

# Construct Validity Checklist

This checklist follows the recommendations made in the paper:

*Measuring what Matters: Construct Validity in Large Language Model Benchmarks*

NeurIPS 2025 Datasets & Benchmarks

<https://openreview.net/pdf?id=mdA51VvNcU>

## Define the phenomenon

- ☐ Provide a precise and operational definition for the phenomenon being measured
- ☐ Specify the scope of the phenomenon being covered and acknowledge any excluded aspects
- ☐ Identify if the phenomenon has sub-components and ensure they are measured separately

## Measure only the phenomenon

- ☐ Control for unrelated tasks that may affect the results
- ☐ Assess the impact of format constraints on model performance
- ☐ Validate any automated output parsing techniques for accuracy, consistency and bias

## Construct a representative dataset for the task

- ☐ Employ sampling strategies to ensure task items are representative of the overall task space
- ☐ Verify the quality and relevance of all task items, especially for large or automatically generated datasets
- ☐ Include task items that test known LLM sensitivities (e.g. input permutations or variations)

## Acknowledge limitations of reusing datasets

- ☐ Document whether the benchmark adapts a previous dataset or benchmark
- ☐ If so, analyse and report the relevant strengths and limitations of the adapted prior work
- ☐ If so, report and compare performance on the new benchmark against the original
- ☐ Explain modifications to reused datasets and how they improve construct validity

## Prepare for contamination

- ☐ Implement tests to detect data contamination and apply them to the benchmark
- ☐ Maintain a held-out set of task items to facilitate ongoing, uncontaminated evaluation
- ☐ Investigate the potential pre-exposure of benchmark source materials or similar data in common LLM training corpora

## Use statistical methods to compare models

- ☐ Report the benchmark's sample size and justify its statistical power
- ☐ Report uncertainty estimates for all primary scores to enable robust model comparisons
- ☐ If using human raters, describe their demographics and mitigate potential demographic biases in rater recruitment and instructions
- ☐ Use metrics that capture the inherent variability of any subjective labels, without relying on single-point aggregation or exact matching.

## Conduct an error analysis

- ☐ Conduct a qualitative and quantitative analysis of common failure modes
- ☐ Investigate whether failure modes correlate with non-targeted phenomena (confounders) rather than the intended construct
- ☐ If so, identify and discuss any potential scoring biases revealed in the error analysis
- ☐ Conduct experiments or propose new directions to improve model scores on the benchmark

## Justify construct validity

- ☐ Justify the relevance of the benchmark for the phenomenon with real-world applications
- ☐ Provide a clear rationale for the choice of tasks and metrics, connected to the operational definition of the phenomenon
- ☐ Compare similarities and differences between the benchmark and existing evaluations of similar phenomena
- ☐ Discuss the limitations and design trade-offs of the benchmark concerning construct validity