Evaluation II: Benchmarks

EN. 601.792.01

Ziang Xiao
Department of Computer Science

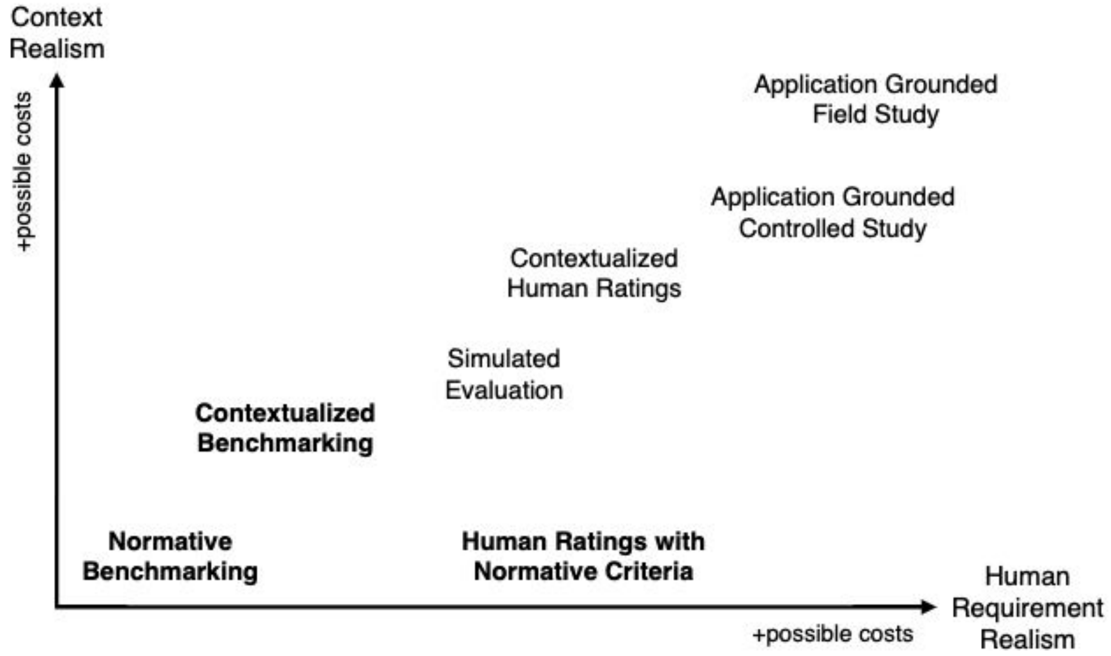
Spring 2024

Announcement

Reading Responses:

- Responding under **other's posts**
- Write about **one** paper only

What is benchmark?



Benchmarks for Conversational Agents

Document Progress

Guide model selection

Benchmarks for Conversational Agents

Empathetic Dialogues

LibriSpeech

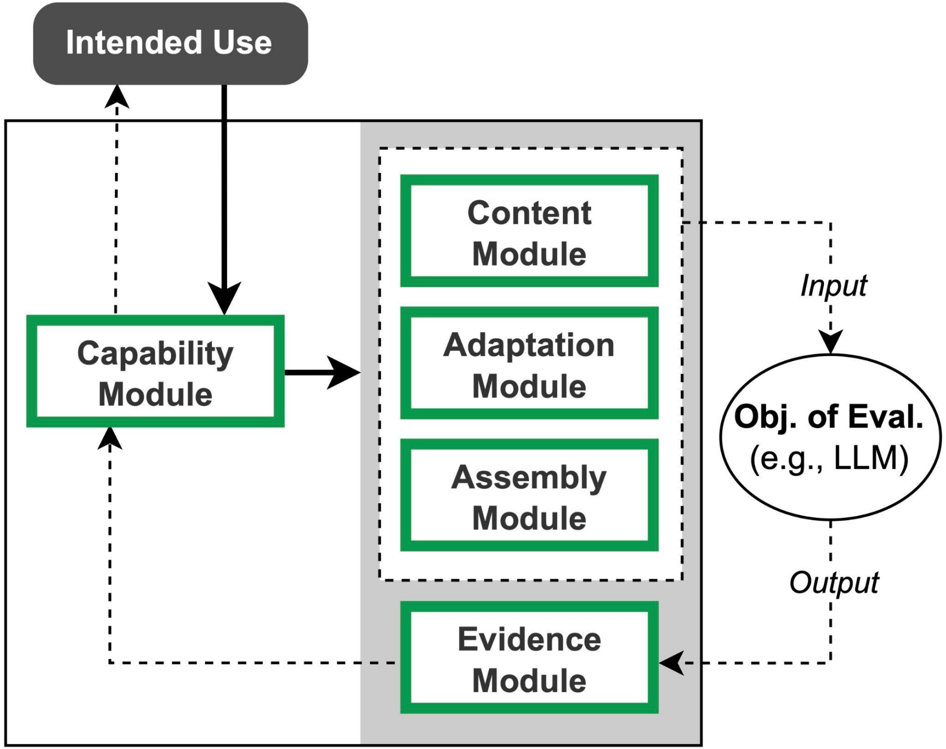
WikiQA

OpenDialKG

OpenbookQA

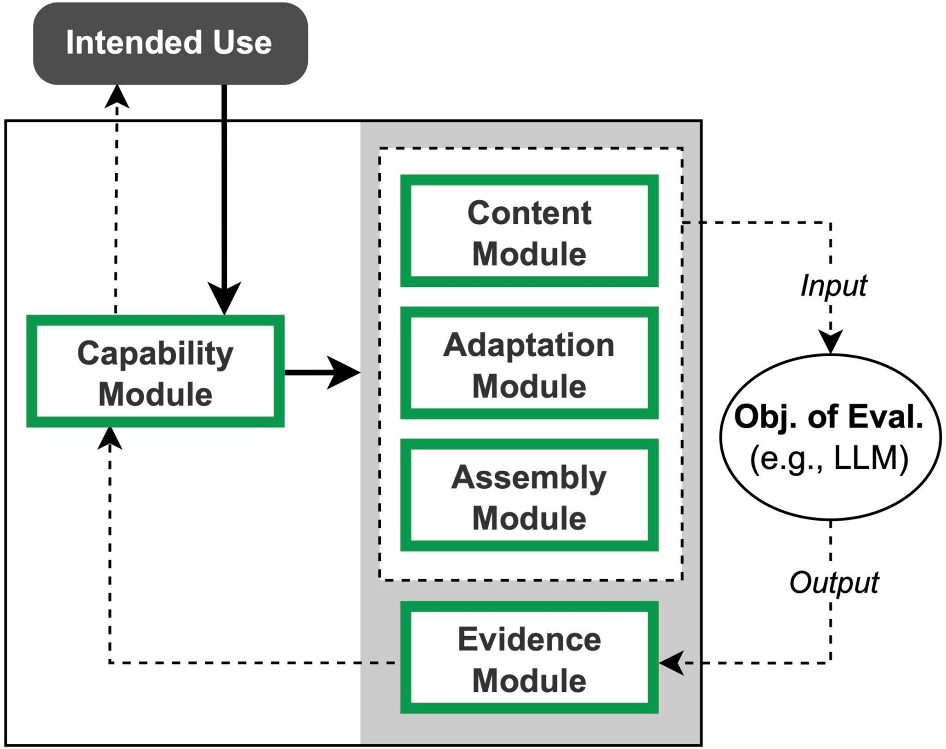
What constitutes a good benchmark?

ECBD: Evidence-centered Benchmark Design

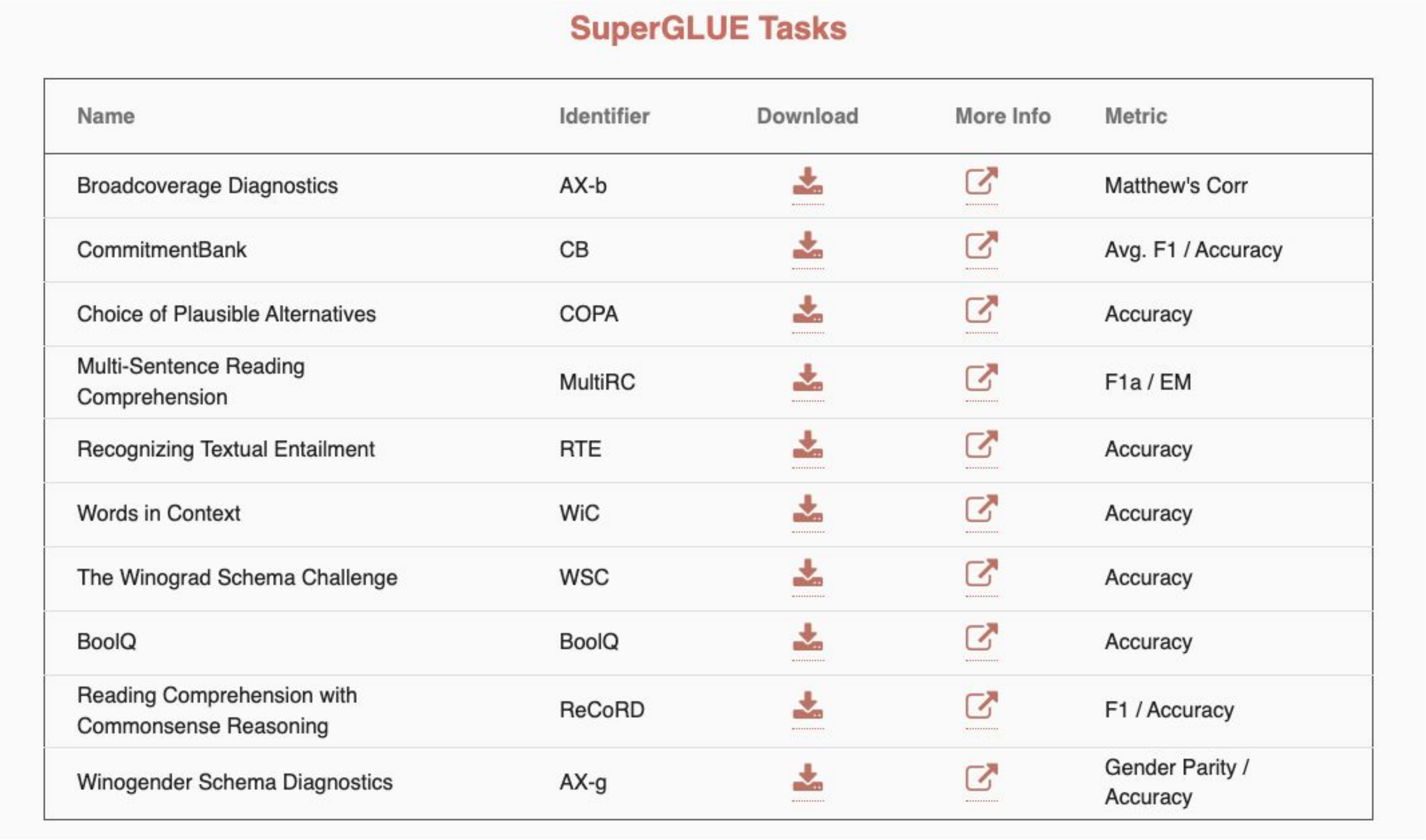


ECBD: Evidence-centered Benchmark Design

ECBD views benchmarking as the process of **gathering**, from objects of evaluation (e.g., LMs), **capability evidence**---i.e., evidence about whether or to what degree said objects have the capabilities of interest.


Intended Use
Capability Module
Content Module
Adaptation Module
Assembly Module
Evidence Module
Input
Obj. of Eval. (e.g., LLM)
Output

SuperGLUE



SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b	⬇	↗	Matthew's Corr
CommitmentBank	CB	⬇	↗	Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA	⬇	↗	Accuracy
Multi-Sentence Reading Comprehension	MultiRC	⬇	↗	F1a / EM
Recognizing Textual Entailment	RTE	⬇	↗	Accuracy
Words in Context	WiC	⬇	↗	Accuracy
The Winograd Schema Challenge	WSC	⬇	↗	Accuracy
BoolQ	BoolQ	⬇	↗	Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD	⬇	↗	F1 / Accuracy
Winogender Schema Diagnostics	AX-g	⬇	↗	Gender Parity / Accuracy

Intended Use

- What are the intended objects of evaluation?
- Who are the intended users of the benchmark?
- How should the users interpret and use the benchmark results?

SuperGLUE

To provide a simple, hard-to-game measure of ***progress*** *toward general-purpose language understanding technologies for English*

Capability Module

Capabilities - constructs that the objects of evaluation are thought to exhibit or posses-that the benchmark aims to measure (i.e., capabilities of interest)

Role: Connection between the benchmark and its intended use.

SuperGLUE

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)

Content Module

Pool of available test examples;

Role: Each example elicits capability evidence about the capabilities it targets

SuperGLUE

BoolQ (Boolean Questions, Clark et al., 2019)
CB (CommitmentBank, De Marneffe et al., 2019)
COPA (Choice of Plausible Alternatives, Roemmele et al., 2011)
MultiRC (Multi-Sentence Reading Comprehension, Khashabi et al., 2018)
ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset, Zhang et al., 2018)
RTE (Recognizing Textual Entailment)
WiC (Word-in-Context, Pilehvar and Camacho-Collados, 2019)
WSC (Winograd Schema Challenge, Levesque et al., 2012)

Adaptation Module

Adapting or instructing the obj. of eval. to respond

Role: Adaptation methods are well-suited for all obj. of eval.

SuperGLUE

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.”

Assembly Module

Selecting test examples to present to obj. of eval.;

Role: Selected set elicits sufficient evidence to measure the capabilities

SuperGLUE

All test datasets

Evidence Module

Evidence Extraction: For each example, capture response from obj. of eval.

and extract evidence about the targeted capabilities

Role: Extracted evidence captures the capabilities targeted by the example.

Evidence Accumulation: Accumulate extracted evidence across all presented

examples, to measure the capabilities of interest.

Role: Accumulated evidence captures the capabilities of interest.

SuperGLUE

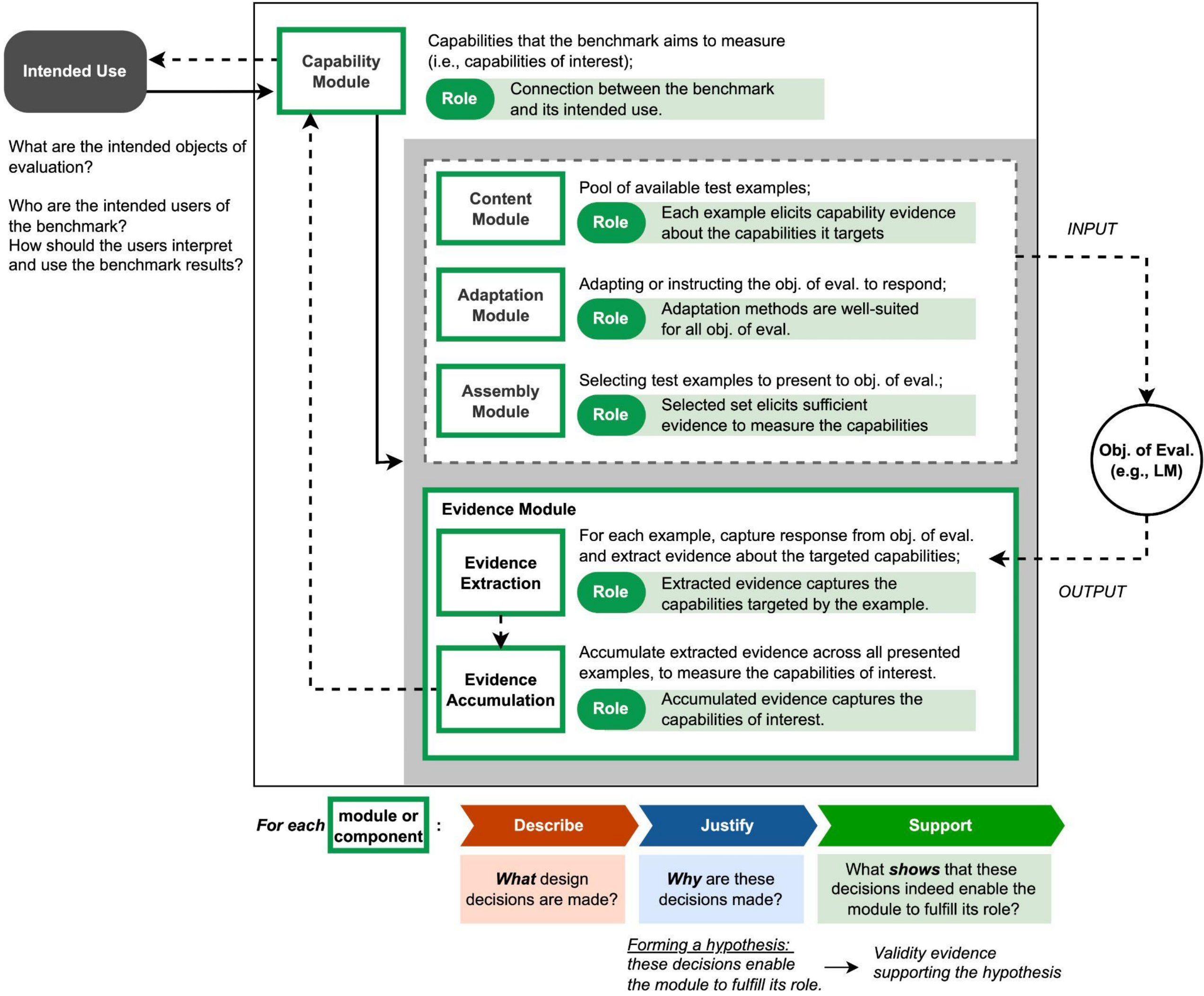
The evaluation method for each corpus differs:

- 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.

SuperGLUE

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

ECBD: Evidence-centered Benchmark Design



Intended Use

What are the intended objects of evaluation?

Who are the intended users of the benchmark?
How should the users interpret and use the benchmark results?

Capability Module

Capabilities that the benchmark aims to measure (i.e., capabilities of interest);

Role: Connection between the benchmark and its intended use.

Content Module

Pool of available test examples;

Role: Each example elicits capability evidence about the capabilities it targets

Adaptation Module

Adapting or instructing the obj. of eval. to respond;

Role: Adaptation methods are well-suited for all obj. of eval.

Assembly Module

Selecting test examples to present to obj. of eval.;

Role: Selected set elicits sufficient evidence to measure the capabilities

INPUT

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

OUTPUT

Evidence Module

Evidence Extraction

For each example, capture response from obj. of eval. and extract evidence about the targeted capabilities;

Role: Extracted evidence captures the capabilities targeted by the example.

Evidence Accumulation

Accumulate extracted evidence across all presented examples, to measure the capabilities of interest.

Role: Accumulated evidence captures the capabilities of interest.

For each **module or component**:

Describe — ***What*** design decisions are made?

Justify — ***Why*** are these decisions made?

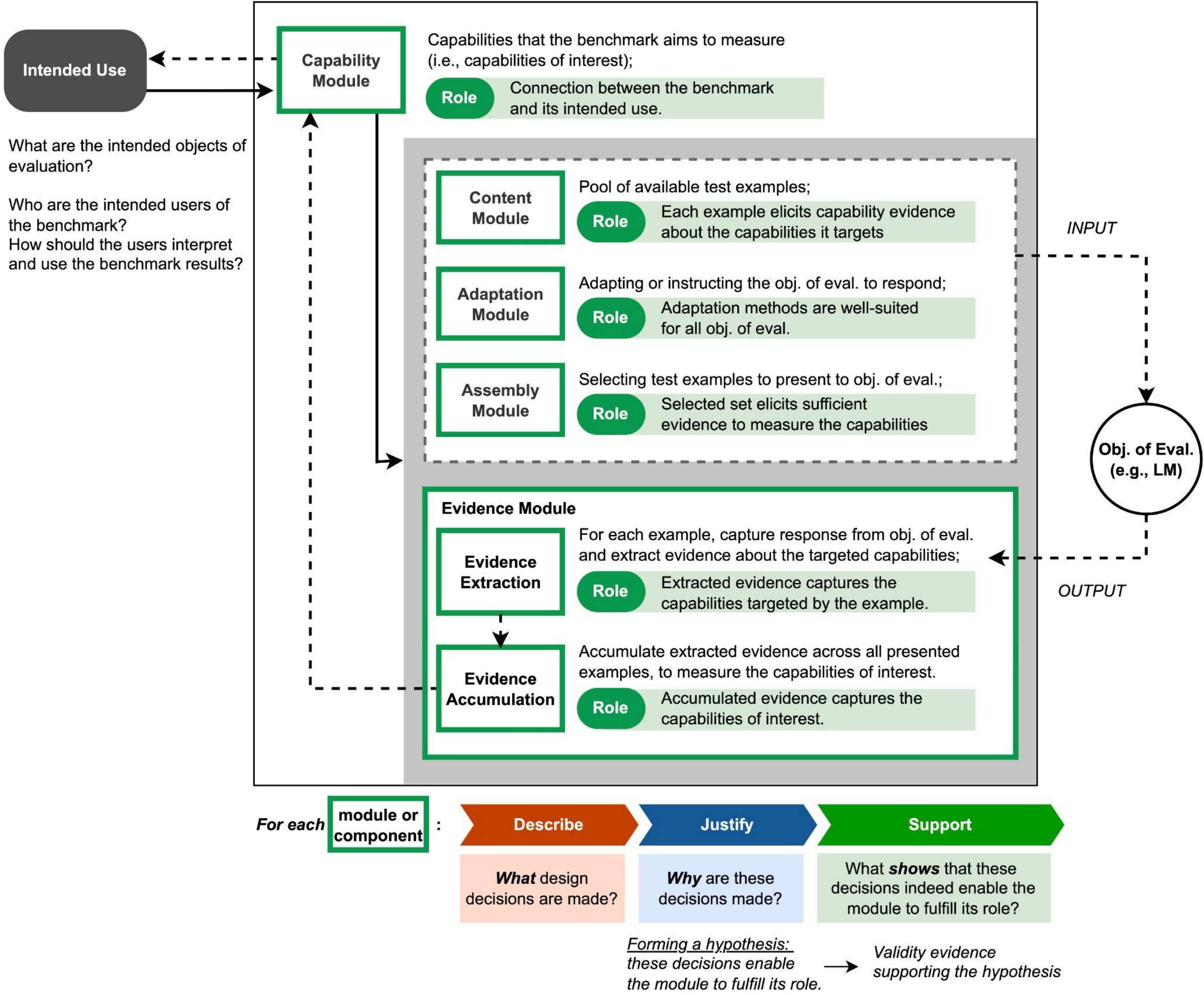
Support — What ***shows*** that these decisions indeed enable the module to fulfill its role?

Forming a hypothesis: these decisions enable the module to fulfill its role. → *Validity evidence supporting the hypothesis*

Presentation

Discussion

According to ECBD, how each module is defined and designed in MMLU?



Intended Use

What are the intended objects of evaluation?

Who are the intended users of the benchmark?
How should the users interpret and use the benchmark results?

Capability Module

Capabilities that the benchmark aims to measure (i.e., capabilities of interest);

Role: Connection between the benchmark and its intended use.

Content Module

Pool of available test examples;

Role: Each example elicits capability evidence about the capabilities it targets

Adaptation Module

Adapting or instructing the obj. of eval. to respond;

Role: Adaptation methods are well-suited for all obj. of eval.

Assembly Module

Selecting test examples to present to obj. of eval.;

Role: Selected set elicits sufficient evidence to measure the capabilities

Evidence Module

Evidence Extraction

For each example, capture response from obj. of eval. and extract evidence about the targeted capabilities;

Role: Extracted evidence captures the capabilities targeted by the example.

Evidence Accumulation

Accumulate extracted evidence across all presented examples, to measure the capabilities of interest.

Role: Accumulated evidence captures the capabilities of interest.

For each module or component:

Describe: ***What*** design decisions are made?

Justify: ***Why*** are these decisions made?

Support: What ***shows*** that these decisions indeed enable the module to fulfill its role?

Forming a hypothesis: these decisions enable the module to fulfill its role. → *Validity evidence supporting the hypothesis*