CURRICULUM: A Broad-Coverage Benchmark for Linguistic Phenomena in Natural Language Understanding

Anonymous ACL submission

Abstract

In the age of large transformer language models, linguistic benchmarks play an important role in diagnosing models' abilities and limitations on natural language understanding. However, current benchmarks show some significant shortcomings. In particular, they do not provide insight into how well a language model captures distinct linguistic phenomena essential for language understanding and reasoning. In this paper, we introduce CURRICULUM, a new large-scale NLI benchmark for evaluation on broad-coverage linguistic phenomena. We show that our benchmark for linguistic phenomena serves as a more difficult challenge for current state-of-the-art models. Our experiments also provide insight into the limitation of existing benchmark datasets. In addition, we find that sequential training on selected linguistic phenomena effectively improves generalizing performance on adversarial NLI under limited training examples.

1 Introduction

With the rising power of pre-trained language models, large-scale benchmarks serves as an important factor driving the future progress of NLP. These benchmarks can provide a tool for analyzing the strengths and weaknesses of pre-trained language models. In recent years, many benchmarks (Wang et al., 2019, 2020; Rajpurkar et al., 2018) have been proposed for diverse evaluation objectives. However, criticisms have been made that these benchmarks do not formulate specific linguistic skills required for understanding(Raji et al., 2021). Thus, they do not explain how well a language model captures distinct linguistic phenomena essential to language understanding and reasoning.

In this paper, we present the CURRICULUM benchmark: a large-scale collection of diverse natural language inference (NLI) datasets for evaluating how well a language model captures reasoning skills for distinct types of linguistic phenomena. Targeted linguistic phenomena in CURRICULUM range from fundamental properties like named entity and coreference to complex ones like commonsense and deductive reasoning. With the CURRICULUM benchmark, we aim to investigate the following research questions:

- **Q1**: Do language models trained on benchmark datasets have the ability to reason over a wide range of linguistic phenomena?
- **Q2**: Are linguistic phenomena missing from the training data recoverable through inoculation?
- **Q3**: Do language models learn a general reasoning skill of a phenomenon through inoculation?
- **Q4**: Can models generalize from linguistic phenomena data to adversarial inference tests with limited training examples?

To address the above questions, we empirically analyze NLI models trained on popular benchmark datasets through a zero-shot diagnostic test, inoculation by fine-tuning, hypothesis-only tests, and cross-distribution generalization tests. In addition, we closely study the low-data generalization performance of models sequentially trained on selected linguistic phenomena datasets.

For **Q1**, we observe that models trained on benchmark datasets, including adversarial data, do not have the reasoning ability for a large set of linguistic phenomena. Our results show that training on more datasets can help the model learn more types of reasoning but does not help the model acquire complex reasoning skills. Our benchmark exposes multiple knowledge gaps in large NLI models regarding diverse linguistic phenomena. For **Q2**, our analysis provides empirical evidence that either exploits the lack of recoverable linguistic phenomena in benchmark datasets or exposes models' inability to learn certain linguistic phenomena. We also show that, on some phenomena, models may rely heavily on superficial cues or artifacts existing in the hypothesis to reach high accuracy.

For **Q3**, Our experiments show that a model's learning performance may not align with its generalization ability. Models fail to generalize across different difficulty distributions on many phenomena, suggesting the lack of a general reasoning skill. Models can generalize across distributions only on a limited number of phenomena. For **Q4**, we find that sequential training on selected linguistic phenomena can help the model efficiently generalize to the adversarial test sets under limited training examples. Compared to models trained on large-scale NLI datasets (MNLI and SNLI), linguistic-phenomena-based sequential training shows a more significant performance gain and is a more efficient method. Overall, our proposed benchmark systematically maps out a wide range of specific linguistic skills required for language understanding and inference. We envision linguistic-phenomena-based evaluation to be an integral component of general linguistic intelligence. We hope CURRICULUM can serve as a useful evaluation tool and learning scaffold for more complex language understanding.

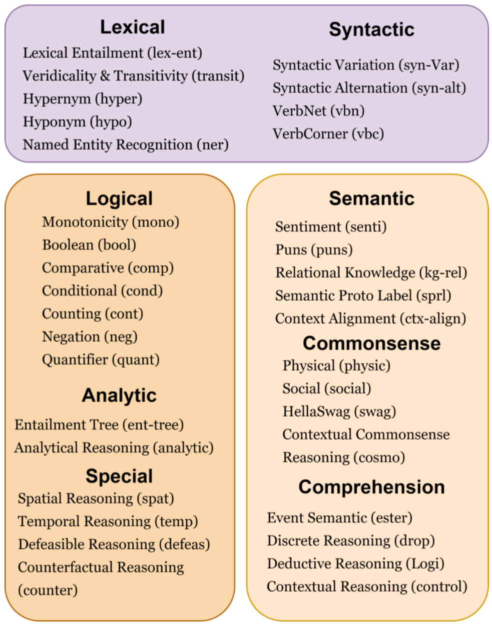

Figure 1: Linguistic Phenomena Ontology for the CURRICULUM benchmark. Abbreviation for each phenomena, used in this paper, is listed in the parenthesis.

2 Related Work

NLU Benchmarks In recent years, multiple large-scale benchmarks for evaluating models' general language understanding performance have been proposed. Similar to our benchmark's task format, SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) are the two common benchmarks for Natural Language Inference (NLI). GLUE and SuperGLUE are the two most popular benchmarks that aim to provide a straightforward comparison between task-agnostic transfer learning techniques. They cover various task formats, task domains, and training volumes, with datasets all collected from publicly available sources. The construction of our benchmark is similar in that we also collect linguistic phenomena datasets from published papers. Adversarial NLI (ANLI) was a new benchmark collected "via an iterative, adversarial human-and-model-in-the-loop procedure." (Nie et al., 2020). ANLI was shown to be a more difficult challenge than previous benchmarks. A part of our study focuses on the low-data generalization performance on ANLI. Different from these benchmarks, our work aims to map out and evaluate specific linguistic skills a model needs for language understanding.

Challenge Datasets for NLU Many challenge datasets have been developed to evaluate models on specific linguistic skills for understanding. These datasets are in different formats such as NLI, Question Answering (QA), and Reading Comprehension (RC). They target a large set of skills including monotonicity (Yanaka et al., 2019a), deductive logic (Liu et al., 2020), event semantics (Han et al., 2021), physical and social commonsense (Sap et al., 2019; Bisk et al., 2019), defeasible reasoning (Rudinger et al., 2020), and more. Our work brings together a set of challenge datasets to build a benchmark covering a large set of specific linguistic skills. We also merge different evaluation methods proposed by these works into a complete evaluation pipeline for our benchmark.

Evaluation on Linguistic Phenomena Our work is mostly related to the DNC (Poliak et al., 2018a) benchmark that also provides a collection of datasets focusing on distinct linguistic phenomena. Several datasets in the syntactical and semantic categories come directly from this collection. DNC includes many different recast NLI datasets (White et al., 2017) which are converted automatically from other NLU datasets with little human effort. We follow their idea and automatically convert several datasets from the QA and RC domain into recast NLI datasets to cover phenomena like commonsense and deductive reasoning. Our benchmark covers a wider range of linguistic phenomena from richer categories than DNC. In particular, our benchmark contains semantic phenomena and

Category	Description
Lexical	Testing a model's Word-level reasoning skill on lexical semantic and direct or transitive lexical relationships.
Syntactic	Testing a model's reasoning skill on syntactic structure and compositionality.
Semantic	Testing a model's reasoning skill on sentence-level reasoning involving diverse semantic properties: entity relations, context, events, subjectivity, and semantic proto roles.
Logical	Testing a model's reasoning skill on logical operations: propositional structure, quantification, and monotonicity.
Analytical	Testing a model's knowledge exploitation ability: drawing accurate conclusions based on domain-specific knowledge, symbolic knowledge, and interpretable reasoning steps.
Commonsense	Testing a model's reasoning skill on commonsense knowledge independent of cultural and educational background.
Comprehension	Testing a model's reasoning skill on complex reasoning types targeted by different reading comprehension challenges.
Special	Testing a model's reasoning skill on non-monotonic and spatial-temporal reasoning.

Table 1: Descriptions of each category in the CURRICULUM benchmark

includes phenomena from fundamental linguistic properties to complex reasoning types. In addition, the evaluation methodology for our benchmark provides more in-depth analysis of model behaviors.

3 The CURRICULUM Benchmark

3.1 Benchmark Construction

Our benchmark aims to map out a specific set of linguistic skills required for language understanding. The targeted linguistic skills should range from fundamental linguistic properties to complex reasoning types. Our linguistic phenomena selection is motivated by three benchmarks: GLUE Diagnostic, Rainbow, and DNC. In addition, we include many more phenomena focusing on complex reasoning types such as deductive logic and analytical thinking. Our finalized benchmark covers eight categories of linguistic phenomena. We briefly describe the types of reasoning skill each category focus on in Table 1. Appendix A and B shows a list of references and dataset details for the train and test datasets used for each linguistic phenomenon.

3.2 Dataset Selection

We collect many challenge NLI or NLU datasets and filter them individually with the following criteria: (1) We focus on datasets that evaluate a specific or a set of specific linguistic phenomena. (2) We focus on English monolingual datasets that are institutional and publicly available. (3) We exclude datasets that require domain-specific knowledge not given by the premise, such as medical knowledge. We finalize our selection with 36 datasets. Figure 1 shows a detailed ontology of our selected linguistic phenomena and their abbreviations.

$\mathcal{P}$	I_v	$\mathcal{P}$	I_v	$\mathcal{P}$	I_v
lex-ent	0.31	transit	0.41	hyper	-0.99
hypo	-0.10	ner	0.19	vbn	0.55
vbc	-0.40	syn-alt	0.10	syn-var	0.11
bool	1.12	cond	1.13	cont	0.75
comp	0.98	negat	1.13	quant	0.78
monot	-1.57	kg-rel	0.05	coref	-0.38
senti	0.42	ctx-align	-0.79	puns	0.14
sprl	-0.11	ent-tree	0.50	analytic	0.00
temp	0.10	spat	0.49	counter	0.47
defeas	-0.39	social	-0.40	physic	-0.17
swag	-0.66	cosmo	-0.57	drop	0.19
ester	-0.10	logi	-0.71	control	-0.07

Table 2: Dataset difficulty measured by the amount of usable information (I_v) from input data instances. The lower I_v is the more difficulty a dataset will be for the model. $\mathcal{P}$ here are the abbreviations of linguistic phenomena listed in Figure 1.

3.3 Unified Task Format

We unified the task formats into a single linguistic task, Natural Language Inference (NLI). We select NLI as the universal task format because NLI often serves as a general evaluation method for models on different downstream tasks. A model would need to handle nearly the full complexity of natural language understanding in order to solve the NLI task (Poliak et al., 2018b). Our benchmark contains two types of NLI problems: (1) the 3-way NLI with `Entailment`, `Contradiction`, and `Neutral`; (2) the 2-way NLI with `Entailed` and `Not-Entailed`. Each example has a premise and a hypothesis with 2-way or 3-way labels.

3.4 Automatic Recast

To convert non-NLI datasets into the NLI task format, we follow the dataset recast procedure (Poliak et al., 2018b): automatically convert from non-NLI datasets with minimum human intervention. We design algorithmic ways to generate sentence pairs from the input text and convert the original labels into the NLI labels. Question Answering (QA) and Reading Comprehension (RC) are the two major tasks we need to convert. In QA datasets, if choices are given as declarative statements, we consider them as hypotheses and the question context as the premise. If choices are given as phrases answering the question, we concatenate the context and question to form a premise and consider the answers as hypotheses. Several datasets are tasks with free-response problems, and an answer can only be converted to an entailed hypothesis. To generate non-entailed hypotheses, we use several techniques during recasting, whose details are described in Appendix C. We randomly sample a subset of examples for each recast dataset and conduct human verification to ensure the conversion does not create artifacts in hypotheses for models to leverage. Our hypothesis-only bias analysis shows that most of our recast datasets have low hypothesis-only bias.

3.5 Dataset Controlled Split

We split each dataset along the pointwise difficulty dimension. The point-wise difficulty is measured by the pointwise $\mathcal{V}$-information (Ethayarajh et al., 2021). The pointwise $\mathcal{V}$-information (PVI) is a framework for measuring the degree of usable information in individual data examples. The higher the PVI, the more usable information a data example contains, the easier that example is for a model. Given input data X, output Y, and the model family $\mathcal{V}$, the PVI is computed as:

$$\mathrm{PVI}(x \rightarrow y) = -\log_2 \mathrm{g}[\varnothing](y) + \log_2 \mathrm{g}'[x](y)$$

We first calculate the PVI for each phenomenon dataset, then we split each dataset into two portions: simple and hard, based on each example's PVI.

3.6 Dataset Difficulty

To enhance our benchmark to provide more information on each dataset for in-depth evaluation and analysis, we provide each phenomenon a difficulty level. The $\mathcal{V}$-information framework can also serve as a difficulty measurement for datasets and can be computed explicitly by averaging over PVI:

$$\mathrm{I}_v(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{1}{n}\sum_i \mathrm{PVI}(x_i \rightarrow y_i)$$

As Table 2 shows, the difficulty level ranges from negative to positive. The higher the $\mathcal{V}$-information is, the easier a dataset is for the model.

Name	Model	Train/Test	Accuracy
roberta-mnli	RoBERTa (Liu et al., 2019c)	MNLI/MNLI	90.2%
bart-mnli	BART (Lewis et al., 2020)	MNLI/MNLI	89.9 %
roberta-anli-mix	RoBERTa	SNLI, MNLI, FEVER, ANLI/ ANLI	53.7 %
xlnet-anli-mix	XLNet (Yang et al., 2019)	SNLI, MNLI FEVER, ANLI/ ANLI	55.1 %

Table 3: Details on models used in our experiments. All four models are large models and publicly available.

4 Evaluation Methodology

We define an evaluation process for the CURRICULUM benchmark that aims to bring different types of evaluation and diagnosing methods used by previous challenge NLI datasets. Following Raji et al. (2021)'s suggestion, we want our evaluation process to both to analyze the model output in detail and explore which aspects of the inference problem space remain challenging to current models.

Zero-shot Diagnostic Test This test is motivated by the diagnostic test in GLUE. We focus on providing fine-grained analysis of zero-shot system performance on a broad range of linguistic phenomena.

Inoculation by Fine-tuning We use inoculation (Liu et al., 2019a) to further analyze model failures on target linguistic phenomena. This method fine-tunes the model on a down-sampled training section of a phenomenon dataset (inoculation). One can interpret inoculation performance in two ways:

1. Good performance: the original training set of the model, prior to inoculation, did not sufficiently cover the target phenomenon, but it is recoverable through inoculation.
2. Poor performance: there exists a model weakness to handle the target phenomenon.

Hypothesis-only Bias Analysis We train a hypothesis-only baseline (Poliak et al., 2018b) for each phenomenon to verify whether the model's good performance is from leveraging artifacts in the hypotheses. We want to ensure that models' improved performance after inoculation is due to their ability to reason about a hypothesis and the given context together. We also use the baselines to assure dataset quality by observing the amount of hypothesis-only bias each dataset contains.

Cross-Distribution Generalization We conduct the cross-distribution generalization test Rozen

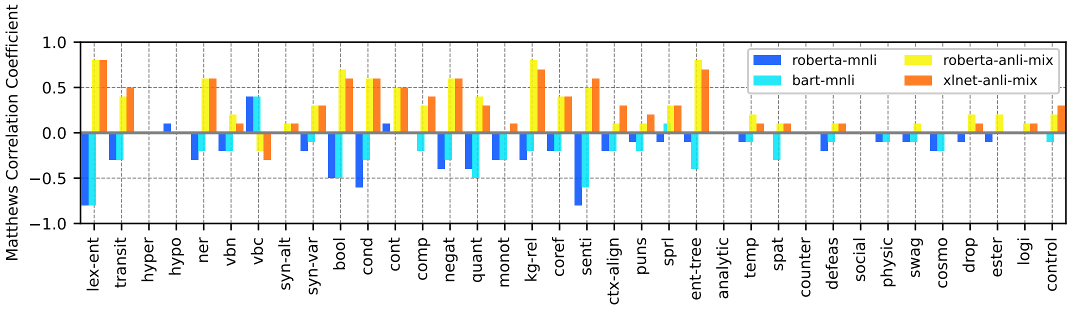

Figure 2: Zero-shot system performance on the CURRICULUM benchmark.

et al. (2019) to verify if the model learns a general reasoning skill from inoculation. The good inoculation performance does not ensure that the model's learned skill is generalizable. The model can likely over-fit the dataset distribution by adopting superficial cues. We evaluate the model's generalization ability by training and testing the model on different distributions within the same phenomenon.

4.1 Experiment Setup

For the zero-shot test, we test a model on each test set without additional fine-tuning. We select NLI models with top performance on NLI benchmarks MNLI and ANLI. We list these models in Table 3. We follow the GLUE diagnostic dataset and use the Matthews correlation coefficient as the evaluation metric. For inoculation, we fine-tune models on training examples with a size ranging from 10 to 1000 examples per label. For the cross-distribution generalization test, we first create variant data distributions for train and test sets using the controlled split method from Section 3.5. We split each dataset into two portions (simple and hard) according to the point-wise $\mathcal{V}$ information. Next, we either train and test the model on the same difficulty distribution or train it on one portion and test it on a different portion. In the inoculation, hypothesis-only, and generalization experiments, we all use roberta-anli-mix as our NLI model because its training set covers all the major NLI training datasets: SNLI, MNLI, FEVER (Thorne et al., 2018), and ANLI. We use accuracy as our evaluation metric for all these three experiments.

5 Empirical Analysis

5.1 Zero-shot Linguistic Phenomena Diagnose

First, we report the results on zero-shot diagnostic evaluation for each baseline model. From Figure 2, we observe that both contextualized and generative models trained on MultiNLI show a negative correlation in the majority of linguistic phenomena. Meanwhile, anli-mix models (roberta-anli-mix, xlnet-anli-mix) are positively correlated on most (77.8 %) of the phenomena and they show high correlation (> 0.50) on 27.8 % of the phenomena. On average, models trained on the large dataset mixture show better performance than models trained on MultiNLI alone, suggesting that training on more datasets help models capture more types of linguistic phenomena. However, most of the phenomena captured by the anli-mix models are easier to learn (higher $\mathcal{V}$ information). On harder phenomena, models did not benefit from the training dataset mixture. For instance, both the anli-mix models have a low correlation on deductive and analytical reasoning. Overall, the zero-shot evaluation shows that a benchmark with a wide range of linguistic phenomena can evaluate a model's specific linguistic skills.

5.2 Inoculation

Based on Figure 3, the model can reach high accuracy on about 64 % of the phenomena as the training examples accumulate. Most of these phenomena have higher $\mathcal{V}$ information (> 0.0) that should relatively be easier to learn. We are surprised that for some hard phenomena (≤ 0.0) such as commonsense contextual reasoning (cosmo, -0.67), the model's performance improved after inoculation. The improvement shows an gap in the original training data mixture.

On 25 % of the phenomena, the model's performance did not improve significantly after inoculation, meaning that it fails to learn the reasoning skills for these phenomena. Most of these phenomena are difficult, with a low $\mathcal{V}$ information, such as monotonicity(mono) and deductive (logi) reasoning. The accuracy is consistently low when train-

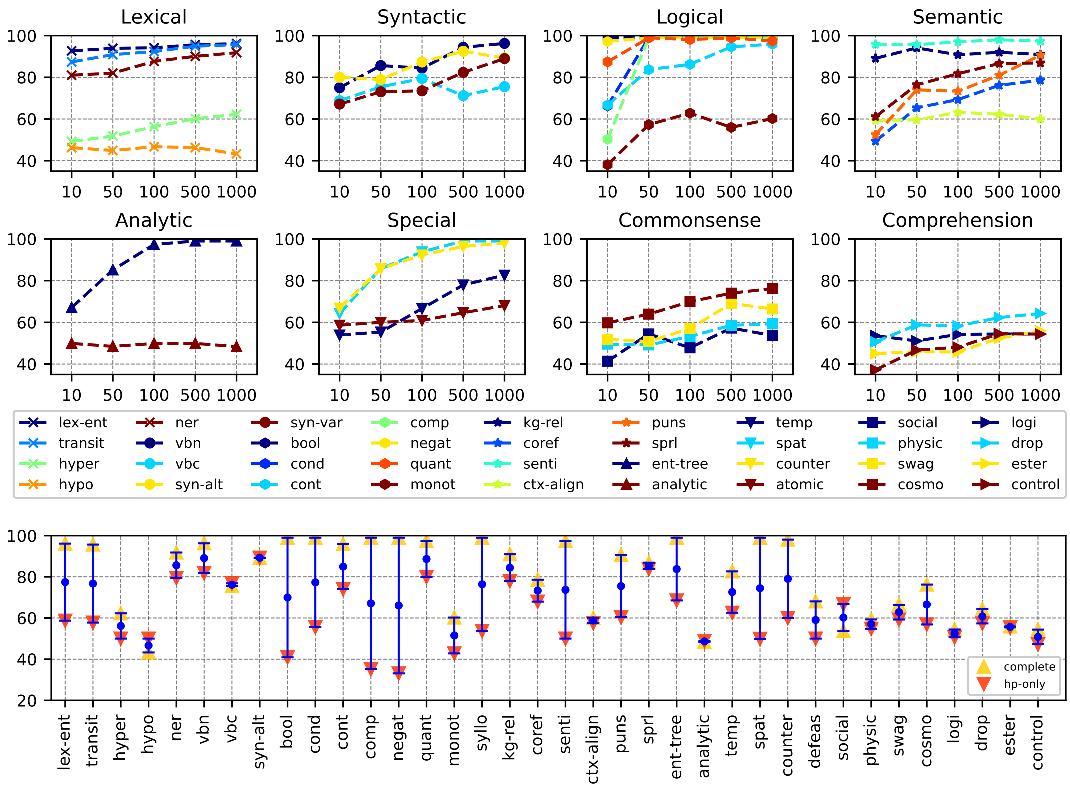

Figure 3: Inoculation by fine-tuning (top) vs Hypothesis-only analysis (bottom). The X-axis of the top plot represents training examples per label. Both plots' Y-axis show the accuracy. Models used in these two experiments are both the roberta-anli-mix model, introduced in Section 4.1.

ing examples accumulate. We also observe that model struggles to learn phenomena that require complex reasoning, such as phenomena from the comprehension category. This trends show inherent weaknesses in the model or its training strategy that cause its failure to learn complex and hard phenomena. Overall, results from this experiment, combined with the zero-shot evaluation, suggest that many linguistic phenomena are missing from different large-scale NLI datasets but are recoverable through additional training examples. However, the model fails to learn the skills for hard and complex phenomena.

5.3 Hypothesis-only Bias

To determine if models can leverage spurious artifacts in the hypotheses of each phenomenon, we compare full models to hypothesis-only baselines. From Figure 3, we observe that hypothesis-only baseline performs poorly on a majority of the phenomena. This indicates that our benchmark generally requires the model to learn an inference process between contexts and hypotheses for good performance. We observe that on 30.6% of the phenomena, the full-model can reach a high accuracy while the baseline has low accuracy, suggesting the model can learn the phenomenon without relying on hypothesis artifacts. On 36 % of the phenomena, the model does not show a significant performance gain compared to the baseline. Most of these are complex reasoning phenomena like deductive and analytical reasoning. The result validates that the model struggles more with complex linguistic phenomena. On 33.3 % of the phenomena, both the full-model and the baseline achieve high accuracy showing the possibility that the model exploits spurious artifacts from the hypothesis to reach high accuracy. Overall, this experiment shows that the hypothesis-only baseline effectively verifies the performance from inoculation. These results also assure the quality of our benchmark.

5.4 Generalization

As Figure 4 show, the model can adapt between different distributions only on 22.2 % of the phenomena. The model achieves high accuracy consistently for all four categories in the generalization matrix suggesting the learned skills are general-

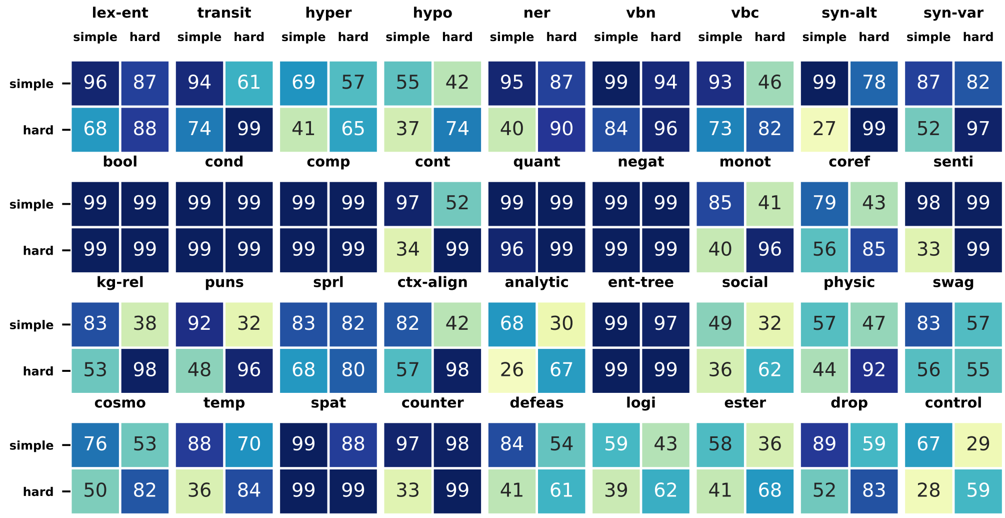

Figure 4: Generalization between controlled dataset splits. Here each heat-map shows the generalization performance of the model fine-tuned and evaluated on different distributions within each linguistic phenomenon.

izable. On 58.3 % phenomena, models can not generalize between different difficulty distributions. They show higher accuracy when trained and tested on the same distribution but low accuracy when the test distribution shifted. For example, on relational knowledge reasoning (kg-rel), the model achieves 83% for simple $\rightarrow$ simple and 98 % for hard $\rightarrow$ hard. Nevertheless, the performance drops to 53 % for hard $\rightarrow$ simple and 38 % for simple $\rightarrow$ hard. Notice that model's good performance on inoculation does not align with its generalization ability. For example, the model reaches 90.9 % accuracy on kg-rel, but its generalization performance is poor. This behavior highlights a model weakness: can over-fit to a particular distribution but fail to learn a general reasoning skill for the target phenomenon.

We observe an interesting behavior that models struggle to generalize from hard to simple distribution on about 14 % of the phenomena while showing good generalization from simple to hard distribution. We think the possible reason is that the hard distribution contains data with relatively low $\mathcal{V}$ information. A low amount of usable information makes it hard for the model to learn the phenomena well enough to generalize to the simple distribution.

6 Sequential Training On Curriculum

This section studies the effectiveness of sequential training on linguistic phenomena for low-data generalization to a target dataset. Sequential training (Liu et al., 2019b) first conducts multi-task training on multiple datasets (excluding the target dataset) and then continues to fine-tune on the target dataset. The goal is to transfer from intermediate datasets to the target task to improve the performance. We want to investigate whether a combination of linguistic phenomena data can transfer well to the ANLI dataset and thus improve a model's low-data generalization performance.

Setup We conduct a random search by sampling a combination of phenomenon datasets from the benchmark. We select RoBERTa-large as our model following the ANLI paper. We first train the model on the data combination of selected phenomena. Next, we fine-tune the model on each round of ANLI with limited data examples ($\leq$ 2000) per label. Through a random search, we have the *ling* model in Figure 5 which shows the best performance among other random selections. The phenomena selected for *ling* include ester, drop, temporal, and all the semantic phenomena. We create an additional selection by adding lexical entailment and syntactic variation to *ling*'s selection. Many NLI datasets (Bowman et al., 2015; Marelli et al., 2014) have covered these two phenomena, which could potentially improve the performance. The model trained using this selection is *ling+* in Figure 5. We select three baseline strategies for comparison: *direct*, *mnli*, and *snli*. The direct strategy fine-tunes the model on ANLI without sequential training on any intermediate tasks. The other two baselines are first trained on MNLI or SNLI before fine-tuning on ANLI.

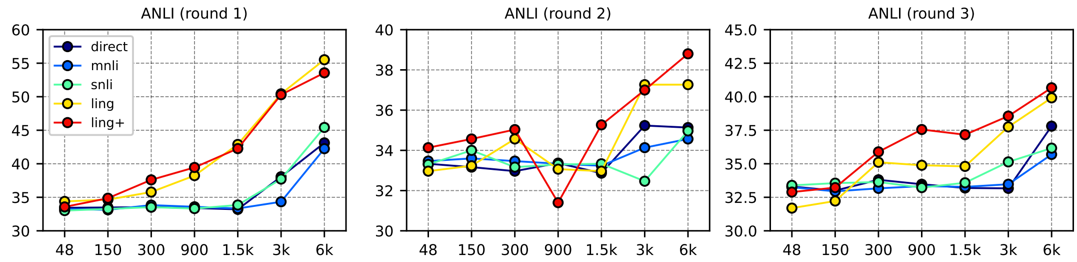

Figure 5: Generalization performance on ANLI under low-data regime, with different sequential training strategies. The model used is a pre-trained RoBERTa-large model. The X-axis represents the number of training examples used. The Y-axis shows the accuracy. The accuracy reported here are the average of three trial runs.

Result As Figure 5 shows, sequential training on selected linguistic phenomena first (*ling* and *ling+*) indeed improve the low-data generalization performance on all three rounds of ANLI. The learning curves show that the performance of these two models improves much faster, and overall they have higher accuracy than the baselines. This trend is more significant on ANLI round 1, where the data efficiency of *ling* and *ling+* can increase at a much faster rate than the baselines.

Note that the *mnli* and *snli* baselines are both trained with an extensive amount of examples before fine-tuning on ANLI. However, they only show trivial improvements over *direct*. Models trained on linguistic phenomena perform better than baselines, even with fewer intermediate training examples. This suggests that sequential training on selected linguistic phenomena is more efficient than pre-training on large-scale benchmark datasets.

Overall, our experiment highlights the benefit of sequential training on selected linguistic phenomena for learning adversarial NLI examples under a low-data regime. Many factors play an important role in sequential training, such as task selections, training strategies, and hyperparameters. Due to computational constraints, our random search cannot cover all possible settings. We encourage future work to examine a wide range of scenarios. That being said, we believe that linguistic phenomena can be potential learning scaffolds for NLI models.

7 Conclusion and Future Work

In this paper, we provide a comprehensive study on how well language models capture specific linguistic skills essential for understanding. We also explore the potential of linguistic phenomena as learning scaffolds to improve models' generalization performance in the low-data regime. We introduce the CURRICULUM benchmark that covers 36 types of linguistic phenomena ranging from fundamental properties to complex reasoning types. We then defined an evaluation methodology that can analyze model behavior in different aspects. Our major findings include:

- Models trained on benchmark NLI datasets fail to reason over a diverse set of linguistic phenomena.
- Good inoculation performance on some phenomena results from the model leveraging superficial artifacts in the hypothesis.
- The model tends to over-fit the dataset distribution without learning a general reasoning skill on a majority of phenomena.
- Sequential training on selected linguistic phenomena can effectively improve the model's generalization performance on adversarial NLI under low-data settings.

Overall, our benchmark effectively evaluates a model on specific linguistic skills. We hope that our benchmark and empirical findings can encourage the development of new datasets that cover richer types of linguistic phenomena and language models to handle more types of complex reasoning. We plan to study more on phenomena selection methods and training strategies that can improve the few-shot performance on adversarial tests for future work. We also plan to add more linguistic phenomena and evaluation methods into our benchmark.

References

BIG-bench collaboration. 2021. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

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.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

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.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2021. Information-theoretic measures of dataset difficulty.

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.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. Ester: A machine reading comprehension dataset for event semantic relation reasoning.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context - investigating contextual reasoning over long texts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13388–13396. AAAI Press.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. Inoculation by fine-tuning: A method for analyzing challenge datasets. 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 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Rajaswa Patil and Veeky Baths. 2020. CNRL at SemEval-2020 task 5: Modelling causal reasoning in language with multi-head self-attention weights based counterfactual detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 451–457, Barcelona (online). International Committee 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.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

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

Kyle Richardson and Ashish Sabharwal. 2020. What does my QA model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Martin Schmitt and Hinrich Schütze. 2021. Language models for lexical inference in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1267–1280, Online. Association for Computational Linguistics.

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv: Artificial Intelligence*.

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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. 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.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural NLI models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 920–934, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text.

A Linguistic Phenomena in CURRICULUM

Phenomena	Train Reference	Test Reference
	Lexical Phenomena	
Lexical Entailment	Schmitt and Schütze 2021	Schmitt and Schütze 2021; Glockner et al. 2018
Hypernymy	Richardson and Sabharwal 2020	Richardson and Sabharwal 2020
Hyponymy	Richardson and Sabharwal 2020	Richardson and Sabharwal 2020
Named Entity	Poliak et al. 2018a	Poliak et al. 2018a
Veridicality and Transitivity	Poliak et al. 2018a; Yanaka et al. 2021	Poliak et al. 2018a; Yanaka et al. 2021
	Syntactic Phenomena	
VerbNet	Poliak et al. 2018a	Poliak et al. 2018a
VerbCorner	Poliak et al. 2018a	Poliak et al. 2018a
Syntactic Variation	Dolan and Brockett 2005	Dolan and Brockett 2005
Syntactic Alternations	Kann et al. 2019	Kann et al. 2019
	Semantic Phenomena	
Coreference & Anaphora	Sakaguchi et al. 2019; Wang et al. 2019; Webster et al. 2018	Sakaguchi et al. 2019; Wang et al. 2019; Webster et al. 2018
Sentiment	Poliak et al. 2018a	Poliak et al. 2018a
Relational Knowledge	Poliak et al. 2018a	Poliak et al. 2018a
Puns	Poliak et al. 2018a	Poliak et al. 2018a
Semantic Proto Label	White et al. 2017	White et al. 2017
Context Alignment	White et al. 2017	White et al. 2017; BIG-bench collaboration 2021
	Logical Phenomena	
Boolean	Richardson et al. 2019	Richardson et al. 2019
Conditional	Richardson et al. 2019	Richardson et al. 2019
Comparative	Richardson et al. 2019	Richardson et al. 2019
Counting	Richardson et al. 2019	Richardson et al. 2019
Quantifier	Richardson et al. 2019	Richardson et al. 2019
Negation	Richardson et al. 2019	Richardson et al. 2019
Monotonicity	Yanaka et al. 2019b	Yanaka et al. 2019a; Richardson et al. 2019
	Analytic Phenomena	
Entailment Tree	Dalvi et al. 2021	Dalvi et al. 2021
Analytical Reasoning	Zhong et al. 2021	Zhong et al. 2021
	Commonsense Phenomena	
Physical	Bisk et al. 2019	Bisk et al. 2019
Social	Sap et al. 2019	Sap et al. 2019
HellaSwag	Sap et al. 2018	Sap et al. 2018
Contextual Commonsense Reasoning	Huang et al. 2019	Huang et al. 2019
	Comprehension Phenomena	
Deductive Reasoning	Liu et al. 2020	Liu et al. 2020
Contextual Reasoning	Liu et al. 2021	Liu et al. 2021
Event Semantic Reasoning	Han et al. 2021	Han et al. 2021
Discrete Reasoning	Dua et al. 2019	Dua et al. 2019
	Special Reasoning Phenomena	
Defeasible Reasoning	Rudinger et al. 2020	Rudinger et al. 2020
Temporal Reasoning	Weston et al. 2016	Weston et al. 2016
Spatio Reasoning	Weston et al. 2016	Weston et al. 2016
Counterfactual Reasoning	Patil and Baths 2020	Patil and Baths 2020

Table 4: A detailed list of training datasets and test datasets used for each linguistic phenomenon in our benchmark.

B CURRICULUM Dataset Details in CURRICULUM

Name	|Train|	|Dev|	Original task
Lexical Entailment	6398	2964	NLI
Hypernymy	20000	8500	QA
Hyponymy	20000	8500	QA
Named Entity	50000	30000	NLI
Veridicality and Transitivity	20000	8788	NLI
VerbNet	1398	160	NLI
VerbCorner	110898	13894	NLI
Syntactic Variation	3668	408	SC
Syntactic Alternations	19990	8739	SC
Coreference & Anaphora	12135	5799	NLI/SC
Sentiment	4800	600	NLI
Relational Knowledge	21905	761	NLI
Semantic Proto Label	14038	1756	NLI
Puns	14038	1756	NLI
Context Align	14038	1756	NLI
Boolean	3000	1000	NLI
Conditional	3000	1000	NLI
Comparative	3000	1000	NLI
Counting	3000	1000	NLI
Quantifier	3000	1000	NLI
Negation QA	3000	1000	NLI
Monotonicity	35891	5382	NLI
Entailment Tree	1314	340	TG
Analytical Reasoning	3260	922	SC
Physical	10000	1838	QA
Social	6003	6003	QA
HellaSwag	20000	8518	QA
Contextual Commonsense Reasoning	9046	5452	RC
Deductive Reasoning	14752	2604	RC
Contextual Reasoning	6719	1604	RC
Event Semantics Reasoning	2800	662	RC
Discrete Reasoning	20000	13148	RC
Defeasible Reasoning	39036	9860	SC
Temporal Reasoning	4248	1174	NLI
Spatial Reasoning	10000	10000	QA
Counterfactual Reasoning	6062	3364	SC

Table 5: Overview of all the linguistic phenomena datasets in our benchmark. QA is short for Question Answering. NLI is short for Natural Language Inference. SC is short for Sentence Classification. TG is short for Text Generation. RC is short for Reading Comprehension.

C Data Recasting Details

Here we provide more details on the major techniques we used to convert Question Answering (QA) and Reading Comprehension (RC) datasets into recast NLI datasets.

C.1 Entity Swapping

<Original>
Context: ...The Buccaneers tied it up with a 38-yard field goal by Connor Barth, ... The game's final points came when Mike Williams of Tampa Bay caught a 5-yard pass...
Q: Who caught the touchdown for the fewest yard?
Answer: Mike Williams

<Recast>
Premise: ...The Buccaneers tied it up with a 38-yard field goal by Connor Barth, ... The game's final points came when Mike Williams of Tampa Bay caught a 5-yard pass...
Hypothesis: Mike Williams caught the touchdown for the fewest yard
Label: Entailed
Hypothesis: Connor Barth caught the touchdown for the fewest yard
Label: Not-Entailed

Table 6: Example of converting an RC example from DROP (Dua et al., 2019) to NLI format. The entailed hypothesis is a concatenation of question and answer. The non-entailed hypothesis is created by entity swapping on the entailed one (`Mike Williams` $\rightarrow$ `Connor Barth`).

C.2 Question/Answer Concatenation

<Original>
Context: The flash in the room that followed was proof of that assumption. The man grabbed his arm again. "Please let go of my arm." He requested, his voice low. "Look."
Q: Why did the man grabbed his arm?
Choice 1: The man wanted to dance with him.
Choice 2: ***The man wanted to get his attention.***
Choice 3: The man wanted to pull him closer so he can cry on this shoulder.
Choice 4: The man was angry with him and wanted to push him outside.

<Recast>
Premise: The flash in the room that followed was proof of that assumption. The man grabbed his arm again. "Please let go of my arm." He requested, his voice low. "Look."
Hypothesis: The man wanted to get his attention.
Label: Entailed
Hypothesis: The man wanted to dance with him.
Label: Not-Entailed

Table 7: Example of converting an QA example from Cosmos QA (Huang et al., 2019) to NLI format. The entailed hypothesis is the correct answer from the given choices. The non-entailed hypothesis is one of the false answers, excluding the choice "None of the above choices".

D Reproducibility

Implementation. All our experiments are implemented with models publicly available from Huggingface Transformers (Wolf et al., 2020).

Hyper-parameters We mainly follow the practice in (Nie et al., 2020). For all the experiments excluding the zero-shot test in Section 5.1, we use a learning rate of $1e-5$ with a batch size of 8. We set the number of warmup updates to be 1000. We set the epoch number to be 2. We evaluate the model on D_{dev} every 200 steps for the inoculation and generalization experiments, and 500 steps for the hypothesis-only experiment. For the low-data generalization on ANLI, we evaluate on the full-test set according to the number of training examples listed in Figure 5. We use the AdamW (Loshchilov and Hutter, 2019) as our optimizer.

Infrastructure All experiments are done with one single Geforce RTX 3090 (24GB). A single inoculation or generalization job finishes within 0.5 hours on average. A single hypothesis-only job finishes within 1-2 hours on average. A single job on sequential training and low-data fine-tuning finishes within approximately 1.5 hours on average.

Number of Parameters. RoBERTa-large model contains 355 million parameters. BART-large model contains 139 million parameters. BART-Large model contains 406 million parameters. XLNet-large model contains 340 million parameters.