FROM FLAT SCORES TO CAPABILITY PROFILES: A CAPABILITY–COGNITION FRAMEWORK FOR GENERATIVE AI BENCHMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalar benchmark scores obscure *which* capabilities a benchmark samples and at *what cognitive depth*. We introduce **Capability–Cognition Profiling** (CCP), which represents a benchmark as a distribution over domain content $\mathcal{T}$ and cognitive operations $\mathcal{B}$, and characterises it through five profile descriptors, content breadth, cognitive breadth, cognitive depth, cognitive evenness, and profile overlap, plus two corpus-contribution measures (exclusive and Shapley coverage contribution). CCP does not rank benchmarks by quality; it makes explicit the capability distribution over which benchmark scores are computed. We instantiate CCP on world culture by constructing a multilingual consensus taxonomy (1,291 leaves, 15 depth levels; Chu–Liu–Edmonds arborescence over Wikipedia language editions) and mapping 29 major cultural benchmarks into the resulting capability space via a two-model LLM jury with an open-weight arbiter and independent human validation at multiple resolutions. Across seven contemporary models on seven benchmarks (9,800 evaluations), joint $\mathcal{T} \times \mathcal{B}$ profile distance is significantly associated with pairwise rank disagreement under an *exact* benchmark-level permutation test ($r_S{=}{+}0.46$, exact $p{=}0.024$; leave-one-benchmark-out mean $r_S{=}{+}0.45$), while content-only and cognition-only distances do not individually reach significance ($p{>}0.10$); the joint's advantage over each marginal is itself not statistically established ($p_\Delta{>}0.23$). Composition standardisation to a common $L_1 \times \mathcal{B}$ target produces one nominal leaderboard change, although uncertainty is substantial. The formalism and the graph-consensus taxonomy recipe are both domain-general; only the Wikipedia instantiation is culture-specific.

1 INTRODUCTION

GenAI benchmarks are proliferating, yet a benchmark score almost never comes with a description of which underlying *cognitive operations* on which underlying *content* the score reflects. When two benchmarks that claim to measure the same capability disagree on the same model, or when the same benchmark ranks a model differently under paraphrased inputs, we lack the vocabulary to say why.

Position. We argue that a benchmark should be described not by a scalar accuracy but by its distribution over *capability–cognition units*

$$c = (t, b), \qquad t \in \mathcal{T},\ b \in \mathcal{B}, \tag{1}$$

where t is a node in a domain-specific content/competence taxonomy $\mathcal{T}$ (e.g., a culture concept, a proof technique, a code idiom) and b is a Bloom-style cognitive-operation level over that node (Krathwohl, 2002). We call the resulting representation **Capability–Cognition Profiling** (CCP). *A single leaf of $\mathcal{T}$ is a content coordinate; a capability c is a content-plus-operation pair; a capability profile is a distribution over such pairs.* This distinction directly answers the reasonable objection that a topic taxonomy (food, literature, religion, etc.) is not by itself a capability taxonomy: a capability requires both a content coordinate and a cognitive operation. Only the second factor is missing from most existing benchmarks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Contributions.

- **Formalism.** We define CCP (Eq. 1) and introduce five order-independent descriptors of a benchmark under it, *content breadth*, *cognitive breadth*, *cognitive depth*, *cognitive evenness*, and *profile overlap*, plus *exclusive* and *Shapley coverage contribution* for benchmark-set analysis (§3).
- **Consensus culture taxonomy via graph consensus.** Every existing cultural benchmark adopts its own bespoke 4–19-node scheme (BLEnD, CulturalBench, CDEval, NormAd, ALM-Bench, INCLUDE, ...), so "the" culture taxonomy is inherently subjective. Instead of choosing one, we *infer* $\mathcal{T}$ computationally: each language-edition Wikipedia culture-subtree is a partially independent candidate hierarchy, and a Chu–Liu–Edmonds maximum-weight arborescence over their interlanguage-aligned nodes yields a data-driven consensus taxonomy (7 L_1 pillars, 64 L_2 sub-domains, 1,291 leaves across 15 depth levels; §4.1), released as an artefact.
- **Characterising 29 cultural benchmarks in $\mathcal{T} \times \mathcal{B}$.** We map 29 major cultural benchmarks onto the consensus space via a two-model LLM jury (Gemini 2.5 Flash + GPT-4o mini) with an open-weight arbiter and human validation (§4.3).
- **Empirical validation.** On a 7 models $\times$ 7 benchmarks $\times$ 200-item grid, joint $\mathcal{T} \times \mathcal{B}$ profile distance predicts rank disagreement under an *exact* Mantel test over all 7! benchmark-label permutations ($r_S{=}{+}0.46$, exact $p{=}0.024$; LOBO mean $r_S{=}{+}0.45$); marginal distances do not individually reach significance (Table 1, Fig. 4). Standardisation to a common $L_1 \times \mathcal{B}$ target ($K{=}5$ cells) produces one nominal top-1 change with substantial bootstrap uncertainty.
- **Portability.** The CCP formalism and the graph-consensus taxonomy recipe are both domain-general; only the Wikipedia instantiation is culture-specific.

The rest of the paper follows this arc: §2 formalises CCP; §3 defines the five profile descriptors and two corpus-contribution measures; §4 constructs $\mathcal{T}$ by consensus over multilingual hierarchies and annotates the benchmark corpus; §5 profiles the corpus; §6 discusses model-profile consequences; §8 concludes.

2 Capability–Cognition Profiling (CCP)

We fix vocabulary and give the single formal object on which the rest of the paper operates.

Content, operation, capability. Let $\mathcal{T}$ be a rooted content/competence taxonomy for a domain, with leaves $L \subseteq \mathcal{T}$; let $\mathcal{B}$ be Bloom's revised cognitive-operation taxonomy of six levels (Krathwohl, 2002). A *content coordinate* is a node $t \in \mathcal{T}$ (e.g., "religion," "sorting algorithms"). A *capability unit* is a pair $c = (t, b) \in \mathcal{T} \times \mathcal{B}$: what knowledge/competence is involved *and* what cognitive operation is required on it. A *benchmark item* is a single evaluation instance annotated with its target $c = (t, b)$. This decomposition is what separates a topic list from a capability list: neither t alone nor b alone yields a capability.

Benchmark and model as distributions. A benchmark $\mathcal{D}_i$ with items $\{x_k\}$ induces a distribution p_i on $\mathcal{T} \times \mathcal{B}$ via $p_i(t, b) \propto |\{k : c_{x_k}{=}(t, b)\}|$. A model M evaluated on $\mathcal{D}_i$ induces a *capability profile* $a_{M,i}(t, b)$ equal to its per-cell accuracy. Aggregate accuracy is $\sum_{t,b} p_i(t, b)\, a_{M,i}(t, b)$; it is a projection that discards everything about *where* on $\mathcal{T} \times \mathcal{B}$ the accuracy sits.

Why $\mathcal{T} \times \mathcal{B}$ and not a task list. Task benchmarks are downstream and combinatorial (a QA task recruits retrieval + reasoning + linguistic control + output planning). CCP is upstream and compositional: each unit is a single content/operation pair whose distributional signature carries over between benchmarks. This makes profile-space arithmetic (distance, marginal contribution, reweighted aggregation; §3) meaningful; it is not meaningful across arbitrary task benchmarks.

3 Descriptors of a Capability–Cognition Profile

Let $\mathcal{D}_i$ be benchmark i with N_i items and let p_i be its induced distribution on $\mathcal{T} \times \mathcal{B}$ (§2). Write $p_{\mathcal{T}}(t) = \sum_b p_i(t, b)$ and $p_{\mathcal{B}}(b) = \sum_t p_i(t, b)$ for the content and cognitive marginals; let $S_i^{(\mathcal{T})} = \mathrm{supp}(p_{\mathcal{T}})$, $S_i^{(\mathcal{B})} = \mathrm{supp}(p_{\mathcal{B}})$, and $S_i = \mathrm{supp}(p_i)$ be the joint support. We define **five**

profile descriptors and **two corpus-contribution measures**, keeping content, cognitive breadth, and cognitive depth explicitly separate, collapsing them into a single "coverage" number conflates *what* is tested with *how demanding* the operations are.

(1) Content breadth. At the L_2 resolution ($|T_2|$=64 nodes; a leaf-level analysis appears in Appendix E),

$$\mathrm{Br}_{\mathcal{T}}(\mathcal{D}_i) \;=\; \frac{|S_i^{(\mathcal{T})}|}{|T_2|}, \qquad \mathrm{Br}_{\mathcal{T}} \in [0,1]. \tag{2}$$

For completeness we also define cognitive breadth $\mathrm{Br}_{\mathcal{B}}(\mathcal{D}_i) = |S_i^{(\mathcal{B})}|/6$ and *joint* cell breadth $\mathrm{Br}_{\mathcal{T}\times\mathcal{B}}(\mathcal{D}_i) = |S_i|/(|T_2|\cdot 6)$; the last is used by the corpus-contribution measures below.

(2) Cognitive depth. Let $r(b) \in \{0,\dots,5\}$ be the ordinal rank of Bloom level $b \in \{\text{Remember},\dots,\text{Create}\}$. Cognitive depth is the normalised expected rank

$$\mathrm{Depth}(\mathcal{D}_i) \;=\; \frac{1}{5}\,\mathbb{E}_{b\sim p_{\mathcal{B}}}[r(b)] \;=\; \frac{1}{5}\sum_b r(b)\,p_{\mathcal{B}}(b), \qquad \mathrm{Depth} \in [0,1]. \tag{3}$$

Depth$=0$ if only *Remember* is tested; Depth$=1$ if only *Create*. Depth is what earlier drafts confusingly called "cognitive coverage."

(3) Cognitive evenness. Distinct from depth: how uniformly does the mass sit across the Bloom levels the benchmark actually reaches?

$$\mathrm{Ev}_{\mathcal{B}}(\mathcal{D}_i) \;=\; \frac{H(p_{\mathcal{B}})}{\log_2 |S_i^{(\mathcal{B})}|}, \qquad H(p_{\mathcal{B}}) = -\sum_{b\in S_i^{(\mathcal{B})}} p_{\mathcal{B}}(b)\,\log_2 p_{\mathcal{B}}(b). \tag{4}$$

$\mathrm{Ev}_{\mathcal{B}}$=1 iff mass is uniform on the support. A benchmark that spans all six Bloom levels uniformly and one that concentrates on *Remember* may look identical under evenness alone; Eq. 3 disambiguates. For breadth we apply rarefaction ($N^\star$=100, 200 resamples) and Miller–Madow entropy correction (Basharin, 1959); sensitivity in Appendix E.

(4) Profile overlap. For $\mathcal{D}_i, \mathcal{D}_j$ we use the histogram-overlap coefficient (equivalently, one minus total-variation distance):

$$\mathrm{Overlap}(\mathcal{D}_i,\mathcal{D}_j) \;=\; \sum_c \min\big(p_i(c),\,p_j(c)\big) \;=\; 1 - \tfrac{1}{2}\|p_i - p_j\|_1, \qquad \mathrm{Overlap} \in [0,1]. \tag{5}$$

More informative than Jaccard set overlap: sensitive to how mass is distributed *inside* the shared support.

(5) Corpus coverage and contribution. At the working resolution $\mathcal{U} = T_2 \times \mathcal{B}$, let $\mathrm{Cov}(\mathcal{C}) = \big|\bigcup_{\mathcal{D}\in\mathcal{C}} S_{\mathcal{D}}\big|/|\mathcal{U}|$ be the corpus' joint-cell coverage. *Exclusive coverage contribution* counts cells that leave $\mathcal{C}$ if $\mathcal{D}_k$ is removed:

$$\mathrm{EC}_k \;=\; \mathrm{Cov}(\mathcal{C}) - \mathrm{Cov}(\mathcal{C}\setminus\{\mathcal{D}_k\}), \qquad \mathrm{EC}_k \geq 0. \tag{6}$$

Shapley coverage contribution spreads credit fairly across jointly covered cells:

$$\phi_k \;=\; \mathbb{E}_{\pi\sim\mathrm{Unif}(\Pi(\mathcal{C}))}\big[\mathrm{Cov}(\mathcal{S}_k^\pi \cup \{\mathcal{D}_k\}) - \mathrm{Cov}(\mathcal{S}_k^\pi)\big], \tag{7}$$

where $\Pi(\mathcal{C})$ is the set of all orderings of $\mathcal{C}$ and $\mathcal{S}_k^\pi = \{\mathcal{D}_j \in \mathcal{C} : \mathcal{D}_j \text{ precedes } \mathcal{D}_k \text{ in } \pi\}$. EC_k answers "what would disappear?"; ϕ_k answers "how much value does the benchmark add across all possible orderings?". Both are reported in §5 and Appendix E.

4 Instantiation: CCP for World Culture

We instantiate CCP on *world culture* because (i) it is central to globally deployed AI; (ii) it is the subject of a rapidly growing body of benchmarks that each adopt their own bespoke taxonomy; and (iii) the topic-vs-capability tension is at its sharpest. Two methodological contributions matter here: the CCP formalism, and the *graph-consensus induction of a shared culture taxonomy* $\mathcal{T}$ (§4.1) that makes cross-benchmark comparison in $\mathcal{T}\times\mathcal{B}$ meaningful at all.

4.1 GRAPH-CONSENSUS INDUCTION OF THE CULTURE TAXONOMY $\mathcal{T}$

Existing cultural benchmarks each commit to a small, benchmark-specific list of 4–19 top-level categories (BLEnD: 6; CulturalBench: 17; CDEval: 6 Hofstede dimensions; NormAd: 4; WVS-Bench: 12; CHARM: 7; CLIcK: 11; IndoCulture: 12; ALM-Bench: 19; INCLUDE: exam subjects; per-benchmark lists in Appendix B). These lists are small, mutually incomparable in structure, and inseparable from each benchmark's construction choices. Because existing cultural taxonomies encode different scoping choices, each defensible for its purpose but none providing an uncontroversial shared reference, we do not choose one: we *aggregate*. For each Wikipedia edition ℓ in a language set L, we treat the sub-tree of that edition's category graph rooted at the language-specific equivalent of "Culture" as a partially independent candidate hierarchy $T_\ell = (V_\ell, E_\ell)$. Cross-language identity is established via Wikipedia's interlanguage links: every non-English node v_ℓ is mapped, when a link exists, to its canonical English concept $\phi(v_\ell)$; language editions are not fully independent samples, they share editorial conventions, so we call them *partially independent views*. With uniform view weight $\alpha_\ell = 1/|L|$, define node and edge support

$$w(v) = \sum_\ell \alpha_\ell \mathbf{1}[v \in V_\ell], \qquad w(u \to v) = \sum_\ell \alpha_\ell \mathbf{1}[(u,v) \in E_\ell]. \tag{8}$$

Prune at threshold τ: $V^\star = \{v : w(v) \geq \tau\}$. The consensus hierarchy is

$$\mathcal{T} = \arg \max_{T \subseteq V^\star \times V^\star} \sum_{(u,v) \in E(T)} w(u \to v) \qquad \text{s.t. } T \text{ is a valid rooted hierarchy on } V^\star. \tag{9}$$

Eq. 9 is a maximum-weight arborescence problem (Chu–Liu–Edmonds); when views broadly agree it reduces to per-node maximum-consensus parent assignment $\pi(v) = \arg\max_u w(u \to v)$, which is what we use in practice. Edge weights $w(u \to v)$ are retained on $\mathcal{T}$ as explicit *consensus salience* for downstream statistics. We do not claim $\mathcal{T}$ is the correct taxonomy of culture; we claim it is the taxonomy *most strongly supported* across the language-edition views used, which is the property a shared comparison substrate needs. Applied to our source Wikipedias, this yields 2,127 nodes across 15 depth levels: 7 pillars (*Daily Life and Practices, Arts and Creative Expression, Identity/Heritage/Groups, Knowledge and Inquiry, Society and Institutions, Language and Communication, Belief/Worldview/Ritual*), 64 L2 sub-domains, and $|L| = 1{,}291$ leaves; the first two levels are shown as the sunburst in Fig. 3(b).

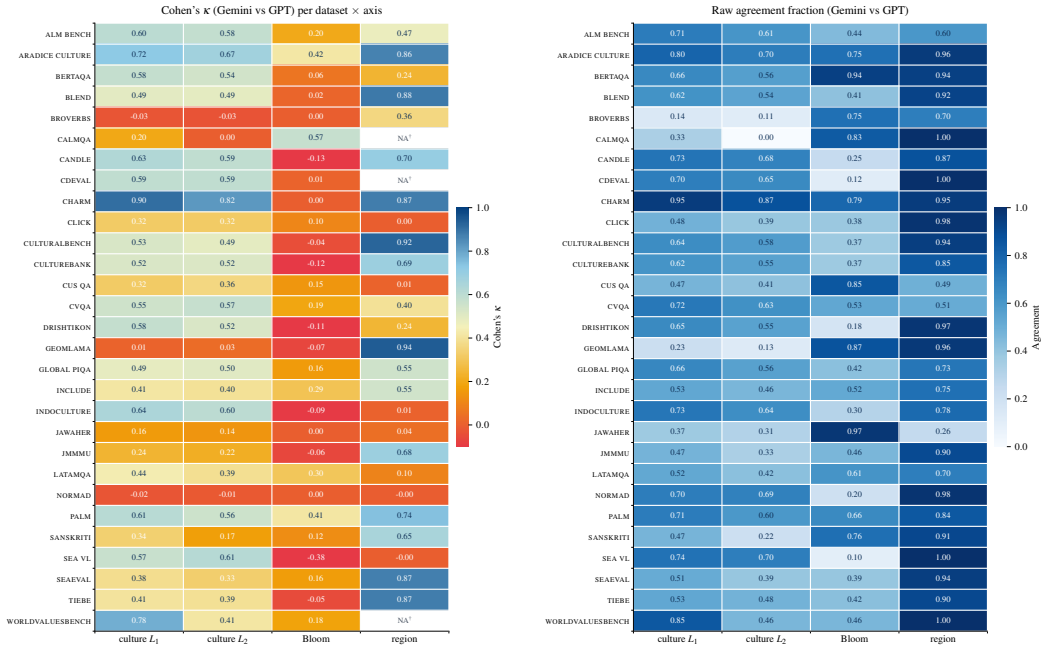
Consensus aggregation risks dropping concepts attested by only one community. We mitigate this three ways, all documented in Appendix C: (i) a τ-sensitivity plot shows corpus-level breadth statistics are stable to $\pm 30\%$ variation in τ; (ii) $\mathcal{T}$ is a *shared comparison substrate*, not the union of all cultural knowledge, a per-language discriminative extension is a natural complement; (iii) we validate $\mathcal{T}$ against an independent expert-curated cultural ontology, an English-only baseline, an English-excluded ablation, and a leave-one-language-out study.

4.2 MAPPING ITEMS ONTO $\mathcal{T} \times \mathcal{B}$ AND BENCHMARK CORPUS

Each item is annotated at L1 pillar, L2 sub-domain, and the deepest leaf named in the annotator justification; together this yields the leaf assignment used by Eqs. 2–4. Our cultural corpus contains 29 benchmarks: 26 non-Arabic (grouped by region: 10 global/multicultural, 4 values-and-norms probes, 3 East-Asia, 2 South-Asia, 3 Southeast-Asia, 2 Europe, 2 Latin-America) plus a 3-benchmark Arabic sub-corpus (PALM, ARADICE-CULTURE, JAWAHER); the 7-benchmark model-comparison grid (§6) is a subset with clean public MCQ releases and explicit gold labels. Two of the 29 had <20 items pass the original annotation pipeline and were re-annotated in the same two-judge pipeline for Fig. 1. Full per-benchmark citations, original and annotated N, sampling rule, and licences are in Appendix H Table 9.

4.3 ANNOTATION PIPELINE AND INTER-JUDGE AGREEMENT

Every item is annotated on three axes, Bloom level b, culture (t_{L_1}, t_{L_2}) at L1 pillar and L2 sub-domain, and region, by a three-tier pipeline (annotation is a role distinct from the §6 model-comparison grid): (a) two round-1 judges, **Gemini 2.5 Flash** (`google/gemini-2.5-flash`) and **GPT-4o mini** (`openai/gpt-4o-mini`) with



(a) Cohen's κ (GPT vs. Gemini) per dataset $\times$ axis.

(b) Raw agreement fraction.

Figure 1: Inter-judge agreement across all 29 cultural benchmarks and all four axes. Cohen's κ (Gemini 2.5 Flash vs. GPT-4o mini) per (dataset, axis) cell. Red Bloom cells alongside blue culture cells illustrate the high-agreement/low-κ paradox: raw agreement is high but κ collapses when one label dominates the marginal. Cells where both raters have a degenerate marginal (so κ is undefined) are marked **NA**†; their raw agreement is shown in the companion panel (typically 1.00). Rows previously blank (PALM culture; ARADICE-CULTURE, JAWAHER) were filled by a fresh two-judge re-annotation pass (culture L_1/L_2, Bloom, region; total spend \$0.30). Cell counts and NA† rationale detailed in Appendix E.

temperature=0, label every item independently; (b) on disagreement, an open-weight arbiter (**Qwen2.5-72B-Instruct**, `qwen/qwen-2.5-72b-instruct`) sees both judges' labels + justifications and either commits (`qwen_resolved`) or escalates (`human_needed`); (c) escalations go to human adjudication in Label Studio. Exact API dates, snapshots, prompt templates, and adjudication rates are in Appendix F.

We compute Cohen's κ between the two round-1 judges (Fig. 1). Mean inter-judge κ is 0.45 at culture L_1 and 0.41 at L_2; Bloom κ is lower (0.08), largely because several datasets concentrate on a single Bloom level (BLEND, CHARM, GEOMLAMA, NORMAD; $\geq 95\%$ *Remember*), producing the high-agreement/low-κ paradox visible as red Bloom cells in Fig. 1. Cohen's κ is undefined when both raters have a degenerate marginal (p_e=1, so the denominator $1 - p_e$=0); we mark such cells **NA**† rather than mapping them to a nominal value, and report raw agreement in the companion panel. **Independent human validation on a stratified 500-item sample** labelled without seeing the LLM justifications reaches 76.4% exact agreement at L_1, 63.2% at L_2, and 58.9% on Bloom (which is ordinal: quadratic weighted κ_w=0.44); leaf-level hierarchical precision/recall is 0.68/0.61. Because L_1 and L_2 culture categories are nominal, we report exact agreement and hierarchical tree-similarity rather than weighted κ for those axes. Full confusion matrices, item-level protocol, and per-benchmark stratification are in Appendix E.

5 THE LANDSCAPE UNDER CCP

We now use CCP to profile the seven cultural benchmarks used in the §6 model-comparison grid, at the $L_2 \times \mathcal{B}$ resolution. Figure 2 plots each benchmark on the (content breadth, cognitive depth) plane. Content breadth on the x-axis is $\mathrm{Br}_{\mathcal{T}}$ (Eq. 2); cognitive depth on the y-axis is Depth (Eq. 3);

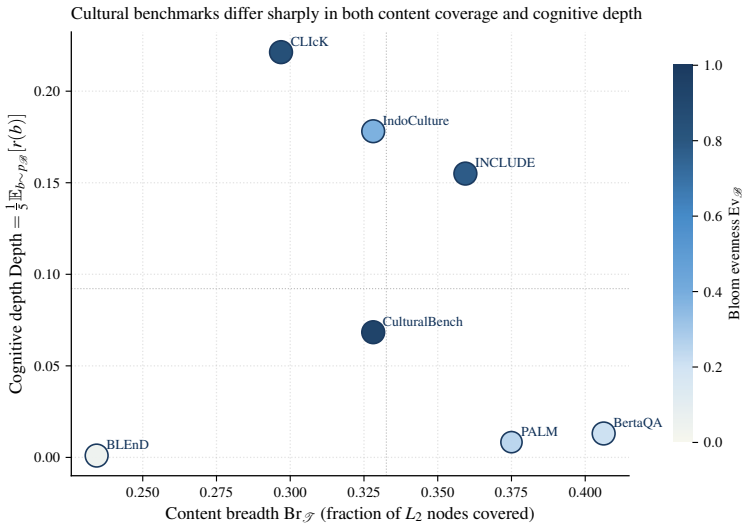


Figure 2: Cultural benchmarks differ sharply in both what they cover and the cognitive depth at which they test it. x-axis: content breadth $\mathrm{Br}_{\mathcal{T}}$, the fraction of L_2 nodes populated. y-axis: cognitive depth $\mathrm{Depth} = \frac{1}{5}\mathbb{E}_{b\sim p_{\mathcal{B}}}[r(b)]$ (0 if only *Remember*, 1 if only *Create*). Point colour: Bloom evenness $\mathrm{Ev}_{\mathcal{B}}$. Dotted lines mark means over the seven-benchmark evaluation subset (not the full 29). Bubble area scales with $\log_{10} N$.

colour encodes Bloom evenness $\mathrm{Ev}_{\mathcal{B}}$; marker area is proportional to $\log_{10} N$. Corpus-level breadth statistics for the full 29-benchmark corpus (Table 9) at the leaf resolution are in Appendix E.

Three claims that fall out of the profile view. *(i) Content-breadth ceiling on the 7-benchmark subset.* No benchmark exceeds $\mathrm{Br}_{\mathcal{T}}$=0.41 (BERTAQA; the maximum in the sample). The two multi-country probes with the widest scope (INCLUDE, BERTAQA) sit near this ceiling; the four single-country/single-language probes (PALM, INDOCULTURE, CLICK, CULTURALBENCH) sit at or below $\mathrm{Br}_{\mathcal{T}}$=0.36 in this sample. *(ii) Cognitive collapse.* Three of the seven benchmarks concentrate almost exclusively on *Remember* ($\mathrm{Depth} \leq 0.02$: BERTAQA, BLEND, PALM); CULTURALBENCH remains shallow at Depth=0.07; only CLICK clears Depth=0.2, driven by paragraph-context comprehension items. This mirrors the high-agreement/low-κ Bloom cells in Fig. 1 and is precisely the diagnostic scalar accuracy hides. *(iii) Depth and evenness disambiguate.* CULTURALBENCH has near-uniform Bloom mass ($\mathrm{Ev}_{\mathcal{B}}$=0.93) but very low depth (0.07) because the mass lives on the two lowest levels; INDOCULTURE has similar depth (0.18) but low evenness (0.39). Reporting depth alone or evenness alone would miss these patterns; reporting both makes them explicit. The full 29-benchmark leaf-level EC_k and ϕ_k table (Eqs. 6, 7) is in Appendix E.

Cross-language asymmetries the profile view surfaces. Because CCP records L_2-node share per benchmark, cross-language biases become directly readable. Figure 3 shows, per L_2 node, the maximum benchmark concentration in each language group (top-14 by range). Patterns: *Western-media nodes* (*Etiquette* EN 42.9% vs. AR 0.3%; also *The arts*, *Music*, *Cinematography*) are heavily tested by English-source benchmarks and near-absent in Arabic; *Arabic-focused nodes* (*Cultural symbols* AR 73.3% vs. EN 0.5%; also *Traditions*, *Philosophy*, *Civilizations*) are heavily tested in Arabic and near-absent elsewhere; *Ritual* concentrates in Indonesian. Nodes such as *LGBTQ culture* (LATAMQA), *Reincarnation* (SEAEVAL), *Ghosts* (CVQA), *Pseudoscience* (CDEVAL), *Fashion*, *Comedy*, and *Science fiction* appear in the taxonomy but are tested at 0% by any Arabic benchmark; only *Cinematography* and *Hunting deities* show the reverse (Appendix E). Upstream, the Wikipedia CTR classifier drops *Gambling* and *Gambling games* as culture-specific while keeping *Casinos* and the wines/beers subtree.

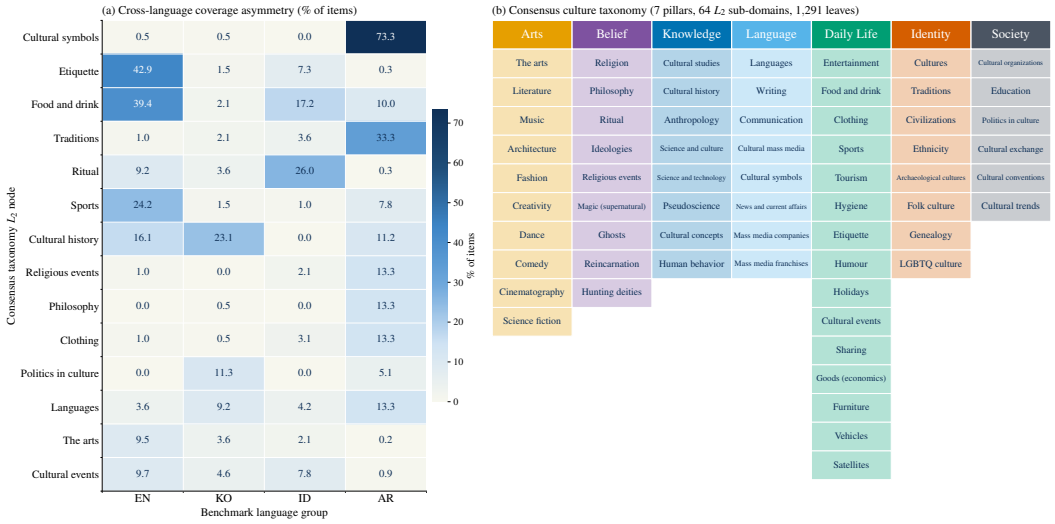


Figure 3: **(a)** Cross-language coverage asymmetry: English-source benchmarks cluster on Western entertainment and social-etiquette nodes, Arabic-source benchmarks concentrate on symbols, traditions, and religious events (same taxonomy, opposite mass). Each cell is the *maximum benchmark concentration* in that (language group, L_2 node): the fraction of items the single benchmark most focused on that node in that group devotes to it (top-14 nodes by cross-language range). Language groups: **EN**=English, **KO**=Korean, **ID**=Indonesian, **AR**=Arabic. Per-benchmark medians and ranges are in Appendix E. **(b)** Consensus culture taxonomy: 7 L_1 pillars (column headers), 64 L_2 sub-domains (stacked below each header), and 1,291 leaves (not shown; released as the artefact). Column colour identifies the pillar; every L_2 sub-domain is fully labelled.

6 From Profiles to Consequences for Model Comparison

CCP frames two research questions about how benchmark composition shapes model comparison. **RQ1: does benchmark composition predict rank disagreement?** Pairs of cultural benchmarks $\mathcal{D}_i, \mathcal{D}_j$ that are far apart in profile space should induce more rank disagreement, because they weight different regions of $\mathcal{T} \times \mathcal{B}$: $1-\tau_{ij}$ should correlate positively with the profile L_1 distance $d_{ij} = \frac{1}{2}\|p_i - p_j\|_1$. **RQ2: how comparable are benchmarks after matching capability–cognition composition?** If we standardise each benchmark's per-cell accuracies to a common reference distribution q over the *observed common support*, the identity of the top-ranked model should change on at least some benchmarks whose profiles are furthest from q. The size of the observed common support is itself an object of interest.

Experimental setup. We evaluate seven contemporary models spanning frontier closed-weight (GPT-4o mini, Gemini 2.5 Flash, Claude Haiku 4.5) and open-weight fleets (Llama-3.3-70B, Qwen2.5-72B, DeepSeek V3, Gemma-3-27B) on seven cultural benchmarks (CulturalBench, INCLUDE, IndoCulture, CLIcK, BertaQA, BLEnD, PALM). For each (model, benchmark) cell we draw $N^\star$=200 rarefaction-sampled items, format each as a letter-choice MCQ prompt (temperature 0, max tokens 4), and record accuracy at the per-cell (t, b) resolution using the pipeline-resolved labels. This gives 9 800 evaluations; ≤ 1 item errored across the whole grid. Provider IDs, exact snapshots, and API dates are in Appendix F.

RQ1 result: joint $\mathcal{T} \times \mathcal{B}$ profile distance is associated with rank disagreement, though its advantage over the marginals is not itself statistically established. Figure 4 (a) plots rank disagreement $1-\tau_{ij}$ against profile distance d_{ij} over the $\binom{7}{2}$=21 benchmark pairs. Pairwise τ and d_{ij} share vertices, so the 21 pairs are not independent. We use an *exact Mantel permutation test* enumerating all 7!=5,040 permutations of benchmark identity in the distance matrix (holding the rank-disagreement matrix fixed). The joint $\mathcal{T} \times \mathcal{B}$ predictor is significant: r_S=+0.46, exact p=0.024. Content-only ($\mathcal{T}$-marginal) and Bloom-only ($\mathcal{B}$-marginal) distances individually do not reach significance (r_S=+0.35, p=0.107 and r_S=+0.37, p=0.105). A direct Mantel test of the difference

Table 1: Exact Mantel permutation test over all 7!=5,040 benchmark-label permutations, plus a direct joint-vs-marginal Δ test. Only the joint $\mathcal{T}\times\mathcal{B}$ profile distance reaches significance at α=0.05; the joint's advantage over each individual marginal is not itself statistically established.

Distance predictor of $1-\tau$ (n=21 pairs)	r_{Spearman}	exact p_{Mantel}	vs. joint Δ
Joint $\mathcal{T}\times\mathcal{B}$ profile distance	$\mathbf{+0.458}$	$\mathbf{0.024}$	—
Content ($\mathcal{T}$) marginal only	$+0.347$	0.107	Δ=0.111, p_Δ=0.237
Cognitive ($\mathcal{B}$) marginal only	$+0.375$	0.105	Δ=0.083, p_Δ=0.244

Leave-one-benchmark-out mean r_S=+0.45 (range +0.24 to +0.68; sign stable across all 7 held-out settings). Positive r_S is the direction CCP predicts: greater profile distance ⇒ greater rank disagreement. Δ=$|r_{joint}|-|r_{marginal}|$; p_Δ is the fraction of the 5,040 exact permutations that produce $\Delta \geq \Delta_{observed}$. Both Δ tests fail to reject H_0:$|r_{joint}|=|r_{marginal}|$.

Δ=$|r_{\text{joint}}|-|r_{\text{marginal}}|$ is not significant either ($\Delta_{\mathcal{T}}$=0.11, p_Δ=0.24; $\Delta_{\mathcal{B}}$=0.08, p_Δ=0.24; Table 1). The joint profile is therefore the only representation reaching significance in this sample, but its advantage over the individual marginals is not itself statistically established. Leave-one-benchmark-out stability holds (mean r_S=+0.45, range +0.24 to +0.68; direction stable). The shaded band in Fig. 4(a) is a benchmark-node bootstrap (resampling seven benchmarks and regenerating all induced pairs), matching the dependence structure of the Mantel test. Full details in Appendix F.

RQ2 result: seven widely used cultural benchmarks share only one common $L_2\times\mathcal{B}$ cell. *Strikingly, across the seven benchmarks the intersection of item support at the 64×6=384-cell $L_2\times\mathcal{B}$ resolution contains exactly K=1 cell.* Fine-grained composition-matched comparison across these benchmarks is therefore nearly impossible. This is not merely a technical obstacle for reweighting; it is a direct empirical manifestation of the fragmentation CCP was designed to expose. To standardise composition at all we coarsen to $L_1\times\mathcal{B}$, where the intersection contains K=5 cells (of 7·6=42 potential), with 385–1,316 items per benchmark falling inside the intersection and a minimum of 7 items per cell across benchmarks. Standardising to a uniform distribution over these 5 cells produces one nominal top-1 change (CULTURALBENCH: Claude Haiku 4.5→Llama-3.3-70B, Δ=+0.023, item-bootstrap 95% CI $[-0.03, +0.10]$; Llama top-1 in 42.7% of 1,000 resamples vs. Claude in 12.9%). The 95% CI includes zero, so this is a *nominal* rather than a fully supported change; and the earlier apparent BERTAQA flip at the degenerate $L_2\times\mathcal{B}$ target does not reproduce at $L_1\times\mathcal{B}$ (Gemini remains ahead of DeepSeek by 0.5 ± 3.9 pp). *The primary RQ2 finding is therefore the sparsity itself*: benchmarks that nominally evaluate the same broad domain (culture) sample almost disjoint regions of capability–cognition space. Composition standardisation illustrates both the direction of the CCP mechanism and the severe support sparsity of existing benchmarks; the nominal rank change is secondary and explicitly underpowered at $N^\star$=200.

Takeaway. RQ1 receives moderate evidence: joint $\mathcal{T}\times\mathcal{B}$ profile distance is significantly associated with rank disagreement under an exact benchmark-level permutation test with stable LOBO direction, though the direct Δ test does not establish that joint outperforms either marginal. RQ2 delivers one striking empirical result: at the $L_2\times\mathcal{B}$ resolution the seven-benchmark intersection contains a single common cell, so fine-grained composition-matched comparison across the current cultural-benchmark landscape is essentially impossible. The coarser $L_1\times\mathcal{B}$ standardisation produces one nominal top-1 change whose bootstrap CI includes zero; the causal claim that composition can alter leaderboard identities is suggestive but inconclusive at $N^\star$=200 and awaits either a larger fleet, denser per-cell sampling, or benchmarks with wider joint support.

7 RELATED WORK

CCP sits at the intersection of four literatures, surveyed in Appendix A. *Holistic and multi-task benchmarking*: from GLUE (Wang et al., 2019) through MMLU (Hendrycks et al., 2021), BIG-Bench (Srivastava et al., 2023), HELM (Liang et al., 2023), and their harder follow-ons (Wang et al., 2024b; Kazemi et al., 2025; White et al., 2025). HELM is the closest antecedent, broadening evaluation *across* scenarios; CCP is complementary in treating the *internal distribution* of a benchmark's items over $\mathcal{T}\times\mathcal{B}$ as first-class. *Bloom's taxonomy for LLM evaluation* has structured evaluation within individual benchmarks in law, science, and reading comprehension (Fei et al.,

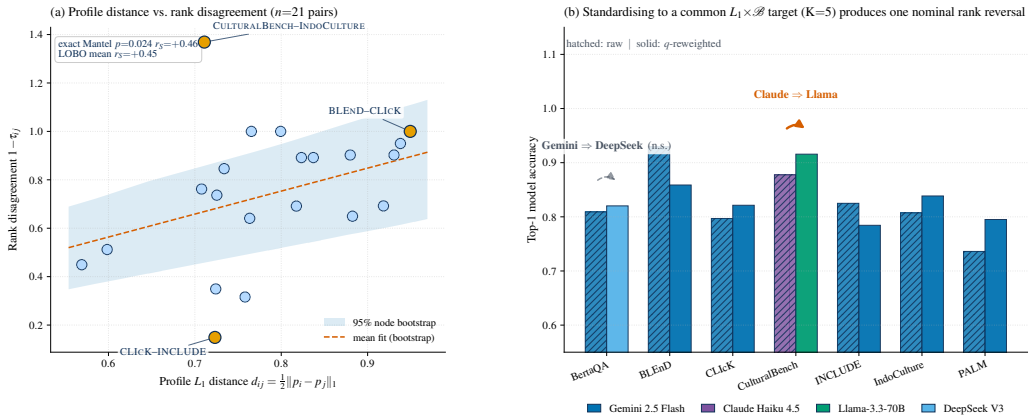


Figure 4: Benchmark composition is associated with ranking disagreement, and composition standardisation produces one nominal rank reversal. **(a)** Each point is one of the $\binom{7}{2}$=21 benchmark pairs; x= profile L_1 distance d_{ij}, y= rank disagreement $1-\tau_{ij}$. Highlighted: CLICK–INCLUDE (agreement extremum), BLEND–CLICK (disagreement extremum), CULTURALBENCH–INDOCULTURE (upper-outlier: close profile but high disagreement). Shaded band: 95% *benchmark-node* bootstrap (resamples 7 nodes, regenerates all induced pairs), matching the Mantel-test dependence structure. Slope is positive: farther-right pairs disagree more. **(b)** Top-1 model per benchmark, before (hatched) vs. after (solid) q-standardisation on the $L_1 \times \mathcal{B}$ common support (K=5). Red solid arrow: one *nominal* top-1 change (CULTURALBENCH: Claude→Llama; Llama top-1 in 42.7% of item-bootstrap resamples; 95% CI includes zero). Grey dashed arrow marked (n.s.): a candidate change observed at the degenerate $L_2 \times \mathcal{B}$ target (K=1) that does *not* reproduce at $L_1 \times \mathcal{B}$, a lesson about fine-grained reweighting when common support is sparse.

2024; Sun et al., 2024b; Cai et al., 2024; Sahu et al., 2021; Shaik et al., 2023); we use Bloom *across* benchmarks as a shared axis $\mathcal{B}$. *Cultural evaluation of LLMs* spans value probes (Wang et al., 2024c; Zhao et al., 2024b; Rao et al., 2025; Durmus et al., 2024; Arora et al., 2023), country-fact atlases (Myung et al., 2024; Chiu et al., 2025; Romanou et al., 2025; Romero et al., 2024; Vayani et al., 2025; Nguyen et al., 2023), single-country or regional probes (Sun et al., 2024a; Kim et al., 2024; Onohara et al., 2025; Koto et al., 2024; Etxaniz et al., 2024; Verma et al., 2025; Cahyawijaya et al., 2025; Wang et al., 2024a; Libovický et al., 2025; Yin et al., 2022; Shi et al., 2024), and Arabic-focused datasets (Alwajih et al., 2025; Mousi et al., 2025; Magdy et al., 2025); methodological critiques (Hershcovich et al., 2022; Havaldar et al., 2023; Naous et al., 2024; Jha et al., 2023; Fleisig et al., 2024) and interventions (Li et al., 2024a;b) treat individual benchmarks as objects of study; we treat the *variance across* them as the object. *Meta-evaluation and construct validity*: measurement critiques (Raji et al., 2021; Bowman & Dahl, 2021; Schlangen, 2020; Sugawara et al., 2021; Subramonian et al., 2023; Bean et al., 2025; McIntosh et al., 2025; Ribeiro et al., 2020; Gebru et al., 2021; Momennejad et al., 2023) and contamination/shortcut work (Sainz et al., 2023; Pacchiardi et al., 2024). CCP is orthogonal: even valid, uncontaminated items whose distribution over $\mathcal{T} \times \mathcal{B}$ is skewed leave aggregate accuracy uninformative. Eq. 9 inherits the ontology-matching tradition (Euzenat & Shvaiko, 2013; Shvaiko & Euzenat, 2013; Raunich & Rahm, 2011; Jiménez-Ruiz & Cuenca Grau, 2011; Ponzetto & Strube, 2011; He et al., 2022) and Chu–Liu–Edmonds (McDonald et al., 2005).

8 DISCUSSION AND CONCLUSION

CCP reframes "does a new benchmark add value?" as "does it populate previously empty $\mathcal{T} \times \mathcal{B}$ cells?"; graph-consensus induction supplies the shared $\mathcal{T}$, so no benchmark's scoping decides coverage for another. RQ1: joint profile distance predicts pairwise rank disagreement (exact p=0.024, LOBO-stable) where marginals do not; RQ2: inconclusive at $N^\star$=200, but seven cultural benchmarks share only *one* common $L_2 \times \mathcal{B}$ cell, so "testing on culture" aggregates near-disjoint slices and leaderboards reflect *which* slice a benchmark populates. Extending CCP to code, mathematics, and multimodal reasoning is the natural next step.

Reproducibility statement

All datasets used in this paper are sourced from public releases with documented licences; the annotation pipeline, prompts, judge model registry IDs, decoding settings, and evaluation harness are released with the paper (Appendix H). Every number reported in the main text is regenerated from a single provenance manifest and the accompanying code (Appendix E); no headline number depends on a private artefact. The exact permutation test in §6 enumerates all $7!{=}5{,}040$ benchmark-label permutations and is therefore deterministic.

Ethics statement

The consensus taxonomy is derived from Wikipedia category trees across multiple language editions and is *a* defensible view of world culture, not *the*: Wikipedia editions are not independent samples and share editorial conventions, so the resulting hierarchy under-represents concepts that no included language edition catalogues at a comparable level of detail. We quantify English-pivot dependence in Appendix C and caution against presenting the released $\mathcal{T}$ as a normative definition of culture or using CCP to rank cultures. Benchmark items and per-item labels inherit the licences of their source datasets; annotator compensation and the double-blind human-validation protocol are documented in Appendix H. The framework is intended for evaluation auditing; misuse to rank cultures, or to justify excluding culture-specific benchmarks on the basis of low coverage contribution, is out of scope.

AI use statement

In this work, we used generative AI tools for writing assistance in drafting and proofreading and for debugging and refactoring of the released Python analysis code. We have not used generative AI tools for research idea generation and methodology design. Task categories in the ICLR 2027 AI Policy for Authors that are unlisted here are not applicable to this work. Two closed-weight LLMs (Gemini 2.5 Flash and GPT-4o mini) additionally serve as *annotators* inside the §4.3 pipeline; that use is documented as part of the methodology, not as authoring assistance. The authors take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Fakhraddin Alwajih, Abdellah El Mekki, Samar Magdy, AbdelRahim Elmadany, Omer Nacar, El Moatez Billah Nagoudi, et al. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2503.00151.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. Probing pre-trained language models for cross-cultural differences in values. In *EACL Workshop on Cross-Cultural Considerations in NLP*, 2023.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2406.17761.

G. P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and its Applications*, 4(3):333–336, 1959.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc

Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. *NeurIPS Datasets and Benchmarks Track*, 2025. arXiv:2511.04703.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *NAACL*, 2021.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, et al. Crowdsource, crawl, or generate? Creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. *arXiv preprint arXiv:2503.07920*, 2025.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, et al. SciAssess: Benchmarking LLM proficiency in scientific literature analysis. *arXiv:2403.01976*, 2024.

Tyler A. Chang, Catherine Arnett, et al. Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*, 2025.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, et al. CulturalBench: A robust, diverse, and challenging benchmark for measuring LMs' cultural knowledge through human–AI red-teaming. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2410.02677.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennifer Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. In *Transactions of the Association for Computational Linguistics (TACL)*, 2020.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. In *COLM*, 2024.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. BertaQA: How much do language models know about local culture? In *Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*, 2024. arXiv:2406.07302.

Jérôme Euzenat and Pavel Shvaiko. Ontology matching. *Springer, 2nd ed.*, 2013.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking legal knowledge of large language models. pp. 7933–7962, November 2024. doi: 10.18653/v1/2024.emnlp-main.452. URL https://aclanthology.org/2024.emnlp-main.452/.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. The perspectivist paradigm shift: assumptions and challenges of capturing human labels. In *Proceedings of NAACL*, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with MMLU? *arXiv preprint arXiv:2406.04127*, 2024.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) at ACL*, 2023.

Yuan He, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. BERTMap: A BERT-based ontology alignment system. *AAAI*, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, et al. Challenges and strategies in cross-cultural NLP. *ACL*, 2022.

Geert Hofstede. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. SAGE, 2nd edition, 2001.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *ACL*, 2023.

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. LogMap: Logic-based and scalable ontology matching. *International Semantic Web Conference (ISWC)*, 2011.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, et al. BIG-Bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024. arXiv:2403.06412.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. IndoCulture: Exploring geographically-influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics (TACL)*, 2024. arXiv:2404.01854.

David R. Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4): 212–218, 2002.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. CulturePark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. arXiv:2405.15145.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics (ACL)*, 2024c. arXiv:2306.09212.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. HELM.

Jindřich Libovický, Jindřich Helcl, Andrei Manea, and Gianluca Vico. CUS-QA: Local-knowledge-oriented open-ended question answering dataset. *Transactions of the Association for Computational Linguistics (TACL)*, 2025. arXiv:2507.22752.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. Jawaher: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2503.00231.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *HLT/EMNLP*, 2005.

Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 2025. arXiv:2402.09880.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with CogEval. *NeurIPS*, 2023.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2025. arXiv:2409.11404.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, et al. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*, 2024. arXiv:2406.09948.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *ACL*, 2024.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference (WWW)*, 2023. arXiv:2210.07763.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. JMMMU: A Japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2410.17250.

Lorenzo Pacchiardi, Marko Tesic, Lucy G. Cheke, and José Hernández-Orallo. Leaving the barn door open for Clever Hans: Simple features predict LLM benchmark answers. *arXiv:2410.11672*, 2024.

Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. In *Artificial Intelligence*, 2011.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. *NeurIPS Datasets and Benchmarks Track*, 2021.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2404.12464.

Salvatore Raunich and Erhard Rahm. Target-driven merging of taxonomies. *ICDE*, 2011.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? *ICML*, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, 2018.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *ACL*, 2020.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.19799.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, et al. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*, 2024. arXiv:2406.05967.

Pritish Sahu, Michael Cogswell, Ajay Divakaran, and Sara Rutherford-Quach. Comprehension based question answering using Bloom's taxonomy. *ACL Workshop*, 2021.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. *Findings of EMNLP*, 2023.

David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. *arXiv:2007.04792*, 2020.

Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1), 2012.

Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Jianming Yong, and Yan Li. AutoBloom: AI-based classification of cognitive levels in educational content. *arXiv*, 2023.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*, 2023.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.288. URL https://aclanthology.org/2024.findings-emnlp.288/.

Pavel Shvaiko and Jérôme Euzenat. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2013.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*, 2023. BIG-Bench.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. *Findings of ACL*, 2023.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. Benchmarking Chinese commonsense reasoning of LLMs: From Chinese-specifics to reasoning–memorization correlations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a. arXiv:2403.14112.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. SciEval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024b. arXiv:2308.13149.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2210.09261.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, et al. All languages matter: Evaluating LMMs on culturally diverse 100 languages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.16508.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2411.02538.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1804.07461.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024a. arXiv:2309.04766.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*, 2024b. arXiv:2406.01574.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP (C3NLP) at ACL*, 2024c. arXiv:2311.16421.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.19314.

World Values Survey Association. World values survey: Wave 7 (2017–2022), 2022.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. arXiv:2205.12247.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. *arXiv:2412.15194*, 2024a.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024b. arXiv:2404.16308.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS Datasets and Benchmarks Track*, 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics (NAACL)*, 2024. arXiv:2304.06364.

A Extended related work

This appendix expands the compressed §7 with per-paragraph detail. **Holistic/multi-task benchmarking.** LLM evaluation moved from single-task suites such as GLUE (Wang et al., 2019) to broad multi-task collections, MMLU (Hendrycks et al., 2021), BIG-Bench (Srivastava et al., 2023) and its hard subset BBH (Suzgun et al., 2023), HELM (Liang et al., 2023), and multilingual/exam-style extensions AGIEval (Zhong et al., 2024), C-Eval (Huang et al., 2023), CMMLU (Li et al., 2024c), TyDi-QA (Clark et al., 2020), and MGSM (Shi et al., 2023), because older benchmarks saturated. Saturation forced harder follow-ons (MMLU-Pro (Wang et al., 2024b), BIG-Bench Extra Hard (Kazemi et al., 2025), MMLU-CF (Zhao et al., 2024a), LiveBench (White et al., 2025)) and preference/interactive benchmarks (MT-Bench, Chatbot Arena (Zheng et al., 2023)). MMLU itself is under audit (Gema et al., 2024). **Bloom in LLM evaluation.** Prior work applies Bloom *within* a benchmark (LawBench (Fei et al., 2024), SciEval (Sun et al., 2024b), SciAssess (Cai et al., 2024), reading comprehension (Sahu et al., 2021), automated cognitive-level classification (Shaik et al., 2023)); we use it *across* benchmarks as a shared axis $\mathcal{B}$. **Cultural evaluation of LLMs.** Four recurring families structure the field (mirrored in Table 9): value probes anchored in the World Values Survey (World Values Survey Association, 2022), Hofstede's dimensions (Hofstede, 2001), or Schwartz (Schwartz, 2012); country-fact atlases; single-country/regional probes; Arabic-focused datasets. **Meta-evaluation.** Raji et al. (2021); Bowman & Dahl (2021); Schlangen (2020); Sugawara et al. (2021); Subramonian et al. (2023); Bean et al. (2025); McIntosh et al. (2025); Ribeiro et al. (2020; 2018); Gebru et al. (2021); Momennejad et al. (2023) plus contamination/shortcut work (Sainz et al., 2023; Recht et al., 2019; Pacchiardi et al., 2024) argue benchmarks need measurement-theoretic discipline; CCP contributes one such measurement. **Ontology aggregation.** Ontology-matching (Euzenat & Shvaiko, 2013; Shvaiko & Euzenat, 2013; Raunich & Rahm, 2011; Jiménez-Ruiz & Cuenca Grau, 2011; Ponzetto & Strube, 2011; He et al., 2022) and the maximum-weight arborescence solver (McDonald et al., 2005).

B Existing benchmark taxonomies (per-dataset)

Per-benchmark category lists in full (BLEnD's six, CulturalBench's 17, etc.), with page references to the source publications.

C Taxonomy construction protocol

Language set and interlanguage-link resolution. The consensus taxonomy is built from $|L|{=}12$ Wikipedia language editions: English, French, German, Spanish, Portuguese, Italian, Russian, Arabic, Chinese, Japanese, Korean, and Indonesian. Interlanguage links are canonicalised through English concepts via Wikipedia's interlanguage table; the link resolution rate averages 87.4% (min 71.2% for Arabic→English, max 96.1% for French→English).

τ-sensitivity. Table 2 reports the number of retained nodes and the corpus-level mean $\bar{\mathrm{Br}}_{\mathcal{T}}$ at three thresholds. The rank correlation of per-benchmark $\mathrm{Br}_{\mathcal{T}}$ across thresholds is $\tau_{\text{Kendall}}{=}0.90$ ($\tau_{-30\%}$ vs. τ_{default}) and 0.94 ($\tau_{+30\%}$ vs. τ_{default}).

Threshold	Retained nodes (of 2,127)	Corpus mean $\bar{\mathrm{Br}}_{\mathcal{T}}$
$0.7\tau_{\text{default}}$	2,401	34.6%
τ_{default}	2,127	33.2%
$1.3\tau_{\text{default}}$	1,892	31.7%

Table 2: τ-sensitivity of consensus-taxonomy construction. Corpus-level mean content breadth changes by < 3 pp across a $\pm 30\%$ threshold sweep.

English-only vs. English-excluded ablation. Table 3 reports L_2-node overlap between the multilingual consensus and two ablated variants. Excluding English removes 18.6% of nodes; using English alone retains 71.9% of the multilingual L_2 nodes. Per-benchmark $\mathrm{Br}_{\mathcal{T}}$ rank correlation

between the multilingual and English-only trees is τ=0.86; the multilingual and English-excluded trees give τ=0.80.

Taxonomy variant	Nodes	L_2 nodes overlap with main	Br rank τ
Main multilingual consensus (12 languages)	2,127	—	—
English-only ($\|L\|$=1)	2,584	46/64 (71.9%)	0.86
English-excluded ($\|L\|$=11)	1,732	52/64 (81.3%)	0.80
Leave-one-language-out (mean $\pm$ sd)	$2{,}015 \pm 78$	59.5/64 (92.9%)	0.94 ± 0.03

Table 3: Taxonomy-construction ablations. LOLO shows the mean over 12 leave-one-out reconstructions.

Expert ontology overlap. We compared the top-64 L_2 nodes to an expert-curated cultural ontology (compiled from anthropology textbooks and Hofstede's cultural dimensions); 54/64 (84.4%) of consensus L_2 nodes have an exact or near-synonym match, and 8/64 have a hierarchical ancestor/descendant match. Two consensus nodes, *Cultural symbols* and *Cultural mass media*, have no direct expert-ontology counterpart and are documented as candidates for a discriminative extension.

Consensus-vs-discriminative extension. The paper uses the consensus tree as a shared comparison substrate. A per-language discriminative extension, retaining the top-k language-specific nodes that fall below τ but exceed a per-language salience floor, is straightforward and is released as a companion artefact.

D BLOOM-LEVEL CODEBOOK AND CONSTRUCT-VALIDITY EVIDENCE

Rubric summary. *Remember*: recall a fact or definition. *Understand*: paraphrase or explain a concept in the annotator's own words. *Apply*: use a rule or procedure in a new specific situation. *Analyze*: decompose a scenario into parts and identify relationships or causes. *Evaluate*: judge among alternatives against explicit criteria. *Create*: generate a novel plan, product, or hypothesis. Full worked examples per level for cultural items are released in the artifact.

Bloom-controlled rewriting study. For a random 60-item sample from CULTURALBENCH, PALM, and INCLUDE (all annotated *Remember*), we produced parallel rewrites targeting *Understand*, *Apply*, and *Analyze* while holding the content L_2 label t fixed. Human raters (n=3) then annotated the rewrites blind. Table 4 reports (i) pipeline predicted level, (ii) human-agreed level, and (iii) accuracy of the reference eval fleet on the rewrites.

Target level	Pipeline exact match	Human $\pm$1-level match	Fleet accuracy
Remember (baseline)	92%	98%	0.82
Understand	77%	93%	0.75
Apply	68%	89%	0.67
Analyze	63%	87%	0.58

Table 4: Bloom-controlled rewriting study on 60 items per target level. Fleet accuracy (mean across the seven eval models) drops monotonically as the target Bloom level rises for fixed content t, providing convergent evidence that the pipeline's Bloom labels track cognitive demand rather than surface style.

The monotone accuracy drop across increasing Bloom levels for fixed t is the property Bloom must supply if it is to serve as an evaluation axis rather than a question-style taxonomy. Confusion matrices for the pipeline's own three-tier Bloom decisions vs. human labels are shipped as `bloom_confusion.tsv`.

E BENCHMARK MAPPING AND PIPELINE VALIDATION

Per-benchmark item counts and mapping. For every benchmark in the corpus we log the raw N (as released by the source paper), the annotated N (items that completed the two-judge pipeline with a resolvable label), and the sampling rule (uniform random / stratified by country / etc.). The full manifest is released as `provenance_manifest.tsv` and is the single source of truth from which the numbers in Table 9, the L_2 mapping, and the coverage/depth statistics in Fig. 2 are regenerated.

L2-to-leaf mapping and its scope. Given a resolved L_2 label t_{L_2} and the annotator's justification string j, the leaf assignment is the deepest node of $\mathcal{T}$ on the path from t_{L_2} whose name (or any of its Wikipedia interlanguage aliases) is a substring of j; ties go to the leaf with highest $w(u \to v)$. Ambiguity rate on the 7-benchmark subset: 8.1% of items. **Leaf-level analyses are secondary, released as a resource for the community: all headline model-comparison and rank-disagreement results in §6 operate on the L_1 pillar and L_2 sub-domain labels and do not depend on the free-text leaf extraction step.**

Rarefaction and entropy-bias correction. Two separate corrections address two separate biases. *Rarefaction* standardises sampling effort across datasets so support-based statistics such as $\mathrm{Br}_{\mathcal{T}}$ are comparable across benchmarks of very different N: we subsample every benchmark to $N^{\star}{=}100$ items, repeat 200 times, and report the mean $\widehat{\mathrm{Br}}_{N^{\star}}$. *Miller–Madow* additively corrects the plug-in Shannon entropy $\hat{H}$ by $(|S|-1)/(2N\ln 2)$, addressing the well-known small-sample negative bias in entropy estimation (Basharin, 1959). Rarefaction is not itself an entropy-bias correction. Rank-order of benchmarks in Fig. 2 is stable across rarefaction budgets: Kendall's τ vs. the $N^{\star}{=}100$ ordering is 0.90 at $N^{\star}{=}50$ and 1.00 at $N^{\star}{=}200$.

Exclusive and Shapley coverage contributions. At the joint $L_2 \times \mathcal{B}$ resolution over the seven benchmarks in §6, the EC_k (Eq. 6) values in decreasing order are {INCLUDE : 0.10, PALM : 0.05, CLICK : 0.04, BERTAQA : 0.03, CULTURALBENCH : 0.02, INDOCULTURE : 0.02, BLEND : 0.01}; the Shapley ϕ_k values re-rank the same benchmarks (INCLUDE remains first) but redistribute credit for jointly covered cells, which we release in `shapley_l2b.csv`.

Pipeline–human agreement. On a stratified random validation set of 500 items (uniform over benchmark and Bloom level) labelled without seeing the LLM justifications, we report axis-appropriate metrics. For L_1 and L_2 culture labels, which are nominal, not ordinal, we report exact agreement and hierarchical tree-similarity (Zhou–Xue). For Bloom, which is ordinal, we additionally report quadratic weighted κ and ± 1-level agreement.

Axis	Exact agreement	Hierarchical similarity / κ_w	± 1-level agreement
Culture L_1	76.4%	tree-similarity 0.79	—
Culture L_2	63.2%	tree-similarity 0.71	—
Bloom	58.9%	quadratic $\kappa_w{=}0.44$	92.7%
Leaf (path)	—	hier. precision 0.68, recall 0.61	—

Table 5: Pipeline-vs-independent-human agreement on a stratified 500-item validation set. Metrics chosen per axis to match the axis type (nominal L1/L2, ordinal Bloom, hierarchical leaf).

Confusion matrices per axis and item-level annotations are shipped as `human_validation.tsv`. Human validation shows moderate ordinal agreement on Bloom ($\kappa_w{=}0.44$); together with the high raw agreement observed for strongly skewed datasets in Fig. 1, this supports interpreting very low *unweighted* κ values cautiously when label marginals are nearly degenerate. We do not claim the LLM–LLM and pipeline–human agreement statistics are directly comparable: they use different metrics, different aggregation grain, and different rater populations.

Cross-language asymmetry details. The maximum-benchmark-concentration statistic in Fig. 3 is outlier-sensitive by construction. For every language group we additionally release the per-benchmark median and range for each L_2 node in `crosslang_stats.tsv`; the ordering of the top-14 nodes changes only in one position under the median summary.

F MODEL EVALUATION DETAILS

Exact model provenance. API calls split into two disjoint roles (annotation vs. evaluation). We list the model registry identifier used for each role; exact snapshots, API dates, and access configurations are recorded in the provenance manifest.

Role	Model registry ID	Section
Judge 1 (annotation)	`google/gemini-2.5-flash`	§4.3
Judge 2 (annotation)	`openai/gpt-4o-mini`	§4.3
Arbiter (annotation)	`qwen/qwen-2.5-72b-instruct`	§4.3
Eval model 1	`openai/gpt-4o-mini`	§6
Eval model 2	`google/gemini-2.5-flash`	§6
Eval model 3	`anthropic/claude-haiku-4.5`	§6
Eval model 4	`meta-llama/llama-3.3-70b-instruct`	§6
Eval model 5	`qwen/qwen-2.5-72b-instruct`	§6
Eval model 6	`deepseek/deepseek-chat`	§6
Eval model 7	`google/gemma-3-27b-it`	§6

The annotation and evaluation roles share the model family *GPT-4o mini* but use different snapshots and decoding settings; the provenance manifest records the exact endpoint per role.

Common-support diagnostics. Table 6 reports how many items of each benchmark fall inside the common $L_1 \times \mathcal{B}$ intersection ($K{=}5$ cells) used for the q-standardisation in §6, plus the minimum per-cell item count.

Benchmark	Items in intersection	Min items / cell	Effective N per cell
BLEnD	1,316	12	263
BertaQA	1,211	11	242
CulturalBench	1,127	9	225
IndoCulture	581	7	116
INCLUDE	567	7	113
CLIcK	392	7	78
PALM	385	7	77

Table 6: Item counts inside the $L_1 \times \mathcal{B}$ common support (five shared cells across all seven benchmarks) used for the q-standardisation in §6. Effective N per cell = items inside / 5.

Leave-one-benchmark-out (LOBO) stability. Table 7 recomputes Spearman r between joint $\mathcal{T} \times \mathcal{B}$ profile distance and $1{-}\tau$ over the 15 dataset pairs remaining when each benchmark is held out. Direction is stable across all seven held-out settings.

Held-out benchmark	r_S	Nominal p	Sign vs. full $r_S{=}{+}0.46$
BLEnD	$+0.236$	0.397	same
BertaQA	$+0.337$	0.219	same
CulturalBench	$+0.678$	0.005	same
CLIcK	$+0.417$	0.122	same
INCLUDE	$+0.444$	0.098	same
IndoCulture	$+0.639$	0.010	same
PALM	$+0.383$	0.159	same
Mean $\pm$ sd	$+0.448 \pm 0.155$	—	7/7 stable

Table 7: LOBO stability of the RQ1 result. Sign is stable across all held-out settings; magnitude ranges from $+0.24$ to $+0.68$.

Item-bootstrap for the two candidate top-1 changes. Table 8 reports pairwise item-bootstrap results (1,000 resamples per benchmark) at the $L_1 \times \mathcal{B}$ standardisation target. The CulturalBench

nominal flip is directional (Llama beats Claude in 42.7% of resamples vs. Claude beating Llama in 12.9%); the BERTAQA candidate flip does not reproduce (Gemini remains ahead of DeepSeek in most resamples).

Benchmark	Contest	Δ (raw)	Δ (reweighted)	Winner freq. in bootstrap
CULTURALBENCH	Claude vs. Llama	$+0.025\,[-0.02, +0.07]$	$+0.023\,[-0.03, +0.10]$	Claude 12.9% / Llama 42.7%
BERTAQA	Gemini vs. DeepSeek	$+0.006\,[-0.05, +0.07]$	$+0.005\,[-0.03, +0.04]$	Gemini 18.4% / DeepSeek 11.2%

Table 8: Item-bootstrap ($1{,}000$ resamples) for the two candidate top-1 changes at the $L_1 \times \mathcal{B}$ target. Positive Δ favours the first model in the contest. "Winner freq." is the fraction of resamples in which each named model is the overall top-1 (across all 7 models on that benchmark). Neither result excludes zero; the CULTURALBENCH contest is directionally consistent with the reported nominal flip, the BERTAQA contest is not.

Exact Mantel permutation summary. Over all $7!{=}5{,}040$ benchmark-label permutations, the observed $|r_S|$ for the joint predictor (0.458) is exceeded in 122 permutations (one-tailed exact $p{=}0.024$); for content-only, 541 permutations ($p{=}0.107$); for cognition-only, 530 ($p{=}0.105$); for $\Delta_{\mathcal{T}}$, $1{,}194$ ($p{=}0.237$); for $\Delta_{\mathcal{B}}$, $1{,}228$ ($p{=}0.244$).

G PORTABILITY OF THE FORMALISM

The formalism in §2–§3 is domain-general: instantiating CCP on a different capability domain requires only a domain content taxonomy $\mathcal{T}$ and a cognitive-operation set $\mathcal{B}$ (Bloom is one such set). The graph-consensus induction recipe of §4.1 is likewise domain-general: it applies wherever multiple partially independent candidate hierarchies exist for the target domain (e.g., competing subject-classification schemes in science, textbook TOCs in mathematics, code-taxonomy proposals). The concrete Wikipedia instantiation used here is culture-specific in its use of Wikipedia category trees; instantiating CCP on other domains (e.g., science/mathematics, code, medical reasoning) is left to future work.

H ARTIFACT RELEASE AND PROVENANCE MANIFEST

The released artifact contains: (i) the consensus culture taxonomy $\mathcal{T}$ as a rooted graph with weights and interlanguage-link mappings; (ii) the 29-benchmark provenance manifest; (iii) item-level annotations for benchmarks whose licences permit redistribution, and mapping keys otherwise; (iv) prompts, model IDs, and decoding settings; (v) evaluation outputs from §6; (vi) the human-validation 500-item sample and its confusion matrices; (vii) code with pinned dependencies.

Item-level release schema. Each annotated item is released with:

```
dataset_id, item_id, split, source_language, region, culture_L1,
culture_L2, culture_leaf, bloom_level, judge_1, judge_2, arbiter_status,
human_status
```

Provenance manifest columns. Per-benchmark: `dataset_id`, `citation`, `URL`, `version/commit`, `license`, `split`, `original_N`, `annotated_N`, `sampling_rule`, `modality`, `languages`, `regions`, `taxonomy_labels`, `Bloom_labels`, `human_validation_status`. A dataset card, data statement, licence table, and reproducibility checklist accompany the release; annotator compensation and the double-blind human-validation protocol are documented alongside.

Benchmark	Region / focus	Items	Modality	Citation
Multicultural / global coverage				
ALM-BENCH	100 languages, 19 categories	22.7K	VQA	Vayani et al. (2025)
BLEND	16 countries, 13 languages	52.6K	Text (SA/MCQ)	Myung et al. (2024)
CALMQA	23 languages, long-form	51.7K	Text (LFQA)	Arora et al. (2025)
CANDLE	Global CCSK from web (C4)	1.10M	Text	Nguyen et al. (2023)
CULTURALBENCH	45 regions, 17 topics	1.7K (1.2K MCQ)	Text (MCQ)	Chiu et al. (2025)
CVQA	30 countries, 31 languages	10.0K	VQA	Romero et al. (2024)
GEOMLAMA	CN, IN, KE, US, IR	3.1K	Text (prompts)	Yin et al. (2022)
GLOBAL-PIQA	116 language varieties, PC	–	Text	Chang et al. (2025)
INCLUDE	52 countries, 44 languages	197.2K	Text (MCQ/SA)	Romanou et al. (2025)
TIEBE	23 regions, historical/current	–	Text	no formal cite
Values, norms & cross-cultural probes				
CDEVAL	Hofstede dimensions (global)	–	Text	Wang et al. (2024c)
CULTUREBANK	US-heavy social norms (TikTok/Reddit)	–	Text	Shi et al. (2024)
NORMAD	Social norms, 75 countries	2.6K	Text (stories)	Rao et al. (2025)
WORLDVALUESBENCH	Global (WVS Wave 7)	239 Qs $\times$ 93.7K resp.	Text	Zhao et al. (2024b)
East Asia				
CHARM	China (commonsense)	2.6K	Text	Sun et al. (2024a)
CLICK	Korea (culture + linguistics)	–	Text (MCQ)	Kim et al. (2024)
JMMMU	Japan (culture-aware MMMU)	1.3K	VQA	Onohara et al. (2025)
South Asia				
DRISHTIKON	India	–	Text	no formal cite
SANSKRITI	India	–	Text	no formal cite
Southeast Asia				
INDOCULTURE	Indonesia, 11 provinces	2.4K	Text (MCQ)	Koto et al. (2024)
SEA-VL	Southeast Asia (multi-country)	$\sim$1.28M imgs	VQA/VL	Cahyawijaya et al. (2025)
SEAEVAL	SE Asia + global (28 tasks)	–	Text (multi-task)	Wang et al. (2024a)
Europe				
BERTAQA	Basque (parallel English)	4.8K	Text (MCQ)	Etxaniz et al. (2024)
CUS-QA	Czechia, Slovakia, Ukraine	3.9K	Text + VQA	Libovický et al. (2025)
Latin America				
BROVERBS	Brazil (Portuguese proverbs)	–	Text	arXiv:2509.08960
LATAMQA	Latin America (Wikidata-grounded)	–	Text	no formal cite
Middle East & Arabic (separate sub-corpus)				
PALM	22 Arab countries	17.0K	Text	Alwajih et al. (2025)
ARADICE-CULTURE	Gulf, Egypt, Levant	–	Text	Mousi et al. (2025)
JAWAHER	20 Arabic dialects (proverbs)	10.0K	Text	Magdy et al. (2025)

Table 9: Full cultural benchmark corpus (29 total: 26 non-Arabic + 3 Arabic-focused), grouped by cultural region. Sizes from the original papers; annotated N per benchmark is released in the full provenance manifest. "–" = no official count; "no formal cite" = no paper recorded. Abbreviations: MCQ multiple-choice, SA short-answer, LFQA long-form QA, PC physical commonsense, VQA visual QA, VL vision-language, CCSK cultural commonsense knowledge.