Completed Worksheet: SuperGLUE Evidence-Centered Benchmark Design

Evidence-Centered Benchmark Design (ECBD) is a framework that formalizes the benchmark design process. It requires first specifying the **intended use** of the benchmark (including specifying the objects of evaluation). The process is then broken down into five modules:

- **Capability module:** capabilities that the benchmark aims to measure.
- **Content module:** pool of test items that draw out responses from the objects.
- **Adaptation module:** adapting or instructing the objects to complete the tasks.
- **Assembly module:** selecting from the pool of test items to build the set used for evaluation.
- **Evidence module:** extracting and accumulating evidence about the capabilities of interest from responses produced by the objects.

This worksheet provides guidance on how to create a new benchmark or analyze an existing benchmark following ECBD. It can be completed from different perspectives: as the creator of a new benchmark, as the custodian or the user of an existing benchmark, or as a third-party analyzing benchmarks, etc. Each module contains three questions:

- **Describe:** what design decisions did the benchmark creators make for this module?
- **Justify:** why did the benchmark creators make these decisions? This involves forming a hypothesis that the decisions allow the module to accomplish its role in the process of gathering necessary capability evidence.
- **Evidence:** what validity evidence do the benchmark creators have to support the above hypothesis? In other words, what shows that the module indeed accomplishes its role?

This worksheet is not a checklist, and it is not required to answer each question perfectly. These questions are meant to encourage reflection and validation of benchmark design decisions, as well as to guide benchmark documentation.

Table of Contents

Benchmark Name and Reference(s)

The references are the source of information used to complete this worksheet. For example, a third-party analyzing an existing benchmark may choose to use the academic publication introducing said benchmark as the source of information. Other sources of information could be blogposts, official websites, or code repositories accompanying the benchmark.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In
Advances in Neural Information Processing Systems (NeurIPS)

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. GLUE: A multitask benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations (ICLR)

From here onwards, passages from the GLUE paper will be cited as (GLUE, p.X) while passages from the SuperGLUE paper will be cited with only the page number: (p.X).

URL: https://dl.acm.org/doi/10.5555/3454287.3454581

Additionally to the leaderboard evaluation, SuperGLUE includes diagnostic datasets ("tools for model analysis" p. 6). For the purpose of this analysis, we focus on the leaderboard evaluation.

Who is filing the worksheet:

From what perspective is this worksheet completed? In other words, what is the relation between the person(s) completing this worksheet and the benchmark that is the focus of this worksheet?

Third-party: Yu Lu Liu, Q. Vera Liao, Alexandra Olteanu, Jackie Chi Kit Cheung

Intended Use

The validity of a benchmark concerns whether it can be used as intended. It is therefore crucial to first clearly establish the intended use of a benchmark before analyzing it or creating it.

Q1 - Who/What are the intended objects of evaluation? Elaboration on the objects of evaluation (e.g., their assumed capabilities, demographic information for human objects of evaluation, etc.) helps us better understand whether the benchmark is suitable for all intended objects of evaluation.

The main evaluatees are ***“general-purpose language understanding technologies for English.”*** (p.2) Additionally, we reason from the inclusion of *“human performance estimates for all benchmark tasks"* (p.2) that the intended evaluatees of the benchmark also includes **humans** (human performance estimated by the authors involved trained crowdworker annotators; see p.9).

Restrictions on said technologies: *“Any system or method that can produce predictions for the SuperGLUE tasks is eligible for submission to the leaderboard, subject to the data-use and submission frequency policies [...]. There are no restrictions on the type of methods that may be used, and there is no requirement that any form of parameter sharing or shared initialization be used across the tasks in the benchmark”* (p.7)

Q2 - What is the intended use of the benchmark? Who are the intended users of the benchmark? Benchmark results aim to provide insights about the objects of evaluation: how are users meant to use these insights?

“SuperGLUE has the same high-level motivation as GLUE: to provide a simple, hard-to-game measure of ***progress*** *toward general-purpose language understanding technologies for English.”* (p.2)

GLUE has been used for:

- Guiding “research towards general-purpose language understanding technologies”
- “Afford straightforward comparison between [...] task-agnostic transfer learning techniques” and “exhibiting the transfer-learning potential of approaches like OpenAI GPT and BERT”.

So, SuperGLUE inherits these intended uses.

Although the intended users of the benchmark are not explicitly mentioned, the uses described above seem to imply that the users are researchers and developers of general-purpose natural language understanding technologies.

The Capability Module

The capability module specifies the *capabilities* that the benchmark aims to evaluate. The term “capability” refers to a construct (e.g., quality criteria, skill, etc.) that the objects of evaluation are thought to exhibit or possess. Capabilities often cannot be directly observed or directly measured, thus requiring the benchmark to indirectly measure them by gathering necessary evidence about said capabilities.

Q3 - DESCRIBE: i) What are the capabilities of interest? ii) How is each one defined, and under what context is each one defined?

The capability of interest is “**General(-purpose) language understanding**” (GLU), which seems to mean the ability *“to learn to execute a range of different linguistic tasks in different domains”*, inherited from GLUE (GLUE, p.1)

Note that, when GLUE was introduced, “most NLU models above the word level are designed for a specific task and struggle with out-of-domain data” which the authors contrasted to the human ability to understand language which is “general, flexible, and robust” (GLUE, p.1).

Although not explicitly stated so, GLUE seems to be decomposed into sub-capabilities such as “causal reasoning”, “commonsense reasoning”, etc. (See Q6, where we **boldface** potential sub-capabilities in SuperGLUE’s description of test items)

Additional recommended questions to consider so to further clarify and contextualize the definitions (in benchmark analysis: as presented by the benchmark)

- How does the definition used by the benchmark differ from other existing definitions of this capability?

  GLUE mentioned SentEval, decaNLP and other prior benchmarks for “general NLU systems”. However, when mentioning these prior works, SuperGLUE did not elaborate on whether they provide an explicit definition or conceptualization of NLU.

  Note that SuperGLUE also mentioned these benchmarks, with no additional information on them.

- How does this capability differ from other similarly defined capabilities?

  Nothing mentioned.

Q4 - JUSTIFY: How are the capabilities of interest connected to the intended use of the benchmark (specified in Q2)? Are the capabilities theoretically attainable by the objects to be evaluated? Explain the interest in measuring the capabilities in Q3 and question whether it may be impossible for the objects of evaluation to have said capabilities.

They could be directly connected: progress on “general-purpose language understanding technologies” necessarily involves measuring “general-purpose language understanding”. This seems circular, though.

Q5 - SUPPORT: What evidence do you have to support the choice and definition of capabilities of interest?

Nothing mentioned.

The Content Module

The content module specifies test items that the benchmark could require objects of evaluation to perform or to respond to. The test items should elicit evidence about some capability of interest, so that said capability evidence can be later extracted from the responses and aggregated to produce a measurement of said capability.

Q6 - DESCRIBE:

- **Characterize the test items**. Most often, NLP evaluation relies on input data, so this step could involve describing the data that is available to the benchmark to use, how the data is obtained, etc.

  The content instances are from existing work and are organized accordingly:

Prior Work	Description provided in the SuperGLUE paper (p. 5 - 6)
BoolQ (Boolean Questions, Clark et al., 2019)	*“QA task where each example consists of a short passage and a yes/no question about the passage”*
CB (CommitmentBank, De Marneffe et al., 2019)	*“Each example consists of a premise containing an embedded clause and the corresponding hypothesis is the extraction of that clause.”*
COPA (Choice of Plausible Alternatives, Roemmele et al., 2011)	*“**Causal reasoning** task in which a system is given a premise sentence and must determine either the cause or effect of the premise from two possible choices.”*
MultiRC (Multi-Sentence Reading Comprehension, Khashabi et al., 2018)	*“QA task where each example consists of a context paragraph, a question about that paragraph, and a list of possible answers.”* with the following desirable property: *“the questions are designed such that answering each question requires **drawing facts from multiple context sentences**”*
ReCoRD (Reading Comprehension with **Commonsense Reasoning** Dataset, Zhang et al., 2018)	*“Each example consists of a news article and a Cloze-style question about the article in which one entity is masked out. The system must predict the masked out entity from a list of possible entities in the provided passage, where the same entity may be expressed with multiple different surface forms, which are all considered correct.”*
RTE (Recognizing Textual Entailment)	Merging RTE1 (Dagan et al., 2006), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009).

	“All datasets are combined and converted to two-class classification: entailment and not_entailment.”
WiC (Word-in-Context, Pilehvar and Camacho-Collados , 2019)	***“Word sense disambiguation*** *task cast as binary classification of sentence pairs. Given two text snippets and a polysemous word that appears in both sentences, the task is to determine whether the word is used with the same sense in both sentences.”*
WSC (Winograd Schema Challenge, Levesque et al., 2012)	*“**Coreference resolution** task in which examples consist of a sentence with a pronoun and a list of noun phrases from the sentence. The system must determine the correct referrent of the pronoun from among the provided choices.* *Winograd schemas are designed to require **everyday knowledge** and **commonsense reasoning** to solve.”*

- **Which capabilities of interest does each test item aim to capture?** Each instance can aim to capture one or several capabilities amongst those listed in Q3.

Since the benchmark has one capability of interest (GLU), we infer that all test items aim to measure GLU.

An alternative view (of sub-capabilities as explained in Q3) would be that each item aims to measure a type of GLU capability (i.e. sub-capabilities) as inherited from the tasks described in Q6 (e.g., items from the COPA dataset aim to measure “causal reasoning”). All these sub-capabilities, combined together, contribute to the measurement of GLU.

Q7 - JUSTIFY: How does each test item elicit evidence about its target capabilities? Justify via the item descriptions above.

One of SuperGLUE’s criteria for included tasks is that *“tasks should test a system’s ability to understand and reason about texts in English.”* (p. 4) We found no explicit elaboration and justification as to why the included tasks elicit evidence about systems' GLU capability.

Q8 - SUPPORT: What evidence do the benchmark creators offer to support *content validity* of the test items? In other words, we question whether the data captures capabilities of interest. Content validity is often based on analysis by external experts or benchmark users.

Nothing mentioned.

The Adaptation Module

When evaluating humans, the benchmark might instruct them to perform a task by providing instructions, training exercises, demonstrations, etc. When evaluating models/systems, there are also myriad methods that i) modify the models/systems (e.g., fine-tuning), or ii) format or add onto the input (e.g., adding examples in few-shot prompting). These adaptation methods should be chosen carefully so as to not confound evaluation results.

Q9 - DESCRIBE: Given an input, how are the objects of evaluation adapted or instructed to provide the output?

“Any system or method that can produce predictions for the SuperGLUE tasks is eligible for submission to the leaderboard”

The only requirement related to the adaptation method is data-related: *“Systems may only use the SuperGLUE-distributed versions of the task datasets, as these use different train/validation/test splits from other public versions in some cases. Systems also may not use the unlabeled test data for the tasks in system development in any way, may not use the structured source data that was used to collect the WiC labels (sense-annotated example sentences from WordNet, VerbNet, and Wiktionary) in any way, and may not build systems that share information across separate test examples in any way.”* (p. 7)

As for human evaluatees, they are trained before annotating. Task-specific instructions are shown in Appendix C. Example:

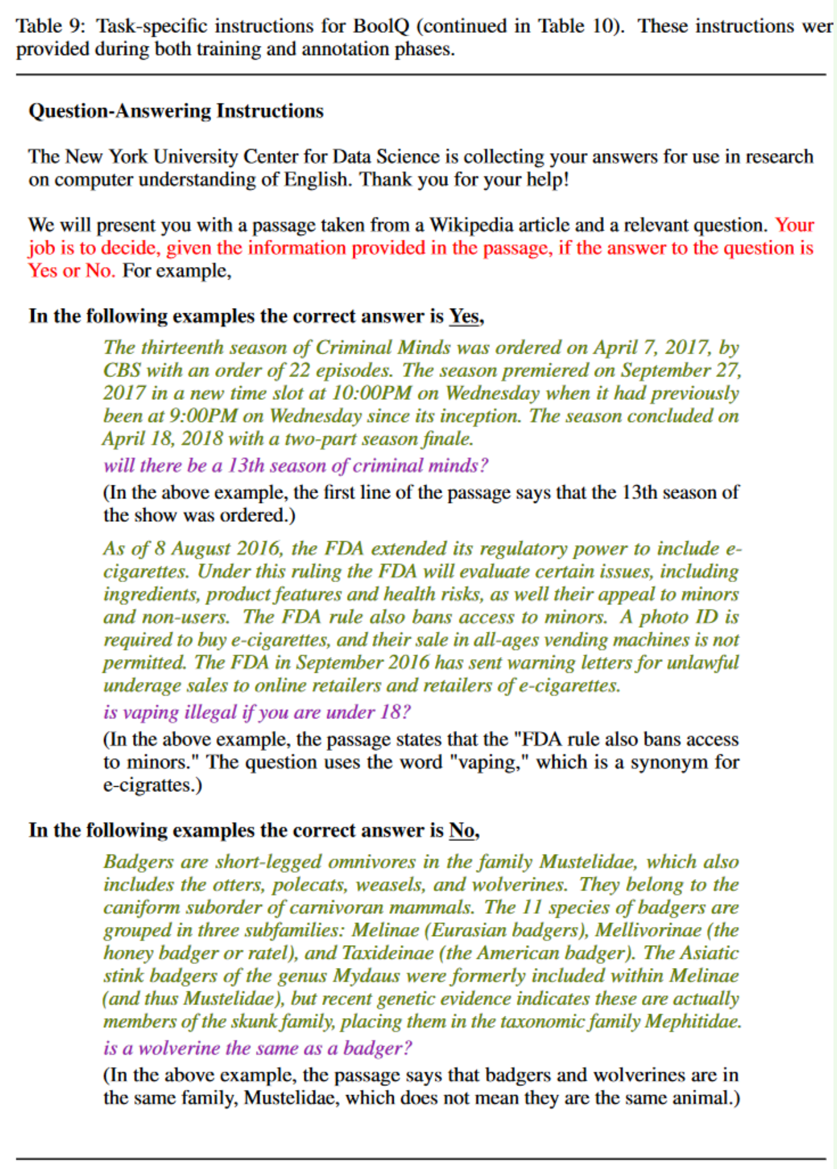


Table 9: Task-specific instructions for BoolQ (continued in Table 10). These instructions wer provided during both training and annotation phases.

Question-Answering Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a passage taken from a Wikipedia article and a relevant question. Your job is to decide, given the information provided in the passage, if the answer to the question is Yes or No. For example,

In the following examples the correct answer is Yes,

> *The thirteenth season of Criminal Minds was ordered on April 7, 2017, by CBS with an order of 22 episodes. The season premiered on September 27, 2017 in a new time slot at 10:00PM on Wednesday when it had previously been at 9:00PM on Wednesday since its inception. The season concluded on April 18, 2018 with a two-part season finale.*
>
> *will there be a 13th season of criminal minds?*
>
> (In the above example, the first line of the passage says that the 13th season of the show was ordered.)
>
> *As of 8 August 2016, the FDA extended its regulatory power to include e-cigarettes. Under this ruling the FDA will evaluate certain issues, including ingredients, product features and health risks, as well their appeal to minors and non-users. The FDA rule also bans access to minors. A photo ID is required to buy e-cigarettes, and their sale in all-ages vending machines is not permitted. The FDA in September 2016 has sent warning letters for unlawful underage sales to online retailers and retailers of e-cigarettes.*
>
> *is vaping illegal if you are under 18?*
>
> (In the above example, the passage states that the "FDA rule also bans access to minors." The question uses the word "vaping," which is a synonym for e-cigrattes.)

In the following examples the correct answer is No,

> *Badgers are short-legged omnivores in the family Mustelidae, which also includes the otters, polecats, weasels, and wolverines. They belong to the caniform suborder of carnivoran mammals. The 11 species of badgers are grouped in three subfamilies: Melinae (Eurasian badgers), Mellivorinae (the honey badger or ratel), and Taxideinae (the American badger). The Asiatic stink badgers of the genus Mydaus were formerly included within Melinae (and thus Mustelidae), but recent genetic evidence indicates these are actually members of the skunk family, placing them in the taxonomic family Mephitidae.*
>
> *is a wolverine the same as a badger?*
>
> (In the above example, the passage says that badgers and wolverines are in the same family, Mustelidae, which does not mean they are the same animal.)

Q10 - JUSTIFY: Elaborate on the suitability of the adaptation methods for all intended objects of evaluation.

Nothing mentioned, but the data requirement might be to ensure “fairness” across submitted systems (avoiding data contamination) (?)

Q11 - SUPPORT: What validity evidence do benchmark designers offer that supports the choice of the adaptation methods?

Nothing mentioned.

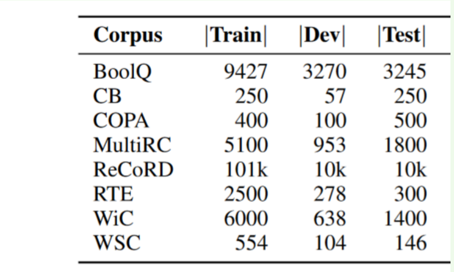
The Assembly Module

Test items specified by the content module are what the benchmark could use. The assembly module concerns what test items from that pool will actually be used by the benchmark for evaluation, and whether this set allows the benchmark to gather sufficient evidence.

Q12 - DESCRIBE: How many items are chosen to assemble the subset used for evaluation? What factors inform this selection?

See sizes of the test sets → (Table 1, p. 3).

There are no mentions of factors informing the assembly of the test sets. We assume that SuperGLUE took the test sets from their original work and omitted to describe how these test sets were originally assembled.



Corpus	|Train|	|Dev|	|Test|
BoolQ	9427	3270	3245
CB	250	57	250
COPA	400	100	500
MultiRC	5100	953	1800
ReCoRD	101k	10k	10k
RTE	2500	278	300
WiC	6000	638	1400
WSC	554	104	146

Exceptions:

- CB: SuperGLUE used *“a subset of the data that had inter-annotator agreement above 80%”* (p. 5)
- WSC: *“The test examples are derived from fiction books and have been shared with us by the authors of the original dataset.”* (p. 6)

For human evaluatees, SuperGLUE randomly sample 100 examples from the task’s test set for them to annotate (p. 17).

Q13 - JUSTIFY: How does the described assembly method ensure that the produced subset elicits sufficient evidence for all capabilities of interest?

Nothing mentioned.

Q14 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of assembly methods?

Nothing mentioned.

The Evidence Module

Evidence Extraction

In response to each presented test item, objects of evaluation produce observable behaviors (referred to as “responses”) which are captured by the benchmark. From these responses, the benchmark extracts evidence about capabilities of interest that said test item targets (referred to as “salient evidence”).

Q15 - DESCRIBE: For each test item…

- **What responses are captured and used for evidence extraction?** When evaluating humans, many types of responses can be captured: selection in multiple-choice questions, long-form answers, response time, etc. Similarly, the benchmark can use the generated text (decoded in a certain way), token probabilities, running time, etc. .

  Predicted class labels

- **How is evidence extracted and represented?**

  The evaluation method for each corpus differs:
  - BoolQ, CB, COPA, RTE, WiC, WSC use exact-match;
  - MultiRC uses F1 over all answer-options and exact match of each question’s set of answers
  - ReCoRD uses max (over all options) token-level F1 and exact match.

  The “gold” labels for each content instance -- we assume -- come from the corpus’ original work. SuperGLUE does not mention exactly how these labels are obtained (e.g., human annotation, use of heuristics, etc.)

Q16 - JUSTIFY: How does the extracted evidence capture the capabilities of interest?

Nothing mentioned.

Q17 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of evidence extraction method?

Nothing mentioned.

Evidence Accumulation

Q18 - DESCRIBE: How is the evidence accumulated to draw insights about the objects of evaluation in terms of capabilities of interest?

For dataset-level scores, ost use accuracy, with the exception of:
- CB: macro-average F1 is also computed →
- MultiRC and ReCoRD: average of item-level score (?)

Average (uniform weights) is computed over dataset-level scores to produce the SuperGLUE score.

Q19 - JUSTIFY: How does the method of accumulating evidence capture capabilities of interest?

For the system performance score, each corpus is weighted equally because “a fair criterion with which to weight the contributions of each task to the overall score” is lacking (p. 6).

The implication is that tasks included in SuperGLUE contribute equally to general language understanding.

Q20 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of evidence accumulation method?

Nothing mentioned.