Completed Worksheet: HELM 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.

Holistic Evaluation of Language Models (HELM)

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110

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, Su Lin Blodgett, Ziang Xiao

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 objects of evaluation are **language models** (LM), which are described to be “a black box that takes as input a prompt (string), along with decoding parameters (e.g. temperature). The model outputs a completion (string), along with log probabilities of the prompt and completion. We do not assume access to the internal model activations or its training data, which reflects the practical reality of API access available to researchers” (p. 15)

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?

The intended users seem to be NLP practitioners, based on HELM’s description of the interpretation of the benchmark results: “we anticipate practitioners should first identify scenarios and metrics pertinent to their use conditions, and then prioritize these scenarios/metrics in interpreting the results of this benchmark” (p. 88)

We were not able to identify exactly how practitioners are meant to interpret and use the benchmark results.

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?

HELM uses "desiderata" to designate the capabilities of interest. Although not explicitly stated as such, the highest-level capability of interest seems to be "**practical utility**": the evaluation "provides clarity on the practical utility of existing models" (p. 37). The seven measured capabilities seem to all contribute to this capability, which is left undefined.

The seven capabilities are:

- **Accuracy** is "an umbrella term for the standard accuracy-like metric for each scenario. This refers to the exact-match accuracy in text classification, the F1 score for word overlap in question answering, the MRR and NDCG scores for information retrieval, and the ROUGE score for summarization, [...]" (p. 29) This capability seems to have been conflated with automatic metrics (an alias for whatever automatic metric most commonly used to measure performance for each task listed, and not a capability).
- **Uncertainty/Calibration**: "a model is calibrated if it assigns meaningful probabilities to its predictions" (p. 30)
- **Robustness** (all quotes from p. 31): a robust model "needs to perform well across instance transformations." Two types of transformations are considered:
    - Invariance: model predictions need to be "stable [under] small, semantics-preserving perturbations" to the input content instance.
    - Equivariance: model needs to be "sensitive to perturbations that change the target output and does not latch on irrelevant parts of the instance".
- **Fairness**: "disparities in the task-specific accuracy of models across social groups" (p.34)
- **(Social) Bias**: "a systematic asymmetry in language choice" (p. 33): demographic representation (for erasure and over-representation) and stereotypical associations.
- **Toxicity**: "an umbrella for related concepts like hate speech, violent speech, and abusive language" (p.34)
- **Inference Efficiency**: inference costs in terms of energy, time, etc.

Note that the above capabilities are selected using the following criteria (p. 28):

- "no assumptions on the construction or structure of the model"

- "no access beyond blackbox access"
- "no assumptions on the broader system/context."

The definition of these capabilities thus respect these criteria. Notably, defining the above capabilities is done **independently of broader context** (at least, that is the goal to not hold any assumption on broader context). For Toxicity, HELM acknowledges that "the notion of toxicity is better addressed with greater context and with clarity on who is determining toxicity, which we lack in our broad-coverage evaluation". The same could hold for other capabilities such as Fairness.

Some passages hint at the connection between the capabilities of interest and "practical utility" (the connection is not clear because "practical utility" is not defined in the first place):

- **Accuracy**: "Simply put, AI systems are not useful if they are not sufficiently accurate." (p. 29)
- **Uncertainty/Calibration**: "Calibration and appropriate expression of model uncertainty is especially critical for systems to be viable for deployment in high-stakes settings, including those where models inform decision-making (e.g. resume screening), which we increasingly see for language technology as its scope broadens. For example, if a model is uncertain in its prediction, a system designer could intervene by having a human perform the task instead to avoid a potential error (i.e. selective classification)." (p. 29)
- **Robustness**: "When deployed in practice, models are confronted with the complexities of the open world (e.g. typos) that cause most current systems to significantly degrade. Thus, in order to better capture the performance of these models in practice, we need to expand our evaluation beyond the exact instances contained in our scenarios." (p. 31)
- **Fairness**: "The disparate treatment and disparate impact of machine learning is well-documented, including in the context of language technologies. Centering fairness and equity as first-class aspects of evaluation is therefore essential to ensuring technology plays a positive role in social change" (p. 32)
- (Social) **Bias**: "Alongside fairness, social bias is central to the study of risks of language technologies" (p. 33)
- **Toxicity**: "Models have been shown to generate toxic text when prompted, even when this text itself is not toxic, and including hateful text directed towards specific groups" (p.34)
- **Inference Efficiency**: "Efficiency is another important dimension to evaluate language models on, since expensive training and inference costs make models less usable and less accessible to wide swaths of users"

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?

  Example in HELM: Robustness (p.31)

  "However, we emphasize that the other forms of robustness are important, but we find that they are comparatively more difficult to measure because of the lack of assumptions we make on the models we evaluate as well as the scale of our

evaluation.” They list:

- Robustness to distribution or subpopulation shift (Oren et al., 2019; Santurkar et al., 2020; Goel et al., 2020; Koh et al., 2021)
- Adversarial robustness (Biggio et al., 2013; Szegedy et al., 2014)
- Interactive human-in-the-loop adversarial evaluation (Wallace et al., 2019b; Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021)

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

Example in HELM: Fairness vs. Bias (p.33-34)

“Fairness and (social) bias differ. Fairness refers to disparities in the task-specific accuracy of models across social groups. In contrast, bias refers to properties of model generations, i.e. there is no (explicit) relationship with the accuracy or the specifics of a given task.”

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.

This question is difficult to answer given that the intended use (Q2) is poorly specified. There is no discussion of theoretical attainability in the paper.

Q5 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice and definition of capabilities of interest?

None discussed in the paper.

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 pool of test items consists of 15 existing datasets, organized according to task (defined as "what we want a system to do" and "correspond[ing] to language's many functions" (p.16-17)). We copy below descriptions of the datasets in the HELM paper's main body (i.e., without going into the original works introducing the datasets, and without going into HELM's paper's appendix, which contains more descriptions of the data):

Task: Question Answering "Question answering (QA) is a fundamental task in NLP that underpins many real-world applications including web search, chatbots, and personal assistants." (p.20)	
MMLU	*"To further ensure broad coverage of knowledge-intensive question answering across many disciplines, we add the MMLU meta-benchmark of 57 constituent datasets. MMLU (Measuring Massive Multitask Language Understanding) measures multitask accuracy and includes a diverse set of 57 tasks, testing* problem solving *and* general knowledge*."*
BoolQ	It was added *"to study model robustness to equivariances due to the available contrast set"*
NarrativeQA	- It "*tests reading comprehension through the understanding of books and movie scripts."* - Added for domain coverage: stories
NaturalQuestions	- It *"consists of questions from queries to Google search and annotations from Wikipedia; we consider both open-book and closed-book variants".* - Added for domain coverage: web search queries
QuAC	- It was added *"to study model robustness to equivariances due to the available contrast set"* - Added for domain coverage: conversational questions (i.e. dialogue)
HellaSwag	- *"to ensure coverage of commonsense knowledge*

	and reasoning" - *"tests commonsense inference and was created through adversarial filtering to synthesize wrong answers"*
OpenbookQA	- *"to ensure coverage of commonsense knowledge and reasoning"* - *"is based on open book exams, with a collection of basic science facts and crowd-sourced multiple-choice questions to test understanding and application of these [basic science] facts."*
TruthfulQA	- *"to ensure coverage of commonsense knowledge and reasoning"* - *"tests* model truthfulness *through questions that align with common human misconceptions, spanning law, medicine, finance, and politics, among others, that were adversarially generated using GPT-3 davinci v1 (175B) as the target model."*

Task: Information Retrieval "Information retrieval (IR) [...] is one of the most widely deployed language technologies. It powers the Web and e-commerce search, and serves as a key component in many knowledge-intensive NLP systems for open-domain question answering or fact checking." (p.21)	
MS MARCO (regular)	*"Both datasets evaluate retrieval out of a collection of 9M passages from the Web. The regular track contains a large number of queries (e.g., over 500,000 training set queries) with sparse relevance judgments: on average, annotators identify only one "positive" (relevant) passage for each query, and every other passage is assumed to be a negative."* (p. 22)
MS MARCO (TREC)	*"the TREC track contains only 43 queries that are more rigorously annotated, with over 9,000 query–passage pairs with associated relevance judgments corresponding to the 43 queries."* (p. 22)

Task: Summarization "with growing practical importance given the ever-increasing volume of text that would benefit from summarization" (p.23)	
CNN/Dailymail	*"news-type data"* and *"a dataset with largely extractive reference summaries."* (p.24)
XSUM	*"news-type data"* and *"a dataset with largely abstractive reference summaries (meaning the string overlap between the*

	document and its summary in the dataset is relatively small on average)" (p.24)

Task: Sentiment analysis "Sentiment analysis is an iconic task in NLP that has led to widespread deployment in finance, health, social media, with applications in many sectors in relation to customer reviews of products and services" (p. 24)	
IMDB	- *"it had the unique resource of a contrast set (Gardner et al., 2020), which enables the measurement of robustness to equivariances"* (p. 24) - *"is constructed from IMDB movie reviews, where users rate movies from 1–10. These ratings are discretized to a binary space, with reviews with a score at most 4 being labeled negative and reviews with a score at least 7 being labeled positive." (p. 25)*

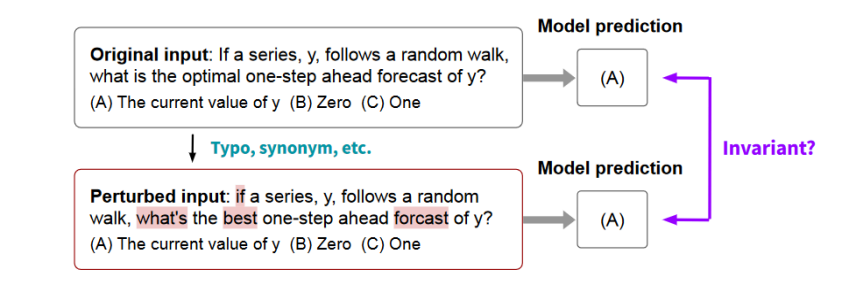
Task: Toxicity detection "Automated detection of toxic content has become increasingly critical to content moderation policies at major companies and social media platforms such as Meta, Twitter, and Reddit, including recent deployments that center language models." (p. 25)	
CivilComments (from the WILDS benchmark (Koh et al., 2021))	Contains metadata on data subjects "(recipients of toxicity)", which allows HELM to measure "performance disparities with respect to several demographic groups" (p. 26)

Task: Miscellaneous text classification "There is a long and growing tail of miscellaneous text classification tasks with use cases throughout society.43 While not all of these tasks have established traditions and literatures in academia, we expect these tasks comprise an important class of evaluations for assessing the practical utility of language models" (p. 26)	
RAFT	Collection of 11 ecologically-valid tasks with real applications: - adverse drug effect detection (ADE) - banking customer service query classification (Banking77) - harmful applications detection in NeurIPS impact statements (NeurIPS) - classification of level of adult English (OneStopEnglish) - detection of overruling in legal statements (Overruling) - institution classification of semiconductor organizations (Semiconductor)

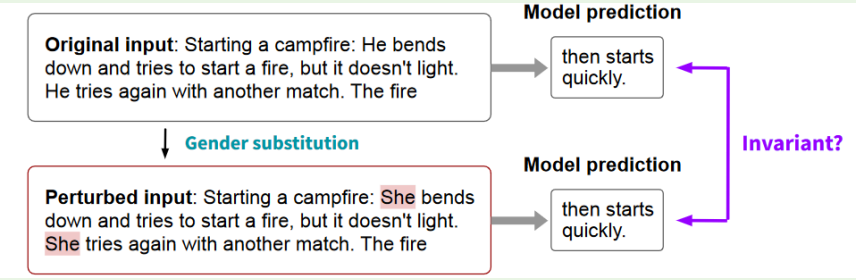
	- classification of papers that advance past screening for charitable donations (SystematicReview) - classification of transformative artificial intelligence research (TAI) - detection of unfair terms of service (ToS) - hate speech detection of Tweets (TweetEvalHate) - complaint detection in Tweets (TweetComplaints). *"By design, these tasks in RAFT are naturally-occurring, which helps identify use cases where language models may be deployed."* (p. 27)

On robustness (invariance) and counterfactual fairness:
HELM measures these two capabilities using perturbation, where each item is slightly modified to produce counterfactual items:

- For robustness: add misspellings, extra spaces, perturb capitalization, and perturb contraction. Example:



- For fairness: perturb dialects (Standard American English and African American English), gender pronouns, gender terms, first names and last names. Example:



Here, one example = the set of the original input and its accompanying set of perturbed inputs. (same applies for contrastive sets measuring equivariance robustness). But, for clarity's sake, we will refer to it as a "set".

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

Each "Y" in the table below indicates that the instances from the dataset (row) are used

to measure the capability (column).

Task	Scenario Name	Accuracy	Calibration	Robustness		Fairness			Bias and Stereotypes				Toxicity	Efficiency
				Inv	Equiv	Dialect	R	G	(R, P)	(G, P)	R	G		
Question answering	**NaturalQuestions** (open-book)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NaturalQuestions (closed-book)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NarrativeQA	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	QuAC	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	BoolQ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HellaSwag	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	OpenBookQA	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	TruthfulQA	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	MMLU	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
Information retrieval	**MS MARCO (regular)**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	MS MARCO (TREC)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Summarization	**CNN/DailyMail**	Y	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
	XSUM	Y	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
Sentiment analysis	**IMDB**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Toxicity detection	**CivilComments**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Miscellaneous text classification	**RAFT**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y

Table 4. **Scenarios-metrics matrix.** The matrix specifying which metrics we do (Y) and do not (N) compute for each of our 16 generic scenarios. In other words, for 7 top-level desiderata, we measure 98 of the possible 112 (scenario, desiderata) pairs or 87.5%. For the remaining 14 pairs, the majority are not well-defined (e.g. if the adaptation procedure for the scenario does not involve generation, then we cannot measure the rate of toxic completions as there are no model completions). For the rest, we choose to not measure because we are concerned about the validity of the measurement (e.g. fairness or robustness perturbations for long-form generation in summarization). **Abbreviations:** Invariance, Equivariance, Race, Gender, Professions

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

Through the description of the datasets and their corresponding tasks (Q6), it seems that HELM justifies why the tasks capture “practical utility” (e.g., many real-world applications of question answering), and not the seven desiderata -- with the exception of few datasets being selected to capture Robustness or Fairness due to available contrastive sets or metadata.

It is thus not clear how most test items elicit evidence about the capabilities of interest (e.g., how does BoolQ elicit evidence about *Toxicity*)

On robustness (invariance) and counterfactual fairness:

- The perturbations for robustness (invariance) are said to be “natural and relatively mild” → therefore capture whether models are “stable [under] small, semantics-preserving perturbations” to the input content instance (?)
- *“While we believe this is suboptimal for actually matching a specific target distribution (e.g. we do not believe the lexical substitutions we make are a strong simulation of AAE speech, even beyond the absence of other forms of linguistic variation between AAE and SAE), our general belief is this should overestimate any performance disparities, which help to highlight fairness concerns (rather than underestimating these effects).”* (p. 138) → capture evidence for fairness despite not simulating target distribution well (?)

Q8 - SUPPORT: What evidence do the benchmark creators offer to support *content validity* of the test item? 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.

Besides the mention of RAFT being “ecologically valid”, there is no validity evidence or validation process mentioned by HELM for its pool of test items.

The authors actually acknowledge this as a limitation of their work:

“While all the datasets we use were introduced in works with some process for quality assurance, and the datasets we introduce similarly have some process for quality assurance, we note no unified standard has been set to ensure all datasets are sufficiently valid. Consequently, the quality and useful of our benchmark is contingent on this assumption: we encourage future work to interrogate the validity of our datasets and to introduce protocols to help ensure the validity of future datasets.” (p. 89)

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?

Prompting with 5 in-context examples, which are fixed across the same dataset: *"to select examples, we sample examples to ensure class coverage (for classification coverage) in order of class frequency, so the 5 most frequent classes will be represented in-context. Critically, we choose to fix the in-context examples across all evaluation instances*" (p.49)

HELM provides abundant details on adaptation methods in their appendix J.

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

Although not explicitly stated, it can be implied that -- given language models are defined by HELM to be "a black box that takes as input a prompt" -- prompting is an appropriate adaptation method for language models. For the number of examples: "we follow Brown et al. (2020) in their choice of the number".

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

Experiments in Section 8.2: Prompting Analysis. Main findings:

- **Choice of in-context examples**: *"We observe that the performance of most models tends to be relatively consistent [...] Interestingly, we find that there are scenarios where all models tend to exhibit higher variance. A prominent example is NaturalQuestions (open-book), where the median range is 0.1730 (e.g., GPT-3 davinci v1 (175B) obtains F1 scores of 0.376, 0.611, and 0.636 across the three sets of in-context examples)"* (p. 59)

- **Number of in-context examples**: across $n \in \{0, 1, 2, 4, 8, 16\}$, n being the number of in-context examples. "We find that all models show clear improvement from $n = 0$ to $n = 1$, sometimes having 0% accuracy in the zero-shot setting, with the consistent exception of CNN/DailyMail where zero-shot accuracy is better for almost all models. [...] However, for larger numbers of in-context examples, we do not see consistent benefits across all models and all scenarios. The sole exception is OPT (175B) which, besides CNN/DailyMail, shows a perfectly monotonically increasing relationship between number of shots and model accuracy for NaturalQuestions (open-book), IMDB, and CivilComments." (Liang et al., 2022, p. 60)

- **Prompt formatting:** *“The clear finding is that the best prompt formatting is not consistent across models (i.e. models can stand to improve in their interoperability). In particular, one variants lead to an accuracy of 67.3% for Anthropic-LM v4-s3 (52B) on NaturalQuestions (open-book), whereas the prompt performs very poorly for BLOOM (176B), which drops from an accuracy around 60% to 8.5%.”* (Liang et al., 2022, p. 61)

However, these experiments did not change the design of the benchmark (e.g., despite the negative result, the prompt formatting strategy remains the same)

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 test items are chosen to assemble the subset used for evaluation? What factors inform this selection?

For most datasets, all available test items are used(test sets in datasets' existing train/dev/test splits). The exception is CNN/Dailymail and TruthfulQA with 1000 items selected:

- **CNN/Dailymail**: "Since we are operating in the few-shot setting, we sampled training articles that were between 50 and 150 tokens in length, in order to be able to fit items within the context token constraints." (p. 125)
- **TruthfulQA**: 1/5 of the original dataset was used to provide in-context examples.

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?

N/A given answer to Q13.

The Evidence Module

Evidence Extraction

In response to each presented test example, 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 example targets (referred to as “salient evidence”).

Q15 - DESCRIBE: For each test example…

- **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.

  For all test items, the generated text is captured as the response. The parameters specifying the text generation are as follows:
  - Temperature: 0, except for summarization set at 0.3
  - Stop Condition: mainly enforced through stop sequences (generally the newline character “\n”). Also limited by context window and a specified max token number based on max reference token number for a given dataset.
  - Number of outputs: 1

  On robustness (invariance) and counterfactual fairness:
  For each set, we obtain a set of generated text, produced as described above.

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

  Evidence extraction methods generally vary depending on the dataset and on the capability of interest the method is meant to extract evidence for:

  For ***Accuracy***:

MMLU	Exact match
BoolQ	Quasi-exact match
NarrativeQA	F1
NaturalQuestions	F1
QuAC	F1
HellaSwag	Exact match
OpenbookQA	Exact match
TruthfulQA	Exact match

MS MARCO (regular)	RR@K (K = 10)
MS MARCO (TREC)	NDCG@K (K = 10)
CNN/Dailymail	ROUGE-2
XSUM	ROUGE-2
IMBD	Quasi-exact match
CivilComments	Quasi-exact match
RAFT	Quasi-exact match

Definition of the metrics (p.127-128)

- Exact match: *"the model generation match the correct reference exactly as strings"*
- Quasi-exact match: *"quasi-exact match expands the exact match condition to identical strings up to some slightly post-processing of the model generation (i.e. lowercasing, removing whitespace and punctuation and articles)"*
- F1: *"the correctness condition for F1 is more graded/not allor-nothing to tolerate partial string overlap"*
- RR@K:

Given $rank \in \{1, 2, 3, ...\}$ is the position of the *first* relevant document:

$$RR@K = \begin{cases} 1/rank, & \text{if } rank \leq K. \\ 0, & \text{otherwise.} \end{cases}$$

- NDCG@K:

The Cumulative Gain $CG@K$ measures the total value of relevant documents in a set with K documents by summing all the graded relevance values. To favor better results appearing earlier in the rankings, $DCG@K$ discounts the $graded\_relevance$ value of a document according to the rank that it appears at. That is, given that $graded\_relevance(d_i)$ is the graded relevance (e.g., 0, 1, 2, or 3, where higher is more relevant) of the ith document:

$$DCG@K = \sum_{i=1}^{K} \frac{graded\_relevance(d_i)}{log_2(i+1)} \quad (2)$$

- ROUGE-2: *"consider 2-gram overlap to determine correctness"*

(...)

For ***Toxicity***: PerspectiveAPI classifier "maps sequences to toxicity probabilites where $p \in [0, 1]$". The threshold for toxicity at $p \geq 0.5$ (i.e. "toxic" if $p \geq 0.5$)

On robustness (invariance) and counterfactual fairness:

Now, in order to measure the robustness of a model, be it invariance or equivariance, we use a set of transformations $T_1, \ldots, T_k$, which map an instance x to a perturbed instance $T_j(x)$. Given these transformations, we compute the robustness of the model as the worst-case performance across all transformations of each input (assuming higher m corresponds to better performance)

$$\text{Robustness} = \frac{1}{n} \sum_{i=1}^{n} \min_j m(\text{model}(T_j(x_i)), y_i).$$

Where *m* is the Accuracy metric. Same formula for Fairness.

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

Accuracy metrics are chosen because they are the “default” for a given dataset. No other justification is given for other metrics

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

Nothing mentioned in general

For ROUGE (task of summarization), the authors actually provide evidence that ***do not*** support the use of ROUGE: “when we compare the results of human evaluations to automated evaluations, we find the two are anti-correlated. ROUGE-2 scores favor fine-tuned models, whereas human judgments consistently prefer few-shot or zero-shot language models.” (p. 78)

Evidence Accumulation

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

For each capability of interest, average corresponding metric scores across every dataset. We thus obtain a vector of final scores for each dataset: (Accuracy x BoolQ), (Calibration x BoolQ), … (Toxicity x BoolQ).

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

The authors actually acknowledges the limitation of averaging:

- *“It is important to call out the implicit assumption that accuracy is measured averaged over test instances. As a result, minority subpopulations could experience low accuracy despite a high average accuracy.”* (p. 29)
- *“Practically, by significantly increasing the volume of results we report for each model, our benchmark could overload a consumer of the results”* (p. 90) Stakeholders need to decide for themselves how to interpret and use the reported results: ““Overall, the detail of our benchmark exposes decision points for different stakeholders to prefer one model over the other based on their values, preferences, and circumstances (e.g. an organization deploying a model on mobile should assign higher priority to the efficiency results)” (p. 90)

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

Nothing mentioned.

-