

It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance

Arjun Subramonian

University of California, Los Angeles
arjunsub@cs.ucla.edu

Xingdi Yuan

Microsoft Research
Eric.Yuan@microsoft.com

Hal Daumé III

University of Maryland
Microsoft Research
me@hal3.name

Su Lin Blodgett

Microsoft Research
SuLin.Blodgett@microsoft.com

Abstract

Progress in NLP is increasingly measured through benchmarks; hence, contextualizing progress requires understanding when and why practitioners may disagree about the validity of benchmarks. We develop a taxonomy of disagreement, drawing on tools from measurement modeling, and distinguish between two types of disagreement: 1) how tasks are conceptualized and 2) how measurements of model performance are operationalized. To provide evidence for our taxonomy, we conduct a meta-analysis of relevant literature to understand how NLP tasks are conceptualized, as well as a survey of practitioners about their impressions of different factors that affect benchmark validity. Our meta-analysis and survey across eight tasks, ranging from coreference resolution to question answering, uncover that tasks are generally not clearly and consistently conceptualized and benchmarks suffer from operationalization disagreements. These findings support our proposed taxonomy of disagreement. Finally, based on our taxonomy, we present a framework for constructing benchmarks and documenting their limitations.

1 Introduction

Claims of progress in NLP are often premised on how models perform on benchmarks for various NLP tasks¹ (e.g., coreference resolution, question answering) (Wang et al., 2018, 2019; Hu et al., 2020; Gehrmann et al., 2021). Benchmarks instantiate a task with a specific format, dataset of correct input-output pairs, and an evaluation metric (Bowman and Dahl, 2021), and they are intended to serve as measurement models for performance on the task. On the one hand, benchmarks allow for performance results to be easily compared across a rapidly-rising number of NLP

models (Schlangen, 2021; Ruder, 2021). Additionally, many NLP benchmarks are easily accessible via open-source platforms (Lhoest et al., 2021), which reduces the need of practitioners to construct new evaluation datasets and metrics from scratch. However, prior research has identified numerous threats to the validity of benchmarks (i.e., how well benchmarks assess the ability of models to correctly perform tasks). These threats include spurious correlations and poorly-aligned metrics (refer to Table 4 in the appendix).

However, little literature has surfaced sources of *disagreement* among NLP practitioners about benchmark validity, which is paramount to contextualize progress in the field. Hence, we develop a taxonomy of disagreement based on measurement modeling (from the social sciences (Adcock and Collier, 2001; Jacobs and Wallach, 2021)). Our taxonomy critically distinguishes between disagreement in how tasks are conceptualized and how measurements of model performance are operationalized (Blodgett et al., 2021). It thereby goes beyond prior examinations of NLP benchmarking methodology, which assume that tasks are generally clearly and consistently understood from person to person (Schlangen, 2021; Bowman and Dahl, 2021). This is important because our taxonomy captures that practitioners may perceive a benchmark for a task to have poor validity because they conceptualize the task differently than the benchmark creators do, and not simply because of the creators’ oversights or mechanistic failures when constructing the benchmark. (We validate this hypothesis empirically in § 5.1.) Furthermore, our taxonomy addresses that benchmarks can shape practitioners’ conceptualization of a task.

Ultimately, our taxonomy equips practitioners with a language to structure their thinking around and communicate their perceptions of benchmark validity. To provide evidence for our taxonomy, we conduct a survey of practitioners ($N = 46$) about

¹We disambiguate “benchmarks” and “tasks” in Appendix A.

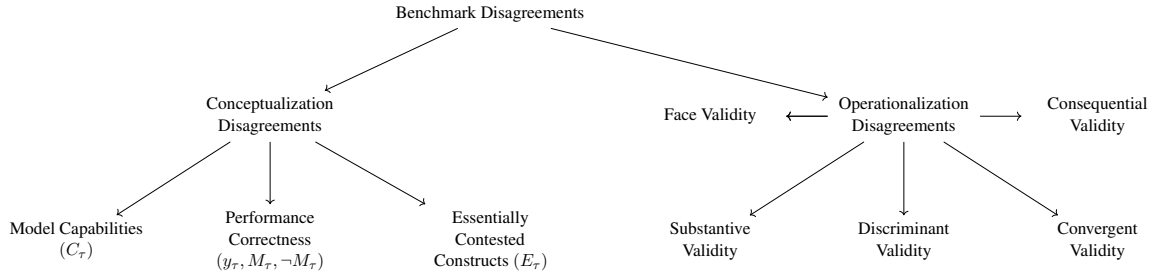


Figure 1: Bird’s eye view of our taxonomy comprising conceptualization and operationalization disagreements.

their opinions on different factors that affect benchmark validity: how contested tasks are and the quality of common benchmark formats, datasets, and metrics for tasks. We further conduct a meta-analysis of relevant literature to understand how tasks are conceptualized. Our meta-analysis and survey across eight tasks, ranging from coreference resolution to question answering, uncover that tasks are generally not clearly and consistently conceptualized and benchmarks suffer from operationalization disagreements. These findings support our taxonomy of disagreement. Finally, based on our taxonomy, we present a framework for constructing benchmarks and documenting their limitations.

2 Related Work

Community surveys Researchers have conducted community surveys of NLP evaluation practices, often to surface perceptions that are not stated in related literature. [Michael et al. \(2022\)](#) survey NLP practitioners to “elicit opinions on controversial issues” around benchmarking. [Zhou et al. \(2022\)](#) survey NLG practitioners to uncover “goals, community practices, assumptions, and constraints that shape NLG evaluations.” [Dev et al. \(2021\)](#) survey non-binary individuals to understand how they are not included in NLP model bias evaluations. We survey NLP practitioners to excavate perceptions of how contested tasks are and how well benchmarks measure model performance on tasks.

Benchmark validity A few previous works have studied benchmark validity through a measurement modeling lens ([Jacobs and Wallach, 2021](#)). [Blodgett et al. \(2021\)](#) analyze NLP bias evaluation benchmarks to inventory conceptualization and operationalization disagreements that threaten their validity as measurement models for stereotyping. [Liao et al. \(2021\)](#) review papers from various machine learning subfields to characterize benchmarks from the angles of internal and external validity.

[Raji et al. \(2021\)](#) argue that benchmarks cannot measure “progress towards general ability on vague tasks such as [...] ‘language understanding’,” and hence lack construct validity. We draw from measurement modeling to navigate how perceptions of validity issues with NLP benchmarks arise.

3 Taxonomy of Disagreement

We present our taxonomy of disagreement about the validity of NLP benchmarks (displayed in Figure 1). Drawing from measurement modeling ([Jacobs and Wallach, 2021](#)), our taxonomy critically distinguishes between disagreement in: 1) how a task τ is conceptualized, and 2) how a benchmark B_τ operationalizes measurements of model performance on τ . We provide evidence for our taxonomy in § 5, via our survey results and a meta-analysis of relevant literature.

3.1 Task Conceptualization

τ is *contested* when it lacks consistency or clarity in how it is conceptualized. In this case, because B_τ operationalizes measurements for model performance on B_τ ’s creators’² conceptualization of τ , there will necessarily be disagreement about the content validity of B_τ ([Jacobs and Wallach, 2021](#)). Disagreement in τ ’s conceptualization can stem from the following constructs with which τ is inextricably entangled:

- **Model capabilities:** Practitioners may disagree or lack clarity on the set of model capabilities C_τ that they assume τ involves ([Gardner et al., 2019](#); [Ribeiro et al., 2020](#); [Schlangen, 2021](#)). Our conceptualization of C_τ is broader than “cognitive

²By “creators,” we refer to all individuals involved in the construction of B_τ , including crowdworkers. We do not claim that all the creators of B_τ necessarily have nor does B_τ necessarily encode a consistent conceptualization of τ . For example, the Universal Dependencies Treebank attempts to consolidate different conceptualizations of dependency parsing ([Nivre et al., 2016](#)); hence, it likely fails to exactly match any individual linguist’s conceptualization of syntax.

capabilities” (Paullada et al., 2021), encompassing e.g., handling various genres of text. However, C_τ can also include the coarse-grained capability of performing τ correctly. In contrast to Schlangen (2021), we argue that practitioners may determine C_τ in a top-down or bottom-up manner. They may first conceptualize τ as a specific real-world application and identify C_τ required to meet the needs of application users (Cao et al., 2022). Alternatively, practitioners may first identify C_τ that they believe to be linguistically interesting or crucial to general-purpose language systems, and subsequently devise τ such that C_τ is necessary to perform τ correctly (Pericliev, 1984; Schlangen, 2021; Mahowald et al., 2023). In both cases, we gauge the extent to which a model possesses C_τ by proxy, by attempting to measure its performance on τ .

- **Performance correctness:** Practitioners may disagree or lack clarity on what constitutes performing τ correctly (Jamison and Gurevych, 2015; Baan et al., 2022; Plank, 2022). This could include different perspectives on correct outputs y_τ , as well as acceptable methods M_τ and unacceptable methods $\neg M_\tau$ for performing τ correctly (Teney et al., 2022).
- **Essentially contested constructs:** Practitioners often disagree or lack clarity on essentially contested constructs E_τ entangled with τ . A construct is essentially contested when its significance is generally understood, but there is frequent disagreement on what it looks like (e.g., language understanding, justice) (Gallie, 1955). Developing criteria for whether a construct is essentially contested has been a subject of philosophical study for decades. For instance, Gallie (1955) posited that essentially contested constructs must have “reciprocal recognition of their contested character among contending parties” and “an original exemplar that anchors conceptual meaning,” among other characteristics (Collier et al., 2006).

Model capabilities, performance correctness, and essentially contested constructs are mutually-building. C_τ (capabilities assumed to be involved to perform τ correctly) rely on a particular understanding of y_τ . Similarly, M_τ (acceptable methods for performing τ correctly) may overlap with C_τ .

3.2 Perceptions of Benchmark Validity

Our taxonomy connects disagreement in how τ is conceptualized to impressions of the validity of B_τ (i.e., how well B_τ operationalizes measurements of model performance on τ). In particular, there are two reasons for perceptions of poor benchmark validity: disagreements in how the task is conceptualized, and operationalization disagreements. We delve into these reasons, with examples, in § 5.

- **Conceptualization disagreements:** Disagreements in how practitioners fundamentally conceptualize an aspect of τ (e.g., C_τ , y_τ , M_τ , $\neg M_\tau$, E_τ) necessarily yields disagreements about the content validity of B_τ . For example, Williams et al. (2018) construct MNLi because they conceptualize natural language inference as requiring models to handle various text genres, which they perceive SNLI “falls short of providing a sufficient testing ground for” because “sentences in SNLI are derived from only a single text genre.” Additionally, practitioners’ conceptualizations of tasks can evolve over time, and even be influenced by the benchmarks with which they work. For instance, SQuAD arguably radically shifted practitioners’ conceptualizations of QA from open-ended information retrieval to reading comprehension-style questions (Rajpurkar et al., 2016). As such, constructing valid benchmarks for a task can be a game with a shifting goalpost.
- **Operationalization disagreements:** Consider a set P_{B_τ} of practitioner(s) whose conceptualization of an aspect of τ aligns with that of the creators of B_τ . Operationalization disagreements are choices made by the creators of B_τ (with respect to task format, dataset, and metric) that even within P_{B_τ} , engender divergent perceptions of B_τ ’s validity. As an example, consider practitioners P_{B_τ} who believe that metrics for machine translation quality should “yield judgments that correlate highly with human judgments” (Pillutla et al., 2021). Pillutla et al. (2021), motivated by their impression that popular automatic evaluation metrics in NLG (e.g., BLEU, ROUGE) “weakly” operationalize how humans judge machine translations, propose a new metric MAUVE.

We provide an extended discussion of conceptualization and operationalization disagreements in Appendix C.

4 Survey Methodology

With our taxonomy in mind, we conduct a survey of NLP practitioners³ ($N = 46$) to surface and understand for various NLP tasks, practitioners’ perceptions of: (1) the extent to which the tasks appear to have a clear and consistent conceptualization and (2) the quality of benchmarks (with respect to task format, dataset, and metric). We ultimately chose to include the following tasks in our survey: Sentiment Analysis (SENT), Natural Language Inference (NLI), Question Answering (QA), Summarization (SUM), Machine Translation (MT), Named-Entity Recognition (NER), Coreference Resolution (COREF), and Dependency Parsing (DEP). We detail our task selection protocol in Appendix D.

Survey topics In our survey, we begin by asking participants about their background (i.e., occupation and experience with NLP) to understand the demographics of our sample. We then inquire into participants’ initial impressions of how current state-of-the-art NLP models perform on various NLP tasks; we do this prior to asking participants to engage more critically with task definitions and benchmarks, so as not to sway their responses. Subsequently, for each task, we ask participants about their familiarity with the task, and if they are familiar, their perceptions of the (a) clarity and consistency of the task’s definition or conceptualization, (b) extent to which common task formats capture the underlying language-related skill, (c) quality of benchmark datasets and metrics, and (d) progress on the task. We utilize perceptions of (a) as a proxy for how contested tasks are across practitioners. We do this because it is not feasible to collect and compare participants’ raw task conceptualizations in a quantifiable manner. Furthermore, we collect perceptions of (b) and (c) to capture conceptualization and operationalization disagreements across benchmarks generally. We do not inquire into participants’ impressions of (b) and (c) for specific benchmarks in order to keep the survey reasonably long and have a sufficient sample size.

For all survey questions that ask participants to rate their perception, we provide them with a scale that ranges from 1 to 6 with articulations of what

³Following Zhou et al. (2022), by “practitioners,” we refer to academic and industry researchers, applied scientists, and engineers who have experience with NLP tasks or evaluating NLP models or systems.

Role	#
Works on deployed systems	6
Industry practitioner (not researcher)	7
Industry researcher	10
Academic researcher	32

Table 1: Demographics of survey participants. Some participants identified with more than one role.

1 and 6 mean in the context of the question. We include the entirety of our survey and survey results in Appendix H, and discuss participant guidance in Appendix E.

Survey recruitment and quality control As seen in Table 1, our sample is heavily skewed towards academic researchers; we detail our participant recruitment protocol and IRB approval in Appendix G. We additionally document our quality control measures in Appendix F.

5 Results

5.1 Task Conceptualization

Figure 2 shows how survey participants perceive the clarity and consistency with which various NLP tasks are conceptualized. We observe that:

- **Tasks are not perfectly clearly or consistently conceptualized.** No task in Figure 2 received a score of 6 from all participants.
- **Tasks are conceptualized with varying levels of clarity and consistency.** The tasks in Figure 2 exhibit a range of average and median conceptualization scores. NLI and SENT appear to have objectives that are less clearly and consistently understood by practitioners, while COREF and MT seem to be better defined.
- **Practitioners diverge in their impressions of how clearly and consistently the NLP community conceptualizes a task.** Many tasks in Figure 2 have a large interquartile range, and for NLI and SENT, scores span from 2 to 6.

To further provide evidence for these observations, we leverage our taxonomy (in particular, the sources of disagreement in task conceptualization described in § 3.1) and relevant literature.

Model capabilities In order to understand disagreement about involved capabilities C_τ for the tasks, we meta-analyze benchmarks that survey participants mention. Specifically, for each task, we first select the 2–4 most frequently mentioned benchmarks; we then perform light open coding⁴

⁴Open coding refers to “labeling concepts, defining and de-

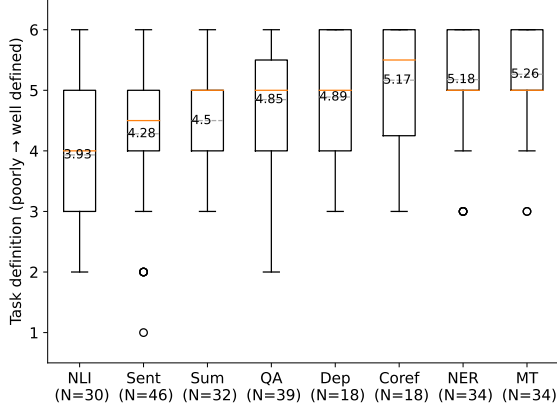


Figure 2: Perceived clarity and consistency of task definition. Orange lines indicate median score, while dashed lines indicate average score.

on the papers that initially proposed these benchmarks in order to identify model capabilities⁵ that the authors claim the benchmark assesses.

We find that for each task, stated capabilities overlap but often vary across benchmarks, suggesting disagreement in task conceptualization. For instance, for SUM, the authors of XSum claim that the benchmark assesses whether models can generate novel language, handle linguistic phenomena, and handle various domains (Narayan et al., 2018), while the authors of CNN/Daily Mail claim that this benchmark gauges whether models possess benchmark-external knowledge (Nallapati et al., 2016); however, authors of both benchmarks intend to test language understanding. We present all our meta-analysis results in Appendix J.

Performance correctness Correct outputs y_τ for a task may be inherently disagreed upon or unclear. For instance, in MT, the adequacy of translations in y_τ is subjective (White and O’Connell, 1993); further, it can be unclear how to translate lexical and syntactic ambiguity in the source language (Pericliev, 1984; Baker et al., 1994), or translate from a language without to with grammatical gender (Gonen and Webster, 2020). We present additional examples in Table 2.

Practitioners can also disagree about acceptable methods M_τ and unacceptable methods $\neg M_\tau$ for

veloping categories based on their properties and dimensions” without a predefined list of categories (Khandkar).

⁵We restrict our attention to stated capabilities that lie below the surface of the capability of performing the task correctly (Schlangen, 2021). Some annotated datasets when proposed, were not intended for model evaluation, but were later repurposed as benchmarks (e.g., Penn Treebank (Marcus et al., 1993)).

performing a task correctly. For example, Sugawara et al. (2020) expect models to take certain actions when performing reading comprehension, e.g., {recognize word order, resolve pronoun coreferences} $\subset M_\tau$. On the other hand, numerous works have raised concerns about models exploiting annotation artifacts in NLI (Gururangan et al., 2018; Poliak et al., 2018) and QA benchmarks (Si et al., 2019; Kavumba et al., 2019; Chen and Durrett, 2019), which suggests that they view exploiting artifacts as part of $\neg M_\tau$.

Essentially contested constructs E_τ pose an issue when practitioners incorrectly presuppose that E_τ have clear and consistent conceptualizations, thus failing to communicate how they personally understand E_τ . We present examples of essentially contested constructs E_τ entangled with various tasks in Table 2, elaborating on a few in this section. SENT presupposes that the essentially contested construct “sentiment:” 1) has a clear and consistent definition (e.g., falls on a spectrum between “positive” and “negative”); 2) can be gleaned from text alone; and 3) admits expressions that are universally or predominantly interpreted the same way from person to person. However, there exists “divergences of sentiments about different concepts” across cultures (Heise, 2014), and hence “sentiment” $\in E_\tau$. Furthermore, COREF, in asking if two expressions refer to the same entity, presupposes that the essentially contested construct “identity” is clearly and consistently understood, and thus “identity is never adequately defined” (Recasens et al., 2010).

Often, C_τ and E_τ overlap. As revealed by our meta-analysis of model capabilities, practitioners may believe that performing certain tasks involves:

- **Possessing benchmark-external knowledge:** But, what constitutes benchmark-external knowledge is left ambiguous. For example, in QA, questions may involve “commonsense knowledge” (Talmor et al., 2019; Schlegel et al., 2020), whose constitution is essentially unclear and inconsistently understood (Mueller, 2015). It is also unclear how much external knowledge and context NER requires to disambiguate entities (Ratinov and Roth, 2009).
- **Being on par with humans:** However, practitioners often do not specify which humans (e.g., crowdworkers, trained syntacticians) with which they would like models to be on par, or use vague or problematic language in their specifi-

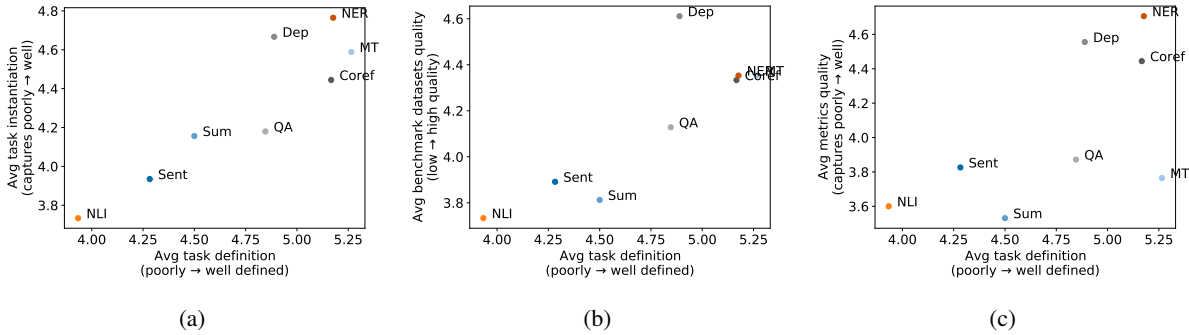


Figure 3: Perceived quality of common task instantiations, benchmark datasets, and benchmark metrics vs. perceived clarity and consistency of task definition.

cations (e.g., “normally-abled adults whose first language is English” (Levesque et al., 2011)).

We discuss additional examples of essentially contested constructs in Appendix K.

5.2 Perceptions of Benchmark Validity

Figure 3 depicts for various tasks, how survey participants’ perceptions of the quality of common benchmark task instantiations, datasets, & metrics (which are central to benchmark validity) vary in relation to their perceptions of the clarity and consistency of how the task is defined. These plots show that there is generally a positive association between perceptions of benchmark validity and task contestedness. This observation indicates that benchmarks suffer from conceptualization disagreements. However, this observation could also reflect that NLP practitioners collapse task contestedness onto their perceptions of benchmark validity.

The plots also demonstrate that the association (especially between perceptions of metric quality and task contestedness) is weak, with seemingly well-defined tasks like MT facing impressions of low-quality metrics. This association weakness suggests that benchmarks suffer from operationalization disagreements. To provide evidence for our findings, we leverage relevant literature.

Conceptualization disagreements We describe some disagreements in the conceptualization of NLP tasks and provide examples of resultant conceptualization disagreements in Table 2.

Operationalization disagreements Operationalization disagreements can be attributed to various factors. Measurement modeling naturally provides us with a language to categorize and discuss these factors, and in the process, theorize about the real world. Hence, we taxonomize operationalization disagreements through the lens of different threats

to validity in the measurement modeling literature.

- **Face validity:** Benchmarks can have surface characteristics (e.g., incorrect or incomplete annotations) that affect perceptions of their quality. For instance, QA, COREF, and NER benchmarks often contain incorrect or incomplete annotations (Jie et al., 2019; Schlegel et al., 2020; Blodgett et al., 2021). Many SUM benchmarks have unfaithful reference summaries (Zhang et al., 2022; Tang et al., 2022; Goyal and Durrett, 2021). MT benchmarks often contain incorrect reference translations (Castilho et al., 2017).
- **Substantive validity:** A benchmark may not exhaustively assess a model capability (Schlangen, 2021). For example, practitioners may conceptualize a task as involving the capability to handle phenomena in real-world data, but benchmark datasets (e.g., from “constrained social media platforms”) can fail to “reflect broader real-world phenomena” (Olteanu et al., 2016; Hupkes et al., 2022). For example, despite having saturated SST-2 (Wang et al., 2019), NLP models struggle with domain shift, bi-polar words, negation (Hussein, 2018; Hossain et al., 2022). Furthermore, QA benchmarks are often restricted to a single format (e.g., multiple-choice reading comprehension, story-cloze queries (Schlegel et al., 2020)), which does not substantively instantiate QA. Moreover, the format of MT benchmarks (e.g., of WMT shared tasks) often precludes sufficient intersentential context for substantively assessing translations (Toral, 2020).
- **Discriminant validity:** Benchmarks may inadvertently assess undesired model capabilities or “unacceptable” methods of performing a task (e.g., picking up on spurious cues) (Jacobs and Wallach, 2021). For instance, despite having saturated SuperGLUE NLI benchmarks (Wang

Task	Disagreement in conceptualization?	Disagreement examples
NLI	C_τ : yes (Table 6).	SNLI, MNLI datasets operationalize validity of natural language inferences with single gold label (Bowman et al., 2015; Williams et al., 2018).
	y_τ : yes; inherent disagreement in validity of natural language inferences (Pavlick and Kwiatkowski, 2019); lack of clarity and disagreement about y_τ when premise or hypothesis is question (Figure 13).	
	E_τ : yes; {understand language, possess benchmark-external knowledge} $\subset E_\tau$ (Table 6).	
QA	C_τ : yes (Table 7).	HotpotQA, ReCoRD, MultiRC datasets operationalize reference answers with arbitrary precision (Schlegel et al., 2020).
	y_τ : yes; appropriate adequacy of answers in y_τ is subjective (Schlegel et al., 2020).	
	E_τ : yes; {understand language, reason over a context, possess benchmark-external knowledge, be on par with humans} $\subset E_\tau$ (Table 7).	
COREF	C_τ : yes (Table 8).	OntoNotes dataset does not capture near-identity coreferences (Recasens et al., 2010; Zeldes, 2022).
	y_τ : yes; inherent anaphoric ambiguity induces lack of clarity and disagreement about y_τ (Poesio and Artstein, 2005).	
	E_τ : yes; {identity, be on par with humans, possess benchmark-external knowledge} $\subset E_\tau$ (Table 8).	
SUM	C_τ : yes (Table 9).	benchmark datasets contain single gold summaries with varying levels of adequacy (Kano et al., 2021).
	y_τ : yes; "goodness" and adequacy of summaries in y_τ are subjective (Nallapati et al., 2016; Li et al., 2021; Ter Hoeve et al., 2022).	
	E_τ : yes; {understand language, possess benchmark-external knowledge} $\subset E_\tau$ (Table 9).	

Table 2: Disagreements in the conceptualization of NLP tasks and relevant examples.

et al., 2019), NLP models fail on a controlled evaluation set where it is not possible to rely on syntactic heuristics (McCoy et al., 2019).

- **Convergent validity:** Benchmarks may not “match other accepted measurements” of performance on a task. For example, practitioners may consistently conceptualize SUM and MT as involving “being on par with humans”; however, automatic evaluation metrics like ROUGE and BLEU are poorly aligned with human judgments of summarization (Deutsch and Roth, 2021; Deutsch et al., 2022) and translation (Reiter, 2018; Toral, 2020; Marie et al., 2021; Amrhein et al., 2022) quality, respectively. This is reflected in Figure 3c, which shows that SUM and MT noticeably deviate from the positive trend; in particular, although these tasks are more consistently and clearly conceptualized, practitioners perceive their metrics to be low-quality (i.e., SUM and MT benchmarks have poor convergent validity).
- **Consequential validity:** Practitioners may be concerned that the use of a benchmark has societal harms. For example, SENT benchmarks can reinforce hegemonic conceptions of emotion and be culturally discriminatory (Crawford, 2021).

Benchmark issues may threaten more than a single aspect of validity.

5.3 Progress

Figure 4 (and Figure 5 in the appendix) suggest that perceptions of better task conceptualization

and benchmark validity are associated with perceptions of stronger progress on the task. In reality, impressions of progress in NLP (especially for non-practitioners) may be disconnected from the validity of the benchmarks used to make claims about progress (Bender et al., 2021). This is important because claims of progress shape the social and academic capital of NLP, and are implicitly embedded in every research artifact, including scientific publications. Thus, towards proper science and accountability, NLP practitioners ought to make realistic and tenable claims about progress, and not over-hype NLP models. Furthermore, progress is neither monolithic nor does it increase monotonically; it is critical to be transparent about benchmark validity issues and their implications for claims of progress.

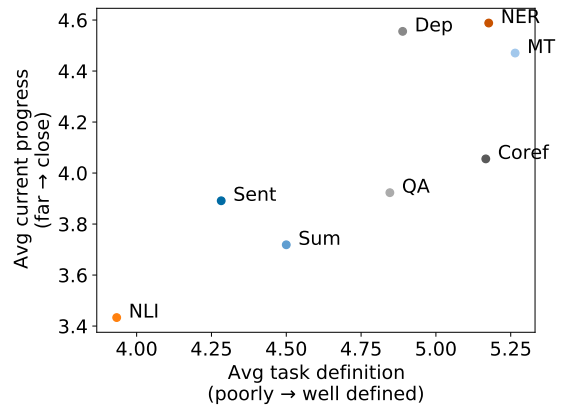


Figure 4: Perceived quality of common task instantiations, benchmark datasets, and benchmark metrics vs. perceived current progress on task.

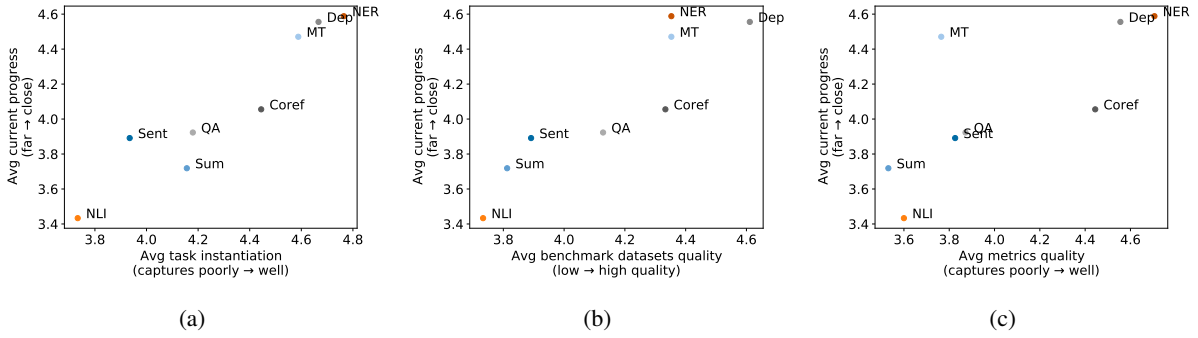


Figure 5: Quality of common task instantiations, benchmark datasets, and benchmark metrics vs. perceived current progress on task among all responding practitioners.

We must simultaneously re-imagine “progress” in NLP to encapsulate measuring, alleviating, and communicating benchmark validity issues.

6 A Framework for NLP Benchmarks

Towards better documenting benchmarks’ conceptualization and operationalization, we encourage benchmark creators to answer the questions in Table 3 in their future directions or limitations section when they propose a new benchmark B_τ for a task τ . This framework is not a post-hoc intervention. We intend for benchmark creators to answer these questions before, during, and after they construct benchmarks; this framework should be grounded in care for and facilitating collective progress in NLP. Furthermore, creators should share their answers to these questions, so that this framework becomes normalized and shapes people’s thinking about their own contributions. Moreover, this framework is intended to supplement processes like Datasheet for Datasets and Data Statements for NLP (Gebu et al., 2021; Bender and Friedman, 2018), which enable comprehensive documentation for benchmarks, but do not ask benchmark creators to reflect in a way that distinguishes between: 1) how they conceptualize a task (and how others may disagree with their conceptualization), and 2) how well the benchmark operationalizes a measurement model for model performance on their conceptualization of the task. This framework is also complementary to technical solutions (e.g., human-in-the-loop approaches) to resolving task ambiguity (Tamkin et al., 2022).

We hope that this reflection will benefit the NLP community in the following ways:

- **Reduce overhyping:** By being transparent about and defining the model capabilities that benchmarks are intended to assess, as well as docu-

menting benchmark validity issues, benchmark creators will: 1) not misrepresent model capabilities, and 2) remind people to be careful about extrapolating benchmark performance results.

- **Encourage reflexivity and engagement with the politics of benchmarks:** By clarifying how they conceptualize tasks and considering how others may disagree with their conceptualization, benchmark creators will: 1) assess how their social context and power influences task conceptualization and benchmark construction (Collins, 2017), 2) reflect on which groups of people benchmarks represent, and 3) include people from diverse communities during benchmark construction towards alleviating disagreement. Towards considering historical and social context, we urge practitioners to not neutralize disagreements in conceptualization by valuing all “sides” equally, as this inevitably invalidates marginalized people’s lived experiences and perpetuates the power relations in which benchmark construction participates (Collins, 2017; Denton et al., 2021). In particular, the widespread adoption, presumed validity, and inertia of benchmarks influence the direction of NLP, shaping funding landscapes and the domains in which NLP systems are deployed (Blili-Hamelin and Hancox-Li, 2022; Bommasani, 2022). As such, we encourage practitioners to prioritize the perspectives of marginalized people.
- **Provide actionable insights to address benchmark validity issues:** Distinguishing between conceptualization and operationalization disagreements in a benchmark will better enable the creators of the benchmark, as well as creators of future benchmarks, to address benchmark validity issues. For example, to address perceptions that a benchmark does not exhaus-

<p>Conceptualization questions</p> <p>Model capabilities: Which C_τ do you believe τ involves and why? (e.g., Table 1 in Ribeiro et al. (2020)) How does C_τ differ from the capabilities that other benchmarks for τ are intended to assess?</p> <p>Performance correctness: How may $y_\tau, M_\tau, \neg M_\tau$ be contested? How did you involve relevant communities to co-create B_τ? How would you accurately characterize “solving” τ?</p> <p>Essentially contested constructs: Do you define any E_τ (e.g., model capabilities) entangled with τ? (e.g., “universality” in Bhatt et al. (2021)) How did you come up with the name of τ and B_τ? Do you avoid employing overloaded or overclaiming terminology in your τ’s name (Shanahan, 2022)?</p> <p>Overarching questions: How may B_τ limit “progress” to only working on one conceptualization of τ? Do you hold space for others to propose alternatives?</p>
<p>Operationalization questions</p> <p>Validity: How well does B_τ operationalize a measurement model for model performance on your conceptualization of τ? What kinds of validity may B_τ lack and why? If B_τ were to indicate that a model performs exceptionally well on it, what can the NLP community conclude?</p>

Table 3: Documentation questions to facilitate the creation of NLP benchmarks.

tively assess whether models can “handle real-world phenomena,” benchmark creators can decide if this is a conceptualization disagreement (e.g., “real-world” is too open-ended, in which case creators should clearly explain which domains they foreground in their conceptualization of “real-world”) or operationalization disagreement (e.g., acquiring real-world data is difficult.)

7 Conclusion

We develop a taxonomy of disagreement (based on measurement modeling) which distinguishes between how tasks are conceptualized and how measurements of model performance are operationalized. To provide evidence for our taxonomy, we conduct a survey of practitioners and meta-analysis of relevant literature. Based on our taxonomy, we propose a framework for the creation of benchmarks and the documentation of their limitations. Future work includes studying task conceptualization via benchmark inter-annotator disagreement.

Limitations

Survey limitations Our survey sample size over-represents English-speaking NLP practitioners, and likely practitioners from the United States. While we would like to study the demographic skews in our sample (e.g., seniority) and its implications for the results in our paper, we could not collect demographic data due to privacy concerns. Nevertheless, our results still highlight that even within skewed samples, there exists weak agreement on how tasks are conceptualized. Additionally, we assume that survey participants do not base their perceptions of task conceptualization on surface characteristics of tasks, or task ethos (e.g., task longevity, task popularity, rhetoric associated with the task). Furthermore, while we provide some justification for the 6-point scale in Appendix E, the scale is not optimal, as not many participant judgments are below 4; we had not run a similar survey previously, nor did our pilot responses indicate that many judgments would be ≥ 4 . Finally, while we would like to provide a qualitative analysis of participants’ free responses, the majority of participants did not answer the “Additional Thoughts” questions.

Meta-analysis limitations We largely focus on static textual single-task English-language benchmarks. Furthermore, we assume that the capabilities stated by authors generally represent the primary capabilities that they believe the task involves; however, authors may refrain from including particular information due to space limits or reviewing incentives.

Framework limitations While our proposed framework for creating benchmarks has not been explicitly tested, we have confidence in its efficacy as it was borne out of our systematic analysis of NLP practitioners, literature, and benchmarks. We ultimately wish to implement the framework, but doing so is beyond the scope of this paper (whose primary focus is a systematic perspective on disagreements on evaluative practices in NLP), and leave it to future work.

Ethics and Broader Impact

We obtained informed consent from all survey participants, and the survey was IRB-approved. In administering the survey, we did not collect any personally identifiable information that could be traced back to participants’ responses, and we transparently communicated our data privacy, usage,

and retention policies (refer to Appendix H.1). Furthermore, we shared our survey with artificial intelligence affinity groups to increase the diversity of our sample. We detail our participant recruitment protocol and IRB approval in Appendix G. Additionally, in our paper, we discuss our taxonomy and benchmark documentation guidelines in the context of scientific accountability, power relations, and path dependence in NLP.

Acknowledgements

We thank Alexander M. Hoyle and Eve Fleisig for their feedback on early versions of our survey. We further thank Li Lucy, Kai-Wei Chang, Jieyu Zhao, Swabha Swayamdipta, and the reviewers for their feedback on the writing of this paper.

References

- Lauren Ackerman. 2015. Influences on parsing ambiguity.
- Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95:529–546.
- Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutiérrez, and Krys J. Kochut. 2017. Text summarization techniques: A brief survey. *ArXiv*, abs/1707.02268.
- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. Aces: Translation accuracy challenge sets for evaluating machine translation metrics. *arXiv preprint arXiv:2210.15615*.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. *ArXiv*, abs/2210.16133.
- Kathryn L. Baker, Alexander Franz, Pamela W. Jordan, Teruko Mitamura, and Eric Nyberg. 1994. Coping with ambiguity in a large-scale machine translation system. In *COLING*.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. 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? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Shaily Bhatt, Poonam Goyal, Sandipan Dandapat, Monojit Choudhury, and Sunayana Sitaram. 2021. On the universality of deep contextual language models. In *ICON*.
- Borhane Blili-Hamelin and Leif Hancox-Li. 2022. Making intelligence: Ethics, iq, and ml benchmarks. *ArXiv*, abs/2209.00692.
- Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna M. Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *ACL*.
- Rishi Bommasani. 2022. Evaluation for change. *ArXiv*, abs/2212.11670.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.
- Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *ArXiv*, abs/2104.02145.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. [\(meta-\) evaluation of machine translation](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.
- Yang Trista Cao, Kyle Seelman, Kyungjun Lee, and Hal Daum’e. 2022. What’s different between visual question answering for machine “understanding” versus for accessibility? In *AACL*.
- Sheila Castilho, J. Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sasoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of pbsmt and nmt using professional translators.
- Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. *ArXiv*, abs/1904.12106.
- David Collier, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. 2006. Essentially contested concepts: Debates and applications. *Journal of Political Ideologies*, 11:211 – 246.
- Patrícia Hill Collins. 2017. Black feminist thought in the matrix of domination from.
- 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.
- Kate Crawford. 2021. [Artificial intelligence is misreading human emotion](#).
- Aida Mostafazadeh Davani, Mark D’iaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *ArXiv*, abs/2107.07002.

- Emily L. Denton, Mark D’iaz, Ian D. Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *ArXiv*, abs/2112.04554.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. *ArXiv*, abs/2204.10216.
- Daniel Deutsch and Dan Roth. 2021. [Understanding the extent to which content quality metrics measure the information quality of summaries](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.
- Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. [Harms of gender exclusivity and challenges in non-binary representation in language technologies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graça, and Fernando C Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In *EMNLP*.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar R Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? *ArXiv*, abs/2204.07931.
- 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.
- Dan Friedman, Alexander Wettig, and Danqi Chen. 2022. Finding dataset shortcuts with grammar induction. *ArXiv*, abs/2210.11560.
- W. B. Gallie. 1955. [Essentially contested concepts](#). *Proceedings of the Aristotelian Society*, 56:167–198.
- Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful? *ArXiv*, abs/1909.11291.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ArXiv*, abs/1803.07640.
- Timnit Gebru, Jamie H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64:86–92.
- Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjana Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. [The GEM benchmark: Natural language generation, its evaluation and metrics](#). In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.
- Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and William B. Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *TAC*.
- Hila Gonen and Kellie Webster. 2020. [Automatically identifying gender issues in machine translation using perturbations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.
- Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *ArXiv*, abs/2104.04302.
- 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.
- Gerhard Hagerer, Dávid Szabó, Andreas Koch, Maria Luisa Ripoll Dominguez, Christian Widmer, Maximilian Wich, Hannah Danner, and Georg Groh. 2021. End-to-end annotator bias approximation on crowd-sourced single-label sentiment analysis. *ArXiv*, abs/2111.02326.
- David R. Heise. 2014. Cultural variations in sentiments. *SpringerPlus*, 3.

- Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *ACL*.
- Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Iris Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *EMNLP*.
- Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90% solution. In *NAACL*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080.
- Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella J. Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in nlp: a taxonomy and review. *ArXiv*, abs/2210.03050.
- Doaa Mohey El Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*.
- Abigail Z. Jacobs and Hanna M. Wallach. 2021. Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Emily Jamison and Iryna Gurevych. 2015. [Noise or additional information? leveraging crowdsourcing annotation item agreement for natural language tasks](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.
- Hetvi Jethwani, Arjun Subramonian, William Agnew, MaryLena Bleile, Sarthak Arora, Maria Ryskina, and Jeffrey Xiong. 2022. [Queer in ai](#). *XRDS*, 28(4):18–21.
- Nan Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *ArXiv*, abs/2209.03392.
- Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. [Better modeling of incomplete annotations for named entity recognition](#). 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 729–734, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniec, Piotr Miłkowski, Jan Kocon, and Przemysław Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. *Proceedings of the 1st Workshop on Perspective Approaches to NLP LREC2022*.
- Ryuji Kano, Takumi Takahashi, Toru Nishino, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2021. [Quantifying appropriateness of summarization data for curriculum learning](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1395–1405, Online. Association for Computational Linguistics.
- Divyansh Kaushik and Zachary Chase Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *EMNLP*.
- Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reiser, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. *ArXiv*, abs/1911.00225.
- Katherine Keith, Su Lin Blodgett, and Brendan O’Connor. 2018. [Monte Carlo syntax marginals for exploring and using dependency parses](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 917–928, New Orleans, Louisiana. Association for Computational Linguistics.
- Shahedul Huq Khandkar. [Open coding](#).
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MTSUMMIT*.
- 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, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *KR*.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario vSavsko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu,

- Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clement Delangue, Th'eo Matussiere, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, Francois Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. *ArXiv*, abs/2109.02846.
- Lei Li, Wei Liu, Marina Litvak, Natalia Vanetik, Jiacheng Pei, Yinan Liu, and Siya Qi. 2021. Subjective bias in abstractive summarization. *ArXiv*, abs/2106.10084.
- Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. [Are we learning yet? a meta review of evaluation failures across machine learning](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, A. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*.
- Kyle Mahowald, Anna A. Ivanova, Idan Asher Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19:313–330.
- 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.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *ArXiv*, abs/1902.01007.
- Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, , and Samuel R. Bowman. 2022. [What do nlp researchers believe? results of the nlp community metasurvey](#).
- Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2022. Extrinsic evaluation of machine translation metrics.
- Erik T. Mueller. 2015. [Chapter 19 - acquisition of commonsense knowledge](#). In Erik T. Mueller, editor, *Commonsense Reasoning (Second Edition)*, second edition edition, pages 339–363. Morgan Kaufmann, Boston.
- Mahin Naderifar, Hamideh Goli, and Fereshteh Ghaljaie. 2017. Snowball sampling: A purposeful method of sampling in qualitative research.
- Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *ArXiv*, abs/2010.03532.
- Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *ArXiv*, abs/1907.07355.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. [The CoNLL 2007 shared task on dependency parsing](#). In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic. Association for Computational Linguistics.
- Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *ArXiv*, abs/2103.14749.
- Alexandra Olteanu, Carlos Castillo, Fernando D. Diaz, and Emre Kıcıman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. *SSRN Electronic Journal*.
- Cecilia Ovesdotter Alm. 2011. [Subjective natural language problems: Motivations, applications, characterizations, and implications](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.
- Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. [A case for a range of acceptable annotations](#). In *Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing*.

- Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don’t blame the annotator: Bias already starts in the annotation instructions. *ArXiv*, abs/2205.00415.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton, and A. Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Vladimir Pericliev. 1984. Handling syntactical ambiguity in machine translation. In *ACL*.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. [Mauve: Measuring the gap between neural text and human text using divergence frontiers](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.
- Barbara Plank. 2022. The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. [Linguistically debatable or just plain wrong?](#) In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Massimo Poesio and Ron Artstein. 2005. [The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account](#). In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. [CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes](#). In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.
- Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily L. Denton, and A. Hanna. 2021. Ai and the everything in the whole wide world benchmark. *ArXiv*, abs/2111.15366.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.
- Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *NAACL*.
- Lev-Arie Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.
- Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2010. [A typology of near-identity relations for coreference \(NIDENT\)](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, Just Accepted:1–8.
- 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.
- Sebastian Ruder. 2021. [Challenges and opportunities in nlp benchmarking](#).
- Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *CoNLL*.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *ArXiv*, abs/2111.07997.

- 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 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. [The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.
- Nikil Rooshan Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2022. [The tail wagging the dog: Dataset construction biases of social bias benchmarks](#). *ArXiv*, abs/2210.10040.
- Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. [Quantifying social biases using templates is unreliable](#). *ArXiv*, abs/2210.04337.
- Murray Shanahan. 2022. [Talking about large language models](#).
- Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. [What does bert learn from multiple-choice reading comprehension datasets?](#) *ArXiv*, abs/1910.12391.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Shane Storks, Qiaozhi Gao, and Joyce Yue Chai. 2019. [Recent advances in natural language inference: A survey of benchmarks, resources, and approaches](#). *arXiv: Computation and Language*.
- Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. [Assessing the benchmarking capacity of machine reading comprehension datasets](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8918–8927.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [Commonsenseqa: A question answering challenge targeting commonsense knowledge](#). *ArXiv*, abs/1811.00937.
- Alex Tamkin, Kunal Handa, Ava Shrestha, and Noah D. Goodman. 2022. [Task ambiguity in humans and language models](#).
- Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryscinski, Justin F. Rousseau, and Greg Durrett. 2022. [Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors](#). *ArXiv*, abs/2205.12854.
- Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. 2022. [Predicting is not understanding: Recognizing and addressing underspecification in machine learning](#). In *ECCV*.
- Maartje Ter Hoeve, Julia Kiseleva, and Maarten Rijke. 2022. [What makes a good and useful summary? Incorporating users in automatic summarization research](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 46–75, Seattle, United States. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Antonio Toral. 2020. [Reassessing claims of human parity and super-human performance in machine translation at wmt 2019](#). In *EAMT*.
- Kiri L. Wagstaff. 2012. [Machine learning that matters](#). *ArXiv*, abs/1206.4656.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *NeurIPS*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. [Glue: A multi-task benchmark and analysis platform for natural language understanding](#). In *BlackboxNLP@EMNLP*.
- John S. White and Theresa A. O’Connell. 1993. [Evaluation of machine translation](#). In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

- Ka Wong, Praveen Paritosh, and Kurt Bollacker. 2022. [Are ground truth labels reproducible? an empirical study](#). In *ML Evaluation Standards Workshop at ICLR 2022*.
- 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.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). 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 7347–7360, Online. Association for Computational Linguistics.
- Amir Zeldes. 2022. Can we fix the scope for coreference? problems and solutions for benchmarks beyond ontonotes. *ArXiv*, abs/2112.09742.
- Shiyue Zhang, David Wan, and Mohit Bansal. 2022. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization.
- Ziqi Zhang. 2013. Named entity recognition : challenges in document annotation, gazetteer construction and disambiguation.
- Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. 2021. [Are larger pretrained language models uniformly better? comparing performance at the instance level](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, Online. Association for Computational Linguistics.
- Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daum'e, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing nlg evaluation: Evaluation practices, assumptions, and their implications. *ArXiv*, abs/2205.06828.

Appendix

Contents in Appendices:

- In Appendix [A](#), we disambiguate the definition of “benchmarks” and “tasks.”
- In Appendix [B](#), we provide examples of issues that threaten the validity of NLP benchmarks.
- In Appendix [C](#), we supply an extended discussion of how to distinguish between conceptualization and operationalization disagreements.
- In Appendix [D](#), we explain how we selected the tasks in our survey.
- In Appendix [E](#), we detail the guidance we provided to survey participants.
- In Appendix [F](#), we detail the quality control protocol we followed for our survey.
- In Appendix [G](#), we detail how we recruited and compensated survey participants.
- In Appendix [H](#), we provide the full script of our survey, including the consent form.
- In Appendix [I](#), we provide plots that summarize the responses to our survey questions.
- In Appendix [J](#), we provide our qualitative analyses of the model capabilities that papers claim NLP benchmarks assess.
- In Appendix [K](#), we offer additional examples of common essentially contested constructs in NLP.
- In Appendix [L](#), we offer additional examples of conceptualization disagreements for different NLP tasks.

A Disambiguating Benchmarks and Tasks

- **Benchmarks:** We refer to benchmarks for a specific NLP task rather than a benchmark suite (Dehghani et al., 2021). We further only consider benchmarks for evaluation and do not make assumptions about how models are trained.
- **Tasks:** In NLP, “task” has been used to refer to a “format” or “language-related skill” (Gardner et al., 2019). A format is typically a behavior specification, including a “way of posing a particular problem to a machine” along with what is expected as output (Bowman and Dahl, 2021). Consider summarization, which can vary in format: given a long passage of text, extractive summarization is about directly copying the most important spans from the passage, while abstractive summarization permits the generation of new sentences (Narayan et al., 2018). Some formats may be more amenable to certain real-world use cases or domains (e.g., clinical text, legal documents) than others. However, these various formats often capture a common language-related skill: capturing the main points from a longer passage of text using a few statements. Formats may capture the language-related skill underlying the task to varying degrees.

In this paper, we consider tasks that the NLP community has largely decided form a category (e.g., coreference resolution, question answering). These tasks can refer to a “format,” “language-related skill,” or both, and often have benchmarks specifically dedicated to them. Tasks may also overlap in “format” or “language-related skill.” For example, many consider the Winograd Schema Challenge to fall under the task of commonsense reasoning (Levesque et al., 2011), but the benchmark also assesses the ability of an NLP model to perform coreference resolution. Tasks also “can exist at varying granularities” (Liao et al., 2021).

B Issues that Threaten the Validity of NLP Benchmarks

Benchmark issue	Prior research
data noise and errors	(Schlegel et al., 2020; Blodgett et al., 2021; Northcutt et al., 2021; Dziri et al., 2022, <i>inter alia</i>)
superficial cues (e.g., annotation artifacts) in the data	(Schwartz et al., 2017; Gururangan et al., 2018; Poliak et al., 2018; Kaushik and Lipton, 2018; McCoy et al., 2019; Kavumba et al., 2019; Niven and Kao, 2019; Si et al., 2019; Ramponi and Tonelli, 2022; Friedman et al., 2022, <i>inter alia</i>)
inherent annotator disagreement	(Ovesdotter Alm, 2011; Plank et al., 2014; Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Basile et al., 2021; Davani et al., 2022; Wong et al., 2022; Sap et al., 2022; Kanclerz et al., 2022; Jiang and de Marneffe, 2022, <i>inter alia</i>)
poor linguistic diversity	(Hossain et al., 2020, 2022; Parmar et al., 2022; Selvam et al., 2022; Seshadri et al., 2022, <i>inter alia</i>)
task format unsuitability	(Kaushik and Lipton, 2018; Chen and Durrett, 2019, <i>inter alia</i>)
insufficiently fine-grained evaluation	(Lalor et al., 2016; Rodriguez et al., 2021; Zhong et al., 2021, <i>inter alia</i>)
poorly aligned metrics	(Wagstaff, 2012; Ethayarajh and Jurafsky, 2020; Marie et al., 2021; Deutsch et al., 2022; Moghe et al., 2022, <i>inter alia</i>)

Table 4: Prior research has surfaced issues with NLP benchmarks that call into question their validity as measurements of model performance.

C Extended Discussion of Conceptualization and Operationalization Disagreements

There often exists a blurry line between conceptualization and operationalization disagreements. This is because it can be difficult to ascertain that everyone in P_{B_τ} truly conceptualizes an aspect of τ in the same way. As such, every practitioner could conceivably conceptualize τ differently. However, [Palomaki et al. \(2018\)](#) argue that, while tasks are often “inherently subjective,” there exists “acceptable variation” in task conceptualization (e.g., in the case of y_τ , “there may be divergent annotations that are truly of unacceptable quality”).

Moreover, it is often challenging to impute how the creators of B_τ conceptualize τ solely from their stated goals (e.g., in the paper that proposes B_τ), due to incomplete statements of C_τ , M_τ , and $\neg M_\tau$; ambiguous specifications of y_τ ; and unclear explanations (if any) of how the creators understand E_τ ([Jiang and de Marneffe, 2022](#); [Tamkin et al., 2022](#)). However, distinguishing between conceptualization and operationalization disagreements is critical to contextualize progress in NLP. As such, we argue for benchmark documentation practices wherein the creators of B_τ clearly and comprehensively delineate their conceptualization of τ (§ 6).

D Survey Task Selection

- **Task selection:** We first sourced well-recognized tasks to potentially include in our survey from a variety of sources including the AllenNLP demo⁶ (Gardner et al., 2018), NLP-Progress⁷, Papers With Code⁸, and popular benchmark suites such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and GEM (Gehrmann et al., 2021). To keep our survey at a reasonable length, we shortlisted tasks that we perceive, based on a cursory literature review (described below), to fall on a spectrum with respect to factors (1) and (2). Our survey results suggest that our perceptions generally agree with those of the broader community.
- **Literature selection:** To analyze factors (1) and (2) for various tasks, we identified relevant literature by inputting the search queries “[TASK] survey” and “[TASK] challenges” into the Semantic Scholar search engine⁹. We considered the top 50 returned papers (sorted by “Relevance”), for each also considering the papers it cites and that cite it.

E Participant Guidance

To ground participants’ responses, we disambiguate “task” (Appendix A) and provide task definitions (Appendix H). However, because we are interested in participants’ perceptions, we purposely do not prescribe definitions for terms like “performance,” “progress,” and “state-of-the-art.” For each task, we also first ask participants to list associated benchmark datasets and metrics with which they are familiar to further ground their responses to questions about general benchmark quality. Moreover, for all questions where participants are asked to rate their perception, we provide a scale that ranges from 1 to 6 with articulations of what 1 and 6 mean in the context of the question. We do this to: a) capture the distribution of participants’ responses with sufficient granularity, b) impel participants to lean towards one side of the scale, and c) improve the consistency of how participants interpret answer choices. Finally, we do not specify language(s) for any tasks, including machine translation.

⁶<https://demo.allennlp.org/>

⁷<http://nlpprogress.com/>

⁸<https://paperswithcode.com/>

⁹<https://www.semanticscholar.org/>

F Survey Quality Control

Before releasing our survey, we piloted it with a few industry practitioners ($N = 4$) in order to identify potential problems with the clarity of our questions. We further provided participants with the opportunity to optionally justify their responses or indicate disagreement or a lack of clarity with any definitions or questions. We intentionally included a few free-response questions (e.g., description of their NLP work) to deter and remove spammers from our sample. After filtering out spammers, we ultimately had $N = 46$ responses.

G Survey Participant Recruitment and IRB

We recruited survey participants who identify as NLP practitioners by sharing our survey as a Microsoft form on Twitter, NLP-focused Slack workspaces, and mailing lists or Slack channels for artificial intelligence (AI) affinity groups like Queer in AI (Jethwani et al., 2022), Widening NLP¹⁰, and Women in Machine Learning¹¹. We additionally shared the survey at a tech company via internal NLP mailing lists and an internal communication platform. In all cases, we requested participants to share the survey with other relevant groups in order to perform snowball sampling (Naderifar et al., 2017). Survey participants could optionally enter a raffle to win one of ten \$50 Amazon gift cards or virtual visa cards (depending on location of residence¹²). Given that many participants receive no money, the raffle is not adequate payment in all countries of residence. We obtained informed consent (refer to Appendix H.1) from all survey participants, and the survey was IRB-approved.

¹⁰<https://www.winlp.org/>

¹¹<https://winlworkshop.org/>

¹²While practitioners from anywhere in the world were welcome to participate in our survey, participants from certain countries were not eligible to enter the raffle due to local laws or gift card supplier rules.

H Survey Questions and Responses

* Indicates required questions.

Perceptions of Conceptualization and Evaluation of Natural Language Processing (NLP) Tasks

We are interested in understanding how NLP practitioners and researchers perceive how well conceptualized and evaluated NLP tasks are. We hope that by understanding such perceptions, we will be able to better unpack validity issues with existing NLP benchmarks.

What is an NLP task? In NLP, “task” has been used to refer to a “format” or “language-related skill” (Bowman and Dahl, 2021). A format is typically a behavior specification, including a “way of posing a particular problem to a machine” along with what is expected as output (Gardner et al., 2019). Consider summarization, which can vary in format: given a long passage of text, extractive summarization is about directly copying the most important spans from the passage, while abstractive summarization permits the generation of new sentences. Some formats may be more amenable to certain real-world use cases or domains (e.g., clinical text, legal documents, etc.) than others. However, these various formats are often designed to capture a common language-related skill: capturing the main points from a longer passage of text using a few statements. Different formats may capture the language-related skill underlying the task to varying degrees.

If you have any questions, please feel free to contact us at: [REDACTED FOR ANONYMITY].

H.1 Consent Form

H.1.1 Introduction

Thank you for taking the time to consider volunteering in a [REDACTED FOR ANONYMITY] research project. This form explains what would happen if you join this research project. Please read it carefully and take as much time as you need. Ask the study team about anything that is not clear. You can ask questions about the study any time. Participation in this study is voluntary and you will not be penalized if you decide not to take part in the study or if you quit the study later.

Project Name: Perception of Formulation and Evaluation of Natural Language Processing (NLP) Tasks

Principal Investigator: [REDACTED FOR ANONYMITY]

Other Investigators: [REDACTED FOR ANONYMITY]

H.1.2 Purpose

The purpose of this project is to audit popular benchmark datasets that are commonly used to assess natural language models’ performance on a range of NLP tasks, with a focus on issues related to validity. To select a subset of NLP tasks and associated datasets for our study, we would like to run an online survey of individuals who are working on these tasks and/or are familiar with evaluating NLP models to collect their opinions on the ambiguity, simplicity, and popularity of NLP tasks, as well as of the perceived quality of benchmark datasets and metrics associated with each task.

H.1.3 Procedures

During this project, you will complete a 20-25-minute MS forms survey.

[REDACTED FOR ANONYMITY] may document and collect information about your participation through the answers you provide in the forms. No third parties will be involved in the transcription, processing, or analysis of the data. Approximately 100 participants will be involved in this study. You can copy or print this consent form for your own records, or you can email us at [REDACTED FOR ANONYMITY] for a copy of this form.

H.1.4 Study Information and Confidentiality

[REDACTED FOR ANONYMITY] is ultimately responsible for determining the purposes and uses of your study information.

How we use study information. The study information and other data collected during this project will be used primarily to perform research for purposes described in the introduction above. Such information and data, or the results of the research may eventually be used to develop and improve our commercial products, services or technologies.

Personal information we collect. During this project, if you choose to enter the sweepstakes and provide the required personal information, we will collect details such as first name, last name, email address, and country of residence.

How we store and share your study information. Your name and other personal information will not be on the study information we receive about you or from you; the personal information will be identified by a code (e.g., a key phrase you provide) and this personal information will be kept separate from your study information, in a secured, limited access location. We will use this code only to ensure that those signing up for the sweepstakes have answered the survey and are not spammers. If you chose not to enter the sweepstakes, no personally identifiable information will be collected about you.

Your study information will stored for a period of up to 18 months.

Aside from the researchers of this study, your study information may be shared with study team members outside of [REDACTED FOR ANONYMITY], applicable individuals within [REDACTED FOR ANONYMITY], but confidentiality will be maintained, as allowed by law.

How you can access and control your personal information. If you wish to review or copy any personal information you provided during the study, or if you want us to delete or correct any such data, email your request to the research team at: [REDACTED FOR ANONYMITY].

For additional information or concerns about how [REDACTED FOR ANONYMITY] handles your personal information, please see the [REDACTED FOR ANONYMITY] Privacy Statement ([REDACTED FOR ANONYMITY]).

H.1.5 Benefits and Risks

Benefits: There are no direct benefits to you that might reasonably be expected as a result of being in this study. We seek to audit popular benchmark datasets that are commonly used to assess NLP model performance on a range of NLP tasks, particularly focusing on issues related to validity. In doing so, we will bring light to issues with current evaluation practices in NLP and their implications for the claims made about NLP model performance. We hope to publish a paper and develop guidance or tools for how practitioners could audit their benchmark datasets.

Risks: The risks of participating in this study are no greater than those encountered in everyday life. To help reduce such risks, all identifiers will be removed from the survey responses. The primary contact and investigator have completed IRB training, including safe data handling practices. We will update survey respondents with research outcomes.

H.1.6 Future Use of Your Identifiable Information

Identifiers might be removed from your identifiable private information, and after such removal, the information could be used for future research studies or distributed to another investigator for future research studies without your (or your legally authorized representative's) additional informed consent.

H.1.7 Payment for Your Participation

At the end of the survey, you will be asked if you want to participate in a raffle for one of ten \$50 USD Amazon gift cards or equivalent virtual visa cards (depending on location of residence). Your odds of winning depend on the total number of participants but are no less than 1 in 10. Your data may be used to make new products, tests or findings. These may have value and may be developed and owned by [REDACTED FOR ANONYMITY] and/or others. If this happens, there are no plans to pay you. For Official Rules, see the PDF [REDACTED FOR ANONYMITY].

H.1.8 Participation

Taking part in research is always a choice. If you decide to be in the study, you can change your mind at any time without affecting any rights including payment to which you would otherwise be entitled. If you decide to withdraw, you should contact the person in charge of this study, and also inform that person if you would like your personal information removed as well.

[REDACTED FOR ANONYMITY] or the person in charge of this study may discontinue the study or your individual participation in the study at any time without your consent for reasons including:

- your failure to follow directions
- it is discovered that you do not meet study requirements
- it is in your best interest medically
- the study is canceled
- administrative reasons

If you leave the study, the study staff will still be able to use your information that they have already collected, however, you have the right to ask for it to be removed when you leave. Significant new findings that develop during the course of this study that might impact your willingness to be in this study will be given to you.

H.1.9 Contact Information

Should you have any questions concerning this project, or if you are injured as a result of being in this study, please contact: [REDACTED FOR ANONYMITY].

Should you have any questions about your rights as a research subject, please contact the [REDACTED FOR ANONYMITY].

H.1.10 Consent

By completing this form, you confirm that the study was explained to you, you had a chance to ask questions before beginning the study, and all your questions were answered satisfactorily. At any time, you may ask other questions. By completing this form, you voluntarily consent to participate, and you do not give up any legal rights you have as a study participant.

Please confirm your consent by completing the bottom of this form. If you would like to keep a copy of this form, please print or save one. On behalf of [REDACTED FOR ANONYMITY], we thank you for your contribution and look forward to your research session.

[Q1] Do you understand and consent to these terms? *

- Yes
- No [if selected, survey branches to final section]

H.2 Background

[Q2] Do you have any experience with NLP tasks? *

- Yes
- No

[Q3] Briefly describe the type of NLP work that you do. *

- Open text field

[Q4] Please select all the options that apply to you. *

- I work on deployed systems
- I am an industry practitioner (not researcher)
- I am an industry researcher
- I am an academic researcher

H.3 Perceived Performance

[Q5] In general, how well do you think current state-of-the-art NLP models perform on the following tasks? Please select “I don’t know” if you have never heard of the task or have little to no knowledge about it. *

6 (high performance) means that you think current state-of-the-art NLP models tend to perform very well on this task, with little to no area for improvement. In contrast, 1 (low performance) means that current state-of-the-art models perform poorly on this task, including because the task is new or the task has been neglected by the community.

1 2 3 4 5 6 I don’t know

-

- Sentiment Analysis
- Natural Language Inference
- Question Answering
- Coreference Resolution
- Summarization
- Named-Entity Recognition
- Dependency Parsing
- Machine Translation

H.4 NLP Task: Sentiment Analysis

Given some input text, a model must correctly identify opinions, sentiments, and subjectivity in the text (read more: https://en.wikipedia.org/wiki/Sentiment_analysis) (Pang and Lee, 2007).

Reminder: “Task” has been used to refer to a “format” or “language-related skill” (Bowman and Dahl, 2021). A format is typically a behavior specification, including a “way of posing a particular problem to a machine” along with what is expected as output (Gardner et al., 2019). While NLP tasks can sometimes vary in format or domain, these various formats are often designed to capture a common language-related skill (e.g., summarization tasks try to capture the key points in a long passage of text using a few statements). Different formats may capture the language-related skill underlying the task to varying degrees.

[Q6] Are you familiar with this NLP task (including associated benchmark datasets and metrics)? *

- Yes, I am an expert (e.g., I have developed, deployed, researched, or evaluated NLP models on this task)
- Yes, but I only have passing knowledge (e.g., I only have read, studied, or heard about this task)
- No [if selected, survey branches to next section]

[Q7] **Task definition:** How well defined or conceptualized do you think this task is? *

6 (well defined) means that the task has an objective that is clearly and consistently articulated and

understood by the NLP community. In contrast, 1 (poorly defined) means that the task has an objective that is understood differently from person to person in the community.

1 2 3 4 5 6

[Q8] **Task instantiation:** In general, how well do you think common formats of this task capture the underlying language-related skill? *

6 (captures the skill well) means that common formats of this task perfectly capture the underlying language-related skill, while 1 (captures the skill poorly) means that common formats of this task do not capture the underlying language-related skill at all.

1 2 3 4 5 6

[Q9] Write in any performance metrics for this task that you have experience with, if any. If none, please write "N/A." *

- Open text field

[Q10] **Metrics quality:** In general, how well do you think common metrics (considering a broad range of metrics) capture NLP models' performance on this task? *

6 (captures performance well) means that metrics generally capture everything about performance on this task that we want it to capture, without capturing extraneous information. In contrast, 1 (captures performance poorly) means that metrics generally do not capture any valuable information about task performance or is highly influenced by extraneous signals.

1 2 3 4 5 6

[Q11] Write in any benchmark datasets for this task that you have experience using, if any. If none, please write "N/A." *

- Open text field

[Q12] **Benchmark datasets quality:** In general, how would you assess the quality of benchmark datasets that are commonly used to evaluate NLP models on this task? *

6 (high dataset quality) means that the datasets generally are free of errors, and correctly and consistently capture the language-related skill underlying the task. In contrast, 1 (low dataset quality) means that the datasets generally contain significant errors or fail to capture the language-related skill underlying the task correctly and consistently.

1 2 3 4 5 6

[Q13] **Current progress:** How close do you think current state-of-the-art NLP models are to learning the language-related skill underlying this task? * 6 (close) means that you think current state-of-the-art NLP models have successfully learned the language-related skill underlying this task. In contrast, 1 (not close) means that you think current state-of-the-art NLP models are still far from learning this skill.

1 2 3 4 5 6

[Q14] **Potential progress:** How likely do you think NLP models are to ever learning the language-related skill underlying this task?

6 (highly likely) means that you think current state-of-the-art NLP models have learned or will surely learn the language-related skill underlying this task, while 1 (highly unlikely) means that you think NLP models will likely never learn this skill.

1 2 3 4 5 6

[Q15] Do you have additional thoughts in response to the definitions and questions for this task?

This can include justifications of your responses or a lack of clarity on any definitions or questions. For example, do you agree with the task definition?

- Open text field

H.5 NLP Task: Natural Language Inference

Given a pair of input sentences, a model must correctly determine if the sentences satisfy a certain semantic relationship (e.g., textual entailment) (read more: <https://paperswithcode.com/task/natural-language-inference>) (Storks et al., 2019).

Rest of section is same as Appendix H.4.

H.6 NLP Task: Question Answering

Given some knowledge source (e.g., a passage, image, knowledge base), a model must correctly answer given questions (read more: https://en.wikipedia.org/wiki/Question_answering) (Gardner et al., 2019).

Rest of section is same as Appendix H.4.

H.7 NLP Task: Coreference Resolution

Given some input text, a model must correctly identify expressions that refer to the same entity (read more: https://en.wikipedia.org/wiki/Coreference#Coreference_resolution) (Pradhan et al., 2011).

Rest of section is same as Appendix H.4.

H.8 NLP Task: Summarization

Given some input text, a model must output a shorter summary that preserves key information from the input text (read more: https://en.wikipedia.org/wiki/Automatic_summarization) (Allahyari et al., 2017).

Rest of section is same as Appendix H.4.

H.9 NLP Task: Named-Entity Recognition

Given some input text, a model must correctly identify named entities (people, locations, organizations) in the text (read more: https://en.wikipedia.org/wiki/Named-entity_recognition) (Tjong Kim Sang and De Meulder, 2003).

Rest of section is same as Appendix H.4.

H.10 NLP Task: Dependency Parsing

Given some input text, a model must correctly identify head words in the text and the dependent words which modify those heads (read more: <https://paperswithcode.com/task/dependency-parsing>) (Nivre et al., 2007).

Rest of section is same as Appendix H.4.

H.11 NLP Task: Machine Translation

Given some input content (e.g., text, video), a model must correctly translate the content from the source language to a target language (read more: https://en.wikipedia.org/wiki/Machine_translation) (White and O'Connell, 1993; Yin et al., 2021).

Rest of section is same as Appendix H.4.

H.12 Raffle Entry

[Q86] [OPTIONAL] If you would like to enter the raffle drawing for one of the ten \$50 Amazon gift cards or equivalent virtual visa cards (depending on location of residence), for anonymity purposes, after submitting this form you will be provided with a link to another form to fill in your email address and enter the raffle. For this, **please also write down a key phrase here, which you will also be asked to re-enter on the raffle form.** We will only use this key phrase to validate that the raffle participants have completed the survey. Please don't use a key phrase that is associated with any accounts.

- Open text field

H.13 Feedback

[Q87] Do you have any comments or feedback on the questions in this survey?

Please be mindful not to bring up any identifying or sensitive information about yourself or third-parties.

- Open text field

I Comprehensive Survey Results

I.1 Perceived performance

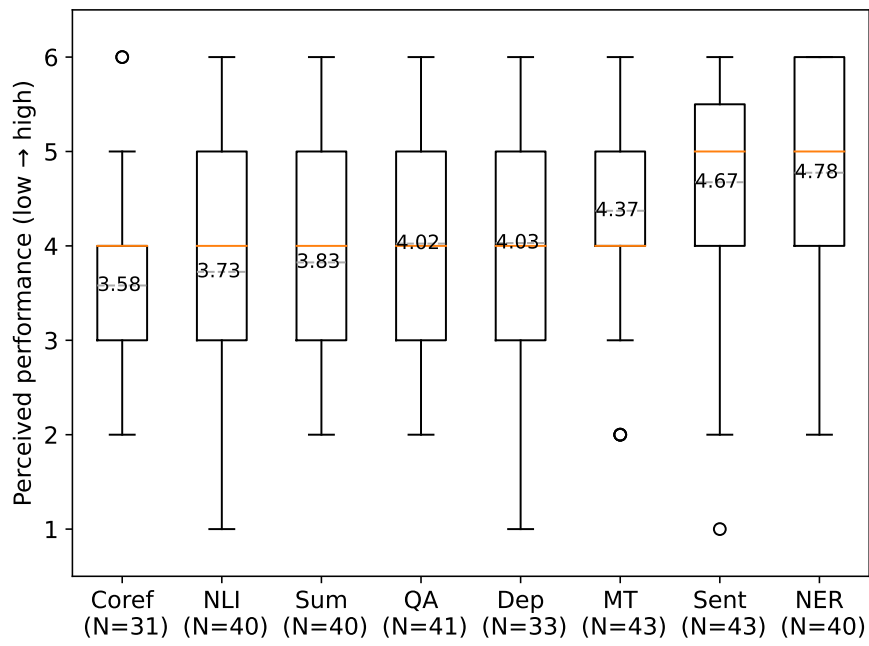
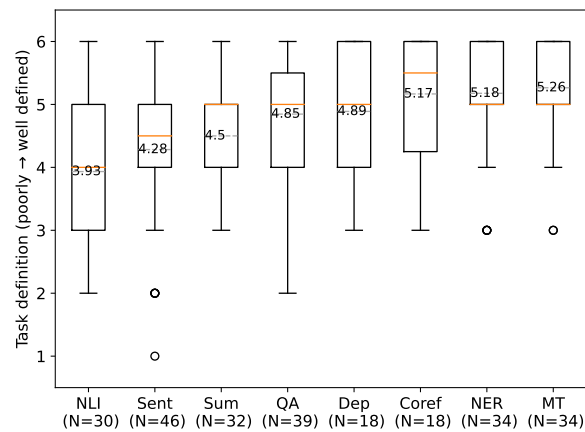
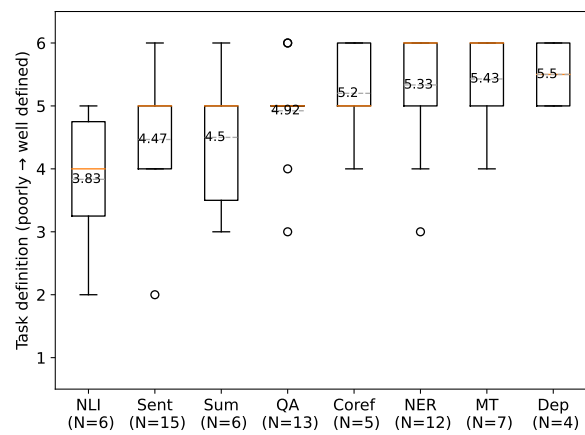


Figure 6: Perceived performance for all tasks.

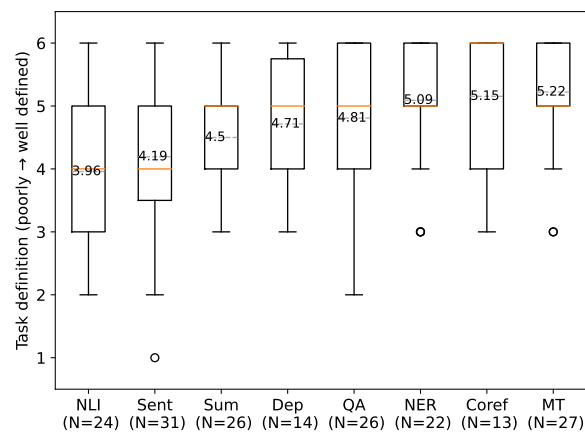
I.2 Task definition



(a) Perceived clarity and consistency of task definition among all responding practitioners for each task.



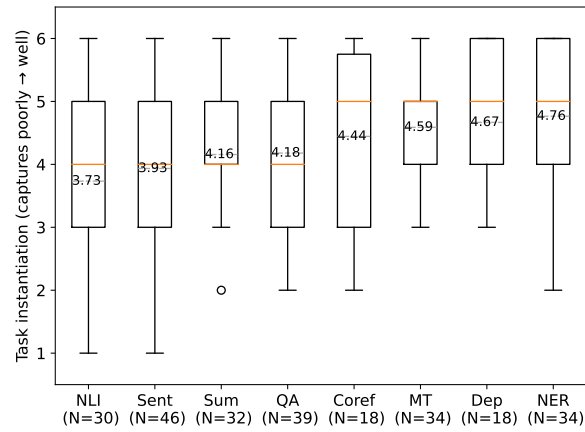
(b) Perceived clarity and consistency of task definition among survey participants who consider themselves an “expert” at each task.



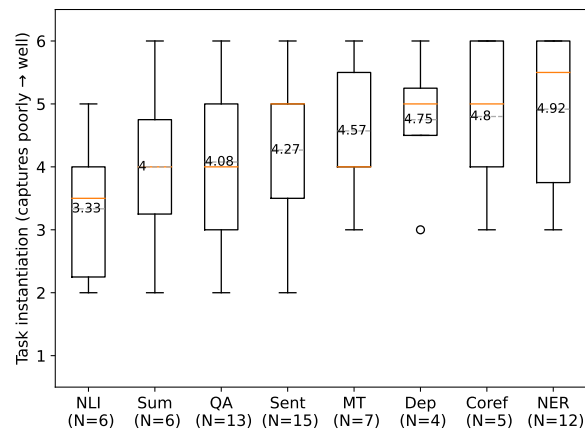
(c) Perceived clarity and consistency of task definition among survey participants who consider themselves to have “passing knowledge” about each task.

Figure 7

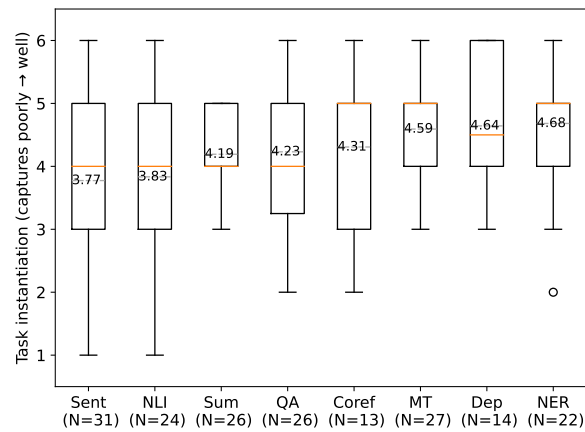
I.3 Task instantiation



(a) Perceived task instantiation quality among all responding practitioners for each task.



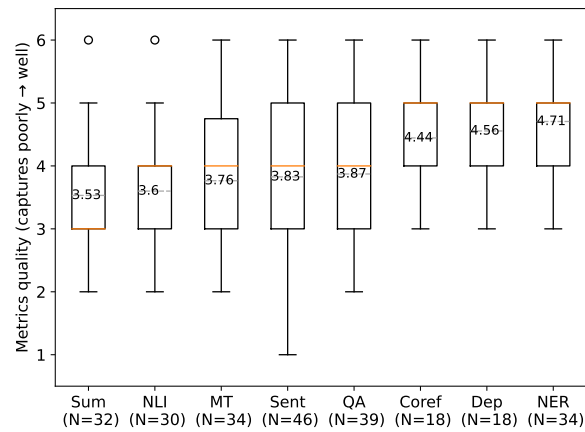
(b) Perceived task instantiation quality among survey participants who consider themselves an “expert” at each task.



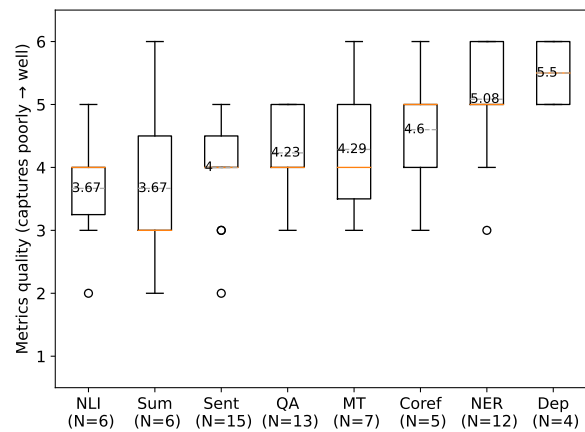
(c) Perceived task instantiation quality among survey participants who consider themselves to have “passing knowledge” about each task.

Figure 8

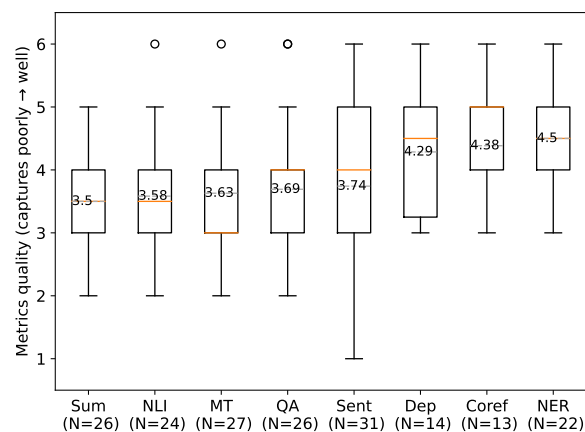
I.4 Metrics quality



(a) Perceived metrics quality among all responding practitioners for each task.



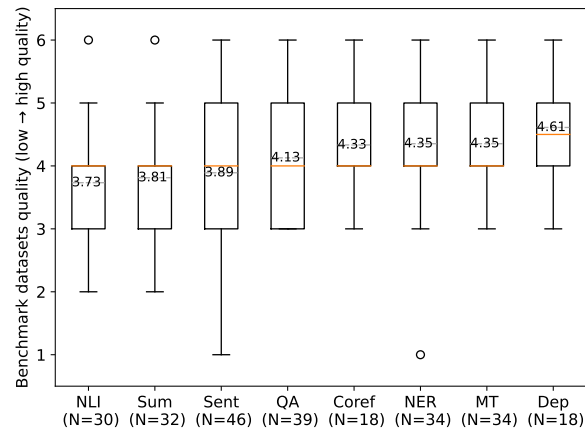
(b) Perceived metrics quality among survey participants who consider themselves an “expert” at each task.



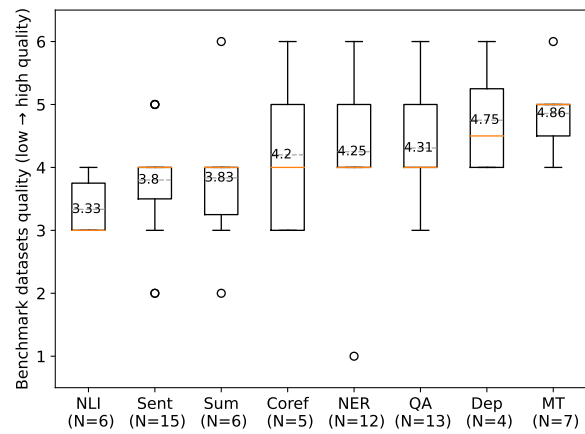
(c) Perceived metrics quality among survey participants who consider themselves to have “passing knowledge” about each task.

Figure 9

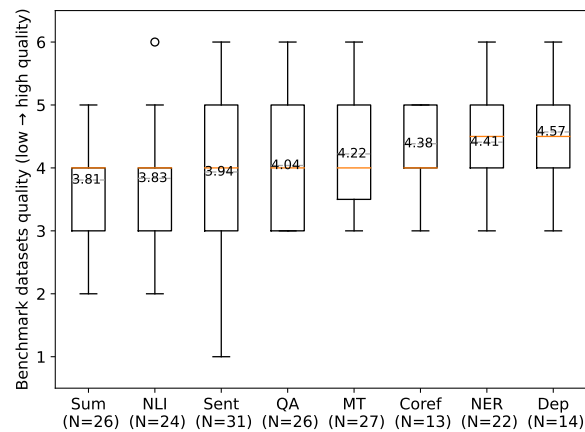
I.5 Benchmark datasets quality



(a) Perceived benchmark datasets quality among all responding practitioners for each task.



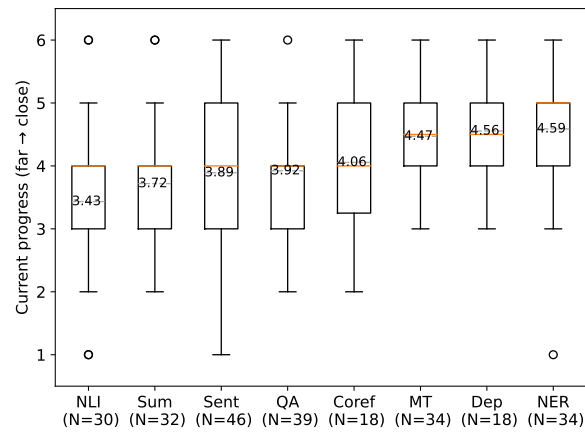
(b) Perceived benchmark datasets quality among survey participants who consider themselves an “expert” at each task.



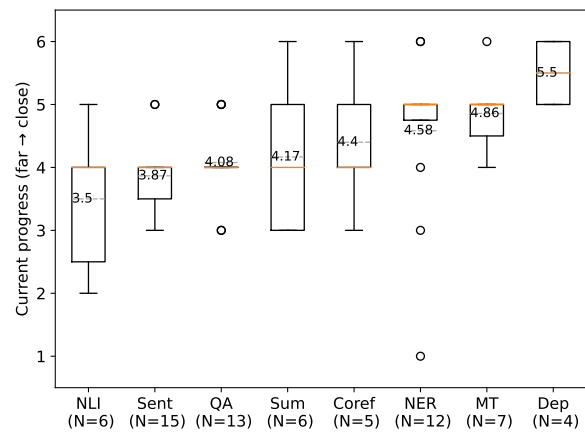
(c) Perceived benchmark datasets quality among survey participants who consider themselves to have “passing knowledge” about each task.

Figure 10

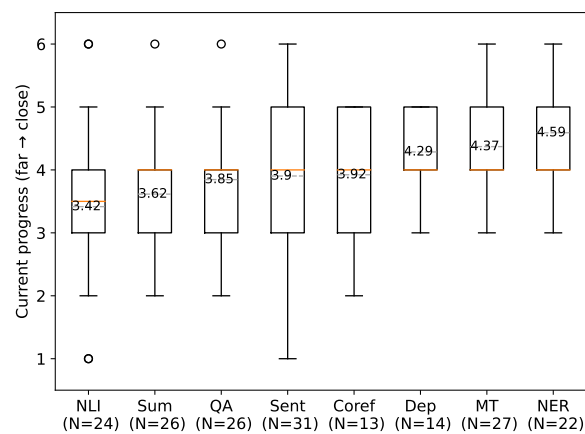
I.6 Current progress



(a) Perceived current progress among all responding practitioners for each task.



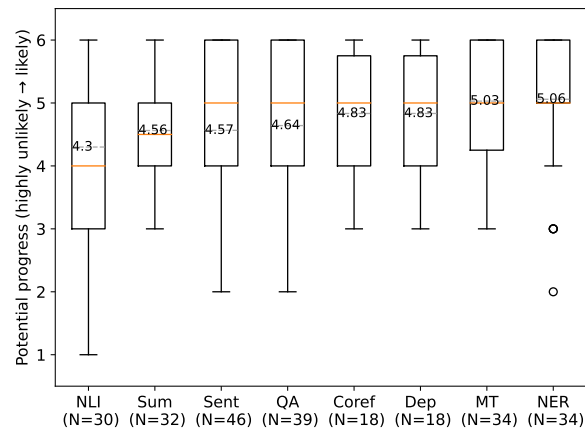
(b) Perceived current progress among survey participants who consider themselves an “expert” at each task.



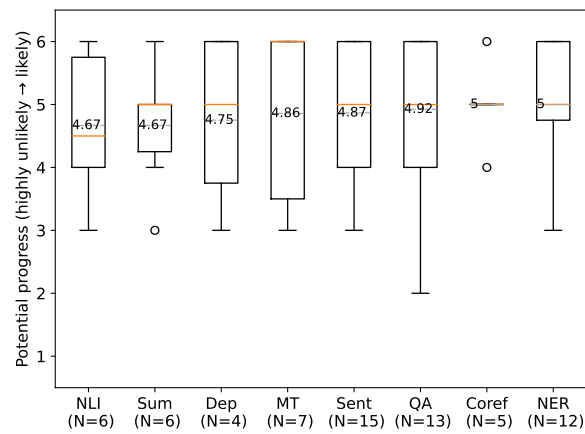
(c) Perceived current progress among survey participants who consider themselves to have “passing knowledge” about each task.

Figure 11

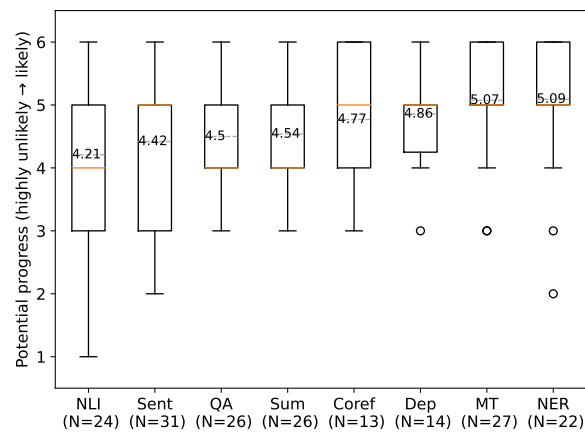
I.7 Potential progress



(a) Perceived potential progress among all responding practitioners for each task.



(b) Perceived potential progress among survey participants who consider themselves an “expert” at each task.



(c) Perceived potential progress among survey participants who consider themselves to have “passing knowledge” about each task.

Figure 12

J Model Capabilities Purportedly Assessed by Benchmarks

J.1 Benchmark Review Protocol

Our use of all the benchmarks below was reviewed by an IRB. The IRB deemed that our analysis of the benchmarks was not prohibited by their license nor the terms of use of the benchmark data sources. The IRB also confirmed that the benchmarks do not contain any information that names or uniquely identifies individual people or offensive content.

J.2 Benchmark Analysis

Capability	Benchmarks
Capture meaning	SST (Socher et al., 2013) : “capture the meaning of longer phrases” IMDb (Maas et al., 2011) : “capture both semantic and sentiment similarities among words,” “learns [...] nuanced sentiment information”
Outperform humans	Cornell movie reviews (Pang et al., 2002) : “outperform human-produced baselines”
Handle linguistic phenomena	SST (Socher et al., 2013) : “presents new challenges for sentiment compositionality,” “capture the effects of negation,” “capture complex linguistic phenomena,” “learn that sentiment of phrases following the contrastive conjunction ‘but’ dominates,” “from a linguistic or cognitive standpoint, ignoring word order in the treatment of a semantic task is not plausible” Cornell movie reviews (Pang et al., 2002) : handle “thwarted-expectations rhetorical device”
Handle phenomena in real-world data	SST (Socher et al., 2013) : “there is a need to better capture sentiment from short comments, such as Twitter data”

Table 5: Model capabilities that SENT benchmarks are intended to measure.

Capability	Benchmarks
Understand language	<p>MNLI (Williams et al., 2018): “evaluation of methods for sentence understanding”</p> <p>SNLI (Bowman et al., 2015): “understanding entailment and contradiction is fundamental to understanding natural language,” “models’ attempts to shortcut this kind of inference through lexical cues can lead them astray”</p> <p>XNLI (Conneau et al., 2018): “[test bed for] crosslingual language understanding”</p>
Handle linguistic phenomena	<p>MNLI (Williams et al., 2018): “handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lexical and syntactic ambiguity”</p> <p>SNLI (Bowman et al., 2015): “NLI is an ideal testing ground for theories of semantic representation,” “additional attention to compositional semantics would pay off”</p> <p>RTE (Giampiccolo et al., 2008): “finding equivalences and similarities at lexical, syntactic and semantic levels”</p>
Handle various domains	<p>MNLI (Williams et al., 2018): “corpus [...] meant to approximate full diversity of ways in which modern standard American English is used,” “represents both written and spoken speech in a wide range of styles, degrees of formality, and topics,” “benchmark for cross-genre domain adaptation”</p> <p>SNLI (Bowman et al., 2015): “evaluation of domain-general approaches”</p>
Possess benchmark-external knowledge	<p>MNLI (Williams et al., 2018): “using only [...] what you know about the world”</p> <p>SNLI (Bowman et al., 2015): “data collected draws fairly extensively on commonsense knowledge”</p>
Aid in other NLP tasks	<p>XNLI (Conneau et al., 2018): “evaluation of pretrained general-purpose language universal sentence encoders”</p> <p>RTE (Giampiccolo et al., 2008): “captures major semantic inference needs across many natural language processing applications, such as Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and multi-document summarization (SUM)”</p>

Table 6: Model capabilities that NLI benchmarks are intended to measure.

Capability	Benchmarks
Understand language	<p>SQuAD (Rajpurkar et al., 2016): “requiring both understanding of natural language and knowledge about the world,” “towards the end goal of natural language understanding”</p> <p>HotpotQA (Yang et al., 2018): “test their understanding of both language and common concepts such as numerical magnitude”</p>
Handle linguistic phenomena	<p>SQuAD (Rajpurkar et al., 2016): “all examples have some sort of lexical or syntactic divergence between the question and the answer in the passage”</p> <p>TriviaQA (Joshi et al., 2017): “has relatively complex, compositional questions,” “has considerable syntactic and lexical variability”</p>
Reason over a context	<p>SQuAD (Rajpurkar et al., 2016): “multiple sentence reasoning”</p> <p>HotpotQA (Yang et al., 2018): “questions require finding and reasoning over multiple supporting documents to answer,” “test the reasoning ability of intelligent systems,” numerous “types of multi-hop reasoning required to answer questions”</p> <p>TriviaQA (Joshi et al., 2017): “requires more cross sentence reasoning to find answers”</p>
Possess benchmark-external knowledge	<p>SQuAD (Rajpurkar et al., 2016): “requires both understanding of natural language and knowledge about the world”</p> <p>HotpotQA (Yang et al., 2018): “the questions are [...] not constrained to any pre-existing knowledge bases or knowledge schemas”</p> <p>TriviaQA (Joshi et al., 2017): “17% of the examples required some form of world knowledge”</p>
Handle various domains	<p>HotpotQA (Yang et al., 2018): “the questions are [...] not constrained to any pre-existing knowledge bases or knowledge schemas,” “our dataset covers a diverse variety of questions centered around entities, locations, events, dates, and numbers, as well as yes/no questions directed at comparing two entities”</p> <p>TriviaQA (Yang et al., 2018): models “should be able to deal with large amount of text from various sources such as news articles, encyclopedic entries and blog articles”</p>
Handle phenomena in real-world data	<p>SQuAD (Rajpurkar et al., 2016): “existing datasets for RC [...] that are large [...] are semi-synthetic and do not share the same characteristics as explicit reading comprehension questions”</p> <p>TriviaQA (Joshi et al., 2017): “first dataset where full-sentence questions are authored organically”</p>
be on par with humans	<p>SQuAD (Rajpurkar et al., 2016): “these results are still well behind human performance”</p> <p>HotpotQA (Yang et al., 2018): “if the baseline model were provided with the correct supporting paragraphs to begin with, it achieves parity with the crowd worker in finding supporting facts”</p> <p>TriviaQA (Joshi et al., 2017): “neither approach comes close to human performance”</p>

Table 7: Model capabilities that QA benchmarks are intended to measure.

Capability	Benchmarks
Think	Winograd Schema Challenge (Levesque et al., 2011): “thinking is required to get a correct answer with high probability”
Handle linguistic phenomena	Winograd Schema Challenge (Levesque et al., 2011): “question involves determining the referent of the pronoun or possessive adjective”
be on par with humans	Winograd Schema Challenge (Levesque et al., 2011): “required to achieve human-level accuracy in choosing the correct disambiguation”
Possess benchmark-external knowledge	Winograd Schema Challenge (Levesque et al., 2011): “you need to have background knowledge that is not expressed in the words of the sentence”
Aid in other NLP tasks	Winograd Schema Challenge (Levesque et al., 2011): “it is sometimes possible to find sentences in natural text that can easily be turned into Winograd schemas”
Handle various genres	OntoNotes (Hovy et al., 2006): “annotation will cover [...] multiple genres (newswire, broadcast news, news groups, weblogs, etc.), to create a resource that is broadly applicable”

Table 8: Model capabilities that COREF benchmarks are intended to measure.

Capability	Benchmarks
Understand language	XSum (Narayan et al., 2018): “posing several challenges relating to understanding (i.e., identifying important content),” “high-level document knowledge in terms of topics and long-range dependencies is critical for recognizing pertinent content and generating informative summaries” CNN/DM (Nallapati et al., 2016): “capturing the ‘meaning’ of complex sentences”
Generate novel language	XSum (Narayan et al., 2018): “generation (i.e., aggregating and rewording the identified content into a summary),” “there are [...] novel [<i>n</i> -]grams in the XSum reference summaries”
Handle linguistic phenomena	XSum (Narayan et al., 2018): “displays multiple levels of abstraction including paraphrasing, fusion, synthesis, and inference”
Handle various domains	XSum (Narayan et al., 2018): “collected 226,711 Wayback archived BBC articles ranging over almost a decade (2010 to 2017) and covering a wide variety of domains”
Possess benchmark-external knowledge	CNN/DM (Nallapati et al., 2016): “potentially using vocabulary unseen in the source document”

Table 9: Model capabilities that SUM benchmarks are intended to measure.

Capability	Benchmarks
Handle various domains	OntoNotes (Hovy et al., 2006) : “annotation will cover [...] multiple genres (newswire, broadcast news, news groups, weblogs, etc.), to create a resource that is broadly applicable”
Handle various languages	CoNLL-2003 (Sang and Meulder, 2003) : “language-independent named entity recognition”
Aid in real-world applications of task	CoNLL-2003 (Sang and Meulder, 2003) : “named entity recognition is an important task of information extraction systems”
Possess benchmark-external knowledge	CoNLL-2003 (Sang and Meulder, 2003) : “interested in approaches that made use of resources other than the supplied training data”

Table 10: Model capabilities that NER benchmarks are intended to measure.

Capability	Benchmarks
Understand language	Penn Treebank (Marcus et al., 1993) : “progress can be made in both text understanding and spoken language understanding”
Handle phenomena in real-world data	Penn Treebank (Marcus et al., 1993) : “in naturally occurring unconstrained materials”
Handle linguistic phenomena	Penn Treebank (Marcus et al., 1993) : “evaluation and comparison of the adequacy of parsing models”
Handle various languages	Universal Dependencies (Nivre et al., 2016) : “facilitate multilingual natural language processing”
Handle various genres	Universal Dependencies (Nivre et al., 2016) : “most treebanks are constituted of different genres”

Table 11: Model capabilities that DEP benchmarks are intended to measure.

Capability	Benchmarks
Handle various languages	Europarl (Koehn, 2005) : “parallel text in 11 languages” WMT-2007 (Callison-Burch et al., 2007) : “translating French, German, Spanish, and Czech to English and back”
Handle linguistic phenomena	Europarl (Koehn, 2005) : “reason for the difficulty of translating into a language is morphological richness”
Generate fluent and adequate language	WMT-2007 (Callison-Burch et al., 2007) : “fluency and adequacy” OpenMT¹³ : “goal is for the output to be an adequate and fluent translation of the original”

Table 12: Model capabilities that MT benchmarks are intended to measure.

¹³<https://www.nist.gov/itl/iad/mig/open-machine-translation-evaluation>

K Additional Examples of Essentially Contested Constructs

- **Understanding language:** Practitioners often do not explain how they conceptualize language understanding, nor do they address disagreement about whether models are capable of understanding language or language can be understood from text alone (Michael et al., 2022).
- **Handling real-world phenomena:** Practitioners often leave the definition of “real-world” open-ended, despite foregrounding certain domains in their conceptualization of “real-world,” and do not address contention around the feasibility of capturing “everything in the whole wide world” (Raji et al., 2021).

L Additional Disagreements in NLP Task Conceptualization

Task	Disagreement in conceptualization?	Conceptualization disagreement examples
SENT	C_τ : yes (Table 5)	SST, IMDb, Cornell movie reviews datasets operationalize sentiment with single gold label (Socher et al., 2013; Maas et al., 2011; Pang et al., 2002)
	y_τ : yes; lack of “real ‘ground truth’” due to differing perceptions of sentiment (Ovesdotter Alm, 2011; Hagerer et al., 2021)	
	E_τ : yes; {sentiment, capture meaning, outperform humans} $\subset E_\tau$ (Table 5)	
NER	C_τ : yes (Table 10)	OntoNotes, CoNLL-2003 datasets operationalize type of ambiguous entities with single gold label (Hovy et al., 2006; Sang and Meulder, 2003)
	y_τ : yes; inherent semantic ambiguity induces disagreement about y_τ (Zhang, 2013)	
	E_τ : yes; { possess benchmark-external knowledge } $\subset E_\tau$ (Table 10)	
DEP	C_τ : yes (Table 11)	benchmarks make use of inconsistent annotation formats due to differing conceptualizations of parsing (Dredze et al., 2007; Nivre et al., 2016)
	y_τ : yes; syntactic ambiguity (Ackerman, 2015; Keith et al., 2018) and systematic disagreement about parts of speech (Plank et al., 2014) yield differing y_τ	
	E_τ : yes; { understand language, handle phenomena in real-world data } $\subset E_\tau$ (Table 11)	
MT	C_τ : yes (Table 12)	Europarl, WMT-2007 datasets contain single reference translations (Koehn, 2005; Callison-Burch et al., 2007)
	y_τ : adequacy of translations in y_τ is subjective (White and O’Connell, 1993); can be unclear how to translate lexical and syntactic ambiguity in source language (Pericliev, 1984; Baker et al., 1994), or translate from language without to with grammatical gender (Gonen and Webster, 2020)	
	E_τ : yes; { fluency, adequacy } $\subset E_\tau$ (Table 12)	

Table 13: Additional disagreements in the conceptualization of NLP tasks and examples of resultant conceptualization disagreements.

Example	<i>Sentences</i>
<i>Premise</i>	Isn't a woman's body her most personal property?
<i>Hypothesis</i>	Isn't a woman's body sacred property?
<i>Annotator labels</i>	E, E, E, N, N
<i>Gold label</i>	Entailment
Issues	<i>Description</i>
<i>Conceptualization</i>	<ul style="list-style-type: none"> • unclear whether a question can entail or contradict any hypothesis¹⁴ • unclear whether any premise can entail or contradict a question

Figure 13: Example test instance from the MNLi benchmark (Williams et al., 2018), accompanied by issues with the conceptualization of MNLi that the instance reflects. The NLI task captures whether the premise entails (E), contradicts (C), or is neutral (N) with respect to the hypothesis.

¹⁴Although not stated in the paper, according to Williams et al. (2018), a question entails the set of its possible answers.