nature machine intelligence

Analysis

https://doi.org/10.1038/s42256-023-00729-y

A taxonomy and review of generalization research in NLP

Received: 22 December 2022

Accepted: 5 September 2023

Published online: 19 October 2023

Check for updates

Dieuwke Hupkes[1]✉, Mario Giulianelli[2]✉, Verna Dankers[3]✉, Mikel Artetxe[4], Yanai Elazar[5,6], Tiago Pimentel[7], Christos Christodoulopoulos[8], Karim Lasri[9], Naomi Saphra[10], Arabella Sinclair[11], Dennis Ulmer[12,13], Florian Schottmann[14,15], Khuyagbaatar Batsuren[16], Kaiser Sun[17], Koustuv Sinha[17], Leila Khalatbari[18], Maria Ryskina[19], Rita Frieske[18], Ryan Cotterell[14] & Zhijing Jin[14,20]

The ability to generalize well is one of the primary desiderata for models of natural language processing (NLP), but what 'good generalization' entails and how it should be evaluated is not well understood. In this Analysis we present a taxonomy for characterizing and understanding generalization research in NLP. The proposed taxonomy is based on an extensive literature review and contains five axes along which generalization studies can differ: their main motivation, the type of generalization they aim to solve, the type of data shift they consider, the source by which this data shift originated, and the locus of the shift within the NLP modelling pipeline. We use our taxonomy to classify over 700 experiments, and we use the results to present an in-depth analysis that maps out the current state of generalization research in NLP and make recommendations for which areas deserve attention in the future.

Good generalization, roughly defined as the ability to successfully transfer representations, knowledge and strategies from past experience to new experiences, is one of the primary desiderata for models of natural language processing (NLP), as well as for models in the wider field of machine learning[1,2]. For some, generalization is crucial to ensure that models behave robustly, reliably and fairly when making predictions about data different from the data on which they were trained, which is of critical importance when models are employed in the real world. Others see good generalization as intrinsically equivalent to good performance and believe that, without it, a model is not truly able to conduct the task we intended it to. Yet others strive for good generalization because they believe models should behave in a human-like way, and humans are known to generalize well. Although the importance of generalization is almost undisputed, systematic generalization testing is not the status quo in the field of NLP.

At the root of this problem lies the fact that there is little understanding and agreement about what good generalization looks like, what types of generalization exist, how those should be evaluated, and which types should be prioritized in varying scenarios. Broadly speaking, generalization is evaluated by assessing how well a model performs on a test dataset, given the relationship of this dataset with the data on which the model was trained. For decades, it was common to exert only one simple constraint on this relationship: that the train and test data are different. Typically, this was achieved by randomly splitting the available data into training and test partitions. Generalization was thus evaluated by training and testing models on different

[1]FAIR, Amsterdam, The Netherlands. [2]University of Amsterdam, Amsterdam, The Netherlands. [3]University of Edinburgh, Edinburgh, UK. [4]Reka AI, Zarautz, Spain. [5]Allen Institute for AI, Seattle, WA, USA. [6]University of Washington, Seattle, WA, USA. [7]University of Cambridge, Cambridge, UK. [8]Amazon, Edinburgh, UK. [9]École Normale Supérieure, Paris, France. [10]Harvard University, Cambridge, MA, USA. [11]University of Aberdeen, Aberdeen, UK. [12]IT University of Copenhagen, Copenhagen, Denmark. [13]Pioneer Centre for Artificial Intelligence, Copenhagen, Denmark. [14]ETH Zürich, Zurich, Switzerland. [15]Textshuttle, Zurich, Switzerland. [16]National University of Mongolia, Ulaanbaatar, Mongolia. [17]FAIR, New York, NY, USA. [18]Hong Kong University of Science and Technology, Hong Kong, Hong Kong SAR, China. [19]MIT, Cambridge, MA, USA. [20]Max Planck Institute for Intelligent Systems, Tübingen, Germany. ✉e-mail: dieuwkehupkes@meta.com; m.giulianelli@uva.nl; vdankers@ed.ac.uk

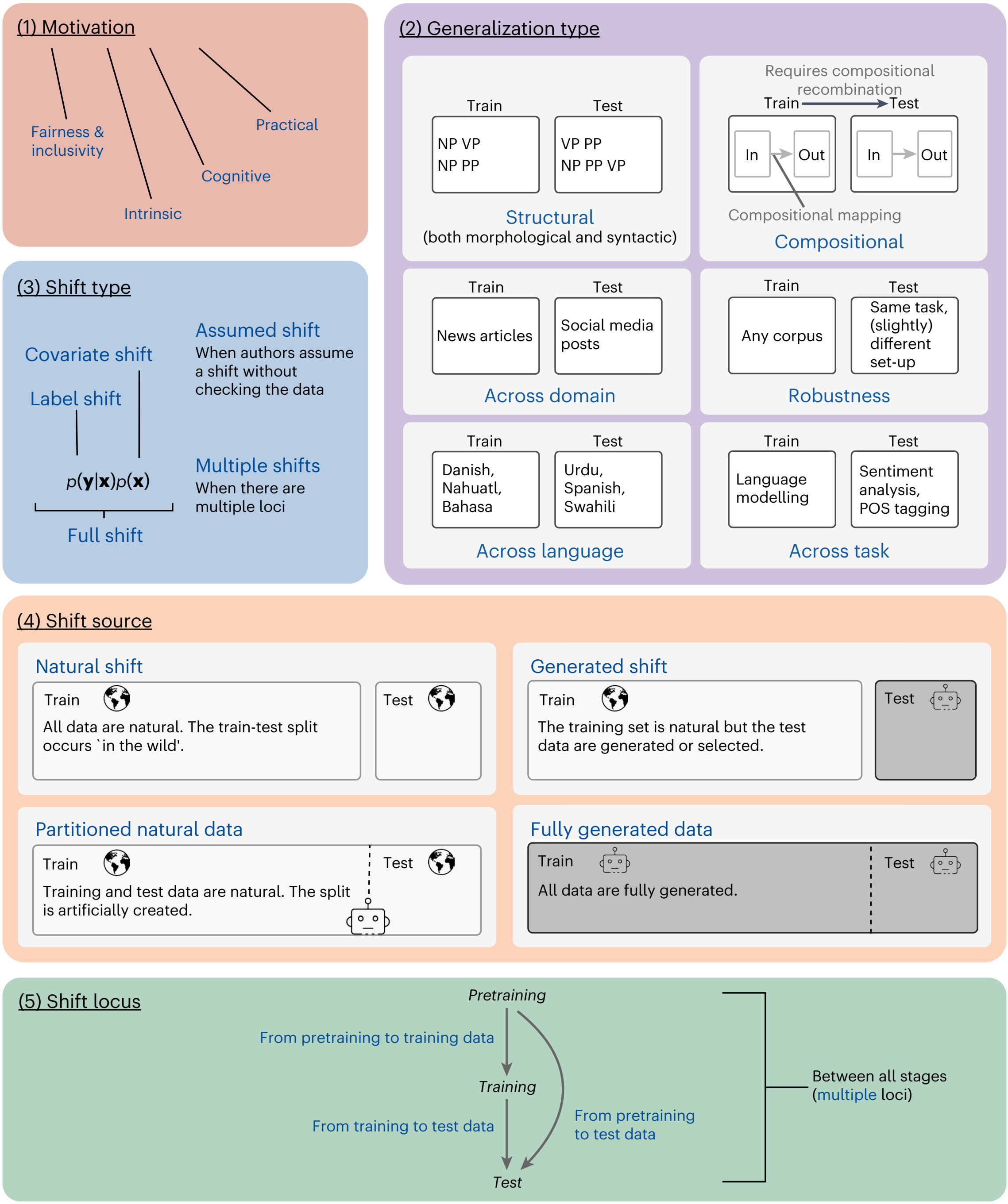

(1) Motivation
Fairness & inclusivity
Intrinsic
Cognitive
Practical
(2) Generalization type
Train
Test
NP VP
NP PP
VP PP
NP PP VP
Structural
(both morphological and syntactic)
Requires compositional recombination
Train
Test
In
Out
In
Out
Compositional mapping
Compositional
Train
Test
News articles
Social media posts
Across domain
Train
Test
Any corpus
Same task, (slightly) different set-up
Robustness
Train
Test
Danish, Nahuatl, Bahasa
Urdu, Spanish, Swahili
Across language
Train
Test
Language modelling
Sentiment analysis, POS tagging
Across task
(3) Shift type
Covariate shift
Label shift
Assumed shift
When authors assume a shift without checking the data
$p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$
Multiple shifts
When there are multiple loci
Full shift
(4) Shift source
Natural shift
Train
All data are natural. The train-test split occurs `in the wild'.
Test
Generated shift
Train
The training set is natural but the test data are generated or selected.
Test
Partitioned natural data
Train
Training and test data are natural. The split is artificially created.
Test
Fully generated data
Train
All data are fully generated.
Test
(5) Shift locus
Pretraining
From pretraining to training data
Training
From training to test data
From pretraining to test data
Test
Between all stages (multiple loci)

Fig. 1 | Graphical representation of our proposed taxonomy of generalization in NLP. The generalization taxonomy we propose consists of five different (nominal) axes that describe (1) the high-level motivation of the work, (2) the type of generalization the test is addressing, (3) what kind of data shift occurs between training and testing and (4) what the source and (5) locus of this shift are. NP, noun phrase; VP, verb phrase; PP, prepositional phrase.

but similarly sampled data, assumed to be independent and identically distributed (i.i.d.). In the past 20 years, we have seen great strides on such random train–test splits in a range of different applications (for example, refs. 3,4).

With this progress, however, came the realization that, for an NLP model, reaching very high or human-level scores on an i.i.d. test set does not imply that the model robustly generalizes to a wide range of different scenarios. We have witnessed a tide of different studies pointing out generalization failures in neural models that have state-of-the-art scores on random train–test splits (as in refs. 5–10, to give just a few examples). Some show that when models perform well on i.i.d. test splits, they might rely on simple heuristics that do not robustly generalize in a wide range of non-i.i.d. scenarios[8,11], over-rely on stereotypes[12,13], or bank on memorization rather than generalization[14,15]. Others, instead, display cases in which performances drop when the evaluation data differ from the training data in terms of genre, domain or topic (for example, refs. 6,16), or when they represent different subpopulations (for example, refs. 5,17). Yet other studies focus on models' inability to generalize compositionally[7,9,18], structurally[19,20], to longer sequences[21,22] or to slightly different formulations of the same problem[13].

By showing that good performance on traditional train–test splits does not equal good generalization, these examples bring into question what kind of model capabilities recent breakthroughs actually reflect, and they suggest that research on the evaluation of NLP models is catching up with the fast recent advances in architectures and training regimes. This body of work also reveals that there is no real agreement on what kind of generalization is important for NLP models, and how that should be studied. Different studies encompass a wide range of generalization-related research questions and use a wide range of different methodologies and experimental set-ups. As of yet, it is unclear how the results of different studies relate to each other, raising the question of how should generalization be assessed, if not with i.i.d. splits? How do we determine what types of generalization are already well addressed and which are neglected, or which types of generalization should be prioritized? Ultimately, on a meta-level, how can we provide answers to these important questions without a systematic way to discuss generalization in NLP? These missing answers are standing in the way of better model evaluation and model development—what we cannot measure, we cannot improve.

Here, within an initiative called GenBench, we introduce a new framework to systematize and understand generalization research in an attempt to provide answers to the above questions. We present a generalization taxonomy, a meta-analysis of 543 papers presenting research on generalization in NLP, a set of online tools that can be used by researchers to explore and better understand generalization studies through our website—https://genbench.org—and we introduce GenBench evaluation cards that authors can use to comprehensively summarize the generalization experiments conducted in their papers. We believe that state-of-the-art generalization testing should be the new status quo in NLP, and we aim to lay the groundwork for facilitating that.

The GenBench generalization taxonomy

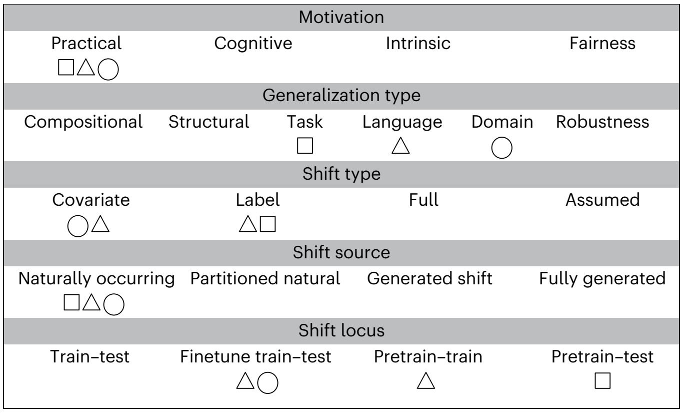
The generalization taxonomy we propose—visualized in Fig. 1 and compactly summarized in Extended Data Fig. 2—is based on a detailed analysis of a large number of existing studies on generalization in NLP. It includes the main five axes that capture different aspects along which generalization studies differ. Together, they form a comprehensive picture of the motivation and goal of the study and provide information on important choices in the experimental set-up. The taxonomy can be used to understand generalization research in hindsight, but is also meant as an active device for characterizing ongoing studies. We facilitate this through GenBench evaluation cards, which researchers can include in their papers. They are described in more detail in Supplementary section B, and an example is shown in Fig. 2. In the following, we give a brief description of the five axes of our taxonomy. More details are provided in the Methods.

Motivation

The first axis of our taxonomy describes the high-level motivation for the study. The motivation of a study determines what type of generalization is desirable, as well as what kind of conclusions can be drawn from a model's display or lack of generalization. Furthermore, the motivation of a study shapes its experimental design. It is therefore important for researchers to be explicitly aware of it, to ensure that the experimental set-up aligns with the questions they seek to answer. We consider four different types of motivation: practical, cognitive, intrinsic, and fairness and inclusivity.

Generalization type

The second axis in our taxonomy indicates the type of generalization the test is addressing. This axis describes on a high level what the generalization test is intended to capture, rather than considering why or how, making it one of the most important axes of our taxonomy. In the literature, we have found six main types of generalization: compositional generalization, structural generalization, cross-task generalization, cross-lingual generalization, cross-domain generalization and robustness generalization. Figure 1 (top right) further illustrates these different types of generalization.



Motivation					
Practical	Cognitive	Intrinsic	Fairness		
□△○					
Generalization type					
Compositional	Structural	Task	Language	Domain	Robustness
		□	△	○	
Shift type					
Covariate	Label	Full	Assumed		
○△	△□				
Shift source					
Naturally occurring	Partitioned natural	Generated shift	Fully generated		
□△○					
Shift locus					
Train–test	Finetune train–test	Pretrain–train	Pretrain–test		
	△○	△	□		

Fig. 2 | Example of a GenBench evaluation card. This example GenBench evaluation card describes a hypothetical paper with three different experiments. As can be seen in the first two rows, all experiments are practically motivated and test different types of generalization: cross-task generalization (square), cross-lingual generalization (triangle) and cross-domain generalization (circle). To do so, they use different data shifts and different loci. The task generalization experiment (square) involves a label shift from pretrain to test, the domain-generalization experiment (circle) a covariate shift in the finetuning stage, and the cross-lingual experiment (triangle) considers multiple shifts (covariate and label) across different stages of the modelling pipeline (pretrain–train and finetune train–test). All experiments use naturally occurring shifts. The LaTeX code for this card was generated with the generation tool at https://genbench.org/eval_cards.

Shift type

The third axis in our taxonomy describes what kind of data shift is considered in the generalization test. This axis derives its importance from the fact that data shift plays an essential formal role in defining and understanding generalization from a statistical perspective, as well as from the fact that different types of shift are best addressed with different kinds of experimental set-up. On the data shift axis we consider three shifts, which are well-described in the literature: covariate, label and full shift. We further include assumed shift to denote studies that assume a data shift without properly justifying it. In our analysis, we mark papers that consider multiple shifts between different distributions involved in the training and evaluation process as having multiple shifts.

Shift source

The fourth axis of the taxonomy characterizes the source of the data shift used in the experiment. The source of the data shift determines how much control the experimenter has over the training and testing data and, consequently, what kind of conclusions can be drawn from an experiment. We distinguish four different sources of shift: naturally occurring shifts, artificially partitioned natural corpora, generated shifts and fully generated data. Figure 1 further illustrates these different types of shift source.

Shift locus

The last axis of our taxonomy considers the locus of the data shift, which describes between which of the data distributions involved in the modelling pipeline a shift occurs. The locus of the shift, together with the shift type, forms the last piece of the puzzle, as it determines what part of the modelling pipeline is investigated and thus the kind of generalization question that can be asked. On this axis, we consider shifts between all stages in the contemporary modelling pipeline—pretraining, training and testing—as well as studies that consider shifts between multiple stages simultaneously.

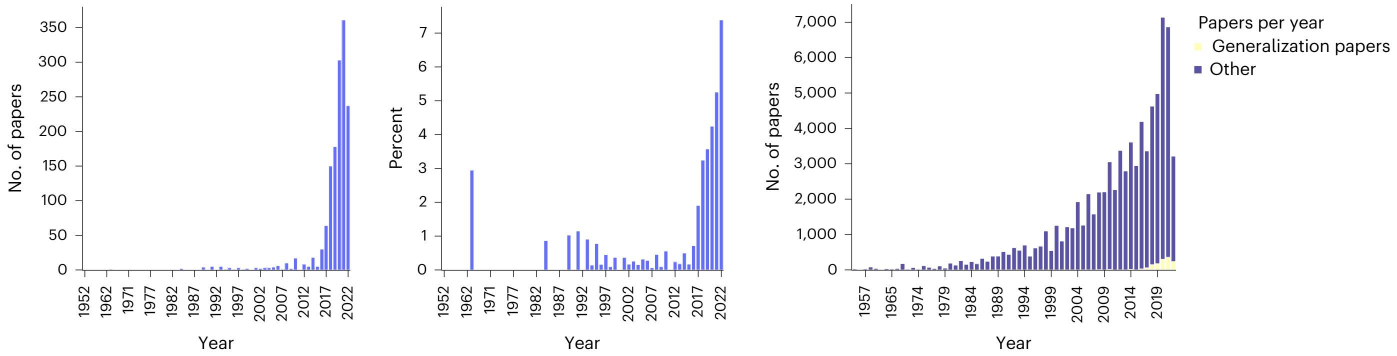


Fig. 3 | Papers about generalization in the ACL anthology. Visualization of the number of papers in the ACL anthology that contain the (sub)words 'generalisation', 'generalization', 'generalise' or 'generalize' in their title or abstract, over time, in absolute terms (left), percentually (middle) and compared to all papers (right). We see how both the absolute number of papers and the percentage of papers about generalization have starkly increased over time. On the right, we visualize the total number of papers and generalization papers published each year.

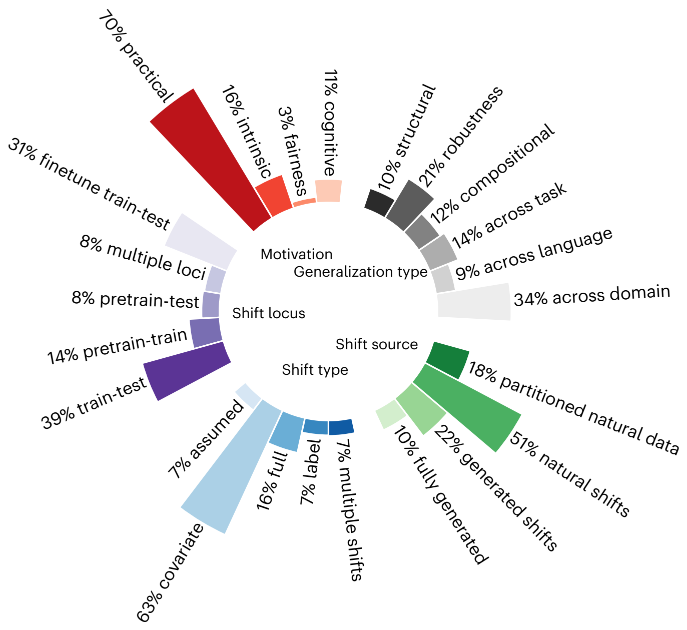


Fig. 4 | Relative occurrences of axes values across all analysed papers. Visualization of the percentage of times each axis value occurs, across all papers that we analysed. Starting from the top left, shown clockwise, are the motivation, the generalization type, the shift source, the shift type and the shift locus.

A review of generalization research in NLP

Using our generalization taxonomy, we analysed 752 generalization experiments in NLP, presented in a total of 543 papers from the anthology of the Association for Computational Linguistics (ACL) that have the (sub)words 'generali(s|z)ation' or 'generali(s|z)e' in their title or abstract. Aggregate statistics on how many such papers we found across different years is available in Fig. 3. For details on how we selected and annotated the papers, see Supplementary section A. A full list of papers is provided in Supplementary section G, as well as on our website (https://genbench.org). On the same website, we also present interactive ways to visualize the results, a search tool to retrieve relevant citations, and a means to generate GenBench evaluation cards, which authors can add to their paper (or appendix) to comprehensively summarize the generalization experiments in their paper (for more information, see Supplementary section B). In this section, we present the main findings of our analysis.

Overall trends on different axes

We begin by discussing the overall frequency of occurrence of different categories on the five axes, without taking into account interactions between them. We plot the relative frequencies of all axis values in Fig. 4 and their development over time in Fig. 5. Because the number of generalization papers before 2018 that are retrieved is very low (Fig. 3a), we restricted the diachronic plots to the last five years. All other reported statistics are computed over our entire selection of papers.

Motivations. As we can see in Fig. 4 (top left), by far the most common motivation to test generalization is the practical motivation. The intrinsic and cognitive motivations follow, and the studies in our Analysis that consider generalization from a fairness perspective make up only 3% of the total. In part, this final low number could stem from the fact that our keyword search in the anthology was not optimal for detecting fairness studies (further discussion is provided in Supplementary section C). We welcome researchers to suggest other generalization studies with a fairness motivation via our website. However, we also speculate that only relatively recently has attention started to grow regarding the potential harmfulness of models trained on large, uncontrolled corpora and that generalization has simply not yet been studied extensively in the context of fairness. Overall, we see that trends on the motivation axis have experienced small fluctuations over time (Fig. 5, left) but have been relatively stable over the past five years.

Generalization type. We find that cross-domain is the most frequent generalization type, making up more than 30% of all studies, followed by robustness, cross-task and compositional generalization (Fig. 4). Structural and cross-lingual generalization are the least commonly investigated. Similar to fairness studies, cross-lingual studies could be undersampled because they tend to use the word 'generalization' in their title or abstract less frequently. However, we suspect that the low number of cross-lingual studies is also reflective of the English-centric disposition of the field. We encourage researchers to suggest cross-lingual generalization papers that we may have missed via our website so that we can better estimate to what extent cross-lingual generalization is, in fact, understudied.

Shift type. Data shift types (Fig. 4) are very unevenly distributed over their potential axis values: the vast majority of generalization research considers covariate shifts. This is, to some extent, expected, because covariate shifts are more easily addressed by most current modelling techniques and can occur between any two stages of the modelling pipeline, whereas label and full shifts typically only occur between pretraining and finetuning. More unexpected, perhaps, is the relatively high amount of assumed shifts, which indicate studies that claim to

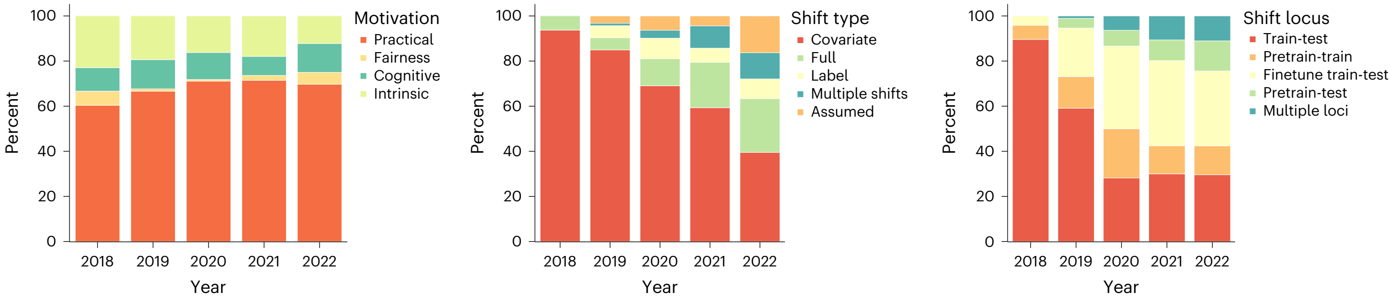


Fig. 5 | Percentual occurrences of different motivations, shift types and shift loci over time. Trends from the past five years for three of the taxonomy's axes (motivation, shift type and shift locus), normalized by the total number of papers annotated per year.

test generalization but do not explicitly consider how their test data relate to their training data. The percentage of such assumed shifts has increased over the past few years (Fig. 5, middle). We hypothesize that this trend, which signals a movement of the field in the wrong direction, is predominantly caused by the use of increasingly large, general-purpose training corpora. Such large corpora, which are often not in the public domain, make it very challenging to analyse the relationship between the training and testing data and, consequently, make it hard to determine what kind of conclusions can be drawn from evaluation results. More promising, instead, is the fact that several studies consider multiple shifts, thus assessing generalization throughout the entire modelling pipeline.

Shift source. On the shift source axis (Fig. 4) we see that almost half of the reviewed generalization studies consider naturally occurring shifts: natural corpora that are not deliberately split along a particular dimension. As discussed later in this section, this type of data source is most prevalent in cross-task and cross-domain generalization studies, for which such naturally different corpora are widely available. The next most frequent categories are generated shifts, where one of the datasets involved is generated with a specific generalization property in mind, and artificially partitioned natural data, describing settings in which all data are natural, but the way it is split between train and test is controlled. Fully generated datasets are less common, making up only 10% of the total number of studies.

Shift locus. Finally, for the locus axis (Fig. 4), we see that the majority of cases focus on finetune/train–test splits. Much fewer studies focus on shifts between pretraining and training or pretraining and testing. Similar to the previous axis, we observe that a comparatively small percentage of studies considers shifts in multiple stages of the modelling pipeline. At least in part, this might be driven by the larger amount of compute that is typically required for those scenarios. Over the past five years, however, the percentage of studies considering multiple loci and the pretrain–test locus—the two least frequent categories—have increased (Fig. 5, right).

Interactions between axes

Next we consider interactions between different axes. Are there any combinations of axes that occur together very often or combinations that are instead rare? We discuss a few relevant trends and encourage the reader to explore these interactions dynamically on our website.

What data shift source is used for different generalization types? In Fig. 6 (top left), we show the relative frequency of each shift source per generalization type. We can see that the shift source varies widely across different types of generalization. Compositional generalization, for example, is predominantly tested with fully generated data, a data type that hardly occurs in research considering robustness, cross-lingual or cross-task generalization. Those three types of generalization are most frequently tested with naturally occurring shifts or, in some cases, with artificially partitioned natural corpora. Structural generalization is the only generalization type that appears to be tested across all different data types. As far as we are aware, very few studies exist that directly compare results between different sources of shift—for example, to investigate to what extent results on generated shifts or fully generated data are indicative of performances on natural corpora (such as refs. 23,24). Such studies could provide insight into how choices in the experimental design impact the conclusions that are drawn from generalization experiments, and we believe that they are an important direction for future work.

For which loci of shift are different generalization types studied? Another interesting question to ask is for which locus different generalization types are considered (Fig. 6, top right). We observe that only cross-task generalization is frequently investigated in the pretrain–train and pretrain–test stages. For all other types of generalization, the vast majority of tests are conducted in the train–test or finetune train–test stage. In some cases, these differences are to be expected: as general-purpose pretrained models are usually trained on very large, relatively uncontrolled corpora, investigating how they generalize to a different domain without further finetuning is hardly possible, and neither is evaluating their robustness, which typically also requires more detailed knowledge of the training data. The statistics also confirm the absence of studies that consider compositional generalization from pretraining to finetuning or from pretraining to training, which is philosophically and theoretically challenging in such set-ups because of their all-encompassing training corpora and the fact that in (large) language models, form and meaning are conflated in one space. A final observation is the relative underrepresentation of studies with multiple loci across all generalization types, especially given the large number of studies that consider generalization in the finetuning stage or with the pretrain–train locus. Those studies have included multiple training stages but considered generalization in only one of them. We hope to see this trend change in the future, with more studies considering generalization in the entire modelling pipeline.

Which types of data shift occur across different loci? Another interesting interaction is the one between the shift locus and the data shift type. Figure 6 (centre left) shows that assumed shifts mostly occur in the pretrain–test locus, confirming our hypothesis that they are probably caused by the use of increasingly large, general-purpose training corpora. When such pretrained models are further finetuned, experiments often consider either a shift between pretraining and finetuning where new labels are introduced, or a covariate shift in the finetuning stage; as such, they do not require an in-depth understanding of the pretraining corpus. The studies that do investigate covariate or full shifts with a pretrain–train or pretrain–test are typically not studies considering large language models, but instead multi-stage processes for domain adaptation.

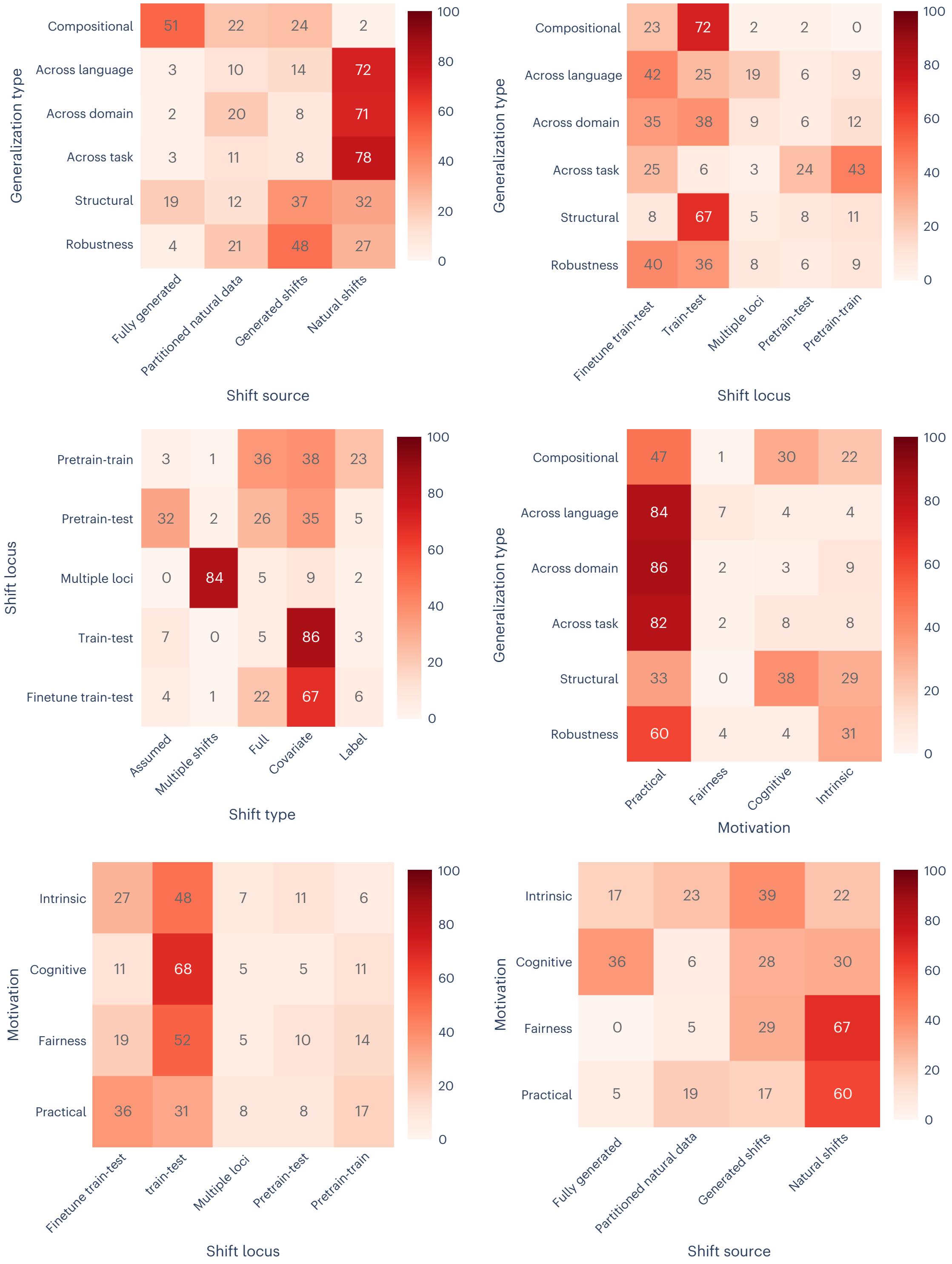


Fig. 6 | Interactions between the various axes of our taxonomy. The interaction between occurrences of values on various axes of our taxonomy, shown as heatmaps. The heatmaps are normalized by the total row value to facilitate comparisons between rows. Different normalizations (for example, to compare columns) and interactions between other axes can be analysed on our website, where figures based on the same underlying data can be generated.

How does motivation drive generalization research? To discuss the relationship between the motivation behind a study and the other axes, we focus on its interactions with generalization type, shift locus and shift source, as shown in the bottom right half of Fig. 6. Considering first the relationship between motivation and generalization type (Fig. 6, centre right), we see that cross-domain, robustness, cross-task and cross-lingual generalizations are predominantly motivated by practical considerations; robustness generalization studies are also frequently motivated by an interest in understanding how models work intrinsically. We find that compositional and structural generalization studies are both frequently driven by cognitive motivations, which is to be expected given the importance of these concepts in human cognition and intelligence (for example, ref. 25). The motivation given most frequently for compositional generalization, however, is a practical one. Although in human learning, compositionality is indeed often associated with important practical properties—speed of learning, powerful generalization—as far as we know, there is little empirical evidence that compositional models actually perform better on natural language tasks. A similar apparent mismatch can be observed in Fig. 6 (bottom right) when focusing on the practical motivation. Practical generalization tests are typically aimed at improving models or at being directly informative of a model's applicability. Nonetheless, more than 20% of the practically motivated studies use either artificially partitioned natural data or even fully generated data. To what extent could their conclusions then actually be informative of models applied in practical scenarios? These apparent mismatches between the motivation and the experimental set-up illustrate the importance of the motivation axis in our taxonomy—being aware of and explicit about a study's motivation ensures that its conclusions are indeed informative with respect to the underlying research question.

Another interesting observation that can be made from the interactions between motivation and shift locus is that the vast majority of cognitively motivated studies are conducted in a train–test set-up. Although there are many good reasons for this, conclusions about human generalization are drawn from a much more varied range of 'experimental set-ups'. For example, any experiments done with adults can be thought of as more similar to tests with a finetune train–test or pretrain–test locus than to the train–test locus, as adults have a life-long experience over which the experimenter has little control beyond participant selection. On the one hand, this suggests that generalization with a cognitive motivation should perhaps be evaluated more often with those loci. On the other hand, it begs the question of whether the field could take inspiration from experiments on human generalization for the challenging effort of evaluating the generalization of large language models, trained on uncontrolled corpora, in a pretrain–test setting. Although there are, of course, substantial differences between the assumptions that can reasonably be made about the linguistic experiences of a human and the pretraining of a language model, we still believe that input from experts that have extensively considered human generalization would be beneficial to improve generalization testing in these more challenging set-ups.

Conclusion

In this Analysis we have presented a framework to systematize and understand generalization research. The core of this framework consists of a generalization taxonomy that can be used to characterize generalization studies along five dimensions. This taxonomy, which is designed based on an extensive review of generalization papers in NLP, can be used to critically analyse existing generalization research as well as to structure new studies. The five nominal axes of the taxonomy describe why a study is executed (the main motivation of the study), what the study intends to evaluate (the type of generalization it aims to solve) and how the evaluation is conducted (the type of data shift considered, the source of this data shift, and the locus in which the shift is investigated).

To illustrate the use and usefulness of our taxonomy, we analysed 543 papers from the ACL anthology about generalization. Through our extensive analysis, we demonstrated that the taxonomy is applicable to a wide range of generalization studies and were able to provide a comprehensive map of the field, observing overall patterns and making suggestions for areas that should be prioritized in the future. Our most important conclusions and recommendations are as follows:

- The goal of a study is not always perfectly aligned with its experimental design. We recommend that future work should be more explicit about motivations and should incorporate deliberate assessments to ensure that the experimental set-up matches the goal of the study (for example, with the GenBench evaluation cards, as discussed in Supplementary section B).
- Cross-lingual studies and generalization studies motivated by fairness and inclusivity goals are underrepresented. We suggest that these areas should be given more attention in future work.
- Papers that target similar generalization questions vary widely in the type of evaluation set-up they use. The field would benefit from more meta-studies that consider how the results of experiments with different experimental paradigms compare to one another.
- The vast majority of generalization studies focus on only one stage of the modelling pipeline. More work is needed that considers generalization in all stages of training, to prioritize models whose generalizing behaviour persists throughout their training pipeline.
- Recent popular NLP models that can be tested directly for their generalization from pretraining to testing are often evaluated without considering the relationship between the (pre)training and test data. We advise that this should be improved, and that inspiration might be taken from how generalization is evaluated in experiments with human participants, where control and access to the 'pretraining' data of a participant are unattainable.

Along with this Analysis we also launch a website, with (1) a set of visualization tools to further explore our results; (2) a search tool that allows researchers to find studies with specific features; (3) a contributions page, allowing researchers to register new generalization studies; and (4) a tool to generate GenBench evaluation cards, which authors can use in their articles to comprehensively summarize their generalization experiments. Although the review and conclusions presented in this Analysis are necessarily static, we commit to keeping the entries on the website up to date when new papers on generalization are published, and we encourage researchers to engage with our online dynamic review by submitting both new studies and existing studies we might have missed. By providing a systematic framework and a toolset that allow for a structured understanding of generalization, we have taken the necessary first steps towards making state-of-the-art generalization testing the new status quo in NLP. In Supplementary section E, we further outline our vision for this, and in Supplementary section D, we discuss the limitations of our work.

Methods

In this Analysis we propose a novel taxonomy to characterize research that aims to evaluate how (well) NLP models generalize, and we use this taxonomy to analyse over 500 papers in the ACL anthology. In this section, we describe the five axes that make up the taxonomy: motivation, generalization type, shift type, shift source and shift locus. A list of examples for every axis value is provided in Supplementary section C. More details about the procedure we used to annotate papers is available in Supplementary section A.

Motivation—what is the high-level motivation for a generalization test?

The first axis we consider is the high-level motivation or goal of a generalization study. We identified four closely intertwined goals of generalization research in NLP, which we refer to as the practical motivation, the cognitive motivation, the intrinsic motivation and the fairness motivation. The motivation of a study determines what type of generalization is desirable, shapes the experimental design, and affects which conclusions can be drawn from a model's display or lack of generalization. It is therefore crucial for researchers to be explicit about the motivation underlying their studies, to ensure that the experimental set-up aligns with the questions they seek to answer. We now describe the four motivations we identified as the main drivers of generalization research in NLP.

Practical. One frequent motivation to study generalization is of a markedly practical nature. Studies that consider generalization from a practical perspective seek to assess in what kind of scenarios a model can be deployed, or which modelling changes can improve performance in various evaluation scenarios (for example, ref. 26). We provide further examples of research questions with a practical nature in Supplementary section C.

Cognitive. A second high-level motivation that drives generalization research is a cognitive one, which can be separated into two underlying categories. The first category is related to model behaviour and focuses on assessing whether models generalize in human-like ways. Human generalization is a useful reference point for the evaluation of models in NLP because it is considered to be a hallmark of human intelligence (for example, ref. 25) and, perhaps more importantly, because it is precisely the type of generalization that is required to successfully model natural language. The second, more deeply cognitively inspired category embraces work that evaluates generalization in models to learn more about language and cognition (for example, ref. 27). Studies in this category investigate what underlies generalization in computational models, not to improve the models' generalization capabilities, but to derive new hypotheses about the workings of human generalization. In some cases, it might be difficult to distinguish cognitive from practical motivations: a model that generalizes like a human should also score well on practically motivated tests, which is why the same experiments can be motivated in multiple ways. In our axes-based taxonomy, rather than assuming certain experiments come with a fixed motivation, we rely on motivations provided by the authors.

Intrinsic. A third motivation to evaluate generalization in NLP models, which cuts through the two previous motivations, pertains to the question of whether models learned the task we intended them to learn, in the way we intended the task to be learned. We call this motivation the intrinsic motivation. The shared presupposition underpinning this type of research is that if a model has truly learned the task it is trained to do, it should also be able to execute this task in settings that differ from the exact training scenarios. What changes, across studies, is the set of conditions under which a model is considered to have appropriately learned a task. Some examples are provided in Supplementary section C. In studies that consider generalization from this perspective, generalization failures are taken as proof that the model did not—in fact—learn the task as we intended it to learn it (for example, ref. 28).

Fairness and inclusivity. A last yet important motivation for generalization research is the desire to have models that are fair, responsible and unbiased, which we denote together as the fairness and inclusivity motivation. One category of studies driven by these concepts, often ethical in nature, asks questions about how well models generalize to diverse demographics, typically considering minority or marginalized groups (for example, ref. 5), or investigates to what extent models perpetuate (undesirable) biases learned from their training data (for example, ref. 17). Another line of research related to both fairness and inclusivity focuses on efficiency, both in terms of the amount of data that is required for a model to converge to a solution as well as the necessary amount of compute. In such studies, efficiency is seen as a correlate of generalization: models that generalize well should learn more quickly and require less data (for example, ref. 29). As such, they are more inclusively applicable—for instance to low-resource languages or demographic groups for which little data are available—they are more accessible for groups with smaller computational resources, and they have a lower environmental impact (for example ref. 30).

Generalization type—what type of generalization is a test addressing?

The second axis in our taxonomy describes, on a high level, what type of generalization a test is intended to capture, making it an important axis of our taxonomy. We identify and describe six types of generalization that are frequently considered in the literature.

Compositional generalization. The first prominent type of generalization addressed in the literature is compositional generalization, which is often argued to underpin humans' ability to quickly generalize to new data, tasks and domains (for example, ref. 31). Although it has a strong intuitive appeal and clear mathematical definition[32], compositional generalization is not easy to pin down empirically. Here, we follow Schmidhuber[33] in defining compositionality as the ability to systematically recombine previously learned elements to map new inputs made up from these elements to their correct output. For an elaborate account of the different arguments that come into play when defining and evaluating compositionality for a neural network, we refer to Hupkes and others[34].

Structural generalization. A second category of generalization studies focuses on structural generalization—the extent to which models can process or generate structurally (grammatically) correct output—rather than on whether they can assign them correct interpretations. Some structural generalization studies focus specifically on syntactic generalization; they consider whether models can generalize to novel syntactic structures or novel elements in known syntactic structures (for example, ref. 35). A second category of structural generalization studies focuses on morphological inflection, a popular testing ground for questions about human structural generalization abilities. Most of this work considers i.i.d. train–test splits, but recent studies have focused on how morphological transducer models generalize across languages (for example, ref. 36) as well as within each language[37].

Cross-task generalization. A third direction of generalization research considers the ability of individual models to adapt to multiple NLP problems—cross-task generalization. Cross-task generalization in NLP has traditionally been strongly connected to transfer and multitask learning[38], in which the goal was to train a network from scratch on multiple tasks at the same time, or to transfer knowledge from one task to another. In that formulation, it was deemed an extremely challenging topic. This has changed with the relatively recent trend of models that are first pretrained with a general-purpose, self-supervised objective and then further finetuned, potentially with the addition of task-specific parameters that learn to execute different tasks using the representations that emerged in the pretraining phase. Rather than evaluating how learning one task can benefit another, this pretrain–finetune paradigm instead gives a central role to the question of how well a model that has acquired some general knowledge about language can successfully be adapted to different kinds of tasks (for example, refs. 4,39), with or without the addition of task-specific parameters.

Cross-lingual generalization. The fourth type of generalization we include is generalization across languages, or cross-lingual generalization. Research in NLP has been very biased towards models and technologies for English[40], and most of the recent breakthroughs rely on amounts of data that are simply not available for the vast majority of the world's languages. Work on cross-lingual generalization is thus important for the promotion of inclusivity and democratization of language technologies, as well as from a practical perspective. Most existing cross-lingual studies focus on scenarios where labelled data is available in a single language (typically English) and the model is evaluated in multiple languages (for example, ref. 41). Another way in which cross-lingual generalization can be evaluated is by testing whether multilingual models perform better than monolingual models on language-specific tasks as a result of being trained on multiple languages at the same time (for example, ref. 42).

Generalization across domains. The next category we include is generalization across domains, a type of generalization that is often required in naturally occurring scenarios—more so than the types discussed so far—and thus carries high practical relevance. Although there is no precise definition of what constitutes a domain, the term broadly refers to collections of texts exhibiting different topical and/or stylistic properties, such as different genres or texts with varying formality levels. We include also temporal generalization, where the training data are produced in a specific time period and the model is tested on data from a different time period, either in the future or in the past (for example, ref. 43), in the category of domain generalization. In the literature, cross-domain generalization has often been studied in connection with domain adaptation—the problem of adapting an existing general model to a new domain (for example, ref. 44).

Robustness generalization. The last category of generalization research we consider on the generalization type axis is robustness generalization, which concerns models' ability to learn task solutions that abstract away from spurious correlations that may occur in the training data and that are aligned with the underlying generalizing solution that humans associate with the task (for example, ref. 28). Research on robustness generalization usually focuses on data shifts that stem from varying data collection processes, which are generally unintended and can be hard to spot. Current work therefore focuses on characterizing such scenarios and understanding their impact. Many of these studies show that models do not generalize in the way we would expect them to, because the training data was in some subtle manner not representative of the true task distribution. For example, they may focus on how models generalize in the face of annotation artefacts (for example, ref. 45), across static and non-static splits (for example, ref. 46) and when certain demographics are under- or over-represented in the training data (for example, ref. 17).

Shift type—what kind of data shift is considered?

We have seen that generalization tests differ in terms of their motivation and the type of generalization that they target. What they share, instead, is that they all focus on cases in which there is a form of shift between the data distributions involved in the modelling pipeline. In the third axis of our taxonomy, we describe the ways in which two datasets used in a generalization experiment can differ. This axis adds a statistical dimension to our taxonomy and derives its importance from the fact that data shift plays an essential role in formally defining and understanding generalization from a statistical perspective.

We formalize the differences between the test, training and potentially pretraining data involved in generalization tests as shifts between the respective data distributions:

$$p\left(\mathbf{x}_{\mathrm{tst}}, \mathbf{y}_{\mathrm{tst}}\right) \qquad \texttt{test} \tag{1}$$

$$p\left(\mathbf{x}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}}\right) \qquad \texttt{training/finetuning/adaptation} \tag{2}$$

$$p\left(\mathbf{x}_{\mathrm{ptr}}, \mathbf{y}_{\mathrm{ptr}}\right) \qquad \texttt{pretraining} \tag{3}$$

These data distributions can be expressed as the product of the probability of the input data $p(\mathbf{x})$ and the conditional probability of the output labels given the input data $p(\mathbf{y}|\mathbf{x})$:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x}) \tag{4}$$

This allows us to define four main types of relation between two data distributions, depending on whether the distributions differ in terms of $p(\mathbf{x})$, $p(\mathbf{y}|\mathbf{x})$, both or none. Note that, for clarity, we focus on train–test shifts, as this is the most intuitive setting, but the shift types we describe in this section can be used to characterize the relationship between any two data distributions involved in a modelling pipeline. One of the four shift types constitutes the case in which there is no shift in data distributions—both $p(\mathbf{x}_{\mathrm{tr}}) = p(\mathbf{x}_{\mathrm{tst}})$ and $p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}}) = p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}})$. This matches the i.i.d. evaluation set-up traditionally used in machine learning. As discussed earlier, this type of evaluation, also referred to as within-distribution generalization, has often been reported not to be indicative of good performance for the more complex forms of generalization that we often desire from our models. We will not discuss this further here, but instead focus on the other three cases, commonly referred to as out-of-distribution (o.o.d.) shifts. In the following, we discuss the shift types we include in our taxonomy.

Covariate shift. The most commonly considered data distribution shift in o.o.d. generalization research is the one where $p(\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{x}_{\mathrm{tr}})$ but $p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) = p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}})$. In this scenario, often referred to as the covariate shift[47,48], the distribution of the input data $p(\mathbf{x})$ changes, but the conditional probability of the labels given the input—which describes the task—remains the same. Under this type of shift, one can evaluate if a model has learned the underlying task distribution while only being exposed to $p(\mathbf{x}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}})$.

Label shift. The second type of shift corresponds to the case in which the focus is on the conditional output distributions: $p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}})$. We refer to this case as the label shift. Label shift can happen within the same task when there are inter-annotator disagreements, when there is a temporal shift in the data, or a change of domain (for example, the phrase 'it doesn't run' can lead to different sentiment labels depending on whether it appears in a review for software or one for mascara). Label shift also occurs when there is a change in task, where it may even happen that not only the meaning of the labels, but the labels themselves change, for example, when shifting from language modelling (where the set of labels is the language vocabulary) to part-of-speech (POS) tagging.

Full shift. The most extreme type of shift corresponds to the case in which $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ change simultaneously: $p(\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{x}_{\mathrm{tr}})$ and $p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}})$. We refer to this case as full shift. Full shifts may occur in language modelling tasks, where changes in the $p(\mathbf{x})$ directly translate into changes in $p(\mathbf{y}|\mathbf{x})$, when adapting to new language pairs in multilingual experiments (for example, ref. 49) or when entirely different types of data are used either for pretraining (for example, ref. 50) or for evaluation (for example, ref. 51).

Assumed shift. When classifying shifts in our Analysis, we mainly focus on cases where authors explicitly consider the relationship between the data distributions they use in their experiments, and the assumptions they make about this relationship are either well-grounded in the literature (for example, it is commonly assumed that switching between domains constitutes a covariate shift) or empirically verified. Nevertheless, we identify numerous studies that claim to be about generalization where such considerations are absent: it is assumed that there is a shift between train and test data, but this is not verified or grounded in previous research. We include this body of work in our Analysis and denote this type of shift with the label 'assumed shift'.

Multiple shifts. Note that some studies consider shifts between multiple distributions at the same time, for instance to investigate how different types of pretraining architecture generalize to o.o.d. splits in a finetuning stage[52] or which pretraining method performs better cross-domain generalization in a second training stage[53]. In the GenBench evaluation cards, both these shifts can be marked (Supplementary section B), but for our analysis in this section, we aggregate those cases and mark any study that considers shifts in multiple different distributions as multiple shift.

Shift source—how are the train and test data produced?

We have discussed what types of shift may occur in generalization tests. We now focus on how those shifts originated. Our fourth axis, graphically shown in Fig. 1, concerns the source of the differences occurring between the pretraining, training and test data distributions. The source of the data shift determines how much control an experimenter has over the training and testing data and, consequently, what kind of conclusions can be drawn from a generalization experiment.

To formalize the description of these different sources of shift, we consider the unobserved base distribution, which describes all data considered in an experiment:

$$p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}}, \boldsymbol{\tau}) \qquad \texttt{base} \tag{5}$$

In this equation, the variable $\boldsymbol{\tau}$ represents a data property of interest, with respect to which a specific generalization ability is tested. This can be an observable property of the data (for example, the length of an input sentence), an unobservable property (for example, the timestamp that defines when a data point was produced) or even a property relative to the model (architecture) under investigation (for example, $\boldsymbol{\tau}$ could represent how quickly a data point was learned in relation to the overall model convergence). The base distribution over $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{\tau}$ can be used to define different partition schemes to be adopted in generalization experiments. Formally, such a partitioning scheme is a rule $f:\mathcal{T}\rightarrow\{\texttt{true},\ \texttt{false}\}$ that discriminates data points according to a property $\boldsymbol{\tau}\in\mathcal{T}$. To investigate how a partitioning scheme impacts model behaviour, the pretraining, training and test distributions can be defined as

$$p\left(\mathbf{x}_{\text{ptr}}, \mathbf{y}_{\text{ptr}}\right) = p\left(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} | f_{\texttt{pretrain}}\left(\boldsymbol{\tau}\right) = \texttt{true}\right) \tag{6}$$

$$p\left(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}\right) = p\left(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} | f_{\texttt{train}}\left(\boldsymbol{\tau}\right) = \texttt{true}\right) \tag{7}$$

$$p\left(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}\right) = p\left(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} | f_{\texttt{test}}\left(\boldsymbol{\tau}\right) = \texttt{true}\right) \tag{8}$$

Using these data descriptions, we can now discuss four different sources of shifts.

Naturally occurring shifts. The first type of shift we include comprises the naturally occurring shifts, which naturally occur between two corpora. In this case, both data partitions of interest are naturally occurring corpora, to which no systematic operations are applied. For the purposes of a generalization test, experimenters have no direct control over the partitioning scheme $f(\boldsymbol{\tau})$. In other words, the variable $\boldsymbol{\tau}$ refers to properties that naturally differ between collected datasets.

Artificially partitioned natural data. A slightly less natural set-up is one in which a naturally occurring corpus is considered, but it is artificially split along specific dimensions. In our taxonomy, we refer to these with the term 'partitioned natural data'. The primary difference with the previous category is that the variable $\boldsymbol{\tau}$ refers to data properties along which data would not naturally be split, such as the length or complexity of a sample. Experimenters thus have no control over the data itself, but they control the partitioning scheme $f(\boldsymbol{\tau})$.

Generated shifts. The third category concerns cases in which one data partition is a fully natural corpus and the other partition is designed with specific properties in mind, to address a generalization aspect of interest. We call these generated shifts. Data in the constructed partition may avoid or contain specific patterns (for example, ref. 18), violate certain heuristics (for example, ref. 8) or include unusually long or complex sequences (for example, ref. 54), or it may be constructed adversarially, generated either by humans[55] or automatically using a specific model (for example, ref. 56).

Fully generated. The last possibility is to use fully generated data. Generating data is often the most precise way of measuring specific aspects of generalization, as experimenters have direct control over both the base distribution and the partitioning scheme $f(\boldsymbol{\tau})$. Sometimes the data involved are entirely synthetic (for example, ref. 34); other times they are templated natural language or a very narrow selection of an actual natural language corpus (for example, ref. 9).

Locus of shift—between which data distributions does the shift occur?

The four axes that we have discussed so far demonstrate the depth and breadth of generalization evaluation research, and they also clearly illustrate that generalization is evaluated in a wide range of different experimental set-ups. They describe high-level motivations, types of generalization, data distribution shifts used for generalization tests, and the possible sources of those shifts. What we have not yet explicitly discussed is between which data distributions those shifts can occur—the locus of the shift. In our taxonomy, the shift locus forms the last piece of the puzzle, as it determines what part of the modelling pipeline is investigated and, with that, what kind of generalization questions can be answered. We consider shifts between all stages in the contemporary modelling pipeline—pretraining, training/finetuning and testing—as well as studies that consider shifts between multiple stages at the same time, as expressed by the data distributions that we have considered (for a graphical representation, see Extended Data Fig. 1).

We describe the loci of shift and how they interact with different components of the modelling pipeline with the aid of three modelling distributions. These modelling distributions correspond to the previously described stages—testing a model, training it, and potentially pretraining it:

$$p\left(\mathcal{Y}_{\text{tst}} | \mathcal{X}_{\text{tst}}, \boldsymbol{\theta}^{*}\right) \qquad \texttt{model} \tag{9}$$

$$p\left(\boldsymbol{\theta}^{*} | \mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}}, \boldsymbol{\phi}_{\text{tr}}, \hat{\boldsymbol{\theta}}\right) \qquad \texttt{training/finetuning/adaptation} \tag{10}$$

$$p\left(\hat{\boldsymbol{\theta}} | \mathcal{X}_{\text{ptr}}, \mathcal{Y}_{\text{ptr}}, \boldsymbol{\phi}_{\text{pr}}, \boldsymbol{\theta}_{0}\right) \qquad \texttt{pretraining} \tag{11}$$

In these equations, $\boldsymbol{\phi}$ broadly denotes the training and pretraining hyperparameters, $\boldsymbol{\theta}$ refers to the model parameters, and $\mathcal{X}$, $\mathcal{Y}$ indicate sets of inputs ($\mathbf{x}$) and their corresponding output ($\mathbf{y}$). Equation (9) defines a model instance, which specifies the probability distribution over the target test labels $\mathcal{Y}_{\text{tst}}$, given the model's parameters $\boldsymbol{\theta}^*$ and a set of test inputs $\mathcal{X}_{\text{tst}}$. Equation (10) defines a training procedure, by specifying a probability distribution over model parameters $\boldsymbol{\theta}^*\in\mathbb{R}^d$ given a training dataset $\mathcal{X}_{\text{tr}}$, $\mathcal{Y}_{\text{tr}}$, a set of training hyperparameters $\boldsymbol{\phi}_{\text{tr}}$ and a (potentially pretrained) model initialization $\hat{\boldsymbol{\theta}}$. Finally, equation (11) defines a pretraining procedure, specifying a conditional probability over the set of parameters $\hat{\boldsymbol{\theta}}$, given a pretraining dataset, a set of pretraining hyperparameters $\boldsymbol{\phi}_{\text{pr}}$ and a model initialization. Between which of these stages a shift occurs impacts which modelling distributions can be evaluated. We now discuss the different potential loci of shifts.

The train–test locus. Probably the most commonly occurring locus of shift in generalization experiments is the train–test locus, corresponding to the classic set-up where a model is trained on some data and then directly evaluated on a shifted (o.o.d.) test partition. In some cases, researchers investigate the generalization abilities of a single model instance (that is, a set of parameters $\boldsymbol{\theta}^*$, as described in equation (9)). Studies of this type therefore report the evaluation of a model instance—typically made available by others—without considering how exactly it was trained, or how that impacted the model's generalization behaviour (for example, ref. 57). Alternatively, researchers might evaluate one or more training procedures, investigating if the training distribution results in model instances that generalize well (for example, ref. 58). Although these cases also require evaluating model instances, the focus of the evaluation is not on one particular model instance, but rather on the procedure that generated the evaluated model instances.

The finetune train–test locus. The second potential locus of shift—the finetune train–test locus—instead considers data shifts between the train and test data used during finetuning and thus concerns models that have gone through an earlier stage of training. This locus occurs when a model is evaluated on a finetuning test set that contains a shift with respect to the finetuning training data. Most frequently, research with this locus focuses on the finetuning procedure and on whether it results in finetuned model instances that generalize well on the test set. Experiments evaluating o.o.d. splits during finetuning often also include a comparison between different pretraining procedures; for instance, they compare how BERT models and RoBERTa models behave during finetuning, thus investigating both a pretrain–train shift and a finetune train–test shift at the same time.

The pretrain–train locus. A third possible locus of shift is the pretrain–train locus, between pretraining and training data. Experiments with this locus evaluate whether a particular pretraining procedure (equation (11)) results in models (parameter sets $\hat{\boldsymbol{\theta}}$) that are useful when further trained on different tasks or domains (for example, ref. 59).

The pretrain–test locus. Finally, experiments can have a pretrain–test locus, where the shift occurs between pretraining and test data. This locus occurs when a pretrained model is evaluated directly on o.o.d. data, without further training (that is, $\mathcal{X}_{\mathrm{tr}}, \mathcal{Y}_{\mathrm{tr}} = \varnothing, \varnothing$)—as frequently happens in in-context learning set-ups (for example, ref. 60)—or when a pretrained model is finetuned on examples that are i.i.d. with respect to the pretraining data and then tested on out-of-distribution instances. The former case ($\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}$) is similar to studies with only one training stage in the train–test locus, but distinguishes itself by the nature of the (pre)training procedure, which typically has a general-purpose objective, rather than being task-specific (for example, a language modelling objective).

Multiple loci. In some cases, one single study may investigate multiple shifts between different parts of the modelling pipeline. Multiple-loci experiments evaluate all stages of the modelling pipeline at once: they assess the generalizability of models produced by the pretraining procedure as well as whether generalization happens in the finetuning stage (for example, ref. 61). Although those can be separately annotated in GenBench evaluation cards, in the analysis section of this Analysis we take them all together in a single category and denote those studies to have multiple loci.

Data availability

The full annotated list of articles included in our survey is available through the GenBench website (https://genbench.org/references), where articles can be filtered through a dedicated search tool. This is an evolving survey: we encourage authors to submit new work and to request annotation corrections through our contributions page (https://genbench.org/contribute). The exact list used at the time of writing can be retrieved from https://github.com/GenBench/GenBench.github.io/blob/cea0bd6bd8af6f2d0f096c8f81185b1dfc9303b5/taxonomy_clean.tsv. We also release interactive tools to visualize the results of our survey at https://genbench.org/visualisation. Source data are provided with this paper.

References

1. Marcus, G. F. Rethinking eliminative connectionism. *Cogn. Psychol.* **37**, 243–282 (1998).
2. Kirk, R., Zhang, A., Grefenstette, E. & Rocktäschel, T. A survey of generalisation in deep reinforcement learning. *J. Artif. Intell. Res.* https://doi.org/10.1613/jair.1.14174 (2023).
3. Chowdhery, A. et al. PaLM: scaling language modeling with pathways. *J. of Mach. Learn. Res.* **24**, 1–113 (2023).
4. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Burstein, J. et al eds) 4171–4186 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/N19-1423
5. Blodgett, S. L., Green, L. & O'Connor, B. Demographic dialectal variation in social media: a case study of African-American English. Jian Su, Kevin Duh, Xavier Carreras (eds). In *Proc. 2016 Conference on Empirical Methods in Natural Language Processing* (Su, J. et al eds) 1119–1130 (Association for Computational Linguistics, 2016); https://doi.org/10.18653/v1/D16-1120. https://aclanthology.org/D16-1120
6. Plank, B. What to do about non-standard (or non-canonical) language in NLP. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1608.07836 (2016).
7. Lake, B. & Baroni, M. Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks. In *Proc. 35th International Conference on Machine Learning* (*ICML*) 4487–4499 (International Machine Learning Society, 2018).
8. McCoy, T., Pavlick, E. & Linzen, T. Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A. et al eds.) 3428–3448 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/P19-1334, https://aclanthology.org/P19-1334
9. Kim, N. & Linzen, T. COGS: a compositional generalization challenge based on semantic interpretation. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*) (Webber, B. et al eds.) 9087–9105 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.emnlp-main.731, https://aclanthology.org/2020.emnlp-main.731
10. Khishigsuren, T. et al. Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* 2798-2807 (European Language Resources Association, 2022); https://aclanthology.org/2022.lrec-1.299
11. Kaushik, D., Hovy, E. & Lipton, Z. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations* (2019).
12. Parrish, A. et al. BBQ: a hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics* (Muresan, S. et al eds.) 2086–2105 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.findings-acl.165, https://aclanthology.org/2022.findings-acl.165
13. Srivastava, A. et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.04615 (2022).
14. Razeghi, Y., Logan, R. L. IV, Gardner, M. & Singh, S. Impact of pretraining term frequencies on few-shot reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022* 840-854 (Association for Computational Linguistics, 2022); https://aclanthology.org/2022.findings-emnlp.59.pdf
15. Lewis, P., Stenetorp, P. & Riedel, S. Question and answer test-train overlap in open-domain question answering datasets. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Merlo, P. et al eds.) 1000–1008 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.eacl-main.86, https://aclanthology.org/2021.eacl-main.86
16. Michel, P. & Neubig, G. MTNT: a testbed for machine translation of noisy text. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* (Riloff, E. et al eds.) 543–553 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/D18-1050, https://aclanthology.org/D18-1050
17. Dixon, L., Li, J., Sorensen, J., Thain, N. & Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proc. 2018 AAAI/ACM Conference on AI, Ethics and Society* 67–73 (Association for Computing Machinery, 2018); https://doi.org/10.1145/3278721.3278729

18. Dankers, V., Bruni, E. & Hupkes, D. The paradox of the compositionality of natural language: a neural machine translation case study. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*) (Muresan, S. et al eds.) 4154–4175 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.acl-long.286, https://aclanthology.org/2022.acl-long.286
19. Wei, J., Garrette, D., Linzen, T. & Pavlick, E. Frequency effects on syntactic rule learning in transformers. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (Moens, M.-F. et al eds.) 932–948 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.emnlp-main.72, https://aclanthology.org/2021.emnlp-main.72
20. Weber, L., Jumelet, J., Bruni, E. & Hupkes, D. Language modelling as a multi-task problem. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume* (Merlo, P. et al eds.) 2049–2060 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.eacl-main.176, https://aclanthology.org/2021.eacl-main.176
21. Raunak, V., Kumar, V., Metze, F. & Callan, J. On compositionality in neural machine translation. In *NeurIPS 2019 Context and Compositionality in Biological and Artificial Neural Systems Workshop* (2019); https://arxiv.org/abs/1911.01497
22. Dubois, Y., Dagan, G., Hupkes, D. & Bruni, E. Location attention for extrapolation to longer sequences. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. et al eds.) 403–413 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.acl-main.39, https://aclanthology.org/2020.acl-main.39
23. Chaabouni, R., Dessì, R. & Kharitonov, E. Can transformers jump around right in natural language? Assessing performance transfer from SCAN. In *Proc. Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* 136–148 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.blackboxnlp-1.9, https://aclanthology.org/2021.blackboxnlp-1.9
24. Sun, K., Williams, A. & Hupkes, D. A replication study of compositional generalization works on semantic parsing. *In ML Reproducibility Challenge 2022.* (2023); https://openreview.net/pdf?id=MF9uv95psps
25. Marcus, G. F. *The Algebraic Mind*: *Integrating Connectionism and Cognitive Science* (Linzen, T. et al eds.) (MIT Press, 2003).
26. Zhou, X., Elfardy, H., Christodoulopoulos, C., Butler, T. & Bansal, M. Hidden biases in unreliable news detection datasets. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume* (Merlo, P. et al eds.) 2482–2492 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.eacl-main.211, https://aclanthology.org/2021.eacl-main.211
27. Lakretz, Y. et al. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition* **213**, 104699 (2021).
28. Talman, A. & Chatzikyriakidis, S. Testing the generalization power of neural network models across NLI benchmarks. In *Proc. 2019 ACL Workshop BlackboxNLP*: *Analyzing and Interpreting Neural Networks for NLP* (Linzen, T. et al eds.) 85–94 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/W19-4810, https://aclanthology.org/W19-4810
29. Marcus, G. Deep learning: a critical appraisal. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1801.00631 (2018).
30. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A. et al eds.) 3645–3650 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/P19-1355, https://aclanthology.org/P19-1355
31. Fodor, J. A. & Pylyshyn, Z. W. Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71 (1988).
32. Montague, R. Universal grammar. *Theoria* **36**, 373–398 (1970).
33. Schmidhuber, J. Towards compositional learning in dynamic networks. Technical report (Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), 1990).
34. Hupkes, D., Dankers, V., Mul, M. & Bruni, E. Compositionality decomposed: how do neural networks generalise? *J. Artif. Intell. Res.* **67**, 757–795 (2020).
35. Jumelet, J., Denic, M., Szymanik, J., Hupkes, D. & Steinert-Threlkeld, S. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics* (Zong, C. et al eds.) 4958–4969 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.findings-acl.439, https://aclanthology.org/2021.findings-acl.439
36. Pimentel, T. et al. SIGMORPHON 2021 shared task on morphological reinflection: generalization across languages. In *Proc. 18th SIGMORPHON Workshop on Computational Research in Phonetics*, *Phonology and Morphology* (Nicolai, G. et al eds.) 229–259 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.sigmorphon-1.25, https://aclanthology.org/2021.sigmorphon-1.25
37. Liu, L. & Hulden, M. Can a transformer pass the wug test? Tuning copying bias in neural morphological inflection models. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*) (Muresan, S. et al eds.) 739–749 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.acl-short.84, https://aclanthology.org/2022.acl-short.84
38. Collobert, R. & Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proc. Twenty-Fifth International Conference on Machine Learning* (*ICML 2008*) Vol. 307 of *ACM International Conference Proceeding Series* (eds Cohen, W. W., McCallum, A. & Roweis, S. T.) 160–167 (ACM, 2008).
39. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
40. Bender, E. M. On achieving and evaluating language-independence in NLP. *Ling. Issues Lang. Technol.* https://doi.org/10.33011/lilt.v6i.1239 (2011).
41. Wu, S. & Dredze, M. Beto, Bentz, Becas: the surprising cross-lingual effectiveness of BERT. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*) (Inui, K. et al eds.) 833–844 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/D19-1077, https://aclanthology.org/D19-1077
42. Zhang, B., Williams, P., Titov, I. & Sennrich, R. Improving massively multilingual neural machine translation and zero-shot translation. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. et al eds.) 1628–1639 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.acl-main.148, https://aclanthology.org/2020.acl-main.148
43. Lazaridou, A. et al. Mind the gap: assessing temporal generalization in neural language models. *Adv. Neural Inf. Process. Syst.* **34**, 29348–29363 (2021).
44. Daumé, H. III. Frustratingly easy domain adaptation. In *Proc. 45th Annual Meeting of the Association of Computational Linguistics* (Zaenen, A. et al eds.) 256–263 (Association for Computational Linguistics, 2007); https://aclanthology.org/P07-1033
45. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. & Van Durme, B. Hypothesis only baselines in natural language inference. In *Proc. Seventh Joint Conference on Lexical and Computational Semantics* (Nissim, M. et al eds.) 180–191 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/S18-2023, https://aclanthology.org/S18-2023

46. Gorman, K. & Bedrick, S. We need to talk about standard splits. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A. et al eds.) 2786–2791 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/P19-1267, https://aclanthology.org/P19-1267
47. Storkey, A. When training and test sets are different: characterizing learning transfer. *Dataset Shift Mach. Learn.* **30**, 3–28 (2009).
48. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, Rocío, Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. *Pattern Recogn.* **45**, 521–530 (2012).
49. Kodner, J. et al. SIGMORPHON–UniMorph 2022 shared task 0: generalization and typologically diverse morphological inflection. In *Proc. 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology and Morphology* (Nicolai, G. et al eds.) 176–203 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.sigmorphon-1.19, https://aclanthology.org/2022.sigmorphon-1.19
50. Papadimitriou, I. & Jurafsky, D. Learning music helps you read: using transfer to study linguistic structure in language models. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*) (Webber, B. et al eds.) 6829–6839 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.emnlp-main.554, https://aclanthology.org/2020.emnlp-main.554
51. De Varda, A. & Zamparelli, R. Multilingualism encourages recursion: a transfer study with mBERT. In *Proc. 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP* (Vylomova, E. et al eds.) 1–10 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.sigtyp-1.1, https://aclanthology.org/2022.sigtyp-1.1
52. Li, B. et al. Quantifying adaptability in pre-trained language models with 500 tasks. In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Carpuat, M. et al eds.) 4696–4715 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.naacl-main.346, https://aclanthology.org/2022.naacl-main.346
53. Wang, B., Lapata, M. & Titov, I. Meta-learning for domain generalization in semantic parsing. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Toutanova, K. et al eds.) 366–379 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.naacl-main.33, https://aclanthology.org/2021.naacl-main.33
54. Lakretz, Y., Desbordes, T., Hupkes, D. & Dehaene, S. Causal transformers perform below chance on recursive nested constructions, unlike humans. In *Proceedings of the 29th International Conference on Computational Linguistics* 3226–3232 (International Committee on Computational Linguistics, 2022); https://aclanthology.org/2022.coling-1.285
55. Kiela, D. et al. Dynabench: rethinking benchmarking in NLP. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (Toutanova, K. et al eds.) 4110–4124 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.naacl-main.324, https://aclanthology.org/2021.naacl-main.324
56. Zellers, R., Bisk, Y., Schwartz, R. & Choi, Y. SWAG: a large-scale adversarial dataset for grounded commonsense inference. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* (Riloff, E. et al eds.) 93–104 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/D18-1009, https://aclanthology.org/D18-1009
57. Lakretz, Y. et al. The emergence of number and syntax units in LSTM language models. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume 1* (*Long and Short Papers*) (Burstein, J. et al eds.) 11–20 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/N19-1002, https://aclanthology.org/N19-1002
58. Rae, J. W. et al. Scaling language models: methods, analysis and insights from training gopher. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2112.11446 (2021).
59. Artetxe, M. et al. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y. et al eds.) 11699-11732 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.emnlp-main.804, https://aclanthology.org/2022.emnlp-main.804/
60. Lin, Xi Victoria et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y. et al eds.) 9019–9052 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.emnlp-main.616, https://aclanthology.org/2022.emnlp-main.616/
61. Yanaka, H., Mineshima, K. & Inui, K. Exploring transitivity in neural NLI models through veridicality. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume* (Merlo, P. et al eds.) 920–934 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.eacl-main.78, https://aclanthology.org/2021.eacl-main.78

Acknowledgements

We thank A. Williams, A. Joulin, E. Bruni, L. Weber, R. Kirk and S. Riedel for providing feedback on the various stages of this paper, and G. Marcus for providing detailed feedback on the final draft. We also thank the reviewers of our work for providing useful comments. We thank E. Hupkes for making the app that allows searching through references, and we thank D. Haziza and E. Takmaz for other contributions to the website. M.G. was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 819455). V.D. was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant no. EP/S022481/1) and the University of Edinburgh. N.S. was supported by the Hyundai Motor Company (under the project Uncertainty in Neural Sequence Modeling) and the Samsung Advanced Institute of Technology (under the project Next Generation Deep Learning: From Pattern Recognition to AI).

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00729-y.

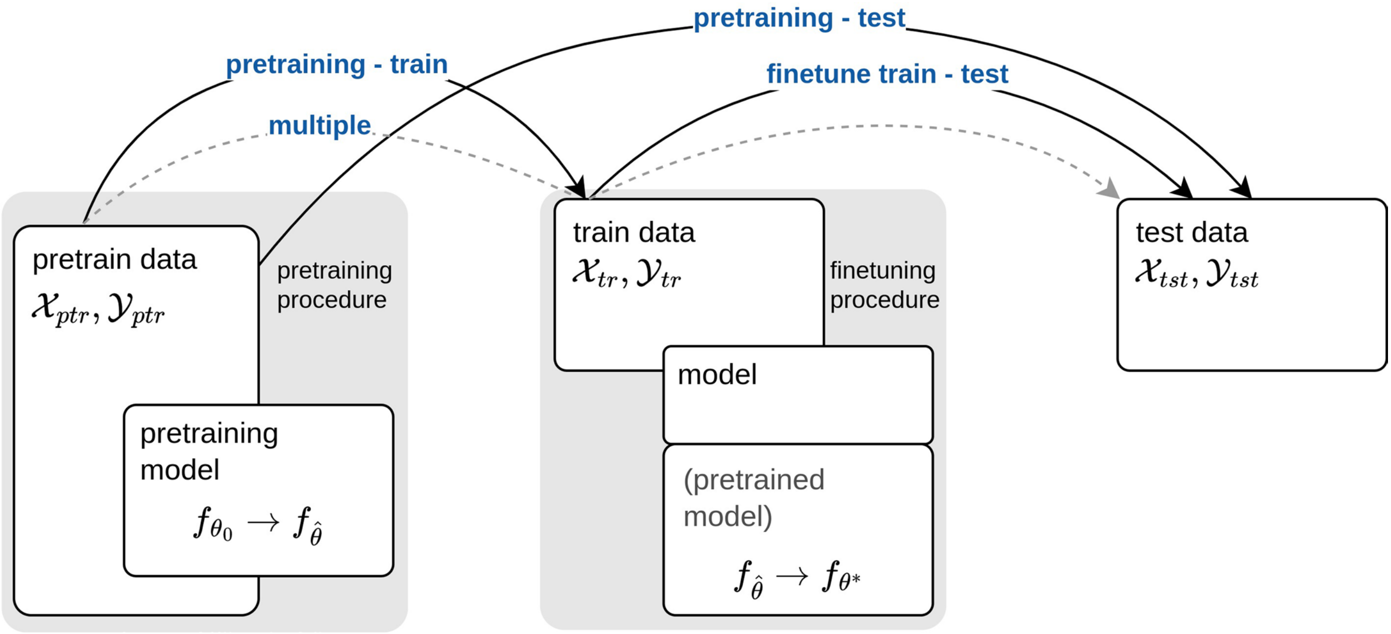
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00729-y.

Correspondence and requests for materials should be addressed to Dieuwke Hupkes, Mario Giulianelli or Verna Dankers.

Peer review information *Nature Machine Intelligence* thanks Karin Verspoor and Raphaël Millière for their contribution to the peer review of this work. Primary Handling Editor: Jacob Huth, in collaboration with the *Nature Machine Intelligence* team.

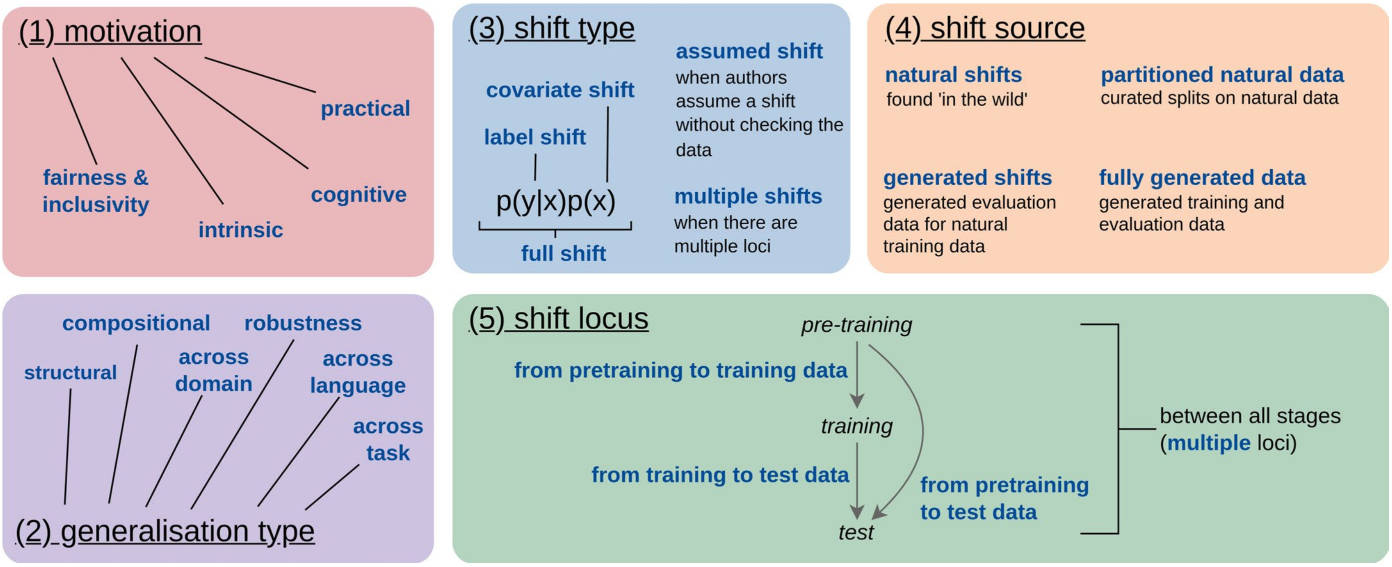
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


pretraining - test
pretraining - train
finetune train - test
multiple
pretrain data
$\mathcal{X}_{ptr}, \mathcal{Y}_{ptr}$
pretraining procedure
pretraining model
$f_{\theta_0} \rightarrow f_{\hat{\theta}}$
train data
$\mathcal{X}_{tr}, \mathcal{Y}_{tr}$
finetuning procedure
model
(pretrained model)
$f_{\hat{\theta}} \rightarrow f_{\theta^*}$
test data
$\mathcal{X}_{tst}, \mathcal{Y}_{tst}$

Extended Data Fig. 1 | Different loci of splits, and the parts of the modelling pipeline for which they may investigate generalisation. The shifts that characterise generalisation experiments in NLP can occur in different places in the modelling pipeline. In this figure, we visualise the three stages of the contemporary modelling pipeline: the pretraining stage, consisting of pretraining data as well as a pretraining procedure; the training stage, which involves training data, a pretrained model, and a training procedure; and finally, the test stage, in which an already trained model is tested on a test dataset. As visualised in this figure, shifts can occur between all or multiple of those stages, which allows to investigate different parts of the modelling pipeline.

https://doi.org/10.1038/s42256-023-00729-y



Extended Data Fig. 2 | A compact graphical representation of our proposed taxonomy of generalisation in NLP. The generalisation taxonomy we propose consists of five different (nominal) axes, that describe the high-level motivation of the work (top, left), the *type* of generalisation the test is addressing (bottom, left); what kind of *data shift* occurs between training and testing (top, middle), and what the *source* (top, right) and *locus* of this shift (bottom, right) are.