

# State-of-the-art generalisation research in NLP

a taxonomy and review

Author 1, author 2, author 3, ...

## Abstract

The ability to generalise well is one of the primary desiderata of natural language processing (NLP). Yet, how ‘good generalisation’ should be defined and how it should be evaluated is not well understood, nor are there any common standards to evaluate generalisation. As a consequence, newly proposed models are not usually systematically tested for their ability to generalise. In this paper, we lay the groundwork for making generalisation-testing the new status-quo in evaluation: we develop a taxonomy for characterising and understanding generalisation research in NLP, and we present a comprehensive map of NLP generalisation research presented in the past 5 years. Our taxonomy is based on an extensive literature review of generalisation research, and contains five different (nominal) axes along which generalisation research can differ: their main *motivation*, the *type* of generalisation they aim to solve, the type of *data shift* they are considering, the *locus* of this shift and the *source* by which this data shift is obtained. We explain the axes of our taxonomy by providing ample examples from the literature and then use it to classify N previous papers that test generalisation. We use the results of this survey to visualise what the field of generalisation research in NLP looks like, to more generally assess where we are when it comes to evaluating generalisation in NLP, identify areas that are over- or underrepresented, and make recommendations for what questions should be addressed in the future. Along with this paper, we release a webpage where the results of our review can be dynamically viewed, and which we intend to update as new NLP generalisation studies come out. With this work, we aim to make steps towards *state-of-the-art* generalisation evaluation in NLP becoming the new standard for any new model that gets proposed.

## 1 Introduction

Good generalisation, 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) (Elangovan et al., 2021; Lake et al., 2017; Linzen, 2020; Plank, 2016; Schmidhuber, 1990; Wong and Wang, 2007; Yogatama et al., 2019, i.a.), as well as in the wider field of machine learning (e.g. Kirk et al., 2021; Shen et al., 2021). There is, however, little agreement about what kind of generalisation behaviour modern-age NLP models should exhibit, and under what conditions they should be evaluated. Broadly speaking, generalisation is evaluated by assessing how well a model performs on a test dataset, given the relationship of this dataset with the data the model was trained on. For decades, it was common to put only one very simple constraint on this relationship: that the train and test data be different. This was achieved by randomly splitting a corpus into a training and a test partition. Generalisation was thus evaluated by training and testing models on different but similarly sampled data – or more precisely, 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. Since the first release of the Penn Treebank (Marcus et al., 1993), F1 scores went from values in the

high 80's at the end of the previous century (Collins, 1996; Magerman, 1995) and the first ten years of the current one (e.g. Petrov and Klein, 2007; Sangati and Zuidema, 2011) to scores up to 96 in the most recent past (Mrini et al., 2020; Yang and Deng, 2020). On the same corpus, performance for language modelling went from perplexity scores well above 100 (Kneser and Ney, 1995; Rosenfeld, 1996) to a score of 20.5 in 2020 (Brown et al., 2020). Progress in many areas of NLP has become even faster in the very last years. Scores for the popular evaluation set GLUE went from values between 60 and 70 at its release (Wang et al., 2018), to scores exceeding 90 less than a year after (most famously, Devlin et al., 2019), with performances on a wide range of tasks reaching and surpassing human level (e.g. Devlin et al., 2019; Liu et al., 2019b; Wang et al., 2019, 2018). Yet more recently, strongly scaled-up models (e.g. Chowdhery et al., 2022) showed astounding performances on almost all existing i.i.d. natural language understanding benchmarks.

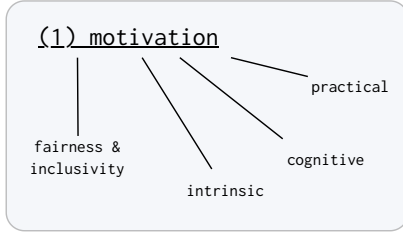
With this impressive progress, however, also came the realisation that for a neural network to reach very high or human-level scores on an i.i.d. test set does not imply that the model in fact robustly generalises to a wide range of different scenarios. In the recent past, we witnessed a surge of different studies pointing out generalisation failures in neural models that have state-of-the-art scores on random train-test splits (Blodgett et al., 2016; Khishigsuren et al., 2022; Kim and Linzen, 2020; Lake and Baroni, 2018; McCoy et al., 2019; Plank, 2016; Razeghi et al., 2022; Sinha et al., 2021, 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 generalise in a wide range of non-i.i.d. scenarios (Gardner et al., 2020; Kaushik et al., 2019; McCoy et al., 2019), that models over-rely on stereotypes (Parrish et al., 2022; Srivastava et al., 2022), or bank on memorisation rather than generalisation (Razeghi et al., 2022). Others, instead, discuss cases in which performances drop when the evaluation data differs from the training data in terms of genre, domain or topic (e.g. Malinin et al., 2021; Michel and Neubig, 2018; Plank, 2016), or when it is produced by different subpopulations (e.g. Blodgett et al., 2016; Dixon et al., 2018). Yet others focus on models' inability to generalise compositionally (Dankers et al., 2022; Kim and Linzen, 2020; Lake and Baroni, 2018; Li et al., 2021b), structurally (Sinha et al., 2021; Weber et al., 2021; Wei et al., 2021), to longer sequences (Dubois et al., 2020; Raunak et al., 2019), or to slightly different task formulations of the same problem (Srivastava et al., 2022).

The examples above are just a few examples in a long list of studies that aim to investigate the generalisation abilities of NLP models, focussing in particular on models and training regimes that score well on traditional train-test splits. Taken together, this body of work brings into question the kind of generalisation capabilities recent breakthroughs actually reflect, and how generalisation should be tested for – if not with i.i.d. splits. At the same time, these works differ amply in the definitions they give of generalisation, the assumptions they make about when and how models should generalise, and the evaluation settings they use. They encompass a wide range of generalisation-related research questions, and they use a wide range of different methodologies and experimental setups. Such differences make it difficult to understand how results in this area relate to each other, what types of generalisation are being addressed and which are neglected, what types of generalisation we should prioritise in NLP, and how we can adequately assess generalisation in the first place.

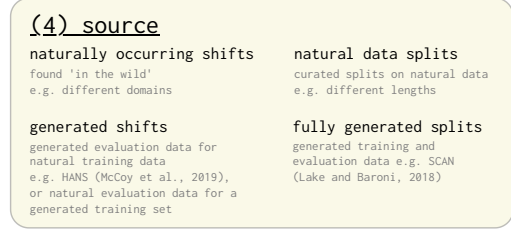
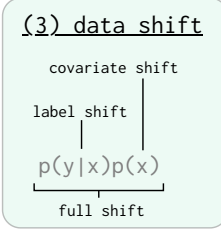
With this work, we aim to provide structure to the field of generalisation evaluation, to critically analyse the work that has been done so far, and to set the grounds for systematic generalisation testing to become the standard in any future modelling efforts. By carefully surveying existing work on generalisation evaluation, we identify five main axes of variation along which those studies differ. We incorporate those five axes in a taxonomy that can be used to better understand the heterogeneous landscape of generalisation testing, with as ultimate goal to help researchers better design and understand generalisation evaluation research in the future. The different axes in our taxonomy target the following five questions:

- What is the high-level *motivation* for a generalisation test? (Section 2)
- What is the *type* of generalisation the test is addressing? (Section 3)

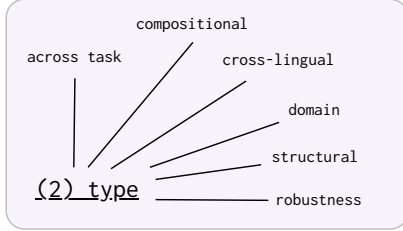
Generalisation studies have various motivations (1)...



They involve data shifts (3), where the data can come from natural or synthetic sources (4).



...and can be categorised into types (2).



These data shifts can occur in different stages of the modelling pipeline (5).

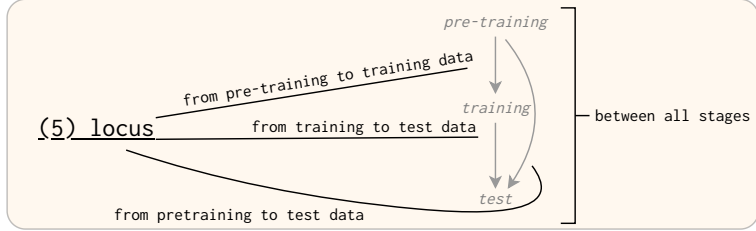


Figure 1: A graphical representation of the NLP generalisation taxonomy we present in this paper. The taxonomy consists of five different (nominal) axes, that describe the high-level *motivation* of the work (§ 2); the *type* of generalisation the test is addressing (§ 3); what kind of *data shift* occurs between training and testing (§ 4), what the *source* is of the data shift considered in the test (§ 5) and what the *locus* of the data shift is (§ 6)

- What kind of *data shift* occurs between training and testing? (Section 4)
- What is the *source* of the data shift considered? (Section 5)
- What is the *locus* of the data shift in the modelling pipeline? (Section 6)

We describe the meaning of these axes and the possible (nominal) values that generalisation studies can take on these axes, providing representative examples for each. Then, in Section 7, we use our axis-based taxonomy to review N studies. We present the results of the survey in comprehensive figures, which we use to describe the current landscape of generalisation testing in NLP, and to identify areas where more work is needed. We conclude by summarising our main findings from this extensive literature review and make concrete recommendations, outlining a vision of what generalisation testing should look like in the future.

In summary, our contributions are the following:

- i) We present an axis-based *generalisation taxonomy* that can be used to characterise generalisation studies in NLP;
- ii) We review N generalisation studies in NLP, along the five main axes of variation in this taxonomy;
- iii) With these survey results, we discuss the status of generalisation research in NLP;
- iv) We provide suggestions to steer the field towards more sound and exhaustive generalisation tests.

Along with this paper, we also present a website [www.genbench.github.io/taxonomy](http://www.genbench.github.io/taxonomy), where our survey results can be visualised dynamically, and where we encourage readers to add (new) generalisation studies that are not yet included.

## 2 High-level motivations for evaluating generalisation

The first axis we consider in our taxonomy is the high-level motivation of a generalisation study. We identified four closely intertwined goals of generalisation research in NLP, which we refer to as the *practical*, the *cognitive*, the *intrinsic*, and the *fairness* motivation.<sup>1</sup> We discuss each of them below.

**Practical: in what settings can the model be used?** One frequently posed motivation to study generalisation is of a highly practical nature; it concerns in what kind of scenarios a trained model can be successfully used. Questions with a primarily practical motivation often relate to how well models generalise to different domains or differently collected data. For instance, Michel and Neubig (2018) consider how well machine translation models trained on canonical text can generalise to noisy data from an internet platform; Lazaridou et al. (2021) investigate language model generalisation to different time periods; and Talman and Chatzikyriakidis (2019) investigate how well natural language inference (NLI) models generalise from one NLI dataset to another. Other questions that are frequently addressed from a practical perspective concern biases in the training data, and whether models robustly generalise to datasets that do not share these (spurious) biases (e.g. Behnke et al., 2022; Zhou et al., 2021)

**Cognitive: does the model generalise like a human?** A second high-level motivation that drives generalisation research is cognitively oriented, and can be separated into two underlying categories. The first category is related to model behaviour: human generalisation is a useful reference point for the evaluation of model generalisation in NLP, because human generalisation is known to be particularly powerful (e.g. Lake et al., 2017). Humans learn quickly, from fewer data than models (Linzen, 2020), and they easily (compositionally) recombine concepts they already know to understand concepts they have never before encountered. These feats are arguably also important for models; they therefore provide a good point of reference for generalisation testing and a compelling motivation for research toward better generalising models.<sup>2</sup> There is an evident overlap between cognitively-inspired and practical motivations: assuming human generalisation is strong, a model that generalises like a human should score well also on practically motivated tests. In our axes-based taxonomy, the difference between *cognitive* and *practical* resides mostly in the types of scenario that are considered in tests: are the scenarios artificially created to get a higher-level, isolated impression of how their behaviour compares to human-like generalisation, or are the scenarios realistic and practically relevant?

The second, more deeply cognitively inspired category contains work that evaluates generalisation in models to learn more about cognition and language (Baroni, 2021; Hupkes, 2020). Studies in this category investigate whether a particular model generalises primarily in order to derive new hypotheses about how human generalisation might work. For instance, Lakretz et al. (2021b) perform a detailed study of how LSTM models generalise to specific kinds of nested syntactic constructions, which they then use to inform a human experiment on the same syntactic constructions.

**Intrinsic: does the model capture the task correctly?** A third motivation to evaluate generalisation in NLP models, which cuts through the two previous motivations, appertains to the question “*did a model learn the task we intended it to learn, as we intended it to learn it?*”. The assumption underpinning this type of research as a whole is that if a model has truly learned the task it is trained to do, it should be able

---

<sup>1</sup>As we will see in what follows, it is at times difficult to tease apart the exact motivation of a generalisation study. Some studies genuinely stem from two or more motivations, and we mark them accordingly in our survey. More often, however, even for generalisation tests that may inform research along all four directions, it is possible to identify a main guiding motive.

<sup>2</sup>We do not always expect from a model the same type or level of generalisation a human exhibits. There are several cases in which models already generalise better than humans – consider, for instance, calculators, which since long outperform humans when it comes to arithmetic operations – and would be useless if they did not, as well as cases in which it is desirable for models to generalise better, for example across languages – something humans above a certain age typically do not excel at.

to execute this task also 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. For instance, researchers studying compositional generalisation (see § 3.1) assume that a correct understanding of language implies that the assumed compositional structure of language is captured. Under that assumption, a model should not have trouble to generalise to new inputs that are generated using the same compositional system. Others instead assume that true language understanding implies being able to use language across a wide variety of tasks (see Section 3.3). Yet others argue that if a model truly captures the relationship between two sentences in NLI tasks (e.g. Bowman et al., 2015a; Marelli et al., 2014; Williams et al., 2018), it should be able to do so across different data sets, even if those were sampled in a slightly different way. In studies that consider generalisation from this perspective, generalisation failures are taken as a proof that the model – in fact – did not learn the task as we intended it to learn it (but instead showed behaviour that made us think it did, for instance by relying on spurious patterns or non-generalisable heuristics).

**Fairness and inclusivity: does the model generalise in a fair and responsible way?** A last yet very important motivation for generalisation research is the desire to have models that are fair, responsible and unbiased. One category of studies driven by these concepts, often ethical in nature, asks questions about how well models generalise to diverse demographics, typically considering minority or marginalised groups (e.g. Bender et al., 2021; Blodgett et al., 2016; Koh et al., 2021), or investigates to what extent models perpetuate (undesirable) biases learned from their training data (e.g. Dixon et al., 2018; Hutchinson et al., 2020; Sheng et al., 2019).

Another line of research related to both fairness and inclusivity focusses 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 generalisation: models that generalise well should learn more quickly and require fewer data. The relationship of efficiency with fairness, inclusivity and responsibility stems from the idea that models that generalise well from small amounts of data are more inclusively applicable – for instance for low-resource languages or demographic groups for which little data is available. Furthermore, models that require less compute are more accessible for groups with smaller computational resources, and have a lower environmental impact (see, e.g. Strubell et al., 2019).

### 3 Types of generalisation

A second important consideration when it comes to characterising generalisation research, is what *type* of generalisation a test aims to evaluate. We identify and describe five types of generalisation that are frequently considered in the literature. Some generalisation tests are rooted in knowledge about human generalisation, such as those that target compositional (§ 3.1) or structural generalisation (§ 3.2). Others, instead, are motivated by more practical concerns, such as tests focussing on generalisation across tasks (§ 3.3), languages (§ 3.4) and domains (§ 3.5), or on the sensitivity of models to the exact data they are trained on (§ 3.6).

#### 3.1 Compositional generalisation

The first prominent type of generalisation that can be found in the literature is *compositional generalisation*, which is often argued to underpin human’s ability to quickly generalise to new data, tasks and domains (Fodor and Pylyshyn, 1988; Lake et al., 2017; Schmidhuber, 1990). Because of this strong connection with humans and human language, work about compositional generalisation often has a primarily cognitive motivation, although practical concerns such as sample efficiency, quick adaptation and



Figure 2: [DH](#): Infographic that illustrates how train and test differ for different generalisation types.

good generalisation in low-resource scenarios are frequently mentioned as additional or alternative motivations to study compositional generalisation (Chaabouni et al., 2021; Linzen, 2020, to give just a few examples). While it has a strong intuitive appeal and clear mathematical definition (Montague, 1970), compositional generalisation is not easy to pin down empirically. Here, we follow Schmidhuber (1990) 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 et al. (2020).

Compositionality involves mapping forms (e.g. phrases, sentences, larger pieces of discourse) to their meaning. It is therefore usually evaluated in sequence-to-sequence domains such as sequence classification (e.g. Bowman et al., 2015b; Hupkes et al., 2018; Veldhoen et al., 2016), machine translation (e.g. Dankers et al., 2022; Liu et al., 2021; Raunak et al., 2019), semantic parsing (e.g. Finegan-Dollak et al., 2018; Keysers et al., 2019; Kim and Linzen, 2020; Shaw et al., 2021) or other kinds of generation tasks (e.g. Hupkes et al., 2020; Lake and Baroni, 2018). As far as we know, there have been no explicit systematic attempts to evaluate compositionality in language models (LMs), or in an in-context learning setup.<sup>3</sup> If and how compositionality can be adequately evaluated in such a setup, where the domains of form and meaning are conflated in one space, is a question that is yet to be answered.<sup>4</sup>

### 3.2 Structural generalisation

Another category of cognitively-inspired generalisation instead focusses on the extent to which models can generate structurally (grammatically) correct forms, rather than on whether they can *understand* them (i.e. whether they can compositionally assign a correct interpretation to inputs). Because of this, structural generalisation is most straightforwardly evaluated in form-only models (i.e. language models).

<sup>3</sup>There are, however, several studies that focus on *structural* generalisation in such models. Contrary to compositional generalisation, structural generalisation does not focus on the ability of models to correctly interpret new inputs, or assign meanings to them, but only on whether they can generalise to their correct form. We will discuss structural generalisation in the next subsection.

<sup>4</sup>An interesting example to consider in this context is the qualitative study conducted by Brown et al. (2020) to test if GPT-3 can use novel words correctly in a sentence; as another example, a bit further away from traditional forms of compositionality, Talmor et al. (2020) finetune pretrained masked language models on multi-hop composition in question answering.

Furthermore, since evaluating structural generalisation requires understanding only the input domain, it is much more easily evaluated in completely natural setups. We will therefore focus only on work that considers structural generalisation in models trained on natural language. Structural generalisation studies typically focus on two broad categories: syntactic generalisation, and morphological generalisation.

**Syntactic generalisation** One category of structural generalisation studies focusses on *syntactic generalisation*: they consider whether models can generalise to novel syntactic structures, or novel elements in known syntactic structures. For instance, Jumelet et al. (2021) and Weber et al. (2021) filter out from the training data specific licensing environments for negative polarity items, and they test whether models nevertheless learn to generalise to such environments. It is unfortunately difficult to conduct this type of studies, which involve several different training corpora, using very large language models. On the one hand, their high training cost makes the necessary experiments computationally extremely expensive. On the other hand, generating specific test splits given knowledge of what is in the training data is often also not possible for such models, because their training data is not in the open domain. These limitations prevent researchers from controlling the relationship between the evaluation and training data, making it hard to assess to what extent the incidental examples reported for the large language models (most notably, in their respective release papers) are reflective of generalisation. Interesting exceptions are a few studies that do explicitly consider shifts between training and testing in the context of syntactic generalisation, such as those presented by Wei et al. (2021), Razeghi et al. (2022), and Elazar et al. (2022). Wei et al. (2021), in particular, investigate how the performance of pretrained language models in tests targeting syntactic rule learning is affected by a term’s training data frequency, by varying those frequencies in the training corpus. Razeghi et al. (2022), instead, focus on a larger model trained on more data, and while they do not systematically vary the training corpus, they do an elaborate analysis of how test performance in their trained model (GPT-Z) is affected by absolute and relative frequencies of specific terms in the model’s training data. Even more recently, Elazar et al. (2022) studies the causal effect of simple statistics from the training data, such as co-occurrences, on models’ prediction.

Note that the vast majority of other studies focussing on the syntactic abilities of language models (e.g. Giulianelli et al., 2018; Jumelet and Hupkes, 2018; Linzen et al., 2016; Warstadt et al., 2019, 2020) focus on whether and how models recognise, represent, and process syntactic information, or they try to discern the causal mechanisms by which models use such abilities (Amini et al., 2022; Elazar et al., 2021a; Feder et al., 2021). These works do not (explicitly) consider the relationship between the data they test on and the data that a model was trained on, and as such they do not directly study the models’ *generalisation* abilities across syntactic structures. We will not further discuss these studies, but in our map of generalisation literature (Section 7), we will include a few papers in which there is an implicit yet clear assumption that the test data substantially differs from the training data, for instance because it includes sentences created with semantically nonsensical words (Gulordava et al., 2018), or unusually deep levels of recursion (Lakretz et al., 2021a,b) that are not likely to naturally occur in corpora.

**Morphological generalisation** A second category of structural generalisation studies focusses on morphological inflection, a popular testing ground for questions about human generalisation. Papers focussing on morphological inflection (e.g. Corkery et al., 2019; Dankers et al., 2021; Kirov and Cotterell, 2018; Liu and Hulden, 2022; Malouf, 2017; McCurdy et al., 2020) are typically rooted in strong cognitive motivations. While most of this work considers i.i.d. train-test splits, recent studies have focussed on how morphological transducer models generalise across languages (e.g. McCarthy et al., 2019; Pimentel et al., 2021a; Vylomova et al., 2020) as well within each language (Calderone et al., 2021; Li and Wilson, 2021; Liu and Hulden, 2022; Pimentel et al., 2021b; Szolnok et al., 2021; Wilson and Li, 2021), taking inspiration from *wug* tests which are used in psycholinguistics to probe morphological generalisation to novel words in humans (Berko, 1958; Marcus et al., 1995). In principle, such studies could also be conducted with large language models but the lack of access to their training data

is, again, a complication for determining whether the supposedly novel words were truly never seen by the models.

### 3.3 Generalisation across tasks

A third and completely different direction of generalisation research considers the ability of a single model to adapt to multiple NLP problems. We refer to this ability as generalisation across tasks, or *cross-task* generalisation. Along with the great advancements in NLP models, in the past ten years, the nature of cross-task generalisation tests has quite substantially changed; we discuss this evolution in the present section.

**Multitask learning** Initially, cross-task generalisation was strongly connected to transfer and multitask learning (Collobert and Weston, 2008). In multitask learning, a model is either trained on a set of tasks and evaluated on those same tasks, or pretrained on some tasks and then adapted to others. As this setup favours approaches that benefit from positive transfer across tasks, it implicitly studies forms of cross-task generalisation.<sup>5</sup> Examples of benchmarks that were originally meant to address this kind of cross-task transfer – although they are not used as such any longer – are multitask benchmarks such DecaNLP (McCann et al., 2018), GLUE (Wang et al., 2018) and its successor SuperGLUE (Wang et al., 2019). In recent times, a common approach has been to formulate all tasks as sequence-to-sequence problems, a direction explored in the DecaNLP benchmark (McCann et al., 2018), as well as in modelling, by T5 (Raffel et al., 2020), exT5 (Aribandi et al., 2022) and UnifiedSKG (Xie et al., 2022), among others.

**The pretrain-finetune paradigm** In the context of multitask learning, cross-task generalisation was deemed an extremely challenging topic. This has changed with the relatively recently introduced *pretrain-finetune paradigm*, which has also shifted thoughts on how to evaluate cross-task generalisation. Rather than evaluating how learning one task can benefit to another, this paradigm instead gives a central role to the question of how well a model that has acquired some general knowledge about language during pretraining can be used to generalise to different kinds of tasks in a finetuning stage – i.e. a second round of training which involves task specific parameters (e.g. Devlin et al., 2019; Howard and Ruder, 2018; Liu et al., 2019b; Peters et al., 2018). Interestingly, in the finetuning stage, performance on the tasks themselves is typically evaluated with random train-test splits, and thus generalisation within individual tasks is not necessarily considered.

**Zero-shot and few-shot learning** The focus of cross-task generalisation studies has more recently shifted even further, to scenarios which consider how well pretrained language models fare in different tasks *without* any additional parameters.<sup>6</sup> In the most extreme case, this implies evaluating a language model directly on a range of tasks without any further training. To do so, tasks are reformulated as text-completion problems, such that language models can be *prompted* directly with a question representing a specific task (*zero-shot learning*), potentially preceded by a few examples (*few-shot learning*) (Radford et al., 2019). Datasets to do so are typically created by adapting conventional multitask datasets, where prompting templates are (often manually) designed for each task (e.g. Mishra et al., 2022; Wang et al., 2022; Weller et al., 2020). Unfortunately, studies that investigate the relationship between the training

---

<sup>5</sup>Notably, as illustrated by the work of Weber et al. (2021), the definition of *task* can be taken liberally in this context, ranging from traditional notions of NLP tasks, to considering subproblems of a single classic NLP task. For instance, while language modelling constitutes its own task, learning how to handle negative polarity items such as *any* or *ever* in a grammatically correct way can be considered one subtask.

<sup>6</sup>If the pretraining corpus is seen as a large collection of different uncontrolled task, this scenario is more similar to the original multitask learning scenario than the pretrain-finetune paradigm.



and test data are rare, which leaves many open questions in this area. Where Brown et al. (2020) report that data leakage from training had a small impact on their results, other recent work suggests that the impressive capabilities of large language models on zero- or few-shot learning tasks can largely be attributed to the presence of similar or identical examples in the training corpus (Han and Tsvetkov, 2022; Razeghi et al., 2022). Moreover, models have been reported to be sensitive to exact task formulation (Jiang et al., 2020; Schick and Schütze, 2021) and even to the order of the examples given in the few-shot setting (Lu et al., 2022), to some extent contradicting the intuitive idea of task understanding – and thus being considered as evidence against models’ generalisation ability.

**In-context finetuning** A different class of studies that considers task evaluation in the prompting setup are those that finetune a pretrained model with prompts from one set of tasks and then evaluates them on another set of tasks (e.g. Sanh et al., 2022; Wei et al., 2022; Zhong et al., 2021). Here, the relationship between task performance and generalisation is clearer than in the zero- and few-shot learning setups. While also in this case the pretraining corpus is uncontrolled, at least the relationship between the finetuning training and test data can be clearly monitored, and the performances on the test data with and without finetuning easily compared. Nevertheless, there are few studies that actually do so.

### 3.4 Generalisation across languages

A fourth type of generalisation, which has recently gained in popularity thanks to the strong improvements in English models, is generalisation across languages, or *cross-lingual* generalisation. Cross-lingual generalisation is highly relevant from a practical perspective: while the data for a selected amount of languages (English in particular) is plentiful, for many others, resources are much more scarce or virtually inexistent. Furthermore, strong generalisation across languages can contribute to the democratisation and inclusiveness of NLP, by increasing the coverage over languages of the world for which adequate models are available.

**Cross-lingual finetuning** There are several ways in which cross-lingual generalisation can be evaluated. Most existing cross-lingual studies focus on the scenario where labelled data is available in a single language (typically English), and the model is evaluated in multiple languages. A common approach to address this problem is to finetune a multilingually pretrained language model on the English labelled data, and to then transfer to the rest of the languages in a zero-shot fashion (e.g. Papadimitriou et al., 2021; Pires et al., 2019; Wu and Dredze, 2019).<sup>7</sup> For instance, Pires et al. (2019) show that Multilingual BERT (Devlin et al., 2019) finetuned on English generalises well even to languages with different scripts, but exhibits some systematic deficiencies that affect specific language pairs. Papadimitriou et al. (2021), instead, investigate how grammatical features generalise across languages for the same Multilingual BERT model.

**Multilingual learning** A second way in which cross-lingual generalisation can be evaluated, is by considering whether models trained on multiple languages at the same time (multilingual models) perform better than models trained on only one language. In multitask learning, approaches that are simultaneously trained on multiple tasks can be seen as an implicit evaluation of generalisation across tasks. Similarly, multilingual models trained on multiple languages can be seen as implicitly evaluating generalisation across languages. There is a large number of papers that investigate and proposes multilingual models, mostly in the domains of language modelling and machine translation (e.g. Aharoni et al., 2019; Al-Shedivat and Parikh, 2019; Costa-jussà et al., 2022; Fan et al., 2021; Zhang et al., 2020).

---

<sup>7</sup>Other approaches instead use machine translation to translate test sets into English and directly use an English model; or to translate the training data into another language and finetune a multilingual model on the augmented data. As this setup does not focus on generalisation per se, but rather depends on the quality of the translation model, we will not further discuss it.

Most of these papers have as main aim to introduce improved models, and they are not motivated by generalisation questions. Some, however, do include explicit generalisation experiments in their setup. For instance, Zhou et al. (2018) investigate how generalisation depends on the amount of data added for different languages; whereas Aharoni et al. (2019) investigate how zero-shot generalisation changes depending on the amount of different languages that a model is trained on.

**Multilingual benchmarks** As pointed out before, while the field of multilingual modelling is vast and associated with many interesting generalisation questions, papers in this area do not often focus explicitly on generalisation. We would, therefore, like to end this subsection by discussing the most important available multilingual benchmarks which can be used to evaluate cross-lingual generalisation. Multilingual benchmarks or datasets are created in a variety of ways. Several benchmarks are created by translating monolingual benchmarks into different languages, usually through a professional translation service (Artetxe et al., 2020; Conneau et al., 2018; Ebrahimi et al., 2022; FitzGerald et al., 2022; Lewis et al., 2020; Li et al., 2021a; Lin et al., 2021; Longpre et al., 2021; Mostafazadeh et al., 2016; Ponti et al., 2020; Williams et al., 2018; Xu et al., 2020; Yang et al., 2019; Zhang et al., 2019). Other multilingual benchmarks, instead, have been built by separately annotating each language via its native speakers (e.g. Adelani et al., 2021; Asai et al., 2021; Clark et al., 2020; Muller et al., 2021). Yet another way to construct multilingual benchmarks is to leverage existing resources that cover multiple languages. For instance, Wikipedia has been used as a resource to derive multilingual benchmarks (Botha et al., 2020; Liu et al., 2019a; Pan et al., 2017; Rahimi et al., 2019), and several multilingual summarisation datasets have been created by extracting article-summary pairs from online newspapers or how-to guides (e.g. Hasan et al., 2021; Ladhak et al., 2020; Nguyen and Daumé III, 2019; Scialom et al., 2020; Varab and Schluter, 2021). Various linguistic resources have also been exploited: for instance, the Universal Dependencies treebank (Nivre et al., 2020) has been used to evaluate cross-lingual part-of-speech tagging, and multilingual WordNet and Wiktionary have been used to build XL-WiC (Raganato et al., 2020), an extension of WiC (Pilehvar and Camacho-Collados, 2019) which reformulates word sense disambiguation in 12 languages as a binary classification task. Finally, in the same spirit of GLUE and SuperGLUE for English, there are also several aggregated benchmarks that include selected sets of benchmarks previously proposed by others (e.g. Hu et al., 2020; Liang et al., 2020; Ruder et al., 2021; Wang et al., 2022), which allow to evaluate cross-task and cross-language generalisation simultaneously.

### 3.5 Generalisation across domains

The next category considers a type of generalisation that is required in naturally occurring scenarios (more so than the types discussed so far), and is thus very important in practice: generalisation across different domains. As examples of the practical relevance of cross-domain generalisation, consider: a sentiment analysis model trained to classify the sentiment of reviews for certain products which then needs to generalise to newly commercialised products, necessarily not represented in its training data (Ryu et al., 2018; Tan et al., 2019); a model trained on data collected from one demographic which is then asked to generalise to the entire population (Blodgett et al., 2016); or a machine translation model trained on canonical text and then expected to generalise to noisy data from an internet platform (Blodgett et al., 2017; Michel and Neubig, 2018) or to data from a different real-world domain (Malinin et al., 2021). While there is not a precise definition of what constitutes a domain, different domains broadly refer to collections of texts exhibiting different topical and/or stylistic properties, such as different genres or formality levels. Again, examples help us clarify this definition. MultiNLI (Williams et al., 2018), for instance, collects training corpora from five different genres (e.g. fiction and telephone conversations), and includes both an in-domain evaluation set with corpora from those five genres, as well as an out-of-domain evaluation set with corpora from five more sources (e.g. face-to-face conversations and the 9/11 public report). Blodgett et al. (2016) consider how language tools trained on data collected from white

African-American speakers generalises to text from non-white ones. Fried et al. (2019) compare how neural and non-neural constituency parsers generalise on out-of-domain treebanks (e.g. on a treebank of biomedical texts), whereas Artetxe et al. (2021) compare how sparse and dense language models generalise in-domain and out-of-domain (on texts from ArXiv, Github, OpenSubtitles, among many other sources). Kamath et al. (2020) study the problem of selective question answering under domain shift, where the test distribution includes both in-domain and out-of-domain questions and the model must abstain from answering when not confident. Connected to this last type of study, there is a substantial body of work in out-of-domain *detection* (Hendrycks et al., 2020; Lane et al., 2007; Ryu et al., 2017, 2018; Tan et al., 2019).

Domain generalisation has often been studied in connection with domain adaptation, the problem of adapting an existing general model to a new domain (Daumé III, 2007). This has been a very active research area in machine translation (Axelrod et al., 2011; Bertoldi and Federico, 2009; Chu et al., 2017; Chu and Wang, 2018; Freitag and Al-Onaizan, 2016; Hu et al., 2019; Joty et al., 2015; Koehn and Schroeder, 2007; Luong and Manning, 2015; Wang et al., 2017a,b), with several standard datasets (Malinin et al., 2021; Michel and Neubig, 2018) and dedicated tracks in popular shared tasks like WMT (Bojar et al., 2019; Specia et al., 2020). In addition to machine translation, domain adaptation has also been studied in part-of-speech tagging (Blitzer et al., 2006), sentiment analysis (Blitzer et al., 2007) and language model pre-training (Gururangan et al., 2020), among others.

Finally, domain generalisation is closely related to temporal generalisation, where the training data is 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. This problem has been studied in an as yet limited range of tasks, including language modelling (Lazaridou et al., 2021), named entity recognition in social media (Derczynski et al., 2016; Fromreide et al., 2014; Rijhwani and Preotiuc-Pietro, 2020), named entity disambiguation (Agarwal et al., 2018), document classification (He et al., 2018; Huang and Paul, 2018, 2019) and sentiment analysis (Lukes and Søgaard, 2018).

### 3.6 Generalisation in the context of robustness

One last category of generalisation research considers how *robust* models are to changes with respect to their exact training data. Studies of this kind consider train-test shifts that stem from the data collection process. Different from most of the previous categories discussed in Section 3, such shifts are generally unintended and can be hard to spot. Existing research therefore focusses on characterising such scenarios and understanding their impact. Oftentimes, studies in this category intend to show that models do not generalise in the way we would expect them to, because the training data was in some very subtle manner not representative of the true target distribution. This line of work is based on the idea that models should learn task solutions that abstract away over specific, often spurious correlations that may occur in the training data, i.e. models should learn the underlying generalising solution that humans associate with the task (e.g. Gururangan et al., 2018; McCoy et al., 2019; Talman and Chatzikyriakidis, 2019). We refer to tests that assess whether model performance is independent from the exact training data with the term *robustness evaluation*. Robustness evaluation is very important from a practical perspective. If a model has a strong sensitivity to spurious patterns in the training data, and is then tested on data collected with the same bias, this can result in overestimating its performance – either generally or on specific test cases – with potentially harmful consequences, for instance when a model does not generalise well to particular population demographics. Below, we discuss three common scenarios associated with robustness evaluation.

**Annotation artefacts** A scenario that frequently occurs in robustness studies is one where there are *annotation artefacts* in the training data, which may result in overestimation of a model’s performance on a particular task. Artefacts occur particularly frequently when datasets are collected through crowd-

sourcing. Crowdsourced datasets often depend strongly on how exactly the annotation procedure was set up, with subtle artefacts as a consequence. For instance, annotators may naturally tend to minimise their cognitive effort, resorting to patterns that models learn to exploit. Popular NLI datasets like SNLI (Bowman et al., 2015a) and MultiNLI (Williams et al., 2018) have been found to be particularly susceptible to such artefacts. For instance, Gururangan et al. (2018) and Poliak et al. (2018) showed that a hypothesis-only baseline performs better than chance, due to its exploitation of spurious patterns in word choice and grammatical features (e.g. negation being indicative of the *contradiction* class). d that NLI models do not generalise well across different datasets. Besides NLI, other tasks like question answering have also been reported to suffer from annotation artifacts (Jia and Liang, 2017; Kaushik and Lipton, 2018), even when such artifacts were deliberately and consciously avoided during the annotation process (Elazar et al., 2021b). Finally, Lewis et al. (2021) showed that open-domain question answering datasets have a high-overlap between train and test instances, revealing that memorisation plays a bigger role in these benchmarks than previously assumed.

**Standardised splits** Another line of work questions the way we use data splits in general, and in particular the extent to which scores on standardised splits that stay static over time are reflective of a model’s generalisation abilities. For instance, Gorman and Bedrick (2019) show that models perform much worse on random train-test splits than the reported state-of-the-art performances on a standardised split. Sogaard et al. (2021) go even further, and advocate for the use of heuristic and adversarial splits, where a model’s capability for generalisation is challenged directly – for instance by putting all longer sentences in the test set, or by splitting the data to maximise the difference between train and test set.

**Subpopulation bias** A third scenario in which robustness and performance overestimation play a role is the case where certain demographics are under- or over-represented in the training data. As this may result in models that generalise poorly to specific demographic groups, it is a particularly harmful case of overestimation. For instance, Dixon et al. (2018) show that toxicity classifiers suffer from unintended bias, caused by certain identity terms being disproportionately represented in the training data (e.g. “*I am a gay man*” being assigned high toxicity scores because of “*gay*” being often used in toxic comments). Similarly, Park et al. (2018) show that abusive language detection models exhibit gender bias, which is caused by the training data being imbalanced. Blodgett et al. (2016) show that dependency parsing and language identification tools perform poorly on text from non-white African-American speakers. As a way to detect such imbalances and thus systematically avoid such cases of overestimation, Koh et al. (2021) propose to evaluate models by their worst-group accuracy, rather than the average accuracy across all demographic groups, in their CivilComments-Wilds dataset (a variant of the CivilCommons toxicity classification dataset released by Borkan et al., 2019).

The examples above demonstrate that evaluating generalisation in the context of robustness can be driven by several different motivations. Some studies are motivated by more practical concerns, or conducted to gain a better perspective on intrinsic task understanding, but robustness evaluation is also particularly important when the goal is to have fair and unbiased NLP models.

## 4 Shift type: what kind of shift is considered?

As we have seen in the previous section, tests to evaluate generalisation may differ in terms of their *motivation* and the *type* of generalisation that they target. What they instead share, is that they all focus on cases in which there is a form of *shift* between the data a model was (pre)trained on and the data that was used for evaluation. In the third axis of our taxonomy, we discuss how shifts between the datasets used in a generalisation experiment can be characterised.

We formalise the differences between the test, training and potentially pretraining data involved in generalisation tests as shifts between the respective *data distributions*:

$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) \quad \text{test} \quad (1)$$

$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) \quad \text{training / finetuning} \quad (2)$$

$$p(\mathbf{x}_{\text{ptr}}, \mathbf{y}_{\text{ptr}}) \quad \text{pretraining} \quad (3)$$

By expressing these data distributions as the product of the probability of the input data  $p(\mathbf{x})$  and the conditional probability of the output labels given the input  $p(\mathbf{y}|\mathbf{x})$  –

$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = p(\mathbf{x}_{\text{tr}}) p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}}) \quad (4)$$

$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) = p(\mathbf{x}_{\text{tst}}) p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \quad (5)$$

we can define four main types of relations between any two data distributions.<sup>8</sup> One of these four types constitutes the case in which there is no shift in data distributions – i.e. both  $p(\mathbf{x}_{\text{tr}}) = p(\mathbf{x}_{\text{tst}})$  and  $p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}}) = p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}})$ . This matches the i.i.d. evaluation setup traditionally used in machine learning. As discussed earlier, this type of evaluation, also referred to as *within-distribution generalisation*, has frequently been reported not to be indicative of good performance for the more complex forms of generalisation that we often desire from our models. We will therefore not further discuss it here, but instead focus on the other three cases, commonly referred to as *out-of-distribution* (o.o.d.) evaluation.

**Covariate shift** The most commonly considered data distribution shift in o.o.d. generalisation research is one where  $p(\mathbf{x}_{\text{tst}}) \neq p(\mathbf{x}_{\text{tr}})$  and  $p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) = p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}})$ . In this scenario, often referred to as *covariate shift* (Moreno-Torres et al., 2012; Storkey, 2009), 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  $p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}})$ .

Most research in NLP on evaluating generalisation focuses on covariate shift. For example, challenge test sets such as HANS (McCoy et al., 2019), PAWS (Yang et al., 2019), or the COGS (Kim and Linzen, 2020) test set contain deliberately unusual, out-of-distribution examples, selected or generated to violate invalid heuristics in assigning labels to data samples. Less deliberate cases of covariate shift are evaluated in out-of-domain detection or robustness evaluation studies, such as those conducted by Ryu et al. (2018) and Tan et al. (2019) on real-world datasets. Tan et al. (2019), e.g., assume that the process by which the sentiment of a sentence is to be computed does not change, but the data that this process needs to be applied to does. Of the three o.o.d. shifts we discuss in this section, covariate shift is the shift that can be more easily solved without performing additional training or pre- or post-processing. As we will see in the next paragraphs, a common approach to address other, more complex shifts, is to turn them into covariate shifts.

**Label shift** The second type of shift corresponds to the case in which there is no difference between the input distributions,  $p(\mathbf{x}_{\text{tst}}) = p(\mathbf{x}_{\text{tr}})$ , but instead in the conditional distributions of the labels/output:  $p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \neq p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}})$ . We refer to this case as *label shift* but it is also known as *concept shift* (Moreno-Torres et al., 2012). Label shift can happen within the same task when there is a change of domain – e.g. 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; when there are inter-annotator disagreements; or when there is a temporal shift in the data (see § 3.5). Another common case of label shift is a change in

---

<sup>8</sup>For clarity, we leave pretraining distributions aside and focus on train-test shifts, as this is the most intuitive setting. However, the shifts described in this section can be used to describe the relation between any two data distributions involved in a modelling pipeline.

	$P(\mathbf{x})$	$P(\mathbf{y} \mathbf{x})$
No shift		
Covariate shift	✓	
Label shift		✓
Full shift	✓	✓

Table 1: Types of data distribution shifts. [DH: replace with figure](#)

task (as in § 3.3), where the meaning of the labels themselves changes as well. For example, the same sentence may need to be binarily classified for sentiment in some cases, and for toxicity in others. In even more extreme cases, the labels themselves might change, for example when shifting from language modelling (where the set of labels is the language vocabulary) to POS-tagging. In NLP studies, label shift is more often seen as an obstacle that needs to be overcome than as a setting in which models are directly evaluated: if the same example has contradictory labels in training and test data, it is unclear whether a correct decision at test time should be considered good generalising behaviour.

There are two main ways in which label shift is typically addressed, and framed as a simpler generalisation problem. The first is by adding an additional finetuning step (Devlin et al., 2019; Howard and Ruder, 2018; Peters et al., 2018, i.a.), or continual learning phase (Biesialska et al., 2020; Sun et al., 2020). In that scenario, there is a label shift between the pretraining and finetuning training data, but not between the finetuning training and testing data. The level at which generalisation is (somewhat implicitly) evaluated in that case, is then the pretraining level: does my pretrained model adapt well to different conditional label distributions when further trained? The second way to address label shift is to augment the input data with domain or task indicators (e.g. Brown et al., 2020; Raffel et al., 2020). We saw before that the phrase ‘*it doesn’t run*’ can be both positive and negative, depending on what it describes. Without further information, it is impossible for a model to infer the correct meaning. However, if we add an indicator that specifies the domain (`review for mascara:...`, `review for software:...`), the problem is converted into a covariate shift (or potentially even no shift, if both indicators are represented in the two distributions at hand), which then can be solved by correctly generalising. Something similar happens in the in-context learning setup: by adding a *prompt* that describes what needs to be done with the input, label shifts caused by a change of task are turned into shifts that can be solved without further finetuning (see e.g. Brown et al., 2020; Schick and Schütze, 2021; Bach et al., 2022).

**Full shift** The most extreme case of shift is the case in which both  $p(\mathbf{x})$  and  $p(\mathbf{y}|\mathbf{x})$  change simultaneously:  $p(\mathbf{x}_{\text{tst}}) \neq p(\mathbf{x}_{\text{tr}})$ ,  $p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \neq p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}})$ . We may encounter such a situation when switching languages in sequence-to-sequence or classification tasks (as described in § 3.4); when changing modality, as from linguistic to visual processing (Lu et al., 2021); or when switching data types completely from language to gameplay (Ciolino et al., 2020), robotics (Jang et al., 2021), and other non-linguistic (Papadimitriou and Jurafsky, 2020) or non-textual data (Kao and Lee, 2021). Similarly to label shift, these *full shifts* are often turned into a different type of shift which can be more easily addressed without retraining, and they are not directly used to evaluate generalisation. There also exist less extreme cases, where both the input and the conditional label distribution vary less drastically. Consider again the example sentence ‘*it doesn’t run*’, which can be both positive and negative depending on what it describes. We have described this before as label shift, under the assumption of a stable input distribution (e.g. in the case where training and test dataset contain reviews of all product types). However, if the training and test dataset were domain-specific (e.g. *body care* vs. *software*), then the label shift described above would be accompanied by covariate shift, thus becoming a case of full shift. Such cases can more easily be used for direct generalisation evaluation.



Figure 3: [DH](#): Figure that illustrates different sources of splits/shifts.

#### 4.1 On detecting shift type

We conclude this section by pointing out that while from a formal perspective the shifts that we describe are well-defined, they may difficult to tell apart in practice, because the base distributions by which natural languages are ‘generated’ are rarely fully known. As a consequence, it is often extremely difficult to determine what the relationship between two different natural datasets is. While in some cases there is nevertheless little discussion on the type of shift that occurs between two datasets, in other cases, it might be unclear if there is an actual shift, or what its nature is. When classifying shifts in our survey, we will focus on cases where authors (i) explicitly consider the relationship between the data distributions they use in their experiments and (ii) the assumptions they make about this relationship are either well-grounded in the literature (e.g. it is commonly assumed that switching between domains constitutes a covariate shift) or actually empirically verified. Nevertheless, we identify numerous studies that claim to be about generalisation 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. Sometimes, the assumed shift is not explicitly checked because it is considered plausible given general (linguistic) knowledge about language. Consider, for instance, how Gulordava et al. (2018) and Lakretz et al. (2021b), as discussed earlier in Section 3.2, regard sentences with semantically non-sensical words and unusually deep levels of recursion as out-of-distribution with respect to the training data. Other times, the relationship between training and test data is not investigated because the researchers do not have access to the training data. The BigBench benchmark (Srivastava et al., 2022), for instance, contains several tasks that might measure generalisation, but the training datasets of the models investigated are not in the public domain. Yet in other cases, the training data is available to the authors of the paper, but simply no extensive analysis is presented (e.g. Brown et al., 2020; Chowdhery et al., 2022). In our survey, we also consider this entire body of work, which we mark *assumed shift*.

## 5 Data sources: how are the train and test data produced?

In the previous section, we considered what kind of shifts may occur in generalisation tests. We now focus on a related relevant dimension, that expresses how shifts are produced or found, or, in other words, what the *source* is of the differences occurring between the pretraining, training and test data

distributions. Do shifts naturally occur between existing corpora, or are they the result of deliberate splitting of a corpus? Is the test set generated or selected with a particular kind of shift in mind, or is all data generated? In the fourth axis of our taxonomy, we consider how the pretraining, training and test data distributions – and the shifts between them – are produced. We distinguish four different sources of shifts: (i) *naturally occurring shifts*, shifts occurring naturally between different corpora; (ii) *splits of natural corpora*, in which both the training and pretraining data are fully natural, but they are partitioned along a specific dimension; (iii) *generated shifts*, where the training data is natural, but the test data is designed with a specific distribution shift in mind; and (iv) *fully generated datasets*, where all data involved is generated.

To formalise the description of these different sources of shift, we consider the unobserved *base distribution* which describes all data considered in an evaluation test:

$$p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}}, \tau) \quad \text{base} \quad (6)$$

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

$$p(\mathbf{x}_{\text{ptr}}, \mathbf{y}_{\text{ptr}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} | f_{\text{pretrain}}(\tau) = \text{true}) \quad (7)$$

$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} | f_{\text{train}}(\tau) = \text{true}) \quad (8)$$

$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} | f_{\text{test}}(\tau) = \text{true}) \quad (9)$$

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

**Naturally occurring shifts** The first option we consider is the scenario in which shifts naturally occur between different corpora. In such cases, the variable  $\tau$  refers to properties that naturally differ between collected datasets. What characterises this type of shift source, is that both the data partitions of interest are naturally occurring corpora, to which no systematic operations are applied: for the purposes of a generalisation test, experimenters have no direct control over the partitioning scheme  $f(\tau)$ . Examples of naturally occurring shifts emerge from splits containing data from different annotators (Geva et al., 2019), sources or domains (e.g. Artetxe et al., 2021; Talman and Chatzikyriakidis, 2019), data sampled from different populations (e.g. Dixon et al., 2018; Talat et al., 2018) data from different points in time (e.g. Lazaridou et al., 2021), or separately collected corpora targeting the same task, such as MNLI (Williams et al., 2018) and WNLI (Wang et al., 2018). In this category we also include cross-task and cross-lingual generalisation studies in which all corpora involved are natural corpora (e.g. FitzGerald et al., 2022; Mishra et al., 2022).

**Splits of natural corpora** A slightly less natural setup is the one in which a natural corpus is considered, but it is artificially split along specific dimensions. The primary difference with the previous category is that the variable  $\tau$  refers to data properties along which data would not naturally be split – such as the length or complexity of a sample – and thus that experimenters have control over the partitioning scheme  $f(\tau)$ . Raunak et al. (2020), for instance, split naturally occurring machine translation corpora such that longer sentences occur in the test data, and Weber et al. (2021) split a language modelling corpus such that the training data does not contain specific types of negative polarity item



licensors. Other examples of natural data splits could be splits that maximise compound divergence to investigate compositionality (Keysers et al., 2019).<sup>9</sup>

**Generated shifts** The third category on our source of shift axis concerns the case in which one data partition (usually the *training* set) is a fully natural corpus, but the other partition is designed with specific properties in mind, to address a generalisation aspect of interest. Not only do the experimenters control the partitioning scheme, but they can also influence the underlying base distributions (Eq. 6) by arbitrarily constructing one of the partitions. Data in the constructed partition may avoid or contain specific (syntactic) patterns (Bhargava et al., 2021; Cui et al., 2022), violate heuristics about gender (Dayanik and Padó, 2021; Libovický et al., 2022), or include unusually long or complex sequences (Lakretz et al., 2021a; Raunak et al., 2019). As an example of this shift source, Dankers et al. (2022) investigate compositionality in MT models trained on fully natural corpora by constructing test data that addresses compositional generalisation given the specific properties of the training corpus. For NLI, McCoy et al. (2019) design a test set that cannot be solved with models that rely on specific heuristics. Another category of studies that fit into this type are those with *adversarial* test sets, generated either by humans (Kiela et al., 2021) or automatically using a specific model (e.g. Sakaguchi et al., 2021; Zellers et al., 2018). In the examples above, all of the constructed data occurs in the test data; note that the opposite – where instead the *training data* is synthetic or generated and the test data natural – is also possible, yet less common.

**Fully generated or selected splits** The last category we consider are splits that use only generated, or even fully synthetic data. Generating data is often the most precise way of measuring specific aspects of generalisation, as experimenters have direct control over both the base distribution and the partitioning scheme. Sometimes the data involved is entirely synthetic (e.g. Hupkes et al., 2020; Lake and Baroni, 2018), other times it is templated natural language, or a narrow selection of an actual natural language corpus (e.g. Keysers et al., 2019; Kim and Linzen, 2020). Generated splits can vary in a number of different dimensions. Sometimes,  $\tau$  is a simple observable data property. For instance, Hupkes et al. (2020) split their corpus based on the presence of particular function pairs  $\mathcal{P}$ , implicitly setting  $\tau = \mathcal{P} \in x$ . In some cases,  $\tau$  may also be defined relative to the  $\tau$  of other examples, and can only be computed globally, such as in the case of *maximum compound divergence* splitting (Keysers et al., 2019).

## 6 Locus of shift: between which data distributions does the shift occur?

In the previous sections, we discussed high-level motivations for studying generalisation in neural NLP models, types of generalisation that have been frequently evaluated in the literature, kinds of data distribution shifts, and possible sources of data shift. These four axes demonstrate the depth and breadth of generalisation evaluation research, and they also clearly illustrate that generalisation is evaluated in a wide range of different experimental setups. What we have not yet explicitly discussed is between which data distributions those shifts can occur (the *locus* of the shift), and how that impacts which part of the modelling pipeline is evaluated.

Given the three data distributions that we have considered in § 4, there are four possible loci of shifts: shifts only between the *training and the test data*, shifts only between the *pretraining and the training data*, shifts only between the *pretraining and the test data*, and shifts between *all data distributions*. The locus of shift determines what component of the modelling pipeline can be assessed by a generalisation test, and thus impacts what kind of generalisation questions can be asked. For instance, a shift between pretraining and training allows to investigate if a particular pretraining procedure is successful, whereas a shift between train and test instead can be used to evaluate a model instance or training procedure. We

<sup>9</sup>Keysers et al. (2019) themselves do not apply this split to fully natural data



Figure 4: **DH:** Figure that illustrates different shift loci.

describe the loci of shift as well as how they interact with the different components in the modelling pipeline with the aid of three *modelling distributions*. These modelling distributions correspond to the different stages in contemporary machine learning pipelines – testing a model, training it, and potentially pretraining it:

$$p(\mathcal{Y}_{\text{tst}} \mid \mathcal{X}_{\text{tst}}, \theta^*) \quad \text{model} \quad (10)$$

$$p(\theta^* \mid \mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}}, \phi_{\text{tr}}, \hat{\theta}) \quad \text{training/finetuning} \quad (11)$$

$$p(\hat{\theta} \mid \mathcal{X}_{\text{ptr}}, \mathcal{Y}_{\text{ptr}}, \phi_{\text{pr}}, \theta_0) \quad \text{pretraining} \quad (12)$$

In these equations,  $\phi$  broadly denotes training and pretraining hyperparameters,  $\theta$  refers to model parameters, and  $\mathcal{X}, \mathcal{Y}$  indicate sets of inputs ( $\mathbf{x}$ ) and their corresponding output ( $\mathbf{y}$ ).

In short, Equation 10 defines a model instance, which specifies the probability distribution over the target test labels  $\mathcal{Y}_{\text{tst}}$ , given the model’s parameters  $\theta^*$  and a set of test inputs  $\mathcal{X}_{\text{tst}}$ . Equation 11, instead, defines a training procedure, specifying a probability distribution over model parameters  $\theta^* \in \mathbb{R}^d$  given a training dataset  $\mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}}$ , a set of training hyperparameters  $\phi_{\text{tr}}$ , and a (potentially pretrained) model initialisation  $\hat{\theta}$ . Lastly, Equation 12 defines a pretraining procedure, specifying a conditional probability over the set of parameters  $\hat{\theta}$ , given a pretraining dataset, a set of pretraining hyperparameters  $\phi_{\text{pr}}$ , and a model initialisation.<sup>10</sup> Between which of these stages a shift occurs impacts which of these modelling distributions can be evaluated. We discuss the different potential loci of shifts below.

**The (finetune) train-test locus** Probably the most commonly occurring locus of shift in generalisation experiments is the one between (finetuning) train and test data. This locus occurs in the classic setup where a model is trained on some training data and then directly evaluated on a shifted (out-of-distribution) test partition, or when a model is evaluated on a finetuning test set that contains a shift with respect to the finetuning training data. Experiments of the former category are, for example, those testing compositional (see § 3.1) and structural generalisation (§ 3.2), and frequently also domain generalisation (§ 3.5). An example of the latter category would be a test that investigates how well one pretrained model generalises to an o.o.d. finetune train/test set (Damonte and Monti, 2021; Kavumba et al., 2022).

<sup>10</sup>Note that this formalisation generalises to the *training from scratch* paradigm when  $\mathcal{X}_{\text{ptr}}, \mathcal{Y}_{\text{ptr}} = \emptyset, \emptyset$ , and to the *in-context-learning* setup when  $\mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}} = \emptyset, \emptyset$ .

Note that very frequently, studies evaluating o.o.d. splits during finetuning, include also a comparison between different pretraining procedures (e.g. they investigate whether BERT or RoBERTa generalises better to an o.o.d. finetuning test set). Such studies usually investigate a shift from the pretraining to finetuning training data (typically a label shift), as well as a shift in the finetuning stage, and we will mark them as having *multiple loci*, as further discussed in the last paragraph of this section.

Studies with the (finetune) train-test locus can assess two different parts of the modelling pipeline. In some cases, researchers investigate the generalisation abilities of a particular *model instance* (i.e. a set of parameters  $\theta^*$ , as described in Equation 10). The study then focuses on the evaluation of a single model instance – typically made available by others – without considering how exactly it was trained, and how that impacted the model’s generalisation behaviour. For example, someone might investigate how OPT (Zhang et al., 2022), given its training data, generalises to different test-sets, without knowing the details about how this model was trained. Alternatively, researchers might evaluate one or more training procedures, by considering if the *training distribution* results in model instances that generalise well – for example to study whether training with different optimisers results in model instances with different generalisation behaviour. While also this case requires evaluating model instances, the focus of evaluation is not on one particular model instance, but rather on the procedure that generated multiple model instances.

**The pretrain-train locus** A second potential locus of shift is between the pretraining and training corpus. Experiments with this locus evaluate whether a particular pretraining procedure, as described in Equation 12, results in models (or: parameter sets  $\hat{\theta}$ ) that are useful when further trained on different tasks or domains. For instance, Artetxe et al. (2021) consider which pretraining procedure shows best downstream generalisation to a number of different tasks, Tian et al. (2021) investigate how well pretrained models generalise to a newly proposed first-order-logic dataset, and Freitag and Al-Onaizan (2016) test how well a pretrained NMT model can adapt to different domains. Crucially, we classify studies to have a pretrain-train locus only when in their second training stage – that is necessarily required to have this locus – they use i.i.d. splits. If also the finetuning stage contains a shift, the study has multiple loci (as described below).

**The pretrain-test locus** The third potential locus of shift is between pretraining and testing data. This locus occurs when a pretrained model is not further updated but evaluated directly ( $\mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}} = \emptyset, \emptyset$ ) – as frequently happens in in-context learning setups (e.g. Lin et al., 2021; Zhang et al., 2022) – 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 ( $\theta^* = \hat{\theta}$ ) 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 (e.g. a language modelling objective). Furthermore, while generalisation studies with a train-test locus almost always explicitly consider the relationship between training and testing data, this is frequently not the case with pretrain-test studies in the in-context learning setup: often, they do not explicitly consider the relationship between training and test data, but merely assume a shift occurs between those stages (e.g. Radford et al., 2019).

**Multiple loci** The last option on our locus axis, which we already mentioned above, is the *multiple loci* class, which we use for works that consider, in a single study, multiple shifts between different parts of the modelling pipeline. In other words, experiments of this type present shifts both between the pretraining and training data, as well as between the training and test data.<sup>11</sup> Crucially, multiple-loci

<sup>11</sup>We do not distinguish cases where the test data is shifted with respect to the pretraining data from cases where it is not, as the latter are very uncommon. It is, however, possible to set up an experiment where the pretraining and test data are drawn

experiments evaluate all stages of the modelling pipeline at once: they consider both how generalisable the models produced by the pretraining procedure are, as well as whether generalisation happens in the finetuning stage itself. For instance, some studies compare how well models with different pretraining procedures (e.g. BERT vs RoBERTa) generalise to o.o.d. splits during finetuning (e.g. Tu et al., 2020), others how different multilingual pretraining procedures perform across-language task generalisation in a finetuning stage (e.g. FitzGerald et al., 2022; Hu et al., 2020; Yanaka et al., 2021). Because multiple-loci experiments necessarily also contain multiple shifts, we mark them as *double shifts* in the shift type axis. The nature of these shifts may not be the same: the shift from pretraining to training may be of any type, while the shift from training to test is often a less extreme covariate shift.

## 7 A review of existing generalisation research

In the previous sections, we have presented a taxonomy containing five categorical axes along which generalisation research can be characterised, providing examples for each of the different values studies might take on those axes. Here, in this last section, we use our taxonomy to characterise a large amount of existing generalisation research, with the aim to create a comprehensive map of existing research, as well as to identify gaps. On our website<sup>12</sup>, we present several interactive plots of results, that the reader can use to get a more in-depth view of how generalisation research in NLP is structured, to generate different kinds of plots to support their own work, to understand how their own work fits in with the rest of the literature or which areas might be promising to address, or to get relevant citations for their related work section. We also provide instructions to contribute to the review, by proposing to add new studies or studies we may have missed, or by proposing corrections to studies that might have been misqualified on one of their axes values. In this section, instead, we present our main findings, illustrated by several figures.

### 7.1 Setup

Before we arrive at our main conclusions, we first briefly describe the procedure we used for both selection and annotation.

**Paper selection** For our literature review, we annotated 200 papers in total. Part of this selection was made through a substantive literature review of the main authors, resulting in . To ensure that the selected papers were not biased to their expertise, however, we also did a search through the acl anthology, selecting any papers that had the word *generalisation*, *generalise*, *generalization* or *generalize* in their title, and for from the abstract and title could be inferred that they were in fact addressing a generalisation question. This resulted in a total of 601 papers to annotate. While the conclusions in this paper pertain this specific selection, we intend to keep expanding the amount of papers on our website, either when new generalisation papers come out, or when existing generalisation papers are sent to us.

**Annotation setup** We annotated all papers along the five axes in our taxonomy, asking for every paper the following five questions:

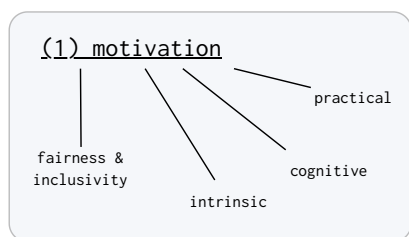
- What is the main motivation for the paper to study generalisation?
- What type of generalisation does the paper consider?
- What is the locus of the shift that the study focuses on?

---

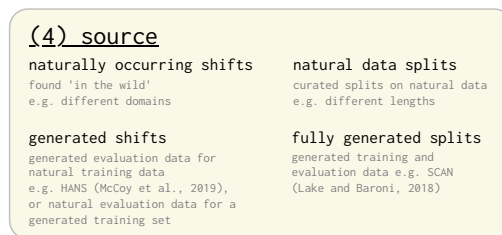
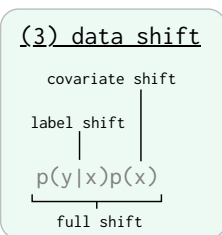
from the same distribution, for example to test whether a finetuning procedure results in catastrophic forgetting.

<sup>12</sup><https://genbench.github.io/survey>

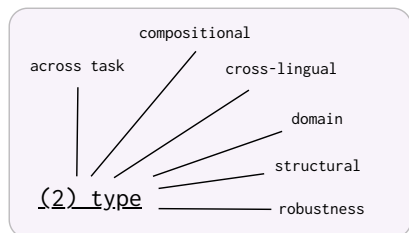
Generalisation studies have various motivations (1)...



They involve data shifts (3), where the data can come from natural or synthetic sources (4).



...and can be categorised into types (2).



These data shifts can occur in different stages of the modelling pipeline (5).

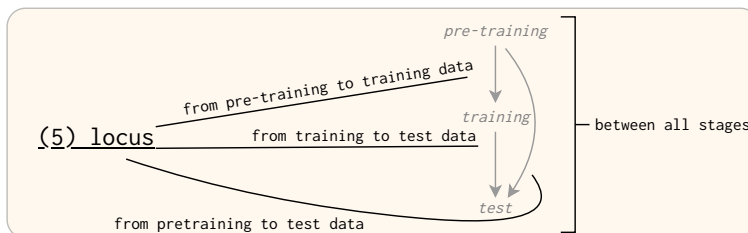


Figure 5: A graphical representation of the NLP generalisation taxonomy we present in this paper. The taxonomy consists of five different (nominal) axes, that describe the high-level *motivation* of the work (§ 2); the *type* of generalisation the test is addressing (§ 3); what kind of *data shift* occurs between training and testing (§ 4), what the *source* is of the data shift considered in the test (§ 5) and what the *locus* of the data shift is (§ 6) **DH: Potentially update with a flow-diagram with annotation instructions?**

- What type of shift is there between the data distributions of interest?
- How was the considered shift produced?

We also annotated which task(s) the studies considered, marking papers that considered multiple tasks at the same time *multitask*, or by the overarching category that the tasks belong to (e.g. *NLU*). If a paper contained multiple studies, with different values (e.g. a paper considers both across-domain and compositional generalisation), we registered those separately, resulting in a total of 694 entries. Every entry was first annotated by a first annotator, and then double checked by a second annotator. For convenience of the reader, we repeat the reference figure representing the axes of our taxonomy below.

## 7.2 Results

In this section, we report the results of our review.

### 7.2.1 Overall trends on different axes

First, we discuss the overall occurrences of values on all axes, without taking into account interactions between them. We plot the (relative) occurrences of all values in Figure 6. **DH: Would like to add something on how this recently changed (potentially), with some over time plots)**

**Motivations** As we can see in Figure 6, by far the most common motivation to test generalisation is a practical motivation. The intrinsic and cognitive motivation follow, whereas the studies in our review that consider generalisation from a fairness perspective make up only 8% of the total. We hypothesise that one of the reasons that this percentage is so low stems from the fact that our keywords search in the anthology was not optimal for detecting fairness studies, and we welcome researchers to suggest other

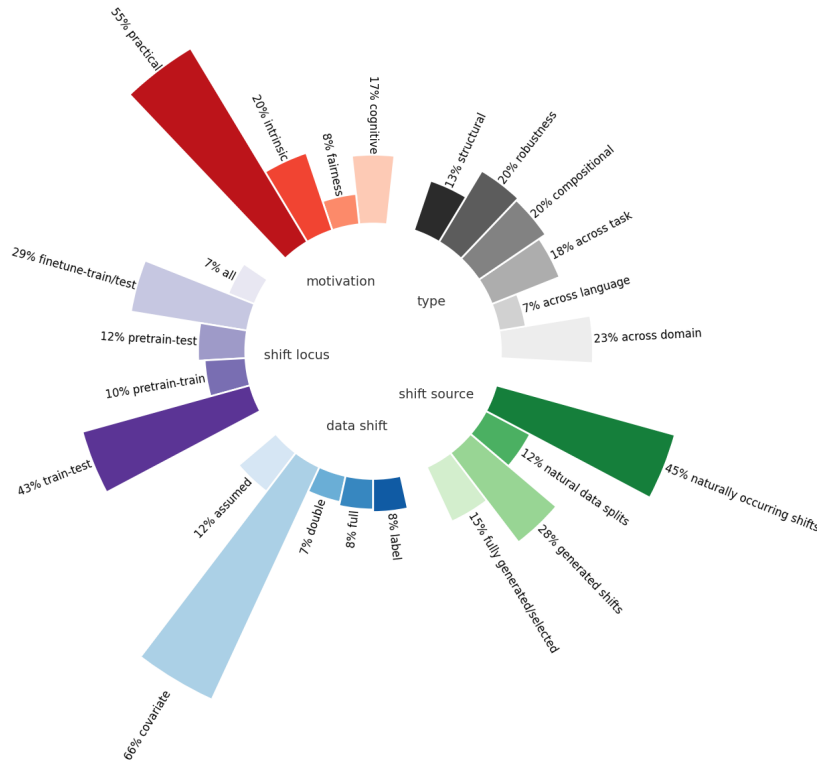


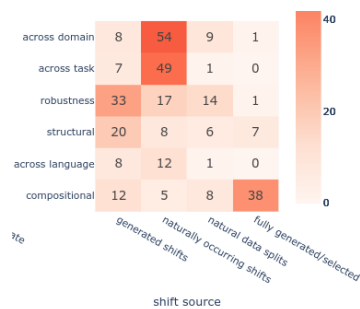
Figure 6: Relative occurrences of the different values on the axes of our taxonomy.

generalisation studies with a fairness motivation for review; we will include them in a new version of this paper.

**Shift type** As the motivations, also the shift types are very unevenly distributed over their potential values: the vast majority of generalisation research considers covariate shifts. Given the fact that covariate shifts can occur between any two axes, and label and full shift typically only occur between pretraining and finetuning, this is – to some extent – to be expected. More unexpected, perhaps, is the relatively high amount of *assumed* shifts, that correspond with studies that claim to test generalisation, but do not actually explicitly consider how the test data they consider relates to the training data they have used. A more promising number, instead, is that several studies consider *double shifts*, implicating that they consider generalisation in the entire modelling pipeline, rather than only in one level.

**Shift locus** At the locus access, we see the double shift studies appear again, with as label *all*: they consider shifts in multiple levels in the modelling pipeline. On this axis, however, we also see that there are many more studies that consider shifts only between pretrain en train, without considering if models that generalise well from pretrain en train also generalise well in the finetuning stage. There is, however, a relatively large number of studies that consider generalisation in the finetuning stage. The vast majority, still, considers generalisation in models that are not pretrained: directly from train to test. [DH: Something about pretrain-test?](#)

**Shift source** Almost half of the generalisation studies that we have reviewed consider naturally occurring shifts: natural corpora, that are not deliberately split along a particular dimension. As we will see later, this type of data source is most prevalent in across task and and domain generalisation studies,



(a) Data source and generalisation type



(b) TBD



(c) TBD

for which such naturally different corpora are widely available. The next most occurring category are generated shifts, where one of the datasets involved is generated, with a specific generalisatino property in mind. Natural data splits – corpora split along ‘unnatural’ dimensions – and fully generated datasets are less common, but each make up more than 10% of the total number of tests as well.

**Generalisation type** Lastly, in terms of generalisation type, studies are distributed relatively evenly over different types, with the exception of across-language generalisation, which makes up a much smaller part of the studies than the other types of generalisation. [DH: Why? Is this just because we haven’t annotated those studies yet?](#)

### 7.2.2 Interactions between axes

Another interesting thing to consider are the interactions between different axes. Are there any combinations of axes that occur together very often, or combinations that are instead rare? Here, we discuss the most important trends.

**What type of data is used when?** [DH: Include information about which types of data are used when, and where there are potential gaps \(e.g. compositional with natural data\)](#)

### 7.2.3 Trends across different subfields in NLP

Report overall statistics: which are the most frequent forms of generalisation research, for which tasks is generalisation research well researched? I imagine a line/bar plot here with the numbers per axis.

### 7.2.4 Trends over time

## 7.3 Conclusion

I imagine that we might want to have a conclusion already here, but maybe it is not needed and we can put everything together in the last section?

## 8 Discussion

[DH: In this section we will recap and summarise our work, and also make recommendations for future work. This will also include a description of our the website, and a commitment to add new tests that will be sent to us.](#)

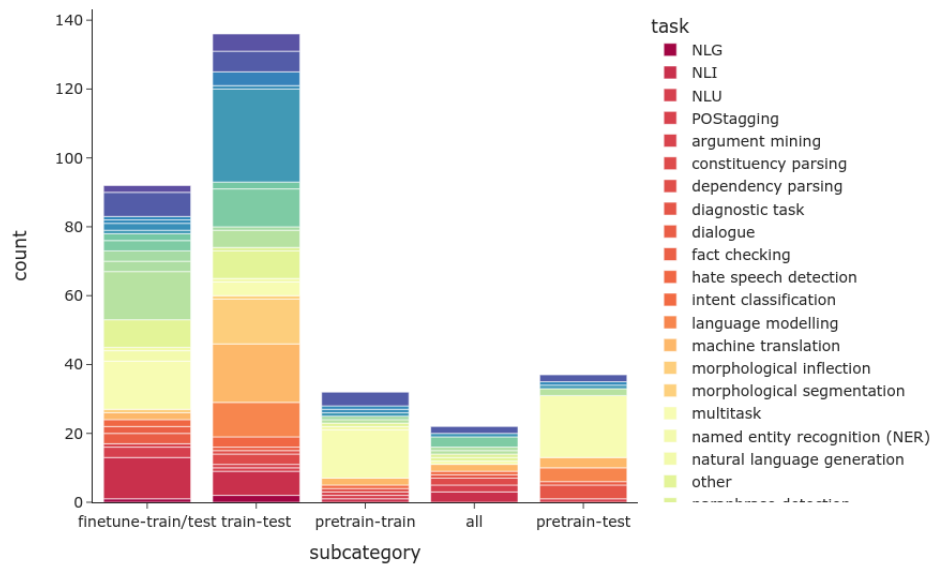
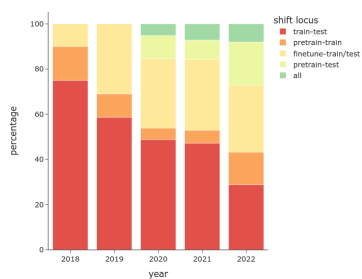


Figure 8: Trends across different subfields / tasks in NLP (plot will be updated)



(a) Shift locus



(b) TBD



(c) TBD

Figure 9: Trends over time



## References

- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabi Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.
- Prabal Agarwal, Jannik Strötgen, Luciano del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-aware named entity disambiguation for diachronic corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Melbourne, Australia. Association for Computational Linguistics.
- Roe Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.
- Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2022. Naturalistic causal probing for morpho-syntax. *arXiv preprint arXiv:2205.07043*.
- Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.
- Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. 2021. Efficient large scale language modeling with mixtures of experts.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference*

- of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.
- Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.
- Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *CoRR*, abs/2106.08694.
- Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487, 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 *FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.
- Jean Berko. 1958. The child’s learning of English morphology. *Word*, 14(2-3):150–177.
- Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.
- Prajwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.
- Su Lin Blodgett, Johnny Wei, and Brendan O’Connor. 2017. A dataset and classifier for recognizing social media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névél, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW ’19*, page 491–500, New York, NY, USA. Association for Computing Machinery.
- Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. 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.
- Samuel R Bowman, Christopher D Manning, and Christopher Potts. 2015b. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches*, pages 37–42.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 274–282, Online. Association for Computational Linguistics.
- Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can transformers jump around right in natural language? assessing performance transfer from SCAN. In *Proceedings of the Fourth Black-boxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob

- Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.
- Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.
- Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Matthew Ciolino, David Noever, and Josh Kalin. 2020. The Go transformer: Natural language modeling for game play. *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pages 23–26.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Santa Cruz, California, USA. Association for Computational Linguistics.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Maria Corkery, Yevgen Matushevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah

- Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.
- Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.
- Marco Damonte and Emilio Monti. 2021. One semantic parser to parse them all: Sequence to sequence multi-task learning on semantic parsing datasets. In *Proceedings of \*SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 173–184, Online. Association for Computational Linguistics.
- Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.
- Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to German plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108, Online. Association for Computational Linguistics.
- Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.
- Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.
- Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.
- 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.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, page 67–73, New York, NY, USA. Association for Computing Machinery.
- Yann Dubois, Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. 2020. Location Attention for Extrapolation to Longer Sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 403–413, Online. Association for Computational Linguistics.

- Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. 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.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model’s ‘factual’ predictions.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021a. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021b. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.
- Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.
- Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.
- Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.
- Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation.
- Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

- Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2544–2547, Reykjavik, Iceland. European Language Resources Association (ELRA).
- 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.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.
- Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- 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.
- Xiaochuang Han and Yulia Tsvetkov. 2022. Orca: Interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

- Yu He, Jianxin Li, Yangqiu Song, Mutian He, and Hao Peng. 2018. Time-evolving text classification with deep neural networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2241–2247. International Joint Conferences on Artificial Intelligence Organization.
- Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedziec, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.
- Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.
- Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.
- Dieuwke Hupkes, Sara , and Willem Zuidema. 2018. Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.
- Dieuwke Hupkes. 2020. Hierarchy and interpretability in neural models of language processing.
- Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.
- Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.
- Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2021. BC-Z: Zero-shot task generalization with robotic imitation learning. *ArXiv*, abs/2202.02005.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.



- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. 2015. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270, Lisbon, Portugal. Association for Computational Linguistics.
- Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.
- Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.
- Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.
- Wei-Tsung Kao and Hung-yi Lee. 2021. Is BERT a cross-disciplinary knowledge learner? a surprising finding of pre-trained models’ transferability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2195–2208, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.
- Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.
- Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are prompt-based models clueless? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2333–2352, Dublin, Ireland. Association for Computational Linguistics.
- Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.
- Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2022. Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. *CoRR*, abs/2204.05049.
- 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.
- Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.
- Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2021. A survey of generalisation in deep reinforcement learning. *CoRR*, abs/2111.09794.
- Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.
- Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.
- Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.
- Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4487–4499.
- Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.
- Yair Lakretz, Theo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2021a. Causal transformers perform below chance on recursive nested constructions, unlike humans. *CoRR*, abs/2110.07240.
- Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021b. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition’s founding editor.
- Ian Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2007. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):150–161.

- Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.
- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.
- Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.
- Jane S.Y. Li and Colin Wilson. 2021. Leveraging paradigmatic information in inflection acceptability prediction: The JHU-SFU submission to SIGMORPHON shared task 0.2. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 289–294, Online. Association for Computational Linguistics.
- Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021b. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780, Online. Association for Computational Linguistics.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.
- Jindřich Libovický, Helmut Schmid, and Alexander Fraser. 2022. Why don’t people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland. Association for Computational Linguistics.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models.
- Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

- Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.
- Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.
- Linqing Liu, Patrick S. H. Lewis, Sebastian Riedel, and Pontus Stenetorp. 2021. Challenges in generalization in open domain question answering. *CoRR*, abs/2109.01156.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Kevin Lu, Aditya Grover, P. Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *ArXiv*, abs/2103.05247.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.
- Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.
- David M. Magerman. 1995. Statistical decision-tree models for parsing. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, Massachusetts, USA. Association for Computational Linguistics.
- Andrey Malinin, Neil Band, Yarin Gal, Mark J. F. Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.
- Gary F Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. 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.
- Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there’s no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.
- Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.
- Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.
- Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

- Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. 2021. Cross-lingual genqa: Open-domain question answering with answer sentence generation.
- Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.
- Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.
- Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

- Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud’hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021a. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.
- Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud’hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021b. Sigmorphon 2021 shared task on morphological reinflection part 2: Are we there yet? A shared task on cognitively plausible morphological inflection.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

- Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.
- Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, Online. Association for Computational Linguistics.
- Vikas Raunak, Vaibhav Kumar, Florian Metze, and Jaimie Callan. 2019. On compositionality in neural machine translation. In *NeurIPS 2019 Context and Compositionality in Biological and Artificial Neural Systems Workshop*.
- Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *CoRR*, abs/2202.07206.
- Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.
- Roni Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. 2017. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32.
- Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, Brussels, Belgium. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Federico Sangati and Willem Zuidema. 2011. Accurate parsing with compact tree-substitution grammars: Double-DOP. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 84–95, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin



- Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.
- Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.
- Jürgen Schmidhuber. 1990. Towards compositional learning in dynamic networks.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.
- Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.
- Zheyang Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.
- Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.
- Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*.
- Amos Storkey. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28.

- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. LAMOL: LAnguage MOdeling for Lifelong Language Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Gábor Szolnok, Botond Barta, Dorina Lakatos, and Judit Ács. 2021. Bme submission for sigmorphon 2021 shared task 0. a three step training approach with data augmentation for morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 268–273, Online. Association for Computational Linguistics.
- Zeera Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.
- Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.
- Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.
- Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572, Hong Kong, China. Association for Computational Linguistics.
- Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. 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.
- Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. 2016. Diagnostic classifiers: Revealing how neural networks process hierarchical structure. In *Proceedings of the NIPS2016 Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.
- Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina

- Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- 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.
- Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.
- Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks.
- Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT’s knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2021. Language modelling as a multi-task problem. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2049–2060, Online. Association for Computational Linguistics.

- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.
- Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.
- 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.
- Colin Wilson and Jane S.Y. Li. 2021. Were we there already? applying minimal generalization to the sigmorphon-unimorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 283–291, Online. Association for Computational Linguistics.
- Francis CK Wong and William SY Wang. 2007. Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks. In *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems*, pages 139–144. IEEE.
- Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.
- Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models.
- Weijia Xu, Batoool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.
- Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural NLI models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 920–934, Online. Association for Computational Linguistics.
- Kaiyu Yang and Jia Deng. 2020. Strongly incremental constituency parsing with graph neural networks. *Advances in Neural Information Processing Systems*, 33:21687–21698.
- Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods*

in *Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492, Online. Association for Computational Linguistics.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2018. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 232–243, Brussels, Belgium. Association for Computational Linguistics.