CURRICULUM: Diagnosing Neural Language Models through Broad-Coverage Linguistic Phenomena

Zeming Chen Qiyue Gao

Advisors: Michael Wollowski, Larry Moss (Indiana University)

Motivation

Existing large-scale NLP benchmarks and evaluation methods show significant limitation:

- Do not map out specific linguistic skills required for understanding.
- Do not evaluate language models' reasoning ability on linguistic phenomena.
- Do not verify language models' generalization ability across distributions.

Contributions

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?

We propose **CURRICULUM**, a broad-coverage diagnostic framework for linguistic-phenomena-driven probing and evaluation. We make the following contributions:

- A large empirical study on Models' specific linguistic skills and their ability yo learn theses skills.
- A linguistic-phenomena-driven diagnostic framework including (1) a dataset collection covering 36 types of linguistic phenomena, (2) an evaluation procedure for probing and analysis.
- A meta in-context learning paradigm based on linguistic phenomena for in-context generalization to complex linguistic tasks.

CURRICULUM Diagnostic Dataset Collection

Category	Description	Phenomena
Lexical	Testing a model's Word-level reasoning skill on lexical semantic, direct and transitive lexical relationships.	Lexical Entailment, Veridicality, Transitivity Hypernymy, Hyponymy, Named Entity
Syntactic	Testing a model's reasoning skill on syntactic structure and compositionality.	Syntactic Alternation, Syntactic Variation VerbNet, VerbCorner
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, Relational Knowledge, Puns, Semantic Proto Label Context Alignment
Logical	Testing a model's reasoning skill on logical operations: propositional structure, quantification, and monotonicity.	Boolean, Comparative, Conditional, Counting Monotonicity, Negation, Quantifier
Analytical	Testing a model's knowledge exploitation ability: drawing accurate conclusions based on domain-specific knowledge, symbolic knowledge, and interpretable reasoning steps.	Entailment Tree Analytical Reasoning
Commonsense	Testing a model's reasoning skill on commonsense knowledge independent of cultural and educational background.	Physical, Social, HellaSwag Contextual Commonsense Reasoning
Comprehension	Testing a model's reasoning skill on complex reasoning types targeted by different reading comprehension challenges.	Event Semantics, Discrete Reasoning Deductive Reasoning, Contextual Reasoning
Special	Testing a model's reasoning skill on non-monotonic and spatial-temporal reasoning.	Spatial Reasoning, Counterfactual Reasoning Defeasible Reasoning, Temporal Reasoning

Dataset Difficulty Split

- We split each dataset(X, Y) by the Point-wise V-information (PVI) into simple and hard parts.
- The PVI of a data example is defined as:

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

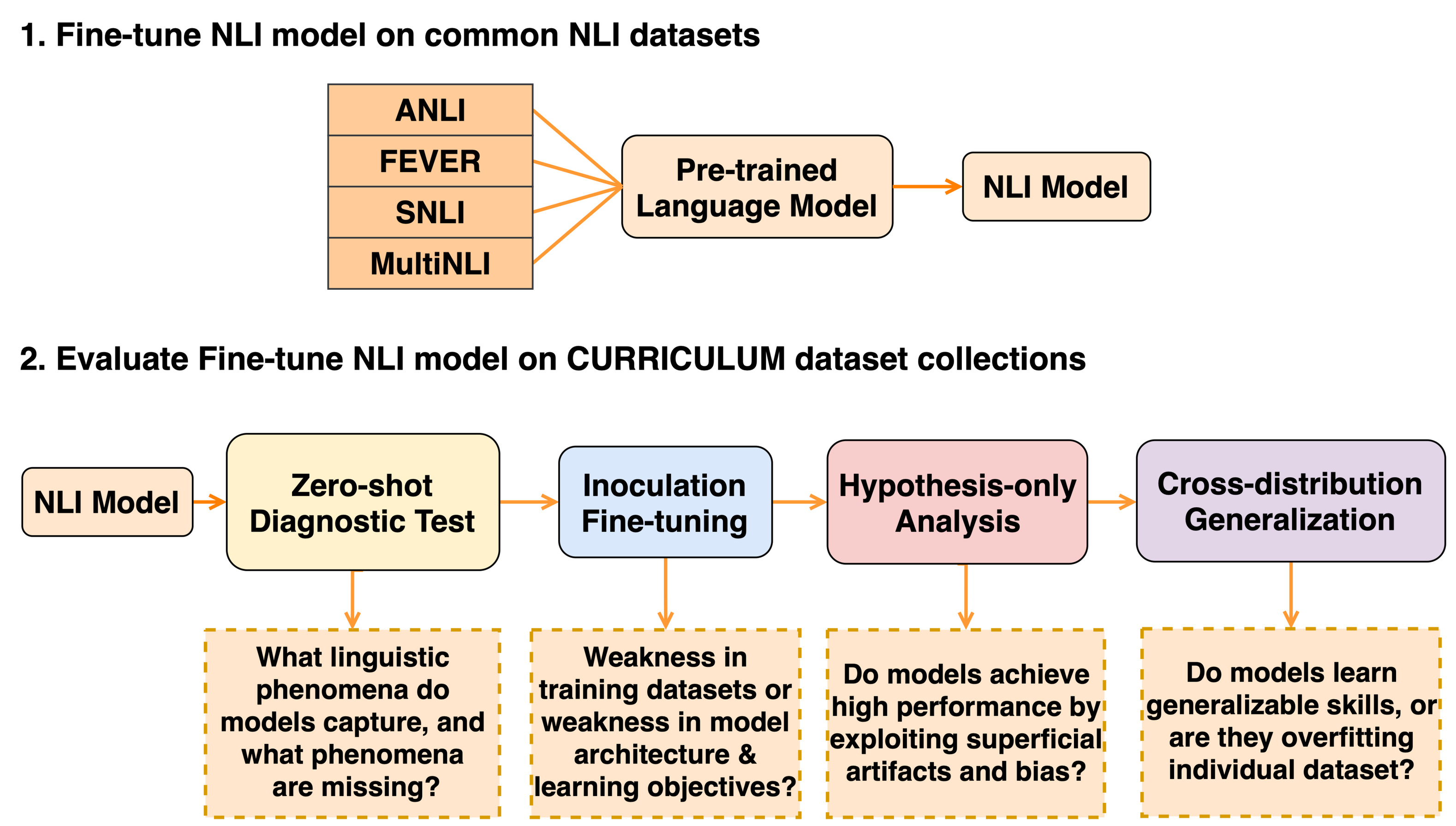
- The overall difficulty of a dataset is the average PVI over all the examples (x_i, y_i):

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

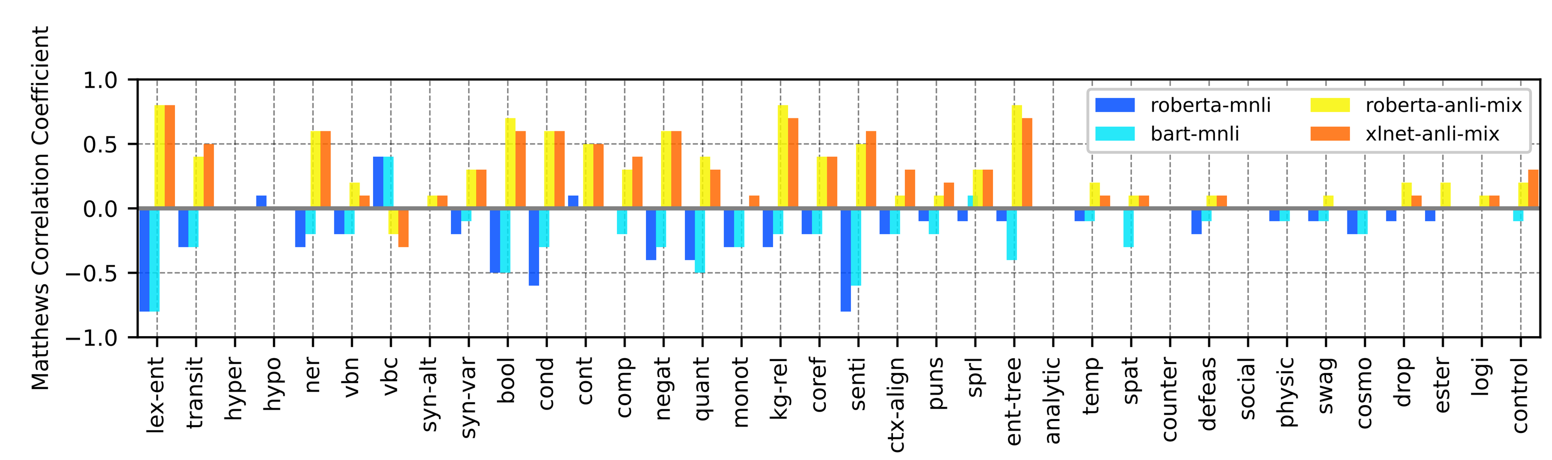
Targeted Pre-trained Language Models

Name	Model	Train/Test	Accuracy
roberta-mnli	RoBERTa	MNLI/MNLI	90.2%
bart-mnli	BART	MNLI/MNLI	90.9%
roberta-mnli-mix	RoBERTa	SNLI,MNLI,FEVER,ANLI/ANLI	53.7%
xlnet-anli-mix	XLNet	SNLI,MNLI,FEVER,ANLI/ANLI	55.1%

CURRICULUM Evaluation Procedure



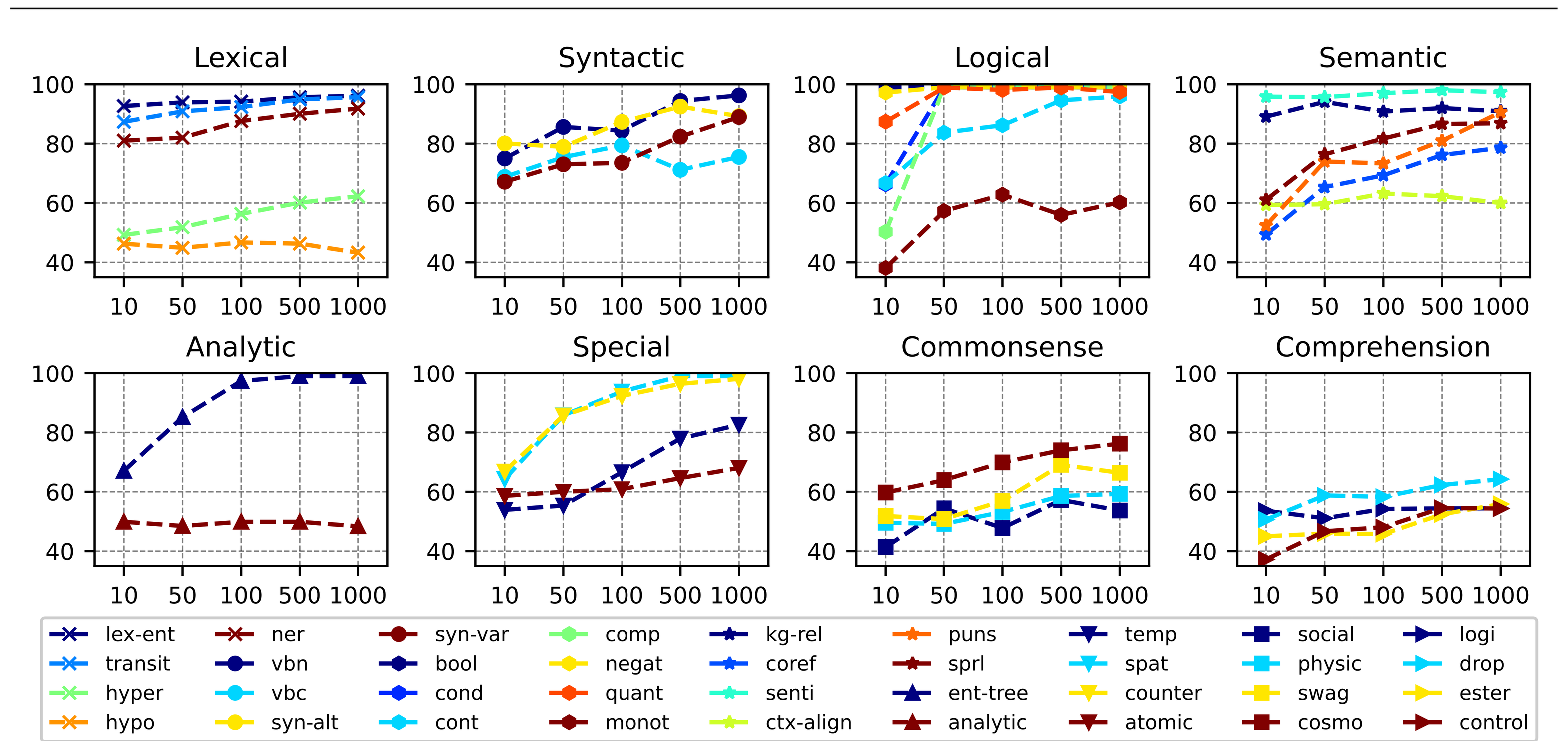
Zero-shot Diagnostic Test



- Models trained on more datasets capture more types of linguistic skills
- More training datasets fail to help models capture hard phenomena

Finding 1
Common benchmark datasets lack examples of a diverse set of reasoning skills.

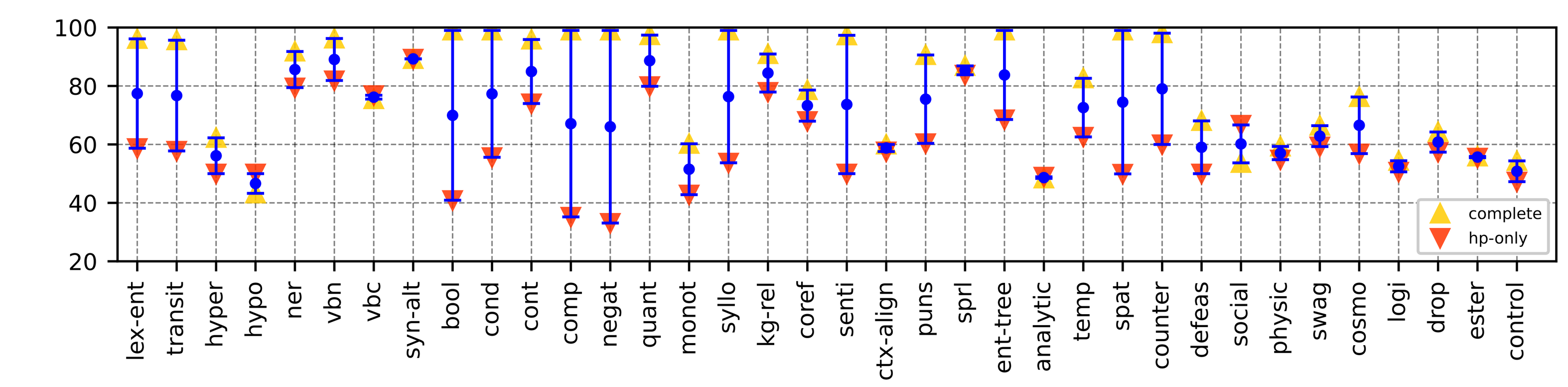
Inoculation Fine-tuning



- Most of the fundamental and simple phenomena are recoverable through inoculation fine-tuning on additional data suggesting dataset incompleteness of the original training data.
- Model did not benefit from inoculation for hard and complex phenomena meaning that there exists model weaknesses on handling hard reasoning tasks.

Finding 2
Our diagnostic study exposes a diverse set of dataset and model weaknesses.
They can drive future research on new model architectures and learning objectives.

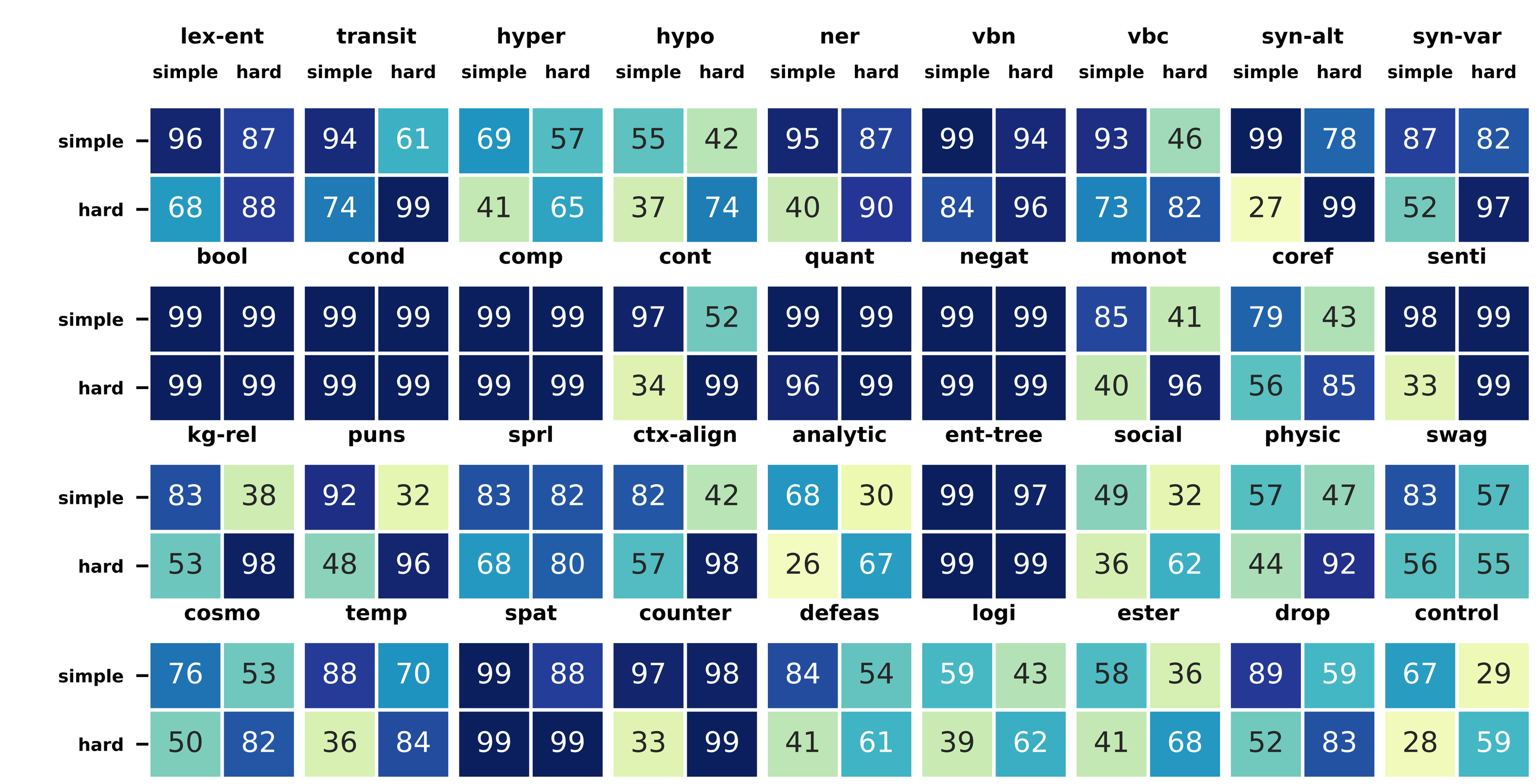
Hypothesis-only Bias Analysis



- Hypothesis-only baselines perform poorly on most phenomena
- 30.6% high performance gain and 36% trivial performance gain

Finding 3
Good performance on a dataset does not mean the model captured the associated phenomena.
The model can learn short-cuts through hypothesis-only bias and artifacts.

Cross-Distribution Generalization

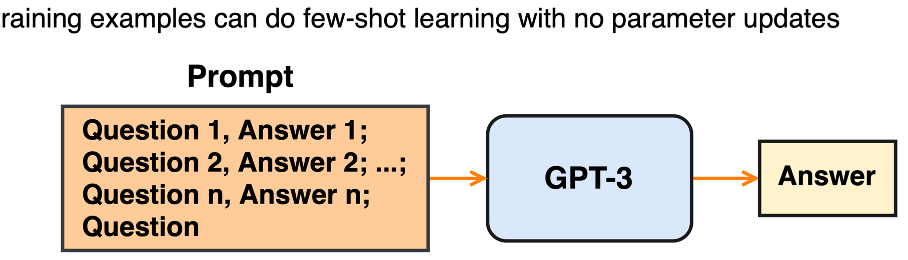


- On 87.8% Phenomena: model fails to learn a general reasoning skill
- On 22.2% Phenomena: Learned a general reasoning skill

Finding 4
Language Model tends to overfit a distribution and fail to learn a general reasoning skill.

Future Work (In progress)

- **In-context Learning:** a language model conditioned on a concatenation of training examples can do few-shot learning with no parameter updates



- **Meta In-context Learning**: train a model on many tasks using in-context learning to learn how to in-context learn on out-of-distribution tasks.
- **Neural Task Scheduler**: selects the optimal set of meta training tasks, updated by bi-level optimization with gradient approximation from the meta model.
- **Example Selector**: select optimal examples for in-context learning, updated by a data-manipulation model learned with the model simultaneously,