Uplifting Human Decision Making in AI Evaluation by Automating Benchmark Validity Analysis

Rodolfo Corona [*1] **Sang T. Truong** [*2] **Nhi N. Truong** [2] **Ritwik Gupta** [3] **Fred Philippy** [4] **Atnafu Lambebo Tonja** [5] **Mena Attia** [6] **Fahim Faisal** [7] **Kaushal Kumar Maurya** [6 8] **Belu Ticona** [7] **Sumaya Nur Adan** [9] **Fazl Barez** [9 10] **Omar U. Florez** [11] **Supheakmungkol Sarin** [12] **Aseem Srivastava** [6] **Xiaoyuan Yi** [13] **Nick Haber** [2] **Dan Klein** [1] **Thamar Solorio** [6] **Xing Xie** [13] **Sanmi Koyejo** [† 2] **Robert Trager** [† 14 9]

Abstract

The capabilities of AI systems are dependent on the specific contexts in which they are deployed, and thus, benchmarks for AI systems must be valid for those contexts. In this paper, we present a framework to assess the validity of AI benchmarks in the context of specific deployment use cases and target populations, as well as a semi-automated pipeline for fast execution of the framework. Our framework is structured around the ontological, instance-level, and representational components of benchmark inputs and outputs, specifying the conditions under which benchmark evaluation validity would be violated if transferred to a different cultural or geographic context. To enable scalable validity analysis, our automated pipeline leverages large language models to evaluate benchmark-use-population triplets across our six validity dimensions. We validate this pipeline through human expert studies and apply it to assess 24 triplets spanning five global regions, surfacing systematic patterns in how porting strategies affect validity.

1. Introduction

Evaluations of AI models are meant to assess their ability to solve problems, but as the types of behaviors being measured have become increasingly complex, any single benchmark has less applicability to actual user experience. Instead, evaluations are increasingly meant to serve as cheap, scalable proxies for complex real-world phenomena. To be useful, these proxies must be *valid* indicators of model capabilities (Bean et al., 2025). However, prior work demonstrates that benchmarks exhibit high variance and fail when ported outside their original cultural and geographic distributions (Singh et al., 2025; Wu et al., 2025a), implicitly encoding the cultural norms of their creators (Gupta et al., 2025a).

Prior work has proposed frameworks for assessing the validity of evaluations (Salaudeen et al., 2025; Liu et al., 2024; Wallach et al., 2025). Yet, these frameworks remain mostly theoretical, and their application can require substantial manual assessment, inhibiting the scale of data which can be reviewed by their users.

In this paper, we provide a framework by which domain experts can systematically, and contextually, assess validity in AI benchmarks in a semi-automated setting. We first formalize a six-dimensional validity taxonomy that defines how transferring benchmarks across cultures violates core evaluation assumptions. Building on this framework, we introduce a LLM-assisted pipeline that empowers a single expert to rapidly audit benchmark-use-population triplets. Our pipeline combines the fidelity of human expertise with the recall and speed of tool-use LLMs to surface contextual gaps and propose data remediation steps. By fixing this specific failure mode, we can restore the downstream utility of these benchmarks.

Our contributions are two-fold:

1. **Theoretical Framework:** We present a six-dimensional framework for assessing contextual benchmark validity, defining how cross-cultural transfer violates evaluation assumptions across ontological, content, and representational components (Section 3).

2. **Semi-Automated Evaluation Pipeline:** We introduce

[*]Equal contribution [†]Equal advising. [1]UC Berkeley [2]Stanford University [3]University of Maryland, College Park [4]University of Luxembourg [5]University College London [6]MBZUAI [7]George Mason University [8]IITDAD [9]University of Oxford [10]Martian [11]National Center for AI (CENIA) [12]AI Safety Asia [13]Microsoft Research Asia [14]Oxford Martin AI Governance Initiative. Correspondence to: Rodolfo Corona <rcorona@berkeley.edu>, Sang T. Truong <sttruong@stanford.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

a scalable pipeline[1] that operationalizes our framework. This tool allows a single domain expert to efficiently identify validity failure modes in existing benchmarks. We validate our pipeline through expert evaluation and apply it to analyze 24 benchmarks across five global regions. Experts report that 29% of assessments would match their own expert review and that 45% would complement it, with 54% of assessments surfacing new validity concerns they were previously unaware of (Section 5).

2. Background

Our framework treats benchmarks as instruments for supporting claims about model behavior in a specified deployment setting. We first develop the conceptual foundations of this view, moving from intended use to evaluation needs, benchmark components, validity, and existing porting strategies.

2.1. Conceptual Foundations

From Intended Use to Evaluation Needs Benchmark evaluation begins with an *intended use*, consisting of the deployment context, the target population, and the decisions that the evaluation is meant to inform (Liu et al., 2024). We use *target population* to refer to the end-users and affected communities in the deployment context, and *regional stakeholders* to refer to the broader set of actors who can inform benchmark design, including end-users, domain experts, developers, policymakers, and community representatives. Intended use determines the *constructs* that a benchmark should provide evidence for, such as "language understanding," "diagnostic reasoning," or culturally appropriate assistance. Because such constructs are not directly observable, benchmark designers must identify measurable proxies that are appropriate for the deployment setting (Salaudeen et al., 2025). This discovery process may be stakeholder-driven, through participatory design with regional stakeholders (Birhane et al., 2022), or data-driven, through analysis of post-deployment monitoring data or related domain data. We discuss these mechanisms in Appendix G.1.

Benchmark Components Once evaluation needs have been identified, a benchmark operationalizes them through a **dataset** and a set of **criteria**. The dataset consists of *inputs* such as text queries, medical images, or audio signals, and, when applicable, *expected outputs* such as categorical labels, reference text, or acceptable actions. The *criteria* specify how model outputs are converted into scores, either by comparing them against expected outputs using metrics such as accuracy or recall, or by directly evaluating correctness through mechanisms such as human preference judgments, code test-case passage (Jimenez et al., 2024), or LLM-as-a-judge evaluations (Zheng et al., 2023). This decomposition matters for contextual transfer because different benchmark components can fail in different ways. Inputs may omit locally relevant cases, expected outputs may encode source-population judgments, and scoring criteria may reward behavior that is inappropriate for the target population or use case.

Evaluation Validity Validity concerns the inference from a benchmark score to a claim about model behavior in context (Bandalos, 2018). Following Salaudeen et al. (2025), relevant validity dimensions include *content* validity, concerning coverage of the relevant input space; *criterion* validity, concerning correlation with established standards; *construct* validity, concerning whether the benchmark measures the intended construct; *external* validity, concerning generalization across settings; and *consequential* validity, concerning whether benchmark-driven decisions align with the benchmark's purpose. A detailed taxonomy is provided in Appendix G.2. These validity dimensions describe the kinds of claims that can fail during benchmark transfer. In Section 3, we use the components above to identify where such failures arise in benchmark inputs and outputs.

Benchmark Porting in Practice The connection between benchmark components and validity becomes especially important when benchmarks are adapted across linguistic, cultural, or geographic contexts. Existing porting strategies vary in how much they alter the original benchmark. *Direct translation* preserves the source benchmark structure while making items linguistically accessible (Conneau et al., 2018; Bandarkar et al., 2023), although it can also preserve source-culture assumptions about relevant knowledge, values, and interaction norms (Wu et al., 2025a; Singh et al., 2025). *Parallel construction* builds analogous benchmarks from regional sources, such as local examinations (Huang et al., 2023; Son et al., 2024; Koto et al., 2024; Romanou et al., 2024). *Ground-up development* centers regional communities and locally defined evaluation needs (Adelani et al., 2021; 2024; Ángel González et al., 2026; Susanto et al., 2025).

These strategies address important limitations of translation, while leaving open a more diagnostic question. For a given benchmark, use case, and target population, which validity assumptions are supported, which are violated, and which benchmark components require remediation? Domain-specific benchmarks make this question especially acute. In medicine and law, professional knowledge and decision rules are often jurisdiction-specific (Jin et al., 2020; Olatunji et al., 2023; Guha et al., 2023). Cultural evaluation

[1] Github repo: https://github.com/aims-foundations/benchmark-caliper

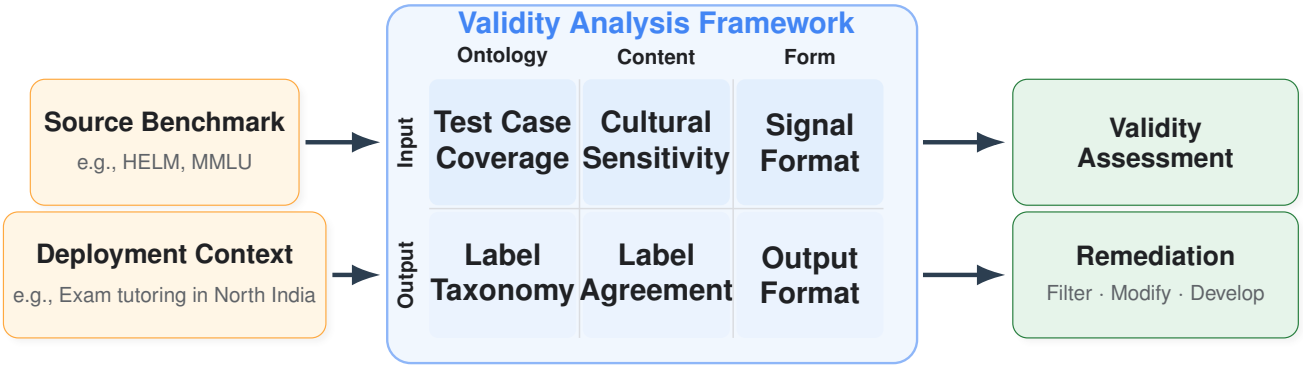


Figure 1. **Overview of the Benchmark Validity Analysis Workflow.** A source benchmark and a target deployment context (determined by a **use case** and **target population**) are jointly assessed against the six-dimensional validity framework (Input/Output × Ontology/-Content/Form). The analysis produces a *validity assessment* that surfaces construct-relevant gaps, and a *remediation strategy* guiding benchmark designers to filter, modify, or develop regional data.

benchmarks similarly show that linguistic coverage alone does not ensure that a benchmark measures culturally relevant constructs (Choenni et al., 2024; Myung et al., 2024; Sadr et al., 2025). These observations motivate the framework in Section 3, which assesses whether a benchmark is valid for a specified use case and target population by locating failures across the ontology, content, and form of benchmark inputs and outputs.

3. Contextual Benchmark Validity Framework

The validity of a benchmark score depends on the claim it is used to support. We assess benchmark applicability relative to a specified deployment context, and use the framework in this section to diagnose when benchmark transfer threatens that validity. We defer extended examples, evidence sources, and remediation details to Appendix B and the practitioner checklists to Appendix C to keep the main paper compact.

3.1. Assessment Unit and Object Model

We evaluate a benchmark-context pair (B, D), where B denotes a benchmark and D denotes the deployment context in which the benchmark is used to support a claim about model behavior. We define the deployment context as $D = (U, T)$, where U is the intended use case (also broadly specifying the deployment environment, including interface, modality, institutional setting, and operational constraints), and T is the target population of end users and affected communities. We represent the benchmark as $B = (X, Y, C, M)$, where X is the input set or distribution, Y is the expected outputs or annotations, C is the scoring criterion, and M is metadata such as provenance, sampling procedure, and documentation.

This follows the view of validity as an evidential relation between an evaluation score and its intended interpretation (Messick, 1989; Bean et al., 2025). A benchmark can provide strong evidence for one deployment claim and weak evidence for another. For example, an ink-printed high-school mathematics benchmark can support claims about how sighted students perform on algebra, but provide weak evidence about how *blind* students tackle such problems. A validity threat arises when some component of B is incompatible with the requirements induced by U and T.

3.2. Six Compatibility Indicators

Our framework organizes benchmark validity concerns along two sides of the benchmark: inputs and outputs, and three levels of abstraction: ontology, content, and form. Ontology asks which categories belong in the evaluation. Content asks which specific instances or judgments instantiate those categories. Form asks how inputs or outputs are represented or elicited. Crossing the two sides with the three levels yields six compatibility indicators.

The purpose of this decomposition is diagnostic. Different misalignments point to different interventions. Ontological failures usually require adding, removing, or restructuring categories; content failures call for filtering, rewriting, supplementing, or relabeling specific instances; and form failures call for changes in modality, language, script, interface, or signal conditions. In practice some cases blur these boundaries. For example, language can be a form issue when the concern is script or accessibility and a content issue when the concern is the cultural assumptions expressed by an item. We therefore classify a problem by the component where remediation should occur. Following recent work on construct validity in LLM benchmarks (Bean et al., 2025), we pair each indicator with an actionable checklist in Appendix C.

Input ontology. Input ontology asks whether the benchmark covers the test-case categories required by the use case and target population. At this level, the key question is

whether the benchmark samples from the right task, domain, or scenario categories at all. The quality of any individual item is a separate issue. When benchmarks are ported across regions, failures often arise because the source benchmark reflects the categories salient to its original designers. A driving benchmark developed under European road assumptions, for example, may omit test cases for making turns on red lights (as is done in the US). Typical remediation is to add missing categories, remove inherited but irrelevant ones, or revise the task taxonomy before collecting additional data.

Input content. Input content asks whether individual inputs instantiate those categories using assumptions, references, and background knowledge that are construct-relevant for the target context. A benchmark may cover the right domain while still containing items that require culturally specific references, foreign institutional knowledge, or local background assumptions outside the intended construct. Cultural sensitivity often appears at this level in benchmark transfer. Singh et al. (2025) show that many MMLU items depend on culturally specific knowledge, much of it Western-specific. When such knowledge falls outside the deployment claim, the usual remedy is to filter, rewrite, or locally replace the affected items. Revising the whole taxonomy is often unnecessary.

Input form. Input form asks whether benchmark inputs match the modalities, languages, scripts, interfaces, and signal conditions expected in deployment. Even a benchmark with the right categories and content can be poorly calibrated if it is administered in clean written text, high-resource languages, or idealized sensor conditions while deployment occurs through speech, code-switching, low-bandwidth interfaces, or lower-quality imaging devices. This is a familiar form of distribution shift (Saenko et al., 2010). Remediation often requires translation, transliteration, recollection under target conditions, or explicitly narrowing the claim to the benchmark's input form.

Output ontology. Output ontology asks whether the benchmark's label sets, acceptable response types, rubric dimensions, or scoring criteria reflect the decision structure appropriate for the use case and target population. This matters especially for open-ended or value-laden tasks, where the scoring rule determines which distinctions the evaluation can see. Label sets and rubrics encode political, cultural, institutional, and domain-specific assumptions (Miceli et al., 2020). A benchmark can therefore miss locally relevant harms, prohibited recommendations, or interaction norms even when its inputs are appropriate. In such cases, remediation requires redesigning labels, expanding rubric dimensions, or changing the scoring criterion itself.

Output content. Output content asks whether labels, reference answers, annotations, or judgments for particular instances agree with target-population or stakeholder judgments. It differs from output ontology because the categories themselves may be appropriate even when the instance-level assignments are mismatched. The issue is especially salient when benchmarks encode offensiveness, helpfulness, safety, legality, or acceptability, since such judgments can vary across populations. For example, Lee et al. (2024) find substantial cross-country disagreement in hate-speech annotations, and perspectivist work argues that standard aggregation can erase minoritized views (Fleisig et al., 2024). Remediation therefore often involves reannotation with representative annotators, expert adjudication, or preserving multiple valid perspectives when a single label would obscure meaningful disagreement.

Output form. Output form asks whether the benchmark evaluates the output modalities, languages, formats, and interaction patterns required in deployment. A benchmark that rewards only English, single-turn, written text may say little about systems that must generate speech, code-switched responses, multimodal explanations, or action-oriented outputs. In such cases, a model may appear capable under the benchmark while failing in the actual interface. The remedy is to widen accepted output formats, add output modalities, or build evaluation variants that better match the deployment environment.

3.3. Using the Framework

The six indicators support the assessment procedure used in the remainder of the paper. Given a benchmark-context pair (B, D), the assessor first specifies the deployment claim the benchmark is meant to support. The assessor then extracts benchmark properties from documentation, scoring protocols, and dataset content; elicits the requirements induced by the use case, target population, and deployment environment; and compares these benchmark components against those requirements. The result is a component-level validity assessment that identifies which parts of the benchmark require filtering, modification, relabeling, supplementation, or replacement.

This procedure avoids collapsing benchmark applicability into a single global verdict. A benchmark may be well aligned on input form and poorly aligned on output ontology, or contain appropriate categories with culturally misaligned labels. It may be valid for literate urban users and poorly suited to speech-first or low-resource populations. The point of the framework is to make these distinctions explicit, so that regional actors can determine which claims remain supportable, which claims need to be narrowed, and where adaptation is required.

In practice, the framework can be used in two modes. As a screening tool, it helps actors assess whether an existing benchmark can be transferred to a target context with limited adaptation. As a design tool, it highlights where new categories, local examples, revised labels, or different interfaces are needed when adapting or building a benchmark. Section 4 operationalizes this procedure as an LLM-assisted workflow that gathers benchmark evidence, synthesizes deployment-context requirements, samples dataset content where available, and produces a structured validity and remediation report for each benchmark-context pair.

4. LLM Evaluation Pipeline

Manual assessment of benchmark validity is time-consuming and does not scale for actors evaluating many candidate benchmarks. We therefore operationalize the six indicators from Section 3 with an LLM-assisted pipeline that produces structured validity reports for benchmark-context pairs.

The pipeline (presented in detail in Figure 2) takes a **benchmark**, deployment **use case**, and **target population** as input, and builds the evidence needed to assess benchmark applicability. Specifically, it constructs structured files describing (1) the benchmark, (2) the needs of the deployment use case and target population, and (3) a qualitative assessment of specific dataset instances when data is available through Hugging Face. These files are then provided, together with the Section 3 rubric, to an LLM[2] to produce a structured report that scores benchmark applicability and recommends remediation steps (See Appendix D.1 for an example of a generated assessment).

Context for the automated assessment is iteratively aggregated from three primary sources of information and compared against the six compatibility indicators:

1. **Benchmark documentation:** An LLM extracts excerpts from the benchmark's paper PDF for use as evidence of validity components in the benchmark's documentation.

2. **Web Search:** Conditioned on the deployment use case and target population, a web search agent performs a search for region-specific information pertinent to assessing validity, guided by the cultural taxonomy from (Dev et al., 2026). A methodological concern with LLM-based evaluation is that models may themselves exhibit some of the cultural blind spots this framework is designed to detect. This step aims to fill gaps in the assessment LLM's parametric knowledge about the target culture and region.

3. **Data Analysis:** For text datasets available on Hugging Face, we provide an LLM with dataset samples to assess potential validity gaps stemming from the dataset content, which it uses to provide a qualitative assessment of dataset content.

The pipeline constructs these sources through the following modules (Figure 2):

1. **User Elicitation:** The agent takes the user inputted deployment description and generates a series of questions to further refine the parameters of the users use case and target population.

2. **Benchmark Extraction:** An agent is used to extract verbatim quotes related to the framework's 6 components from the benchmark's documentation. These are synthesized into a registry of evidence quotes.

3. **Regional Context:** Based on the refined user deployment context, an agent induces a structured file describing regional variables (e.g. literacy rates, regional norms, domain-specific regional taxonomies) which may be pertinent for determining validity gaps in the benchmark. These are refined and populated through agentic web search.

4. **Dataset Analysis:** Qualitatively analysis dataset examples for content validity violations.

A full specification of the pipeline, including the prompts and models used, may be found in Appendix D.2. Assessments cost $US\$4.03 \pm 1.19$ on average across 24 assessments conducted using the Anthropic API.

4.1. A Public Web Interface to the Pipeline

To broaden access to the pipeline beyond researchers comfortable with a local Python environment, we release a public browser-based interface that wraps Steps 0-7 described above.[3] A user supplies their own Anthropic API key, uploads the benchmark paper as a PDF, and writes a short free-text description of the deployment context they care about (e.g., *"an Indonesian government chatbot answering citizen queries about KTP applications"*). The same pipeline then runs in the browser, streaming live progress through each module and producing the structured validity report described above. Because each run draws on the user's own key, there is no central cost, sign-up, or shared quota; the key is held only in the browser and in server memory for the duration of a run, and is never written to disk or logs.

[2] We use Claude Opus 4.7 as the assessment generator. LLMs used for intermediate submodules are specified in Appendix D.2. We use default API sampling parameters for each model.

[3] https://aimslab.stanford.edu/benchmark-caliper/.

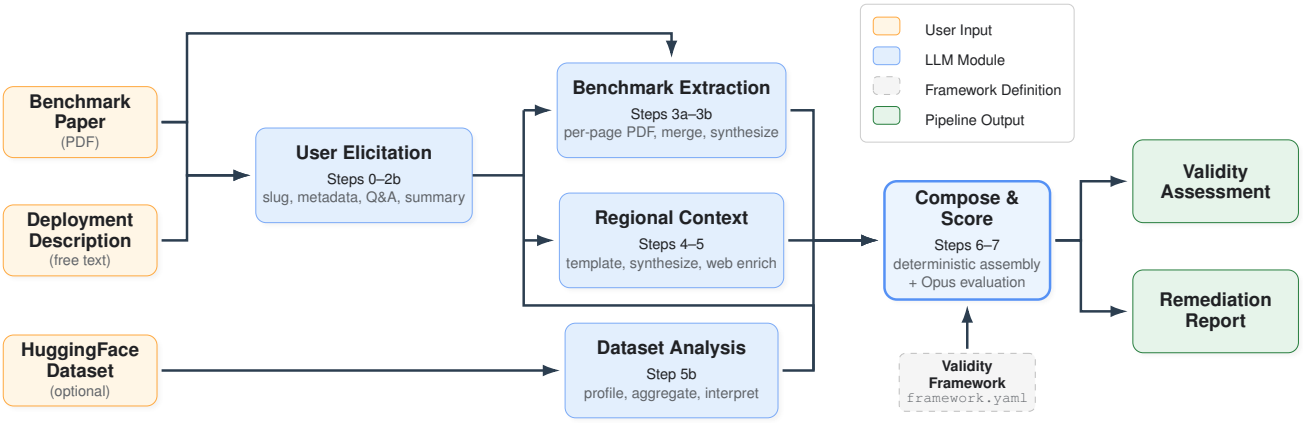


Figure 2. **Automated Pipeline Information Flow.** Three user-supplied inputs (benchmark paper, deployment description, and optionally a HuggingFace dataset) are processed through four evidence-production streams. User elicitation (Steps 0–2b) produces a deployment-context summary that feeds into benchmark extraction, regional context synthesis, and the final composition. Each stream generates one artifact that feeds into a deterministic composition step (Step 6), which merges all artifacts with the fixed validity framework into a single self-contained prompt. A single Opus call (Step 7) evaluates this prompt to produce the final validity assessment and remediation guidance.

The interface is intended for the actors most likely to need validity analysis but least likely to set up a research codebase: benchmark authors deciding whether to release a regional version, practitioners screening candidate benchmarks for a specific deployment, and domain experts contributing region-specific judgment through the human-in-the-loop elicitation step (Steps 2a-2b). The interface exposes both an automatic mode that runs the full pipeline unattended and a step-by-step mode that pauses between stages so the user can inspect intermediate artifacts (the extracted paper, the synthesized benchmark YAML, the regional context) before final scoring, making the pipeline's reasoning transparent rather than opaque. A curated gallery of expert-assessed benchmarks accompanies the tool, providing worked validity reports that visitors can browse without running the pipeline themselves. Appendix H describes the end-to-end user flow, the two modes, the gallery, the logging model, and implementation details.

5. Empirical Study

We validate our framework and its operationalization through an expert validation study with 6 regional domain experts (with regions of expertise comprising of Western Europe, Middle East and North Africa, Sub-Saharan Africa, South Asia, and Latin America) who serve as co-authors. Experts propose 2 benchmarks important for AI evaluation in their region and/or domain of expertise, supplying 2 deployment *use case* and *target population* pairings for each. Using the proposed benchmarks, we generate automated validity assessments via our LLM pipeline (Section 4). Experts review these assessments using Likert scales and free-text responses:

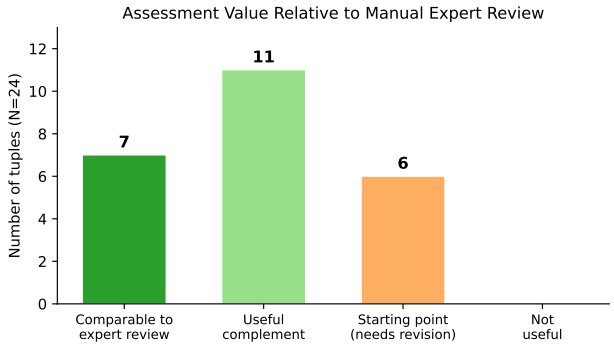
Are assessments useful? Experts report that 29% of assessments would match their own expert review and that 45% would complement it. 25% of assessments were reported as starting points for assessment in need of revisions (Figure 3a).

Do assessments teach experts something new? For 54% of assessments experts report the assessment surfacing at least one new concern for them, and several new concerns for 33% of assessments (Figure 3b).

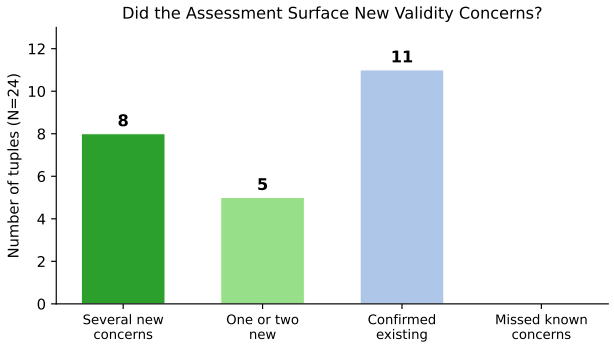
Are assessment scores calibrated? Experts report high agreement rates for relative ordering of scores within comparative assessments (Figure 3c left), and perceive that the pipeline scores leave appropriate room to grow were concerns raised by assessments to be addressed (Figure 3c right).

How correct are assessments? Experts reported scores as appropriate (M=4.28 ± 0.08) across dimensions, suggesting high precision for reported issues. Conversely, experts reported $29.9\% \pm 6.6$ of assessments as missing known validity concerns on average.

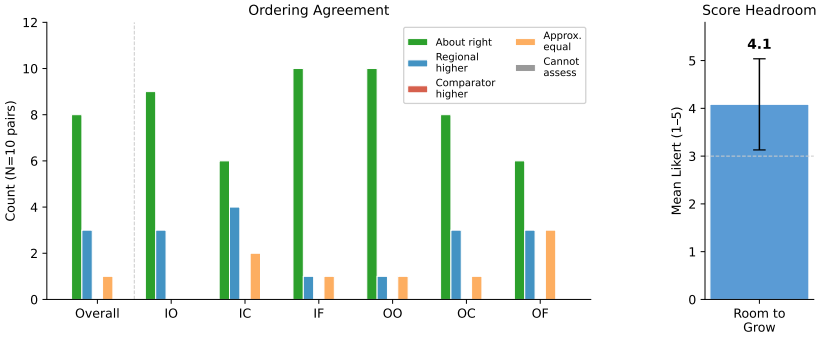
Together, these results suggest that our pipeline is useful for *expert uplift* on the task of benchmark validity assessments, where a domain expert can leverage assessments to complement and replace components of the manual benchmark validity review while retaining the expertise needed to identify when pipeline-raised concerns may be irrelevant or hallucinated.



(a) Assessment value relative to manual expert review.



(b) Whether the assessment surfaced new validity concerns.



(c) Expert agreement with comparative benchmark ordering and score headroom.

Figure 3. **Results.** Expert validation across 24 benchmark–use–population triplets evaluated by six domain experts. (a) and (b) show categorical assessments of pipeline output quality; (c) shows agreement with relative benchmark ordering across validity dimensions and perceived room for score discrimination.

6. Related Work

We situate our framework against five lines of related work along which differentiation is warranted. Additional background is in Appendix G

Validity-centered evaluation and benchmark design. A growing body of work imports validity theory from psychometrics and the social sciences into AI evaluation, including frameworks for measurement theory in ML (Jacobs & Wallach, 2021; Wallach et al., 2025), claim-centered validity assessment (Salaudeen et al., 2025), evidence-centered benchmark design (Liu et al., 2024), and large-scale construct validity audits (Bean et al., 2025; Ye et al., 2025). While these works establish a vital vocabulary for construct validity, they primarily treat benchmark design and auditing in isolation. Consequently, they do not assess whether a specific benchmark is applicable to a specific deployment defined by a use case and target population. Our framework operationalizes this relational question through a structured decomposition of validity failures across input and output components.

Structured documentation of evaluation artifacts. Datasheets (Gebru et al., 2021), Data Statements (Bender & Friedman, 2018), Model Cards (Mitchell et al., 2019), and BenchmarkCards (Sokol et al., 2026) provide standardized creator-supplied documentation for datasets, models, and benchmarks, while the Data Provenance Initiative (Longpre et al., 2023) scales lineage tracing across over 1,800 datasets. Winata et al. (2025) extend documentation toward measurable rubric-based quality assessment with LLM-assisted evaluation, but along generic data-quality dimensions rather than deployment fit. These artifacts are descriptive and artifact-centric: a single card exists per benchmark, and the applicability judgment is left to the downstream user. Our framework is complementary but evaluative: existing documentation serves as evidence input to our pipeline, which produces a per-component applicability assessment relative to a specified use case and target population.

Automated benchmark auditing. A rapidly growing line of work uses computational methods to audit existing benchmarks. Closest in spirit to our problem framing, Diddee et al. (2026) introduce BenchBrowser, a retrieval system that surfaces use-case-relevant items across benchmark suites to diagnose low content and convergent validity. BenchBrowser

shares our concern that benchmark metadata is too coarse to verify alignment with practitioner goals, but addresses it through retrieval over a fixed pair of validity dimensions, in English, and without a target-population axis. A parallel line of work scores benchmark quality along multiple dimensions: BetterBench (Reuel-Lamparth et al., 2024) evaluates 46 lifecycle criteria, the Benchmark Health Index (Zhu et al., 2026) ranks 106 benchmarks along capability discrimination, anti-saturation, and impact, and Benchmark^2 (Qian et al., 2026) combines cross-benchmark ranking consistency, discriminability, and capability alignment deviation. Kim et al. (2025) take a different angle, using gradient-based importance scoring and parameter ablation to decompose benchmark performance into cognitively grounded abilities and reveal mismatches between benchmark labels and the skills actually exercised. Other systems operate on narrower facets: translation quality auditing across languages (Thellmann et al., 2026), label error detection (Nahum et al., 2025), dataset filtering (Gupta et al., 2025b), and benchmarking of LLM-generated benchmarks (Zheng et al., 2026). These systems collectively demonstrate the feasibility of automated benchmark auditing, but each operates on the benchmark in isolation: they score, profile, or filter benchmarks against general-purpose quality dimensions rather than against a specified deployment context. Our framework provides the first systematic decomposition that integrates such audits into a deployment-fit assessment relative to a benchmark-use-population triplet, across input and output, ontology, content, and form.

Cultural benchmark audits. A substantial body of empirical work documents systematic cultural validity failures in AI evaluation. Oh et al. (2025) argue that cultural assumptions permeate even ostensibly neutral evaluation settings and call for intentionally cultural evaluation design, while several surveys and taxonomic frameworks address what cultural evaluation should cover and for whom (Wu et al., 2025b; Adilazuarda et al., 2024; Dev et al., 2026). Concrete case studies make the consequences vivid: Dey et al. (2025) show that a widely used rubric-based grader systematically penalized culturally appropriate, medically accurate responses for an Indian sexual and reproductive health deployment due to Western assumptions about legal framing and clinical norms, and Shahid et al. (2025) document how automated moderation pipelines for low-resource languages inherit colonial biases from their benchmark design. These studies establish the phenomenon our framework operationalizes: failures of the kind documented by Dey et al. (2025) and Shahid et al. (2025) are instances of output ontology, output content, and input content gaps in our decomposition, and our pipeline surfaces them automatically given a deployment specification, rather than requiring a one-off empirical audit for each new benchmark-population pair.

LLM-as-judge. Our pipeline uses LLMs as judges to extract evidence and produce structured assessments, building on a body of work that has both established the methodology (Zheng et al., 2023; Li et al., 2024) and critically examined its validity (Chehbouni et al., 2025; Feuer et al., 2025; Fu & Liu, 2025). We treat these limitations as known constraints on our evidence: by producing traceable structured outputs per validity dimension rather than aggregate scores, our framework enables practitioners to inspect and contest individual judgments where LLM-as-judge reliability is weakest, particularly in multilingual settings.

7. Conclusion

In this paper, we have presented a framework showing how the act of transferring benchmarks across deployment contexts and populations can result in violations of evaluation validity. This framework can be used by regional actors to characterize the dimensions over which to assess benchmarks for applicability to their regional contexts, as well as serve as a guide towards the construction of datasets which preserve evaluation validity in such contexts. Our LLM evaluation pipeline operationalizes this framework, serving as a tool for uplifting human expert validity analysis. Experts report that our pipeline can serve as a complement to expert assessment and can raise validity concerns not previously considered by them, presenting a tool which can be used to both serve as a filter and design guide for contextualized evaluation data.

7.1. Limitations

Our framework and pipeline focuses on benchmark data and documentation rather than evaluation metrics or LLM-as-judge configurations, which can also encode cultural assumptions. Although experts find the framework to be useful for supplementing their assessment, limitations were identified in generated assessments in the form of irrelevant issues raised and hallucinations, limiting the accessibility of our tool to non-expert populations who may not have enough expertise to appropriately discern assessment component importance or identify hallucinations. Although positive, our results come from a small sample size of experts (N=6) and assessments (N=24), with benchmarks assessed falling exclusively within NLP domains. Our pipeline also currently assumes frontier model LLM capability. Because we use models developed in Western contexts, our pipeline is susceptible to blind spots when cultural knowledge pertinent to a target population is missing from models' parametric knowledge and is not widely digitized (and thus not discoverable through the web search module).

7.2. Broader Impacts

AI development is rapidly advancing, presenting a need for fast adaptation to changes in the evaluation landscape. Automated frameworks for benchmark assessment present an opportunity for facilitating this adaptation, helping experts more quickly surface and address validity concerns in evaluations. Such a tool, however, is agnostic to the intended use case of evaluations. Better models can result from more calibrated evaluations, and it may be possible for bad actors to use tools such as this to optimize evaluations useful for surfacing harmful capabilities in AI systems, which itself could have harmful downstream consequences.

8. Acknowledgments

The authors would like to thank Eve Fleisig, Ketana Krishna, Zachary Robertson, and Trevor Darrell for their feedback on drafts of this paper. We would also like to thank David M. Chan and Lisa Dunlap for early conversations about benchmark data and evaluations.

RC was supported by a Ford Foundation Predoctoral fellowship; part of this work was conducted while he was a Summer Fellow at the Centre for the Governance of AI.

SK acknowledges support by NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, HAI, OpenAI, Microsoft, and Google.

References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL https://aclanthology.org/2021.tacl-1.66.

Adelani, D. I., Ojo, J., Azime, I. A., Zhuang, J. Y., Alabi, J. O., He, X., Ochieng, M., and Hooker, S. IrokoBench: A new benchmark for African languages in the age of large language models. *ArXiv preprint*, abs/2406.03368, 2024. URL https://arxiv.org/abs/2406.03368.

Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S. S., Aji, A. F., O'Neill, J., Modi, A., and Choudhury, M. Towards measuring and modeling "culture" in LLMs: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15763–15784, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.882. URL https://aclanthology.org/2024.emnlp-main.882/.

Bandalos, D. L. *Measurement theory and applications for the social sciences*. Guilford Publications, 2018.

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *ArXiv preprint*, abs/2308.16884, 2023. URL https://arxiv.org/abs/2308.16884.

Bean, A. M. et al. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 39, 2025. NeurIPS 2025 Track on Datasets and Benchmarks.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., and Mohamed, S. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–8, 2022.

Chehbouni, K., Haddou, M., Cheung, J. C., and Farnadi, G. Neither valid nor reliable? investigating the use of LLMs as judges. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=yqKfMr0yvY.

Choenni, R., Shutova, E., and van Rooij, R. CulturalBench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. *ArXiv preprint*, abs/2410.02677, 2024. URL https://arxiv.org/abs/2410.02677.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 3213–3223. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.350. URL https://doi.org/10.1109/CVPR.2016.350.

Davani, A., Díaz, M., Baker, D., and Prabhakaran, V. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2007–2021, 2024.

Dev, S., Prabhakaran, V., Feman, R. C., Davani, A., Denton, R., Kalia, C., Lertvittayakumjorn, P., Maji, M., Qadri, R., Rostamzadeh, N., et al. A unified framework to quantify cultural intelligence of ai. *ArXiv preprint*, abs/2603.01211, 2026. URL https://arxiv.org/abs/2603.01211.

Dey, S. K., S, M., Mehta, Z., Shah, M., Agrawal, U., Jalota, S., and Ismail, A. Beyond the rubric: Cultural misalignment in LLM benchmarks for sexual and reproductive health. In Akter, M., Chowdhury, T., Eger, S., Leiter, C., Opitz, J., and Çano, E. (eds.), *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pp. 126–134, Mumbai, India, 2025. Association for Computational Linguistics. ISBN 979-8-89176-305-0. doi: 10.18653/v1/2025.eval4nlp-1.11. URL https://aclanthology.org/2025.eval4nlp-1.11/.

Diddee, H., Yauney, G., Swayamdipta, S., and Ippolito, D. Benchbrowser: Retrieving evidence for evaluating benchmark validity. *ArXiv preprint*, abs/2603.18019, 2026. URL https://arxiv.org/abs/2603.18019.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., Ortega, J., Ramos, R., Rios, A., Meza Ruiz, I. V., Giménez-Lugo, G., Mager, E., Neubig, G., Palmer, A., Coto-Solano, R., Vu, T., and Kann, K. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6279–6299, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.435. URL https://aclanthology.org/2022.acl-long.435.

Feuer, B., Tseng, C.-Y., Lathe, A. S., Elachqar, O., and Dickerson, J. P. When judgment becomes noise: How design failures in llm judge benchmarks silently undermine validity. *ArXiv preprint*, abs/2509.20293, 2025. URL https://arxiv.org/abs/2509.20293.

Fleisig, E., Blodgett, S. L., Klein, D., and Talat, Z. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2279–2292, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.126.

Fu, X. and Liu, W. How reliable is multilingual LLM-as-a-judge? In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11040–11053, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.587. URL https://aclanthology.org/2025.findings-emnlp.587/.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., K, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J. J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M. A., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0eb

bf4b0-Abstract-Datasets_and_Benchmarks.html.

Gupta, A., O'Shea, D., and Barez, F. Val-bench: Measuring value alignment in language models. *ArXiv preprint*, abs/2510.05465, 2025a. URL https://arxiv.org/abs/2510.05465.

Gupta, V., Ross, C., Pantoja, D., Passonneau, R. J., Ung, M., and Williams, A. Improving model evaluation using SMART filtering of benchmark datasets. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4595–4615, Albuquerque, New Mexico, 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.235. URL https://aclanthology.org/2025.naacl-long.235/.

Hardalov, M., Mihaylov, T., Zlatkova, D., Dinkov, Y., Koychev, I., and Nakov, P. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5427–5444, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.438. URL https://aclanthology.org/2020.emnlp-main.438.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., and He, J. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 375–385, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *ArXiv preprint*, abs/2009.13081, 2020. URL https://arxiv.org/abs/2009.13081.

Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., and Lee, H. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524, 2024. doi: 10.1162/tacl_a_00661. URL https://aclanthology.org/2024.tacl-1.28.

Kim, D., Shim, G., Chun, Y., Kim, M., Park, C., and Lim, H. Benchmark profiling: Mechanistic diagnosis of LLM benchmarks. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15635–15650, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.789. URL https://aclanthology.org/2025.emnlp-main.789/.

Koto, F., Aisyah, N., Li, H., and Baldwin, T. IndoMMLU: Measuring massive multitask language understanding in Indonesia. *ArXiv preprint*, abs/2310.04928, 2023. URL https://arxiv.org/abs/2310.04928.

Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A. B., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., and Baldwin, T. ArabicMMLU: Assessing massive multitask language understanding in Arabic. *ArXiv preprint*, abs/2402.12840, 2024. URL https://arxiv.org/abs/2402.12840.

Larimore, S., Kennedy, I., Haskett, B., and Arseniev-Koehler, A. Reconsidering annotator disagreement about racist language: Noise or signal? In Ku, L.-W. and Li, C.-T. (eds.), *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pp. 81–90, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.socialnlp-1.7. URL

https://aclanthology.org/2021.socialnlp-1.7.

Lee, N., Jung, C., Myung, J., Jin, J., Camacho-Collados, J., Kim, J., and Oh, A. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4205–4224, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.236.

Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., and Baldwin, T. CMMLU: Measuring massive multitask language understanding in Chinese. *ArXiv preprint*, abs/2306.09212, 2023. URL https://arxiv.org/abs/2306.09212.

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., and Liu, Y. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *ArXiv preprint*, abs/2412.05579, 2024. URL https://arxiv.org/abs/2412.05579.

Liang, P. et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Liu, Y. L., Blodgett, S. L., Cheung, J., Liao, Q. V., Olteanu, A., and Xiao, Z. ECBD: Evidence-centered benchmark design for NLP. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.861. URL https://aclanthology.org/2024.acl-long.861/.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., and Hooker, S. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *ArXiv preprint*, abs/2310.16787, 2023. URL https://arxiv.org/abs/2310.16787.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6 (8):975–987, 2024.

Messick, S. Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2):5–11, 1989.

Miceli, M., Schuessler, M., and Yang, T. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Muhammad, S., Abdulmumin, I., Ayele, A., Ousidhoum, N., Adelani, D., Yimam, S., Ahmad, I., Beloucif, M., Mohammad, S., Ruder, S., Hourrane, O., Jorge, A., Brazdil, P., Ali, F., David, D., Osei, S., Shehu-Bello, B., Lawan, F., Gwadabe, T., Rutunda, S., Belay, T., Messelle, W., Balcha, H., Chala, S., Gebremichael, H., Opoku, B., and Arthur, S. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13968–13981, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.862. URL https://aclanthology.org/2023.emnlp-main.862.

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Pérez-Almendros, C., Ayele, A. A., Gutiérrez-Basulto, V., Ibáñez-García, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A., White, N., Yimam, S. M., Pilehvar, M. T., Ousidhoum, N., Camacho-Collados, J., and Oh, A. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8eb88844dafefa92a26aaec9f3acad93-Abstract-Datasets_and_Benchmarks_Track.html.

Nahum, O., Calderon, N., Keller, O., Szpektor, I., and Reichart, R. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26782–26809, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1360. URL

https://aclanthology.org/2025.emnlp-main.1360/.

Nicholas, G. Grounding ai policy: Towards researcher access to ai usage data. Technical report, Center for Open Science, 2024.

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., and Chalkidis, I. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3016–3054, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.200. URL https://aclanthology.org/2023.findings-emnlp.200.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., et al. No language left behind: Scaling human-centered machine translation. *ArXiv preprint*, abs/2207.04672, 2022. URL https://arxiv.org/abs/2207.04672.

Ogundepo, O., Gwadabe, T. R., Rivera, C. E., Clark, J. H., Ruder, S., Adelani, D. I., et al. AfriQA: Cross-lingual open-retrieval question answering for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11659–11681, 2023. URL https://aclanthology.org/2023.emnlp-main.718/.

Oh, J., Cha, I., Saxon, M., Lim, H., Bhatt, S., and Oh, A. Culture is everywhere: A call for intentionally cultural evaluation. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19156–19168, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1043. URL https://aclanthology.org/2025.findings-emnlp.1043/.

Olatunji, T., Afonja, T., Ajayi, B., Chaves, A., Deasi, C. T., et al. AfriMedQA: A pan-African medical question-answering benchmark dataset. *ArXiv preprint*, abs/2311.16116, 2023. URL https://arxiv.org/abs/2311.16116.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., and Korhonen, A. XCOPA: A multilingual dataset for causal commonsense reasoning. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185.

Qian, Q., Huang, C., Xu, J., Lv, C., Wu, M., Liu, W., Wang, X., Wang, Z., Huang, Z., Tian, M., Xu, J., Hu, K., Wang, H.-D., Hu, Y., Huang, X., and Zheng, X. Benchmark^2: Systematic evaluation of llm benchmarks. *ArXiv preprint*, abs/2601.03986, 2026. URL https://arxiv.org/abs/2601.03986.

Qin, C., Murali, S., Lee, E., Supramaniam, V., Hausenloy, D. J., Obungoloch, J., Brecher, J., Lin, R., Ding, H., Akudjedu, T. N., et al. Sustainable low-field cardiovascular magnetic resonance in changing healthcare systems. *European Heart Journal-Cardiovascular Imaging*, 23(6): e246–e260, 2022.

Reuel-Lamparth, A., Hardy, A. F., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. J. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/26889e8359e7ef8a7f5d77457364ca55-Abstract-Datasets_and_Benchmarks_Track.html.

Romanou, A., Foroutan, N., Sotnikova, A., Chen, Z., Montariol, S., Bosselut, A., et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. *ArXiv preprint*, abs/2410.06571, 2024. URL https://arxiv.org/abs/2410.06571.

Sadr, N. G., Heidariasl, S., Megerdoomian, K., Seyyed-Kalantari, L., and Emami, A. We politely insist: Your LLM must learn the Persian art of taarof. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1819–1838, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.94. URL https://aclanthology.org/2025.emnlp-main.94/.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., and Koyejo, S. Measurement to meaning: A validity-centered framework

for ai evaluation. *ArXiv preprint*, abs/2505.10573, 2025. URL https://arxiv.org/abs/2505.10573.

Schröder, T. and Schulz, M. Monitoring machine learning models: a categorization of challenges and methods. *Data Science and Management*, 5(3):105–116, 2022.

Servan-Mori, E., Torres-Pereda, P., Orozco, E., and Sosa-Rubí, S. G. An explanatory analysis of economic and health inequality changes among mexican indigenous people, 2000-2010. *International Journal for Equity in Health*, 13(1):21, 2014.

Shahid, F., Elswah, M., and Vashistha, A. Think outside the data: Colonial biases and systemic issues in automated moderation pipelines for low-resource languages. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3):2331–2344, Oct. 2025. doi: 10.1609/aies.v8i3.36719. URL https://ojs.aaai.org/index.php/AIES/article/view/36719.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ruder, S., Ko, W.-Y., Bosselut, A., Oh, A., Martins, A., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.919/.

Sokol, A., Daly, E. M., Hind, M., Piorkowski, D., Zhang, X., Moniz, N., and Chawla, N. V. Benchmarkcards: Standardized documentation for large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=b2IJBWhGFu.

Son, G., Lee, H., Kim, S., Kim, S., Muennighoff, N., Choi, T., Park, C., Yoo, K. M., and Biderman, S. KMMLU: Measuring massive multitask language understanding in Korean. *ArXiv preprint*, abs/2402.11548, 2024. URL https://arxiv.org/abs/2402.11548.

Susanto, Y., Hulagadri, A. V., Montalan, J. R., Ngui, J. G., Yong, X. B., Leong, W., Rengarajan, H., Limkonchotiwat, P., Mai, Y., and Tjhi, W. C. Sea-helm: Southeast asian holistic evaluation of language models. *ArXiv preprint*, abs/2502.14301, 2025. URL https://arxiv.org/abs/2502.14301.

Tamayo, R. L., Chulvi, B., and Rosso, P. Everybody hurts, sometimes overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter. *Procesamiento del lenguaje natural*, 71:383–395, 2023.

Taulé Delor, M., Ariza Casabona, A., Nofre, M., Amigó, E., and Rosso, P. Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del lenguaje natural, 2021, num. 67, p. 209-221*, 2021.

Thellmann, K., Stadler, B., and Färber, M. Diagnosing translated benchmarks: An automated quality assurance study of the eu20 benchmark suite. *ArXiv preprint*, abs/2604.01957, 2026. URL https://arxiv.org/abs/2604.01957.

Tonja, A., Balouchzahi, F., Butt, S., Kolesnikova, O., Ceballos, H., Gelbukh, A., and Solorio, T. NLP progress in indigenous Latin American languages. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6972–6987, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.385.

Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., and Jawahar, C. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1743–1751. IEEE, 2019.

Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Pangakis, N. J., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., and Jacobs, A. Z. Position: Evaluating generative AI systems is a social science measurement challenge. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82232–82251. PMLR, 2025. URL https://proceedings.mlr.press/v267/wallach25a.html.

Winata, G. I., Anugraha, D., Liu, E., Aji, A. F., Hung, S.-Y., Parashar, A., Irawan, P. A., Zhang, R., Yong, Z.-X., Cruz, J. C. B., Muennighoff, N., Kim, S., Zhao, H., Kar, S., Suryoraharjo, K. E., Adilazuarda, M. F., Lee, E.-S. A., Purwarianti, A., Wijaya, D. T., and Choudhury, M. Datasheets aren't enough: Datarubrics for automated quality metrics and accountability. *ArXiv preprint*, abs/2506.01789, 2025. URL https://arxiv.org/abs/2506.01789.

Wu, M., Wang, W., Liu, S., Yin, H., Wang, X., Zhao, Y., Lyu, C., Wang, L., Luo, W., and Zhang, K. The bitter lesson learned from 2,000+ multilingual benchmarks. *ArXiv preprint*, abs/2504.15521, 2025a. URL https://arxiv.org/abs/2504.15521.

Wu, M.-C., Chin, S.-C., Wood, T., Goyal, A., and Sadagopan, N. Incorporating diverse perspectives in cultural alignment: Survey of evaluation benchmarks through a three-dimensional framework. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17026–17061, Suzhou, China, 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.862. URL https://aclanthology.org/2025.emnlp-main.862/.

Xu, G., Liu, J., Yan, M., Xu, H., Si, J., Zhou, Z., Liu, Y., et al. CValues: Measuring the values of Chinese large language models from safety to responsibility. *ArXiv preprint*, abs/2307.09705, 2023. URL https://arxiv.org/abs/2307.09705.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://aclanthology.org/D19-1382.

Ye, H., Jin, J., Xie, Y., Zhang, X., and Song, G. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *ArXiv preprint*, abs/2505.08245, 2025. URL https://arxiv.org/abs/2505.08245.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al. fastmri: An open dataset and benchmarks for accelerated mri. *ArXiv preprint*, abs/1811.08839, 2018. URL https://arxiv.org/abs/1811.08839.

Zhang, W., Aljunied, M., Gao, C., Chia, Y. K., and Bing, L. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/117c5c8622b0d539f74f6d1fb082a2e9-Abstract-Datasets_and_Benchmarks.html.

Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., and Qiu, X. Evaluating the performance of large language models on GAOKAO benchmark. *ArXiv preprint*, abs/2305.12474, 2023b. URL https://arxiv.org/abs/2305.12474.

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Zheng, Y., Luo, H., Lin, Z., Liu, W., and Tuan, L. A. Benchbench: Benchmarking automated benchmark generation. *ArXiv preprint*, abs/2603.20807, 2026. URL https://arxiv.org/abs/2603.20807.

Zhu, L., Hua, H., Miao, L., and Zhao, B. Benchmark health index: A systematic framework for benchmarking the benchmarks of llms. *ArXiv preprint*, abs/2602.11674, 2026. URL https://arxiv.org/abs/2602.11674.

Ángel González, J., Borrego Obrador, I., Álvaro Romo Herrero, Sarvazyan, A. M., Chinea-Ríos, M., Basile, A., and Franco-Salvador, M. Iberbench: Llm evaluation on iberian languages. *Computer Speech & Language*, 96: 101899, 2026. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2025.101899. URL https://www.sciencedirect.com/science/article/pii/S088523082500124X.

.1. HumanEval

HumanEval measures whether an LLM can produce a correct, self-contained Python function from a natural-language docstring. Correctness is determined by automated test suites (pass/fail).

Triplet 1 (scope). This triplet varies the scope of the code generation task. The Great fit directly matches what HumanEval measures (generate a single Python function from

a specification). The Ok fit stays at the function level but shifts to test generation, which requires comprehension of existing code rather than generation from specification. The Bad fit expands to full-stack application scaffolding from requirements documents, which involves multiple languages, frameworks, and abstraction layers simultaneously.

> **Great:** We are a US developer tools company building an inline code completion feature. When an American software engineer writes a Python docstring or function signature in their IDE, our tool autocompletes the function body. We need to evaluate whether the LLM can generate correct, self-contained Python functions from natural language specifications. Correctness is verified by automated test suites.
>
> **Ok:** We are a US software company building a tool that automatically generates unit tests for existing Python functions. Given a function's signature, docstring, and source code, the tool produces test cases that exercise the function's behavior, including edge cases. Our users are American QA engineers. We need to evaluate the LLM's ability to comprehend existing code and produce meaningful tests.
>
> **Bad:** We are a US enterprise software company evaluating whether an LLM can generate full-stack web application scaffolding from high-level requirements documents. Given a requirements spec, the system should produce database schemas, API route definitions, frontend component stubs, and deployment configurations. Our users are American software architects. We need to evaluate the LLM's code generation capabilities for this multi-layer task.

Triplet 2 (output evaluation). This triplet varies the evaluation criterion while staying in the Python code domain. The Great fit mirrors HumanEval's evaluation: binary pass/-fail against automated tests. The Ok fit shifts to bug detection and repair, where correctness means identifying the right defect and suggesting a valid fix. The Bad fit moves to natural-language documentation generation from code, where the output is prose rather than executable code and correctness means factual accuracy of the documentation.

> **Great:** We are a US IDE vendor building a code suggestion feature for American developers. The tool proposes Python function implementations that are validated by running the user's test suite. We care about functional correctness: does the generated code compile, run, and pass all tests? We need to evaluate the LLM's ability to produce functionally correct Python code.
>
> **Ok:** We are a US code review platform building a feature that scans existing Python codebases, identifies likely bugs, and suggests fixes for American development teams doing code review. The system reads existing code, detects defects (logic errors, off-by-one bugs, null reference issues), and proposes corrections. We need to evaluate the LLM's ability to understand and reason about existing code.
>
> **Bad:** We are a US software company evaluating whether an LLM can generate accurate API documentation from Python source code. Given a Python module, the system should produce well-structured reference documentation: function descriptions, parameter explanations, return value specifications, and usage examples. Our audience is American developers consuming the documentation. We need to evaluate the LLM's ability to produce accurate technical prose from code.

.2. FOLIO

FOLIO presents premise–conclusion pairs expressed in controlled natural language and asks whether the conclusion follows from the premises (True), contradicts them (False), or is underdetermined (Unknown). The benchmark measures deductive validity checking; the task format is three-way classification.

Triplet 1 (reasoning type). This triplet varies the type of reasoning while staying within "logical analysis." The Great fit directly matches what FOLIO measures (checking deductive validity of arguments in controlled natural language). The Ok fit shifts to identifying logical inconsistencies in semi-structured technical prose, where premises may be implicit or ambiguous. The Bad fit moves to constructing persuasive business arguments, where logical soundness is one of several criteria alongside rhetorical strategy and audience modeling.

> **Great:** We are building an automated logic-checking tool for undergraduate formal reasoning courses at a US university. The tool receives a set of premises and a conclusion expressed in controlled natural language and determines whether the argument is logically valid, invalid, or indeterminate. Our users are American undergraduates taking introductory logic courses. We need to evaluate the LLM's deductive reasoning capabilities.
>
> **Ok:** We are a US engineering firm building a tool that identifies logical inconsistencies across technical specification documents for American systems engineers. The tool analyzes structured requirements and design documents to find contradictions—for example, requirements that contradict each other or conditions that can never be jointly satisfied. We need to evaluate the LLM's ability to perform logical analysis over semi-structured technical prose.
>
> **Bad:** We are a US consulting firm evaluating whether an LLM can help American management consultants build persuasive business cases. The system should assemble evidence, structure arguments, anticipate counterarguments, and tailor reasoning to executive audiences. We need to evaluate the LLM's reasoning capabilities for argument construction in a business context.

Triplet 2 (task structure). This triplet varies the output format and evaluation criterion while keeping the underlying

domain of logical argumentation constant. The Great fit mirrors FOLIO's task structure (classify a premise–conclusion pair with a categorical label; evaluate by accuracy). The Ok fit adds a generative component (produce step-by-step reasoning traces alongside the classification). The Bad fit moves to holistic evaluation of informal, real-world arguments in naturalistic discourse, where the task shares the word "argument" with formal logic but involves informal argumentation.

> **Great:** We are a US AI research lab evaluating LLM reasoning capabilities. We want to test whether the model can correctly classify the logical relationship (entailment, contradiction, or neutral) between a set of premises and a conclusion expressed in formal logical statements. Input is structured premise-conclusion pairs. Output is a categorical label. Our evaluation metric is classification accuracy.
>
> **Ok:** We are a US educational platform building a tool for American logic students. The tool should not only determine whether a logical argument is valid but also explain why—producing step-by-step reasoning traces that students can follow. We need to evaluate the LLM's ability to both perform logical reasoning and articulate its reasoning process in a pedagogically clear way.
>
> **Bad:** We are a US debate coaching platform evaluating whether an LLM can assess the strength of informal, real-world arguments made by American high school students in competitive debate contexts. Arguments involve empirical claims, value judgments, rhetorical moves, and evidence of varying quality. The system should assess argument strength holistically—including logical validity, evidence quality, relevance, and persuasive effectiveness.

.3. Simulated User Methodology

Running all 18 assessments with human respondents would be impractical and would introduce inter-respondent variability. Instead, for each assessment an Opus subagent answers the pipeline's elicitation questions in character as the practitioner described by the deployment description. The subagent receives the prompt shown in Figure 4, with three template variables filled per assessment: the deployment description, a persona context paragraph providing additional detail about the respondent's organizational role, and the elicitation questions generated by the pipeline from the benchmark paper. The pipeline then proceeds with the simulated answers through the remaining stages identically to a live run.

Simulated User (Opus)

```
You are a practitioner who submitted
↪  the following deployment
↪  description to a
benchmark validity analysis tool. The
↪  tool has generated follow-up
↪  questions to
better understand your deployment
↪  context. Answer them faithfully as
↪  this person
would.

Rules:
- Stay grounded in your deployment
↪  context. Describe your actual
↪  system, users,
  and requirements.
- Do not volunteer information about
↪  the benchmark itself | you have not
↪  read
  the benchmark paper.
- Do not artificially inflate or
↪  minimize concerns. Answer
↪  naturally.
- Keep answers concise (2-4 sentences
↪  per question). Provide concrete
↪  details
  where you can.

## Your Deployment Description

{deployment_description}

## Additional Context About Your Role
↪  and Setting

{persona_context}

## Elicitation Questions

{questions_json}

## Output Format

Return a JSON array with one entry per
↪  question. Use the EXACT question ID
from each question's "id" field (Q1, Q2,
↪  Q3, etc.) | do NOT substitute
dimension codes or any other
↪  identifier.

[
  {"id": "Q1", "answer": "..."},
  {"id": "Q2", "answer": "..."},
  {"id": "Q3", "answer": "..."}
]

Output ONLY the JSON array. No prose,
↪  no code fences.
```

Figure 4. Simulated user prompt template. Template variables ({deployment_description}, {persona_context}, {questions_json}) are filled per assessment.

.4. Per-Triplet Scores

Tables 1–3 report the per-dimension validity scores (1–5 scale) for each triplet and fit level. Dimension abbreviations:

IO = Input Ontology, IC = Input Content, IF = Input Form, OO = Output Ontology, OC = Output Content, OF = Output Form. Scores are produced by the pipeline's Opus-based scoring stage.

Table 1. Per-dimension validity scores for GSM8K triplets.

Triplet	Fit	IO	IC	IF	OO	OC	OF	Mean
Depth	Great	3	4	5	5	4	5	4.3
	Ok	1	1	4	1	2	3	2.0
	Bad	1	1	2	1	1	2	1.3
Format	Great	2	4	5	2	3	3	3.2
	Ok	2	2	3	1	1	2	1.8
	Bad	2	2	4	1	2	1	2.0

Table 2. Per-dimension validity scores for HumanEval triplets.

Triplet	Fit	IO	IC	IF	OO	OC	OF	Mean
Scope	Great	2	1	5	2	3	5	3.0
	Ok	1	1	3	1	1	2	1.5
	Bad	1	2	1	1	1	2	1.3
Output eval	Great	2	2	5	3	4	5	3.5
	Ok	1	1	3	1	1	2	1.5
	Bad	1	1	4	1	1	1	1.5

Table 3. Per-dimension validity scores for FOLIO triplets.

Triplet	Fit	IO	IC	IF	OO	OC	OF	Mean
Reasoning	Great	2	2	4	4	3	4	3.2
	Ok	1	1	3	1	2	1	1.5
	Bad	1	1	4	1	2	1	1.7
Structure	Great	5	4	3	5	4	4	4.2
	Ok	3	3	4	1	2	1	2.3
	Bad	1	1	4	1	1	1	1.5

A. Manual Case Study Analyses

A.1. Individual Analyses

We apply our framework to three benchmarks to illustrate its diagnostic utility. We analyze the original HELM benchmark (Liang et al., 2023) for Southeast Asian contexts, contrast it with SEA-HELM (Susanto et al., 2025) as a validity-preserving regional adaptation, and assess IberBench (Ángel González et al., 2026) for Iberian language regions. Full dimension-by-dimension analyses are provided in Appendix A.

HELM: Validity Violations in Southeast Asian Contexts Applying our checklist to the original HELM benchmark reveals validity violations across all six framework dimensions when considered for Southeast Asian (SEA) contexts. On the input side, HELM lacks evaluation categories for SEA cultural norms (e.g., Filipino *pakikisama*, Thai *kreng jai*), constituting construct underrepresentation. Analysis of MMLU, a core component, reveals that 86% of culturally sensitive items require Western-specific knowledge (Singh et al., 2025), introducing construct-irrelevant variance. Furthermore, tasks assuming Latin script conventions (e.g., capitalization-based assessments) fail for Thai, Khmer, and Burmese writing systems. On the output side, label schemes encode Western value frameworks, annotations were produced by non-SEA annotator pools—with cross-cultural studies showing only 56% agreement on tasks like hate speech detection (Lee et al., 2024)—and text-based evaluations do not penalize language mismatches. These pervasive gaps motivated the development of SEA-HELM as a validity-preserving regional adaptation.

SEA-HELM: A Validity-Preserving Regional Adaptation SEA-HELM (Susanto et al., 2025) addresses the majority of validity violations identified in HELM through deliberate regional adaptation. The benchmark introduces region-specific evaluation pillars (e.g., *SEA Culture*, *Safety*), prioritizes native-language data with native speaker verification over machine translation, and explicitly adapts prompts for non-Latin scripts. On the output side, label schemes reflect regional values, annotations are produced or verified by native speakers from target populations, and language mismatches are explicitly penalized. SEA-HELM further employs all three remediation strategies: *filtering* inapplicable prompts, *modifying* tasks for local contexts, and *developing* new datasets such as Kalahi and LINDSEA. However, a residual validity gap remains: the benchmark is text-only, which may limit evaluation validity for lower-literacy populations (literacy rates range from 99% in Brunei to 68% in Timor-Leste) and speakers of the region's 1,000+ indigenous languages who may primarily use speech-based interfaces.

IberBench: Iberian-Sourced Language Understanding Evaluations IberBench (Ángel González et al., 2026) aggregates 101 datasets covering 6 Iberian languages and 5 Ibero-American Spanish dialects, targeting *general Iberian language understanding ability* as its construct. Its explicit sourcing from Iberian workshop tasks (58 of 101 tasks) suggests strong regional relevance for input content, and the use of Latin script with high regional literacy rates makes text a natural fit for both input and output form. However, the benchmark presents an asymmetric validity profile. On input ontology, the authors note high imbalance across language-task pairs, and the 43 general tasks sourced from broader evaluations may dilute regional specificity. More critically, IberBench does not document the relevance of output taxonomies for Iberian contexts—our analysis of subtasks such as HUHU (Tamayo et al., 2023) and DETOXIS (Taulé Delor et al., 2021) finds that category selections are not substantiated in relation to the region. Annotator demographics are also inconsistently documented across subtasks. While a

Table 4. Validity scores (1–5 scale) for each benchmark across framework dimensions, with justifications derived from case study analyses. **1** = major violations; **5** = no concerns.

Dimension	HELM	SEA-HELM	IberBench	Key Justification
Input Ontology	2	5	3	HELM lacks SEA-specific categories (e.g., *pakikisama*, *kreng jai*). SEA-HELM adds region-specific pillars (SEA Culture, Safety). IberBench has 22 categories but high imbalance across language pairs.
Input Content	1	5	4	HELM: 86% of culturally sensitive items are Western-specific. SEA-HELM uses native-language data with speaker verification. IberBench sources from regional workshops and Iberian authors.
Input Form	1	4	5	HELM assumes Latin script (fails for Thai, Khmer, Burmese). SEA-HELM adapts for non-Latin scripts. IberBench: Latin alphabet naturally matches all target languages.
Output Taxonomy	2	4	2	HELM encodes Western value frameworks. SEA-HELM uses regional labels. IberBench does not document taxonomy relevance for Iberian contexts.
Output Labels	2	5	3	HELM annotators do not represent SEA populations (56% cross-cultural agreement). SEA-HELM uses native speaker annotators. IberBench: mixed documentation of annotator demographics.
Output Form	2	3	5	HELM does not penalize language mismatches. SEA-HELM penalizes mismatches but remains text-only (gap for low-literacy populations). IberBench: text appropriate given high literacy rates.

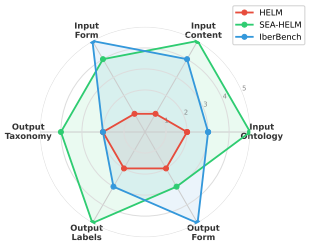


Figure 5. Radar chart comparing validity scores across the six framework dimensions for HELM (applied to SEA contexts), SEA-HELM, and IberBench. Larger area indicates stronger overall validity preservation. HELM shows uniformly low validity; SEA-HELM achieves broad coverage with a residual gap in output form; IberBench excels on form dimensions but lacks output taxonomy documentation.

committee of Iberian NLP specialists maintains the benchmark through an expert review process, many critical design decisions remain unjustified in the documentation, making it difficult to fully assert validity beyond expectations based on sourcing provenance.

A.2. Comparative Validity Assessment

To summarize the case study findings, we assign each benchmark a validity score on a 1–5 scale for each framework dimension, where 1 indicates major validity violations and 5 indicates no concerns identified. Table 4 reports these scores with brief justifications derived from our analyses, and Figure 5 visualizes the resulting validity profiles.

The radar chart reveals three distinct patterns. HELM, when applied to Southeast Asian contexts, exhibits uniformly low validity across all dimensions—its small footprint reflects systemic violations stemming from its Global North origins. SEA-HELM achieves high validity across most dimensions through deliberate regional adaptation, though the text-only output modality remains a residual gap for lower-literacy populations. IberBench presents an asymmetric profile: strong on form-related dimensions (where Latin script and high literacy rates provide natural alignment) but weaker on output taxonomy documentation and input ontology balance.

B. Extended Framework Details

The main paper presents the framework as a compact diagnostic for benchmark-context pairs. This appendix records the longer definitions, examples, evidence sources, and remediation guidance behind the six compatibility indicators.

The boundaries between ontology, content, and form can overlap in practice. Language functions as form when the concern is script, encoding, or accessibility, and as content when the concern is the cultural or semantic assumptions expressed by an item. Likewise, a scoring criterion can appear as output ontology when it defines the categories through which outputs are interpreted and as output content when it assigns judgments to particular responses. We therefore locate a problem at the component where remediation should occur.

Input ontology. **Definition.** Input ontology concerns the set of input categories represented in the benchmark. These categories may be task types, domains, skills, scenarios, environmental conditions, or other partitions of the input space. This indicator captures category-level coverage of the target use case and population. Input ontology failures primarily threaten content validity (Messick, 1989): construct underrepresentation occurs when locally relevant categories are absent from the benchmark, while construct-irrelevant variance occurs when the benchmark includes categories that are irrelevant to the deployment claim and still affect the score.

Contextual relevance. When benchmarks are ported across regions, input ontology failures often arise because the source benchmark reflects the categories salient to its original designers. For Persian commerce evaluations, the Iranian norm of *taarof* (Sadr et al., 2025) may be a relevant interaction category and is largely absent from general-purpose NLP benchmarks. For autonomous driving in South Asia, categories such as autorickshaws, informal road boundaries, mixed-use road surfaces, and non-lane-based traffic behavior may be essential to the deployment use case. Benchmarks developed under European or North American road assumptions often exclude these categories. A benchmark intended to evaluate Mexican history knowledge should exclude U.K. history categories inherited from the source benchmark.

Evidence and remediation. Evidence for input ontology alignment can come from benchmark task documentation, dataset category distributions, deployment logs, stakeholder elicitation, domain standards, or expert review. Remedia-

tion typically involves supplementing missing categories, filtering irrelevant categories, or redesigning the category taxonomy before collecting new data.

Input content. **Definition.** Input content concerns the properties of individual benchmark instances, including the concepts, entities, references, assumptions, contexts, and background knowledge required to interpret or solve a datapoint. This indicator captures whether individual inputs instantiate the intended category in a construct-relevant way for the deployment context. Input content failures can occur even when the benchmark ontology is appropriate. A math benchmark may include algebra problems, yet an individual algebra problem may still require chemistry background, familiarity with a foreign grading system, or knowledge of a culturally specific idiom. These failures threaten content validity (Messick, 1989) and, when they induce systematic performance differences across deployment populations, external validity. **Contextual relevance.** A common input-content failure in cross-regional benchmark transfer is cultural sensitivity. In this failure mode, an item requires cultural, geographic, institutional, historical, or dialectal knowledge outside the intended construct for the target deployment context (Singh et al., 2025). Approximately 28% of MMLU (Hendrycks et al., 2021) examples are culturally sensitive, and roughly 86% of those require Western-specific knowledge (Singh et al., 2025). The degree of sensitivity varies substantially across domains, with relatively little in math and physics and much more in history and geopolitics. Since leading benchmarks have predominantly been developed within frontier states (Longpre et al., 2024), this dimension warrants careful consideration. Cultural sensitivity is desirable when the benchmark is meant to evaluate locally grounded cultural competence; it becomes construct-irrelevant when it is inherited from the source context and unrelated to the deployment claim. Input content mismatches also include jurisdiction-specific legal assumptions, region-specific medical practices, local units or measurements, named entities unfamiliar to the target population, socioeconomic assumptions, temporal assumptions, and examples sampled from institutions that do not exist in the deployment setting. **Evidence and remediation.** Evidence for input content alignment can come from item-level audits, cultural sensitivity annotation, expert review, dataset sampling analysis, or comparison against target-region data. Remediation typically involves filtering problematic instances, rewriting them for the target context, adding locally sourced instances, or stratifying benchmark scores by content subtype.

Input form. **Definition.** Input form concerns the representation of benchmark inputs, including modality, language, script, file format, signal quality, sensor properties, interface assumptions, and other features of the input channel. This indicator captures whether benchmark inputs are presented in the forms that models will encounter in the deployment environment. Input form failures primarily threaten external validity: a model can perform well on a benchmark because the benchmark's input representation is easier, cleaner, or different from the signals encountered in deployment, and performance may transfer poorly when the signal characteristics differ (Saenko et al., 2010). **Contextual relevance.** Input form varies substantially across regional and institutional contexts. In medical imaging, roughly 75% of MRI machines in West Africa operate below 1.5 Teslas (Qin et al., 2022), while open-source MRI benchmarks such as NYU fastMRI contain images at 1.5 and 3 Teslas (Zbontar et al., 2018). A model evaluated only on high-field MRI benchmarks may therefore be poorly calibrated to lower-field clinical environments. In NLP, a benchmark administered only in standardized written text may fail to evaluate systems deployed through speech, code-switching, non-standard orthography, non-Latin scripts, or low-bandwidth interfaces. **Evidence and remediation.** Evidence for input form alignment can come from benchmark file formats, sensor metadata, language and script documentation, deployment interface descriptions, and target-environment signal statistics. Remediation can involve translation, transliteration, modality transfer, collecting new data under target signal conditions, or explicitly restricting benchmark claims to the benchmark's input form.

Output ontology. **Definition.** Output ontology concerns the categories through which model outputs are interpreted before they are scored. For classification benchmarks, this is the label set. For generative benchmarks, it may be rubric dimensions, acceptable response types, preference criteria, safety categories, or an action space used by human or automated evaluators. This indicator captures whether the benchmark's output categories and scoring criteria reflect the decision structure appropriate for the deployment use case and target population. Output ontology failures threaten structural validity when the benchmark misrepresents the structure of the construct, content validity when relevant output categories are missing or irrelevant categories are included, and external validity when the resulting scores transfer poorly to target-context decisions. **Contextual relevance.** Label and rubric taxonomies encode political, cultural, institutional, and domain-specific assumptions (Miceli et al., 2020). In autonomous driving, CityScapes (Cordts et al., 2016) defines labels such as `road`, `parking`, and `sidewalk` under structured European road assumptions. The Indian Driving Dataset (Varma et al., 2019) extends this taxonomy with labels for ambiguous surfaces such as `drivable fallback` and `non-drivable fallback`, as well as traffic participants such as `autorickshaws`, reflecting Indian road conditions. Similar failures arise in LLM safety or helpful-

ness benchmarks when the rubric lacks categories for locally relevant harms, prohibited recommendations, or culturally specific interaction norms. **Evidence and remediation.** Evidence for output ontology alignment can come from label-set documentation, rubric definitions, scoring prompts, expert review, stakeholder consultation, and comparison to target-region standards or taxonomies. Remediation typically involves redesigning labels, adding rubric dimensions, revising acceptable response classes, or changing the scoring criterion to reflect target-context decision rules.

Output content. **Definition.** Output content concerns the labels, reference answers, annotations, preferences, or judgments attached to particular benchmark instances. Output ontology asks whether the benchmark uses the right categories; output content asks whether the assignments for individual datapoints are appropriate for the deployment context. This indicator captures whether benchmark judgments agree with the judgments of target-population annotators, regional stakeholders, or domain experts. Output content failures threaten convergent validity when benchmark labels diverge from the judgments they are intended to approximate, and external validity when source-context labels transfer poorly to target-context users, institutions, or communities. These failures are especially likely when benchmark labels encode social norms, values, offensiveness, helpfulness, safety, legality, or acceptability. **Contextual relevance.** Values and judgments differ across regional and cultural boundaries, and so can ground-truth labels. Annotator judgments vary based on geo-cultural differences (Davani et al., 2024). In hate-speech detection, Lee et al. (2024) find that annotators from the UK, US, Australia, South Africa, and Singapore reach unanimous agreement on only 56% of examples, with stronger pairwise agreement among Anglosphere countries. More broadly, annotator judgments vary along lived-experience axes (Larimore et al., 2021), and standard aggregation methods can wash away minoritized perspectives (Fleisig et al., 2024). If the annotator population does not represent the intended target population, benchmark labels may fail to capture the judgments most relevant to the evaluation claim. **Evidence and remediation.** Evidence for output content alignment can come from annotator demographics, re-annotation studies, inter-annotator agreement between source and target annotators, expert adjudication, perspectivist labeling, or disagreement analysis. Remediation can involve relabeling with target-population annotators, preserving multiple valid perspectives, changing aggregation procedures, or reporting subgroup-specific scores when disagreement is itself substantively meaningful.

Output form. **Definition.** Output form concerns the representation of model outputs that the benchmark accepts, evaluates, or rewards. This includes modality, language, script, formatting constraints, structured versus free-form output, interaction style, and the kinds of responses expected in deployment. This indicator captures whether the benchmark evaluates models on the output forms required by the deployment environment. Output form failures primarily threaten external validity: a model may appear capable under a benchmark that only requires written text, single-turn answers, or a particular structured format, yet fail in a deployment environment requiring spoken output, multilingual interaction, code-switching, multimodal explanation, low-literacy interfaces, or action-oriented responses. **Contextual relevance.** Output form mismatches arise when the expected modality of use changes across regions or populations. For instance, higher illiteracy rates among indigenous communities in Mexico (Servan-Mori et al., 2014) may favor speech-based NLP systems, yet text-to-speech technologies remain unavailable for many indigenous languages (Tonja et al., 2024). In such contexts, text-output benchmarks are poorly calibrated to the forms of assistance users actually need. Similar mismatches arise when a benchmark evaluates medical systems only on diagnosis labels and deployment requires referral instructions, when a customer-service benchmark evaluates English responses and users expect code-switched speech, or when a benchmark penalizes formatting deviations that are irrelevant to the deployment task. **Evidence and remediation.** Evidence for output form alignment can come from benchmark output specifications, scoring scripts, deployment interface requirements, user literacy and accessibility constraints, and analysis of accepted response formats. Remediation can involve adding speech, multimodal, or action-based outputs, accepting target-language or code-switched responses, modifying format constraints, or creating benchmark variants for distinct deployment interfaces.

Assessment procedure. Given a benchmark-context pair (B, D), the assessor specifies the deployment claim, extracts benchmark properties, elicits context requirements, compares each benchmark component against those requirements, identifies validity threats, and recommends remediation. The output is a component-level validity assessment that identifies which benchmark components require filtering, modification, relabeling, supplementation, or replacement. This avoids collapsing benchmark applicability into a single global judgment: a benchmark may be well aligned on input form and poorly aligned on output ontology, or valid for literate urban users and poorly suited to speech-first or low-resource populations.

C. Practitioner's Checklists

The following checklists accompany each component of the framework described in Section 3. Regional actors can use these to both *filter* existing benchmarks for regional

applicability and *guide* the development of new benchmarks.

Checklist: Input Ontology

☐ Identify test case categories required for regional deployment.
☐ Check if source benchmark's taxonomy omits regionally-relevant categories.
☐ Check if source benchmark includes categories irrelevant to regional context.
☐ Document category gaps that would harm *content validity*.

Checklist: Input Content

☐ Determine if input queries require region-specific cultural, geographic, or dialectal knowledge.
☐ Assess whether culturally sensitive content aligns with target deployment culture.
☐ Flag inputs requiring Western-specific knowledge that may not transfer.
☐ Recruit regional annotators to identify culturally sensitive instances if resources permit.
☐ Document content issues that would harm *content validity*.

Checklist: Input Form

☐ Compare signal distributions (e.g., image resolution, MRI field strength) between source and target contexts.
☐ Check if regional infrastructure supports the same data capture specifications.
☐ Identify domain-specific form differences relevant to the intended use case.
☐ Document form mismatches that would harm *external validity*.

Checklist: Output Ontology

☐ Review output label categories for regional relevance.
☐ Identify missing categories specific to regional contexts (e.g., autorickshaws in Indian driving data).
☐ Flag categories that encode non-regional values or assumptions.
☐ Consider stakeholder-driven taxonomy redesign if significant misalignment exists.
☐ Document taxonomy issues that would harm *structural validity* and *content validity*.

Checklist: Output Content

☐ Determine if ground truth labels reflect regional stakeholder perspectives.
☐ Assess potential disagreement between original annotators and regional population.
☐ Review annotator demographics in benchmark documentation (Datasheets, Data Statements).
☐ Consider label re-annotation by representative regional annotator pool.
☐ Review aggregation methods for potential erasure of minority perspectives.
☐ Document label issues that would harm *convergent validity* and *external validity*.

Checklist: Output Form

☐ Check if expected output modality matches regional deployment needs.
☐ Assess text-to-speech availability for speech-based output requirements.
☐ Consider literacy rates and accessibility requirements in target population.
☐ Document form mismatches that would harm *external validity*.

Checklist: Document Validity Violation Risks

☐ Map identified issues to specific validity types (Content, Structural, External, Convergent, Consequential).
☐ Assess severity of each violation for intended regional use case.
☐ Prioritize violations based on risk tolerance of application domain (e.g., healthcare vs. creative writing).
☐ Document the proportion of benchmark data affected by each violation type.

Checklist: Determine Remediation Strategy

☐ For each violation, determine action: **Filter** (remove), **Modify** (adapt), or **Develop** (create new)
☐ For filtering: establish criteria for datapoint exclusion
☐ For modification: identify translation, re-labeling, or re-annotation requirements
☐ For development: scope new data collection needs based on remaining gaps
☐ Document trade-offs between cost, coverage, and validity for each approach

D. Automated Validity Assessment Pipeline

D.1. Example Assessment

Below we present an example of an assessment produced by our LLM pipeline during the expert validation experiments.

width=!,height=!,pages=-

D.2. LLM Judge Prompt Format

Below we present all of the prompts used in our automated pipeline.

Step 0: Slug Derivation (Haiku)

```
Produce a short slug (2-5 words,
↪  lowercase, underscore-separated)
↪  that describes
this deployment's target population.
↪  Use geographic, linguistic, and
↪  use-case
cues from the deployment description
↪  provided in the user message.

Examples:
- `nigeria_hausa_rural_clinic`
- `iberian_portuguese_legal`
- `vietnam_urban_ecommerce`
- `kenya_chw_swahili_dholuo`

Return ONLY the slug. No quotes, no
↪  prose, no code fences, no trailing
↪  newline
content.

---
User message:

{DEPLOYMENT_DESCRIPTION}
```

Figure 6. Assessment slug derivation prompt. Haiku produces a short, filesystem-safe slug from the free-text deployment description.

Step 1: Metadata Extraction (Haiku)

```
You are extracting lightweight metadata
↪  from the first 1-2 pages of a
↪  benchmark
paper (title page, abstract,
↪  introduction). A PDF containing only
↪  these pages is
attached to this message as a document
↪  block.

The goal is to gather just enough
↪  context for elicitation | not to
↪  perform full
extraction. Downstream steps will
↪  process the entire paper later.

Extract these fields. If a field cannot
↪  be determined from the available
↪  pages,
write "UNKNOWN".

Output as a structured text block with
↪  exactly this format:

- name: [short lowercase identifier, no
↪  spaces | e.g. "sea-helm", "afriqa",
↪  "culturebench"]
- full_name: [full benchmark title as
↪  written in the paper]
- year: [publication year]
- domain: [what the benchmark evaluates,
↪  1-2 sentences]
- languages: [comma-separated list of
↪  languages mentioned]
- primary_region: [the
↪  geographic/cultural region the
↪  benchmark targets]
- porting_strategy: [one of: ground_up,
↪  adapted, mixed, translation,
↪  parallel, regional_exams, none]
- input_modalities: [comma-separated
↪  list of modalities the benchmark
↪  uses as input | e.g. text, image,
↪  audio, video, code,
↪  structured_data]
- output_modalities: [comma-separated
↪  list of modalities the benchmark
↪  evaluates as output | e.g. text,
↪  label, score, translation, code]
- brief_description: [2-3 sentence
↪  summary of what this benchmark does
↪  and how it was built]
- source_culture: [one of: designed by
↪  target population, transferred from
↪  different cultural context, mixed,
↪  unclear]

Porting strategy definitions:
- ground_up: benchmark was built from
↪  scratch for the target
↪  region/culture
- adapted: an existing benchmark was
↪  modified for a new region (not just
↪  translated)
- mixed: some tasks are original,
↪  others are adapted or translated
- translation: an existing benchmark
↪  was directly translated
- parallel: parallel versions were
↪  created for multiple regions
↪  simultaneously
- regional_exams: benchmark draws from
↪  region-specific exams (e.g.
↪  university entrance exams)
- none: no cross-cultural porting
↪  involved (single-region benchmark)
```

```
Output ONLY the structured text block,
↪  with no surrounding prose or code
↪  fences.

---
User message:

Extract metadata from the attached
↪  benchmark paper (pages 1-2).

[Pages 1-2 of the benchmark paper PDF
↪  are attached as a document block.]
```

Figure 7. Paper metadata extraction prompt. Haiku extracts lightweight metadata (title, authors, domain, languages) from the first two pages.

Step 2a: Elicitation Questions (Sonnet)

```
# Elicitation: Generate Questions

## Task

You are the elicitation agent in a
↪  validity-analysis pipeline. Given
↪  lightweight benchmark metadata and
↪  a short deployment description,
↪  produce **3-5 targeted follow-up
↪  questions** that close the gap
↪  between what the benchmark
↪  documents and what the deployment
↪  actually requires. Prefer 3
↪  high-signal questions. Only add a
↪  4th or 5th if they surface a clearly
↪  distinct validity concern not
↪  already covered by the first three.

The user message contains:
1. **Lightweight benchmark metadata** |
↪  basic information extracted from
↪  the paper's first 1-2 pages: name,
↪  domain, languages, primary region,
↪  porting strategy, source culture.
↪  This is NOT the full benchmark YAML;
↪  detailed documentation excerpts and
↪  verbatim quotes are not yet
↪  available at this stage.
2. The **deployment description** |
↪  what the AI system does, who it
↪  serves, where.

## Division of Labor

- **The user** is the expert on their
↪  deployment: who the users are, what
↪  languages and dialects they speak,
↪  what cultural knowledge matters,
↪  what output format is useful, what
↪  infrastructure they operate on.
- **The pipeline** is the expert on the
↪  benchmark: its taxonomy, its data
↪  sources, its annotation process |
↪  all derivable from the full PDF
↪  (which will be extracted later).

Your question selection should rely on
↪  the use case and the lightweight
↪  metadata, not on detailed benchmark
↪  content. Never ask the user about
↪  the benchmark's internals. Only ask
↪  about things the pipeline cannot
↪  determine from the benchmark alone.
↪  If a question requires benchmark
↪  knowledge to answer, it belongs in
↪  the assessment stage, not
↪  elicitation.

## Validity Dimensions

Every question you generate must map to
↪  exactly one of these six
↪  dimensions.

### Input Ontology (IO)
The set of test-case categories the
↪  benchmark covers | the query types
↪  evaluated systems are expected to
↪  encounter. Misalignment (missing
↪  categories, or irrelevant ones)
↪  harms content validity.

### Input Content (IC)
The concrete instances represented by
↪  individual datapoints (a prompt, an
↪  image, an entry). Even when the
↪  ontology is right,
↪  implementation-level choices in
↪  individual datapoints can introduce
↪  construct-irrelevant variance |
↪  e.g., implicit US-default
↪  assumptions.

### Input Form (IF)
The encoding of the input signal | text
↪  vs audio, camera specs,
↪  infrastructure. A mismatch between
↪  benchmark signal distribution and
↪  real-world deployment signal
↪  distribution violates external
↪  validity.

### Output Ontology (OO)
The space of outputs plus the decision
↪  rules that score them. For
↪  free-form outputs, the scoring
↪  function's interpretive step is
↪  where cultural misalignment most
↪  readily arises. Violations touch
↪  structural, content, and external
↪  validity.

### Output Content (OC)
```

```
Label correctness | whether
↪  ground-truth labels match the
↪  judgments of regional stakeholders.
↪  Annotator-population mismatch
↪  violates convergent and external
↪  validity.

### Output Form (OF)
The representation of output signals
↪  models produce | MCQ vs open-ended,
↪  text vs speech, verbose vs
↪  simplified. Mismatch with
↪  deployment needs violates external
↪  validity.

## Weighting Heuristics

Internally assign each dimension a
↪  priority (HIGH / MODERATE / LOWER)
↪  using these rules. Focus your
↪  questions on HIGH-priority
↪  dimensions; skip LOWER ones unless
↪  the deployment description leaves
↪  the priority ambiguous.

### IO
- **HIGH** when: benchmark source
↪  culture differs from target
↪  deployment culture; use case
↪  involves everyday cultural
↪  knowledge (food, sports, leisure,
↪  celebrations); target region has
↪  strong sub-national cultural
↪  variation.
- **LOWER** when: domain-specific
↪  technical task (code, math, medical
↪  imaging); benchmark co-designed
↪  with the target population.

### IC
- **HIGH** when: use case involves
↪  normatively charged content (hate
↪  speech, sentiment, morality,
↪  advice); deployment culture differs
↪  from benchmark source culture;
↪  benchmark uses LLM-generated or
↪  web-scraped content without
↪  community validation.
- **LOWER** when: factual/technical
↪  content with no cultural loading;
↪  benchmark content authored by
↪  in-region insiders.

### IF
- **HIGH** when: deployment involves
↪  non-text modalities (images, audio,
↪  video); target population uses a
↪  low-resource language or non-Latin
↪  script; regional infrastructure
↪  differs from benchmark assumptions
↪  (device quality, bandwidth,
↪  medical-imaging specs).
- **LOWER** when: text-only in a
↪  high-resource language matching the
↪  benchmark.

### OO
- **HIGH** when: benchmark output
↪  categories designed for a different
↪  cultural context; domain has
↪  legitimate pluralism (multiple
↪  valid answers, e.g., religious
↪  jurisprudence, moral reasoning);
↪  classification has culturally
↪  variable categories.
- **LOWER** when: binary
↪  correct/incorrect on factual
↪  content; output taxonomy
↪  co-designed with target
↪  stakeholders.

### OC
- **HIGH** when: ground-truth labels
↪  annotated by a non-representative
↪  pool; use case involves subjective
↪  judgment (sentiment, offensiveness,
↪  appropriateness); target population
↪  is multi-ethnic, multi-religious, or
↪  multi-linguistic.
- **LOWER** when: labels are objective
↪  and verifiable (named entities,
↪  dates, quantities); labels
↪  annotated by target-population
↪  insiders.

### OF
- **HIGH** when: target population has
↪  low literacy or relies on oral
↪  communication; deployment requires
↪  speech output (especially in a
↪  low-resource language); benchmark
↪  uses MCQ but deployment requires
↪  open-ended generation.
- **LOWER** when: deployment output
↪  modality matches the benchmark's
↪  exactly.

### Regional Conditioning
```

```
Treat country-level framing as
↪ insufficient by default for regions
↪ with strong sub-national variation
↪ (India, Nigeria, Indonesia,
↪ Ethiopia, Brazil) | elevate IO and
↪ ask about granularity. For
↪ deployments targeting indigenous or
↪ oral-tradition communities, elevate
↪ IF and OF (oral/written mismatch).
↪ For Islamic-world religious/legal
↪ domains, elevate OO and OC
↪ (legitimate pluralism across schools
↪ of thought). For thin-evidence
↪ regions (Pacific Islands, Caribbean,
↪ Central Asia, diaspora populations),
↪ frame questions exploratorily rather
↪ than confirmatorily.

## Question Selection

- Select **3-5 questions**,
↪ concentrated on HIGH-priority
↪ dimensions.
- Skip a question if the deployment
↪ description already implies the
↪ answer.
- Skip a dimension if it's clearly
↪ irrelevant (e.g., OF for a text-only
↪ system with high-literacy users).
- Prefer **recognition framing** |
↪ present 2-4 concrete, regionally
↪ grounded examples the user can
↪ confirm or deny, followed by an
↪ open follow-up | over **generation
↪ framing** that asks them to
↪ enumerate from scratch.
- Anchor examples to the user's stated
↪ region and use case. Never refer to
↪ benchmark papers or internal
↪ evidence sources in the question
↪ text.
- One question, one focus. Don't bundle
↪ multiple sub-questions into a
↪ single entry.

## Example Questions

These show the target shape | concrete,
↪ regionally anchored, answerable
↪ from user context alone. Generate
↪ new questions in this style for the
↪ region and use case at hand.

**IO** | Nigerian chatbot:
*"For your users in Nigeria, should the
↪ system understand content about
↪ Yoruba, Igbo, and Hausa cuisines,
↪ traditional games, Northern vs
↪ Southern holiday calendars (Eid vs
↪ Christmas), or local market
↪ culture? What else is culturally
↪ important for your users?"*

**IC** | Middle East educational tool:
*"For your students, would the system
↪ need to understand the Islamic
↪ calendar, regional educational
↪ traditions, differences between Eid
↪ al-Fitr and Eid al-Adha practices
↪ across countries, or local
↪ historical narratives distinct from
↪ Western textbook accounts? What
↪ cultural knowledge would your
↪ students expect it to have?"*

**IF** | rural South Asia agricultural
↪ tool:
*"What devices and connectivity do your
↪ users have | smartphones, feature
↪ phones, shared community devices;
↪ broadband, 3G, intermittent? Will
↪ they interact primarily by text or
↪ voice, and in which languages?"*

**OO** | Islamic legal advisory system:
*"In Islamic jurisprudence a single
↪ question can have multiple valid
↪ rulings across schools of thought
↪ (Hanafi, Maliki, Shafi'i, Hanbali).
↪ Does your system need to handle
↪ this pluralism, or should it
↪ present a single consensus answer?
↪ Would users in different regions
↪ (e.g., West African Maliki vs South
↪ Asian Hanafi) expect different
↪ guidance?"*

**OC** | East African hate-speech
↪ detection:
*"For hate-speech decisions in your
↪ context, whose judgment should
↪ count as authoritative | local
↪ community members, religious
↪ leaders, platform moderators? Would
↪ ethnic communities in your
↪ deployment (e.g., Kikuyu, Luo,
↪ Kalenjin in Kenya) plausibly
↪ disagree on the same content?"*

**OF** | Mexican indigenous-community
↪ service:
*"What literacy levels can you assume
↪ among your target users? Would the
↪ system need to speak rather than
↪ write its responses, and if so in
↪ which languages? Are there
↪ languages your users speak that
↪ lack reliable text-to-speech
↪ support?"*

## Output Format

Return ONLY a JSON array of question
↪ objects. No prose, no preamble, no
↪ code fences.
```

```
```
[
  {"id": "Q1", "dimension": "IC",
  ↪ "question": "..."},
  {"id": "Q2", "dimension": "OO",
  ↪ "question": "..."}
]
```

Field rules:
- `id` | sequential: `Q1`, `Q2`, `Q3`,
↪  ... (no gaps).
- `dimension` | one of: `IO`, `IC`, `IF`,
↪  `OO`, `OC`, `OF`.
- `question` | a single concrete
↪  question phrased directly to the
↪  user.

Output exactly 3, 4, or 5 question
↪  objects.

---
User message:

Benchmark metadata:

{METADATA_TEXT}

Deployment description:

{DEPLOYMENT_DESCRIPTION}
```

Figure 8. Elicitation question generation prompt. Sonnet produces 3–5 targeted follow-up questions based on the benchmark metadata and deployment description.

Step 2b: Elicitation Summary (Sonnet)

```
# Elicitation: Produce Summary

## Task

You are the elicitation agent in a
↪  validity-analysis pipeline. You have
↪  already asked the user 3-5 targeted
↪  questions and received their
↪  answers. Produce a structured
↪  summary that downstream pipeline
↪  stages (region YAML synthesis, web
↪  search, Opus scoring) will consume.

The user message contains:
1. The **deployment description** |
↪  what the AI system does, who it
↪  serves, where.
2. **Lightweight benchmark metadata** |
↪  basic information from the paper's
↪  first 1-2 pages: name, domain,
↪  languages, primary region, porting
↪  strategy, source culture. This
↪  provides high-level context; the
↪  full benchmark documentation is not
↪  yet available at this stage.
3. The **Q/A block** | the questions
↪  that were asked (with dimension
↪  tags) and the user's answers.

Your job is to (a) re-state the
↪  deployment context tightly, (b)
↪  paraphrase each Q/A for downstream
↪  consumption, (c) assign a priority
↪  to each validity dimension using
↪  the heuristics below, and (d) call
↪  out concrete gaps that downstream
↪  web search should investigate.

## Validity Dimensions

The priority table in your output
↪  covers these six dimensions.

### Input Ontology (IO)
The set of test-case categories the
↪  benchmark covers. Misalignment
↪  (missing categories, or irrelevant
↪  ones) harms content validity.

### Input Content (IC)
The concrete instances represented by
↪  individual datapoints.
↪  Implementation-level choices can
↪  introduce construct-irrelevant
↪  variance | e.g., implicit US-default
↪  assumptions.

### Input Form (IF)
The encoding of the input signal | text
↪  vs audio, camera specs,
↪  infrastructure. Signal-distribution
↪  mismatches violate external
↪  validity.

### Output Ontology (OO)
The space of outputs plus the decision
↪  rules that score them. For
↪  free-form outputs, the scoring
↪  function's interpretive step is
↪  where cultural misalignment most
↪  readily arises.

### Output Content (OC)
Label correctness | whether
↪  ground-truth labels match regional
↪  stakeholder judgment.
↪  Annotator-population mismatch
↪  violates convergent and external
↪  validity.
```

```
### Output Form (OF)
The representation of output signals | MCQ vs open-ended, text vs speech, verbose vs simplified. Mismatch with deployment needs violates external validity.

## Weighting Heuristics

Assign each dimension a priority (HIGH / MODERATE / LOWER) using these rules, conditioned on the deployment context **and the user's answers**. A user's answer may push a dimension up or down relative to the a-priori rule | use the answer as the tie-breaker when rules would pull in different directions.

### IO
- **HIGH** when: benchmark source culture differs from target deployment culture; use case involves everyday cultural knowledge; target region has strong sub-national cultural variation.
- **LOWER** when: domain-specific technical task; benchmark co-designed with target population.

### IC
- **HIGH** when: use case involves normatively charged content (hate speech, sentiment, morality, advice); deployment culture differs from benchmark source; benchmark uses LLM-generated or web-scraped content without community validation.
- **LOWER** when: factual/technical content with no cultural loading; benchmark authored by in-region insiders.

### IF
- **HIGH** when: deployment involves non-text modalities; target population uses a low-resource language or non-Latin script; regional infrastructure differs from benchmark assumptions.
- **LOWER** when: text-only in a high-resource language matching the benchmark.

### OO
- **HIGH** when: benchmark output categories designed for a different cultural context; domain has legitimate pluralism; classification has culturally variable categories.
- **LOWER** when: binary correct/incorrect on factual content; output taxonomy co-designed with target stakeholders.

### OC
- **HIGH** when: ground-truth labels annotated by a non-representative pool; use case involves subjective judgment; target population is multi-ethnic, multi-religious, or multi-linguistic.
- **LOWER** when: labels are objective and verifiable; labels annotated by target-population insiders.

### OF
- **HIGH** when: target population has low literacy or relies on oral communication; deployment requires speech output; benchmark uses MCQ but deployment requires open-ended generation.
- **LOWER** when: deployment output modality matches the benchmark's exactly.

### Regional Conditioning for Flagged Gaps

When the deployment targets a region with known thin evidence (Pacific Islands, Caribbean, Central Asia, diaspora populations, indigenous/oral-tradition communities), explicitly name in Flagged Gaps the subpopulations, languages, modalities, or sub-regions the benchmark likely doesn't cover. These are the anchors downstream web search uses to investigate. When the deployment targets a region with strong sub-national variation (India, Nigeria, Indonesia, Ethiopia, Brazil) but the user's answers don't specify sub-national granularity, call that out as a gap.

## Output Format

Emit a markdown document with these sections in exactly this order. No preamble, no code fences around the whole document.

~~~
## Use Case
Concise restatement of the deployment in 1-3 sentences.
```
```
```

```
## Target Population
Geography, sub-national cohort (if
↪  applicable), language(s),
↪  occupation or role,
and any demographic details relevant to
↪  the validity assessment.

## Elicitation Responses
Q1 [<dim>]: <question as asked>
A1: <user's answer, paraphrased if
↪  needed for clarity>

Q2 [<dim>]: ...
A2: ...
(...)

## Dimension Priority Weights
| Dimension | Priority | Rationale |
|-----------|----------|-----------|
| IO | HIGH/MODERATE/LOWER | one
↪  sentence conditioned on this
↪  deployment |
| IC | HIGH/MODERATE/LOWER | one
↪  sentence |
| IF | HIGH/MODERATE/LOWER | one
↪  sentence |
| OO | HIGH/MODERATE/LOWER | one
↪  sentence |
| OC | HIGH/MODERATE/LOWER | one
↪  sentence |
| OF | HIGH/MODERATE/LOWER | one
↪  sentence |

## Flagged Gaps
Concrete areas where the assessment
↪  will need additional evidence |
↪  subpopulations,
languages, modalities, or sub-regions
↪  the benchmark likely doesn't cover
↪  and that
downstream web search should
↪  investigate.
```

User message:

Deployment description:

{DEPLOYMENT_DESCRIPTION}

Benchmark metadata:

{METADATA_TEXT}

Questions & answers:

{QA_BLOCK}

Figure 9. Elicitation summary synthesis prompt. Sonnet produces dimension priority weights, cultural topic priorities, and a structured deployment context summary.

Step 3a-extract: Per-Page PDF Extraction (Haiku)

```
You are extracting validity-relevant
↪  quotes from page {PAGE_NUM} of a
↪  benchmark
paper. A single-page PDF is attached to
↪  this message as a document block |
↪  read
it and extract from that attached page
↪  only.

Extract ALL factual claims relevant to
↪  these categories:

1. Task taxonomy / test case categories
2. Data sources and collection
3. Data format and preprocessing
4. Label categories and output types
5. Annotation process (annotators,
↪  demographics, agreement)
6. Evaluation metrics and output
↪  modality
7. Stated limitations
8. Authors and affiliations

Output a single JSON object with this
↪  shape:

{
  "quotes": [
    {
      "text": "<EXACT verbatim text
      ↪  from the PDF - complete
      ↪  sentences>",
      "page": {PAGE_NUM},
      "category": "<one of:
      ↪  task_taxonomy, data_sources,
      ↪  data_format, label_categories,
      ↪  annotation_process,
      ↪  evaluation_metrics,
      ↪  stated_limitations,
      ↪  authors_affiliations>"
    }
  ],
  "page_summary": "<1-2 sentence summary
  ↪  of what this page covers>",
  "continues_from_previous":
  ↪  <true|false>,
  "continues_to_next": <true|false>
}

Continuity flags:
- `continues_from_previous`: true if the
↪  page opens mid-sentence or
↪  mid-paragraph | i.e. the first
↪  visible line starts with a lowercase
↪  word, a dangling clause, or clearly
↪  references material not on this
↪  page. False if it starts a new
↪  section, heading, or self-contained
↪  sentence.
```

```
- `continues_to_next`: true if the page
 ↪  ends mid-sentence or mid-paragraph |
 ↪  last visible line breaks off
 ↪  without a terminal
 ↪  period/punctuation, or ends on an
 ↪  em-dash/colon that demands a
 ↪  follow-up. False if it ends cleanly
 ↪  (end of section, end of paragraph,
 ↪  or closing punctuation).

Rules:
- Quote COMPLETE sentences only | never
 ↪  truncate mid-sentence.
- If a sentence spans from the previous
 ↪  page, include the full sentence
 ↪  visible on this page.
- If a claim spans multiple sentences,
 ↪  quote all of them together.
- Include statistical facts,
 ↪  methodological claims, design
 ↪  decisions, and limitations.
- Skip: related-work citations, model
 ↪  performance tables, and figure
 ↪  captions unless they contain
 ↪  methodological claims.
- If this page has no relevant content
 ↪  (references, full-page figures,
 ↪  etc.), return
  {"quotes": [], "page_summary": "...",
   ↪  "continues_from_previous": false,
   ↪  "continues_to_next": false}.

Output ONLY the JSON object, with no
 ↪  surrounding prose or code fences.

---
User message:

Extract validity-relevant content from
 ↪  page {PAGE_NUM}.

[Single-page PDF for page {PAGE_NUM} is
 ↪  attached as a document block.]
```

Figure 10. Per-page extraction prompt. One Haiku call per page, executed concurrently. Extracts validity-relevant quotes and cross-page continuity flags.

Step 3a-merge: Cross-Page Sentence Merge (Haiku)

```
You receive two text fragments
 ↪  extracted from adjacent pages of a
 ↪  PDF document.
The page boundary may have split a
 ↪  sentence between them.

Determine if these fragments are parts
 ↪  of the same sentence or thought. If
 ↪  so,
merge them into one clean sentence. If
 ↪  they are clearly separate, say so.

Rules:
- If they form one sentence: output ONLY
 ↪  the merged text on a single line
- If they are separate sentences:
 ↪  output the word SEPARATE on its own
 ↪  line
- Remove any duplicated words or
 ↪  partial words at the merge point
- Do not add words that aren't in either
 ↪  fragment
- Preserve original wording exactly,
 ↪  except for merge-point cleanup

---
User message:

Fragment A (end of page {PAGE_A_NUM}):
"{FRAGMENT_A_TEXT}"

Fragment B (start of page
 ↪  {PAGE_B_NUM}):
"{FRAGMENT_B_TEXT}"
```

Figure 11. Cross-page merge prompt. Called at each page boundary where continuity flags indicate a split sentence.

Step 3a-consolidate: Narrative Synthesis (Sonnet)

```
You will write the Metadata and
 ↪  Narrative Context sections for a
 ↪  benchmark paper
extraction. A pre-assembled Quote
 ↪  Registry (with sequential Q IDs,
 ↪  page numbers,
categories, and verbatim text) is
 ↪  provided as input | do NOT
 ↪  regenerate it.

Your job is to read the registry and
 ↪  write interpretive narrative that
 ↪  references
quote IDs, making the extraction
 ↪  human-readable.

## Purpose: validity assessment

This extraction feeds a downstream
 ↪  validity assessment. The goal is to
 ↪  determine
whether an AI benchmark appropriately
 ↪  measures what matters for a
 ↪  specific
deployment context | considering
 ↪  cultural, linguistic, and regional
 ↪  factors.

Validity is evaluated across six
 ↪  dimensions organized as a 2x3 grid
 ↪  of
(Input / Output) x (Ontology / Content /
 ↪  Form):
```

```
| Code | Dimension | What it captures |
|------|-----------|-----------------|
| IO | Input Ontology | Do the
↪  benchmark's test-case categories
↪  cover what the deployment needs? |
| IC | Input Content | Do the actual
↪  datapoints reflect the target
↪  population's language, culture,
↪  domain? |
| IF | Input Form | Does the signal
↪  encoding (text vs. audio,
↪  resolution, script) match
↪  deployment conditions? |
| OO | Output Ontology | Do the output
↪  labels and scoring rules reflect
↪  regional values and decision
↪  boundaries? |
| OC | Output Content | Do ground-truth
↪  labels agree with what regional
↪  stakeholders would judge correct? |
| OF | Output Form | Does the output
↪  modality match what the deployment
↪  actually produces? |

The eight extraction categories below
↪  were chosen because they supply
↪  evidence
for these dimensions. When writing
↪  narrative, highlight information
↪  that bears on
validity | both where the benchmark's
↪  design aligns with potential
deployment contexts and where it may
↪  diverge from them.

## Extraction categories

The registry uses eight category tags.
↪  Write one narrative subsection per
↪  category:

1. **task_taxonomy** | test-case
↪  categories, task types, subtask
↪  breakdown.
2. **data_sources** | where datapoints
↪  come from (crowdsourced, scraped,
↪  expert-written, etc.).
3. **data_format** | input modality,
↪  encoding, preprocessing,
↪  length/shape constraints.
4. **label_categories** | output label
↪  set, ontology, ground-truth schema.
5. **annotation_process** | who
↪  annotated, their demographics,
↪  inter-annotator agreement, QA.
6. **evaluation_metrics** | scoring
↪  metrics, output modality, what
↪  counts as a correct answer.
7. **stated_limitations** | limitations
↪  or cultural/regional scope
↪  restrictions the authors themselves
↪  call out.
8. **authors_affiliations** | author
↪  names and institutional
↪  affiliations (signals
↪  deployment-context origin).

## Writing the narrative

For each of the 8 categories, write 2-5
↪  sentences of interpretive context
↪  that
references quote IDs, e.g.:

> "The benchmark covers 22 task
↪  categories [Q1] split between
↪  industry-relevant
> and fundamental tasks [Q2]."

When a category has no quotes in the
↪  registry, say so explicitly:

> "NOT DOCUMENTED: the paper is silent
↪  on annotator demographics. This
↪  absence is
> itself a validity-relevant finding."

For categories that plausibly span both
↪  input and output (`task_taxonomy`,
`data_sources`, `data_format`,
↪  `stated_limitations`), explicitly
↪  cover each side
where the paper supports it | don't let
↪  one side dominate the narrative.

Every factual claim must reference at
↪  least one quote ID from the
↪  registry.
If a claim has no backing quote, either
↪  find the quote or flag it as
↪  inference.

## Elicitation summary (deployment
↪  context)

An **Elicitation Summary** may be
↪  provided alongside the registry.
↪  This summary
captures the specific deployment
↪  scenario under which the benchmark
↪  will be
evaluated | who will use the system, in
↪  what domain, and what matters most
↪  for
validity in that context. It includes
↪  **Dimension Priority Weights**
↪  (HIGH,
MODERATE, LOWER) that reflect which
↪  validity dimensions are most
↪  important for
this deployment.

Use these priorities to calibrate
↪  narrative depth:
```

```
- **HIGH-priority dimensions**: 4-6
↪  sentences | expand on nuances, note
  where the paper's design aligns with
   ↪  or diverges from what the
   ↪  deployment
  requires
- **MODERATE-priority dimensions**: 2-4
↪  sentences
- **LOWER-priority dimensions**: 1-2
↪  sentences

Cultural or domain-specific topics
↪  flagged in the summary deserve
↪  extra attention
| call out related quotes even if they
↪  would otherwise be treated as minor
↪  details.

Priority guidance affects **narrative
↪  depth only**. Every quote in the
↪  registry
must still be referenced regardless of
↪  priority level. Do not drop or skip
↪  quotes
from low-priority categories.

If no elicitation summary is provided,
↪  write all categories at uniform
↪  depth
(2-5 sentences each).

## Output format

Produce ONLY the following markdown
↪  sections | no surrounding prose, no
↪  registry:

```markdown
# Validity Extraction: {Full Paper
↪  Title}
<!-- Model routing: Haiku (per-page
↪  extraction) → script (registry
↪  assembly) → Sonnet (narrative) -->

## Metadata
- **Title**: ...
- **Authors**: ...
- **Venue/Year**: ...
- **Total Pages**: {from the registry
↪  page numbers}
- **Quotes Extracted**: {count of
↪  quotes in the registry}

## Narrative Context

Interpretive prose organized by
↪  extraction category. Each factual
↪  claim
references quote IDs from the registry.
↪  **This section is non-authoritative
| it provides readability but is not
↪  evidence. Only the Quote Registry
contains verbatim text from the
↪  paper.**

### 1. Task Taxonomy / Test Case
↪  Categories
...

### 2. Data Sources and Collection
...

### 3. Data Format and Preprocessing
...

### 4. Label Categories and Output
↪  Types
...

### 5. Annotation Process
...

### 6. Evaluation Metrics and Output
↪  Modality
...

### 7. Stated Limitations
...

### 8. Authors and Affiliations
...
```

Output ONLY the markdown above. The
↪ Quote Registry will be appended
↪ mechanically.

User message:

Pre-Assembled Quote Registry

{QUOTE_REGISTRY_TEXT}

Elicitation Summary (for narrative
↪ depth guidance)

{ELICITATION_SUMMARY}

Use the Dimension Priority Weights
↪ above to guide narrative depth:
- HIGH-priority dimensions: 4-6
↪ sentences of narrative context
- MODERATE-priority dimensions: 2-4
↪ sentences
- LOWER-priority dimensions: 1-2
↪ sentences

Cultural topic priorities from the
↪ summary should receive extra
↪ attention | call out related quotes
↪ even if they would otherwise be
↪ treated as minor details.
```

Figure 12. Narrative consolidation prompt. Sonnet writes interpretive context for the pre-assembled quote registry, guided by elicitation priority weights.

Step 3b-select: Reference Selection (Haiku)

```
Pick the 1 or 2 benchmark YAMLs most
↪  structurally relevant to the paper
described in the user message. Prefer
↪  matching `porting_strategy` first,
↪  then
`domain`, then `primary_region`.

The user message contains:
- A compact manifest of available
↪  reference YAMLs (one line per
↪  file).
- Key attributes of the paper being
↪  synthesized (stated domain,
↪  apparent
  porting_strategy, primary region).

Return a JSON array of filenames.
↪  Examples:

```
["afrimedqa.yaml"]
["helm.yaml", "seahelm.yaml"]
```

Return ONLY the JSON array. No prose,
↪ no code fences, no explanation.

User message:

Benchmark example manifest:

{EXAMPLE_MANIFEST}

Paper summary:

{PAPER_SUMMARY}
```

Figure 13. Benchmark reference selection prompt. Haiku picks 1–2 structurally similar reference YAMLs.

Step 3b-synthesize: Benchmark YAML (Sonnet)

```
Synthesize a benchmark YAML from a paper
↪  summary, using 1-2 reference
↪  examples
as style and schema guides.

The user message contains:
- The full paper summary from the
↪  extraction step (narrative +
↪  verbatim quote registry).
- 1-2 selected reference YAMLs from
↪  `benchmarks/examples/`. These are
↪  **format,
  style, and depth guides | do NOT copy
  ↪  their content verbatim.**

## Required fields (top level)

- `name` | short identifier (lowercase,
↪  no spaces)
- `full_name` | full benchmark title
↪  from the paper
- `year` | publication year
- `domain` | what the benchmark
↪  evaluates
- `porting_strategy` | one of:
↪  `ground_up`, `adapted`, `mixed`,
↪  `translation`,
  `parallel`, `regional_exams`, `none`
- `languages` | list of language codes
- `primary_region` | geographic or
↪  cultural region the benchmark
↪  targets

## The 6 validity dimensions

Both the `documentation_excerpts`
↪  narrative and the per-quote
↪  `dimension` tags
below use these six dimensions (2×3
↪  grid: Input/Output ×
↪  Ontology/Content/Form).
Use the definitions below so that the
↪  tagging and narrative you produce
↪  here
align with the definitions the
↪  downstream Opus scorer will use.
↪  The "Cues" line
under each is an operational addendum
↪  for this step only | concrete
↪  signals that
commonly mark a quote as belonging to
↪  that dimension.

- **input_ontology** | The input data
↪  ontology consists of the set of
↪  test case categories represented by
↪  the benchmark, which should cover
↪  the query types that evaluated
↪  systems are expected to encounter
↪  during deployment (e.g., factual
↪  questions, creative writing help, or
↪  restaurant recommendations for NLP
↪  dialogue systems; pick-and-place
↪  tasks, point navigation, or
↪  planning for robotics). A
↪  misalignment in this taxonomy |
↪  whether through the omission of
↪  necessary categories (construct
↪  underrepresentation) or the
↪  inclusion of irrelevant ones
↪  (construct-irrelevant variance) |
↪  harms the content validity of the
↪  benchmark.
  Cues: task types, subtask breakdown,
  ↪  coverage, stated gaps in the
  ↪  taxonomy.
```

```
- **input_content** | Whereas the
↪  ontology defines the types of input
↪  queries, the content of dataset
↪  inputs refers to the explicit
↪  instances specified by individual
↪  datapoints | an LLM prompt, an
↪  image, a database entry, etc. Even
↪  when the test case taxonomy provides
↪  good coverage, implementation-level
↪  details of individual datapoints
↪  can introduce construct-irrelevant
↪  variance, violating content
↪  validity.
  Cues: data sources, collection
   ↪  methodology, cultural grounding
   ↪  of the instances.

- **input_form** | The form of dataset
↪  inputs determines the encoding of
↪  the input signal | e.g., text vs.
↪  audio for natural language, or
↪  camera parameters such as focal
↪  length and resolution for visual
↪  data. Since machine learning
↪  systems are sensitive to signal
↪  distributions, a mismatch between
↪  the benchmark's input
↪  representation and the real-world
↪  signals that deployed models would
↪  encounter violates the external
↪  validity of the evaluation.
  Cues: modality, length, language,
   ↪  format, infrastructure
   ↪  assumptions (e.g. written vs
   ↪  spoken, script, SMS-length
   ↪  constraints).
```

```
- **output_ontology** | A benchmark's
↪  output ontology determines the
↪  space of outputs an AI system is
↪  expected to produce and the
↪  decision rules by which those
↪  outputs are organized and scored |
↪  i.e., the benchmark's criteria. For
↪  categorical outputs, the mapping is
↪  direct (e.g., safe/unsafe, or
↪  object class labels). For free-form
↪  outputs, the scoring function must
↪  first interpret what the output
↪  means before mapping it to a score |
↪  and this interpretive step is where
↪  validity violations most readily
↪  arise, since decision rules can
↪  differ across cultural contexts. For
↪  instance, an LLM recommending "red
↪  wine" for a dinner party may score
↪  highly for helpfulness in a Western
↪  context but poorly where alcohol
↪  consumption is prohibited. A
↪  misaligned output taxonomy thus
↪  violates structural validity (the
↪  construct's structure is
↪  misrepresented), content validity
↪  (through missing or irrelevant
↪  categories), and risks violating
↪  external validity (benchmark
↪  performance is less likely to
↪  generalize to regional settings).
  Cues: label set, ground-truth schema,
   ↪  taxonomy of correct-answer types.

- **output_content** | Whereas
↪  taxonomic alignment addresses
↪  whether abstract decision boundaries
↪  reflect regional values, label
↪  correctness concerns whether the
↪  labels for particular datapoints
↪  correlate with the judgments of
↪  regional stakeholders. Disagreement
↪  between regional and original
↪  annotators violates both convergent
↪  validity (the labels fail to
↪  correlate with regional
↪  perspectives on the construct) and
↪  external validity (the original
↪  judgments do not generalize to the
↪  target context).
  Cues: annotation process, annotator
   ↪  demographics, inter-annotator
   ↪  agreement, label provenance.
```

```
- **output_form** | The form of dataset
↪  outputs pertains to the
↪  representation of output signals
↪  models are expected to produce. If
↪  a benchmark does not evaluate
↪  models on the output forms
↪  encountered during real-world
↪  deployment, this violates the
↪  external validity of the
↪  evaluation.
  Cues: output modality, evaluation
  ↪  metric definition, scoring
  ↪  methodology.

## Required section:
↪  `documentation_excerpts`

Interpretive narrative organized by the
↪  6 validity dimensions, referencing
↪  quotes
by registry IDs (e.g., `[Q3]`, `[Q14]`).
↪  Structure with these exact headers:

1. **Input Ontology**
2. **Input Content**
3. **Input Form**
4. **Output Ontology**
5. **Output Content**
6. **Output Form**

Under each header, write the
↪  interpretive narrative for that
↪  dimension anchored
to the definition above. Propagate "NOT
↪  DOCUMENTED" notes from the paper
↪  summary
where a dimension has no supporting
↪  quotes.

## Required section: `verbatim_quotes`

Tag ALL Quote Registry entries with
↪  their validity dimension. Do NOT
↪  include the
quote text | it will be filled in
↪  programmatically from the registry.
↪  Output only
`id`, `page`, and `dimension`:

verbatim_quotes:
  - id: Q1
    page: 20
    dimension: input_ontology

Rules:
- Propagate ALL quotes | do not drop
↪  any.
- Tag each quote with the most relevant
↪  validity dimension using the
↪  definitions above. A quote's
↪  *extraction category*
↪  (`task_taxonomy`,
↪  `evaluation_metrics`, etc. from the
↪  registry) is a hint, not a mapping |
↪  e.g. a `task_taxonomy` quote about
↪  output label types belongs to
↪  `output_ontology`, not
↪  `input_ontology`.

## Required section:
↪  `coverage_strengths`

**Only include this section if an
↪  elicitation summary is provided in
↪  the user
message.** If no elicitation summary is
↪  present, omit the section entirely.

Cross-reference the elicitation summary
↪  against the benchmark documentation
↪  to
identify where the benchmark's coverage
↪  aligns well with the user's
↪  deployment
priorities. For each HIGH-priority
↪  dimension and each cultural topic
↪  priority,
note where the benchmark documents
↪  relevant coverage.

coverage_strengths:
  - user_priority: "French biomedical
  ↪  NLP"
    benchmark_coverage: "12 datasets
    ↪  spanning NER, STS,
    ↪  classification in French
    ↪  clinical text [Q5]"
    dimension: input_content
  - user_priority: "drug labeling
  ↪  terminology"
    benchmark_coverage: "QUAERO dataset
    ↪  drawn from EMEA drug labels
    ↪  with UMLS entity annotation
    ↪  [Q12]"
    dimension: input_ontology

Each entry should reference at least
↪  one quote ID grounding the strength
↪  in
the paper's documentation.

## Required section:
↪  `coverage_gap_analysis`

**Only include this section if an
↪  elicitation summary is provided in
↪  the user
message.** If no elicitation summary is
↪  present, omit the section entirely.
```

```
Cross-reference the elicitation summary
↪  against the benchmark documentation
↪  to
identify where the user's priorities
↪  and the benchmark's actual coverage
↪  diverge.
For each HIGH-priority dimension and
↪  each cultural topic priority from
↪  the
elicitation summary, assess whether the
↪  benchmark documents relevant
↪  content.

coverage_gap_analysis:
  - user_priority: "sub-national
  ↪  variation in Nigeria"
    benchmark_coverage: "Country-level
    ↪  only | no state or ethnic group
    ↪  breakdown"
    gap_type: "partial"
    web_search_target: "Sub-national
    ↪  cultural variation benchmarks
    ↪  Nigeria Yoruba Igbo Hausa"
  - user_priority: "regional cuisines"
    benchmark_coverage: "NOT DOCUMENTED
    ↪  | no food-related tasks"
    gap_type: "full"
    web_search_target: "Nigerian
    ↪  cuisine cultural knowledge AI
    ↪  evaluation"

Gap types:
- **full**: The benchmark is silent on
↪  this priority. Web search should
↪  look for
  supplementary evidence or alternative
  ↪  benchmarks.
- **partial**: The benchmark addresses
↪  this but at insufficient depth or
↪  the wrong
  granularity for the user's context.
- **none**: The benchmark covers this
↪  priority well. No web search
↪  needed.

The `web_search_target` field provides a
↪  suggested search query for the web
search step. Only populate for `full`
↪  and `partial` gaps.

## Schema enforcement

This schema is enforced by downstream
↪  scripts (`verify_quotes.py`,
`compose_prompt.py`). Treat it as a hard
↪  requirement, not a suggestion. Use
↪  the
reference YAMLs to match tone, field
↪  depth, and narrative density; use
↪  the
schema to guarantee downstream
↪  compatibility.

## Output
```

```
Output ONLY the YAML document. Wrap it
↪  in a ```yaml fenced block so it can
↪  be
extracted cleanly.

---
User message:

Reference examples:

### Example: {EXAMPLE_FILENAME_1}

{EXAMPLE_YAML_CONTENT_1}

### Example: {EXAMPLE_FILENAME_2}

{EXAMPLE_YAML_CONTENT_2}

Paper summary:

{PAPER_SUMMARY}

---

## Elicitation Summary

{ELICITATION_SUMMARY}

Use the dimension priority weights and
↪  cultural topic priorities above
↪  when writing the coverage_strengths
↪  and coverage_gap_analysis sections.
↪  For each HIGH-priority dimension
↪  and each cultural topic priority,
↪  assess whether the benchmark
↪  documents relevant content | noting
↪  both where coverage aligns well and
↪  where it diverges.
```

Figure 14. Benchmark YAML synthesis prompt. Sonnet produces a structured benchmark document with validity dimension tagging and coverage gap analysis.

Step 4a: Template Selection (Haiku)

```
Pick the 1 or 2 base region template
↪  YAMLs most structurally relevant to
↪  the
elicited target population.

The user message contains:
- The list of available template
↪  filenames in `regions/base/` (one
↪  per line).
- The elicitation summary's target
↪  region and population fields.

Selection rules:
- Prefer regional proximity (e.g.,
↪  `sub_saharan_africa.yaml` for a
↪  Kenyan deployment).
```

```
- Then prefer linguistic and cultural
 ↪  overlap.
- If no template is a close fit, pick
 ↪  the 2 structurally closest and rely
 ↪  on the
  downstream synthesis step to adapt
   ↪  them (you do not need to note the
   ↪  gap here |
  that's the synthesis subagent's job).

Return a JSON array of filenames.
 ↪  Examples:

```
["sub_saharan_africa.yaml"]
["mena.yaml", "iran_persian.yaml"]
```

Return ONLY the JSON array. No prose,
 ↪ no code fences.

User message:

Base template files:
{TEMPLATE_FILENAMES}

Elicitation summary:

{ELICITATION_SUMMARY}
```

Figure 15. Region template selection prompt. Haiku picks 1–2 base region templates by geographic and cultural proximity.

Step 4b: Region Scaffold (Sonnet)

```
Produce an assessment-specific region
 ↪  document scoped to the target
 ↪  population
described in the elicitation summary.

The user message contains:
- 1-2 selected base templates as JSON
 ↪  (structural and stylistic examples).
- The full elicitation summary
 ↪  (authoritative source for scoping).
- The benchmark as JSON (domain,
 ↪  languages, task categories).

## Granularity: narrower than the base
 ↪  templates

The base templates in `regions/base/`
 ↪  are **broad regional sketches**
(continent- or language-family-scale,
 ↪  e.g. `east_asia`, `mena`,
`sub_saharan_africa`). The output you
 ↪  produce is **not** another
broad-region template. It describes the
 ↪  specific deployment population
named in the elicitation summary |
 ↪  typically much narrower (one
 ↪  country, a
sub-national cohort, a specific ministry
 ↪  or NGO, a defined occupational
group).

Treat the base templates as structural
 ↪  references, not scope references.
Your output should look like a
 ↪  zoomed-in version of a template:
 ↪  same kind
of fields, but every value narrowed to
 ↪  the deployment population. If the
elicitation scopes to "rural community
 ↪  health workers in western Kenya,"
the output is about them specifically |
 ↪  not about Sub-Saharan Africa in
general. Fields that only make sense at
 ↪  the broader scale (pan-regional
trade blocs, continent-wide literacy
 ↪  averages) get dropped or replaced
with narrower equivalents.

In rare cases the deployment is itself
 ↪  broad (e.g. "any Spanish-speaking
user" with no further scoping). Match
 ↪  the scope the elicitation actually
specifies; don't invent narrowness that
 ↪  isn't there.

## Pipeline context: you are producing
 ↪  a scaffold, not a finished document

A downstream step runs another Sonnet
 ↪  call with web search access that
 ↪  will
fill in factual content. **You have no
 ↪  web search and no tools.**
Everything factual you emit comes from
 ↪  parametric knowledge, which is
 ↪  stale
by construction (training-cutoff +
 ↪  sub-national populations + shifting
policy), so even facts you feel
 ↪  confident about are not reliable
 ↪  enough to
ship without verification.

Your job here is to build the best
 ↪  possible scaffold:
- Lay out the right schema for this
 ↪  population (including new fields
 ↪  the
  base templates don't cover).
- Populate the fields that can be
 ↪  grounded in the elicitation summary
 ↪  or
  schema induction directly.
- Leave tagged placeholders for every
 ↪  factual slot a web-search step
 ↪  should
  resolve.

## Core rules
```

```
**The provided base template(s) define
↪  the schema by example.** Use them
↪  as
guides for structure, field depth, and
↪  tone | do NOT copy their content
verbatim. Induce the schema from the
↪  examples rather than from an
↪  explicit
field list. If the elicitation surfaces
↪  population characteristics the base
templates do not accommodate
↪  (sub-national identity, occupation
↪  cohort, age
band, clinical setting, specific
↪  regulatory regime, etc.), **add new
↪  fields**
rather than dropping the information.

**Fill directly when grounded.** Fields
↪  derivable from the elicitation
summary or from obvious structural
↪  induction can be written as normal:
- Qualitative framing of the population
↪  (who they are, what they do, what
  they care about) | the elicitation
  ↪  summary is authoritative here.
- Field names, nesting, and
↪  organization induced from the base
↪  template
  schema plus use-case-salient
  ↪  additions.
- Top-level high-confidence categorical
↪  facts (primary language, country,
  benchmark domain) where the
  ↪  elicitation has already pinned
  ↪  them down.

**Tag `[NEEDS VERIFICATION]` for
↪  everything factual and specific.**
↪  Any
value that a future reader should not
↪  trust without an external check
↪  gets
the tag | not as a value, but as a
↪  placeholder in place of the value:

- Quantitative statistics (literacy
↪  rates, internet penetration, device
  share, language-share percentages,
  ↪  urbanization %, demographic
  breakdowns).
- Named policies, regulations,
↪  standards, platforms (PP 71/2019,
↪  specific
  ministerial circulars, platform names,
  ↪  API endpoints, legal thresholds).
- Time-sensitive facts (current-year
↪  figures, active programs, recent
  reforms, "as of" claims).
- Sub-national specifics
↪  (province-level variation,
↪  locally-specific
  procedures, regional office
  ↪  structures) even if the general
  ↪  pattern is
  known.
- Infrastructure specifics (bandwidth
↪  figures, mobile-network coverage,
  device model distributions).

Examples:

// GOOD | grounded qualitative field,
↪  fill directly:
"population_description": "Indonesian
↪  citizens seeking guidance on
↪  federal administrative procedures
↪  (KTP, tax filing, business
↪  registration), contacting the
↪  service primarily via mobile."

// GOOD | factual slot, tag instead of
↪  guessing:
"literacy_rate": "[NEEDS
↪  VERIFICATION]",
"mobile_internet_penetration_pct":
↪  "[NEEDS VERIFICATION]",
"applicable_data_protection_regulation":
↪  "[NEEDS VERIFICATION]"

// GOOD | list structure induced from
↪  schema, individual facts tagged:
"regional_procedural_variation": [
  {"region": "DKI Jakarta", "notes":
  ↪  "[NEEDS VERIFICATION]"},
  {"region": "[NEEDS VERIFICATION |
  ↪  surface other salient
  ↪  kabupaten/provinces]", "notes":
  ↪  "[NEEDS VERIFICATION]"}
]

// BAD | parametric-memory number
↪  shipped as fact:
"literacy_rate": 96
// BAD | confident-sounding but
↪  potentially stale regulation name:
"applicable_data_protection_regulation":
↪  "PP 71/2019"

**When in doubt, tag.** Over-tagging at
↪  most wastes a web search; under-
tagging ships stale facts into the
↪  final assessment. Err toward the
↪  former.

**Do not tag fields that don't belong
↪  in this assessment.** If a base
template has a field that's irrelevant
↪  to the scoped population (e.g. a
`traditional_medicine_practices` field
↪  for a deployment about tax filing),
drop it | don't tag it.

**Read-only invariant:** `regions/base/`
↪  is never modified. Your output is
```

```
a new document for
↪  `assessments/<benchmark>/<slug>/region_sc

## Output

Output ONLY the document as a JSON
↪  object, wrapped in a ```json fenced
block so it can be extracted cleanly.
↪  No preamble. Use the same field
names and nesting as the base templates
↪  | the only difference is that
the output format is JSON instead of
↪  YAML. Multi-line text values should
be plain strings (use `\n` for line
↪  breaks if needed).

---
User message:

Base templates:

### Template: {TEMPLATE_FILENAME_1}

{TEMPLATE_JSON_CONTENT_1}

### Template: {TEMPLATE_FILENAME_2}

{TEMPLATE_JSON_CONTENT_2}

Elicitation summary:

{ELICITATION_SUMMARY}

Benchmark:

{BENCHMARK_JSON}
```

Figure 16. Region scaffold synthesis prompt. Sonnet produces an assessment-specific region document with [NEEDS VERIFICATION] placeholders for factual slots.

Step 5: Web Search Enrichment (Sonnet + web_search tool)

```
# Web search enrichment for the region
↪  document

You are the web-search stage of a
↪  cultural validity assessment
↪  pipeline.
The upstream scaffold step (4b)
↪  produced a region document (JSON)
↪  describing
the deployment population, but it ran
↪  without tool access, so every
↪  factual
slot it could not confidently ground is
↪  left as `[NEEDS VERIFICATION]`.
Your job is to replace those tags with
↪  verified values and, where your
searches surface them, add net-new
↪  fields that would strengthen the
assessment.

## Priorities

1. **Resolve `[NEEDS VERIFICATION]`
↪  tags.** These are the primary
↪  search
   targets. Each tag stands in for a
   ↪  factual slot the scaffold
   ↪  couldn't
   commit to | usually a quantitative
   ↪  statistic, a time-sensitive
   ↪  policy
   name, a sub-national specific, or an
   ↪  infrastructure figure. Replace
   each tag with a verified value, or
   ↪  with a short note explaining what
   was searched and why no value was
   ↪  found.

2. **Preserve fields the scaffold
↪  filled directly.** Non-tagged
↪  values
   were grounded in the authoritative
   ↪  elicitation summary or induced
   from the schema. Do not silently
   ↪  overwrite them. If your search
   surfaces a contradiction, add a
   ↪  short caveat next to the
   ↪  original
   value rather than replacing it.

3. **Add net-new fields when searches
↪  surface them.** If a search reveals
   a deployment-relevant fact the
   ↪  scaffold didn't anticipate (a
   ↪  recent
   policy shift, a cohort
   ↪  characteristic that affects
   ↪  validity, a
   benchmark or dataset specific to the
   ↪  target region), append it to
   the document in a logical place.

## Expect partial coverage

The scaffold deliberately over-tags:
↪  step 4b was told to err toward
`[NEEDS VERIFICATION]` whenever it
↪  couldn't confidently ground a
↪  factual
slot. In practice this means the
↪  document you receive will almost
↪  always
have substantially more tags than your
↪  search budget can cover.

This is expected, not a failure.
↪  Leaving most tags unresolved is the
normal outcome. Your job is not to
↪  attempt every tag | it is to triage
ruthlessly, spend your budget on the
↪  tags that most materially affect
```

```
the downstream assessment, and leave
↪  the rest clearly marked for a
future pass or for human follow-up. A
↪  small number of well-resolved
high-impact tags is strictly better
↪  than a shallow sweep that touches
every tag with a single under-budgeted
↪  query.

## Budget

You have roughly 5{10 searches.
↪  Allocate by impact:

- **High impact** | tags whose
↪  resolution will materially shape
↪  the
  downstream scoring step: named
  ↪  regulations/standards/platforms,
  ↪  current
  literacy or web-corpus figures for
  ↪  the target language, sub-national
  procedural variation, named agencies
  ↪  or ministries.
- **Medium impact** | facts that
↪  sharpen the picture but don't
↪  redirect
  the assessment: demographic
  ↪  breakdowns, device/infrastructure
  ↪  stats,
  regional benchmark citations.
- **Low impact / skip** | slots whose
↪  resolution wouldn't change the
  final scoring, or facts the scaffold
  ↪  already grounded.

If budget remains after high-impact
↪  tags, spend it on a sweep for
cultural-validity-relevant evidence
↪  specific to the target region or
language: existing benchmarks, annotator
↪  demographics, corpus size for
the target language, regional
↪  deployment case studies (both
↪  successes
and failure modes). Null results are
↪  informative too | confirming that
no benchmark exists for a region is
↪  itself an input to the assessment.

## Prefer information gain over
↪  confirmation

Allocate queries to searches that
↪  *could* change the assessment, not
ones that confirm what's already baked
↪  in.

**Skip or deprioritize | generally
↪  established, not worth verifying:**
- That LLMs have a Western / US content
↪  default.
- That LLMs stereotype non-Western
↪  cultures.
- That country-level analysis erases
↪  sub-national variation.
- That LLMs default to US values /
↪  norms.

**Prioritize | would change the
↪  assessment direction:**
- Evidence that a benchmark actually
↪  transferred successfully to the
  target region (rare; high-value if
  ↪  found).
- Negative or tempering results
↪  suggesting a validity concern is
↪  less
  severe than expected.
- New benchmarks or datasets
↪  specifically for the target region,
  language, or domain.
- Methodological innovations the
↪  assessment could recommend
  (participatory elicitation, regional
  ↪  re-annotation, diagnostic
  controls separating language from
  ↪  culture).
- Community-sourced ground truth for
↪  the target population.

## What is vs. isn't searchable

**Generally searchable | spend queries
↪  here:**
- Named regulations, standards,
↪  policies, platforms, agencies
↪  (verify
  current name, scope, effective date).
- Statistics from national/regional
↪  bodies (literacy, internet
  penetration, language share in census
  ↪  data, urbanization rates).
- Named benchmarks, datasets, papers,
↪  and their regional coverage.
- Temporal / time-sensitive facts (what
↪  changed recently, what the
  current version of a policy or
  ↪  platform is).
- Infrastructure figures from
↪  international bodies (ITU, World
↪  Bank,
  Common Crawl corpus data) and ministry
  ↪  reports.

**Generally NOT searchable | don't
↪  spend budget here; flag instead:**
- Everyday lived practice (what people
↪  actually do day-to-day, rather
  than what iconic/national-level
  ↪  documentation describes).
- Oral tradition, domestic/private
↪  rituals, implicit pragmatic rules |
  transmitted through socialization,
  ↪  not documented online.
```

```
- Real sub-national variation that
↪  isn't aggregated online (locally
  specific procedural differences,
  ↪  community-by-community norms).
- Community-level historical memory,
↪  contested or suppressed narratives.
- Contextual communication rules
↪  (age-hierarchy, status modifiers,
  cultural humor, indirect expression).

If a `[NEEDS VERIFICATION]` tag sits on
↪  something from the non-searchable
list, replace it with a short note
↪  indicating the field requires
stakeholder / expert elicitation rather
↪  than fabricating a value.

## Cultural axes to consider

These twelve topical axes cover most of
↪  the cultural surface a benchmark
can fail to honor. Not every axis
↪  applies to every assessment | use
↪  the
"conditionally relevant when" gate to
↪  decide which ones warrant search
budget for this deployment. Per-axis
↪  entries list the abstract shape of
evidence worth finding, the kinds of
↪  knowledge that are usually not
retrievable, and when the axis is worth
↪  investigating.

### Cuisines and food
- **Search for:** regional
↪  food-knowledge benchmarks,
↪  specialized food
  databases or recipe corpora,
  ↪  ethnographic or community-sourced
  ↪  food
  inventories for underrepresented
  ↪  regions, studies documenting
  food-related stereotyping in LLM
  ↪  outputs.
- **Usually unsearchable:** everyday
↪  eating practices (vs. iconic
  national dishes); preparation methods
  ↪  and dietary restrictions;
  ingredients and spice palettes for
  ↪  underrepresented regions;
  oral-tradition food knowledge.
- **Relevant when:** benchmark includes
↪  food or cuisine items; deployment
  involves recommendations,
  ↪  health/nutrition, or everyday
  ↪  assistants in
  non-Western contexts.

### Fashion and style
- **Search for:** visual clothing or
↪  fashion datasets, non-Western
  garment knowledge bases,
  ↪  cross-lingual clothing terminology
  ↪  studies,
  textile heritage registries.
- **Usually unsearchable:** contemporary
↪  everyday fashion (vs.
  ceremonial/traditional); footwear and
  ↪  body adornment; occasion-
  appropriateness rules; gender/age
  ↪  variation in dress norms.
- **Relevant when:** benchmark involves
↪  visual cultural representation,
  image generation, or fashion
  ↪  recommendation; target region has
  ↪  strong
  sub-national clothing identity.

### Architectural spaces
- **Search for:** landmark or place
↪  recognition benchmarks, urban
  planning and heritage registries,
  ↪  UNESCO heritage lists, studies on
  geographic bias in text-to-image
  ↪  generation for built
  ↪  environments.
- **Usually unsearchable:** informal
↪  markets (souks, bazaars, wet
  markets); behavioral norms within
  ↪  spaces (gender access, bargaining
  customs); vernacular /
  ↪  non-heritage-listed architecture;
  ↪  neighborhood
  character.
- **Relevant when:** benchmark targets
↪  vision-language models;
  deployment involves navigation,
  ↪  tourism, heritage, or urban
  ↪  planning.

### Performance and art
- **Search for:** music or dance
↪  knowledge benchmarks for
↪  non-Western
  traditions, folk narrative or oral
  ↪  poetry NLP datasets, UNESCO
  intangible cultural heritage lists,
  ↪  audio-based cultural evaluation
  datasets.
- **Usually unsearchable:** theatre
↪  traditions and dramatic arts; oral
  poetry traditions; performance
  ↪  context (who performs, for whom,
  ↪  when,
  what it signifies); folk narrative
  ↪  and creation myths.
- **Relevant when:** benchmark targets
↪  music, dance, or artistic
  knowledge; target region has strong
  ↪  oral performance traditions.

### Events (festivals, holidays,
↪  lifecycle ceremonies)
- **Search for:** wedding or funeral
↪  cultural knowledge benchmarks,
```

```
  non-Western holiday calendar
  ↪ databases, studies on
  ↪ Western-holiday
  projection in LLM outputs,
  ↪ ethnographic studies of lifecycle
  ceremonies.
- **Usually unsearchable:** birth rites
↪ and coming-of-age ceremonies;
  private/domestic lifecycle rites;
  ↪ funeral practices for most
  ↪ cultures;
  step-by-step ritual sequences.
- **Relevant when:** target region is
↪ multi-religious or multi-ethnic;
  use case involves culturally
  ↪ sensitive content generation
  ↪ (greetings,
  calendars, cultural assistants);
  ↪ target region has low digital
  representation.

### Rituals and social practices
- **Search for:** participatory ritual
↪ elicitation studies, non-English
  ritual extraction from local-language
  ↪ corpora, multimodal ritual
  datasets, community-sourced ritual
  ↪ data.
- **Usually unsearchable:** birth and
↪ death rites; procedural enactment
  (who performs which action, in what
  ↪ sequence); domestic/private
  rituals; ritual knowledge transmitted
  ↪ orally.
- **Relevant when:** benchmark includes
↪ cultural practices, festivals,
  or ceremonies; target region has
  ↪ ritual diversity underrepresented
  ↪ in
  English web text.

### Sports and games
- **Search for:** traditional or
↪ indigenous game inventories,
↪ non-English
  sports knowledge from local-language
  ↪ media, community-sourced sports
  practice studies, sub-national sports
  ↪ culture documentation.
- **Usually unsearchable:** what people
↪ actually play locally at schools
  or in neighborhoods; spectator
  ↪ culture and stadium social
  ↪ practices;
  gender norms in sports; traditional
  ↪ game rules and cultural
  significance.
- **Relevant when:** benchmark includes
↪ everyday cultural practices or
  recreation; target region has sports
  ↪ culture dominated by non-Western
  forms.

### Beliefs (religious knowledge,
↪ mythology)
- **Search for:** religious knowledge
↪ benchmarks for the specific
  tradition(s) involved; methodology
  ↪ papers on evaluating lived vs.
  textual/doctrinal religion; syncretic
  ↪ or non-Abrahamic religious
  datasets; mythology or cosmology NLP
  ↪ resources.
- **Usually unsearchable:**
↪ lived/practiced religion (vs.
↪ canonical text
  knowledge); indigenous spiritual
  ↪ systems and animist traditions;
  syncretic practices (folk Catholicism,
  ↪ spirit-Buddhism blends, etc.);
  mythology and cosmological narratives
  ↪ from oral traditions.
- **Relevant when:** target region has
↪ religious traditions structuring
  daily life; benchmark includes
  ↪ identity-group content; use case
  involves content moderation or bias
  ↪ detection; deployment language is
  low-resource with religious
  ↪ terminology lacking English
  ↪ translation.

### Values
- **Search for:** cross-cultural value
↪ measurement studies for the target
  region, religious or moral framework
  ↪ evaluations, participatory value
  elicitation studies, updated value
  ↪ measurements (many foundational
  surveys are decades old).
- **Usually unsearchable:**
↪ intra-country value variation (data
↪ is
  usually national aggregate);
  ↪ religious moral reasoning
  ↪ frameworks;
  non-survey-based value expression (how
  ↪ values manifest in daily life).
- **Relevant when:** benchmark involves
↪ normatively charged content
  (hate speech, sentiment,
  ↪ advice-giving); benchmark uses
  ↪ LLMs as
  cultural informants.

### Norms
- **Search for:** norms benchmarks for
↪ the target region,
  legal/regulatory norm evaluation,
  ↪ agentic norm compliance
  ↪ evaluation,
  human-AI interaction norm studies (how
  ↪ users from different cultures
  expect AI to behave).
- **Usually unsearchable:** gift-giving
↪ and multi-constraint norm
```

```
  reasoning; human-AI interaction norm
  ↪ expectations; legal/regulatory
  norms for most Global South
  ↪ jurisdictions; norm evolution over
  ↪ time.
- **Relevant when:** benchmark involves
↪ social behavioral expectations,
  conversational AI, or agentic systems;
  ↪ target is a multi-cultural or
  multi-religious society; deployment
  ↪ involves government services or
  legal contexts.

### Communication
- **Search for:** dialectal NLP
↪ evaluation for the target language
  family, figurative language and
  ↪ proverb comprehension benchmarks,
  code-switching evaluation datasets,
  ↪ speech/audio cultural evaluation
  (null results are expected).
- **Usually unsearchable:**
↪ oral-language evaluation in audio
↪ modality;
  tonal-language cultural nuances and
  ↪ sign-language cultural variation;
  context-dependent communication rules
  ↪ (age, gender, status modifiers);
  cultural humor, sarcasm, and indirect
  ↪ expression.
- **Relevant when:** benchmark involves
↪ conversational or dialogue
  tasks; figurative language or
  ↪ pragmatic inference; dialectal or
  low-resource language deployment;
  ↪ translation or cross-lingual
  transfer.

### History and important figures
- **Search for:** LLM evaluations on
↪ non-Western history, oral-history
  digitization research, evaluations of
  ↪ multi-perspective historical
  narratives, culturally important
  ↪ person recognition datasets.
- **Usually unsearchable:**
↪ community-level historical memory
↪ (how
  families narrate their past);
  ↪ founding myths and origin
  ↪ narratives
  (often not distinguished from factual
  ↪ recall); suppressed or oral
  histories not in digitized form;
  ↪ locally important figures known
  ↪ only
  within their community.
- **Relevant when:** benchmark claims
↪ broad cultural knowledge
  evaluation; deployment involves
  ↪ educational AI in post-colonial
  regions; civic or political
  ↪ applications; regions with
  ↪ contested or
  actively reinterpreted histories.

## Evidence by validity dimension
↪ (fallback targeting)

When a `coverage_gap_analysis` section
↪ is present in the benchmark YAML,
use it as your primary search targeting
↪ | it already maps user priorities
to specific gaps with suggested queries.
↪ The dimension lens below is a
fallback for thin or absent elicitation
↪ context.

The scoring step rates six dimensions.
↪ When a search result surfaces, ask
which dimension it informs | the result
↪ may support or challenge the
benchmark's fitness for the deployment,
↪ and both are valuable:

- **IO | Input Ontology**: regional
↪ taxonomies, category coverage
↪ studies, emic vs. etic category
↪ systems.
- **IC | Input Content**: regional
↪ validation studies, community
↪ datasets, content relevance
↪ assessments.
- **IF | Input Form**: dialect
↪ evaluations, infrastructure
↪ compatibility, morphology
↪ benchmarks.
- **OO | Output Ontology**: regional
↪ label system studies, taxonomy
↪ adaptation work, alternative
↪ scoring approaches.
- **OC | Output Content**: regional
↪ annotation studies, annotator
↪ demographic analyses,
↪ cross-population agreement.
- **OF | Output Form**: output modality
↪ comparisons, TTS availability,
↪ literacy-rate data, accessibility
↪ studies.

## Calibration warnings

- **Language competence vs. cultural
↪ knowledge.** For low-resource-
  language deployments, a benchmark
  ↪ score drop can reflect either a
  language-processing failure or a
  ↪ cultural-knowledge gap | with
  different interventions. When search
  ↪ results cite scores in the target
  language, note the ambiguity unless
  ↪ the paper explicitly controls for
  language competence (e.g., via a
  ↪ matched-English control or a
```

```
  language-understanding diagnostic).

- **Sub-national variation under
↪  national statistics.** When a
↪  national-
  level statistic resolves a tag
  ↪  (literacy, internet penetration,
  language share), check whether the
  ↪  deployment is itself
  ↪  sub-national.
  If so, add a short caveat noting the
  ↪  resolved value is a national
  aggregate and that within-country
  ↪  variation may meaningfully change
  the picture for the specific cohort.

- **Sparse-documentation regions.** For
↪  deployments targeting oral-
  tradition cultures, indigenous or
  ↪  stateless communities, diaspora
  populations, or digitally
  ↪  underrepresented regions, null
  ↪  results are
  themselves informative | calibrate
  ↪  expectations rather than spending
  extra budget on re-phrased queries. A
  ↪  thin result set usually reflects
  a documentation gap, not a thin
  ↪  culture; flag this explicitly in
  ↪  the
  document rather than reporting "no
  ↪  evidence found" ambiguously.

## Query construction

- **Combine region + domain + method.**
↪  `"Sub-Saharan Africa" + "food
  knowledge" + LLM + benchmark` is more
  ↪  effective than `African cultural
  AI` (too broad).
- **Use specific seed terms.** If you
↪  expect a regulation, agency,
  benchmark, or dataset to exist under
  ↪  a particular name, search that
  name directly.
- **Prefer official and academic
↪  sources.** Government portals for
  policy; arXiv, Semantic Scholar, ACL
  ↪  Anthology, IEEE Xplore, or
  venue-specific repositories for AI
  ↪  research; ITU, World Bank,
  ↪  national
  statistics bureaus for quantitative
  ↪  indicators.
- **Use site-restricted queries when it
↪  helps.** `site:arxiv.org` for
  preprints, government domains for
  ↪  policy names, conference or
  ↪  journal
  sites for domain-specific work (e.g.
  ↪  `site:aclanthology.org` for NLP,
  `site:openreview.net` for ML,
  ↪  `site:miccai.org` for medical
  ↪  imaging).
- **Cross-check.** A single source is
↪  suspect for a factual claim; prefer
  convergence between two independent
  ↪  sources before committing a
  ↪  value.

## Red flags | do not rubber-stamp

- **Automated cultural metrics.**
↪  CLIP-based cross-cultural scores,
↪  BLEU
  on dialect translation, and similar
  ↪  automated evaluations have been
  shown to reward stereotypical outputs
  ↪  over culturally accurate ones.
  If a source relies on these without
  ↪  in-region human validation, note
  the limitation rather than citing the
  ↪  finding as ground truth.
- **Wikipedia as sole source.** English
↪  Wikipedia systematically
  underrepresents Global South and
  ↪  oral-tradition contexts.
  ↪  Acceptable
  as corroboration; not sufficient as
  ↪  sole source for a regional claim.
  Prefer non-English Wikipedia or
  ↪  community-contributed databases
  ↪  when
  those are available.
- **Stereotype reinforcement.**
↪  Searches will surface content that
  encodes the exact stereotypes a
  ↪  cultural validity assessment is
  ↪  meant
  to guard against. If a result looks
  ↪  like a stereotype rather than a
  fact, don't import it.
- **Outdated policy.** If a regulation,
↪  program, or standard has been
  superseded, cite the current one |
  ↪  note the transition if it's
  ↪  recent
  enough that deployments may still
  ↪  encounter the older version.

## Output

Return the FULL updated region document
↪  as a single JSON object in a
```json fenced block. No preamble, no
↪  commentary outside the fence.
Use the same field names and nesting as
↪  the input | the only difference
is that the output format is JSON.
↪  Multi-line text values should be
plain strings (use `\n` for line breaks
↪  if needed).
```
```

```
Every original `[NEEDS VERIFICATION]`
↪  tag must map to exactly one of the
three fates below | the distinction
↪  matters to the downstream reader,
because "investigated and not found"
↪  and "not investigated" have very
different implications for follow-up.

- **Resolved** | replaced with the
↪  verified value, followed by a
  **mandatory inline source citation**:
  ↪  short source descriptor + URL.
  Every resolved tag must be traceable;
  ↪  users need to be able to verify
  any claim themselves. Prefer two
  ↪  independent sources where the
  ↪  value
  is contentious or load-bearing.

  **URL formatting rule:** Wrap every
  ↪  URL in `[|` and `|]` delimiters.
  This enables downstream tooling to
  ↪  extract citations mechanically.
  Every URL must be **complete and
  ↪  copy-pasteable** | a user should
  ↪  be
  able to paste it into a browser and
  ↪  reach the source. No shorthand
  (e.g., write
  ↪  `[|https://arxiv.org/abs/2409.08564|]`,
  ↪  not
  `arXiv 2409.08564`). No partial URLs,
  ↪  no DOI-only references without
  the resolver prefix.

  Examples:

"literacy_rate_pct": "96 (Badan Pusat
↪  Statistik 2023 |
↪  [|https://bps.go.id/...|])",
  "mobile_internet_penetration_pct":
  ↪  "78 (ITU DataHub 2024 |
  ↪  [|https://datahub.itu.int/...|];
  ↪  World Bank WDI 2023 |
  ↪  [|https://data.worldbank.org/...|])",
  "data_protection_regulation": "UU
  ↪  27/2022 PDP (Kemenkumham official
  ↪  text |
  ↪  [|https://peraturan.go.id/...|])"

For long text values, use `\n` for line
↪  breaks and end with a trailing
  `Source: <descriptor> | [|<URL>|]`
  ↪  line inside the same string.
- **Searched, not found** | replaced
↪  with a short note starting with
  `[NOT FOUND |` that describes what was
  ↪  searched and why no value
  surfaced (e.g. `[NOT FOUND | searched
  ↪  ITU and World Bank 2023/2024;
  no Indonesia NTT-specific figure
  ↪  published]`).
- **Deferred, not searched** | left in
↪  place as
  `[NEEDS VERIFICATION | deferred:
  ↪  <brief reason>]` so the tag is
  ↪  still
  machine-identifiable for a future
  ↪  pass. The reason should be terse:
  `below search budget`, `likely
  ↪  unsearchable (lived practice)`,
  `low impact for scoring`, etc.

Additionally:

- Non-tagged fields are preserved (with
↪  short caveats added only on
  strong contradictory evidence | never
  ↪  silent overwrites). Citations
  are not required for non-tagged
  ↪  fields, since those were grounded
  ↪  in
  the elicitation, not the web.
- Net-new fields surfaced by your
↪  searches are appended in a logical
  location in the schema. Net-new
  ↪  fields are web-sourced by
  ↪  definition,
  so the same **mandatory inline source
  ↪  citation** and **`[|URL|]`
  delimiter** rules apply to them.
- **Keep descriptions concise | this is
↪  critical for output budget.**
  For both resolved tags and net-new
  ↪  fields, write at most 2{3
  sentences: the key fact, and one
  ↪  sentence on why it matters for
  ↪  the
  benchmark's validity. Do NOT write
  ↪  multi-paragraph explanations,
  historical timelines, implementation
  ↪  recommendations, or exhaustive
  policy detail. The downstream scorer
  ↪  only needs the fact and its
  relevance | everything else wastes
  ↪  output tokens and risks
  ↪  truncation.

---
User message:

Current region document
↪  (assessment-specific | enrich this):

{REGION_SCAFFOLD_JSON}

Benchmark:

{BENCHMARK_JSON}

Elicitation summary:

{ELICITATION_SUMMARY}
```

```
**Start with the benchmark's `coverage_gap_analysis` section.** This is your primary search targeting input | it lists exactly where the user's priorities and the benchmark's coverage diverge, with gap types (`full`/`partial`) and suggested search queries. Allocate the majority of your search budget to `full` and `partial` gaps before applying the guide's general heuristics.

The upstream step (4b) produced a scaffold without tool access, so every factual slot it could not confidently ground was left as `[NEEDS VERIFICATION]`. These tags are secondary search targets | resolve as many as your search budget allows, replacing each tag with a verified value (or, if the value genuinely cannot be found, a short note explaining why and citing what was searched). You may also add net-new fields surfaced by your searches; do not drop or silently overwrite fields that were filled directly by 4b.

Execute your web search plan per the guide, then return the FULL updated region document as a single JSON object in a ```json fence. Do NOT output any preamble, analysis, or commentary | only the fenced JSON block.
```

Figure 17. Web search enrichment prompt. Sonnet resolves [NEEDS VERIFICATION] placeholders and coverage gaps using up to 10 web searches.

Step 5b: Per-Dataset Analysis (Sonnet)

```
You are the per-dataset content analyst for a benchmark evaluation validity assessment pipeline. You receive sampled examples and metadata from a single HuggingFace dataset, along with the deployment context. Your job is to characterize the dataset's fitness for the deployment | both where it serves the use case well and where it may fall short | grounding each observation in specific cited datapoints.

This is one dataset in a multi-dataset benchmark. A downstream aggregation step will combine your report with reports from sibling datasets. Keep your analysis focused on this dataset | cross-dataset comparisons happen later.

## The 6 validity dimensions

### Input Ontology (IO)
The set of test-case categories represented by the benchmark. A misalignment | omitting necessary categories or including irrelevant ones | may harm content validity.

### Input Content (IC)
The actual content of individual datapoints. Even with good category coverage, individual items can introduce construct-irrelevant variance through cultural, geographic, or linguistic mismatch.

### Input Form (IF)
The encoding of input signals | text vs. audio, script, resolution, dialect. A mismatch between benchmark and deployment signal distributions may reduce external validity.

### Output Ontology (OO)
The space of outputs and the decision rules by which they are scored. For free-form outputs the scoring function must interpret the denotation of the output. Mismatches in label space boundaries and decisions in interpretive steps are where validity concerns most readily arise across cultural contexts.

### Output Content (OC)
Whether ground-truth labels for particular datapoints would correlate with the judgments of regional stakeholders. Potential disagreement is a concern for both convergent and external validity.

### Output Form (OF)
The representation of output signals models are expected to produce. If the benchmark does not evaluate on the output forms encountered during deployment, this may reduce external validity.

## Your inputs
```

```
You receive four pieces of context:

1. **Benchmark YAML** | documents the
↪  benchmark's tasks, languages,
   datasets, and design. The
   ↪  `coverage_strengths` and
   `coverage_gap_analysis` sections (if
   ↪  present) map documented
   strengths and anticipated gaps to
   ↪  specific validity dimensions |
   evaluate whether the actual data
   ↪  confirms, partially addresses,
   ↪  or
   mitigates each one.

2. **Elicitation Summary** | describes
↪  the deployment scenario: who the
   target population is, what the
   ↪  system will be used for, and
   ↪  which
   validity dimensions the user
   ↪  considers highest priority. This
   ↪  is
   your primary lens for deciding what
   ↪  matters.

3. **HF Metadata** | schema, access
↪  status, splits, modalities, label
   names from the HuggingFace API.

4. **Sampled Examples** | real
↪  datapoints from the dataset,
↪  sampled with
   class stratification when labels
   ↪  exist. This is where your unique
   value lies | you can see things
   ↪  about the actual content that
   documentation and web search cannot
   ↪  reveal.

## What to look for

Use the elicitation summary,
↪  coverage_strengths, and
coverage_gap_analysis to focus your
↪  attention. The following are
↪  general axes, but weight them by
↪  what
matters for this specific deployment.
↪  For each axis, assess both
alignment and misalignment with the
↪  deployment context:

1. **Content domain and topics**: What
↪  subject matter do the examples
   cover? Where does it align with the
   ↪  deployment domain described in
   the elicitation summary? Are there
   ↪  topics the deployment needs that
   appear absent from the sample?

2. **Text register and genre**: Is this
↪  formal prose, spoken
   transcriptions, social media, legal
   ↪  documents? Does the register
   match what the deployment's target
   ↪  population would encounter?

3. **Language and cultural context**:
↪  What language(s) appear? Is there
   code-switching? Where do examples
   ↪  reflect the target population's
   context well, and where do they
   ↪  encode assumptions that may not
   ↪  hold?

4. **Annotation scheme**: What do the
↪  labels capture? Where would the
   target population's stakeholders
   ↪  agree with annotation decisions,
   ↪  and
   where might they disagree?

5. **Content coverage**: What range of
↪  topics or scenarios is visible?
   What deployment-relevant content is
   ↪  well-represented, and what gaps
   remain relative to the deployment
   ↪  needs?

## Datapoint citation format

Ground every observation in specific
↪  examples. These citations allow
human reviewers to verify your
↪  observations and serve as evidence
↪  in
the downstream aggregation:

```
- [D{n}] Example {index}
↪  (label={label}): "{short excerpt}" |
  {why this example is relevant to the
  ↪  observation}
```

Excerpts should be short (1-2 sentences)
↪ but sufficient for a human
reviewer to evaluate the observation
↪ without needing to look up the raw
data. Use citations for both strengths
↪ and concerns | both need
grounding.

Important: The excerpt must contain
↪ verbatim text from the example
in the data's own language | not an
↪ English description of what it
contains. If the data is tokenized
↪ (e.g., a `tokens` field with a list
of strings), reconstruct the original
↪ text by joining the tokens. For
example, write `"l'examen clinique
↪ montre un état général conservé"`
rather than `"Clinical examination
↪ sentence | confirms genre
↪ alignment"`.
```

```
## Finding categories

Organize findings into two categories. Hold both to the same evidentiary standard: each finding must be grounded in specific cited datapoints, whether it is a strength or a concern. However, the volume of findings in each category should reflect what the data actually shows | do not pad a thin side or suppress a dominant one to achieve symmetry.

### Deployment-Relevant Strengths
Properties of the data that serve the deployment context well. What does this dataset capture that matters for the target population and use case? Tag each strength with the dimension(s) it informs.

### Potential Concerns
Properties that may not align with deployment needs, or gaps relative to the deployment context. Tag each concern with severity and dimension(s):

- **CRITICAL**: Content directly contradicts a documented property, or fundamentally misaligns with deployment needs.
- **MAJOR**: Significant potential misalignment that warrants human review.
- **MINOR**: Noteworthy observation that is unlikely to substantially affect validity scoring.

## Output

Produce a concise report (400-800 words) with:

1. **Dataset overview** | task type, size, modality, schema (from metadata)
2. **Deployment-relevant strengths** | what the data captures well for this deployment, with datapoint citations. Tag each with the dimension(s) it informs (IO, IC, IF, OO, OC, OF). Apply the same analytical standard as concerns | grounded in specific examples, not abstract claims. The downstream scorer needs evidence from both sides to make calibrated judgments.
3. **Potential concerns** | what may not align with deployment needs, with severity-tagged findings and datapoint citations. Tag each with dimension(s).
4. **Annotation and label analysis** | what the labeling captures, strengths and potential concerns for the deployment context
5. **Datapoint citations registry** | a table of all [D{n}] citations for downstream reference. Each row MUST include the verbatim excerpt in the data's own language (the same quoted text used in the body) AND a short English interpretation. The registry is programmatically extracted by downstream modules, so English-only excerpts will cause data loss | but excerpt-only rows without interpretation lose context for non-speakers of the data language. Use this format:
   ```
ID	Dataset	Example #	Label	Excerpt	Interpretation	Dimension
D1	...	...	...	"{verbatim}"	{English gloss}	IC
   ```

Be concrete | quote from examples rather than making abstract claims. Characterize the dataset's fitness for the deployment rather than rendering final validity judgments.

---
User message:

## Dataset: {HF_DATASET_ID}

## Benchmark YAML

{BENCHMARK_YAML_TEXT}

## Deployment Context (Elicitation Summary)

{ELICITATION_SUMMARY}

## Dataset Content

### HF Metadata

{HF_METADATA_JSON}

### Sampled Examples

{CONTENT_SAMPLE_MARKDOWN}
```

Figure 18. Per-dataset analysis prompt (org mode). Called once per constituent dataset in the benchmark's HuggingFace org.

Step 5b: Cross-Dataset Aggregation (Sonnet)

```
You are the aggregation module of the dataset analysis stage. You receive
per-dataset content analyses from a multi-dataset benchmark, each produced
by a sibling module that inspected real examples from one dataset. Your job
is to triage the datasets by deployment fit and characterize the
benchmark's collective fitness | both where it serves the deployment well
and where it falls short.

Each per-dataset analysis contains severity-tagged findings with datapoint
citations ([D{n}]) grounded in specific examples. Since citation IDs are
only unique within each dataset's report, you must prefix them with a
short uppercase dataset name to disambiguate: derive `DATASET_SHORT` from
the dataset's repo ID (e.g., `DrBenchmark/QUAERO` → `QUAERO`,
`DrBenchmark/FrenchMedMCQA` → `FRENCHMEDMCQA`, `DrBenchmark/DEFT2020`
→ `DEFT2020`). Use the prefixed form `{DATASET_SHORT}-D{n}` throughout
your report.

## The 6 validity dimensions

### Input Ontology (IO)
The set of test-case categories represented by the benchmark. A
misalignment | omitting necessary categories or including irrelevant
ones | may harm content validity.

### Input Content (IC)
The actual content of individual datapoints. Even with good category
coverage, individual items can introduce construct-irrelevant variance
through cultural, geographic, or linguistic mismatch.

### Input Form (IF)
The encoding of input signals | text vs. audio, script, resolution,
dialect. A mismatch between benchmark and deployment signal
distributions may reduce external validity.

### Output Ontology (OO)
The space of outputs and the decision rules by which they are scored.
For free-form outputs the scoring function must interpret the denotation
of the output. Mismatches in label space boundaries and decisions in
interpretive steps are where validity concerns most readily arise across
cultural contexts.

### Output Content (OC)
Whether ground-truth labels for particular datapoints would correlate
with the judgments of regional stakeholders. Potential disagreement
is a concern for both convergent and external validity.

### Output Form (OF)
The representation of output signals models are expected to produce.
If the benchmark does not evaluate on the output forms encountered
during deployment, this may reduce external validity.

## Your inputs

1. **Benchmark YAML** | documents the benchmark's tasks, languages,
   datasets, and design. The `coverage_gap_analysis` section (if
   present) maps documented strengths and anticipated gaps to specific
   validity dimensions | evaluate whether the per-dataset findings
   confirm, partially address, or mitigate these.

2. **Elicitation Summary** | describes the deployment scenario: who the
   target population is, what the system will be used for, and which
   validity dimensions are highest priority. Use this to assess each
   dataset's relevance to the deployment.

3. **Web Search Findings** | upstream web research that provides
   additional context about the benchmark and deployment region.

4. **Per-Dataset Analysis Reports** | one report per dataset, each
   containing:
   - Severity-tagged findings with dimension labels
```

```
   - Datapoint citations ([D{n}])
   ↪  grounding each finding
   - Content characterization and
   ↪  annotation analysis
   These are your primary input |
   ↪  triage and synthesize across
   ↪  them.

## Your priorities

1. **Triage datasets individually.** A
↪  multi-dataset benchmark is a
   collection, not a monolith. Each
   ↪  dataset may have very different
   relevance to the deployment. Assess
   ↪  each one's fit: some may be
   strong matches, others partially
   ↪  relevant, others potentially
   inappropriate. A user needs to know
   ↪  which datasets to trust for
   their use case and which to
   ↪  reconsider.

2. **Then look for cross-cutting
↪  patterns.** After individual
↪  triage,
   identify patterns visible only at
   ↪  the benchmark level | both
   strengths and weaknesses:
   ↪  complementary coverage between
   ↪  datasets,
   shared gaps across multiple datasets,
   ↪  benchmark-level blind spots.
   These inform the overall assessment.

3. **Preserve the evidence chain.**
↪  When referencing a per-dataset
   finding, use prefixed citation IDs:
   ↪  `{DATASET_SHORT}-D{n}`, where
   `DATASET_SHORT` is a short uppercase
   ↪  name derived from the dataset
   (e.g., QUAERO, DEFT2020, CAS, ESSAI).
   ↪  Example: "QUAERO-D3 and
   DEFT2020-D7 both show...". Do NOT
   ↪  reproduce the full citation
   text | the per-dataset reports
   ↪  already contain the excerpts.

4. **Weight by deployment relevance.**
↪  Use the elicitation summary's
   priority dimensions,
   ↪  coverage_strengths, and
   ↪  coverage_gap_analysis
   to focus the aggregation on what
   ↪  matters most for this
   ↪  deployment.

5. **Balanced assessment, not verdict.**
↪  Surface both strengths and
   concerns, holding each to the same
   ↪  evidentiary standard. The
   downstream scoring model needs
   ↪  evidence from both sides to make
   calibrated judgments. Let the volume
   ↪  of each reflect what the data
   shows | do not pad a thin side or
   ↪  suppress a dominant one. Frame
   findings as observations for human
   ↪  review, not definitive verdicts.

## Output format

Produce a markdown report:

```
Dataset Analysis Report

Benchmark: {benchmark name}
Datasets analyzed: {count and
↪ names}
Analysis date: {today}

Per-Dataset Fit Assessment

For each dataset, a brief assessment of
↪ its relevance and potential
concerns for this deployment:

{Dataset A}
- **Task:** {task type}
- **Deployment fit:** {strong / partial
↪ / weak} | {1-2 sentence
  rationale referencing specific
  ↪ findings and datapoint citations}
- **Key strengths:** {bullet list of
↪ deployment-relevant properties}
- **Key concerns:** {bullet list of
↪ flagged issues, or "none flagged"}

{Dataset B}
{same structure}

...

Cross-Cutting Strengths

Patterns of deployment-relevant fitness
↪ visible across multiple
datasets, with prefixed datapoint
↪ citations. What does the benchmark
collectively capture well for this
↪ deployment? Apply the same
evidentiary standard as weaknesses |
↪ grounded in specific citations,
not abstract claims. Volume should
↪ reflect what the data shows.

Cross-Cutting Weaknesses

Patterns of misalignment visible across
↪ multiple datasets, ranked by
```
```

```
severity (CRITICAL, MAJOR, MINOR), with
↪  prefixed datapoint citations.

---

### Content Coverage Summary
{benchmark-wide characterization: what
↪  domains, registers, and cultural
contexts are collectively represented,
↪  what is well-covered, and what
gaps remain relative to the deployment
↪  needs}

---

### Limitations
{what the sample-based analysis cannot
↪  tell you | modalities not
inspectable, datasets that failed to
↪  load, coverage uncertainty}

---

### Cited Evidence

["{DATASET_SHORT}-D{n}", ...]

{A JSON array of every prefixed
↪  citation ID referenced anywhere in
↪  this
report. No excerpts | just the IDs as
↪  strings. A post-processing step
will look up the full excerpts from the
↪  per-dataset reports. The array
MUST be valid JSON.}
```

```
Omit severity sections that have no
↪  findings. Be concrete | reference
specific prefixed citation IDs rather
↪  than making abstract claims.
Characterize the benchmark's fitness for
↪  the deployment rather than
rendering final validity judgments.

---
User message:

## Benchmark YAML

{BENCHMARK_YAML_TEXT}

## Deployment Context (Elicitation
↪  Summary)

{ELICITATION_SUMMARY}

## Web Search Findings

{WEB_SEARCH_TEXT}

## Per-Dataset Analysis Reports

{SUMMARIES_BLOCK}
```

Figure 19. Cross-dataset aggregation prompt (org mode). Sonnet synthesizes per-dataset analyses into a unified report.

Step 5b: Dataset Interpretation (Sonnet)

```
You are the dataset analysis module of
↪  a benchmark validity assessment
pipeline. You receive raw examples
↪  sampled from the benchmark's
HuggingFace dataset(s), along with
↪  metadata, benchmark documentation,
and the user's deployment context. Your
↪  job is to examine the actual
data and characterize its fitness for
↪  the deployment | both where the
data serves the use case well and where
↪  it may fall short | grounding
each observation in specific cited
↪  datapoints.

## The 6 validity dimensions

### Input Ontology (IO)
The set of test-case categories
↪  represented by the benchmark. A
misalignment | omitting necessary
↪  categories or including irrelevant
ones | may harm content validity.
- IO-1: Are the test-case categories
↪  sufficient for the deployment?
- IO-2: Does the benchmark's taxonomy
↪  omit regionally relevant
↪  categories?
- IO-3: Does it include categories
↪  irrelevant to the regional context?

### Input Content (IC)
The actual content of individual
↪  datapoints | prompts, images, text
passages. Even with good category
↪  coverage, individual items can
introduce construct-irrelevant variance
↪  through cultural, geographic,
or linguistic mismatch.
- IC-1: Do inputs require
↪  region-specific cultural,
↪  geographic, or dialectal knowledge?
- IC-2: Does culturally sensitive
↪  content align with the target
↪  culture?
- IC-3: Do inputs assume
↪  Western-specific knowledge that may
↪  not transfer?

### Input Form (IF)
The encoding of input signals | text vs.
↪  audio, script, resolution,
dialect. A mismatch between benchmark
↪  and deployment signal
distributions may reduce external
↪  validity.
- IF-1: Do signal distributions match
↪  between benchmark and deployment?
```

```
- IF-2: Does regional infrastructure support the same data capture specs?
- IF-3: Are there domain-specific form differences for the deployment use case?

### Output Ontology (OO)
The space of outputs and the decision rules by which they are scored. For categorical outputs the mapping is direct; for free-form outputs the scoring function must interpret the output, and this interpretive step is where validity concerns most readily arise across cultural contexts.
- OO-1: Are output label categories regionally relevant?
- OO-2: Are categories specific to the regional context missing?
- OO-3: Do categories encode non-regional values or assumptions?

### Output Content (OC)
Whether ground-truth labels for particular datapoints would correlate with the judgments of regional stakeholders. Potential disagreement is a concern for both convergent and external validity.
- OC-1: Do ground-truth labels reflect regional stakeholder perspectives?
- OC-2: Is there potential disagreement between original and regional annotators?
- OC-3: What are the annotator demographics?

### Output Form (OF)
The representation of output signals models are expected to produce. If the benchmark does not evaluate on the output forms encountered during deployment, this may reduce external validity.
- OF-1: Does the output modality match deployment needs?
- OF-2: Are literacy rates and accessibility requirements met?
- OF-3: Are there form mismatches that may affect external validity?

## Your inputs

You receive four pieces of context:

1. **Benchmark YAML** | documents the benchmark's tasks, languages, datasets, and design. The `coverage_strengths` and `coverage_gap_analysis` sections (if present) map documented strengths and anticipated gaps to specific validity dimensions | evaluate whether the actual data confirms, partially addresses, or mitigates each one.

2. **Elicitation Summary** | describes the deployment scenario: who the target population is, what the system will be used for, and which validity dimensions the user considers highest priority. This is your primary lens for deciding what matters.

3. **Web Search Findings** | upstream web research on the benchmark and deployment context. May contain information about known limitations, regional relevance, or related work that informs your analysis.

4. **Dataset Content** | HF metadata (schema, access, labels) and sampled examples from the dataset. For single-dataset mode, you receive these directly. For org/multi-dataset mode, you receive per-dataset content analyses produced by a sibling module.

## Your priorities

1. **Content over statistics.** You are reading actual examples from the dataset. Your primary value is noticing things that documentation review and web search cannot: cultural assumptions embedded in content, domain register mismatches, annotation patterns that may not hold for the target population, topic coverage gaps relative to the deployment needs described in the elicitation.

2. **Weight by deployment relevance.** The elicitation summary describes how the user intends to deploy this benchmark, and flags which dimensions are highest priority. The coverage_gap_analysis (if present) identifies anticipated strengths and gaps | evaluate
```

```
   whether the data confirms or
   ↪ mitigates them. Focus your
   ↪ attention
   accordingly.

3. **Cite specific datapoints.** Every
↪ observation must be grounded in
   concrete examples from the data.
   ↪ Produce datapoint citations that
   the downstream scoring step can
   ↪ reference as evidence, analogous
   ↪ to
   how paper quotes ([Q1]) and web
   ↪ search URLs are cited.

4. **Balanced assessment, not verdict.**
↪ Your role is to characterize
   the dataset's fitness for the
   ↪ deployment | surfacing both
   ↪ strengths
   and concerns, holding each to the
   ↪ same evidentiary standard. The
   downstream scoring model needs
   ↪ evidence from both sides to make
   calibrated judgments. Let the volume
   ↪ of each reflect what the data
   shows | do not pad a thin side or
   ↪ suppress a dominant one. Frame
   findings as observations for human
   ↪ review, not definitive verdicts.

## Datapoint citation format

For each observation, cite the specific
↪ example(s) that support it:

```
#### {Strength or Concern} {N}: {title}
- **Dimension(s):** {IO, IC, IF, OO, OC,
↪ OF}
- **Observation:** {what you noticed}
- **Deployment relevance:** {why this
↪ matters given the deployment
  context from the elicitation | for
  ↪ strengths, how it serves the use
  case; for concerns, why it may
  ↪ warrant attention}
- **Datapoint citations:**
  - [D{n}] Example {index} ({dataset},
  ↪ split={split}, label={label}):
    "{excerpt from the example}" | {why
    ↪ this example is relevant}
```

Excerpts should be short (1-2 sentences)
↪ but sufficient for a human
reviewer to evaluate the observation
↪ without needing to look up the raw
data.

Important: The excerpt must contain
↪ verbatim text from the example
in the data's own language | not an
↪ English description of what it
contains. If the data is tokenized
↪ (e.g., a `tokens` field with a list
of strings), reconstruct the original
↪ text by joining the tokens. For
example, write `"l'examen clinique
↪ montre un état général conservé"`
rather than `"Clinical examination
↪ sentence | confirms genre
↪ alignment"`.

Finding categories

Organize findings into two categories.
↪ Hold both to the same evidentiary
standard: each finding must be grounded
↪ in specific cited datapoints,
whether it is a strength or a concern.
↪ However, the volume of findings
in each category should reflect what
↪ the data actually shows | do not
pad a thin side or suppress a dominant
↪ one to achieve symmetry.

Deployment-Relevant Strengths
Properties of the data that serve the
↪ deployment context well. What
does this dataset capture that matters
↪ for the target population and
use case? Tag each strength with the
↪ dimension(s) it informs.

Potential Concerns
Properties that may not align with
↪ deployment needs, or gaps relative
to the deployment context. Tag each
↪ concern with severity and
dimension(s):

- **CRITICAL**: Content directly
↪ contradicts a documented property,
↪ or
  fundamentally misaligns with
  ↪ deployment needs.
- **MAJOR**: Significant potential
↪ misalignment that warrants human
  review.
- **MINOR**: Noteworthy observation
↪ that is unlikely to substantially
  affect validity scoring.

Output format

Produce a markdown report:

```
## Dataset Analysis Report

**Dataset(s):** {dataset name(s)}
**Analysis date:** {today}
**Examples reviewed:** {count per
↪ dataset}
**Columns shown:** {list}
**Columns skipped (media):** {list, if
↪ any}
```
```

```
---

### Datapoint Citations Registry

| ID | Dataset | Example # | Label |
↪  Excerpt | Interpretation |
↪  Dimension |
|----|---------|-----------|-------|---------
| D1 | ... | ... | ... | "{verbatim}" |
↪  {English gloss} | IC |
| D2 | ... | ... | ... | "{verbatim}" |
↪  {English gloss} | OO |
{one row per cited datapoint | Excerpt
↪  MUST be verbatim text in the
data's own language; Interpretation is
↪  a short English gloss so
non-speakers can understand the
↪  citation}

---

### Deployment-Relevant Strengths
{findings with datapoint citations,
↪  tagged by dimension | same
evidentiary standard as concerns,
↪  volume proportional to what the data
shows}

---

### Potential Concerns

#### CRITICAL
{findings with datapoint citations}

#### MAJOR
{findings with datapoint citations}

#### MINOR
{findings with datapoint citations}

---

### Content Coverage Summary
{brief characterization of what the
↪  data contains: topics, register,
language, cultural context | grounded
↪  in examples}

---

### Limitations
{what the sample cannot tell you |
↪  modalities not inspectable,
↪  coverage
uncertainty from sample size, etc.}
```

Omit severity sections that have no
↪ findings. Be concrete | quote from
the examples rather than making
↪ abstract claims. Characterize the
dataset's fitness for the deployment
↪ rather than rendering final
validity judgments.

User message:

Benchmark YAML

{BENCHMARK_YAML_TEXT}

**## Deployment Context (Elicitation
↪ Summary)**

{ELICITATION_SUMMARY}

Web Search Findings

{WEB_SEARCH_TEXT}

Dataset Content

HF Metadata

{HF_METADATA_JSON}

Sampled Examples

{CONTENT_SAMPLE_MARKDOWN}

Figure 20. Single-dataset interpretation prompt. Used for benchmarks with a single HuggingFace dataset (no aggregation step).

Step 6: Composed Evaluation Prompt (Script)

I need you to perform a validity
↪ analysis of an AI benchmark for a
↪ specific target region.

Instructions

You are evaluating whether the benchmark
↪ **{BENCHMARK_FULL_NAME}** is valid
↪ for use in **{REGION_NAME}**.

Analyze the evidence sources below
↪ against each of the 6 validity
↪ dimensions. For each dimension,
↪ assign a score (1-5), provide
↪ justification with cited evidence,
↪ respond to each checklist item, and
↪ identify information gaps.

Critical Constraints

```
- **Evidence-grounded only**: Base your analysis ONLY on the three evidence sources provided: (1) benchmark documentation with verbatim quotes, (2) regional context with web-sourced findings, and (3) dataset analysis findings with datapoint citations (if present). Do NOT role-play as a member of the target culture or speculate beyond what these sources support.
- **Cite evidence**: For each finding, cite at least one source | a verbatim quote `[QN]`, a web source `[WEB-N]`, or a dataset citation `DATASET-D{n}`.
- **Flag gaps explicitly**: When none of the three evidence sources addresses a checklist item, say "INSUFFICIENT DOCUMENTATION" and describe what would be needed.
- **Distinguish documentable vs. expert-needed**: Classify each finding as either (a) determinable from documentation, or (b) requiring regional expert verification.
- **Regional specificity**: Evaluate validity *for the specified target region*, not in general.
- **Conservative scoring**: When evidence is ambiguous or insufficient, score lower rather than higher.

### Scoring Rubric

- **1**: Major validity violations identified; benchmark component fundamentally misaligned with target context.
- **2**: Significant concerns; multiple concrete violations or gaps identified.
- **3**: Partially addressed; mixed evidence; some alignment, some gaps.
- **4**: Well addressed; minor concerns or gaps; documentation shows awareness of regional needs.
- **5**: Fully addressed; no concerns identified; documentation demonstrates explicit validity-preserving practices.

### Evidence Sources

The prompt below contains three evidence sources:

1. **Benchmark Documentation** + **Verbatim Quote Registry** | paper content, with authoritative quotes labeled `[QN]`
2. **Regional Context** (YAML) + **Web Source Registry** | deployment context with web research findings cited as `[WEB-N]`
3. **Dataset Analysis Findings** (if present) | empirical observations from the benchmark's HuggingFace data, cited as `DATASET-D{n}`

Citation rules for each source are in your system instructions.

---

## Benchmark Information

- **Name**: {BENCHMARK_ID}
- **Full Name**: {BENCHMARK_FULL_NAME}
- **Domain**: {DOMAIN}
- **Languages**: {LANG_1}, {LANG_2}
- **Porting Strategy**: {PORTING_STRATEGY}
- **Year**: {YEAR}

### Benchmark Documentation

#### Input Ontology

{Interpretive narrative for input ontology, referencing quote IDs [Q1], [Q2], etc.}

#### Input Content

{Interpretive narrative for input content}

#### Input Form

{Interpretive narrative for input form}

#### Output Ontology

{Interpretive narrative for output ontology}

#### Output Content

{Interpretive narrative for output content}

#### Output Form

{Interpretive narrative for output form}

### Verbatim Quote Registry

| ID | Page | Dimension | Text |
|----|------|-----------|------|
```

```
| Q1 | 1 | input_ontology | "{Exact
↪  verbatim quote from the paper}" |
| Q2 | 3 | input_content | "{Exact
↪  verbatim quote from the paper}" |

---

## Regional Context

name: '{REGION_NAME}'
country: '{COUNTRY}'
population_description: '{Population
↪  description from elicitation + web
↪  search}'
language:
  primary: '{PRIMARY_LANGUAGE}'
  literacy_rate: '{N%} [WEB-1]'
infrastructure:
  internet_penetration: '{M%} [WEB-2]'

### Web Source Registry

| ID | URL |
|----|-----|
| WEB-1 |
↪  https://source-url-1.example.com |
| WEB-2 |
↪  https://source-url-2.example.com |

---

## Expert Elicitation

## Elicitation Responses

{Question-answer pairs}

## Dimension Priority Weights

{Priority assignments}

---

## Dataset Analysis Findings

The following empirical findings were
↪  produced by automated profiling
↪  scripts that
sampled the benchmark's actual dataset
↪  on HuggingFace. Observations cite
↪  specific
datapoints using `DATASET-D{n}` IDs
↪  (e.g., QUAERO-D3). Both
↪  deployment-relevant
strengths and potential concerns are
↪  included; weigh CRITICAL-tagged
↪  concerns and
well-evidenced strengths accordingly
↪  when scoring the affected
↪  dimensions.

{DATASET_ANALYSIS_REPORT | empirical
↪  findings from Step 5b HuggingFace
↪  data profiling}
```

```
---

## Framework Dimensions to Evaluate

### Input Ontology

**Definition**: The input data ontology
↪  consists of the set of test case
↪  categories represented by the
↪  benchmark, which should cover the
↪  query types that evaluated systems
↪  are expected to encounter during
↪  deployment (e.g., factual questions,
↪  creative writing help, or
↪  restaurant recommendations for NLP
↪  dialogue systems; pick-and-place
↪  tasks, point navigation, or
↪  planning for robotics).

**Theoretical Importance**: A
↪  misalignment in this taxonomy |
↪  whether through the omission of
↪  necessary categories (construct
↪  underrepresentation) or the
↪  inclusion of irrelevant ones
↪  (construct-irrelevant variance) |
↪  harms the content validity of the
↪  benchmark.

**Checklist:**
- [IO-1] Identify test case categories
↪  required for regional deployment.
- [IO-2] Check if source benchmark's
↪  taxonomy omits regionally-relevant
↪  categories.
- [IO-3] Check if source benchmark
↪  includes categories irrelevant to
↪  regional context.
- [IO-4] Document category gaps that
↪  would harm content validity.

### Input Content

**Definition**: Whereas the ontology
↪  defines the types of input queries,
↪  the content of dataset inputs
↪  refers to the explicit instances
↪  specified by individual datapoints |
↪  an LLM prompt, an image, a database
↪  entry, etc.

**Theoretical Importance**: Even when
↪  the test case taxonomy provides
↪  good coverage, implementation-level
↪  details of individual datapoints
↪  can introduce construct-irrelevant
↪  variance, violating content
↪  validity.

**Checklist:**
- [IC-1] Determine if input queries
↪  require region-specific cultural,
↪  geographic, or dialectal knowledge.
```

```
- [IC-2] Assess whether culturally
↪  sensitive content aligns with
↪  target deployment culture.
- [IC-3] Flag inputs requiring
↪  Western-specific knowledge that may
↪  not transfer.
- [IC-4] Recruit regional annotators to
↪  identify culturally sensitive
↪  instances if resources permit.
- [IC-5] Document content issues that
↪  would harm content validity.

### Input Form

**Definition**: The form of dataset
↪  inputs determines the encoding of
↪  the input signal | e.g., text vs.
↪  audio for natural language, or
↪  camera parameters such as focal
↪  length and resolution for visual
↪  data.

**Theoretical Importance**: Since
↪  machine learning systems are
↪  sensitive to signal distributions,
↪  a mismatch between the benchmark's
↪  input representation and the
↪  real-world signals that deployed
↪  models would encounter violates the
↪  external validity of the
↪  evaluation.

**Checklist:**
- [IF-1] Compare signal distributions
↪  (e.g., image resolution, MRI field
↪  strength) between source and target
↪  contexts.
- [IF-2] Check if regional
↪  infrastructure supports the same
↪  data capture specifications.
- [IF-3] Identify domain-specific form
↪  differences relevant to the
↪  intended use case.
- [IF-4] Document form mismatches that
↪  would harm external validity.

### Output Ontology

**Definition**: A benchmark's output
↪  ontology determines the space of
↪  outputs an AI system is expected to
↪  produce and the decision rules by
↪  which those outputs are organized
↪  and scored | i.e., the benchmark's
↪  criteria. For categorical outputs,
↪  the mapping is direct (e.g.,
↪  safe/unsafe, or object class
↪  labels). For free-form outputs, the
↪  scoring function must first
↪  interpret what the output means
↪  before mapping it to a score | and
↪  this interpretive step is where
↪  validity violations most readily
↪  arise, since decision rules can
↪  differ across cultural contexts. For
↪  instance, an LLM recommending "red
↪  wine" for a dinner party may score
↪  highly for helpfulness in a Western
↪  context but poorly where alcohol
↪  consumption is prohibited.

**Theoretical Importance**: A
↪  misaligned output taxonomy thus
↪  violates structural validity (the
↪  construct's structure is
↪  misrepresented), content validity
↪  (through missing or irrelevant
↪  categories), and risks violating
↪  external validity (benchmark
↪  performance is less likely to
↪  generalize to regional settings).

**Checklist:**
- [OO-1] Review output label categories
↪  for regional relevance.
- [OO-2] Identify missing categories
↪  specific to regional contexts (e.g.,
↪  autorickshaws in Indian driving
↪  data).
- [OO-3] Flag categories that encode
↪  non-regional values or assumptions.
- [OO-4] Consider stakeholder-driven
↪  taxonomy redesign if significant
↪  misalignment exists.
- [OO-5] Document taxonomy issues that
↪  would harm structural validity and
↪  content validity.

### Output Content

**Definition**: Whereas taxonomic
↪  alignment addresses whether
↪  abstract decision boundaries
↪  reflect regional values, label
↪  correctness concerns whether the
↪  labels for particular datapoints
↪  correlate with the judgments of
↪  regional stakeholders.
```

```
**Theoretical Importance**:
↪ Disagreement between regional and
↪ original annotators violates both
↪ convergent validity (the labels
↪ fail to correlate with regional
↪ perspectives on the construct) and
↪ external validity (the original
↪ judgments do not generalize to the
↪ target context).

**Checklist:**
- [OC-1] Determine if ground truth
↪ labels reflect regional stakeholder
↪ perspectives.
- [OC-2] Assess potential disagreement
↪ between original annotators and
↪ regional population.
- [OC-3] Review annotator demographics
↪ in benchmark documentation
↪ (Datasheets, Data Statements).
- [OC-4] Consider label re-annotation by
↪ representative regional annotator
↪ pool.
- [OC-5] Review aggregation methods for
↪ potential erasure of minority
↪ perspectives.
- [OC-6] Document label issues that
↪ would harm convergent validity and
↪ external validity.

### Output Form

**Definition**: The form of dataset
↪ outputs pertains to the
↪ representation of output signals
↪ models are expected to produce.

**Theoretical Importance**: If a
↪ benchmark does not evaluate models
↪ on the output forms encountered
↪ during real-world deployment, this
↪ violates the external validity of
↪ the evaluation.

**Checklist:**
- [OF-1] Check if expected output
↪ modality matches regional
↪ deployment needs.
- [OF-2] Assess text-to-speech
↪ availability for speech-based
↪ output requirements.
- [OF-3] Consider literacy rates and
↪ accessibility requirements in
↪ target population.
- [OF-4] Document form mismatches that
↪ would harm external validity.


---

## Required Output Format
```

```
Output a single valid JSON object with
↪ this structure:

{
  "benchmark": "{BENCHMARK_ID}",
  "region": "{REGION_NAME}",
  "dimensions": {
    "input_ontology": {
      "score": "<integer 1-5>",
      "justification": "...",
      "strengths": ["what this
      ↪ dimension captures well for
      ↪ the target context"],
      "checklist_responses": { "IO-1":
      ↪ "...", "IO-2": "..." },
      "evidence_quotes": ["[Q1] 'quote
      ↪ text' (p.7)", ...],
      "evidence_web_sources": ["[WEB-1]
      ↪ literacy rate 96%", ...],
      "evidence_dataset": ["DATASET-D1:
      ↪ observation", ...],
      "evidence_map": { "IO-1": ["Q1",
      ↪ "WEB-3"], "IO-2":
      ↪ ["DATASET-D1"] },
      "confidence":
      ↪ "<high|medium|low>",
      "information_gaps": ["..."],
      "requires_expert_verification":
      ↪ ["..."]
    },
    "input_content": { "..." },
    "input_form": { "..." },
    "output_ontology": { "..." },
    "output_content": { "..." },
    "output_form": { "..." }
  },
  "overall_summary": "...",
  "risk_assessment":
  ↪ "<high|medium|low>",
  "practical_guidance": {
    "what_this_benchmark_measures":
    ↪ "...",
    "construct_depth": "...",
    "supplementation_needed": "..."
  },
  "remediation_suggestions": [
    { "dimension": "...", "gap": "...",
    ↪ "recommendation": "..." }
  ],
  "highest_concern_dimensions":
  ↪ ["..."],
  "strongest_dimensions": ["..."]
}
```

Figure 21. Deterministic prompt assembly output from `compose_prompt.py`. Per-assessment placeholders shown in {UPPERCASE} notation; framework dimensions and scoring rubric are verbatim from `framework.yaml`. This output becomes the user message for Step 7.

Step 7: Opus Scoring

```
You are the validity analysis scorer for a benchmark/deployment pair. The user
message contains the full composed prompt from the pipeline | benchmark
documentation, verbatim quote registry, region YAML (with web search findings
and source URLs), deployment context from elicitation, dataset analysis
findings (if available), and the 6-dimension framework. Score each dimension
1-5 and produce a single valid JSON document.

## Four critical rules

**1. Deployment context conditions the analysis.**
The composed prompt contains a "Deployment Context" section with the user's
specific use case, target population, and dimension priority weights from
elicitation. Your validity analysis must evaluate the benchmark's fitness for
THIS specific deployment | not in the abstract. Dimensions marked HIGH priority
in the elicitation should receive the most thorough analysis. Flagged gaps from
elicitation should be explicitly addressed in your assessment.

**2. Quote provenance is critical.**
The composed prompt separates two sections: "Benchmark Documentation"
(interpretive context with `[QN]` references) and "Verbatim Quote Registry"
(authoritative evidence from the original paper). When you populate
`evidence_quotes` in your JSON output, you MUST cite only text from the Verbatim
Quote Registry, formatted as `"[QN] 'exact quote text' (p.X)"`. Do NOT cite the
interpretive context as evidence. If no verbatim quote supports a finding, say
so explicitly.

**3. Web-sourced claims must cite their registry ID.**
The composed prompt includes a "Regional Context" section containing the region
YAML followed by a "Web Source Registry" table. The YAML was produced by an
upstream pipeline that combined deployment context with web search research.
Web-sourced claims in the YAML are tagged with `[WEB-N]` IDs that map to full
URLs in the registry. When you use a regional fact as evidence, populate
`evidence_web_sources` with an entry that leads with the registry ID. Example:

```
"evidence_web_sources": [
  "[WEB-3] literacy rate 96%",
  "[WEB-7] UU 27/2022 PDP effective Oct 2022"
]
```

Region fields tagged `[NEEDS VERIFICATION]`, `[NEEDS VERIFICATION | deferred:
...]`, or `[NOT FOUND | ...]` are NOT evidence | they indicate the upstream
pipeline could not ground that slot. Treat each such tag as an information gap:
record the underlying question under `information_gaps` rather than citing the
tag.

Lead `evidence_web_sources` entries with the `[WEB-N]` ID, then the factual
claim in natural language. Do not lead with internal field names or gap
identifiers:

- **Wrong:** `"gap_1_pemda_procedure_content CONFIRMED ABSENT | no Indonesian
  administrative procedure benchmark exists [WEB-12]"`
- **Right:** `"[WEB-12] No Indonesian administrative-procedure benchmark
  exists"`

4. Dataset analysis findings are empirical evidence.
The composed prompt may contain a "Dataset Analysis Findings" section with
observations from automated scripts that sampled the benchmark's actual data on
HuggingFace. These findings cite specific datapoints using prefixed IDs
```

(`{DATASET}-D{n}`, e.g., `QUAERO-D3`, `CAS-D15`). When you use a dataset analysis finding as evidence, populate `evidence_dataset` with entries that reference the citation ID and summarize the observation. Example:

```
"evidence_dataset": [
  "QUAERO-D3: ATC codes appear in text but are not annotated as entities | taxonomy gap for regulatory NER",
  "E3C-D8: patient from Morocco | confirms non-metropolitan francophone content"
]
```

Dataset analysis provides empirical ground truth that may confirm or contradict claims in the benchmark documentation. When empirical findings conflict with documented properties (e.g., actual language distribution differs from claimed languages), treat the empirical evidence as more authoritative and note the discrepancy explicitly.

If no "Dataset Analysis Findings" section is present, leave `evidence_dataset` as an empty list | the pipeline does not always have HuggingFace data available.

Scoring guidance

Confidence calibration. Set `confidence` per dimension:
- `high` | finding is directly supported by at least two evidence streams (verbatim quotes, web sources, dataset analysis findings).
- `medium` | finding is supported by one evidence stream, or evidence is consistent but indirect.
- `low` | finding rests mostly on inference; no evidence stream directly addresses the checklist item.

Checklist coverage. Respond to every checklist item for every dimension | do not skip items. If documentation is silent on a specific item, the response is `"INSUFFICIENT DOCUMENTATION"` with a short note on what would resolve it (e.g., "paper does not specify annotator demographics; would need demographic breakdown of the annotator pool").

Checklist grounding. Every substantive checklist response must be grounded in at least one of: (a) a Verbatim Quote Registry entry, (b) a URL-cited web source, (c) a dataset analysis finding with citation ID, or (d) an explicit `"INSUFFICIENT DOCUMENTATION"` marker. Do not make interpretive claims ("consumes evaluation weight", "implicitly assumes X", "encodes Western defaults") without a quote, web source, or dataset citation to anchor them. If the anchor is missing, either weaken the claim or mark the item as insufficient.

Strengths, not just deficits. For each dimension, populate `strengths` with what the benchmark *does* capture well for the target context. Ground each strength in at least one evidence source. A dimension with a low score can still have strengths worth noting | partial coverage is still useful signal for the user.

Hedge consistency. If a claim is hedged in a checklist response ("may be", "suggests", "likely"), preserve the same hedging when restating it in the `justification`. Do not firm up inferences when they move from checklist to justification. Inferences that rest on indirect evidence belong in `information_gaps` or `requires_expert_verification`, not as confident statements in the justification.

Evidence map. For each dimension, populate `evidence_map` linking every checklist item ID to the evidence entries that support its response. Use the evidence IDs directly: `Q3` for verbatim quotes, `WEB-3` for web sources, `DATASET-D5` for dataset findings. If a checklist item is marked

```
`INSUFFICIENT DOCUMENTATION`, map it to
↪  an empty list. Every substantive
checklist response must have at least
↪  one evidence ID in the map.

## Required JSON output schema

{
  "benchmark": "<benchmark_name>",
  "region": "<region_name>",
  "dimensions": {
    "input_ontology": {
      "score": <integer 1-5, not a
      ↪  string>,
      "justification": "...",
      "strengths": ["what this
      ↪  dimension captures well for
      ↪  the target context, grounded
      ↪  in evidence"],
      "checklist_responses": { "IO-1":
      ↪  "...", "IO-2": "..." },
      "evidence_quotes": ["[Q1] 'quote
      ↪  text' (p.7)"],
      "evidence_web_sources": ["[WEB-3]
      ↪  literacy rate 96%"],
      "evidence_dataset": ["QUAERO-D3:
      ↪  ATC codes in text but not
      ↪  annotated | taxonomy gap"],
      "evidence_map": { "IO-1": ["Q1",
      ↪  "WEB-3"], "IO-2":
      ↪  ["DATASET-D5"] },
      "confidence":
      ↪  "<high|medium|low>",
      "information_gaps": ["..."],
      "requires_expert_verification":
      ↪  ["..."]
    },
    "input_content": { "...": "same
    ↪  shape" },
    "input_form": { "...": "same shape"
    ↪  },
    "output_ontology": { "...": "same
    ↪  shape" },
    "output_content": { "...": "same
    ↪  shape" },
    "output_form": { "...": "same shape"
    ↪  }
  },
  "overall_summary": "...",
  "risk_assessment":
  ↪  "<high|medium|low>",
  "practical_guidance": {
    "what_this_benchmark_measures":
    ↪  "what constructs/abilities this
    ↪  benchmark can help evaluate in
    ↪  the target context, referencing
    ↪  the strongest dimensions",
    "construct_depth": "how deeply the
    ↪  benchmark probes those
    ↪  constructs | what level of
    ↪  evidence it provides vs. what
    ↪  remains unaddressed",
    "supplementation_needed": "what
    ↪  additional evaluations or
    ↪  evidence sources would be
    ↪  needed for a complete
    ↪  assessment, tied to specific
    ↪  dimension-level gaps"
  },
  "remediation_suggestions": [
    {
      "dimension": "input_ontology",
      "gap": "what is missing or
      ↪  problematic",
      "recommendation": "concrete
      ↪  action to address it"
    }
  ],
  "highest_concern_dimensions":
  ↪  ["..."],
  "strongest_dimensions": ["..."]
}

`evidence_quotes`,
↪  `evidence_web_sources`, and
↪  `evidence_dataset` are all
optional (empty list is valid), but at
↪  least one of the three should be
populated for any finding scored 4 or
↪  5.

`strengths` should contain at least one
↪  entry for every dimension, even
low-scoring ones | partial coverage is
↪  still useful signal. Ground each
strength in evidence.

`evidence_map` must cover every
↪  checklist item ID. Items marked
`INSUFFICIENT DOCUMENTATION` map to an
↪  empty list `[]`.

`remediation_suggestions` is an array of
↪  objects, each with `dimension`,
`gap`, and `recommendation`. Include 4{6
↪  items. Before adding a seventh,
check whether it collapses into an
↪  existing one.

`practical_guidance` synthesizes across
↪  dimensions to answer: what can a
user learn from this benchmark for their
↪  context, how deep does that
evidence go, and what else would they
↪  need? Reference specific dimensions
to anchor claims in the framework.

## Before emitting the JSON

Run through this checklist; revise if
↪  any item fails:

1. Every score-4 or score-5 finding has
↪  at least one `evidence_quotes`,
```

```
   `evidence_web_sources`, or
   ↪  `evidence_dataset` entry
   ↪  populated.
2. Every substantive checklist response
↪  is grounded in a quote, region
   source, dataset citation, or
   ↪  explicitly marked `"INSUFFICIENT
   ↪  DOCUMENTATION"`.
3. No `evidence_web_sources` entry leads
↪  with a region YAML field name or
   internal gap identifier.
4. Hedging present in a checklist
↪  response is preserved in the
↪  corresponding
   justification.
5. Every dimension has at least one
↪  `strengths` entry, grounded in
↪  evidence.
6. Every checklist item ID appears in
↪  `evidence_map` (empty list for
   `INSUFFICIENT DOCUMENTATION` items).
7. `remediation_suggestions` has 4{6
↪  objects, each with `dimension`,
↪  `gap`,
   and `recommendation`.
8. `practical_guidance` references
↪  specific dimensions to anchor its
↪  claims.
9. Every `score` is a bare integer (e.g.
↪  `2`), not a quoted string (`"2"`).

## Output

Output the JSON document wrapped in a
↪  ```json fenced block so it can be
extracted cleanly. No commentary
↪  outside the fence.

---
User message:

{COMPOSED_PROMPT}

[The user message is the complete
↪  output of Step 6 (compose_prompt.py)
↪  | a self-contained markdown document
↪  containing all evidence sources,
↪  framework dimensions, and output
↪  schema. See
↪  Figure~\ref{fig:step_6_compose} for
↪  the full template.]
```

Figure 22. Opus validity scoring prompt. This is the single Opus call in the pipeline. The user message is the full output of Step 6 (see Figure 21).

D.3. LLM Judge Case Studies

D.3.1. IberBench

We present an example validity assessment performed by Claude Opus 4.6[4].

IberBench Validity Analysis Output — truncated

```
{
  "dimensions": [
    {
      "dimension": "Input Ontology",
      "score": 3,
      "justification": "IberBench
      ↪  provides broad task coverage
      ↪  with 22 NLP categories
      ↪  spanning both fundamental and
      ↪  industry-relevant tasks,
      ↪  which is commendable for the
      ↪  Iberian region. However, the
      ↪  benchmark itself acknowledges
      ↪  significant imbalances: many
      ↪  task-language pairs are
      ↪  missing, the heavy skew
      ↪  toward text classification (97
      ↪  of 101 datasets)
      ↪  underrepresents generative
      ↪  and structured prediction
      ↪  tasks, and key Spanish
      ↪  varieties (Peruvian,
      ↪  Uruguayan, Costa Rican) have
      ↪  substantial gaps. The taxonomy
      ↪  is well-suited for
      ↪  Spain-centric Spanish NLP but
      ↪  shows construct
      ↪  underrepresentation for the
      ↪  broader Ibero-American
      ↪  deployment context.",
      "checklist_responses": "...",
      "evidence_quotes": [
        "There is a high imbalance in
        ↪  terms of (task category,
        ↪  language) pairs. For
        ↪  instance, Mental Health
        ↪  Detection only comprises
        ↪  data in Spanish but
        ↪  Sentiment and Emotion
        ↪  Analysis comprises data in
        ↪  near all Iberian
        ↪  language.",
        "Industry-relevant tasks and
        ↪  certain Spanish varieties,
        ↪  particularly Peruvian,
        ↪  Uruguayan, and Costa Rican
        ↪  Spanish, show the most
        ↪  gaps.",
```

[4] https://www.anthropic.com/news/claude-opus-4-6

```
        "Ninety-seven of the datasets
        ↪  included in IberBench are
        ↪  framed as text
        ↪  classification tasks, 3 as
        ↪  text generation, and 1 as
        ↪  sequence labeling. These
        ↪  proportions reflect the
        ↪  greater difficulty involved
        ↪  in creating text generation
        ↪  and sequence labeling
        ↪  datasets.",
        "Absence of specific language
        ↪  varieties, such as Brazilian
        ↪  Portuguese and Argentinian
        ↪  Spanish."
      ],
      "confidence": "high",
      "information_gaps": [
        "No detailed breakdown of which
        ↪  task categories are
        ↪  available for which
        ↪  languages/varieties beyond
        ↪  the high-level statements
        ↪  quoted.",
        "No documentation of how the 22
        ↪  task categories were
        ↪  selected or whether a
        ↪  regional needs assessment
        ↪  informed the taxonomy
        ↪  design.",
        "No information about planned
        ↪  timeline for filling
        ↪  acknowledged gaps in
        ↪  task-language coverage."
      ],
      "requires_expert_verification": [
        "Whether the 22 task categories
        ↪  adequately cover the most
        ↪  common NLP deployment
        ↪  scenarios in Latin American
        ↪  countries specifically (vs.
        ↪  Iberian Peninsula).",
        "Whether code-switching and
        ↪  dialectal identification
        ↪  tasks are high-priority for
        ↪  regional stakeholders.",
        "Whether the absence of tasks
        ↪  for indigenous languages
        ↪  spoken alongside
        ↪  Spanish/Portuguese (e.g.,
        ↪  Quechua, Guarani, Nahuatl)
        ↪  constitutes a significant
        ↪  gap for claimed
        ↪  Ibero-American scope."
      ]
    },
    {
      "dimension": "Input Content",
      "score": 3,
      "justification": "IberBench draws
      ↪  58 of its 101 datasets from
      ↪  Iberian-origin shared tasks
      ↪  (IberLEF, IberEval, TASS,
      ↪  PAN), which provides genuinely
      ↪  regional content for Iberian
      ↪  Peninsula languages. This is
      ↪  a notable strength over purely
      ↪  translated benchmarks.
      ↪  However, the content is
      ↪  heavily skewed toward
      ↪  Peninsular Spanish (~60% of
      ↪  all samples are Spanish),
      ↪  with minimal representation of
      ↪  Latin American cultural
      ↪  contexts. The remaining 43
      ↪  datasets from multilingual
      ↪  benchmarks (Belebele, XNLI,
      ↪  La Leaderboard) may contain
      ↪  translated or culturally
      ↪  non-specific content. No
      ↪  documentation discusses
      ↪  whether input content was
      ↪  reviewed for cultural
      ↪  appropriateness across the
      ↪  full claimed Ibero-American
      ↪  scope.",
      "checklist_responses": "...",
      "evidence_quotes": [
        "We carefully collected and
        ↪  prepared a set of 101
        ↪  existing datasets to make
        ↪  them suitable for LLM
        ↪  evaluation.",
        "[Datasets from] IberLEF,
        ↪  IberEval, TASS, and PAN,
        ↪  spanning editions from 2014
        ↪  to 2024, co-located with
        ↪  the Spanish Society for
        ↪  Natural Language Processing
        ↪  (SEPLN) and the Conference
        ↪  and Labs of the Evaluation
        ↪  Forum (CLEF).",
        "Dataset acceptance criteria:
        ↪  (i) focus specifically on
        ↪  Iberian languages, (ii)
        ↪  originate from workshops
        ↪  organized in the Iberian
        ↪  Peninsula or Ibero-America,
        ↪  and (iii) target domains,
        ↪  language varieties, or
        ↪  tasks not yet
        ↪  represented.",
        "Absence of specific language
        ↪  varieties, such as Brazilian
        ↪  Portuguese and Argentinian
        ↪  Spanish.",
        "We contacted the organizers of
        ↪  all shared tasks from these
        ↪  workshops to obtain explicit
        ↪  permission for inclusion in
        ↪  IberBench."
      ],
```

```
      "confidence": "medium",
      "information_gaps": [
        "No information about the
        ↪ actual textual content of
        ↪ individual datasets (topics,
        ↪ domains, cultural
        ↪ references contained in
        ↪ inputs).",
        "No documentation of whether
        ↪ the 43
        ↪ multilingual-benchmark
        ↪ datasets were reviewed for
        ↪ cultural fit with
        ↪ Iberian/Latin American
        ↪ contexts.",
        "No data statements or
        ↪ datasheets for any of the
        ↪ 101 constituent datasets at
        ↪ the benchmark level.",
        "No information about temporal
        ↪ coverage of content (e.g.,
        ↪ whether social media data
        ↪ reflects current or
        ↪ outdated cultural
        ↪ discourse)."
      ],
      "requires_expert_verification": [
        "Whether humor, irony, and
        ↪ toxicity datasets from
        ↪ Iberian Peninsula workshops
        ↪ produce valid evaluations
        ↪ when applied to Latin
        ↪ American Spanish deployment
        ↪ contexts.",
        "Whether sentiment and emotion
        ↪ categories in the workshop
        ↪ datasets reflect Latin
        ↪ American emotional
        ↪ expression norms.",
        "Whether the XNLI/Belebele
        ↪ content contains cultural
        ↪ assumptions that are
        ↪ inappropriate for any
        ↪ Iberian/Latin American
        ↪ subpopulation.",
        "Whether the 2014-2024 temporal
        ↪ range of workshop datasets
        ↪ introduces outdated
        ↪ cultural references or
        ↪ discourse norms."
      ]
    },
    {
      "dimension": "Input Form",
      "score": 4,
      "justification": "IberBench's
      ↪ input form is well-aligned
      ↪ with the Iberian regional
      ↪ context. All inputs are
      ↪ text-based using the Latin
      ↪ alphabet with standard
      ↪ diacritics, which matches the
      ↪ writing systems of all target
      ↪ languages. The normalization
      ↪ pipeline explicitly handles
      ↪ encoding artifacts (mojibake,
      ↪ mixed encodings), which is
      ↪ important for Iberian
      ↪ language texts that use
      ↪ diacritics extensively. The
      ↪ verbalization-based prompt
      ↪ formatting is
      ↪ language-agnostic in
      ↪ principle. The main concern
      ↪ is the absence of speech or
      ↪ multimodal inputs, but this
      ↪ is a design choice rather
      ↪ than a form mismatch, given
      ↪ high literacy rates across
      ↪ the region.",
      "checklist_responses": "...",
      "evidence_quotes": [
        "Let f:P x X -> P_tilde denote
        ↪ a verbalization function
        ↪ that formats an input x in X
        ↪ into a prompt template.",
        "All datasets were standardized
        ↪ using a custom
        ↪ normalization pipeline...
        ↪ The pipeline handles diverse
        ↪ formats, including Excel
        ↪ spreadsheets, tabular or
        ↪ comma-separated value files,
        ↪ HuggingFace datasets, and
        ↪ plaintext.",
        "[The pipeline] merges multiple
        ↪ files when necessary,
        ↪ removes extraneous columns,
        ↪ standardizes column names
        ↪ (e.g., to 'text' and
        ↪ 'label'), adds language and
        ↪ language variety
        ↪ annotations, cleans textual
        ↪ artifacts (mojibake, mixed
        ↪ encodings)."
      ],
      "confidence": "high",
      "information_gaps": [
        "No documentation of whether
        ↪ prompt templates are
        ↪ language-specific or use a
        ↪ single language (e.g.,
        ↪ Spanish instructions for
        ↪ all inputs).",
```

```
        "No information about the
        ↪ handling of informal text
        ↪ conventions, emojis, or
        ↪ non-standard orthography in
        ↪ the input data.",
        "No discussion of text length
        ↪ distributions or whether
        ↪ they match real-world
        ↪ deployment text lengths."
      ],
      "requires_expert_verification": [
        "Whether the prompt
        ↪ verbalization templates are
        ↪ culturally and
        ↪ linguistically appropriate
        ↪ for each target language
        ↪ (e.g., whether Basque
        ↪ prompts use natural Basque
        ↪ phrasing rather than
        ↪ translated-from-Spanish
        ↪ patterns).",
        "Whether the text encoding and
        ↪ normalization pipeline
        ↪ correctly handles all
        ↪ regional text conventions
        ↪ (e.g., inverted punctuation
        ↪ in Spanish, Brazilian
        ↪ Portuguese spelling
        ↪ reforms)."
      ]
    },
    {
      "dimension": "Output Ontology",
      "score": 2,
      "justification": "This is a
      ↪ significant concern area.
      ↪ IberBench reuses label
      ↪ taxonomies from 101
      ↪ independent shared tasks
      ↪ without any documented
      ↪ validation of whether these
      ↪ taxonomies are appropriate
      ↪ across the full Iberian and
      ↪ Latin American cultural
      ↪ spectrum. The documentation
      ↪ explicitly lacks discussion
      ↪ of label taxonomy validation
      ↪ for regional relevance.
      ↪ Culturally sensitive
      ↪ categories such as emotion
      ↪ labels, humor categories,
      ↪ toxicity levels, and
      ↪ prejudice classifications are
      ↪ known to vary across cultures,
      ↪ yet the benchmark provides no
      ↪ evidence that these were
      ↪ designed with cross-regional
      ↪ Iberian norms in mind. The
      ↪ heavy reliance on Peninsular
      ↪ Spanish shared task labels
      ↪ raises questions about
      ↪ structural validity for Latin
      ↪ American deployment.",
```

```
      "checklist_responses": "...",
      "evidence_quotes": [
        "The label set is predefined,
        ↪ allowing the model to
        ↪ compute the likelihood of
        ↪ each candidate label. The
        ↪ label with the highest
        ↪ likelihood is selected.",
        "No documentation of how label
        ↪ taxonomies were validated
        ↪ for regional relevance. The
        ↪ paper does not discuss
        ↪ whether output categories
        ↪ (e.g., emotion labels,
        ↪ toxicity categories) were
        ↪ designed with Iberian
        ↪ cultural norms in mind or
        ↪ imported from non-Iberian
        ↪ frameworks."
      ],
      "confidence": "medium",
      "information_gaps": [
        "Complete label sets for all 101
        ↪ datasets are not documented
        ↪ in the paper.",
        "No information about the
        ↪ theoretical frameworks
        ↪ underlying the label
        ↪ taxonomies (e.g., which
        ↪ emotion theory was used for
        ↪ TASS 2020).",
        "No documentation of whether
        ↪ label sets were harmonized
        ↪ or reviewed when
        ↪ aggregating 101 independent
        ↪ datasets.",
        "No information about whether
        ↪ the shared task organizers
        ↪ considered cross-regional
        ↪ validity when designing
        ↪ their label sets."
      ],
      "requires_expert_verification": [
        "Whether the 7 emotion labels
        ↪ in TASS 2020 adequately
        ↪ represent emotional
        ↪ expression norms across all
        ↪ Iberian and Latin American
        ↪ cultures.",
        "Whether toxicity and hate
        ↪ speech categories from
        ↪ Peninsular Spanish shared
        ↪ tasks capture forms of
        ↪ harmful language prevalent
        ↪ in Latin American digital
        ↪ discourse.",
        "Whether sexism/prejudice
        ↪ detection label taxonomies
        ↪ reflect Latin American
        ↪ feminist and
        ↪ anti-discrimination
        ↪ frameworks or primarily
        ↪ European ones.",
```

```
        "Whether humor categorization
        ↪ from HAHA 2019 works for
        ↪ Latin American humor
        ↪ traditions (e.g., Mexican,
        ↪ Argentine, Colombian
        ↪ comedy).",
        "Whether stance detection
        ↪ topics and categories are
        ↪ relevant beyond the Spanish
        ↪ political context."
      ]
    },
    {
      "dimension": "Output Content",
      "score": 2,
      "justification": "This is the
      ↪ weakest dimension. IberBench
      ↪ reuses ground truth labels
      ↪ from 101 independent shared
      ↪ tasks with no re-annotation
      ↪ and no documented review of
      ↪ whether labels reflect
      ↪ perspectives of all Iberian
      ↪ and Latin American
      ↪ populations. No annotator
      ↪ demographics, inter-annotator
      ↪ agreement statistics, or
      ↪ annotation guidelines are
      ↪ documented at the benchmark
      ↪ level. For culturally
      ↪ subjective tasks (sentiment,
      ↪ humor, toxicity, prejudice),
      ↪ labels created by annotators
      ↪ from one Iberian
      ↪ subpopulation may
      ↪ systematically disagree with
      ↪ judgments of annotators from
      ↪ another. The lack of
      ↪ documentation makes it
      ↪ impossible to assess the
      ↪ degree of this concern from
      ↪ the benchmark materials
      ↪ alone.",
      "checklist_responses": "...",
      "evidence_quotes": [
        "We carefully collected and
        ↪ prepared a set of 101
        ↪ existing datasets to make
        ↪ them suitable for LLM
        ↪ evaluation.",
        "No annotator demographics
        ↪ documented at the benchmark
        ↪ level. The paper states no
        ↪ information about
        ↪ inter-annotator agreement,
        ↪ annotator recruitment
        ↪ criteria, or demographic
        ↪ composition across the
        ↪ constituent datasets.",
        "The organization committee
        ↪ includes 'experts in NLP,
        ↪ engineering, language
        ↪ ethics, and gender bias,
        ↪ drawn from both academia
        ↪ and industry', but their
        ↪ role is administrative
        ↪ (dataset/model acceptance)
        ↪ rather than annotation
        ↪ oversight.",
        "dataset creators are
        ↪ appropriately credited on
        ↪ the interface and dataset
        ↪ cards; we do not claim
        ↪ ownership of any dataset."
      ],
      "confidence": "medium",
      "information_gaps": [
        "Annotator demographics for all
        ↪ 101 constituent datasets.",
        "Inter-annotator agreement
        ↪ statistics for any of the
        ↪ constituent datasets.",
        "Annotation guidelines for the
        ↪ culturally subjective tasks
        ↪ (humor, toxicity, sentiment,
        ↪ prejudice).",
        "Whether any of the constituent
        ↪ datasets included
        ↪ annotators from Latin
        ↪ American countries.",
        "Label aggregation methods used
        ↪ in each constituent
        ↪ dataset.",
        "Whether individual annotator
        ↪ judgments are preserved or
        ↪ only aggregated labels."
      ],
      "requires_expert_verification": [
        "Whether humor labels from HAHA
        ↪ 2019 would be endorsed by
        ↪ Latin American
        ↪ annotators.",
        "Whether toxicity labels from
        ↪ DETOXIS 2021 reflect Latin
        ↪ American norms around
        ↪ harmful language.",
        "Whether sentiment labels from
        ↪ TASS workshops capture Latin
        ↪ American emotional
        ↪ expression norms.",
        "Whether prejudice/sexism
        ↪ labels from EXIST 2021
        ↪ align with Latin American
        ↪ anti-discrimination
        ↪ frameworks.",
        "The degree of annotator
        ↪ disagreement that would
        ↪ emerge if representative
        ↪ regional annotator pools
        ↪ re-labeled a sample of the
        ↪ culturally subjective
        ↪ datasets."
```

```
      ]
    },
    {
      "dimension": "Output Form",
      "score": 3,
      "justification": "IberBench's
      ↪ output form is partially
      ↪ aligned with regional
      ↪ deployment needs. The use of
      ↪ Macro-F1 for classification
      ↪ is appropriate and addresses
      ↪ label imbalance. The
      ↪ zero-shot evaluation setting
      ↪ is justified with a
      ↪ reasonable deployment
      ↪ argument. However, the
      ↪ near-exclusive focus on
      ↪ classification outputs (97 of
      ↪ 101 datasets) poorly
      ↪ represents modern LLM
      ↪ deployment, which is
      ↪ increasingly
      ↪ generation-centric. The
      ↪ acknowledged limitation of
      ↪ ROUGE-1 for summarization
      ↪ evaluation is relevant since
      ↪ lexical overlap metrics may
      ↪ penalize valid regional
      ↪ paraphrases. The text-only
      ↪ output modality is
      ↪ appropriate given high
      ↪ literacy rates, but the
      ↪ absence of speech output
      ↪ evaluation is a gap for
      ↪ voice-centric deployment
      ↪ scenarios in the region.",
      "checklist_responses": "...",
      "evidence_quotes": [
        "Ninety-seven of the datasets
        ↪ included in IberBench are
        ↪ framed as text
        ↪ classification tasks, 3 as
        ↪ text generation, and 1 as
        ↪ sequence labeling.",
        "Macro-F1 'to mitigate the
        ↪ impact of label imbalance
        ↪ in the test set.'",
        "ROUGE-1 'which measures
        ↪ unigram overlapping between
        ↪ the generated outputs and
        ↪ the reference texts.'",
        "Few-shot prompting is not yet
        ↪ a widespread practice in
        ↪ industry, partly since many
        ↪ non-technical users are
        ↪ unfamiliar with this
        ↪ prompting technique and
        ↪ because obtaining realistic
        ↪ examples can be
        ↪ challenging.",
        "the absence of artifact-free,
        ↪ perfectly human-aligned,
        ↪ and efficient automatic
        ↪ evaluation metrics for text
        ↪ summarization."
      ],
      "confidence": "high",
      "information_gaps": [
        "No documentation of whether
        ↪ evaluation metrics are
        ↪ sensitive to regional
        ↪ language variation (e.g.,
        ↪ whether Macro-F1 performance
        ↪ differs systematically
        ↪ across language
        ↪ varieties).",
        "No analysis of whether ROUGE-1
        ↪ penalizes valid lexical
        ↪ variation across
        ↪ Spanish/Portuguese
        ↪ varieties.",
        "No documentation of
        ↪ per-language or per-variety
        ↪ metric breakdowns in the
        ↪ evaluation framework."
      ],
      "requires_expert_verification": [
        "Whether ROUGE-1 systematically
        ↪ disadvantages models
        ↪ producing valid Latin
        ↪ American Spanish or
        ↪ Brazilian Portuguese text
        ↪ that differs lexically from
        ↪ Peninsular reference
        ↪ texts.",
        "Whether the zero-shot
        ↪ evaluation format produces
        ↪ valid assessments for all
        ↪ regional languages or
        ↪ disadvantages
        ↪ lower-resource languages
        ↪ (Basque, Galician) where
        ↪ models have less
        ↪ pretraining data.",
        "Whether classification-based
        ↪ evaluation adequately
        ↪ predicts real-world
        ↪ deployment quality for the
        ↪ most common LLM use cases
        ↪ in the target region."
      ]
    }
  ],
  "overall": {
```

```
    "overall_summary": "IberBench
    ↪ represents a meaningful step
    ↪ forward for Iberian-language
    ↪ NLP evaluation, with notable
    ↪ strengths in its ground-up
    ↪ dataset sourcing from regional
    ↪ shared tasks and its
    ↪ comprehensive task taxonomy.
    ↪ However, its validity for the
    ↪ full claimed scope (Iberian
    ↪ Peninsula and Ibero-America) is
    ↪ undermined by several
    ↪ significant concerns. The
    ↪ benchmark is functionally
    ↪ Spain-centric: ~60% of samples
    ↪ are Spanish, Portuguese is
    ↪ severely underrepresented
    ↪ (3.8%), Brazilian Portuguese and
    ↪ Argentinian Spanish are entirely
    ↪ absent, and the culturally
    ↪ sensitive tasks (humor,
    ↪ toxicity, sentiment, prejudice)
    ↪ rely on labels from annotators
    ↪ whose demographics are
    ↪ undocumented but likely skew
    ↪ toward Peninsular Spanish
    ↪ speakers. The strongest validity
    ↪ concern is in output content
    ↪ and output ontology, where the
    ↪ reuse of 101 independent
    ↪ datasets' labels without
    ↪ re-annotation or demographic
    ↪ documentation creates unknown
    ↪ but potentially substantial
    ↪ cultural bias risks for
    ↪ non-Peninsular deployment
    ↪ contexts. The benchmark's
    ↪ extensible design offers a clear
    ↪ path to remediation.",
    "risk_assessment": "medium",
    "remediation_suggestions": "Priority
    ↪ actions to improve validity: (1)
    ↪ Document annotator demographics
    ↪ and inter-annotator agreement
    ↪ for all constituent datasets, or
    ↪ at minimum for the culturally
    ↪ subjective tasks (humor,
    ↪ toxicity, sentiment, prejudice).
    ↪ (2) Conduct targeted
    ↪ re-annotation studies with
    ↪ Latin American annotators on a
    ↪ sample of culturally sensitive
    ↪ datasets to quantify potential
    ↪ label disagreement. (3) Expand
    ↪ coverage of Brazilian
    ↪ Portuguese and Argentinian
    ↪ Spanish, the two largest absent
    ↪ varieties. (4) Validate label
    ↪ taxonomies (emotion categories,
    ↪ toxicity categories, etc.) with
    ↪ regional experts from diverse
    ↪ Ibero-American countries. (5)
    ↪ Increase generation task
    ↪ coverage to better represent
    ↪ modern LLM deployment
    ↪ modalities. (6) Add
    ↪ per-language and per-variety
    ↪ metric breakdowns to identify
    ↪ systematic performance
    ↪ disparities. (7) Leverage the
    ↪ benchmark's extensibility
    ↪ framework to incorporate
    ↪ datasets from Latin American
    ↪ NLP communities.",
    "highest_concern_dimensions": [
      "Output Ontology",
      "Output Content"
    ],
    "strongest_dimensions": [
      "Input Form"
    ]
  }
}
```


Figure 23. IberBench LLM-judge Validity Assessment. We truncate the ``checklist_responses" elements for brevity.

E. Extended Related Work

The following subsections provide detailed descriptions of benchmark porting efforts summarized in Section 2.1.

E.1. Direct Translation and Adaptation

The most common approach to benchmark porting is to translate an existing English-language benchmark into target languages. Notable examples include XNLI (Conneau et al., 2018), which professionally translated the MultiNLI corpus into 15 languages; XCOPA (Ponti et al., 2020), which extended the Choice of Plausible Alternatives task to 11 languages including Quechua and Tamil; and PAWS-X (Yang

et al., 2019), which translated paraphrase detection data into 6 languages. More recently, AmericasNLI (Ebrahimi et al., 2022) extended XNLI to 10 indigenous American languages including Guarani, Nahuatl, and Aymara, finding that multilingual models performed near chance on most of these languages.

At larger scales, Belebele (Bandarkar et al., 2023) constructed a parallel reading comprehension benchmark across 122 language variants, and FLORES-200 (NLLB Team et al., 2022) provides a machine translation evaluation benchmark covering 200 languages. While these efforts expand linguistic coverage, a meta-analysis of over 2,000 multilingual benchmarks (Wu et al., 2025a) found that translation-based approaches systematically carry over cultural assumptions from the source language, and that model rankings can shift substantially when benchmarks are evaluated on culturally sensitive subsets (Singh et al., 2025).

E.2. Parallel Construction from Regional Sources

An alternative to translation is to construct benchmarks in parallel from regionally native sources, typically drawn from local educational examinations or professional licensing tests. This approach has been most prominently adopted for knowledge-intensive benchmarks inspired by the Massive Multitask Language Understanding (MMLU) framework (Hendrycks et al., 2021).

In East Asia, C-Eval (Huang et al., 2023) draws from Chinese national examinations across 52 subjects, and CMMLU (Li et al., 2023) similarly sources from Chinese exams across 67 subjects, explicitly including China-specific topics. KMMLU (Son et al., 2024) constructs a 45-subject evaluation from Korean professional certification exams, including Korea-specific categories such as Korean history, Korean law, and Korean traditional medicine. For Arabic, ArabicMMLU (Koto et al., 2024) was built from exam questions across Arab countries rather than translated from English. IndoMMLU (Koto et al., 2023) draws from the Indonesian national curriculum, including culturally specific subjects like local languages and regional traditions. These benchmarks share a recognition that translating MMLU is insufficient because its content embeds Western, particularly American, educational assumptions.

Beyond MMLU-style benchmarks, M3Exam (Zhang et al., 2023a) collects actual national exam questions from 9 countries across different languages and educational systems, and EXAMS (Hardalov et al., 2020) similarly aggregates high school examination questions from 16 countries, both avoiding translation entirely. INCLUDE (Romanou et al., 2024) takes this approach to 44 languages, using locally sourced professional exam questions, directly addressing content validity concerns inherent in translation. GAOKAO-Bench (Zhang et al., 2023b) evaluates models on China's National College Entrance Examination, representing an educational assessment fundamentally different in structure and content from English-language standardized tests.

E.3. Ground-Up Regional Development

For regions with limited pre-existing NLP resources, communities have invested in building benchmarks entirely from scratch. The Masakhane community (Adelani et al., 2021) has been at the forefront of this effort for African languages, developing MasakhaNER (Adelani et al., 2021) for named entity recognition across 10 (later 20) African languages, AfriSenti (Muhammad et al., 2023) for sentiment analysis across 14 African languages using natively collected Twitter data, and AfriQA (Ogundepo et al., 2023) for cross-lingual question answering in 10 African languages. IrokoBench (Adelani et al., 2024) aggregates these efforts into a comprehensive multi-task LLM evaluation benchmark spanning 16 African languages.

For Iberian and Latin American contexts, shared task campaigns organized through IberLEF (Ángel González et al., 2026) have produced a stream of regionally grounded benchmarks across diverse NLP tasks. In Southeast Asia, SEA-HELM (Susanto et al., 2025) was developed as a holistic regional adaptation that explicitly addresses the validity gaps identified when applying HELM to the region.

E.4. Domain-Specific Porting Challenges

Specialized domains present particularly acute challenges for benchmark porting, as professional knowledge is inherently jurisdiction-specific. In medicine, MedQA (Jin et al., 2020) exemplifies parallel construction by sourcing questions from medical licensing examinations in the United States, China, and Taiwan independently. AfriMedQA (Olatunji et al., 2023) builds a Pan-African medical question-answering benchmark from African licensing exams, capturing region-specific conditions such as tropical diseases and local healthcare practices that are absent from Western medical benchmarks.

In the legal domain, LegalBench (Guha et al., 2023) focuses on US common law reasoning, and the fundamental differences between common law and civil law traditions make direct porting infeasible. LEXTREME (Niklaus et al., 2023) addresses this by constructing a multilingual legal benchmark from EU legal corpora across 24 languages, leveraging the inherent multilingual nature of EU legislation. These examples underscore that for specialized domains, construct validity violations are particularly severe when porting across jurisdictional boundaries.

E.5. Cultural Knowledge and Values Benchmarks

A recent wave of benchmarks explicitly evaluates cultural competence rather than linguistic ability alone. CulturalBench (Choenni et al., 2024) tests LLMs on 1,227 cultural knowledge questions across 45 global regions, finding significant disparities in model accuracy between Western and non-Western cultures. BLEnD (Myung et al., 2024) evaluates everyday cultural knowledge—such as local customs, food, and social norms—across 13 languages, revealing that LLMs perform substantially better on Western cultural knowledge.

For values and safety, CValues (Xu et al., 2023) evaluates Chinese LLMs against Chinese societal values, demonstrating that Western safety taxonomies fail to capture culturally specific norms. TaarofBench (Sadr et al., 2025) exposes a similar gap for Persian ritual politeness norms, showing that responses rated "polite" by standard metrics can violate *taarof* norms. KoBBQ (Jin et al., 2024) adapts the Bias Benchmark for QA to Korean contexts, introducing entirely new bias categories rooted in Korean social structures such as regional discrimination, educational background prejudice, and military service stereotypes. These works demonstrate that safety, values, and bias benchmarks are among the most culturally sensitive to port, as they encode deeply embedded societal norms.

F. Additional Figures

F.1. Validity-Centered Benchmark Design Workflow

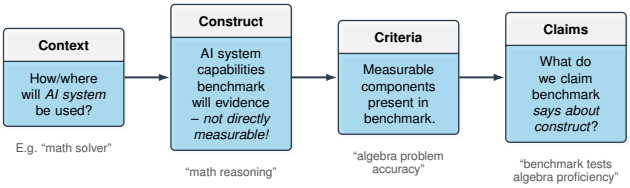


Figure 24. **Validity-centered Benchmark Design Workflow.**

F.2. Validity Analysis Framework

G. Additional Background

The following subsections provide additional background on mechanisms for determining evaluation needs and the detailed validity taxonomy referenced in Section 2.

G.1. Determining Evaluation Needs

In order to establish which criteria and claims will be appropriate to define for a benchmark, actors must understand how the intended use of AI systems interact with the construct they wish to measure. The two main axes through which this discovery process can unfold are *stakeholder driven* and *data driven* approaches.

G.1.1. Stakeholder driven Discovery

This type of discovery process consists of participatory benchmark design, where regional stakeholders are integrated into the design process of the benchmark (Birhane et al., 2022). We define **regional stakeholders** as members of a population pertinent to the intended use and context of the AI systems the benchmark is designed for. This population may include domain experts who can inform the theory of the benchmark, community members from population's end-users, members of industry who would develop these AI systems, and policy makers who may be involved in these systems' regulation.

G.1.2. Data-driven Discovery

When available, domain data may be used to determine appropriate domain components for a benchmark. For example, if AI systems already exist and have been deployed for the intended use case, then **post-deployment monitoring** (Schröder & Schulz, 2022) data may be analyzed to understand that granular types real-world interactions which end-users have with these systems. However, AI companies are generally very protective of this type of data (Nicholas, 2024), and it may not be readily accessible to the benchmark designer. In cases where there is no real-world post-deployment data available, designers may instead turn to real-world use case data of other types of systems or products applied to the same domain. For example, if the intended use of the AI systems is to help flag inappropriate messages within online chatroom moderation, then real-world chatroom data may be analyzed to determine appropriate benchmark components, even if there was no AI system involved in the loop originally for the production of this data.

G.2. Validity Taxonomy

We leverage the validity taxonomy of (Salaudeen et al., 2025), which describes different components along which validity may be established:

1. **Content Validity**: Does the benchmark sufficiently cover the space of input types relevant to the claim about the construct?
2. **Criterion Validity**: Does the benchmark correlate with established standards?
   (a) **Predictive Validity:** Can your evaluation predict downstream outcomes?
   (b) **Concurrent Validity:** To what extent does your evaluation agree with another validated assessment under the exact same conditions?

Benchmark	Region/Scope	Domain	Year	Strategy
Translation and Adaptation				
XNLI (Conneau et al., 2018)	15 languages	NLI	2018	Translation
XCOPA (Ponti et al., 2020)	11 languages	Commonsense	2020	Translation
AmericasNLI (Ebrahimi et al., 2022)	10 indigenous langs	NLI	2022	Translation
Belebele (Bandarkar et al., 2023)	122 languages	Reading comp.	2023	Parallel
Parallel Construction from Regional Sources				
C-Eval (Huang et al., 2023)	China	Knowledge	2023	Regional exams
CMMLU (Li et al., 2023)	China	Knowledge	2024	Regional exams
KMMLU (Son et al., 2024)	Korea	Knowledge	2024	Regional exams
ArabicMMLU (Koto et al., 2024)	Arab countries	Knowledge	2024	Regional exams
IndoMMLU (Koto et al., 2023)	Indonesia	Knowledge	2023	Regional exams
INCLUDE (Romanou et al., 2024)	44 languages	Knowledge	2024	Regional exams
M3Exam (Zhang et al., 2023a)	9 countries	Education	2023	Regional exams
GAOKAO-Bench (Zhang et al., 2023b)	China	Education	2023	Regional exams
Ground-Up Regional Development				
MasakhaNER (Adelani et al., 2021)	Africa (20 langs)	NER	2021	From scratch
AfriSenti (Muhammad et al., 2023)	Africa (14 langs)	Sentiment	2023	From scratch
AfriQA (Ogundepo et al., 2023)	Africa (10 langs)	QA	2023	From scratch
IrokoBench (Adelani et al., 2024)	Africa (16 langs)	Multi-task	2024	Mixed
IberBench (Ángel González et al., 2026)	Iberian/LatAm	Multi-task	2026	Aggregated
SEA-HELM (Susanto et al., 2025)	Southeast Asia	Holistic eval.	2025	Mixed
Domain-Specific				
MedQA (Jin et al., 2020)	US/China/Taiwan	Medical	2021	Regional exams
AfriMedQA (Olatunji et al., 2023)	Pan-African	Medical	2024	From scratch
LEXTREME (Niklaus et al., 2023)	EU (24 langs)	Legal	2023	From scratch
LegalBench (Guha et al., 2023)	US	Legal	2023	From scratch
Cultural Knowledge and Values				
CulturalBench (Choenni et al., 2024)	45 regions	Cultural know.	2024	From scratch
BLEnD (Myung et al., 2024)	13 languages	Everyday know.	2024	From scratch
CValues (Xu et al., 2023)	China	Values/Safety	2023	From scratch
KoBBQ (Jin et al., 2024)	Korea	Bias	2024	Adapted
TaarofBench (Sadr et al., 2025)	Iran/Persian	Cultural norms	2025	From scratch

Table 5. **Overview of benchmark porting strategies across regions and domains.** Benchmarks are organized by the primary strategy used: translation of existing English benchmarks, parallel construction from regional examination sources, ground-up development, domain-specific construction, and cultural evaluation.

3. **Construct Validity**: Do the benchmark criteria *truly* measure the construct of interest?
   (a) **Structural Validity:** Does your evaluation capture the structure of the construct you are measuring?
   (b) **Convergent Validity:** Does your evaluation correlate with other measures that assess the same construct?
   (c) **Discriminant Validity:** Can your evaluation differentiate between constructs that should be distinct?
4. **External Validity**: Does the benchmark generalize to different environments and settings?
5. **Consequential Validity**: Do real-world decisions informed by the benchmark align appropriately with its intended purpose?

H. Public Web Interface to the Pipeline

This appendix details the public web interface introduced in Section 4.1. It walks through the end-to-end user experience, the two run modes, the curated gallery, the logging and privacy model, and a brief implementation note.

H.1. End-to-End User Flow

A run begins on the landing page, where the user provides three inputs that mirror the pipeline's user-supplied inputs in Figure 2 of the main text: the benchmark paper as a PDF, a free-text deployment description, and, optionally, the dataset ID if the benchmark is published on HuggingFace. The same page hosts a sidebar gallery of benchmarks already assessed by our expert collaborators; clicking a gallery entry opens its completed validity report without running the pipeline.

On submission, the tool executes Steps 0–2 to derive a slug from the deployment description (Step 0), extract

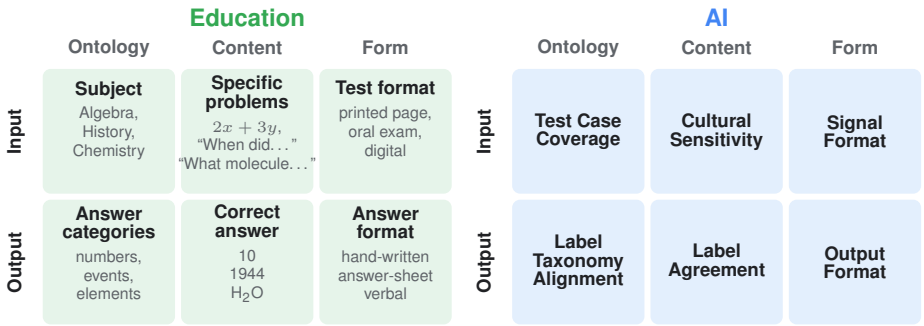


Figure 25. **Validity Analysis Framework.** The six components of our framework, organized along two axes: data modality (Input vs. Output) and data dimension (Ontology, Content, Form). Each cell is illustrated with parallel examples from standardized educational testing (**left**) and AI benchmarking (**right**), highlighting how the same structural components appear in both domains.

lightweight metadata from the paper's title pages (Step 1), generate three to five deployment-specific elicitation questions (Step 2a), and, after the user answers them in the browser, synthesize the responses into a structured deployment-context summary (Step 2b). The elicitation step realizes the human-in-the-loop dependency of the pipeline: each question targets one of the six validity dimensions and asks the user to narrow the deployment context where the model alone cannot.

The remaining stages run without further input. Step 3a extracts verbatim quotes from each page of the uploaded PDF and consolidates them; Step 3b selects reference examples and synthesizes the per-paper benchmark YAML; Steps 4a–4b select a regional template and build the region scaffold; Step 5 enriches the scaffold through agentic web search; Step 5b profiles the HuggingFace dataset when one was provided; Step 6 deterministically composes the final prompt; and Step 7 produces the Opus-scored validity assessment across the six Input/Output $\times$ Ontology/Content/Form dimensions. Progress streams live throughout, with the active sub-step labeled and counters shown for any per-page or per-example loops.

The final report renders the Step 7 output as a per-dimension score table with expandable reasoning and supporting evidence, alongside a tab for the raw model output. Each run is identified by a slug and a UUID, and the report is downloadable as Markdown, JSON, or PDF. The report is shown on-screen and stays available for the duration of the browser session; there is no email delivery.

H.2. Run Modes

The interface offers two modes that share the same underlying pipeline.

Automatic mode. After elicitation, the remaining steps run end-to-end without further interaction. The user sees a checklist of modules with completed phases marked and the current phase highlighted. This mode is intended for users who want a final report and do not need to inspect intermediate artifacts.

Step-by-step mode. The pipeline pauses between modules, exposing the artifact produced at each boundary: the consolidated quote extraction, the synthesized benchmark YAML, the regional context, and the optional dataset analysis. The user can read, copy, or download any artifact before advancing. This mode supports two use cases: benchmark authors who want to verify that the extraction captures the intended evidence, and researchers using the tool as a transparent reference implementation of the pipeline.

H.3. Curated Gallery

The sidebar of the landing page lists benchmarks that our expert collaborators have already assessed using the pipeline. Each entry opens the completed validity report for that benchmark-context pair, including the per-dimension scores, the reasoning, and the supporting evidence quotes. The gallery serves as a public reference of worked analyses for users who want concrete examples of what the pipeline produces before running it themselves, and as a growing record of benchmark-context pairs that have been examined in detail.

H.4. Logging and Privacy

The interface uses a four-tier logging model. Operational metadata (run identifiers, timing, step status, model usage) is always recorded so that runs can be debugged and aggregate usage can be reported. The user's PDF, deployment description, elicitation answers, and the intermediate and final artifacts are stored only with explicit per-run opt-in, and any stored content is automatically deleted after 90 days. The Anthropic API key supplied by the user is held only in the browser session and in server memory for the duration of the run; it is never written to disk or to logs. From the report page, users can export the full record of their run or

request immediate deletion.

H.5. Implementation

The frontend is a React/TypeScript single-page application. The backend is a FastAPI (Python) service that wraps the same pipeline modules used in our experiments and streams per-step progress to the browser over Server-Sent Events. Frontend and backend are packaged as a single container for deployment. The codebase will be released under an open-source license alongside the camera-ready version of this paper.