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

Zeming Chen **Qiyue Gao**
Rose-Hulman Institute of Technology
{chenz16, gaoq}@rose-hulman.edu

Abstract

In the age of large transformer language models, linguistic evaluation play an important role in diagnosing models' abilities and limitations on natural language understanding. However, current evaluation methods show some significant shortcomings. In particular, they do not provide insight into how well a language model captures distinct linguistic skills essential for language understanding and reasoning. Thus they fail to effectively map out the aspects of language understanding that remain challenging to existing models, which makes it hard to discover potential limitations in models and datasets. In this paper, we introduce CURRICULUM as a new format of NLI benchmark for evaluation of broad-coverage linguistic phenomena. CURRICULUM contains a collection of datasets that covers 36 types of major linguistic phenomena and an evaluation procedure for diagnosing how well a language model captures reasoning skills for distinct types of linguistic phenomena. We show that this linguistic-phenomena-driven benchmark can serve as an effective tool for diagnosing model behavior and verifying model learning quality. In addition, our experiments provide insight into the limitation of existing benchmark datasets and state-of-the-art models that may encourage future research on re-designing datasets, model architectures, and learning objectives. [1].

1 Introduction

With the rising power of pre-trained language models, large-scale benchmarks serve 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 that offer a diverse set of evaluation objectives. However, recent criticisms have been made that these benchmarks fail to serve as effective measures of progress in machine learning (Raji et al., 2021). In particular, the task design does not formulate specific linguistic skills required for understanding. They lack the effectiveness in helping researchers understand how certain systems or models work and how they fail. Although many state-of-the-art language models have shown impressive performance on these common benchmarks, their performance degrades considerably on adversarial or out-of-distribution samples (Bras et al., 2020). The performance drop shows that models may not be learning the required linguistic skills for solving the tasks of these benchmarks but exploit spurious dataset biases (Poliak et al., 2018b). Overall, the current benchmark format seems to be more like a contest than a tool that can explain how well a language model captures distinct linguistic skills essential to language understanding and reasoning.

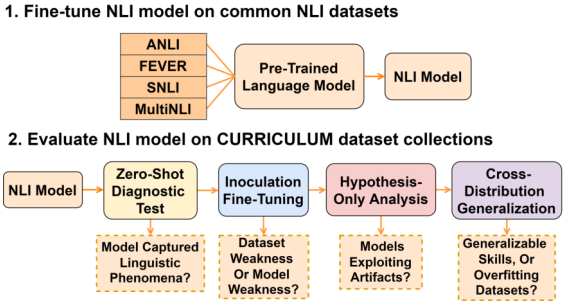


Figure 1: We propose a broad-coverage diagnostic benchmark for linguistic-phenomena-driven evaluation. Our benchmark includes both a dataset collection and an evaluation procedure for evaluating model performance and diagnosing linguistic skills captured by a model. We evaluate models fine-tuned on large NLI datasets through four types of diagnostic tests: zero-shot, inoculation, hypothesis-only, and cross-distribution.

In this paper, we propose a new form of benchmark that serves as a diagnostic evaluation tool for analyzing model linguistic skills. We present CURRICULUM benchmark: a framework for diagnosing

[1] Our code and data are publicly available at https://github.com/eric11eca/curriculum-ling

neural language models through broad-coverage linguistic phenomena. Our benchmark includes (1) a large-scale collection of natural language inference (NLI) datasets covering 36 linguistic phenomena and (2) an evaluation procedure for probing and 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 (i.e., continuing to train models on a small sample of examples) (Liu et al., 2019a)?
- **Q3**: Do language models learn a general reasoning skill of a phenomenon through inoculation?

To address the above questions, we empirically analyze NLI models trained on popular benchmark datasets through a pipeline of evaluations that includes: a zero-shot diagnostic test, inoculation retraining, hypothesis-only sanity check, and cross cross-distribution generalization tests.

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 such as deductive and commonsense reasoning. Our benchmark exposes multiple knowledge gaps in large NLI models regarding diverse linguistic phenomena, particularly in the categories of commonsense and comprehension. For **Q2**, our analysis provides empirical evidence that exposes the lack of recoverable linguistic phenomena in benchmark datasets and models' inability to learn certain linguistic phenomena. We also show that, on some phenomena, models may rely heavily on spurious dataset bias existing in the hypothesis to reach high accuracy. For **Q3**, Our experiments show that models can adapt between distributions with different difficulties only on 22.2% of the phenomena such as Boolean, conditional, and comparative logic. In the majority (58.3 %) of the phenomena, models fail to generalize when the difficulties of the train and test distributions are different, for example, relational knowledge, puns, and contextual commonsense reasoning. A model's learning performance may not align with its generalization ability, suggesting the lack of a general reasoning skill.

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 that can map out which aspects of the problem space remain challenging for existing systems and models.

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 publicly available datasets from peer-reviewed papers. Adversarial NLI (ANLI) is a new benchmark collected "via an iterative, adversarial human-and-model-in-the-loop procedure." (Nie et al., 2020). ANLI is shown to be a more difficult challenge than previous benchmarks. Different from these benchmarks, our work aims to map out and evaluate specific linguistic skills a model needs for language understanding.

Fine-grained NLU Evaluation On top of large-scale benchmarks, there are several works (Joshi et al., 2020; Tarunesh et al., 2021) contributing to the fine-grained analysis of model performance. They collect data examples from existing benchmarks by attaching taxonomic category labels to each data. Or, they build semi-synthetic data allowing analysis on 17 reasoning dimensions. Our data collection and categorization concepts are similar to them. However, our work covers more linguistic phenomena that are difficult but important such as

Category	Description	Phenomena
Lexical	Testing a model's Word-level reasoning skill on lexical semantic, direct, transitive, and compositional lexical relationships.	Lexical Entailment (lex-ent), Named Entity (ner) Hypernymy (hyper), Hyponymy (hypo) Veridicality & Transitivity (transit)
Syntactic	Testing a model's reasoning skill on syntactic structure and compositionality.	Syntactic Alternation (syn-alt), VerbNet (vbn) Syntactic Variation (syn-var), VerbCorner (vbc)
Semantic	Testing a model's reasoning skill on sentence-level reasoning involving diverse semantic properties: entity relations, context, events, subjectivity, and semantic proto roles.	Sentiment (senti), Relational Knowledge (kg-rel), Puns (puns), Semantic Proto Label (sprl) Context Alignment (ctx-align), Coreference (coref)
Logical	Testing a model's reasoning skill on logical operations: propositional structure, quantification, negation, and monotonicity reasoning.	Boolean (bool), Counting (count), Conditional (cond) Comparative (comp), Negation (negat) Monotonicity (monot), Quantifier (quant)
Analytical	Testing a model's knowledge exploitation ability: drawing accurate conclusions based on domain-specific knowledge, symbolic knowledge, and interpretable reasoning steps.	Entailment Tree (ent-tree) Analytical Reasoning (analytic)
Commonsense	Testing a model's reasoning skill on commonsense knowledge independent of cultural and educational background.	Physical (physic), Social (social), HellaSwag (swag) Contextual Commonsense Reasoning (cosmo)
Comprehension	Testing a model's reasoning skill on complex reading comprehension and inference, covering aspects of semantic, context, logic, and numerical	Event Semantics (ester), Discrete Reasoning (drop) Deductive Reasoning (logi) Long Contextual Reasoning (control)
Special	Testing a model's everyday reasoning skill. Including non-monotonic reasoning about valid but defeasible hypothesis from hypothetical context and spatial-temporal reasoning.	Spatial Reasoning (spat), Temporal Reasoning (temp) Defeasible Reasoning (defeas) Counterfactual Reasoning (counter)

Table 1: This table lists the eight categories of linguistic phenomena covered by our dataset collection. We provide a brief introduction for each category describing the types of linguistic skills they intend to evaluate. We also list the dataset names and abbreviations each category contains.

commonsense and non-monotonic reasoning.

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.

Probing Linguistic Knowledge Several works have found evidence that pre-trained models' representations encode knowledge about linguistic phenomena. Tenney et al. (2019) probe contextual representations from four pre-trained language models through the edge-probing method across tasks ranging from syntactic and semantic phenomena. They find that pre-trained models encode rich information on syntactic phenomena but only weakly encode information on semantic tasks compared to non-contextual baselines. Chen and Gao (2021)'s linguistic-information-probing framework extends the edge-probing study by focusing on different semantic phenomena that are important for logical inference in natural language. Their results show that pre-trained contextual embeddings encode more linguistic information on simple semantic phenomena than complex phenomena. Our work is partly motivated by this line of work in which our evaluation is based on the fact that pre-trained models can capture specific linguistic skills from learning.

Other work investigates if models use specific linguistic skills to solve a downstream task. The DNC benchmark (Poliak et al., 2018a) provides a collection of datasets for analyzing if models use distinct linguistic phenomena to conduct natural language inference. Several tasks in our benchmark come directly from this collection. However, our benchmark covers a wider range of linguistic phenomena from more categories than DNC. In particular, our benchmark contains semantic phenomena and includes phenomena from fundamental linguistic properties to complex reasoning types. In addition, our benchmark includes a sys-

tematic evaluation methodology that allows a more in-depth analysis of model behavior.

3 The CURRICULUM Benchmark

3.1 A New Form of Benchmark

Recently, Raji et al. (2021) suggested that good benchmark construction should focus on mapping out a specific set of linguistic skills required for language understanding. They recommend a future benchmark should provide interpretation on how systems work and how they fail on particular aspects of a problem space. Following this suggestion, we propose a new form of benchmark: linguistic-phenomena-driven evaluation. Our main objective is to reformulate the benchmark not simply to be a scoreboard for SOTA model contest but rather as a real measurement and standardization tool for (1) analyzing model performance, (2) exposing model and dataset weakness and (3) providing insights for future research directions.

The curriculum benchmark aims to map out a specific set of linguistic skills required for language understanding. Our benchmark will serve as a diagnostic framework for linguistic-phenomena-driven probing and evaluation. 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. Each linguistic phenomenon is considered one task, and one should train, evaluate, and analyze models on each phenomenon individually. 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

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 tasks that require domain-specific knowledge that we would not expect a model to learn through pretraining, 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. Our motivation for dataset selection is mainly based on the linguistic phenomena categories that we aim to cover which will range from a simple to complex setting.

3.3 Unified Task Format

We unified the task formats into a single linguistic task, Natural Language Inference (NLI). NLI is a task for Natural Language Understanding. The task requires a model to classify the logical relationship between premise and a hypothesis. This logical relationship can either be Entailment (premise is true implies the hypothesis is absolutely true), Contradiction (premise is true implies the hypothesis is absolutely false), and Neutral (one cannot determine if the hypothesis is true or false based on the premise) (Dagan et al., 2013). 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. To convert datasets into NLI format, we follow the standard procedure (Khot et al., 2018). 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

$\mathcal{P}$	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 Table 1

generate non-entailed hypotheses, we use several techniques during recasting. We show more details on our conversion techniques in Appendix C. As a sanity check on our resulting datasets, we empirically find low performance on standard partial-input baselines (Poliak et al., 2018b), suggesting that our conversion yields data of high quality.

3.5 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. We use the predictive $\mathcal{V}$-information (Ethayarajh et al., 2021) as a measurement for dataset difficulty. The $\mathcal{V}$-information can measure how much information an input variable X can provide about Y when constrained to functions $\mathcal{V}$. Intuitively, more usable infromation X can provide, the easier a dataset is for the functions $\mathcal{V}$. Formally, let $\varnothing$ denote a null input that provides no information about Y and $\mathcal{V}$ as a predictive family, we can compute the $\mathcal{V}$-information $I_v(X \to Y)$ as follows:

$$H_v(Y) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[\varnothing](Y)]$$
$$H_v(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[X](Y)]$$
$$I_v(X \to Y) = H_v(Y) - H_v(Y|X)$$

where X, Y denote random variables with sample spaces $\mathcal{X}, \mathcal{Y}$. According to Ethayarajh et al. (2021), $\varnothing$ can be an empty string here as $f[\varnothing]$ models the label entropy. This framework can naturally adapt to the calculation of the point-wise

Name	Model	Train/Test	Accuracy
roberta-mnli	RoBERTa (Liu et al., 2019b)	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.

$\mathcal{V}$-information (PVI) where we measure the difficulty of each data example. Given a training dataset $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^n$, and the predictive family $\mathcal{V}$, the PVI of a data instance $(x, y) \in \mathcal{D}_{train}$ is computed as:

$$\mathrm{PVI}(x \to y) = -\log_2 f[\varnothing](y) + \log_2 f'[x](y),$$

where $\varnothing$ is an empty string (null input) and $\{f, f'\} \subseteq \mathcal{V}$. f' and f are models fine-tuned from $\mathcal{D}_{train}$ and $\{(\varnothing, y_i)|(x_i, y_i) \in \mathcal{D}_{train}\}$ respectively. The $\mathcal{V}$-information framework can also serve as a difficulty measurement for datasets and can be computed explicitly by averaging over PVI:

$$I_v(X \to Y) = \frac{1}{n} \sum_i \mathrm{PVI}(x_i \to y_i)$$

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.

Dataset Controlled Split For our model evaluation pipeline, we are interested in verifying model's ability to learn a generalizable reasoning skill on linguistic phenomena. In particular, we want to check if a model can generalize when its training and testing data distributions have different measurement of difficulty. Thus, we need to conduct controlled split on datasets based on the point-wise difficulty, i.e. the point-wise $\mathcal{V}$-information of their data examples. We first calculate the $\mathrm{PVI}(x \to y)$ for each phenomenon dataset, then we split each dataset into two portions: simple and hard, based on the calculation of each example's PVI.

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. We follow the GLUE diagnostic dataset and use the Matthews Correlation Coefficient (MCC) (Jurman et al., 2012) as the evaluation metric. MCC computes the correlation coefficient of the predicted labels and the true labels. The correlation coefficient value is between -1 and +1. A coefficient of +1 indicates a perfect prediction. A 0 indicates average random prediction A -1 indicates the classifier always miss-classifies. MCC is perfectly symmetric, so it can be used even if the dataset has classes with different sizes.

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 through additional training on a small sample of data.
2. Poor performance: there exists a model weakness to handle the target phenomenon.

Hypothesis-only Bias Analysis We conduct analysis on hypothesis-only bias as (1) a sanity check for our converted datasets and also and (2) a verification on whether model's good performance is from leveraging artifacts in the hypotheses. We train a hypothesis-only baseline (Poliak et al., 2018b) for each phenomenon and compare their performance against the best models from the inoculation experiment. 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. If the hypothesis-only baseline shows good performance, we interpret this as a sign that the datasets contain artifact. If the baseline shows poor performance, it gives evidence that the model is not taking short-cuts.

Cross-Distribution Generalization We conduct the cross-distribution generalization test (Rozen 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 distributions yielding different difficulty levels within the same dataset. For example, we train the model on the simple part of the dataset (data with high $\mathcal{V}$-information) and test it on the hard part (data with low $\mathcal{V}$-information).

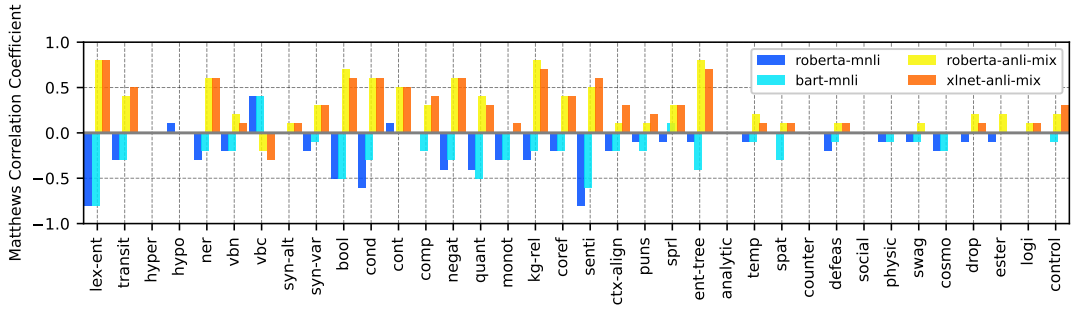
4.1 Experiment Setup

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 are interested in evaluating models with both the single-encoder and the text2text architecture. All models are publicly available from Huggingface (Wolf et al., 2019). 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 $\mathcal{V}$-information-based dataset 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. For all the experiments excluding zero-shot test, we run several turns and select the best performance for analysis.

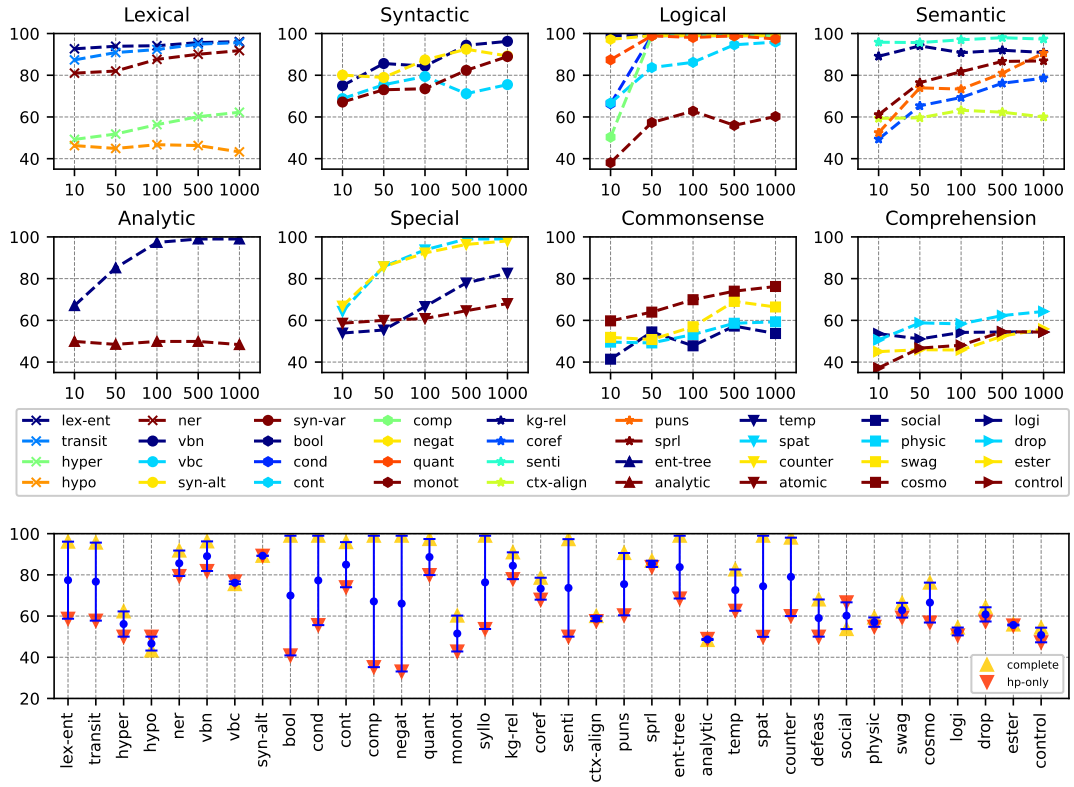
5 Empirical Analysis

5.1 Zero-shot Linguistic Phenomena Diagnose

First, we report the results on zero-shot diagnostic evaluation for each baseline model. From Figure 2a, we observe that both single-encoder and text2text models trained on MultiNLI show a negative correlation in the majority of linguistic phe-



(a) Zero-shot system performance on the CURRICULUM benchmark.



(b) Inoculation by fine-tuning vs. hypothesis-only analysis. 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.

nomena. 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, we find that NLI datasets from common benchmarks lack examples of a diverse set of reasoning skills.

5.2 Inoculation

Based on Figure 2b, 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

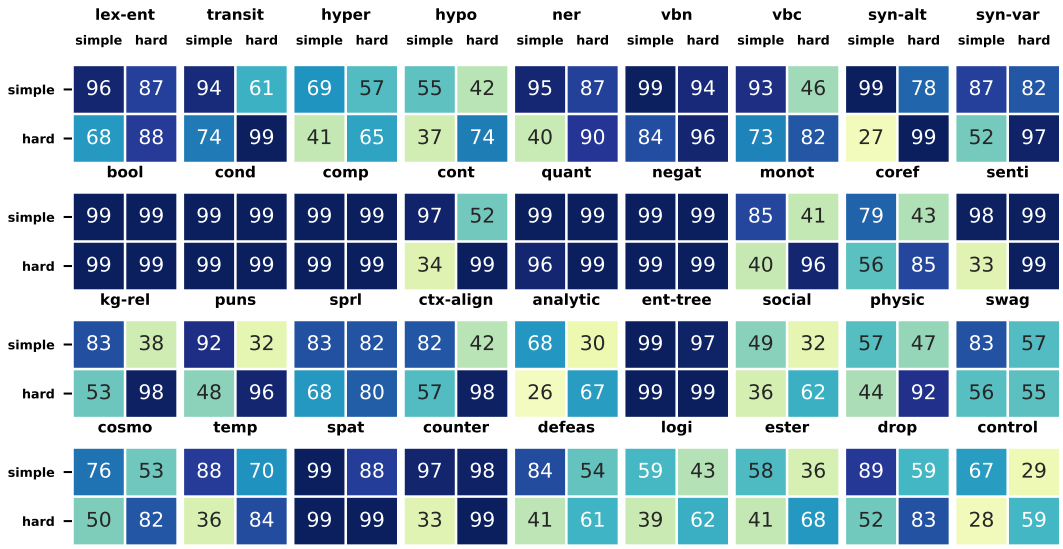


Figure 3: Generalization between controlled dataset splits. Here each heat-map shows the generalization performance of the model fine-tuned and evaluated on different distributions within each linguistic phenomenon.

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 training 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. In summary, our diagnostic study through inoculation exposes a diverse set of dataset and model weaknesses.

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 2b, 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 artifacts from the hypothesis to reach high accuracy.

Also, note that the hypothesis-only baseline performs better for some tasks than the fine-tuned model, which can be interpreted in two ways. When both the baseline and fine-tuned model achieve high accuracy (vbc, syn-alt), higher accuracy on baseline indicates that the hypothesis-only bias is pretty strong in the dataset. When the intervention from the premise is removed (hypothesis-only input), the models can easily exploit the bias to achieve higher accuracy. In contrast, when both the baseline and fine-tuned model achieve low accuracy (hypo, analytic, social, ester), higher accuracy on baseline indicates that the task is very difficult for a model to master successfully. Low baseline accuracy means that the dataset does not contain much bias, so a model must learn the correct reasoning to perform well. However, the fine-tuned model has even worse performance than the baseline, meaning that it fails to learn the skill required for these tasks. Our main finding here is

that good performance on a linguistic phenomenon dataset does not mean the model captured the associated phenomena. The model can learn short-cuts through hypothesis-only bias and artifacts.

5.4 Generalization

As Figure 3 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 generalizable. 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 sufficiently for generalization.

6 Conclusion and Future Work

In this paper, we introduce a new form of benchmark that can serve as an effective tool for evaluating and analyzing model outcomes. We propose a linguistic-phenomena-driven benchmark that aims to diagnose neural language models to discover types of linguistic skills that remain challenging to models. We compiled a dataset collection covering 36 types of linguistic phenomena ranging from fundamental linguistic properties to complex reasoning skills. In addition, we define an evaluation procedure that can provide an in-depth analysis of model and dataset weaknesses. Using our benchmark, we comprehensively study how well language models capture specific linguistic skills essential for understanding. 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.

Overall, our benchmark effectively evaluates a model on specific linguistic skills and exposes a list of model and training data weaknesses. We hope that our benchmark and empirical findings can encourage the communicate to rethink dataset construction and model architecture design. In particular, we hope to encourage the the development of new datasets that cover richer types of linguistic phenomena and language models that can learn essential linguistic skills that are generalizable. For future work, we plan to add more datasets to cover more phenomena such as psycholinguistics (Laverghetta Jr. et al., 2021). We envision our benchmark to be dynamic, meaning that a higher-quality and more difficult dataset for a phenomenon should replace the current ones in the future. For example, the StepGame benchmark (Shi et al., 2022) provides better data for spatial reasoning, which can replace the current spatial reasoning dataset. We also plan to explore new learning methods to help models overcome the weakness of learning non-generalizable skills, such as calibration through symbolic loss functions.

Acknowledgment

We thank the anonymous reviewers for their thoughtful and constructive comments. We thank Kyle Richardson from AI2 for his insights and suggestions on improving our camera-ready version. Thanks also to our advisors Laurence S. Moss and Michael Wollowski for their feedback on earlier drafts of this work. Special thanks to the Machine Learning for Language Group at NYU for their wonderful NLP toolkit, JIANT (Phang et al., 2020).

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.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Zeming Chen and Qiyue Gao. 2021. Probing linguistic information for logical inference in pre-trained language models.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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.

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.

Giuseppe Jurman, Samantha Riccadonna, and Cesare Furlanello. 2012. A comparison of mcc and cen error measures in multi-class prediction. *PLOS ONE*, 7(8):1–8.

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.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Antonio Laverghetta Jr., Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2021. Can transformer language models predict psychometric properties? In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 12–25, Online. Association for Computational Linguistics.

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.

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

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

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.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. http://jiant.info/.

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.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Lonli: An extensible framework for testing diverse logical reasoning capabilities for nli.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

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` → `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. Our model training and testing pipeline is modified from the JIANT toolkit. We mainly adapted several components on classes and functions involving task, dataset, reprocessing, tokenization, model version control, and evaluation metrics. All our experiments are implemented with models publicly available from Huggingface Transformers (Wolf et al., 2020)[2].

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 3 and 5. 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 **??**. We use the AdamW (Loshchilov and Hutter, 2019) as our optimizer.

[2] https://github.com/huggingface/transformers

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.