ECBD: Evidence-Centered Benchmark Design for NLP

Scan for paper!

Yu Lu Liu[5, 1*, 2*], Su Lin Blodgett[3], Jackie Chi Kit Cheung[1, 2, 4], Q. Vera Liao[3], Alexandra Olteanu[3], Ziang Xiao[3*, 5]

[1]Mila – Quebec Artificial Intelligence Institute [2]McGill University
[3]Microsoft Research, Montréal, Canada [4]Canada CIFAR AI Chair [5]Johns Hopkins University
Contact: yliu624@jh.edu *previous affiliation, work done while affiliated

How can we assess benchmark quality?
How can we design better benchmarks?

We take inspiration from **Evidence-Centered Design** (ECD), a framework introduced in the field of education with the goal of guiding the design, evaluation, and interpretation of educational tests (Mislevy et al., 2003).

ECD: ***testing*** *students* as the process of **gathering evidence** from these ***students*** about their ***abilities***.

Evidence-Centered Benchmark Design Framework

We view ***benchmarking*** as the process of **gathering evidence** from ***models*** about their ***capabilities***.

Intended Use
- What are the intended **objects of evaluation**?
- Who are the intended **users** of the benchmark?
- How should the users i**nterpret and use** the benchmark results?

Capability Module
Capabilities that the benchmark aims to measure;
Connection between the benchmark and its intended use.

Content Module
Pool of available test items;
Each item elicits capability evidence about the capabilities it targets.

Adaptation Module
Adapting or instructing the obj. of eval. to respond;
Adaptation methods are well-suited for all intended objects of eval.

Assembly Module
Selecting test items to present to objects of eval.;
Selected set elicits sufficient evidence to measure the capabilities.

Evidence Module

Evidence Extraction
For each test item, capture response from obj. of eval. and extract evidence about the targeted capabilities;
Extracted evidence captures the capabilities targeted by the item.

Evidence Accumulation
Accumulate extracted evidence across all presented items, to measure the capabilities of interest;
Accumulated evidence captures the capabilities of interest.

INPUT / ***OUTPUT***

Object of Eval. (e.g., LM)

Describe
What design decisions are made?

Justify
Why are these decisions made?
Forming a hypothesis: *These decisions enable the module to fulfill its role.*

Support
What **shows** that these decisions indeed enable the module to fulfill its role?
Validity evidence supporting the hypothesis

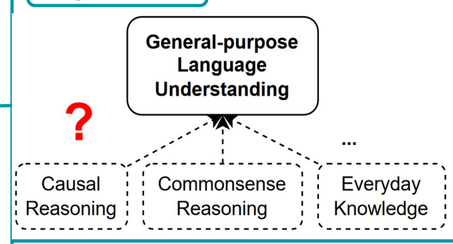
In summary…
- We propose ECBD, a framework guiding practitioners in benchmark creation and analysis;
- We illustrate its usage for benchmark analysis, uncovering issues that threaten benchmark validity.

Case Studies on Existing Benchmarks

BoolQ (Clark et al., 2019) | **SuperGLUE** (Wang et al., 2019) | **HELM** (Liang et al., 2022)

Little description of intended users and how they should interpret and use the benchmark results

HELM
"*Accuracy*" (construct) = *"umbrella term for the standard accuracy-like metric"*
Collapsing capabilities with the way they are measured

SuperGLUE


Unclear decomposition into "intermediate" capabilities

HELM
Using BoolQ dataset to measure "(social) bias", "toxicity", etc.
Repurposing data without appropriate justification

HELM
"(Social) bias," "fairness," "toxicity" measurable without requiring *"knowledge about the broader social context."*
Capabilities lacking appropriate grounding

SuperGLUE **BoolQ**
No prescribed adaptation methods

SuperGLUE **BoolQ** **HELM**
Assembly methods not described, nor justified

SuperGLUE **HELM**
"[ROUGE-2] is the default accuracy metric for CNN-DM and XSUM" (summarization datasets)
Decisions justified by the desire to follow prior work

Benchmarks rarely gather validity evidence to support their design choices

Examples of validity evidence:
- Surveys: what capabilities to measure?
- Prior work conceptualizing capabilities
- Expert panels to examine test items
- Experiments (e.g., on correlation between metric scores and "gold" scores)

Future Directions

Investigating:
- The use of ECBD to guide benchmark creation
- Practitioners use of ECBD through user studies
- How ECBD can be applied to areas beyond the evaluation of NLP models and systems