# On Degrees of Freedom in Defining and Testing Natural Language Understanding

**Saku Sugawara**
National Institute of Informatics
saku@nii.ac.jp

**Shun Tsugita**
Nagoya University
tsugita.shun.u7@f.mail.nagoya-u.ac.jp

## Abstract

Natural language understanding (NLU) studies often exaggerate or underestimate the capabilities of systems, thereby limiting the reproducibility of their findings. These erroneous evaluations can be attributed to the difficulty of defining and testing NLU adequately. In this position paper, we reconsider this challenge by identifying two types of researcher degrees of freedom. We revisit Turing's original interpretation of the Turing test and indicate that an NLU test does not provide an operational definition; it merely provides inductive evidence that the test subject understands the language sufficiently well to meet stakeholder objectives. In other words, stakeholders are free to arbitrarily define NLU through their objectives. To use the test results as inductive evidence, stakeholders must carefully assess if the interpretation of test scores is valid or not. However, designing and using NLU tests involve other degrees of freedom, such as specifying target skills and defining evaluation metrics. As a result, achieving consensus among stakeholders becomes difficult. To resolve this issue, we propose a validity argument, which is a framework comprising a series of validation criteria across test components. By demonstrating that current practices in NLU studies can be associated with those criteria and organizing them into a comprehensive checklist, we prove that the validity argument can serve as a coherent guideline for designing credible test sets and facilitating scientific communication.

## 1 Introduction

Large-scale pretrained language models, also known as foundation models (Bommasani et al., 2021), have advanced significantly, leading to systems that are performing increasingly well at various natural language understanding (NLU) tasks and offering real-world applications (Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022). However, these systems are often erroneously claimed to have human-level understanding (Bender and Koller, 2020). In other cases, their failures in specific situations are presented as systemic inadequacy, while ignoring their excellent performances in certain tasks (Jia and Liang, 2017; Bowman, 2022).

Such exaggerations may stem from unjustified assumptions about the capability of NLU systems. Accordingly, researchers have attempted to improve the benchmarking of NLU through discussions to define language understanding (Bender and Koller, 2020; Bisk et al., 2020; Merrill et al., 2021) and practices, such as using auxiliary tasks for sanity check (Ribeiro et al., 2020), decision boundary evaluation (Gardner et al., 2020), and dataset sourcing (Bowman and Dahl, 2021; Kiela et al., 2021). Such efforts have motivated researchers to reconsider and revise the concept and scientific study of NLU (Lakatos, 1976). Nevertheless, we need a more comprehensive guideline for better scientific communication.

In this study, we rethink this challenge in terms of researcher degrees of freedom, aiming to reframe the definition and evaluation of NLU and provide a pathway to better benchmarking. We begin by revisiting a recent discussion by Bender and Koller (2020), who define language understanding as the link between linguistic form (i.e., symbolic information) and the communicative intent of the speaker. They support their argument by presenting a thought experiment called the Octopus test, which is designed to resemble the Turing test. In the Octopus test, an intelligent deep-sea octopus that can only use symbolic information (an analogy of language models trained only on textual corpora) tries to mimic a person in conversations. The objective of the test is to ascertain whether the octopus can deceive the other person even when that person is defending themselves against an angry bear. Bender and Koller (2020) argue that the octopus cannot achieve language understanding because it



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 13625–13649
July 9-14, 2023 

cannot pass the test in that situation (as a proof of existence). However, we find that the Octopus test may not be a valid thought experiment owing to initial concerns about the Turing test itself (e.g., even humans may sometimes fail the test). Although we agree that symbol grounding plays an important role, several other features contribute to language understanding, which takes various forms (Sahlgren and Carlsson, 2021).

Here, we see that the degrees of freedom in defining NLU are characterized by the *response-dependent interpretation* of the Turing test (Proudfoot, 2020). Thus, interpreting the subject's behavior depends on how it appears to a counterpart interrogator in the test, i.e., we cannot use success in the test as a definite criterion of language understanding, which makes it challenging to construct an NLU test with an operational definition. Instead, using a test merely enables us to obtain inductive evidence to achieve consensus among stakeholders. We elaborate on this observation by revisiting the Turing and Octopus tests in Sections 2 and 3, respectively. Then, we restate the testing of NLU in Section 4 considering that its definition is inevitably arbitrary to observers. In short, a test has to assess distinguishable behaviors in a specific domain rather than general NLU to make it easier for stakeholders to arrive at a consensus about the interpretation of the test results.

Although we can decide what capability we assess in a test, other degrees of freedom are present while designing the test and interpreting its results. This problem can be understood in terms of psychological studies, in which choices arbitrarily taken by researchers heighten the chances of false positive results (Simmons et al., 2011; Wicherts et al., 2016). Several previously proposed NLU practices facilitate detailed evaluation to lessen unjustified claims about NLU systems. However, they are not sufficiently well-organized to be deployed on a universal framework. Therefore, we often choose convenient practices depending on specific situations. Consequently, the degrees of freedom that researchers have in evaluating NLU increase, thereby sacrificing the reproducibility of research (Munafò et al., 2017). In Section 5, we introduce the validity argument, which is a framework used in psychological and educational tests (Kane, 2006; Chapelle et al., 2008; Cook et al., 2015), to set out a guideline for designing, conducting, and using NLU tests. We demonstrate that current practices are associated with inference steps in the validity argument and show that it may serve as a comprehensive, coherent guideline for developing or identifying actionable and beneficial practices to construct a better NLU test and use it properly.

Our major contributions are two-fold:

- Using the response-dependent interpretation of the Turing Test, we rethink the NLU evaluation. In our revised formulation, a test does not provide a concrete definition of NLU but presents inductive evidence, using which stakeholders can concur with the interpretation of the target behavior.
- As a tool for designing and using NLU tests, we introduce the validity argument with a checklist of 16 questions, which helps stakeholders to collect and interpret validity evidence coherently, thereby encouraging more reproducible research.

## 2 Turing Test Revisited

### 2.1 Imitation Game

Turing (1950) proposes a game known as the Imitation Game, which is a conversational test to examine a machine's intelligent behavior:

> A human questioner speaks in natural language to one machine with another human for a certain period. These participants are isolated and can communicate only in text through the display. The topics of conversation and length of questions are unrestricted. The human and machine respond to the interrogator's questions in a way that makes them appear human. The interrogator wins if he identifies the machine as a machine and the human as a human.

According to the Turing test, if a machine wins the Imitation Game more or less reliably, it passes the test, indicating that the machine can be considered intelligent.

Machines may fail the Turing test owing to factors irrelevant to intelligence. For example, a human participant may perform poorly in pretending to be a machine. Similarly, machines may exhibit a process that can be described as thinking but differs significantly from the process performed by humans. To circumvent these problems, the Turing test focuses solely on a sufficient condition for intelligence (Turing, 1950, Copeland 2004, p. 442).

### 2.2 Response-Dependent Interpretation

To evaluate the Turing test, we must clarify the interpretation of intelligence that argues that thinking is performed intelligently if the Imitation Game is played well. However, Turing's interpretation of intelligence is controversial. In the following section, we discuss two interpretations: the standard behaviorist interpretation and the more recent response-dependent interpretation.

According to the standard interpretation, the Turing test is a behaviorist test of intelligent thinking. In general, if a machine behaves as though it is intelligent, it is intelligent (Block, 1981). The Turing test is reminiscent of behaviorism, as it requires evidence for intelligent thinking to be a publicly observable behavior. However, at least two reasons exist to contradict the standard interpretation. First, Turing does not suggest that the mental vocabulary of "intelligence" or "thinking intelligently" is definable based on observable or behavioral terms (Davidson, 1990). Second, and more importantly, the game tests the interrogator's response and not the machine's behavior. In other words, the Turing test does not examine whether a machine can perform a specific cognitive task but how effectively the interrogator is fooled (Proudfoot, 2020).

Proudfoot (2020) provides an alternative interpretation of Turing's view of intelligence, which is considered more promising than the standard behaviorist interpretation. Proudfoot's exegesis begins with the observation that the first version of the Imitation Game appeared in the final section, "Intelligence as an emotional concept" of Turing (1948). In this text, Turing states: *The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration* (Turing, 1948, Copeland 2004, p. 431). Using the phrase "intelligence as an emotional concept," Proudfoot asserts that intelligence is a response-dependent property using current philosophical terminology. Response-dependent properties are those that depend on human responses under certain specified conditions. Secondary qualities, such as color, and values, such as beauty, are examples of response-dependent properties.

A simple response-dependence theory of color may be stated as follows:

> $x$ is red if and only if, in normal conditions, $x$ looks red to normal subjects.

Applying this formulation to intelligence leads to a simple response-dependence theory of intelligence:

> $x$ is intelligent if and only if, in normal conditions, $x$ appears intelligent to normal subjects.

Identifying the "normal" conditions is a typical problem faced by response-dependent theories. However, the setting of the Imitation Game appears to naturally reflect the normal conditions associated with intelligence. Selecting a "normal" interrogator is also challenging from the perspective of response-dependent theories. Therefore, Turing chose average citizens to play the role of interrogator. Thus, Proudfoot suggests that a response-dependent approach that is suited to the Turing test appears as follows:

**Formulation 2.1.** $x$ is intelligent if, in an unrestricted computer-imitates-human game, $x$ appears intelligent to an average interrogator.

We assume that Formulation 2.1 captures the view of intelligence that underlies the Turing test.

### 2.3 Crucial Problems with the Turing Test

We are inclined to consider intelligence as a response-dependent property, as suggested by Turing and Proudfoot. However, whether an unrestricted computer-imitates-human game is an appropriate means for testing intelligence remains debatable. Hayes and Ford (1995) point out several problems with the Turing test.

First, the test design is flawed because it is indeterminate what is being tested. Even if a machine could reliably pass the test, we would not be able to ascertain whether the machine was truly intelligent, or the interrogator was not sufficiently clever to ask informative questions. In short, the object of testing is unclear (Issue 1).

Second, the Turing test indirectly identifies intelligence by distinguishing the participant as a human or machine. This design forces machines to hide their inhuman abilities to impersonate humans. Thus, the test focuses on identifying intelligent behavior that successfully deceives the interrogator. Examples of this type of behavior were observed in the Loebner competition, where the winner sometimes deliberately mistyped a word and subsequently backspaced it to correct it at a human typing speed purely to deceive its interrogator. Therefore, even if a machine could reliably pass the test, it might not be intelligent (Issue 2).

Finally, even humans cannot pass the Turing test under certain conditions. Heuristics that are used to distinguish the behavior of machines from that of humans sometimes misguide interrogators. For instance, judges in the Loebner competition identified a human as a machine because they produced extended, well-written paragraphs of informative text, which tends to be associated with inhuman skills in certain parts of our culture (Issue 3).

In summary, we agree with Hayes and Ford (1995), who argue that passing the Turing test should not be the goal of AI research. Moreover, passing the test is not a necessary condition for using it as a real-world technology for humans or psychologically investigating human intelligence.

## 3 Rethinking the Octopus Test

### 3.1 Octopus Test

Similar to the difficulty of defining intelligence in the Turing Test, the definition of language understanding has also been difficult in the fields of philosophy and linguistics. Bender and Koller (2020) indicate that current hypes relating to natural language processing (NLP) systems are partly because of this confusing and challenging concept of language understanding. They define meaning $M$ as $M \subseteq E \times I$, where $E$ is a set of possible forms and $I$ is a set of possible communicative intents, pragmatically constructed in addition to the conventional meaning.

According to this definition, understanding the forms is inevitably accompanied by associating them with their communicative intents; therefore, Bender and Koller (2020) propose that systems that deal with the forms alone do not understand meaning by definition. They present a thought experiment known as the Octopus test to explain this argument:

> Suppose speakers A and B drift ashore on two separate uninhabited islands. There is a communication device on each island connected by a submarine cable that enables A and B to communicate. At the bottom of the sea, there is an octopus. Although this octopus does not understand their language, by intercepting the cable communication, it finds statistical patterns from their conversation and learns to predict how B would answer A accurately after a certain period. At some point, the octopus cuts the cable and tries to respond to A while pretending to be B. Will the octopus be able to continue to respond to A without raising suspicions?

In this test, Bender and Koller (2020) use the deep-sea octopus as an analogy for pretrained language models that are trained only on textual corpora. The octopus does not have access to sensory data associated with the speaker's communicative intents, which are essential in the authors' definition of meaning (i.e., the link between forms and communicative intents).

Bender and Koller (2020) argue that, in certain situations, the octopus might be incapable of responding to A without arousing suspicion. For example, if A asks how one can build a coconut catapult or what to do if a bear appears, the octopus can offer a convincing answer only if it accurately understands A's situation. However, because the octopus does not have the means to deal with novel information and unforeseen events beyond text, it cannot provide a sufficient answer to questions beyond the scope of what it has learned. Hence, A possibly determines that B is not human.[1]

However, should we use the Octopus test as a test of language understanding in a similar capacity to the original Turing test? Although the Octopus test does not place specific conditions on the speakers A and B, we attempt to reframe the definition of language understanding implied by Bender and Koller (2020), according to Formulation 2.1:

**Formulation 3.1.** $x$ understands language if, in an unrestricted imitation game, $x$ appears to understand language to an average interrogator.

Nevertheless, even if the Octopus test conforms to this formulation, it has severe drawbacks similar to those of the original Turing test.

### 3.2 Issues in Octopus Test

The Octopus test checks whether it understands the language, but Issues 1 to 3 of the Turing test in Section 2.3 also apply to the Octopus test. First, the ability of this test form to evaluate intelligent behavior is questionable, as the inability to converse does not mean that the subject is not intelligent or that the subject cannot understand language. This argument aligns with the singleton fallacy (Sahlgren

[1] See also Appendix A for our brief discussion on symbol grounding.

and Carlsson, 2021): the ability to understand language takes various forms. Pretrained language models may be able to exhibit behavior that can be regarded as language understanding by average interrogators, and no clear evidence for denying such an interpretation is available (Issue 1).

Second, both tests in our formulation define the ability to deceive the average interrogator as a sufficient condition for being intelligent or understanding language; however, no necessary condition has been formulated. Hence, by simply observing that the subject fails in the imitation of humans, we cannot conclude that it does not exhibit NLU because a subject may understand the language even if it does not pass the Octopus test (Issue 2).

Finally, we cannot prove that a situation exists where the octopus fails to impersonate human B. As Sahlgren and Carlsson (2021) argue, in the situation with a bear, the pretrained language model can possibly learn web traffic and generate a meaningful answer to the question of how to use a stick to protect oneself from the bear. Similarly, the octopus may be able to meaningfully communicate with A in any situation. Even if we accept the possibility of the octopus's failure, we cannot use this test owing to Issue 3 in Section 2.3, that is, even humans may fail the test.

## 4 Reframing the Response-Dependent Interpretation of NLU Tests

In this section, we investigate how NLU should be practically tested under the response-dependent interpretation of the Turing test. Considering the three issues outlined in Section 3.2, our goal is to formulate a necessary and sufficient condition of language understanding that does not contradict the actions performed in an NLU study.

First, to achieve our objective, we summarize the corresponding requirements of the formulation:

1. Specify what type of language understanding behavior is to be tested. This specification may include target tasks, skills, domains, and data format (for Issue 1).

2. Impose behavioral tests as a means to evaluate NLU, which provides evaluation metrics that are objective to the observers (for Issue 2).

3. Consider the response-dependent interpretation of the behavioral test to avoid directly defining specific linguistic behavior as language understanding (for Issue 3).

Based on these requirements, we reformulate and elucidate the response-dependent interpretation of language understanding:

**Formulation 4.1.** $x$ understands language under the condition $c$ if and only if the subjective probability of an average observer for the hypothesis, $x$ understands language under $c$, is higher than a threshold, where the hypothesis is supported by evidence obtained by the performance of $x$ on a test under $c$.

According to the three requirements, we elaborate on this formulation as follows.

**1. Specifying Target Linguistic Behavior** Raji et al. (2021) suggest that tests evaluating general language understanding might be impractical owing to their propensity to make false claims. Humans generally identify language understanding in the various behaviors of others, and the success of that behavior is determined depending on stakeholder objectives. Owing to this broad scope, designing a practical, versatile test may be unrealistic. Therefore, we argue that NLU should be decomposed into distinguishable capabilities by specifying a target condition $c$, including skills, tasks, data sources, input and output format, and potential applications.

**2. Using Behavioral Tests** To address the issues of the Turing test, Levesque (2014) argues that a behavioral test should be used to test the commonsense reasoning of machines. Similarly, we can use a behavioral test to evaluate NLU because it provides objective measures (i.e., *evidence obtained by the performance of $x$ on a test*). However, we do not deny the possibility of using a test that inspects a machine's internal properties to effectively evaluate NLU (e.g., using interpretation methods for deep learning models), although such a test would have to define objective criteria for target internal properties to ensure stakeholder consensus.

**3. Response-Dependent Interpretation** This interpretation does not provide an explicit definition of the expected target behavior but requires an average observer to observe the subject's output in a test. The statement *the subjective probability of an average observer [...] is higher than a threshold* indicates that stakeholders or their representative experts agree on their interpretation of test results (i.e., the observed behavior is successful). In other words, (the approximation of) the average speaker

obtains certain information regarding the subject's behavior $x$ and calibrates the subjective probability for the hypothesis. In this formulation, a linguistic test for language understanding does not provide an operational definition; instead, it provides inductive evidence that the speaker can use for their calibration (Moor, 1976, 2001).

Our formulation may be seen as a meta-definition of NLU in the sense that the stakeholders are required to determine their own definition of language understanding in line with their objectives. In other words, it provides an epistemological view of language understanding and avoids making a commitment to what language understanding is.[2] This flexibility in interpretation is possible because the stakeholders can define their own characterizations of language understanding and instantiate intended behavior into target tasks based on their objectives.[3] However, this formulation permits several degrees of freedom in how stakeholders define what language understanding is, which can make our communication about NLU obscure. Therefore, the designers and users of NLU tests have to precisely specify what behavior is evaluated in their tests, including the conditions $c$, and accordingly interpret the test results. This effort improves the validity of resulting claims and thus contributes to avoiding over- and under-claiming.

## 5 Validity Argument for Testing NLU

In Formulation 4.1, the stakeholders employ certain measures to evaluate the behavior of a test subject and interpret those measures to arrive at a decision. To support the interpretation of the measures (i.e., improve the stakeholders' confidence in the subject's target linguistic behavior), the test has to be well-designed to provide sufficient evidence from various perspectives.

Despite significant progress, NLU benchmarks often lack the ability to provide substantial evidence (Bowman and Dahl, 2021; Raji et al., 2021; Dehghani et al., 2021). To date, useful practices have been proposed to better connect observed scores and intended NLU behavior (e.g., McCoy et al., 2019; Gardner et al., 2020). However, without guidelines on design and using a test, these practices cannot be organized coherently. As a result, researchers can arbitrarily select practices most suited to their purpose while, intentionally or not, ignoring others. This freedom of choice allows interpretations of test results that are not justified by reliable evidence. Such unjustified interpretations are reminiscent of a problem in psychology known as researcher degrees of freedom (Simmons et al., 2011; Wicherts et al., 2016). We need to tackle this challenge by developing a unified, comprehensive framework to overview necessary practices that validate the interpretation of test results.

In psychological measurements, Kane (2006, 2013) proposes the validity argument, which is a theoretical framework that guides the collection of evidence to validate the interpretation and use of test scores (Figure 1). It decomposes the design, conduct, and use of a test into the seven components where transitions between them are performed as inferences supported by a warrant and its backing that follow Toulmin's formulation of arguments (Toulmin, 2003). The six inferences constitute the process of collecting and interpreting evidence that the test designers and users follow. The validity argument has been employed in various fields, including linguistic tests (Chapelle et al., 2008) and clinical exams (Cook et al., 2015), but not in NLU. See Appendix B for its background and terminology.

In the following section, we apply the validity argument to the evaluation of NLU. We associate current important practices with the inferences of the validity argument, show that the argument serves as a useful guideline to design and use NLU tests, and provide relevant checklist questions Here, we take SQuAD (Rajpurkar et al., 2016), a question answering (QA) dataset consisting of crowdsourced questions written for Wikipedia articles to instantiate a validity argument and identify missing aspects. Appendix C elaborates on the checklist and relevant practices.

**1. Domain Definition** This inference requires performance observation in the test to reveal relevant knowledge and skills in the target domain, which contributes to providing a means to clarify the achievements and weaknesses of the test subjects at a theoretical level (Doshi-Velez and Kim,

[2]We follow the terminology of Bommasani et al. (2021, Section 2.6.3). Epistemology is related to *how we know that an agent has achieved the relevant type of language understanding*, while metaphysics concerns *what it would mean for an agent to achieve language understanding*.

[3]For example, if stakeholders adopt internalism, their test requires the subject to have consistent internal representations for intended tasks. See also Bommasani et al. (2021, Section 2.6.3) for an overview of their metaphysical characterizations of language understanding that may fit the development and deployment of foundation models.

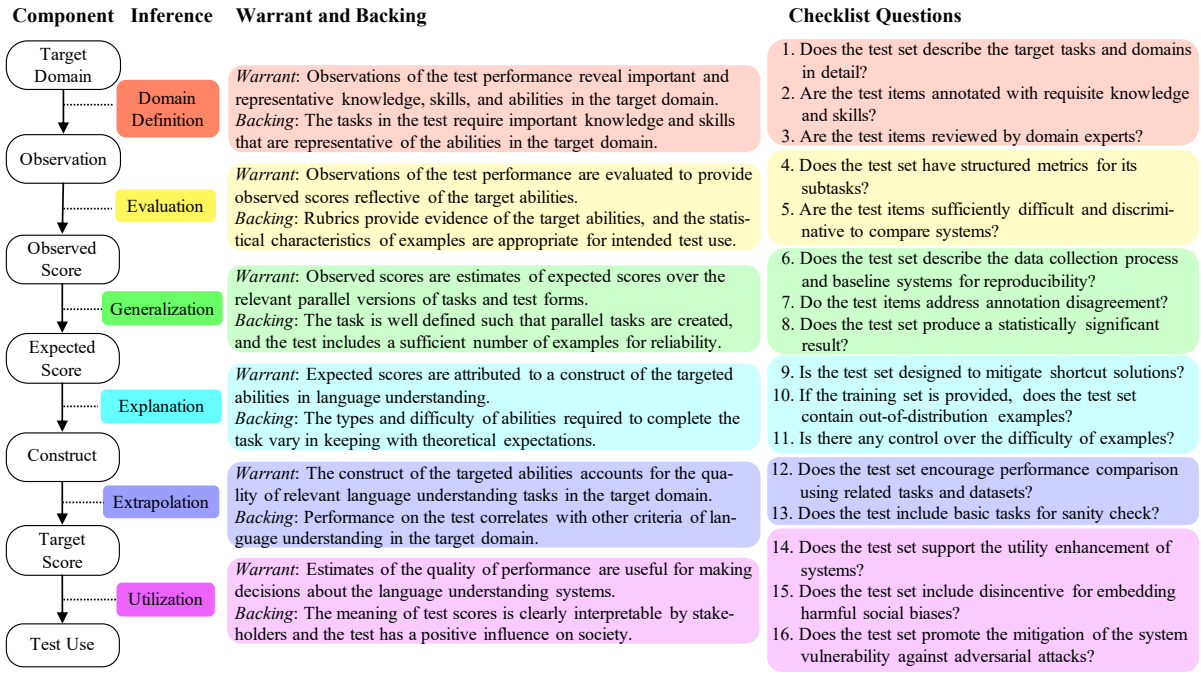


Figure 1: Overview of a validity argument for NLU, including inferences with its warrant and backing.

2017; Rogers et al., 2023). In SQuAD, the authors take 100 questions to annotate necessary reasoning types, including lexical and world knowledge, syntactic variation, and multiple-sentence reasoning. However, given that reading comprehension involves other types of reasoning, such as temporal and causal (Dunietz et al., 2020), this annotation may reveal that SQuAD can evaluate only a limited number of abilities in reading comprehension. This observation implies that, ideally, the test examples need to be reviewed by domain experts (i.e., researchers of human reading comprehension) who may be able to indicate relevant knowledge and reasoning types that are missing from the dataset.

**2. Evaluation** This inference is related to the design of scores and characteristics of test examples. SQuAD provides evaluation metrics that calculate word overlap between predicted and gold text spans. However, it only enables a one-dimensional interpretation of the system performance. Ideally, the task provides multiple metrics that correspond to target abilities or decomposed subtasks identified in the domain definition (e.g., Wolfson et al., 2020). This inference also requires the statistical characteristics of the test examples to be monitored. The test developers have to ensure that collected examples are discriminative and their differences are relevant to the target abilities. Otherwise, stakeholders cannot associate the system performance difference with the target abilities. A promising way to meet this requirement is to use the item response theory, which models the difficulty and discriminability of examples using the responses of test subjects (Rodriguez et al., 2021; Vania et al., 2021).

**3. Generalization** This inference concerns the generalizability of the observed scores, which is characterized by their reproducibility and reliability. For reproducibility, the task specifications must be well-described to ensure that others can construct similar tasks for the same target. Reporting specifications using a template (Gebru et al., 2021) would suffice to cover necessary facets such as the components, collection, and preprocessing of the dataset. For reliability, reporting statistical significance is also an important practice (Benjamin et al., 2018; Dror et al., 2018). The reliability of test scores also has to be supported by reliable annotation of gold labels. In addition, the annotation design should consider inherent disagreements among humans (Pavlick and Kwiatkowski, 2019). SQuAD reports its data collection process and makes the scripts of score calculations publicly available to replicate the baseline systems and reported results. Miller et al. (2020) similarly collect examples for different text sources, such as newspaper articles and product reviews. They find that the average performance largely drops in those text sources compared to the original SQuAD across a broad range of systems.

This finding would help stakeholders assess the generalizability of the observed test scores if the systems need to be tested in terms of general reading comprehension beyond that for Wikipedia articles. Furthermore, the study on SQuAD reported human performance using gold answers by three annotators to ensure reliability of the test scores, but its statistical variation is not calculated.

**4. Explanation** This inference determines if expected scores in the test can be associated with a construct (a conceptual tool used to facilitate understanding of human behavior in psychology) of the target abilities. For instance, systems that exhibit intended behavior despite varying input enable stakeholders to validate the association between their behavior and the target abilities. For this purpose, the test must have examples that can validate various behaviors to cover the range of abilities that the test is intended to assess. References including challenge sets (McCoy et al., 2019), stress tests (Naik et al., 2018), and contrast sets (Gardner et al., 2020) aim to investigate the decision boundary of the systems. Avoiding dataset biases and shortcut solutions is also important for testing intended abilities (Gururangan et al., 2018; Geirhos et al., 2020; Malaviya et al., 2022). For SQuAD, Jia and Liang (2017) find that injecting a manually crafted distracting sentence into the passage causes the systems to predict incorrectly. Such unintended behaviors show that examples in SQuAD may be insufficient to cover a range of inputs necessary for performing the explanation inference. In most NLU tasks, examples are assumed to be easily solvable by humans. This assumption ensures the quality of examples but lacks any control over their difficulty. However, simply collecting examples that are difficult for systems, as in adversarial data collection (Kiela et al., 2021) or dataset cartography (Liu et al., 2022), may be misleading because we need to control how the difficulty contributes to the target abilities in the test (Bowman and Dahl, 2021).

**5. Extrapolation** This inference checks if the system performance on the test successfully correlates with other criteria in related tasks and datasets. For example, when we have a successful system in SQuAD, the system should ideally be able to perform well in a similar dataset such as Natural Questions (Kwiatkowski et al., 2019). Talmor and Berant (2019) and Khashabi et al. (2020) analyze the generalization and transfer of performance across multiple QA datasets. A compilation of tasks provided in the same format or platform is also helpful for users to compare the performance of systems on different tasks (Wang et al., 2019; Srivastava et al., 2022; Liang et al., 2022). A successful system in machine reading comprehension is also expected to pass relevant basic tasks such as semantic role labeling and named entity recognition. Moreover, testing the system on such auxiliary tasks contributes to its sanity checks (Ribeiro et al., 2020). These auxiliary-task practices are optional compared to other inferences because a single test set cannot realistically provide a set of different tasks as well. Nevertheless, referring to existing datasets and aligning the task format with that of those datasets is beneficial.

**6. Utilization** This inference focuses on the utility of the test results and potential social influence of the test. The current convention in machine learning is to train and test models on static datasets. However, this approach does not ensure that the models are properly and fairly deployable in different configurations or real-world applications. For example, Ethayarajh and Jurafsky (2020) and Ma et al. (2021) advocate the reporting of model statistics, such as the size, energy efficiency, and inference latency, on leaderboards to enable more informative comparison of models in terms of utility. Regarding social influence, the test should not motivate the development of systems that have harmful social biases, such as stereotypes (Blodgett et al., 2021) and gender biases (Sun et al., 2019). In addition, depending on the potential applications, we have to monitor the vulnerability of systems towards adversarial attacks (Wallace et al., 2019b). Several methods have been proposed to improve the robustness of systems against adversarial inputs, such as training with diverse data (Tu et al., 2020) and self-debiasing framework (Utama et al., 2020). Although SQuAD does not address these problems, subsequent studies have addressed social biases in QA (Parrish et al., 2022) and the applicability of adversarial examples (Wallace et al., 2019a).

To summarize, our checklist questions help users find useful practices to collect evidence that validates the test interpretations. Although ensuring that a single test set conforms to all criteria may be impractical (e.g., there may be trade-offs between the coverage and diversity of test examples with their reliability and discriminability), knowing

what evidence is missing is helpful to assess the validity of the intended interpretations and develop necessary practices.

## 6 Related Work

The definition of language understanding has been discussed in various NLP tasks, including symbol grounding (Merrill et al., 2021) and reading comprehension (Dunietz et al., 2020). Our work is similar to Bommasani et al. (2021) in that we do not provide concrete definitions and highlight an epistemological perspective.

Messick (1995) introduce six aspects to improve the validity of interpreting results in psychological measurements. Although Sugawara et al. (2021) associate these aspects with the requirements of designing NLU datasets, actionable practices are not proposed for these aspects. Similarly, Raji et al. (2021) discuss the construct validity in benchmarking AI but do not aim to improve evaluation methods. Our validity argument provides a step-by-step guideline for test developers to follow.

The concept of researcher degrees of freedom is originally introduced in psychology as a pertinent factor in severe issues such as HARKing (hypothesizing after the results are known) and publication bias (Munafò et al., 2017). In terms of resolving this problem in NLP, preregistration is a potentially promising direction for research (van Miltenburg et al., 2021). However, it is not suited to all areas of NLU research because certain explanatory and analytic studies do not begin with a clear hypothesis. Nonetheless, clearly stating a research goal, problem definition, data collection method, system statistics, intended use, and potential risks could be the first step towards making credible claims on the capabilities of NLU systems.

## 7 Conclusions

The prevalence of exaggerated claims about the achievements of foundation models motivates us to reconsider how we define and evaluate NLU. Our formulation of NLU using the response-dependent interpretation mitigates the issues of the Turing and Octopus tests; it stipulates that observers and target conditions, including tasks and abilities, must be specified. However, current practices for creating NLU datasets are yet to be aligned, which provides researchers with the freedom to choose convenient strategies. To organize essential practices using a standard guideline, we introduce the validity argument, which guides stakeholders to collect and interpret evidence for validating that the test subject executes its intended behavior. Our proposed checklist helps researchers find relevant practices for benchmarking NLU, but we continually revise it by investigating potential refutation to promote more credible NLU studies.

## Limitations

Although our discussion should be applicable to all NLP tasks, it is mainly pertinent to intellectual linguistic tasks (e.g., natural language inference, reading comprehension, and commonsense reasoning) that may involve language understanding in some sense. The main reason for this limited applicability is that such intellectual tasks are relatively more response-dependent than basic tasks (e.g., syntactic parsing and semantic role labeling) and necessitate well-designed datasets and better evaluation methods.

Our formulation of testing language understanding in Section 4 may be theoretically incomplete and require further discussion with reference to all related fields, including philosophy, psychology, cognitive science, and artificial intelligence. In particular, our formulation only provides an epistemological viewpoint and thus does not provide a concrete definition of language understanding for avoiding confusion, which the community needs to discuss further.

In Section 5, we introduce a framework that deals with current and future practices for better NLU studies. However, the proposed static checklist should be continually revised and improved by incorporating future findings that reveal potential flaws in our methodology to construct an effective set of checklist questions. Furthermore, although our ultimate aim is to create a language-agnostic formulation and checklist that do not depend on specific languages, we have mainly focused on studies that deal with English texts.

Our checklist is designed for NLU and not for other NLP tasks, but it can be modified and extended to other tasks such as machine translation, language grounding, and natural language generation while referring to comprehensive meta-analysis and survey studies (e.g., Marie et al., 2021; Chandu et al., 2021; Clark et al., 2021).

## Acknowledgments

We would like to thank the anonymous reviewers for their insightful and constructive feedback. This work was supported by JSPS KAKENHI Grant Number 22K17954, JST PRESTO Grant Number JPMJPR20C4, and JST Moonshot R&D Program Grant Number JPMJMS2011.

## References

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.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. 2018. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Ned Block. 1981. Psychologism and behaviorism. *Philosophical Review*, 90(1):5–43.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. arXiv preprint 2108.07258.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan

Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 60–70, Melbourne, Australia. Association for Computational Linguistics.

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

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

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

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130, Dublin, Ireland. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Carol A. Chapelle, Mary K. Enright, and Joan M. Jamieson. 2008. Test score interpretation and use. In *Building a Validity Argument for the Test of English as a Foreign Language*, pages 1–25. Routledge, New York, NY.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint 1803.05457.

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.

David A. Cook, Ryan Brydges, Shiphra Ginsburg, and Rose Hatala. 2015. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical education*, 49(6):560–575.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai,

and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning. arXiv preprint 2011.03395.

Donald Davidson. 1990. Turing's test. In K. Said, editor, *Modelling the Mind*. Oxford University Press.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. arXiv preprint 2007.14435.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. arXiv preprint 2107.07002.

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.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint 1702.08608.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Gareth Evans. 1973. The causal theory of names. *Aristotelian Society Supplementary Volume*, 47(1):187–208.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

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

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

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.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

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.

Patrick Hayes and Kenneth Ford. 1995. Turing test considered harmful. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, page 972–977, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

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.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Michael T. Kane. 2006. Validation. In *Educational measurement*, 4th edition, chapter 2, pages 17–64. American Council on Education and Praeger, Westport, CT.

Michael T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. 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 6618–6633, Online. Association for Computational Linguistics.

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.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

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.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Imre Lakatos. 1976. Falsification and the methodology of scientific research programmes. In Sandra G. Harding, editor, *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*, pages 205–259. Springer Netherlands, Dordrecht.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

Hector J. Levesque. 2014. On our best behaviour. *Artificial Intelligence*, 212:27 – 35.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. arXiv preprint 2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

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.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. In *Advances in Neural Information Processing Systems*, volume 34, pages 10351–10367. Curran Associates, Inc.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6540, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. 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 7297–7306, Online. Association for Computational Linguistics.

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

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5033, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 2339–2352, Online. Association for Computational Linguistics.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. 2021. NeurIPS 2020 EfficientQA competition: Systems, analyses and lessons learned. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 86–111. PMLR.

James H. Moor. 1976. An analysis of the Turing test. *Philosophical Studies*, 30(4):249–257.

James H. Moor. 2001. The status and future of the Turing test. *Minds and Machines*, 11(1):77–93.

Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? 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 1221–1235, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. 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.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20.

Diane Proudfoot. 2020. Rethinking Turing's test and the philosophical implications. *Minds and Machines*, 30(4):487–512.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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 4486–4503, Online. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. 2020. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5715–5725, Online. Association for Computational Linguistics.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Magnus Sahlgren and Fredrik Carlsson. 2021. The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4:131.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

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

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366. PMID: 22006061.

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. arXiv preprint 2206.04615.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817, Vancouver, Canada. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. 2021. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1592–1612, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of QA4MRE main task at CLEF 2013. *Working Notes, CLEF*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The uses of argument*. Cambridge university press.

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.

Alan M. Turing. 1948. Intelligent machinery. Reprinted in B. Jack Copeland ed. 2004, *The Essential Turing*, pages 410–432.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX(236):433–460. Reprinted in B. Jack Copeland ed. 2004, *The Essential Turing*, pages 441–464.

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

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. 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 1141–1158, Online. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, Dublin, Ireland. 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, pages 3266–3280. Curran Associates, Inc.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A

survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. 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.

Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. 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 6707–6723, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.

## A Brief Note on Symbol Grounding

According to the observations in Section 3.2, the Octopus test argument cannot support the claim that a system that learns the form alone cannot understand language (i.e., it cannot pass the Octopus test). However, the test remains pertinent because it continues to stimulate the intuition that language models are not symbol-grounded and are therefore unlikely to understand language.

The crucial question is whether this intuition is true; we have two reasons to doubt it. First, we may argue that the corpora used for training current NLU systems do not comprise mere forms. A corpus comprises linguistic expressions that are produced by social interactions among humans who understand the language. Thus, the linguistic expressions in a corpus cannot be regarded as mere physical objects that lack meaning because they have already been assigned meaning. Second, the meaning defined by Bender and Koller (2020) may not agree with our understanding of the language. For example, a layperson has only vague ideas about the satisfaction conditions of words such as "bacteria" or "nicotine" (cf. Evans, 1973). This is the case for many other general terms. Nevertheless, people manage to use these words in their lives, and it is narrow-minded to conclude that they do not actually understand the meanings of these words.

## B Design and Terminology of Validity Argument

Regarding the definition of inferences, Kane (2006, 2013) originally define four inferences (scoring,

generalization, extrapolation, and implication). However, we adopt the extended definition of Chapelle et al. (2008) with six inferences to identify clear and detailed evidence that we consider adequate for testing NLU. We refer readers to Cook et al. (2015) for a concise practical introduction of the validity argument and to Chapelle et al. (2008) for an application example of human language testing.

We briefly mention the terminology. As the validity argument is initially proposed for psychological and educational measurements, several terms have different meanings from those in NLP and machine learning. For example, *inference* is performed by stakeholders (e.g., researchers, model developers, test creators, and users) involved in the test, whereas it is usually performed by a model to make a prediction in NLP. *Domain* in the validity argument includes any condition that specifies target behavior and experimental settings. In contrast, it only indicates data sources or text genres in NLP. *Generalization* in the validity argument is concerned with how the experimental results are generalized to other experimental settings, similar to the reproducibility of findings. Conversely, it mainly refers to the property of machine learning models in NLP, which is related to the explanation inference in the validity argument (i.e., can one explain whether a model shows capabilities that are sufficiently generalizable to the target construct?).

## C Checklist and Relevant Practices for Validity Argument

Table 1 summarizes the checklist for collecting evidence in the validity argument and recent relevant practices that have been proposed in the NLU study. We elaborate on the checklist questions and practices as follows:

### C.1 Domain Definition

**Does the test set clearly describe the target tasks and domains in detail?** This question requires the test description for specifying what task a test aims to evaluate in what domain, rather than aiming for general language understanding (Raji et al., 2021). The test description includes a general goal of the test, test format, intended task, data source, and potential application. To describe target knowledge and skills, test developers can develop their own taxonomy or use existing taxonomies such as for linguistic phenomena (Warstadt et al., 2020), commonsense types (LoBue and Yates, 2011; Sap et al., 2019), science questions (Boratko et al., 2018), reading comprehension (Sugawara et al., 2017; Dunietz et al., 2020), and QA in general (Rogers et al., 2023).

**Are the test items annotated using requisite knowledge and skills?** The detailed annotation of required knowledge and skills with test items helps stakeholders interpret the strengths and weaknesses of the models tested and to associate them with relevant tasks (Doshi-Velez and Kim, 2017; Schlangen, 2021). Annotation can be performed in two main ways: create diagnostic examples with constraints such as keywords and templates (Rogers et al., 2020; Warstadt et al., 2020) or annotate labels after collecting examples with no constraints (Schlegel et al., 2020, inter alia). Nonetheless, the appropriate granularity of the annotation has to be decided depending on how much detail the stakeholders require to analyze the model behavior. This consideration is essential because annotating detailed knowledge and skills unambiguously is difficult, even for experts.

**Are the test items reviewed by domain experts?** Ideally, the appropriateness of test items, given the target task, should be reviewed by experts who are familiar with the task. Parrish et al. (2021) show that involving experts during data collection improves the quality of crowdsourced data by identifying artifacts. This expert-in-the-loop design for collecting examples may be a promising direction to target intended linguistic phenomena. Nevertheless, we must ensure that the distribution of data is not unintentionally biased towards a limited variety of linguistic phenomena (Bowman and Dahl, 2021). Gardner et al. (2020) asked experts who created the source dataset of a target task to modify task examples to ensure that the difference between the original and contrastive examples produced different gold labels. Although hiring experts to craft test examples from scratch is expensive, at least reviewing (and annotating) some of the collected examples by the test developers (as experts) contributes to sourcing test examples that require their target task correctly.

### C.2 Evaluation

**Does the test set have structured metrics for its subtasks?** Because of the possibility of shortcut solutions (Geirhos et al., 2020) that circumvent intended solutions, merely observing the final output

| Inference | Checklist | Relevant Practices |
|---|---|---|
| Domain Definition | ☐ Does the test set describe the target tasks and domains in detail? | Taxonomy of knowledge and skills (LoBue and Yates, 2011; Rogers et al., 2023) |
| | ☐ Are the test items annotated with requisite knowledge and skills? | Diagnostic dataset (Warstadt et al., 2020), qualitative annotation (Schlegel et al., 2020) |
| | ☐ Are the test items reviewed by domain experts? | Data collection with experts (Parrish et al., 2021), contrast sets (Gardner et al., 2020) |
| Evaluation | ☐ Does the test set have structured metrics for its subtasks? | Sub-questions as semi-structured explanation (Wolfson et al., 2020; Geva et al., 2021) |
| | ☐ Are the test items difficult and sufficiently discriminative to compare systems? | Item response theory (Rodriguez et al., 2021; Vania et al., 2021), dataset cartography (Swayamdipta et al., 2020), crowdsourcing protocol design (Nangia et al., 2021) |
| Generalization | ☐ Does the test set describe the data collection process and baseline systems for reproducibility? | Templates of dataset specification (Bender and Friedman, 2018; Gebru et al., 2021), reproducibility checklist (Pineau et al., 2021) |
| | ☐ Do the test items address annotation disagreement? | Taxonomy of disagreement (Jiang and Marneffe, 2022), modeling annotation distribution (Chen et al., 2020; Nie et al., 2020b) |
| | ☐ Does the test set produce a statistically significant result? | Statistical test (Dror et al., 2018; Sadeqi Azer et al., 2020), statistical power (Card et al., 2020), instability analysis (Zhou et al., 2020) |
| Explanation | ☐ Is the test set designed to mitigate shortcut solutions? | Input ablation (Gururangan et al., 2018), competency problems (Gardner et al., 2021), adversarial filtering (Zellers et al., 2019) |
| | ☐ If the training set is provided, does the test set contain out-of-distribution examples? | Diagnosis of heuristics (McCoy et al., 2019), stress tests (Naik et al., 2018; Saha et al., 2020), contrast sets (Gardner et al., 2020) |
| | ☐ Is there any control over the difficulty of examples? | Simplified auxiliary questions (Sutcliffe et al., 2013), human–machine collaboration (Bartolo et al., 2022; Liu et al., 2022) |
| Extrapolation | ☐ Does the test set encourage performance comparison using related tasks and datasets? | Cross-dataset generalization analysis (Talmor and Berant, 2019), compilation of tasks (Wang et al., 2019; Srivastava et al., 2022) |
| | ☐ Does the test include basic tasks for sanity check? | Checklist of basic tests for task-relevant linguistic capabilities (Ribeiro et al., 2020) |
| Utilization | ☐ Does the test set support the utility enhancement of systems? | Reporting practical statistics (Ethayarajh and Jurafsky, 2020; Ma et al., 2021) |
| | ☐ Does the test set include disincentive for embedding harmful social biases? | Underspecified questions (Li et al., 2020), quantifying representational harms (Mehrabi et al., 2021), bias types (Blodgett et al., 2020) |
| | ☐ Does the test set promote the mitigation of the system vulnerability against adversarial attacks? | Universal adversarial triggers (Wallace et al., 2019a), data augmentation (Min et al., 2020), self-debiasing framework (Utama et al., 2020) |

Table 1: Overview of validity inferences, checklist questions, and relevant practices for NLU.

does not necessarily guarantee that the test subject performs the task precisely. Given that a generated explanation about the answering process cannot be evaluated straightforwardly (Clark et al., 2021), asking about a (semi-)structured reasoning path may be a useful approach. For example, several benchmarks require the completion of reasoning process in addition to the main QA task (Bhagavatula et al., 2020; Inoue et al., 2020; Wolfson et al., 2020; Geva et al., 2021; Saha et al., 2021).

**Are the test items sufficiently difficult and discriminative to compare systems?** Item response theory is a standard way to characterize the difficulty and discriminability of test examples while modeling the ability of test subjects (Lalor et al., 2016; Rodriguez et al., 2021; Vania et al., 2021; Byrd and Srivastava, 2022). In the benchmarking of NLU models, a test needs to enable meaningful comparisons between models including the baseline. If all test examples are exceedingly easy or difficult, or if there are many ambiguous examples, no significant difference in evaluation measures can be observed. Item response theory helps test developers analyze test examples and control the distribution of difficulty and discriminability.

### C.3 Generalization

**Does the test set describe the data collection process and baseline systems for reproducibility?** This requirement includes critical aspects to ensure the reproducibility of the study, such as data sources, how the annotators are employed, annotation procedure, annotation instructions, platform or software used for collection, quality control, experimental design, and baseline systems. It is also beneficial to identify potential biases unintentionally embedded by annotators (Geva et al., 2019), although the annotation instructions and examples need to be carefully presented to mitigate such biases (Parmar et al., 2023). To describe the data collection process, using templates of dataset specifications, such as data statements (Bender and Friedman, 2018) and datasheet (Gebru et al., 2021) (especially their data collection part), is helpful. If crowdsourcing is used in the data collection, reporting payment methods is also encouraged to guarantee ethical fairness (Kummerfeld, 2021; Shmueli et al., 2021). If a paper proposing a new dataset includes baseline machine learning systems, ensuring the reproducibility of baseline systems is also important (Pineau et al., 2021).

**Do the test items address annotation disagreement?** The dataset needs to address ambiguous test items to produce reliable results. However, ambiguity may not be noise in annotation but an inherent property of examples (Zhang et al., 2017; Pavlick and Kwiatkowski, 2019). Therefore, in addition to designing a careful procedure to take care of ambiguous and under-specified examples (e.g., Boyd-Graber and Börschinger, 2020), modeling the ambiguity itself can also be a meaningful task in NLU. For example, several studies tackle the task of modeling the distribution of human votes for the labels in the natural language inference task (Chen et al., 2020; Nie et al., 2020b; Meissner et al., 2021; Zhang and de Marneffe, 2021; Zhou et al., 2022). A taxonomy of disagreement (Jiang and Marneffe, 2022) is also a useful reference in the qualitative analysis of ambiguous cases.

**Does the test set produce a statistically significant result?** When comparing the performance between systems, choosing an appropriate statistical test is critical to prove that the observed performance difference is statistically significant (Dror et al., 2018; Sadeqi Azer et al., 2020). In testing, a sufficient number of examples is necessary to detect a true effect of the performance improvement. Card et al. (2020) suggest that statistical power should be analyzed before performing evaluation. Similarly, Zhou et al. (2020) analyze the performance instability in popular benchmarks, suggesting the reporting of decomposed variance measures and use of diverse analysis datasets.

### C.4 Explanation

**Is the test set designed to mitigate shortcut solutions?** In the current standard of machine learning, most NLU datasets are based on the training, validation, and test split. Although this study focuses on contributing to the methodological improvement of the test phase, the distribution between training and test split may affect the test results; machine learning models are generally good at exploiting superficial patterns from training examples. Identifying these patterns helps the models make accurate predictions for test examples; however, this approach does not work well for out-of-distribution (OOD) examples (D'Amour et al., 2020; Geirhos et al., 2020). For example, analysis of spurious correlations between gold labels and tokens reveals potential shortcut solutions (Gururangan et al., 2018; Gardner et al., 2021). However,

Schwartz and Stanovsky (2022) note that controlling balancing methods is difficult for spurious correlations, such as data augmentation (Sharma et al., 2018) and adversarial filtering (Zellers et al., 2019; Bras et al., 2020), because these methods may diminish meaningful signals. Therefore, they suggest alternative methods such as adding rich contexts and stopping large-scale fine-tuning. Analysis of input ablation may also be a useful practice to filter out easy examples, such as by hiding the premise in natural language inference (Gururangan et al., 2018) and removing question tokens in machine reading comprehension (Feng et al., 2018; Kaushik and Lipton, 2018; Yu et al., 2020). Malaviya et al. (2022) find that monitoring heuristic annotation strategies among crowdworkers may improve the quality of collected QA examples.

**If the training set is provided, does the test set contain OOD examples?** Similar to the previous question, to properly associate the target behavior with the intended skill, we have to ensure the generalizability and robustness of the models towards diverse examples in the target task. Given the possibility of shortcut solutions, the test set has to contain OOD examples, that is, ones collected in a different manner to those used for the training examples (Ettinger et al., 2017; Linzen, 2020). This line of research includes adversarial examples (Jia and Liang, 2017; Glockner et al., 2018), diagnosis sets for syntactic heuristics (McCoy et al., 2019), stress test evaluation (Naik et al., 2018; Saha et al., 2020), and contrast sets for probing decision boundaries of the models (Gardner et al., 2020), among others. Adversarial data collection (Bartolo et al., 2020; Nie et al., 2020a; Kiela et al., 2021) can be an effective method of perturbing and expanding the distribution of collected items, but users should take extra care in collected items such that they are properly aligned to the target skills (Bowman and Dahl, 2021; Kaushik et al., 2021). Wallace et al. (2022) find that iterating rounds of adversarial data collection improves the quality of collected data.

**Is there any control over the difficulty of items?** Taking control over the item difficulty is vital to evaluating the proficiency of target skills. However, it appears to be overlooked to capture the degree of skill proficiency in current NLU research. This gap in evaluation can be attributed to the difficulty of instantiating the degree of proficiency of a target skill as examples with different difficulties. In existing datasets, the number of reasoning steps in multi-hop QA may play this role (Yang et al., 2018; Wolfson et al., 2020). Sutcliffe et al. (2013) provide auxiliary questions that are simplified variants of the main questions and require fewer reasoning steps than the main questions. Bartolo et al. (2022) and Liu et al. (2022) have proposed a human–machine collaboration approach, that is, using a generator-in-the-loop data collection method for effectively helping annotators to enhance their creativity. Although it is exceedingly coarse for skill-wise analysis, several datasets provide subsets of the test set with different difficulties (Clark et al., 2018; Lai et al., 2017; Yu et al., 2020; Berzak et al., 2020). Item response theory also contributes to characterizing the item difficulty (refer to the second question in Appendix C.2), but the test creators have to use human responses for attributing the test scores to the target construct.

### C.5 Extrapolation

**Does the test set encourage performance comparison using related tasks and datasets?** The study of model development in NLU usually uses multiple datasets to report the performance of proposed models. However, stakeholders can choose datasets on which the proposed models perform well and refrain from reporting relatively lower scores on other datasets. To prevent this unfair practice, the test set needs to indicate similar datasets for reference and suggest that the test users evaluate their models on those datasets. Beyond a single task, using a collection of tasks in a single platform to facilitate the comparison of system performance across different tasks is informative (Wang et al., 2019; Hendrycks et al., 2021; Srivastava et al., 2022; Liang et al., 2022). Given that the language understanding capabilities may not depend on what language humans speak, cross-lingual applications are worth pursuing (Conneau et al., 2018; Artetxe et al., 2020).

**Does the test set include relevant basic tasks for sanity check?** The previous requirement is also applicable to basic tasks that are expected to be involved in the target NLU task. Ribeiro et al. (2020) have proposed a checklist approach that uses three types of auxiliary tasks: minimal functional test, invariance test, and directional expectation test. Depending on the primary target task, composing subtasks into a checklist enables the system developers to probe their system in detail and detect

unintended errors. This approach is helpful if we can create test cases for requisite knowledge and skills that are presumed in the domain definition.

### C.6 Utilization

**Does the test encourage the reporting of system statistics for utility?** Static leaderboards of benchmark datasets usually tell us which system is better than others in terms of simple evaluation metrics such as accuracy. However, they do not tell us about which system is most useful under conditions such as computational budget and inference time. Therefore, Ethayarajh and Jurafsky (2020) and Ma et al. (2021) advocate the reporting of model statistics, such as the size, energy efficiency, and inference latency. In a similar attempt, Min et al. (2021) propose a shared task for efficient open-domain QA models, comparing the QA performance under limited memory budgets. Bender et al. (2021) discuss the financial and environmental risks of using large-scale pretrained language models. De Vries et al. (2020) review the ecological validity of language user interfaces.

**Does the test set include disincentive for embedding harmful social biases?** Language models and word embeddings are often found to contain harmful social biases, such as stereotypes (Sun et al., 2019; Blodgett et al., 2020, 2021). To date, studies on social biases in NLU datasets have been limited (e.g., Rudinger et al., 2017). However, Li et al. (2020) find that underspecified questions with ambiguity in their answer candidates reveal various stereotypes. Mehrabi et al. (2021) propose quantifying representational harms in commonsense knowledge bases. Although it might not be harmful, falsehood should also be mitigated in foundation models (Lin et al., 2022).

**Does the test set encourage the mitigation of the system vulnerability against adversarial attacks?** Improving the robustness of systems against OOD input is one of the main concerns in the current NLP community. As Wallace et al. (2019a) demonstrate, NLU system predictions are easily changed by adversarial inputs, calling for improvements in the robustness against OOD data including malicious attacks. For this purpose, various methods applicable to NLU systems have been proposed, such as training with diverse data (Tu et al., 2020), systematic data augmentation (Min et al., 2020; Wu et al., 2021), and self-debiasing framework (Utama et al., 2020). Wang et al. (2022) provide a broad survey of datasets and methods for measuring and improving the robustness of NLP models.

## D Potential Arguments and Discussions

In this section, we discuss potential arguments on how we need to deal with our proposed framework.

**"Who should use the proposed framework and for what purpose?"** The framework is related to the entire experimental design including the construction of a dataset and its use for evaluating systems. Thus it should be mainly used by researchers who release datasets and propose their evaluation procedure, but dataset users (i.e., system developers in most cases) can also use this framework to see if the proposed procedure is well-designed, revise it if necessary, and validate their interpretations of system behavior.

**"Should we address all checklist questions to construct the validity argument? It may not be always possible to create such an argument owing to several constraints such as cost and data."** Addressing all questions in the checklist may not be practical. Despite the difficulty in creating thorough frameworks, the checklist contributes to clarifying the potential limitations of a study. Researchers and developers are encouraged to make justified accurate claims about their achievements. In addition, as the community grows and new practices are introduced, including all necessary practices in a single study is expected to become infeasible. Nonetheless, our framework provides a comprehensive reference to collect the necessary evidence for validating NLU evaluation.

**"Why is the discussion of the Octopus test, as well as the response-dependent formulation of language understanding, prerequisite for creating the validity argument?"** Our response-dependent formulation highlights the difficulty of developing a concise definition for NLU. The definition of NLU depends on the goal of stakeholders who use the test results as evidence. We suspect that exaggeration and underestimation in NLP can be attributed to the confusion about this response-dependent property of NLU; therefore, we discuss the problems of the Octopus test while referring to those of the original Turing test. Our focus is language understanding; hence, discussing the Turing test alone is not sufficient.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Limitation section*

☑ A2. Did you discuss any potential risks of your work?
*Limitation section*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☒ Did you use or create scientific artifacts?

*Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*No response.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*No response.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*No response.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*No response.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*No response.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*No response.*

### C ☒ Did you run computational experiments?

*Left blank.*

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*No response.*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*No response.*

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*No response.*

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*No response.*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*No response.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*No response.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*No response.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No response.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*No response.*