Principles from Clinical Research for NLP Model Generalization

Aparna Elangovan[1], **Jiayuan He**[2,1], **Yuan Li**[1,2], **Karin Verspoor**[2,1]
[1]The University of Melbourne, Australia
[2]RMIT University, Australia
aparnae@student.unimelb.edu.au
{jiayuan.he, yuan.li, karin.verspoor}@rmit.edu.au

Abstract

The NLP community typically relies on performance of a model on a held-out test set to assess generalization. Performance drops observed in datasets outside of official test sets are generally attributed to "out-of-distribution" effects. Here, we explore the foundations of generalizability and study the factors that affect it, articulating lessons from clinical studies. In clinical research, generalizability is an act of reasoning that depends on (a) *internal validity* of experiments to ensure controlled measurement of cause and effect, and (b) *external validity* or transportability of the results to the wider population. We demonstrate how learning spurious correlations, such as the distance between entities in relation extraction tasks, can affect a model's internal validity and in turn adversely impact generalization. We, therefore, present the need to ensure internal validity when building machine learning models in NLP. Our recommendations also apply to generative large language models, as they are known to be sensitive to even minor semantic preserving alterations. We also propose adapting the idea of *matching* in randomized controlled trials and observational studies to NLP evaluation to measure causation.

1 Introduction

What factors lead to poor generalizability of models? To understand causes of non-generalization, the notion of generalizability needs to be first clearly defined. One definition of generalizability, in the context of general machine learning, is the ability of a model to perform well on data *unseen during training*, but *drawn from the same distribution or population* (Google, 2022). Data *unseen during training* clearly refers to data that is not part of training data, and is arguably uncontroversial to support sound evaluation. However, the requirement of data *drawn from the same distribution or population* warrants more scrutiny.

It is important to consider the impact of data distribution on generalization, yet this is very challenging for natural language data. In statistics, the concept of distribution indicates "the pattern of variation in a variable or set of variables in the multivariate case", and thus describes the frequency of values of an observed variable (Wild, 2006). Distribution or frequency of observed variable values can be challenging notions to meaningfully adapt to high dimensional, multivariate, and high variability text data. Currently, no formal definition of in-distribution or out-of-distribution (OOD) texts, or how to detect them, exists for NLP (Arora et al., 2021). Where distributions are considered, simple surface linguistic characteristics or lexical-level distributions are emphasized (Verspoor et al., 2009).

Despite the lack of a comprehensive formal definition of OOD in existing NLP literature, OOD has become the most commonly cited reason for generalization failure when a model performs poorly outside an official test set. Due to the black-box nature of deep learning models, it is increasingly difficult to demonstrate if a model has established a robust decision-making process that is generalizable to unseen data. As a consequence, generalization failures are typically ascribed to external factors, i.e. those extraneous to model development practices, and primarily to shifts in data distribution, or OOD.

Recent studies have emerged showing that the robustness of a model can be undermined by inadvertent errors made during development of a model. For example, it has been shown that data leakage from training into test splits can lead to inflated test results (Elangovan et al., 2021). Other works have pointed out spurious correlations in various benchmark datasets that may be leveraged by deep learning models to achieve inflated performances on test sets, while having poor generalization capability to real-world settings (Gururangan et al., 2018; McCoy et al., 2020; Shinoda et al., 2022).

In this paper, we hope to draw attention to the

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2293–2309
June 16-21, 2024 ©2024 Association for Computational Linguistics

causes of NLP model generalization failures, especially those internal factors that are part of the model development process (*e.g.* preparation of training data). We argue that the external validity of a model should only be examined if the internal validity of the model is established. We start with a case study, showing how generalization failures can be caused by internal factors – the model has learned surface patterns in training data. We then propose that a pragmatic notion of model generalizability in the NLP domain can be established through borrowing and adapting practices from a domain seemingly far afield – clinical research.

The contribution of this paper is two-fold. First, we show how OOD may not be the sole cause of generalization failures, via a relation extraction task, highlighting the need for intrinsic investigation of why models fail. Second, we propose to categorize the causes of generalization failures in NLP models, drawing inspiration from clinical studies. Our work provides guidance on how to more systematically analyze generalization failures and adapt the principles behind randomized controlled trials for NLP model evaluation.

2 Relation extraction case study

Through a relation extraction case study over two data sets, we demonstrate that poor generalization in real-world application can result from ineffective modelling, *i.e.* learning of superficial surface patterns, rather than data distribution shift. For completeness, we also study a popular benchmark dataset on natural language inference (NLI) task.

2.1 Approach

Due to the black-box nature of deep learning models, it is difficult to interpret the underlying basis for model predictions. Inspired by recent works in explainable NLP models, such as LIME (Ribeiro et al., 2016), we employ interpretable surrogate models to examine the behavior of deep learning models. Specifically, assume a dataset $\mathcal{D} = \{\langle x, y \rangle\}$, where x represents the input sequence of words and y represents the ground-truth label for x. We train two surrogate models: a model S_g that is fit on the dataset $\langle x_u, y \rangle$ and another model $S_{\hat{m}}$ that is fit on the dataset $\langle x_u, \hat{y}_m \rangle$, where x_u represents x via a representation technique u and $\hat{y}_m$ is the prediction of the deep learning model B_m to be examined, e.g. a BERT-based model (Devlin et al., 2019). For u, we adopt a vector of surface patterns.

We hypothesize that a strong correlation between the predictions of the surrogate $S_{\hat{m}}$ and the corresponding main model predictions indicates that the underlying model B_m has relied on the surface patterns in x_u, and that the model's predictions may not be reliable. The idea here is that if a surrogate model can reproduce the behavior of a comparator model with high fidelity, then that surrogate model is a good approximation of that comparator model and hence there is no evidence that the comparator model has learned anything more than the patterns captured in the surrogate model. This also follows from the principle of Occam's razor related to the law of parsimony (Epstein, 1984; Felsenstein, 1983). We interpret the correlation between the predictions of $S_{\hat{g}}$ and the ground truth labels as the indicator of the extent to which the surface patterns x_u are present in the underlying data. A strong correlation indicates the weakness in the dataset itself and how these patterns can be exploited to achieve highly accurate predictions without deeper linguistic comprehension. We use Cohen's Kappa κ to measure correlations.

2.2 Datasets

We use the following datasets:

- **PTM-PPI (PTM)** is sampled from PubMed abstracts (Elangovan et al., 2022) for relation extraction (REL) task, annotated with 6 types of post-translational modification relationship between two proteins. Out of the 6 positive classes, we only consider the class "phosphorylation" since only this class has a sufficient number (> 100) of training samples. Consequently, the dataset used has 2 classes: "phosphorylation" class and the negative class.
- **ChemProt (CHM)** is sampled from PubMed abstracts, annotated with 5 types of protein–chemical relationships (Krallinger et al., 2017) for REL task. The dataset contains 6 classes in total: 5 positive classes and the negative class.
- **SNLI (SNL)** is a NLI task dataset with 3 classes (Bowman et al., 2015).

For CHM and PTM, we fine-tune a BioBERT model (Lee et al., 2019). For SNLI dataset, we fine-tune a BERT (Devlin et al., 2019) model.

2.2.1 Generalization data sets

To understand the impact of generalization behavior of a model that has relied on spurious factors,

Dataset	Split (Label)	# Pos / # Neg
PTM	TR (GT)	139 / 1116
	TS (GT)	44 / 308
	TS (MP)	24 / 328
	GH (MP)	250 / 5000
CHM	TR (GT)	4172 / 2265
	TS (GT)	3469 / 2275
	TS (MP)	3726 / 2018
	GH (MP)	7500 / 2500
SNL	TR (GT)	366374 / 182764
	TS (GT)	6605 / 3219
	TS (MP)	6462 / 3362

Table 1: Summary of data sets. TR: official training set; GT: ground-truth labels are used; TS: official test set; MP: model prediction labels, GH: generalization set.

beyond the test set, we selected the REL tasks due to the availability of data on PubMed. We select a random subset of PubMed abstracts to create a "generalization" set. We apply the fine-tuned BioBERT model to generate the predictions for the "generalization" set. From this generalization set, we randomly sample from the top 25 percentile high confidence predictions to form set GH for each class (Elangovan et al., 2022) and report results across 10 runs. Selecting only the high confidence predictions follows Hendrycks and Gimpel (2017), which demonstrated that the prediction probabilities of OOD samples tend to be lower than those of correct samples. Table 1 summarizes the datasets.

2.3 Surrogate models and surface patterns

We employ two explainable surrogate models:

- **Multinomial Naive Bayes (NB)**: The Multinomial Naive Bayes approach represents input samples using simple surface patterns (n-grams, $n = 1$). To avoid over-fitting, we select the top k most commonly n-grams per class. Hence, the maximum number of n-grams that the NB model uses is $k\times$ number of classes. We set k as 100 in the experiment.
- **Naive Bayes + Decision Tree (NB-T)**: Here we use model stacking, where a Decision Tree is stacked on top of a NB model. In NB-T, the prediction of NB is used as one feature input in the subsequent decision tree. Additionally, handcrafted rules, detailed in section 2.3.1, are used as surface pattern features in the decision tree. To void over-fitting and to allow the decision process to be explainable, we restrict the tree depth to $<= 4$.

2.3.1 Crafted surface pattern features

To study whether the BioBERT-based model has relied on distance-based surface patterns for relation extraction, we manually analyze BioBERT's predictions on the PTM task and identified 4 surface patterns that potentially explain the model predictions. We then use these hand-crafted surface patterns to represent the inputs to a surrogate model and verify how well the predictions of the surrogate model correlates with BioBERT's predictions (which is given the full-text input). The 4 surface patterns are:

- **Percentage count of participating entities (E1C and E2C) :** Given an input sentence s and a relation $\langle E1, Rel, E2\rangle$ in s, this feature captures the percentage of the total tokens in s corresponding to each entity. For instance, for the input *"GENE_A interacts with CHEMICAL_C and binds to CHEMICAL_C and CHEMICAL_D"* and the relation ⟨CHEMICAL_C, Rel, GENE_A⟩, the features E1C and E2C are $\frac{2}{10} * 100 = 20.0$ and $\frac{1}{10} * 100 = 10.0$, respectively (input contains 10 words, E1 occurs twice and E2 once).
- **Length of shortest span containing the participating entities and a given trigger word T (LSS_⟨T⟩):** This feature represents the length of the shortest span containing the two entities and a specified trigger word. For instance, an input *"GENE_A interacts with CHEMICAL_C and binds to CHEMICAL_C and CHEMICAL_D"* and ⟨CHEMICAL_C, Rel, GENE_A⟩ the length of the shortest span containing the trigger word "interacts", LSS_interacts, is 4, whereas LSS_binds is 6.
- **Length of shortest span that contains the entities and any trigger word (LSS):** For instance, given input *"GENE_A interacts with CHEMICAL_C and binds to CHEMICAL_C and CHEMICAL_D"* and relation ⟨CHEMICAL_C, Rel, GENE_A⟩, the shortest span has length 4.
- **Fraction of sentences containing participating entity pair (SPC):** This feature represents the normalized count of sentences containing the entity pair. For instance, if the input text contains s_n sentences and only k sentences contain both entities E1 and E2 in $\langle E1, Rel, E2\rangle$, then this feature would be $\frac{k}{s_n}$.

For SNLI, we use the spurious factors reported by Gururangan et al. (2018), such as the length of the hypothesis and presence of negation.

2.4 Results

Table 2 reports the Kappa correlation between **(a)** the surrogate models and ground truth; and **(b)** the surrogate models and fine-tuned model's prediction. For the PTM corpus, NB-T achieves better correlation with ground-truth, compared to NB, in all settings, demonstrating that the hand-crafted surface patterns are more likely than n-grams along to be influencing model predictions. In addition, when ground-truth labels are used as targets, i.e. TS (GT) *vs* S_g^{test} predictions and TR (GT) *vs* S_g^{Train} predictions, NB-T correlation κ is 0.55 and 0.54 respectively, indicating that similar surface patterns exist in both the training and test sets of ground-truth labels. This is not surprising given that the test and train sets were obtained using a random split from a single dataset.

To examine if the fine-tuned BioBERT model has relied on those hand-crafted surface patterns, we compare the correlation between NB-T on the test set when fitting to the ground-truth labels, TS (GT), versus the BioBERT model's predictions, TS (MP). We see that the κ correlation increases drastically from 0.55 – weak correlation, to 0.73 – moderate correlation (based on McHugh (2012) ranges), when the target labels are replaced with BioBERT predictions (cf. trees in Appendix A). This demonstrates that NB-T using handcrafted features is more correlated with BioBERT's predictions compared to ground-truth labels themselves, increasing the evidence that the BioBERT model may rely on these features. This phenomenon is further exacerbated on the GH set, where NB-T achieves κ correlation of 0.85 - strong correlation when fitting to BioBERT's predictions. In fact, Elangovan et al. (2022) report that only 6 out of 30 (20%) of the phosphorylation predictions turned out to be accurate when the high confidence predictions were randomly sampled and verified by experts, compared to test set precision of 62.5%. The drop in precision, compared to test set performance, *may be* explained by the model relying on these surface patterns rather than broadly generalizable features.

DS (L)	SM	PTM κ	CHM κ	SNL κ
TR (GT)	NB	0.33	0.45	0.25
	NB-T	0.54	0.46	0.27
TS (GT)	NB	0.26	0.48	0.29
	NB-T	0.55	0.48	0.33
TS (MP)	NB	0.25	0.50	0.32
	NB-T	**0.73**	0.51	0.34
GH (MP)	NB	0.77 (0.3)	0.44 (0.002)	-
	NB-T	**0.85** (0.3)	0.44 (0.002)	-

Table 2: Surrogate model correlations on dataset (DS) and the target label (L): The surrogate model (SM) NB-T correlates better with BioBERT's prediction than the ground truth labels for the PTM dataset. For GH set, we report standard error for 10 runs ($\frac{\sigma}{\sqrt{n}}$, where $n = 10$). The p-value for Cohen's-κ is less than 0.05.

In the CHM dataset, distance based surface patterns do not seem to be a stronger predictor than n-grams, as shown in Table 2. All the surrogate models have a correlation between 0.4 and 0.5 indicating weak correlation (McHugh, 2012).

For SNLI, the surrogate models achieve minimal correlation between 0.21 and 0.39 (McHugh, 2012) indicating there are potentially other features required to improve the surrogate model. The hypothesis length, as reported by Gururangan et al. (2018), is one of the key features that NB-T also identifies, see details in Appendix B.

OOD is NOT always a sufficient explanation for generalization failures. Given the broad and generic definition of OOD, almost any sample can be categorized as OOD. It is difficult to counter the OOD argument, as there is no comprehensive approach to establishing that a given instance is in-distribution (Arora et al., 2021). While in the case of CHM, we were unable to detect clear surface patterns, in the case of the PTM dataset, BioBERT *appears to heavily rely on surface patterns* reflected in our handcrafted distance-based patterns. The strong correlation between the surrogate model and the BioBERT-based model prediction points to the model potentially relying on such surface patterns, which will undoubtedly hinder the model's generalizability to external datasets. Therefore, before concluding OOD as a potential cause of generalization failure, we need to ensure that spurious correlations are NOT the source of a model's high performance. There are cases where some surface patterns might be reasonable for some tasks, we discuss such cases in detail in Section 4.3. While we acknowledge that correlation does not necessarily imply causation, under certain conditions it may indeed (Gardner, 2000). Our results above and several other works suggest that spurious factors might be enabling models to achieve high performance (Gururangan et al., 2018; McCoy et al., 2019), but further (difficult to design) tests would be required to unambiguously establish that *the cause* of high performance is indeed spurious correlations.

Detecting spurious correlations is non-trivial, while simple surrogate models can detect dominant surface patterns provided we know what to look for apriori. Hence, our approach of using surrogate models has two main challenges: **a)** it requires good handcrafted features, and **b)** it assumes only a few dominant patterns exist. Deep learning models, such as transformers, that can potentially learn thousands of low frequency surface patterns (that may not be detectable by simple models), while these patterns may also be present in the test set, leading to inflated test performance.

3 Foundations of generalizability

The core idea behind generalizability is that the conclusions drawn, or a model inferred, from a sample can be applied to a wider population. In this section, we discuss some of the foundations of generalizability from clinical studies and why generalization failures cannot be solely attributed to out-of-distribution factors. Identifying the cause of generalization failures requires several components to be well-defined ahead of the experiments, which are discussed in detail in this section.

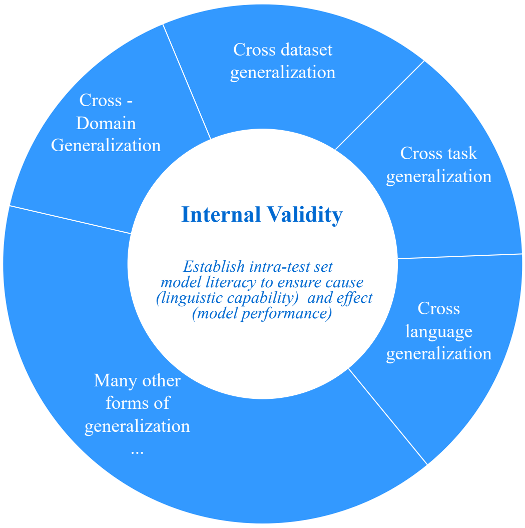
Cross dataset generalization
Cross - Domain Generalization
Cross task generalization
Internal Validity
Establish intra-test set model literacy to ensure cause (linguistic capability) and effect (model performance)
Cross language generalization
Many other forms of generalization ...

Figure 1: Internal validity is a mandatory precursor for any form of external generalization, including cross dataset generalization. Internal validity is required to ensure that the model has learned core linguistic strategies to solve the task within the context of the test set.

3.1 The notion of generalizability

Clinical studies aim to answer questions such as *"is this drug treatment effective?"*, and critically rely on generalizability to establish meaningful evidence (Rothwell, 2005; Guyatt et al., 2011b; Schünemann et al., 2013). A core element of a clinical study is the specification of the study population, referring to a subset of the population selected for research; it is impossible to study the entire population (Kukull and Ganguli, 2012). The implicit assumption is that conclusions drawn from the sample are applicable to the population, requiring the population boundary to be defined. This boundary depends on the aim of the study, can include various factors, including country, insurance memberships or disease status (Kukull and Ganguli, 2012). For instance, for a study that investigates the effects of a drug on a disease, the relevant population would typically be all the people with that disease. To ensure that any conclusions from the study using the samples drawn from the population are confidently generalizable to the entire relevant population, *intrinsic* and *extrinsic validity* of experiments (Guyatt et al., 2011a) must be established. We detail these concepts below.

3.2 Study Aim

In machine learning or more specifically NLP, the aim of a study might seem obvious, for example, to establish whether one model is better than the other for a given task according to a chosen metric. However, even such a straightforward research goal can be ambiguous, as the conclusion drawn depends on the dataset used to evaluate the models. Hence, if the study aim is well-defined or constrained, e.g. referring to performance on a benchmark such as GLUE (Wang et al., 2018), then the objective is clear, allowing the conclusions to be contextualized. In real-world settings, the objective is usually to know if the model can meet certain performance objectives, e.g. will the model's predictions be at least 70% accurate when deployed. As a consequence, it becomes pertinent to define the population boundary or the context to ensure that the model performance is optimal for the task it is designed for in that broader context. Hence, defining the aim of the study requires careful definition of the target population.

3.3 Defining populations

In the context of machine learning, if a test set is a sample that effectively represents a population, then the conclusions based on the test set should apply to the entire population. For instance, if the conclusion is that the model has an accuracy of

90% based on the test set, it should mean that when the model is applied to the entire population, the prediction accuracy ideally should also be $\sim 90\%$.

In NLP, for the datasets used to benchmark model performance, such as GLUE (Wang et al., 2018), the broader "population" corresponding to these datasets is not clearly defined, and hence it is difficult to define the boundary or context of generalizability, *i.e.* the population that these results are meant to apply to or when data is out-of-distribution (OOD). As a result, when a model performs poorly on a different test set, an explanation that the data is OOD is insufficient unless the notion of distribution is clearly defined. We cannot simply attribute poor performance to OOD data. Model performance actually depends on **(a)** what the model has learned, and **(b)** how effective a test set is in measuring its performance. Moreover, OOD data should ideally have lower model confidence scores (Hendrycks and Gimpel, 2017). Thus, claims of OOD as an explanation for poor performance at the very least require the context of population or distribution to be clearly defined.

When we define a new task, we implicitly define the population boundary for the task, such as through the use of data cards (Pushkarna et al., 2022). For instance, if our task is to analyze the sentiment of IMDB movie reviews, IMDB movie reviews are the population of texts that this model applies to. However, defining the boundary of this population is not trivial. It may require additional constraints around language, written vs. spoken style, and more precise specification of the domain, such as -restaurant vs. movie reviews.

An effective, well-informed boundary should consider key factors that can impact the performance of a model in a real-world scenario and constrains the problem space so that the samples can be drawn from the population that the model is meant to serve. Hence, defining the population boundary is a mandatory precursor to collecting training and test data to ensure that the collected samples are representative of the population.

3.4 Internal validity

Internal validity is crucial to ensure that the measurement of the relationship between *cause and effect* is not affected by spurious correlations or bias in the data (Delgado-Rodríguez and Llorca, 2004). This particularly affects what we can infer from a gain in model performance.

To understand *internal validity* in the NLP context, consider a hypothetical example of a customer sentiment analysis text classification task. To collect data, we may randomly select 500 customer emails from organization A (org-A), and another 500 from organization B (org-B). Let's assume that org-A generally provides better customer service than org-B, and that the samples from org-A contain a signature marker, "FROM-ORG-A". Say that a deep learning model that requires no feature engineering has over 90% accuracy, while another simpler model based on carefully curated semantic features achieves a performance of 75%. The software code is well tested, and the researchers also perform statistical significance tests and conclude that the deep learning model is better than the simpler model. However, the deep learning model has, in fact, relied on the signature "FROM-ORG-A" as a key indicator for positive labels, while the simpler model relies on the presence of words such as "great", "mediocre", etc. to differentiate between classes. Is the conclusion that the deep learning model is better than the simple model at customer sentiment analysis internally valid? The same parallels can be drawn from the case study of the PTM dataset discussed above, where the model seems to have relied on spurious correlations. Hence, the performance on a test set need not indicate that the model has the **basic linguistic task level literacy** even within the limited scope of the test set, as depicted in Figure 1.

Internal validity of experiments can also be affected by factors such as sample selection and instrumentation (Wortman, 1983). Internal validity is to ensure that the study and the conclusions are valid within the context of the experiment, where the cause (model has learned the right aspects of the language) and effect (model's performance) is fairly evident. Experimental errors such as bugs in the code, issues with test/training split resulting in data leakage (Elangovan et al., 2021) are obvious examples of errors that invalidate results. Factors such as dataset bias, data splits and test data issues affect reproducibility of experiments (Gundersen et al., 2023) also affect internal validity. The internal validity can also be affected by the selection of participants in the study (Patino and Ferreira, 2018). The participants of a study in NLP can be construed as the data and any human annotators. Lack of careful consideration of details such as training or test sample size, sample selection criteria etc. can make the study lack internal validity.

Performance gains made by large language models can be misunderstood as natural language understanding (Bender et al., 2021). Geirhos et al. (2020) also emphasize the need to differentiate the capability required to perform on a dataset cf. the underlying capabilities of a model. Robustness of experimental design, data selection criteria, well-tested code, careful train-test split, effective test sets, statistical analysis, etc. are core aspects of internal validity in NLP.

3.5 External validity

External validity, associated with *transportability* of results from samples to the wider population, is a heavily debated topic even in clinical research, whereas internal validity is a more established concept in clinical studies (Tipton, 2014; Yarkoni, 2022; Degtiar and Rose, 2023). If a study is not internally valid, then external validity is irrelevant (Patino and Ferreira, 2018). Assuming study results are internally valid, whether the conclusions of a study are generalizable or transportable to the wider population depends on the ability to separate “relevant" from “irrelevant" facts for the study. Importantly, well-designed population-based studies can minimize the risk of *selection factors with unintended consequences* on study results (Kukull and Ganguli, 2012).

Spurious correlations in training data affect internal validly as the model is set up to learn irrelevant facts, while training samples that do not sufficiently represent the underlying population affect external validity (Delgado-Rodríguez and Llorca, 2004). In clinical studies, conclusions drawn from the sample may lack external validity when there are differences between study samples and the target populations, such as subject characteristics or hospital procedures (Degtiar and Rose, 2023). For an equivalent example in NLP, consider sentiment analysis. A trained model with a set of words such as “good” associated with positive sentiment may be internally valid, but may fail to perform well on the wider population when it encounters newer terms such as “heartwarming”.

In NLP, even though there is no single comprehensive formal definition of OOD (Arora et al., 2021), conceptually generalization challenges stemming from OOD can generally be considered external validity challenges. OOD can be a result of domain or distribution shift, due to languages or tasks differing from training data (Hupkes et al., 2023), or even samples from adversarial attacks (Omar et al., 2022).

4 Discussion

4.1 Is the generalization failure due to internal or external validity?

Attributing the right cause of failures enables us to take the most effective corrective action, hence separating internal vs. external factors is important, given internal factors are far more controllable than external ones. Ensuring internal validity requires that we understand cause and effect of a model’s performance. This in turn forces researchers to analyze the data, investigate training methods that are robust against issues in the training data such as noisy labels and spurious correlations. High performance on the test set is clearly not sufficient to ensure that the model is capable of solving the task it is trained for, as similar spurious correlations can exist in both training and test sets. Training data is rarely perfect, as it can contain many problems reflecting annotator bias, incorrect or noisy labels (McCoy et al., 2019; Gururangan et al., 2018).

Inspired by prior works that point to the contributors of poor internal validity, as described in Section 3.4, we propose the following checks to ensure internal validity:

1. How many spurious correlations or noisy labels are present in training and/or test data?
2. How diverse is the training data, and is it sufficient in volume to learn the right features for a given task?
3. How robust is the training procedure against spurious correlations or noisy labels?
4. Were model explainability analyses able to identify the model’s reliance on spurious correlations?
5. Are experiments well-designed and reproducible?
6. How effective is the test set in verifying what the model has learned and/or weaknesses in the model?

Questions 1-4 can be difficult to answer accurately in practice due to deficiencies in the current set of tools and technologies available to analyze the large volumes of data for surface patterns. They may rely on domain knowledge. It may simply be expensive to collect more training data. These challenges are compounded by the fact that neural networks are difficult to understand.

Good test sets, on the other hand, provide a pragmatic way to understand the capabilities of a model, even without access to the underlying model architecture. Ideally, the size of the randomly sampled test set should be sufficiently large, as a large sample size is much more likely to representative of true performance than a smaller one (Faber and Fonseca, 2014). Ribeiro et al. (2020) use the principles of software testing to test models, essentially behavioral testing the models using a *CheckList* of test cases. The *CheckList* tests the model against a set of linguistic capabilities such as negation, replacing named entities etc. This requires careful curation of test examples and requires that these samples are updated as the capabilities of the model improve. Similar strategies have been employed to develop test suites for concept recognition systems (Cohen et al., 2010; Groza and Verspoor, 2014) and negation inference (Truong et al., 2022). Kiela et al. (2021) develop Dynabench to continuously update the test set samples with human-in-the-loop.

4.2 Establishing cause and effect in models

In clinical studies, randomized control trials (RCT) form the gold standard of evidence to establish *cause and effect* of treatment or interventions (Hariton and Locascio, 2018) and their outcomes. In a RCT, participants of the study are randomly assigned to a “experimental" and “control" group, where the experimental group receives the intervention and the control group receives a placebo (Kendall, 2003). The key intuition is that if the only non-random difference between the experimental and control group is the intervention, then the intervention must be the cause of the outcome of the intervention. More specifically, causal effects can be measured by “matching” or balancing the distribution in the case and control groups, an approach that is used in observational studies and RCTs to minimize bias or confounder effects (Stuart, 2010; Paterson and Welsh, 2024; Rubin, 1974).

Adapting this approach to NLP benchmarks would involve curating a counterpart ‘control’ test set for the standard randomly sampled test set. The control test set would be created by making minor perturbations to a sample in the original test set, e.g. through the use of contrast sets (Gardner et al., 2020). The idea here is that if the model has effectively learned the key linguistic aspects required to predict a given label, then the model should also make the correct prediction when the key aspect is perturbed, see Table 3. In Table 3, some of the linguistic aspects that are measured are (a) meaning good vs bad (b) the impact of singular vs plural or swapping nouns on sentiment. Furthermore, we suggest that a model’s prediction should be marked as correct if and only if its prediction on a perturbed counterpart is also correct. This would ensure that the model is judged for its linguistic skills, evaluated in the matched pair, at least within the context of the test set, ensuring internal validity. The examples in Table 3 are simplified to illustrate the key idea to measure causal effects, matching in NLP evaluation needs to be explored further.

Case (O)	**Control (O)**	**S**
The movie is good. ✘	The movie is bad. ✔	0
The flower is pretty. ✔	The flowers are pretty. ✘	0
Tom did a great job. ✔	Jack did a great job. ✔	1

Table 3: Simplified example of matched pairs to measure causal effects. The outcome (O), or the model prediction, can be either correct ✔ or incorrect (✘), and the score (S) for a single test scenario is either 1 or 0. With matched pair evaluation, a model relying on spurious factors would be scored $\frac{1}{3} = 33.3\%$, compared without matched pairs where each sample is treated independently $\frac{4}{6} = 66.7\%$.

4.3 How unreliable is a surface pattern?

While it may be impossible to prevent models from relying on surface patterns, we specifically need to watch out for model’s dependence on spurious correlations or features that tend to be highly unreliable, e.g. the distance between the participating entities in relation extraction discussed above.

Generalizability of deep learning networks depends on whether they learn **(a)** spurious correlations, **(b)** reasonable heuristics, or **(c)** oracle true language meaning. Oracle true meaning refers to true language understanding that takes into account the meanings of the individual words as well as the interplay between them relevant to a target task. Spurious correlations are surface patterns with little or no linguistic backing tied to specific sample characteristics, whereas heuristics are plausible surface patterns that may generally work without the need for deeper comprehension. Generalizability is most adversely impacted by spurious correlations, making the model internally invalid. For instance, a model that associates the presence of the word “good” in a review with positive sentiment has captured a heuristic; it is a reasonable rule but may not produce a correct prediction when used in the context of negation or sarcasm. Spurious correlations can render the model useless beyond the test set.

4.4 Replication studies and generalization

Replication studies in psychology have brought the concerns of unreliable studies from scientific research to the forefront, including the possibility of spurious results to be accepted as genuine effects (Nature Editorial, 2022). While reproducibility of results in machine learning is a challenge (Gundersen and Kjensmo, 2018; Belz et al., 2022), replicating experiments in an independent dataset can help support or challenge the generalizability of an original finding (Kukull and Ganguli, 2012). For instance, if good model performance is only achieved in one test set and not replicable in any other tests, it points to possible issues in internal and/or external validity of the study. As an example, (McCoy et al., 2020) identified that the performance of BERT on the original MNLI (Williams et al., 2018) test set is not consistently replicable on a modified version of the test set, and investigations point to BERT relying on heuristics such as lexical overlap between the premise and hypothesis to achieve high scores on the official MNLI test set (McCoy et al., 2019).

4.5 Generalization in large language models

While pretrained and instruction fine-tuned ultra-large language models (LLMs) with billions of parameters such as GPT-3 (Brown et al., 2020) seem to demonstrate improved in-context zero or few shot generalization capabilities, compared to smaller models with a few hundred million parameters such as BERT (Devlin et al., 2019), the jury is out on whether these results can be explained by improved memorization rather than generalization. This is in part because the official training data / test data from public datasets may not be fully independent – they could have been used to train such large models and the training data used is not publicly disclosed or well documented (Sainz et al., 2023; Magar and Schwartz, 2022). Furthermore, these LLMs tend to be highly sensitive to the prompts used, where characteristics that should *not influence* a prompt's interpretation can result in accuracy varying by over 25 points in LLMs including GPT-4 and LLaMA-2-13B (Gan and Mori, 2023; Sclar et al., 2024). These facts bring into question the linguistic comprehension of such LLMs.

Furthermore, whether fine-tuning such LLMs on smaller target datasets with a few thousand training samples can achieve better generalization, compared to much smaller models like BERT, needs to be studied further. Regardless, the need to verify that it is indeed linguistic capabilities that have resulted in the model's performance and not the model relying on spurious correlations remains. This requires ensuring that the test sets are effective in measuring linguistic capabilities.

5 Related works

There are several reports of spurious correlations in SOTA models and benchmark datasets McCoy et al. (2019); Gururangan et al. (2018); Shinoda et al. (2022), as discussed previously in this paper. Gardner et al. (2021) attempt to formalize the definition of which features can be deemed spurious. Approaches to detect spurious features without apriori handcrafting them have also been studied, e.g. Utama et al. (2020), relying on the observation that pre-trained models exploit simple patterns in the early stage of the training phase. Training a model to be robust against spurious correlations is also an active area of research, e.g. Tu et al. (2020) find that multitask learning improves generalization. Data-driven approaches such as the use of h-adversarial training samples (Elangovan et al., 2023), appropriate sample selection (Schwartz and Stanovsky, 2022), or robust loss function (He et al., 2023) can improve model robustness to spurious correlations. Causal deep learning and inference can also assist in ensuring causality (Luo et al., 2020).

6 Conclusion

Generalization is an act of reasoning that involves drawing broad inferences from specific observations (Polit and Beck, 2010). While generalization of machine learning models in NLP is a complex topic, clinical research offers guiding principles on how to understand generalization, including precise population definitions and contrasting controls. There are various points of failure in NLP modelling that can produce results that may not be generalizable. Out-of-distribution (OOD) effects may be an overly simplistic explanation for generalization failures. Attributing causation to generalization failures requires detailed investigation into some potential pitfalls affecting internal validity. Ensuring internal validity requires at least an effective test set that controls for spurious correlations in the training data. More careful construction of evaluation frameworks will ensure appropriate inferences about the relationship between *cause* (model's linguistic capabilities) and *effect* (performance) from experimental results.

Limitations

Internal validity can be compromised for various reasons, broadly speaking, data vs non-data related problems. Non data related issues can be bugs in code. Data related problems such label noise, poor quality annotation can also make a model internally invalid. In this paper, we have primarily focused on spurious correlations, assuming that the labels themselves are of high quality.

As mentioned in section 2.4, identifying spurious correlations that a model relies on is non-trivial. A simple surrogate model used in our case study can only detect the dominant spurious features, while a deep learning model such as BERT can learn thousands of such spurious features. Hence, if the simple surrogate model has poor performance, it does not mean that the model has not relied on spurious features.

Ethics Statement

This work does not involve collection of new data; all analysis relies on previously published data sets. Our work adheres to the ACL Ethics Policy[1].

References

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

K. Bretonnel Cohen, Christophe Roeder, William A. Baumgartner Jr., Lawrence E. Hunter, and Karin Verspoor. 2010. Test suite design for biomedical ontology concept recognition systems. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Irina Degtiar and Sherri Rose. 2023. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524.

M Delgado-Rodríguez and J Llorca. 2004. Bias. *Journal of Epidemiology & Community Health*, 58(8):635–641.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aparna Elangovan, Estrid He, Yuan Li, and Karin Verspoor. 2023. Effects of human adversarial and affable samples on BERT generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7637–7649, Singapore. Association for Computational Linguistics.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Aparna Elangovan, Yuan Li, Douglas E V Pires, Melissa J Davis, and Karin Verspoor. 2022. Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated BioBERT. *BMC Bioinformatics*, 23(1):4.

Robert Epstein. 1984. The principle of parsimony and some applications in psychology. *The Journal of Mind and Behavior*, pages 119–130.

[1] https://www.aclweb.org/portal/content/acl-code-ethics

Jorge Faber and Lilian Fonseca. 2014. How sample size influences research outcomes. *Dental press journal of orthodontics*, 19:27–9.

Joseph Felsenstein. 1983. Parsimony in systematics: biological and statistical issues. *Annual review of ecology and systematics*, 14(1):313–333.

Chengguang Gan and Tatsunori Mori. 2023. Sensitivity and robustness of large language models to prompt template in Japanese text classification tasks. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

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

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert C Gardner. 2000. Correlation, causation, motivation, and second language acquisition. *Canadian Psychology/Psychologie Canadienne*, 41(1):10.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Google. 2022. Generalization - machine learning crash course.

Tudor Groza and Karin Verspoor. 2014. Automated generation of test suites for error analysis of concept recognition systems. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 23–31, Melbourne, Australia.

Odd Erik Gundersen, Kevin Coakley, Christine Kirkpatrick, and Yolanda Gil. 2023. Sources of irreproducibility in machine learning: A review.

Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Gordon H Guyatt, Andrew D Oxman, Regina Kunz, David Atkins, Jan Brozek, Gunn Vist, Philip Alderson, Paul Glasziou, Yngve Falck-Ytter, and Holger J Schünemann. 2011a. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*, 64(4):395–400.

Gordon H Guyatt, Andrew D Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Yngve Falck-Ytter, Roman Jaeschke, Gunn Vist, et al. 2011b. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *Journal of Clinical Epidemiology*, 64(12):1303–1310.

Eduardo Hariton and Joseph J Locascio. 2018. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG*, 125(13):1716.

Jiayuan He, Yuan Li, Zenan Zhai, Biaoyan Fang, Camilo Thorne, Christian Druckenbrodt, Saber Akhondi, and Karin Verspoor. 2023. Focused contrastive loss for classification with pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. State-of-the-art generalisation research in NLP: A taxonomy and review.

J M Kendall. 2003. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, 20(2):164–168.

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

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jesus Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, J. A. Lopez, Umesh K. Nandal, Erin M. van Buel, Ambika Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the Sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Walter A Kukull and Mary Ganguli. 2012. Generalizability: the trees, the forest, and the low-hanging fruit. *Neurology*, 78(23):1886–1891.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yunan Luo, Jian Peng, and Jianzhu Ma. 2020. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8):426–427.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. *Biochem. Med. (Zagreb)*, 22(3):276–282.

Nature Editorial. 2022. Replication studies hold the key to generalization. *Nature Communications*, 13(1):7004.

Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. 2022. Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*.

Heather Paterson and Brandon C. Welsh. 2024. Is it time for the use of pair-matching in all randomized controlled trials of crime and violence prevention? a review of the research. *Aggression and Violent Behavior*, 74:101889.

Cecilia Maria Patino and Juliana Carvalho Ferreira. 2018. Internal and external validity: can you apply research study results to your patients? *J Bras Pneumol*, 44(3):183.

Denise F Polit and Cheryl Tatano Beck. 2010. Generalization in quantitative and qualitative research: myths and strategies. *Int. J. Nurs. Stud.*, 47(11):1451–1458.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA. Association for Computing Machinery.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

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

Peter M Rothwell. 2005. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93.

Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Holger J Schünemann, Peter Tugwell, Barnaby C Reeves, Elie A Akl, Nancy Santesso, Frederick A Spencer, Beverley Shea, George Wells, and Mark Helfand. 2013. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research synthesis methods*, 4(1):49–62.

Roy Schwartz and Gabriel Stanovsky. 2022. On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2022. Look to the right: Mitigating relative position bias in extractive question answering. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 418–425, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Stat. Sci.*, 25(1):1–21.

Elizabeth Tipton. 2014. How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, 39(6):478–501.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Karin Verspoor, K Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC bioinformatics*, 10(1):1–16.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Christopher Wild. 2006. The concept of distribution. *SERJ EDITORIAL BOARD*, page 10.

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.

P M Wortman. 1983. Evaluation research: A methodological perspective. *Annual Review of Psychology*, 34(1):223–260.

Tal Yarkoni. 2022. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1.

A Appendix: PTM-PPI dataset Decision-tree

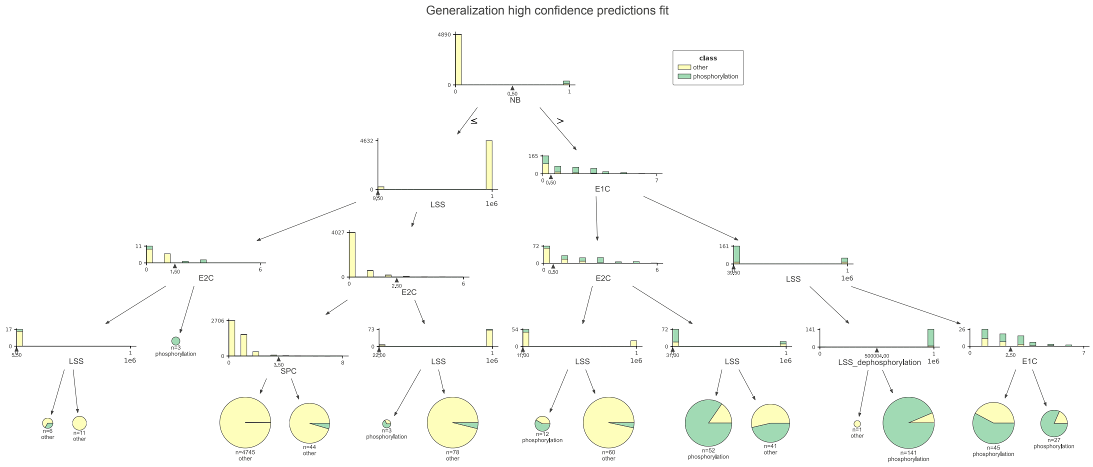

Figure 2: Decision Tree (NB-T) fit in high confidence predictions in the generalization set

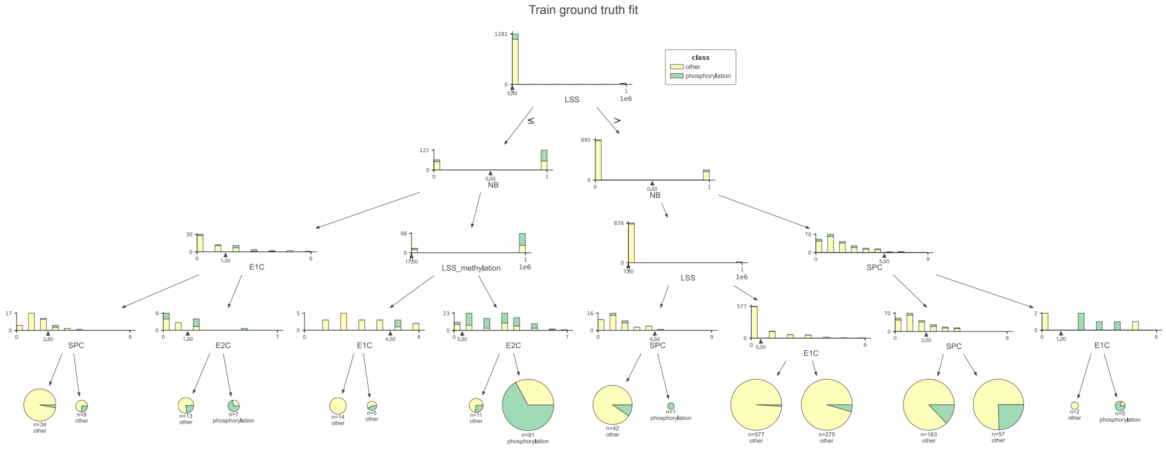

Figure 3: Decision Tree (NB-T) fit in Train ground truth fit

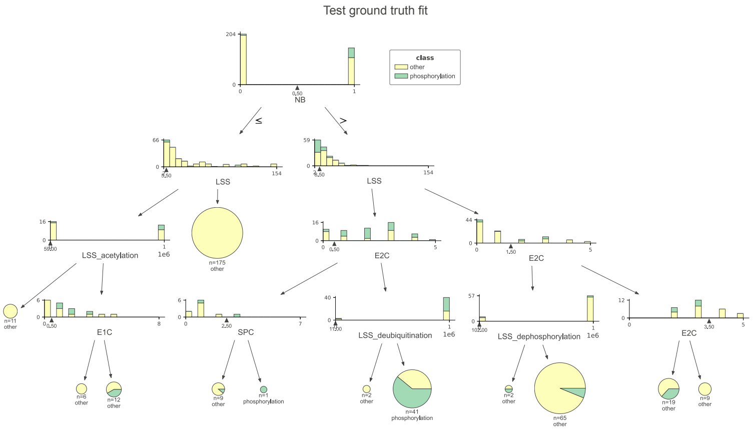

Figure 4: Decision Tree (NB-T) fit in Test ground truth fit

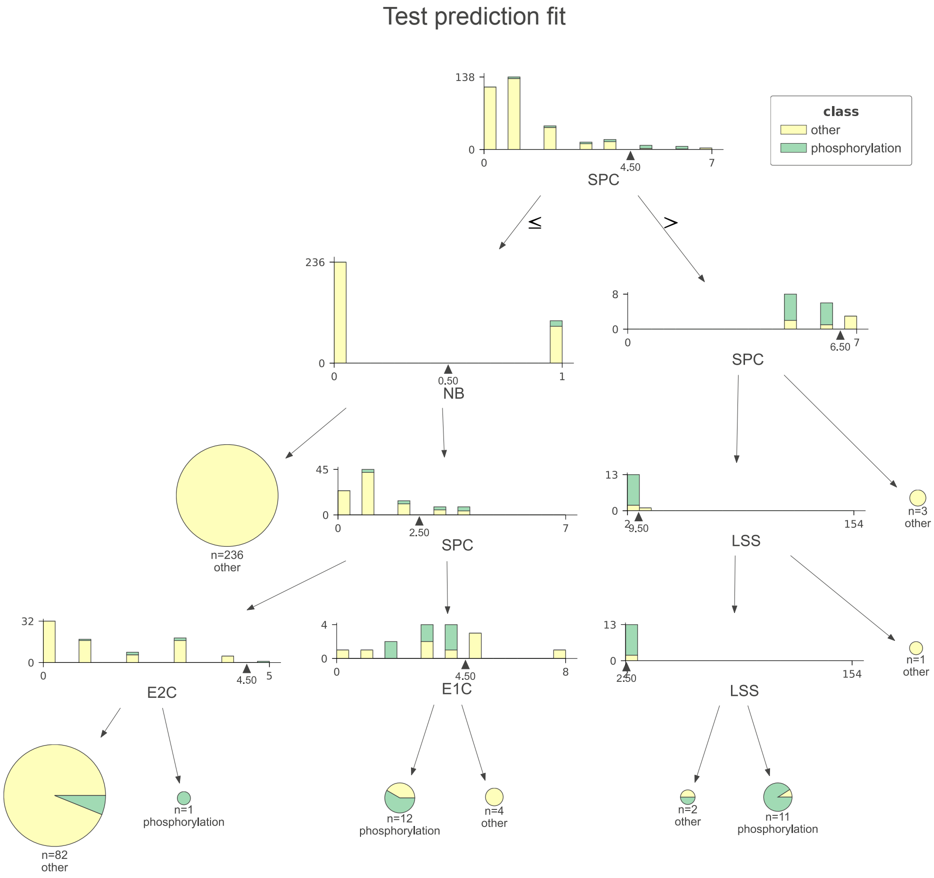

Figure 5: Decision Tree (NB-T) fit in Test set BioBERT predictions fit

B Appendix: SNLI Correlation

We analyze the spurious correlations on the ground truth (GT) as well as the predictions (MP) from BERT (Devlin et al., 2019) using features such as **(a)** the number of words (sentence length) of the hypothesis (HYL) or premise (PRL), **(b)** the presence of negation in the hypothesis (HNEG) and the premise (PNEG). As shown in Table 4, using only the hypothesis achieves the highest κ in the case of both train and test. We also find that hypothesis length (HYL) is a key feature appearing at the top of the decision trees in Figure 7 and Figure 6, which is also identified by Gururangan et al. (2018).

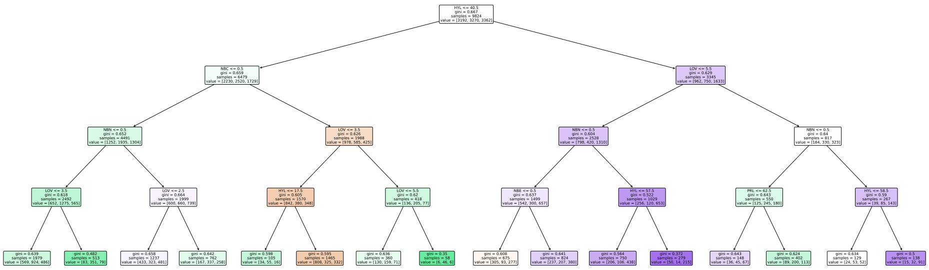

Figure 6: Decision Tree (NB-T) fit in SNLI Test set BERT predictions fit using HYP+PRM

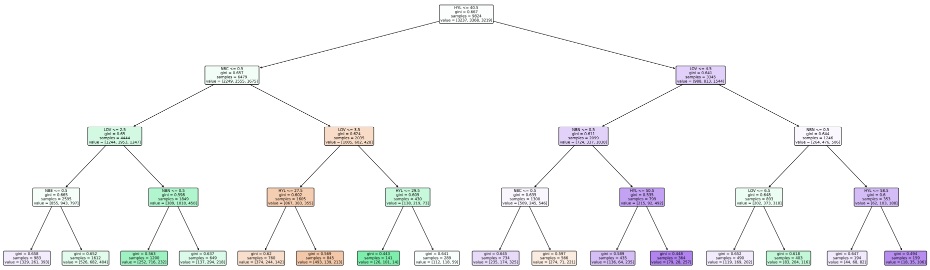

Figure 7: Decision Tree (NB-T) fit in SNLI Test set GT fit using HYP+PRM

Dataset	L	M	κ
SNLI TR PRM	GT	NB	*0.00
SNLI TR PRM	GT	NB-T	0.19
SNLI TR HYP	GT	NB	0.25
SNLI TR HYP	GT	NB-T	**0.27**
SNLI TR HYP+PRM	GT	NB	0.16
SNLI TR HYP+PRM	GT	NB-T	0.22
SNLI TS PRM	GT	NB	0.04
SNLI TS PRM	GT	NB-T	0.20
SNLI TS HYP	GT	NB	0.29
SNLI TS HYP	GT	NB-T	**0.33**
SNLI TS HYP+PRM	GT	NB	0.18
SNLI TS HYP+PRM	GT	NB-T	0.24
SNLI TS PRM	MP	NB	0.05
SNLI TS PRM	MP	NB-T	0.20
SNLI TS HYP	MP	NB	0.32
SNLI TS HYP	MP	NB-T	**0.34**
SNLI TS HYP+PRM	MP	NB	0.19
SNLI TS HYP+PRM	MP	NB-T	0.25

Table 4: Cohen's κ on SNLI Train (SNLI TR) and Test (SNLI TS). The Surrogate models NB and NB-T are used to predict the target label (L) – the ground truth (GT) and the model prediction (MP) of BERT. The features used for Naive Bayes both during NB and stacked NB-T, can either be just the hypothesis (HYP) or just the premise (PRM) or both hypothesis and the premise (PRM). * indicates that Cohen's kappa p-value is 0.10. For all other results, the p-value of Cohen's kappa is less than 0.05.