from particular example idiosyncrasies, remember that examples can be tagged with any number of other inference types and may vary in many other features (e.g., length, vocabulary etc.), so they are not strictly comparable, and more work needs to be done to bolster these conclusions.

Top Level	Second Level	Third Level	Context	Hypothesis	Round	Label	Other Tags
Num.	Cardinal	Dates	Otryadyn Gündegmaa (. . . born **23 May** 1978), is a Mongolian sports shooter. . . .	Otryadyn Gündegmaa was born on May 23rd	A1	E (N)	Ordinal, Dates
		Ages	. . . John Fox probably won't roam an NFL sideline again. . . **the 63-year-old** Fox will now move into an analyst role. . .	John Fox **is under 60 years old**.	A3	C (E)	Ref., Coref.
	Ordinal	Dates	Black Robe. . . is a historical novel by Brian Moore set in New France in **the 17th century**. . .	Black Robe is a novel set in New France in **the mid 1600s**	A2	N (E)	Reasoning, Plaus., Likely, Cardinal
		Ages	John Barnard (**6 July 1794** at Chislehurst, Kent; **died 17 November 1878** at Cambridge, England) was an English amateur cricketer who played first-class cricket from 1815 to 1830. M. . .	John Barnard died before **his fifth birthday**.	A1	C (N)	Cardinal, Dates, Reasoning, Facts
	Counting		. . . The Demand Institute **was founded** in 2012 **by Mark Leiter and Jonathan Spector**. . .	The Demand Institute **was founded by two men**.	A2	E (N)	Ref., Names
	Nominal		Raúl Alberto Osella (born 8 June 1984 in Morteros) is an Argentine association footballer . . . He played **FIFA U-17 World Cup Final** for Argentina national team in **2001**. . . .	Raul Alberto Osella no longer plays for **the FIFA U-17 Argentina team**.	A2	E (N)	Reasoning, Facts, Tricky, Exhaust., Cardinal, Age, Dates
Basic	Lexical		. . . The **dating app Hater**, which matches users based on the things they hate, has compiled all of their data to create a map of **the foods everyone hates**. . .	Hater is an **app designed for foodies in relationships**.	A3	C (N)	
	Comp.& Super.		. . . try to hit your shot onto the upslope because they are **easier** putts to make **opposed to** downhill putts.	Upslope putts are **simple to do**	A3	N (E)	
	Implic.		[DANIDA]. . . provides humanitarian aid . . . to developing countries. . .	**Focusing on developing countries**, DANIDA hopes to improve citizens of different countries lives.	A2	E (N)	
	Idioms		. . . he set to work to hunt for his dear money. . . he found nothing; **all had been spent**. . .	The money **got up and walked away**.	A3	N (C)	Reasoning, Plaus., Unlikely
	Negation		Bernardo Provenzano . . . **was suspected of having been the head of the Corleonesi** . . .	It was **never confirmed that Bernardo Provenzano was the leader of the Corleonesi**.	A2	E (N)	Tricky, Prag.
	Coord.		. . . Dan went home and started cooking a steak. However, Dan accidentally **burned the steak**. . . .	The steak was **cooked for too long or on too high a temperature**.	A3	E (N)	Basic, Lexical, Tricky, Prag.
Ref.	Coref.		. . . Tim was a tutor. . . . His latest student really pushed him, though. Tim could not get through to **him**. He had to give up. . .	Tim gave up on **her** eventually.	A3	C (E)	
	Names		**Never Shout Never** is an EP **by Never Shout Never** which was released December 8, 2009. . . .	**Never Shout Never** has a **self titled** EP.	A1	E (N)	
	Family		Sir Hugh Montgomery . . . was the **son** of Adam Montgomery, the 5th Laird of Braidstane, by his **wife and cousin**.	Sir Hugh Montgomery had at least one **sibling**.	A2	N (E)	Reasoning, Plaus., Likely
Tricky	Syntactic		Gunby. . . **is situated close to the borders with Leicestershire and Rutland**, and 9 mi south from **Grantham**. . .	Gunby **borders Rutland an Grantham**.	A1	C (E)	Imperfect., Spelling
	Prag.		. . . Singh won **the award** for Women Leadership in Industry. . .	. . . Singh won **many awards** for Women in Leadership in Industry.	A3	C (N)	
	Exhaust.		Linguistics . . . **involves an analysis of language form, language meaning, and language in context**. . . .	Form and meaning are **the only aspects** of language linguistics is concerned with.	A1	C (N)	
	Wordplay		. . . Brock Lesnar and Braun Strowman will both be under . . . on **Raw**. . .	**Raw is not an anagram of war**	A3	C (E)	
Reasoning	Plaus.	Likely	**B. Dalton Bookseller. . . founded in 1966 by Bruce Dayton**, a member of the same family that operated the Dayton's department store chain. . .	**Bruce Dayton founded the Dayton's department store chain**.	A1	C (E)	Ref., Names
		Unlikely	The Disenchanted Forest is a 1999 documentary film that follows endangered orphan orangutans . . . returned to their rainforest home. . . .	The Disenchanted Forest is . . . about **orangutans trying to learn how to fly by building their own planes**. . .	A2	C (N)	Reasoning, Facts
		Debatable	The Hitchhiker's Guide to the Galaxy is a 2005 British-American **comic science fiction film**. . .	Hitchhiker's Guide to the Galaxy is **a humorous film**.	A1	N (E)	Basic, Lexical
	Facts		. . . [Joey] decided to make [his mom] **pretend tea**. He got some hot water from the tap and mixed in the herb. But **to his shock**, his mom really drank the tea! **She said the herb he'd picked was chamomile**, a delicious tea!	**Joey knew how to make chamomile tea**.	A3	C (E)	
	Contain.	Parts	Milky Way Farm in Giles County, Tennessee, is **the former estate of Franklin C. Mars** . . . its manor house is now a venue for special events.	**The barn** is occassionaly staged for photo shoots.	A1	N (C)	Plaus., Unlikely, Imperfect., Spelling
		Loc.	Latin Jam Workout is a Latin Dance Fitness Program. . . [f]ounded in 2007 **in Los Angeles, California**, Latin Jam Workout combines . . . music with dance. . .	Latin Jam Workout was not created in **a latin american country**	A2	E (C)	Basic, Negation
		Times	Forbidden Heaven is a 1935 American drama film. . . released **on October 5, 1935** . . .	Forbidden Heaven is . . . film released **in the same month as the holiday Halloween**.	A1		Facts
Imperfect.	Error		Albert Levitt (March 14, 1887 – **June 18, 1968**) was a judge, law professor, attorney, and candidate for political office. . . .	Albert Levitt . . . held several positions in the legal field during his life, (**which ended in the summer of 1978**). . .	A2	N (C)	Num., Cardinal, Dates
	Ambig.		Diablo is a 2015 Canadian-American psychological western . . . starring **Scott Eastwood**. . . It was the first Western starring **Eastwood**, the son of Western icon **Clint Eastwood**.	**It** was the last western starring **Eastwood**	A2	C (N)	Ref., Coref., Label, Basic, Comp.&Sup., Lexical, Num., Ordinal, Family
	Spelling		"Call My Name" is a song **recorded by Pietro Lombardi** from his first studio album "Jackpot". . . It was **written and produced by** "DSDS" jury member Dieter Bohlen. . . .	"Call my Name" was **written and recorded by Pierrot Lombardi** for his album "Jackpot".	A1	C (E)	Tricky, Syntactic, Imperfect., Spelling
	Translat.		Club **Deportivo Dénia** is a Spanish football team. . . it plays in **Divisiones Regionales de Fútbol** . . . holding home games at "**Estadio Diego Mena Cuesta**",. . .	Club **Deportivo Dénia** plays in the Spanish village "**Estadio Diego Mena Cuesta**".	A2	C (E)	Tricky, Syntactic

Table 14: Examples from the full scheme.

	Round	Overall	Cardinal	Ordinal	Counting	Nominal	Dates	Age
Numerical	A1	40.8%	37.8%	6.2%	1.9%	4.2%	27.4%	5.9%
	A2	38.5%	34.7%	6.7%	2.8%	3.5%	24.3%	6.7%
	A3	20.3%	18.6%	2.8%	2.3%	0.4%	7.1%	3.2%
	All	32.4%	29.6%	5.1%	2.3%	2.6%	18.8%	5.1%
	Round	Overall	Lexical	Compr. Supr.	Implic.	Idioms	Negation	Coord.
Basic	A1	31.4%	16.0%	5.3%	1.5%	0.3%	5.6%	5.5%
	A2	41.2%	20.2%	7.6%	2.4%	1.7%	9.8%	4.5%
	A3	50.2%	26.4%	4.9%	4.2%	2.2%	15.8%	6.1%
	All	41.5%	21.2%	5.9%	2.8%	1.4%	10.7%	5.4%
	Round	Overall	Coreference	Names	Family			
Ref. & Names	A1	24.5%	15.8%	12.5%	1.0%			
	A2	29.4%	22.7%	11.2%	1.7%			
	A3	27.5%	25.5%	1.9%	1.3%			
	All	27.1%	21.6%	8.1%	1.3%			
	Round	Overall	Syntactic	Prag.	Exhaustif.	Wordplay		
Tricky	A1	29.5%	14.5%	4.7%	5.5%	2.0%		
	A2	29.1%	8.0%	2.8%	8.6%	5.7%		
	A3	25.6%	9.3%	6.7%	4.8%	5.5%		
	All	27.9%	10.5%	4.8%	6.2%	4.5%		
	Round	Overall	Likely	Unlikely	Debatable	Facts	Containment	
Reasoning	A1	58.4%	25.7%	6.2%	3.1%	19.6%	11.0%	
	A2	62.7%	23.9%	6.9%	6.5%	25.6%	10.3%	
	A3	63.9%	22.7%	10.9%	10.8%	26.5%	5.3%	
	All	61.8%	24.0%	8.2%	7.0%	24.0%	8.7%	
	Round	Overall	Error	Ambiguous	EventCoref	Translation	Spelling	
Imperfections	A1	12.4%	3.3%	2.8%	0.9%	5.7%	5.8%	
	A2	13.5%	2.5%	4.0%	3.4%	6.2%	6.5%	
	A3	16.1%	2.2%	7.6%	1.9%	0.8%	5.5%	
	All	14.1%	2.6%	5.0%	2.1%	4.0%	5.9%	

Table 15: Percent examples in development set with particular tag, per round, on average.

BASIC Round	**Model**	**Basic**	Lexical	Comp.Sup.	ModusPonens	CauseEffect	Idiom	Negation	Coordination
A1	BERT (R1)	0.11 (0.56)	0.12 (0.59)	0.13 (0.66)	0.07 (0.31)	0.15 (0.55)	0.01 (0.45)	0.07 (0.40)	0.10 (0.52)
	RoBERTa Ensemble (R2)	0.69 (0.15)	0.73 (0.14)	0.63 (0.24)	0.43 (0.06)	**0.75** (0.02)	0.35 (0.12)	0.66 (0.17)	0.67 (0.13)
	RoBERTa Ensemble (R3)	0.72 (0.08)	0.78 (0.08)	0.72 (0.15)	0.32 (0.19)	**0.75** (0.01)	0.67 (0.02)	0.67 (0.06)	0.65 (0.08)
	RoBERTa-Large	0.76 (0.10)	**0.80** (0.12)	**0.82** (0.11)	**0.56** (0.08)	0.67 (0.20)	0.66 (0.02)	**0.71** (0.11)	0.74 (0.06)
	BART-Large	0.72 (**0.07**)	0.76 (**0.07**)	0.68 (0.08)	0.29 (**0.01**)	**0.75 (0.00)**	**0.67 (0.00)**	0.65 (**0.07**)	0.76 (0.11)
	XLNet-Large	0.75 (0.09)	0.78 (0.09)	0.77 (0.13)	0.23 (0.35)	**0.75 (0.00)**	0.66 (0.02)	0.64 (0.11)	0.76 (**0.03**)
	ELECTRA-Large	0.68 (0.34)	0.71 (0.34)	0.71 (0.23)	0.39 (0.42)	0.60 (0.20)	0.31 (0.66)	0.61 (0.33)	0.65 (0.43)
	ALBERT-XXLarge	**0.76** (0.20)	**0.80** (0.19)	0.78 (0.24)	0.31 (0.46)	0.64 (0.15)	**0.67** (0.02)	0.63 (0.21)	**0.77** (0.14)
A2	BERT (R1)	0.29 (0.44)	0.31 (0.46)	0.31 (0.56)	0.24 (0.31)	0.29 (0.40)	0.35 (0.44)	0.24 (0.41)	0.20 (0.38)
	RoBERTa Ensemble (R2)	0.20 (0.25)	0.24 (0.23)	0.19 (0.33)	0.33 (0.32)	0.21 (0.35)	0.19 (0.21)	0.17 (0.26)	0.15 (0.29)
	RoBERTa Ensemble (R3)	0.41 (0.14)	0.43 (0.15)	0.49 (0.16)	0.55 (0.18)	0.15 (0.17)	0.28 (0.10)	0.42 (**0.09**)	0.41 (0.21)
	RoBERTa-Large	0.47 (0.17)	0.47 (0.17)	0.49 (0.23)	0.99 (0.07)	0.30 (0.23)	0.37 (0.10)	0.55 (0.12)	0.48 (0.15)
	BART-Large	0.48 (0.14)	0.55 (0.14)	0.48 (0.18)	0.40 (**0.00**)	0.23 (**0.06**)	0.43 (0.21)	0.48 (0.16)	0.44 (**0.09**)
	XLNet-Large	0.53 (**0.13**)	0.54 (**0.13**)	0.51 (**0.13**)	**0.80** (0.02)	0.39 (0.17)	0.53 (**0.02**)	0.56 (0.18)	**0.51 (0.09)**
	ELECTRA-Large	0.52 (0.40)	0.54 (0.46)	0.46 (0.41)	0.47 (0.52)	0.38 (0.42)	0.53 (0.28)	0.56 (0.40)	0.56 (0.26)
	ALBERT-XXLarge	**0.58** (0.28)	**0.61** (0.28)	**0.53** (0.31)	**0.80** (0.04)	**0.48** (0.50)	**0.60** (0.31)	**0.64** (0.22)	0.50 (0.21)
A3	BERT (R1)	0.32 (0.50)	0.33 (0.51)	0.36 (0.59)	0.29 (0.72)	0.25 (0.57)	0.22 (0.47)	0.32 (0.46)	0.34 (0.50)
	RoBERTa Ensemble (R2)	0.26 (0.57)	0.26 (0.57)	0.29 (0.55)	0.25 (0.81)	0.16 (0.58)	0.24 (0.68)	0.25 (0.62)	0.26 (0.56)
	RoBERTa Ensemble (R3)	0.24 (0.53)	0.23 (0.53)	0.21 (0.53)	0.24 (0.57)	0.17 (0.51)	0.19 (0.57)	0.23 (0.57)	0.28 (0.50)
	RoBERTa-Large	0.45 (0.25)	0.44 (0.24)	0.46 (0.38)	0.45 (0.15)	0.39 (0.17)	0.42 (0.22)	0.46 (0.25)	0.49 (0.26)
	BART-Large	0.49 (**0.14**)	0.51 (0.16)	0.49 (**0.11**)	0.29 (**0.14**)	0.42 (**0.10**)	0.46 (**0.13**)	0.49 (**0.15**)	0.52 (0.13)
	XLNet-Large	0.49 (0.15)	0.50 (**0.12**)	0.47 (0.26)	0.34 (0.23)	0.40 (0.14)	0.44 (**0.13**)	0.46 (0.16)	**0.59 (0.08)**
	ELECTRA-Large	0.52 (0.44)	0.56 (0.43)	0.51 (0.50)	0.58 (0.46)	0.43 (0.36)	**0.64** (0.48)	0.52 (0.43)	0.54 (0.44)
	ALBERT-XXLarge	**0.55** (0.36)	**0.55** (0.35)	**0.56** (0.48)	**0.65** (0.33)	**0.48** (0.27)	0.52 (0.44)	**0.56** (0.36)	0.53 (0.33)
ANLI	BERT (R1)	0.26 (0.50)	0.27 (0.51)	0.27 (0.60)	0.21 (0.50)	0.25 (0.52)	0.26 (0.46)	0.26 (0.44)	0.23 (0.48)
	RoBERTa Ensemble (R2)	0.34 (0.37)	0.36 (0.37)	0.35 (0.37)	0.33 (0.46)	0.25 (0.45)	0.23 (0.47)	0.29 (0.44)	0.36 (0.36)
	RoBERTa Ensemble (R3)	0.41 (0.30)	0.42 (0.31)	0.46 (0.27)	0.34 (0.36)	0.23 (0.35)	0.25 (0.36)	0.36 (0.35)	0.43 (0.29)
	RoBERTa-Large	0.53 (0.19)	0.54 (0.19)	0.57 (0.24)	0.61 (0.11)	0.40 (0.19)	0.42 (0.16)	0.53 (0.19)	0.57 (0.17)
	BART-Large	0.54 (**0.13**)	0.58 (0.13)	0.54 (**0.13**)	0.31 (**0.06**)	0.41 (0.08)	0.46 (0.15)	0.51 (**0.14**)	0.57 (0.11)
	XLNet-Large	0.56 (**0.13**)	0.58 (**0.11**)	0.57 (0.17)	0.41 (0.22)	0.44 (0.13)	0.49 (**0.08**)	0.52 (0.16)	**0.62 (0.07)**
	ELECTRA-Large	0.56 (0.40)	0.59 (0.42)	0.54 (0.39)	0.49 (0.46)	0.44 (0.36)	**0.58** (0.42)	0.55 (0.40)	0.58 (0.39)
	ALBERT-XXLarge	**0.61** (0.30)	**0.63** (0.29)	**0.61** (0.34)	**0.58** (0.30)	**0.50** (0.32)	0.56 (0.37)	**0.60** (0.29)	0.60 (0.24)

Table 16: Correct label probability and entropy of label predictions for the BASIC subset: mean probability (mean entropy). BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58.

NUMERICAL Round	Model	Numerical	Cardinal	Ordinal	Counting	Nominal	Dates	Age
A1	BERT (R1)	0.10 (0.57)	0.10 (0.57)	0.11 (0.60)	0.09 (0.64)	0.07 (0.46)	0.10 (0.58)	0.07 (0.41)
	RoBERTa Ensemble (R2)	0.68 (0.13)	0.68 (0.13)	0.71 (0.18)	0.51 (0.23)	0.72 (0.11)	0.69 (0.13)	0.64 (0.11)
	RoBERTa Ensemble (R3)	0.72 (**0.07**)	0.72 (**0.07**)	**0.77** (**0.05**)	0.51 (0.23)	0.69 (0.06)	0.75 (**0.07**)	0.64 (**0.08**)
	RoBERTa-Large	0.73 (0.13)	0.73 (0.13)	0.75 (0.10)	0.58 (**0.10**)	0.76 (0.14)	0.74 (0.14)	0.65 (0.18)
	BART-Large	0.73 (0.10)	0.73 (0.10)	0.72 (0.11)	0.54 (0.12)	0.74 (**0.04**)	**0.77** (0.10)	0.67 (0.12)
	XLNet-Large	0.73 (0.10)	0.74 (0.10)	0.63 (0.08)	0.53 (0.15)	0.70 (0.11)	0.76 (0.09)	**0.71** (0.13)
	ELECTRA-Large	0.71 (0.29)	0.71 (0.28)	0.74 (0.35)	0.69 (0.42)	0.64 (0.23)	0.73 (0.27)	0.68 (0.38)
	ALBERT-XXLarge	**0.74** (0.22)	**0.75** (0.22)	0.72 (0.21)	0.56 (0.19)	**0.78** (0.19)	**0.77** (0.21)	**0.71** (0.32)
A2	BERT (R1)	0.29 (0.53)	0.28 (0.53)	0.33 (0.53)	0.43 (0.49)	0.31 (0.53)	0.25 (0.53)	0.18 (0.48)
	RoBERTa Ensemble (R2)	0.19 (0.28)	0.20 (0.28)	0.19 (0.24)	0.14 (0.30)	0.20 (0.34)	0.19 (0.26)	0.22 (0.25)
	RoBERTa Ensemble (R3)	0.50 (0.18)	0.51 (0.18)	0.50 (0.13)	0.36 (0.20)	0.44 (0.19)	0.55 (0.17)	0.51 (0.15)
	RoBERTa-Large	0.54 (0.22)	0.54 (0.22)	0.49 (0.21)	0.47 (0.17)	0.55 (**0.10**)	0.56 (0.24)	0.51 (0.26)
	BART-Large	0.55 (0.13)	0.54 (0.14)	0.57 (**0.10**)	0.56 (0.08)	0.47 (0.18)	0.56 (0.12)	0.50 (0.15)
	XLNet-Large	0.54 (**0.11**)	0.55 (**0.11**)	0.45 (0.14)	0.51 (0.06)	**0.54** (0.12)	0.57 (**0.11**)	0.54 (**0.14**)
	ELECTRA-Large	0.56 (0.36)	**0.57** (0.35)	0.55 (0.32)	0.52 (0.34)	0.49 (0.22)	**0.60** (0.36)	**0.59** (0.40)
	ALBERT-XXLarge	**0.57** (0.28)	**0.57** (0.28)	**0.60** (0.26)	**0.58** (0.26)	0.53 (0.20)	0.59 (0.30)	0.52 (0.34)
A3	BERT (R1)	0.34 (0.53)	0.34 (0.53)	0.43 (0.49)	0.34 (0.34)	0.41 (0.48)	0.31 (0.48)	0.28 (0.45)
	RoBERTa Ensemble (R2)	0.29 (0.47)	0.29 (0.46)	0.25 (0.47)	0.17 (0.48)	0.35 (0.41)	0.30 (0.34)	0.32 (0.36)
	RoBERTa Ensemble (R3)	0.20 (0.43)	0.20 (0.42)	0.25 (0.52)	0.11 (0.37)	0.20 (0.77)	0.22 (0.30)	0.26 (0.44)
	RoBERTa-Large	0.44 (0.32)	0.44 (0.32)	0.48 (0.29)	0.53 (0.15)	0.36 (0.53)	0.38 (0.33)	0.42 (0.37)
	BART-Large	0.51 (**0.14**)	0.52 (**0.14**)	0.48 (0.12)	0.51 (0.17)	0.59 (**0.07**)	0.51 (**0.10**)	0.46 (**0.14**)
	XLNet-Large	0.52 (0.15)	0.52 (0.16)	0.57 (**0.09**)	0.47 (**0.11**)	0.59 (**0.07**)	0.50 (0.17)	0.42 (0.16)
	ELECTRA-Large (tuned)	0.55 (0.46)	**0.56** (0.44)	0.52 (0.54)	0.58 (0.44)	**0.66** (0.53)	0.54 (0.43)	**0.57** (0.30)
	ALBERT-XXLarge	**0.56** (0.39)	**0.56** (0.38)	**0.61** (0.40)	**0.61** (0.32)	0.46 (0.43)	**0.58** (0.36)	0.56 (0.35)
A3	BERT (R1)	0.22 (0.54)	0.22 (0.55)	0.27 (0.54)	0.31 (0.48)	0.19 (0.49)	0.19 (0.54)	0.16 (0.45)
	RoBERTa Ensemble (R2)	0.41 (0.26)	0.41 (0.26)	0.40 (0.26)	0.25 (0.35)	0.48 (0.22)	0.44 (0.21)	0.39 (0.23)
	RoBERTa Ensemble (R3)	0.52 (0.20)	0.52 (0.19)	0.55 (0.18)	0.30 (0.27)	0.56 (0.16)	0.59 (0.14)	0.50 (0.19)
	RoBERTa-Large	0.59 (0.21)	0.59 (0.21)	0.59 (0.18)	0.52 (0.15)	0.65 (0.15)	0.62 (0.21)	0.54 (0.26)
	BART-Large	0.61 (**0.12**)	0.61 (**0.12**)	0.61 (0.11)	0.54 (0.12)	0.62 (**0.10**)	0.65 (**0.10**)	0.55 (**0.13**)
	XLNet-Large	0.61 (**0.12**)	0.62 (**0.12**)	0.54 (**0.10**)	0.50 (**0.10**)	0.62 (0.11)	0.65 (0.11)	0.57 (0.14)
	ELECTRA-Large	0.62 (0.35)	0.62 (0.34)	0.61 (0.38)	0.58 (0.40)	0.58 (0.25)	0.65 (0.33)	**0.62** (0.37)
	ALBERT-XXLarge	**0.64** (0.28)	**0.64** (0.28)	**0.65** (0.27)	**0.59** (0.26)	**0.66** (0.21)	**0.67** (0.27)	0.60 (0.33)

Table 17: Correct label probability and entropy of label predictions for the NUMERICAL subset: mean probability (mean entropy). BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58.

REASONING Round	Model	Reasoning	Likely	Unlikely	Debatable	Facts	Containment
A1	BERT (R1)	0.13 (0.60)	0.14 (0.57)	0.15 (0.54)	0.16 (0.52)	0.11 (0.64)	0.11 (0.62)
	RoBERTa Ensemble (R2)	0.67 (0.13)	0.64 (0.16)	0.78 (0.13)	0.61 (0.05)	0.65 (0.12)	0.71 (0.14)
	RoBERTa Ensemble (R3)	0.73 (**0.08**)	0.72 (0.09)	0.78 (**0.04**)	0.68 (**0.00**)	0.71 (**0.08**)	0.75 (0.11)
	RoBERTa-Large	0.75 (0.12)	0.74 (0.15)	0.82 (0.09)	0.67 (0.06)	**0.74** (0.15)	0.75 (0.09)
	BART-Large	0.76 (**0.08**)	**0.78** (**0.07**)	0.86 (0.06)	**0.70** (0.04)	0.71 (0.09)	0.72 (0.12)
	XLNet-Large	0.74 (0.09)	0.74 (0.08)	0.82 (0.07)	**0.70** (0.09)	0.72 (0.12)	0.74 (**0.08**)
	ELECTRA-Large	0.66 (0.36)	0.68 (0.35)	0.77 (0.24)	0.66 (0.42)	0.63 (0.37)	0.58 (0.47)
	ALBERT-XXLarge	**0.77** (0.18)	0.77 (0.17)	**0.88** (0.09)	0.67 (0.21)	0.73 (0.21)	**0.76** (0.20)
A2	BERT (R1)	0.30 (0.47)	0.34 (0.44)	0.31 (0.42)	0.36 (0.44)	0.23 (0.49)	0.33 (0.54)
	RoBERTa Ensemble (R2)	0.21 (0.26)	0.27 (0.28)	0.21 (0.33)	0.16 (0.27)	0.18 (0.22)	0.17 (0.19)
	RoBERTa Ensemble (R3)	0.43 (0.16)	0.43 (0.14)	0.45 (0.18)	0.43 (0.16)	0.40 (0.13)	0.38 (0.17)
	RoBERTa-Large	0.51 (0.21)	0.48 (0.19)	0.56 (0.20)	0.43 (0.21)	0.49 (0.23)	0.49 (0.22)
	BART-Large	0.52 (0.13)	0.61 (**0.12**)	0.53 (0.13)	0.48 (0.13)	0.43 (0.14)	0.48 (0.17)
	XLNet-Large	0.53 (**0.12**)	0.57 (0.13)	0.56 (**0.12**)	0.49 (**0.05**)	0.48 (**0.11**)	0.49 (**0.11**)
	ELECTRA-Large	0.53 (0.40)	0.58 (0.39)	0.54 (0.38)	0.52 (0.39)	0.49 (0.39)	0.51 (0.42)
	ALBERT-XXLarge	**0.57** (0.29)	**0.62** (0.27)	**0.65** (0.30)	**0.55** (0.25)	**0.50** (0.30)	**0.53** (0.29)
A3	BERT (R1)	0.34 (0.51)	0.37 (0.47)	0.38 (0.48)	0.35 (0.51)	0.29 (0.54)	0.35 (0.46)
	RoBERTa Ensemble (R2)	0.26 (0.54)	0.25 (0.51)	0.28 (0.58)	0.25 (0.62)	0.25 (0.51)	0.28 (0.38)
	RoBERTa Ensemble (R3)	0.23 (0.50)	0.23 (0.47)	0.25 (0.52)	0.21 (0.56)	0.22 (0.48)	0.20 (0.38)
	RoBERTa-Large	0.44 (0.26)	0.44 (0.25)	0.51 (0.25)	0.47 (0.24)	0.40 (0.27)	0.50 (0.32)
	BART-Large	0.50 (**0.14**)	0.52 (0.14)	0.57 (**0.13**)	0.47 (**0.15**)	0.44 (**0.14**)	0.58 (0.16)
	XLNet-Large	0.49 (**0.14**)	0.47 (**0.13**)	0.56 (0.14)	0.50 (0.16)	0.47 (0.15)	0.51 (**0.13**)
	ELECTRA-Large	0.51 (0.45)	0.49 (0.48)	0.56 (0.39)	0.49 (0.49)	0.48 (0.44)	0.51 (0.48)
	ALBERT-XXLarge	**0.57** (0.33)	**0.59** (0.33)	**0.65** (0.32)	**0.58** (0.37)	**0.50** (0.33)	**0.55** (0.23)
ANLI	BERT (R1)	0.26 (0.52)	0.29 (0.49)	0.31 (0.48)	0.33 (0.49)	0.23 (0.55)	0.25 (0.56)
	RoBERTa Ensemble (R2)	0.37 (0.33)	0.39 (0.32)	0.38 (0.41)	0.28 (0.44)	0.33 (0.32)	0.41 (0.21)
	RoBERTa Ensemble (R3)	0.44 (0.27)	0.46 (0.24)	0.43 (0.32)	0.34 (0.37)	0.41 (0.26)	0.48 (0.19)
	RoBERTa-Large	0.55 (0.20)	0.55 (0.19)	0.60 (0.20)	0.49 (0.21)	0.52 (0.22)	0.60 (0.19)
	BART-Large	0.58 (**0.12**)	0.63 (**0.11**)	0.63 (**0.12**)	0.51 (0.13)	0.50 (**0.13**)	0.60 (0.15)
	XLNet-Large	0.58 (**0.12**)	0.59 (0.12)	0.62 (**0.12**)	0.52 (0.12)	0.54 (**0.13**)	0.59 (**0.10**)
	ELECTRA-Large	0.56 (0.40)	0.58 (0.41)	0.60 (0.35)	0.52 (0.45)	0.52 (0.41)	0.54 (0.46)
	ALBERT-XXLarge	**0.63** (0.27)	**0.66** (0.26)	**0.70** (0.26)	**0.58** (0.31)	**0.56** (0.29)	**0.63** (0.24)

Table 18: Correct label probability and entropy of label predictions for the REASONING subset: mean probability (mean entropy). BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58.

REFERENCE Round	Model	**Reference**	Coreference	Names	Family
A1	BERT (R1)	0.12 (0.59)	0.11 (0.56)	0.12 (0.60)	0.12 (0.56)
	RoBERTa Ensemble (R2)	0.66 (0.15)	0.67 (0.15)	0.68 (0.15)	0.29 (0.19)
	RoBERTa Ensemble (R3)	0.70 (0.08)	0.70 (**0.08**)	0.75 (0.06)	0.44 (0.17)
	RoBERTa-Large	0.75 (0.15)	0.76 (0.15)	0.77 (0.15)	0.52 (0.26)
	BART-Large	0.70 (0.11)	0.73 (0.13)	0.73 (**0.09**)	0.54 (0.10)
	XLNet-Large	0.72 (**0.09**)	0.74 (**0.08**)	0.75 (**0.09**)	0.62 (**0.09**)
	ELECTRA-Large	0.63 (0.41)	0.64 (0.41)	0.66 (0.40)	0.61 (0.35)
	ALBERT-XXLarge	**0.77** (0.18)	**0.78** (0.18)	**0.80** (0.17)	**0.67** (0.12)
A2	BERT (R1)	0.31 (0.47)	0.29 (0.47)	0.33 (0.48)	0.34 (0.41)
	RoBERTa Ensemble (R2)	0.19 (0.24)	0.20 (0.24)	0.16 (0.24)	0.18 (0.24)
	RoBERTa Ensemble (R3)	0.45 (0.14)	0.46 (0.16)	0.42 (0.14)	0.45 (0.17)
	RoBERTa-Large	0.49 (0.20)	0.53 (0.20)	0.42 (0.19)	0.44 (0.16)
	BART-Large	0.50 (0.13)	0.52 (0.13)	0.41 (0.13)	0.40 (**0.14**)
	XLNet-Large	0.50 (**0.10**)	0.52 (**0.10**)	0.43 (**0.08**)	0.48 (0.19)
	ELECTRA-Large	0.53 (0.38)	0.55 (0.39)	0.48 (0.39)	0.38 (0.47)
	ALBERT-XXLarge	**0.56** (0.25)	**0.58** (0.28)	**0.49** (0.22)	**0.58** (0.17)
A3	BERT (R1)	0.32 (0.49)	0.33 (0.48)	0.27 (0.51)	0.25 (0.59)
	RoBERTa Ensemble (R2)	0.27 (0.55)	0.27 (0.53)	0.26 (0.76)	0.39 (0.39)
	RoBERTa Ensemble (R3)	0.25 (0.54)	0.24 (0.54)	0.26 (0.46)	0.47 (0.41)
	RoBERTa-Large	0.46 (0.27)	0.46 (0.27)	0.46 (0.38)	0.47 (0.22)
	BART-Large	0.50 (**0.14**)	0.49 (**0.13**)	0.67 (0.17)	**0.62** (0.23)
	XLNet-Large	0.52 (0.15)	0.50 (0.16)	**0.70** (**0.15**)	0.61 (**0.09**)
	ELECTRA-Large	0.52 (0.48)	0.51 (0.48)	0.66 (0.42)	0.51 (0.45)
	ALBERT-XXLarge	**0.54** (0.32)	**0.53** (0.32)	0.66 (0.31)	**0.62** (0.27)
ANLI	BERT (R1)	0.26 (0.51)	0.27 (0.49)	0.22 (0.54)	0.25 (0.51)
	RoBERTa Ensemble (R2)	0.35 (0.33)	0.34 (0.35)	0.42 (0.24)	0.29 (0.28)
	RoBERTa Ensemble (R3)	0.45 (0.28)	0.42 (0.31)	0.56 (0.13)	0.46 (0.26)
	RoBERTa-Large	0.56 (0.21)	0.55 (0.22)	0.59 (0.19)	0.47 (0.20)
	BART-Large	0.55 (0.13)	0.55 (0.13)	0.59 (**0.12**)	0.51 (0.16)
	XLNet-Large	0.57 (**0.12**)	0.56 (**0.12**)	0.61 (0.09)	0.56 (**0.13**)
	ELECTRA-Large	0.55 (0.43)	0.55 (0.43)	0.59 (0.40)	0.48 (0.43)
	ALBERT-XXLarge	**0.61** (0.25)	**0.60** (0.27)	**0.65** (0.20)	**0.62** (0.20)

Table 19: Correct label probability and entropy of label predictions for the REFERENCE subset: mean probability (mean entropy). BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58.

TRICKY Round	Model	Tricky	Syntactic	Pragmatic	Exhaustification	Wordplay
A1	BERT (R1)	0.10 (0.56)	0.10 (0.54)	0.09 (0.56)	0.11 (0.56)	0.13 (0.72)
	RoBERTa Ensemble (R2)	0.60 (0.18)	0.60 (0.17)	0.60 (0.23)	0.59 (0.17)	0.52 (0.15)
	RoBERTa Ensemble (R3)	0.65 (0.09)	0.67 (**0.09**)	**0.72** (0.08)	0.54 (0.11)	0.51 (0.06)
	RoBERTa-Large	**0.70** (0.14)	0.72 (0.15)	0.68 (0.10)	**0.64** (0.13)	0.65 (0.15)
	BART-Large	**0.70 (0.08)**	**0.73 (0.09)**	0.64 (0.07)	0.62 (**0.08**)	0.75 (**0.02**)
	XLNet-Large	**0.70** (0.10)	**0.73** (0.11)	0.66 (**0.06**)	0.56 (0.10)	**0.78** (0.15)
	ELECTRA-Large	0.62 (0.44)	0.62 (0.49)	0.62 (0.40)	0.56 (0.41)	0.60 (0.45)
	ALBERT-XXLarge	0.65 (0.21)	0.66 (0.19)	0.61 (0.17)	0.58 (0.25)	0.63 (0.23)
A2	BERT (R1)	0.25 (0.48)	0.22 (0.53)	0.20 (0.35)	0.29 (0.47)	0.21 (0.47)
	RoBERTa Ensemble (R2)	0.16 (0.23)	0.19 (0.25)	0.10 (0.13)	0.20 (0.21)	0.09 (0.30)
	RoBERTa Ensemble (R3)	0.44 (0.14)	0.40 (**0.13**)	0.33 (**0.10**)	0.37 (0.16)	0.59 (0.14)
	RoBERTa-Large	0.48 (0.22)	**0.49** (0.20)	0.33 (0.21)	0.40 (0.25)	0.59 (0.16)
	BART-Large	0.48 (0.15)	0.46 (0.14)	0.26 (0.14)	0.45 (0.15)	0.58 (0.13)
	XLNet-Large	**0.52 (0.12)**	0.48 (**0.13**)	0.39 (0.14)	0.50 (**0.14**)	**0.60 (0.07)**
	ELECTRA-Large	0.51 (0.45)	**0.49** (0.52)	0.39 (0.44)	0.47 (0.41)	0.57 (0.45)
	ALBERT-XXLarge	0.50 (0.26)	0.44 (0.25)	**0.40** (0.28)	**0.51** (0.29)	0.42 (0.24)
A3	BERT (R1)	0.29 (0.55)	0.29 (0.50)	0.29 (0.64)	0.28 (0.48)	0.25 (0.58)
	RoBERTa Ensemble (R2)	0.24 (0.58)	0.26 (0.51)	0.24 (0.62)	0.18 (0.53)	0.24 (0.72)
	RoBERTa Ensemble (R3)	0.25 (0.54)	0.29 (0.53)	0.20 (0.57)	0.23 (0.58)	0.24 (0.50)
	RoBERTa-Large	0.49 (0.25)	0.47 (0.28)	0.41 (0.19)	0.46 (0.24)	0.63 (0.25)
	BART-Large	0.53 (0.18)	0.51 (0.17)	0.43 (0.21)	0.46 (0.18)	**0.72** (0.20)
	XLNet-Large	0.51 (**0.14**)	**0.57 (0.14)**	**0.46 (0.13)**	0.36 (**0.11**)	0.57 (**0.16**)
	ELECTRA-Large	**0.54** (0.44)	0.53 (0.44)	0.41 (0.50)	**0.49** (0.45)	**0.72** (0.41)
	ALBERT-XXLarge	0.52 (0.32)	0.53 (0.36)	**0.46** (0.30)	0.44 (0.31)	0.62 (0.32)
ANLI	BERT (R1)	0.21 (0.53)	0.19 (0.52)	0.22 (0.56)	0.24 (0.50)	0.22 (0.55)
	RoBERTa Ensemble (R2)	0.33 (0.34)	0.39 (0.30)	0.32 (0.41)	0.30 (0.29)	0.22 (0.47)
	RoBERTa Ensemble (R3)	0.45 (0.26)	0.48 (0.24)	0.38 (0.34)	0.38 (0.27)	0.42 (0.29)
	RoBERTa-Large	0.56 (0.20)	0.58 (0.21)	0.48 (0.17)	0.48 (0.21)	0.62 (0.20)
	BART-Large	**0.57** (0.14)	0.59 (0.13)	0.46 (0.15)	0.50 (0.14)	**0.67** (0.15)
	XLNet-Large	**0.57 (0.12)**	**0.62 (0.12)**	**0.51 (0.11)**	0.48 (**0.12**)	0.61 (**0.12**)
	ELECTRA-Large	0.56 (0.44)	0.56 (0.48)	0.47 (0.46)	0.50 (0.42)	0.65 (0.43)
	ALBERT-XXLarge	0.56 (0.26)	0.57 (0.26)	0.49 (0.26)	**0.51** (0.28)	0.54 (0.28)

Table 20: Correct label probability and entropy of label predictions for the TRICKY subset: mean probability (mean entropy). BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58.

IMPERFECTIONS Round	Model	Imperfections	Errors	Ambiguity	EventCoref	Translation	Spelling
A1	BERT (R1)	0.13 (0.57)	0.07 (0.38)	0.17 (0.73)	0.12 (0.77)	0.11 (0.59)	0.14 (0.64)
	RoBERTa Ensemble (R2)	0.61 (0.14)	0.38 (0.11)	0.53 (0.19)	0.82 (0.25)	0.67 (0.17)	0.77 (0.12)
	RoBERTa Ensemble (R3)	0.68 (**0.07**)	0.49 (0.12)	0.57 (**0.02**)	**0.89** (**0.00**)	0.71 (0.06)	0.81 (0.07)
	RoBERTa-Large	0.68 (0.13)	0.46 (0.15)	0.65 (0.17)	0.88 (0.04)	**0.75** (0.13)	0.79 (0.14)
	BART-Large	0.71 (0.08)	**0.52** (**0.06**)	**0.73** (0.10)	0.78 (**0.00**)	0.74 (**0.11**)	0.79 (0.11)
	XLNet-Large	0.67 (0.08)	0.49 (0.11)	0.58 (0.18)	0.81 (0.24)	0.72 (0.06)	**0.81** (**0.06**)
	ELECTRA-Large	0.63 (0.40)	0.49 (0.43)	0.65 (0.51)	0.58 (0.50)	0.73 (0.37)	0.70 (0.35)
	ALBERT-XXLarge	**0.69** (0.22)	0.48 (0.24)	0.62 (0.27)	0.78 (0.19)	0.71 (0.19)	0.78 (0.21)
A2	BERT (R1)	0.33 (0.48)	0.42 (0.39)	0.32 (0.47)	0.27 (0.43)	0.29 (0.51)	0.34 (0.45)
	RoBERTa Ensemble (R2)	0.19 (0.27)	0.22 (0.22)	0.19 (0.23)	0.21 (0.33)	0.16 (0.23)	0.21 (0.28)
	RoBERTa Ensemble (R3)	0.33 (0.14)	0.34 (0.17)	0.43 (0.11)	0.40 (**0.11**)	0.46 (0.13)	0.32 (0.12)
	RoBERTa-Large	0.49 (0.19)	0.38 (0.26)	0.50 (0.16)	0.50 (0.20)	0.48 (0.27)	0.56 (0.18)
	BART-Large	0.42 (**0.10**)	0.29 (0.08)	0.48 (**0.10**)	**0.58** (0.12)	0.48 (0.13)	0.45 (0.11)
	XLNet-Large	0.44 (**0.10**)	0.36 (**0.03**)	0.48 (**0.10**)	0.54 (0.13)	0.55 (**0.10**)	0.45 (**0.10**)
	ELECTRA-Large	0.54 (0.39)	0.40 (0.24)	0.63 (0.33)	0.55 (0.38)	0.56 (0.44)	**0.60** (0.46)
	ALBERT-XXLarge	**0.58** (0.32)	**0.60** (0.42)	**0.69** (0.26)	0.54 (0.23)	**0.66** (0.26)	0.54 (0.30)
A3	BERT (R1)	0.31 (0.54)	0.30 (0.57)	0.28 (0.58)	0.24 (0.29)	0.42 (0.76)	0.36 (0.52)
	RoBERTa Ensemble (R2)	0.23 (0.58)	0.22 (0.65)	0.23 (0.58)	0.36 (0.52)	0.26 (0.21)	0.19 (0.46)
	RoBERTa Ensemble (R3)	0.23 (0.52)	0.23 (0.55)	0.17 (0.52)	0.32 (0.48)	0.16 (0.26)	0.22 (0.46)
	RoBERTa-Large	0.40 (0.23)	0.32 (0.14)	0.35 (0.19)	0.70 (0.18)	**0.56** (0.13)	0.39 (0.24)
	BART-Large	0.48 (0.17)	0.37 (0.13)	0.39 (0.17)	0.63 (0.26)	0.30 (**0.03**)	**0.53** (0.15)
	XLNet-Large	0.43 (**0.14**)	**0.41** (**0.10**)	0.40 (**0.15**)	0.64 (**0.13**)	0.52 (0.19)	0.40 (**0.12**)
	ELECTRA-Large	0.47 (0.49)	0.32 (0.42)	0.43 (0.53)	0.63 (0.37)	0.33 (0.40)	0.48 (0.46)
	ALBERT-XXLarge	**0.52** (0.33)	0.39 (0.31)	**0.50** (0.36)	**0.68** (0.32)	0.47 (0.49)	0.49 (0.28)
ANLI	BERT (R1)	0.27 (0.53)	0.24 (0.44)	0.27 (0.58)	0.24 (0.43)	0.22 (0.57)	0.28 (0.53)
	RoBERTa Ensemble (R2)	0.32 (0.37)	0.28 (0.31)	0.27 (0.42)	0.35 (0.39)	0.39 (0.20)	0.38 (0.29)
	RoBERTa Ensemble (R3)	0.39 (0.28)	0.36 (0.27)	0.31 (0.33)	0.44 (0.22)	0.55 (0.11)	0.44 (0.22)
	RoBERTa-Large	0.50 (0.19)	0.39 (0.18)	0.44 (0.18)	0.62 (0.17)	0.60 (0.20)	0.57 (0.19)
	BART-Large	0.52 (0.12)	0.41 (**0.09**)	0.47 (0.14)	0.62 (0.15)	0.58 (0.11)	0.58 (0.12)
	XLNet-Large	0.50 (**0.11**)	0.42 (**0.09**)	0.45 (**0.14**)	0.61 (**0.14**)	0.62 (**0.09**)	0.55 (**0.10**)
	ELECTRA-Large	0.54 (0.44)	0.41 (0.37)	0.52 (0.48)	0.58 (0.40)	0.62 (0.41)	0.59 (0.42)
	ALBERT-XXLarge	**0.59** (0.30)	**0.49** (0.32)	**0.57** (0.32)	**0.62** (0.26)	**0.67** (0.25)	**0.60** (0.27)

Table 21: Correct label probability and entropy of label predictions for the IMPERFECTIONS subset: mean probability (mean entropy). BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58. A3 had no examples of TRANSLATION, so no numbers can be reported.

Genre	Model	Numerical	Basic	Reference	Tricky	Reasoning	Imperfections
Wikipedia	BERT (R1)	0.20 (0.55)	0.23 (0.49)	0.24 (0.51)	0.18 (0.52)	0.23 (0.53)	0.24 (0.52)
	RoBERTa Ensemble (R2)	0.43 (0.21)	0.40 (0.21)	0.40 (0.21)	0.37 (0.22)	0.42 (0.21)	0.37 (0.21)
	RoBERTa Ensemble (R3)	0.58 (0.13)	0.51 (0.12)	0.54 (0.12)	0.52 (0.12)	0.53 (0.13)	0.46 (0.12)
	RoBERTa-Large	0.61 (0.18)	0.57 (0.15)	0.59 (0.18)	0.58 (0.18)	0.60 (0.18)	0.55 (0.18)
	BART-Large	0.63 (**0.11**)	0.57 (**0.11**)	0.58 (0.12)	0.58 (0.12)	0.62 (**0.11**)	0.54 (0.10)
	XLNet-Large	0.62 (**0.11**)	0.61 (0.12)	0.59 (**0.09**)	**0.60** (**0.11**)	0.62 (**0.11**)	0.53 (**0.09**)
	ELECTRA-Large	0.62 (0.33)	0.57 (0.38)	0.57 (0.40)	0.56 (0.44)	0.58 (0.38)	0.57 (0.40)
	ALBERT-XXLarge	**0.65** (0.25)	**0.64** (0.25)	**0.65** (0.22)	0.56 (0.24)	**0.66** (0.24)	**0.63** (0.27)
Fiction	BERT (R1)	0.49 (0.35)	0.28 (0.54)	0.29 (0.52)	0.35 (0.60)	0.29 (0.51)	0.30 (0.62)
	RoBERTa Ensemble (R2)	0.32 (0.73)	0.25 (0.68)	0.26 (0.70)	0.24 (0.71)	0.26 (0.63)	0.24 (0.73)
	RoBERTa Ensemble (R3)	0.35 (0.55)	0.26 (0.70)	0.29 (0.73)	0.26 (0.72)	0.27 (0.64)	0.28 (0.73)
	RoBERTa-Large	0.41 (0.14)	0.46 (0.22)	0.45 (0.26)	0.56 (0.16)	0.45 (0.24)	0.35 (0.15)
	BART-Large	0.14 (0.06)	0.49 (0.17)	0.46 (0.14)	**0.59** (0.12)	0.48 (0.14)	0.47 (0.14)
	XLNet-Large	0.57 (**0.01**)	0.49 (**0.08**)	0.50 (**0.10**)	0.52 (**0.09**)	0.52 (**0.10**)	0.40 (**0.04**)
	ELECTRA-Large	0.23 (0.28)	0.54 (0.36)	**0.56** (0.45)	**0.59** (0.36)	0.51 (0.38)	0.47 (0.45)
	ALBERT-XXLarge	**0.65** (0.27)	**0.55** (0.27)	0.50 (0.23)	0.52 (0.26)	**0.61** (0.28)	**0.62** (0.34)
News	BERT (R1)	0.38 (0.47)	0.32 (0.53)	0.26 (0.48)	0.25 (0.61)	0.40 (0.49)	0.39 (0.46)
	RoBERTa Ensemble (R2)	0.23 (0.40)	0.24 (0.43)	0.16 (0.32)	0.23 (0.49)	0.26 (0.41)	0.14 (0.64)
	RoBERTa Ensemble (R3)	0.19 (0.30)	0.22 (0.37)	0.21 (0.34)	0.26 (0.40)	0.22 (0.39)	0.23 (0.41)
	RoBERTa-Large	0.43 (0.31)	0.46 (0.22)	0.41 (**0.14**)	0.49 (0.15)	0.47 (0.23)	0.50 (0.23)
	BART-Large	0.56 (0.16)	0.49 (0.14)	0.41 (0.18)	0.63 (0.17)	0.54 (0.15)	**0.66** (0.20)
	XLNet-Large	0.56 (**0.14**)	0.51 (**0.13**)	**0.55** (0.18)	0.52 (**0.12**)	0.49 (**0.14**)	0.48 (**0.17**)
	ELECTRA-Large	**0.68** (0.39)	0.53 (0.39)	0.45 (0.33)	0.57 (0.35)	0.48 (0.40)	0.53 (0.45)
	ALBERT-XXLarge	0.67 (0.32)	**0.56** (0.22)	0.52 (0.23)	**0.64** (0.19)	**0.55** (0.24)	0.60 (0.26)
Procedural	BERT (R1)	0.37 (0.43)	0.30 (0.57)	0.38 (0.48)	0.19 (0.46)	0.34 (0.56)	0.30 (0.58)
	RoBERTa Ensemble (R2)	0.28 (0.65)	0.24 (0.67)	0.22 (0.69)	0.21 (0.70)	0.26 (0.70)	0.23 (0.60)
	RoBERTa Ensemble (R3)	0.21 (0.63)	0.24 (0.59)	0.21 (0.68)	0.27 (0.64)	0.25 (0.63)	0.25 (0.51)
	RoBERTa-Large	0.58 (0.23)	0.50 (0.13)	0.65 (0.25)	**0.57** (0.25)	0.45 (0.20)	0.45 (**0.07**)
	BART-Large	0.53 (**0.08**)	0.47 (**0.07**)	0.49 (**0.19**)	0.41 (**0.16**)	0.47 (**0.10**)	0.52 (0.09)
	XLNet-Large	0.57 (0.10)	0.53 (0.14)	0.66 (0.21)	0.53 (0.17)	0.49 (0.15)	**0.57** (0.18)
	ELECTRA-Large	**0.67** (0.35)	0.58 (0.43)	0.58 (0.44)	0.55 (0.41)	**0.58** (0.44)	0.42 (0.52)
	ALBERT-XXLarge	0.66 (0.26)	**0.61** (0.32)	**0.71** (0.29)	**0.57** (0.29)	0.56 (0.31)	0.53 (0.26)

Table 22: Probability of the correct label (entropy of label predictions) for each model on each top level annotation tag. BERT (R1) has zero accuracy, by construction, on A1 because it was used to collect A1, whereas RoBERTas (R2) and (R3) were part of an ensemble of several identical architectures with different random seeds, so they have low, but non-zero, accuracy on their respective rounds. Recall that the entropy for three equiprobable outcomes (i.e., random chance of three NLI labels) is upper bounded by ≈ 1.58.

Subset	Context	Statement	Rationale	Context+Statement
ANLI	film (647), american (588), **known** (377), first (376), (born (365), **also** (355), one (342), new (341), released (296), album (275), **united** (249), **directed** (240), not (236), – (218), **based** (214), **series** (196), **best** (191), may (188), **band** (185), **state** (182), **football** (177), two (175), written (175), **television** (175), **national** (169), **south** (165)	not (252), born (132), years (120), released (107), one (87), film (83), first (82), only (76), **people** (75), year (61), **played** (58), new (58), two (54), **made** (54), album (49), no (46), **died** (46), **won** (46), **less** (44), **last** (42), american (41), **years.** (40), **three** (40), written (38), **used** (37), **john** (37)	not (1306), **system** (753), **statement** (494), **know** (343), **think** (274), **definitely** (268), **context** (261), **correct** (243), **difficult** (228), only (224), **doesn't** (223), may (221), **confused** (218), no (200), **says** (198), **incorrect** (193), **text** (184), **could** (181), **states** (166), born (160), one (155), **say** (147), years (146), **don't** (140), **would** (130), **whether** (129)	film (730), american (629), not (488), first (458), one (429), known (414), released (403), new (399), also (379), (born (368), album (324), united (281), directed (274), based (238), two (229), born (226), series (223), played (221), – (221), best (220), band (219), only (213), written (213), football (208), may (208), state (204)
R1	film (299), american (272), known (175), (born (169), first (158), also (129), released (119), album (115), directed (106), based (104), united (103), new (97), – (93), football (88), one (84), band (77), best (77), south (73), **former** (71), written (70), series (67), played (67), **march** (66), city (65), located (65), television (64)	born (65), film (47), not (46), years (45), released (43), first (36), died (26), only (25), american (24), **population** (23), **old** (23), album (22), won (22), played (21), directed (21), new (19), last (18), football (18), **century.** (18), year (18), united (17), years. (16), **world** (16), written (16), one (16), based (16)	not (392), system (331), know (135), statement (126), think (111), context (105), difficult (93), definitely (86), correct (80), born (80), only (75), may (75), confused (75), incorrect (63), could (62), stated (62), don't (59), says (58), doesn't (57), **information** (54), states (53), no (53), first (52), **probably** (49), used (48), text (47)	film (346), american (296), first (194), known (188), (born (170), released (162), also (140), album (137), directed (127), united (120), based (120), new (116), born (109), football (106), one (100), – (94), band (91), best (89), played (88), written (86), south (81), world (79), city (77), series (77), population (77), name (77)
R2	film (301), american (266), known (166), (born (159), also (146), released (136), new (128), album (127), first (126), directed (114), one (112), series (110), united (97), – (95), television (95), band (87), state (86), based (83), written (82), **song** (79), national (76), played (74), best (69), located (67), city (66), football (66)	not (75), years (54), released (53), born (51), one (32), first (32), film (31), year (29), **ago.** (24), only (24), played (23), album (23), known (22), two (22), new (21), band (19), made (18), city (16), no (16), died (16), john (15), less (15), won (15), written (14), people (14), **lived** (14)	not (387), system (198), statement (125), know (93), doesn't (79), difficult (78), think (77), years (74), context (72), confused (70), may (65), only (63), born (61), states (60), correct (59), no (56), **ai** (55), definitely (55), released (52), text (50), incorrect (49), say (48), year (48), could (45), one (44), says (42)	film (332), american (280), released (189), known (188), (born (161), also (159), first (158), album (150), new (149), one (144), series (124), directed (123), band (106), united (105), television (101), not (98), played (97), – (97), written (96), state (96), song (89), born (88), based (87), national (83), city (82), located (80)
R3	not (197), one (146), said (122), new (116), would (104), first (92), some (91), make (87), people (83), may (83), also (80), **time** (77), no (75), – (75), **like** (74), **get** (74), last (72), only (68), two (68), **police** (66), made (61), think (55), **home** (54), **go** (54), **way** (53), **many** (53)	not (131), people (48), one (39), only (27), no (22), made (21), years (21), speaker (19), two (19), new (18), three (17), used (16), **use** (16), **person** (16), less (16), born (16), **good** (15), make (14), year (14), first (14), played (14), **school** (13), **government** (13), didn't (13), last (13), some (13)	not (527), statement (243), system (224), definitely (127), know (115), correct (104), says (98), no (91), doesn't (87), text (87), think (86), only (86), context (84), incorrect (81), may (81), **model** (75), could (74), confused (73), one (67), said (66), say (63), whether (58), difficult (57), neither (57), incorrect. (56), would (53)	not (328), one (185), new (134), people (131), said (127), would (115), first (106), some (104), make (101), no (97), may (95), only (95), two (87), time (86), last (85), like (83), get (82), made (82), also (80), – (75), **police** (74), use (67), many (66), three (63), **home** (62), **go** (62)
Contra.	american (219), film (216), new (146), (born (129), first (124), also (116), known (115), united (110), one (108), released (94), album (86), – (81), directed (78), series (76), may (72), best (71), television (70), band (69), not (68), based (66), written (65), south (65), national (63), two (62), **song** (60), football (59)	not (63), years (55), born (42), film (37), released (36), first (31), year (30), only (28), one (23), new (23), died (21), people (19), american (19), won (19), years. (19), **world** (18), three (18), played (18), album (17), two (17), less (17), directed (17), **old** (16), made (16), written (15), **lived** (15)	not (471), system (269), statement (174), incorrect (121), think (104), definitely (90), confused (87), difficult (83), only (78), born (71), says (63), context (61), years (57), states (51), one (50), would (49), incorrect. (47), know (42), name (42), **probably** (41), year (41), **ai** (41), could (40), first (38), may (38), model (35)	film (253), american (238), new (169), first (155), not (131), one (131), (born (130), released (130), known (126), also (125), united (119), album (103), directed (95), series (88), – (83), band (82), written (80), two (79), best (79), may (78), television (78), south (77), world (75), based (74), years (74), football (72)
Neut.	film (224), american (198), known (126), first (118), one (116), released (115), (born (112), also (107), album (101), new (97), not (95), directed (93), based (77), united (74), football (67), may (61), band (60), best (60), – (58), city (55), two (55), national (54), played (54), series (53), state (51), **song** (51)	not (63), one (37), born (36), released (29), only (28), never (25), played (24), film (22), people (21), made (19), first (18), no (18), new (17), album (17), won (17), known (16), population (15), john (14), two (14), last (14), name (13), united (13), died (12), best (12), football (11), written (11)	not (608), know (263), system (236), doesn't (157), no (150), context (147), statement (146), may (133), say (125), whether (124), correct (123), could (119), neither (117), don't (117), only (110), definitely (109), text (102), information (89), **nor** (83), mentioned (80), think (80), state (78), says (71), difficult (71), incorrect (69), confused (67)	film (246), american (208), not (158), one (153), released (144), known (142), first (136), album (118), new (114), (born (114), also (112), directed (101), united (87), based (83), played (78), football (78), only (76), best (72), band (70), two (69), made (69), city (66), may (64), born (63), name (63), written (60)
Entail.	film (207), american (171), known (136), first (134), also (132), (born (124), one (118), new (98), album (88), released (87), – (79), state (73), not (73), based (71), directed (69), series (67), united (65), played (61), written (61), best (60), television (60), **former** (60), two (58), band (56), may (55), **located** (53)	not (63), one (37), born (36), released (29), only (28), never (25), played (24), film (22), people (21), made (19), first (18), no (18), new (17), album (17), won (17), known (16), population (15), john (14), two (14), last (14), name (13), united (13), died (12), best (12), football (11), written (11)	not (608), know (263), system (236), doesn't (157), no (150), context (147), statement (146), may (133), say (125), whether (124), correct (123), could (119), neither (117), don't (117), only (110), definitely (109), text (102), information (89), nor (83), mentioned (80), think (80), state (78), says (71), difficult (71), incorrect (69), confused (67)	film (231), not (199), american (183), first (167), known (146), one (145), also (142), released (129), (born (124), new (116), album (103), born (91), state (84), years (82), two (81), based (81), – (79), directed (78), played (77), series (76), united (75), written (73), people (71), best (69), band (67), may (66)

Table 23: Top 25 most common words used by round and gold label. Bolded words are used preferentially in particular subsets.

Subset	Context	Statement	Rationale	Context+Statement+Rationale
ANLI	film (647), american (588), **known** (377), first (376), (born (365), **also** (355), one (342), new (341), released (296), album (275), **united** (249), **directed** (240), not (236), – (218), **based** (214), **series** (196), **best** (191), may (188), **band** (185), **state** (182), **football** (177), two (175), written (175), **television** (175), **national** (169), **south** (165)	not (252), born (132), years (120), released (107), one (87), film (83), first (82), only (76), **people** (75), year (61), **played** (58), new (58), two (54), **made** (54), album (49), no (46), **died** (46), **won** (46), **less** (44), **last** (42), american (41), **years.** (40), **three** (40), written (38), **used** (37), **john** (37)	not (1306), **system** (753), **statement** (494), **know** (343), **think** (274), **definitely** (268), **context** (261), **correct** (243), **difficult** (228), only (224), **doesn't** (223), may (221), **confused** (218), no (200), **says** (198), **incorrect** (193), **text** (184), **could** (181), **states** (166), born (160), one (155), **say** (147), years (146), **don't** (140), **would** (130), **whether** (129)	not (1794), film (802), system (781), american (659), one (584), first (563), statement (511), released (504), known (495), also (467), new (452), only (437), may (429), know (387), born (386), (born (371), album (362), no (337), think (337), based (335), years (332), two (313), states (313), united (308), state (304), directed (301)
Numerical	american (236), film (211), (born (162), first (151), known (138), album (136), new (129), released (126), also (117), united (117), one (109), – (101), band (87), series (83), best (82), television (79), directed (77), football (76), based (75), state (74), played (73), second (72), south (71), world (70), city (69), states (65)	years (114), born (79), released (74), **first** (61), year (52), not (44), **died** (38), **less** (37), two (36), one (35), years. (34), **three** (32), population (30), old (30), film (28), **ago.** (27), album (26), only (24), old. (24), **century.** (23), last (23), won (20), **least** (20), world (20), second (18), played (18)	not (344), system (291), statement (166), years (137), difficult (125), born (115), think (103), definitely (102), year (90), confused (90), only (88), correct (84), know (82), context (77), released (72), may (71), incorrect (70), first (61), text (60), could (59), would (57), one (55), says (51), doesn't (50), **mentioned** (49), died (48)	not (423), system (297), years (278), first (273), released (272), film (265), american (256), born (231), one (199), album (181), year (170), statement (169), (born (166), known (160), only (158), new (158), two (145), may (140), also (137), united (134), difficult (125), based (124), states (117), think (112), band (111), second (109)
Basic	film (238), american (193), one (143), known (138), new (135), first (134), also (132), not (125), released (105), directed (104), (born (100), album (99), state (97), united (90), may (83), song (80), based (78), series (74), best (74), two (73), television (72), – (69), south (68), written (68), said (65), would (64)	not (219), one (51), people (41), no (36), film (31), new (31), released (28), less (28), never (27), played (24), only (24), born (23), two (23), made (22), album (21), last (21), first (21), used (20), least (18), written (18), three (17), directed (17), best (16), years (16), **movie** (16), **good** (16)	not (546), system (290), statement (248), know (125), definitely (120), think (115), context (101), says (101), doesn't (97), correct (92), only (91), confused (89), may (88), incorrect (83), states (78), no (76), text (75), could (69), one (65), difficult (61), whether (58), would (58), say (56), neither (54), said (52), model (50)	not (890), system (303), film (298), one (259), statement (254), american (227), new (196), first (191), known (182), also (181), may (180), only (176), released (165), no (154), think (149), know (146), state (140), would (137), two (134), directed (133), album (132), states (128), based (127), says (127), people (126), said (123)
Reference	film (188), american (163), known (139), (born (128), also (112), first (98), one (85), new (83), directed (72), – (71), not (71), released (70), best (66), united (61), album (57), television (56), south (54), world (54), based (53), may (52), written (52), series (50), band (49),) (45), two (45), national (44)	not (70), born (39), years (33), name (23), film (21), made (20), won (19), one (19), people (19), first (19), only (17), year (17), played (16), released (16), died (16), known (15), band (15), speaker (14), new (14), written (14), three (13), two (12), no (12), man (12), directed (11), album (10)	not (358), system (199), statement (112), know (91), think (71), doesn't (70), confused (67), may (66), context (60), model (60), only (57), says (52), correct (52), could (51), definitely (50), name (50), difficult (49), born (46), one (42), probably (41), would (41), incorrect (40), states (39), don't (38), no (35), understand (34)	not (499), film (230), system (207), known (186), american (171), first (147), one (146), also (139), (born (129), may (126), statement (122), born (122), new (112), only (109), name (105), released (105), know (104), think (100), directed (93), years (89), would (88), written (84), two (83), states (82), based (82), best (80)
Tricky	film (227), american (142), first (110), known (104), one (102), also (99), new (93), (born (88), album (83), released (81), directed (77), based (75), song (71), not (68), series (65), written (61), united (60), band (59),) (55), may (51), – (50), south (48), only (48), two (48), television (46), located (44)	not (82), only (58), born (33), film (32), released (27), one (26), two (22), first (21), made (19), years (19), new (18), three (18), played (16), album (16), american (16), used (16), people (14), series (14), wrote (13), directed (13), written (13), also (13), band (13), known (13), won (13), **starts** (12)	not (386), system (204), statement (129), only (88), know (75), think (73), difficult (69), context (67), confused (66), incorrect (63), definitely (63), may (57), correct (54), says (51), states (49), doesn't (48), one (43), name (42), used (41), text (41), no (40), **ai** (38), don't (37), **words** (36), first (36), could (35)	not (536), film (281), system (208), only (194), one (171), first (167), american (166), also (146), known (141), statement (133), new (124), released (123), album (111), may (110), based (110), directed (99), two (92), (born (89), written (89), series (88), know (87), song (86), used (86), made (86), name (86), think (85)
Reasoning	film (390), american (363), (born (245), first (229), also (227), known (226), new (219), one (203), released (173), album (159), united (154), directed (151), not (147), based (138), – (125), football (124), state (117), national (116), played (111), best (110), band (109), television (108), may (108), series (106), former (105), south (104)	not (131), born (92), released (66), years (60), people (50), first (49), one (49), film (43), played (39), year (36), only (35), new (35), made (30), never (30), two (29), died (27), album (27), won (26), no (26), known (25), last (25), american (24), used (24), united (22), **john** (22), city (22)	not (919), system (466), know (291), statement (279), context (188), definitely (173), correct (172), doesn't (171), think (164), no (162), may (162), could (147), difficult (144), only (126), say (126), whether (123), says (119), confused (119), text (118), don't (114), neither (110), incorrect (110), born (101), one (96), information (95), states (92)	not (1197), system (483), film (481), american (411), one (348), first (335), know (312), released (307), known (306), also (292), new (290), statement (288), may (281), (born (250), only (249), born (249), no (239), state (218), based (213), album (206), think (200), played (196), united (196), context (191), could (184), doesn't (182)
Imperfections	film (87), american (76), also (54), one (52), first (47), known (45), released (45), new (44), album (42), not (36), based (35), directed (35), (born (35), city (34), united (33), written (31), two (30), song (29), – (26), series (25), band (25), people (25), television (24), population (24), name (24), national (24)	not (38), film (18), people (14), born (12), written (12), one (12), only (11), first (11), made (10), released (10), new (10), american (8), city (8), two (7), years (7), **popular** (7), many (6), different (6), united (6), album (6), **street** (6), show (6), also (6), population (6), three (6), **life** (5)	not (168), system (82), statement (70), know (50), correct (38), context (35), think (34), says (32), no (30), definitely (29), doesn't (28), confused (26), could (26), incorrect (26), one (24), states (23), only (23), stated (22), neither (22), may (21), model (21), say (21), text (20), don't (20), difficult (19), state (19)	not (242), film (116), american (94), system (89), one (88), statement (72), also (72), first (71), known (65), released (64), know (63), new (58), written (55), based (54), album (53), only (52), no (50), two (49), people (47), think (46), city (45), may (44), states (44), made (43), directed (42), united (42)

Table 24: Top 25 most common words used by annotation tag. Bolded words are used preferentially in particular subsets.