LARGE-SCALE FACTOR ANALYSIS SHOWS MACHINE INTELLIGENCE IS ONLY PARTIALLY INTERPRETABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

A common assumption in language model development is that cognitive abilities are organized around a general, domain-free intelligence factor, like fluid intelligence in humans. This assumption is rarely tested directly, and prior attempts have done so only at a much smaller scale. We take a latent variable approach to intelligence in language models, similar to how psychometricians study psychological constructs. Performance in every specific problem set is influenced by a domain-specific and a domain-agnostic latent factor. Using factor analysis as a dimension-reduction technique, we analyzed 13,251 published evaluation scores covering 1,618 language models across 456 different text-only benchmarks. Due to the super-sparse nature of the dataset, we triangulate our analysis across different data densifiers and imputation methods. A robust pattern across different modes of bias is that 1. A general intelligence factor accounts for 70.8% of variance in model performance at our most generous estimate, and far less than that in most of our solutions, 2. Content-similar benchmarks do not necessarily cluster together, and 3. The g factor is not dominated by any common theme, and there is a lack of evidence that it is well-proxied by standard "intelligence" benchmarks. Our findings go against current endeavors of defining, identifying, and targeting general intelligence as a tangible construct in language model development. This leaves the strategy of targeting a single conceptual ability without support, since the first-order abilities it would have to reach are often partially idiosyncratic and not identifiable in practice.

1 INTRODUCTION

The field of artificial intelligence is full of theories about what intelligence "is" (Chollet, 2019; Morris et al., 2023; Waterhouse, 2023). This is partly because neural networks tend to overfit their training data, such that their performance across tasks lacks the generalizability of human cognition (Zhang et al., 2021a). In pursuit of this generalizability, models are often implicitly assumed to organize cognitive abilities according to a causal hierarchy, much like humans, where a general-purpose, "content-free" capability underlies and transfers across many specific downstream tasks (Bommasani et al., 2021). A popular label for this hierarchy, borrowed from psychometric research, is "fluid intelligence", defined as the ability to perform well on any task regardless of content (Cattell, 1963). By definition this construct holds a causal relation, where an increase of fluid intelligence leads to an increase in the mastery of every other cognitive task.

The concept of general fluid intelligence is borrowed from human intelligence psychometrics, which discovers the higher-order structure of cognitive abilities by applying dimension reduction techniques to a large set of cognitive measures. A growing body of work in machine learning follows this step by analyzing LLM benchmark scores to ascertain the nature of "general intelligence" in machines. However, a common weakness across these studies is the relatively small number of benchmarks or variables, ranging from 6 to 23, most of which measure "smartness" narrowly (mathematics, academic knowledge, common-sense reasoning) and which risk conflating the extracted capability factor with model scale (Kearns, 2026). These studies also tend to fit a human-informed structure, and do not look for bottom-up patterns specific to LLMs.

Within the context of prior work, our study presents factor analysis of 13,251 evaluation scores from 1,618 models and 456 benchmarks, the widest benchmark coverage to date, to the best of our knowledge. Our pool of benchmarks, unlike previous work, covers a highly diverse set of tasks,

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.

including those quite outside of the mainstream. A useful analogy is that reaction time in humans is correlated with intelligence (Kranzler & Jensen, 1989), showing that no matter how seemingly unimportant a task may be, it may provide information entirely unintuitive to our subjective understanding. Across this analysis, we find that factor patterns are only partially interpretable and often incoherent, and that the evidence for a "general intelligence" factor proxied by commonly-targeted "reasoning" benchmarks is limited, with the highest loadings more often going to miscellaneous tasks with no common theme.

2 Background and Related Work

2.1 The Psychometric Paradigm

A common theoretical ground in psychometric measurement is that observable human behaviors are **causally** influenced by latent variables internal to an individual (Borsboom et al., 2003; 2004), and Federiakin (2025) applies the same distinction to the design of LLM benchmarks. Individuals differ in some latent factor, hence individuals differ in some observable outcomes. If there are no causal latent factors, then the covariance among observed behaviors would have no source at all, which makes little sense.

Personality and intelligence research are prime examples. Allport & Odbert (1936) extracted all or most of the words from the English dictionary that can describe someone's personality, and had a large sample self-report how well each word describes themselves, hence measuring as much of the set of all possible personalities (John et al., 1988). Spearman (1904) did much the same for intelligence, collecting the scores of students across school subjects and finding that their variance overwhelmingly loads on a single wide-breadth latent variable called the g factor (Jensen, 2002), which remains generally accepted (Johnson et al., 2004; 2008). Both follow the same paradigm. Exhaustively measure observable behaviors, subject them to dimension reduction techniques, and draw theories from the resulting latent factor. Subsequent research decomposes the hierarchy further, into facets for personality (Lee & Ashton, 2018; DeYoung et al., 2007) and into specific cognitive abilities for intelligence (Schneider & McGrew, 2018).

2.2 Causal Assumptions in Machine Intelligence

Various LLM benchmarks are known to be intercorrelated, though the origin of this covariance is rarely stated explicitly. We argue that benchmark correlations originate causally from latent variables, and add that existing paradigms of machine intelligence are implicitly causal. The assumed precedence of a content-free intelligence (Chollet, 2019) already implies such a relation, and model development relies on it. In aiming to achieve "general intelligence", developers tend to train in a *targeted* manner. Reasoning-oriented post-training, for instance, is motivated by the expectation that improvements on "pure" logical tasks will transfer to tool-calling, long-horizon agentic tasks, and coding (DeepSeek-AI, 2025). Such an expectation is coherent only if the targeted ability stands in a causal relation to the rest, which we state as follows.

Definition 1. Let F be a set of latent factors and T the set of performance scores over all possible tasks:

$$F = \{\, f_i \mid i \in \mathcal{I} \,\}, \qquad g \longrightarrow F \longrightarrow T, \quad F \not\longrightarrow g.$$

That is, g is a higher-order factor that causally affects the set of latent factors F, through which it in turn influences task performance. No causal path runs from F back to g.

There is a growing assumption in the field that a subset of T, mostly assumed to be reasoning, mathematics, and coding, measures g better than the rest. When a model is fine-tuned to perform better on such tasks, the implicitly expected transfer happens because training improves g and, through it, the factors F:

Definition 2. Let F be a set of latent factors and T the set of performance scores over all possible tasks. Task performance t_i is given by a linear predictor:

$$t_i = \lambda_g g + \lambda_1 f_1 + \cdots + \lambda_k f_k + \epsilon_i, \qquad i \in \mathcal{I},$$

where $\lambda_g, \lambda_1, \ldots, \lambda_k$ are the task's factor loadings (standardized regression coefficients) and ϵ_i is the task's unique variance.

2.3 Interpreting Structural Patterns

Understanding the patterns of benchmark correlation is theoretically relevant to cognitive science, though this relevance does not concern any one specific factor loading pattern. Under a small number of benchmarks the discovered patterns depend on which benchmarks were sampled (Major et al., 2011), and under a large number the missingness pattern becomes too large to trust any single solution. The correct approach is therefore to examine recurring and large-scale patterns. One such pattern is the explanatory power[1] of a general intelligence factor. When a general factor dominates the explained variance of an EFA result, variance in an LLM's capability is predominantly caused by a latent g, supporting the idea that a flexible, "raw" intelligence is a substantial component of performance and that targeting it is a fruitful research program. A weak g factor instead means that LLM abilities are strongly specific, so improvements require brute-force training in as much task diversity as possible, with no silver-bullet construct that parsimoniously describes "general intelligence". A second pattern is which benchmarks load highest. Under the indifference of the indicator (Spearman, 1904), content does not determine loading, so the test is whether a factor recovered from a small battery survives a wider one.

We read these patterns without assuming what a general factor should look like, since factor analysis is a bottom-up, theory-free approach. Authors tend to impose theories of human intelligence on artificial neural networks, as we know of no better model of intelligence than our own. However, it is problematic because these networks present an entirely different form of cognitive process than that of biological minds. Human intelligence research itself began with factor analysis (Spearman, 1904) and continues to apply bottom-up dimension reduction even though its theories have been well established for decades. Therefore, we do not assume any general factor to have a tidy definition, and so we do not discriminate between benchmarks and try to sample as much of the set of all possible tasks as we can. We argue that it is a reasonable position to expect that LLMs work in ways entirely unintuitive to the human mind, and having no priors whatsoever is the correct way to start our inquiry. Purpose-built diversity suites such as BIG-bench (Srivastava et al., 2023) are denser, but they sample one team's construction of task diversity. We are interested in the benchmarks the field uses.

2.4 Related Work

Prior work applying factor analysis to benchmark scores is smaller in scale, and mostly confirmatory. Ilica & Gignac (2024) applied confirmatory factor analysis to hundreds of models and found that their performance fits a human-informed causal structure, which repeats the mistake of fitting an existing theory to model intelligence. A large proportion of their benchmarks are also variants of MMLU (Hendrycks et al., 2021b), posing the risk of common-method bias polluting the model fit. Federiakin (2025) likewise fits a single-factor CFA to the HuggingFace Open LLM Leaderboard. A confirmatory fit can only test the structure its authors specify in advance. Hardy et al. (2026) fit CFA and generalizability theory to over 4,000 models on the same leaderboard, and find local dependence among its items, which is the common-method bias we raise above.

Krakauer (2026) instead applies PCA and finds a first principal component declining from 90% of variance to 64% by 2024, interpreted as a "rotation" in the g factor as models outsource reasoning to external tools. PCA does not partition out systematic from error variance and so cannot be trusted to isolate a genuine g factor from noise. Only two studies run EFA. Burnell et al. (2023) found three major factors, but did not do higher-order factor analysis on the resulting loadings, so they cannot say to what extent those factors are influenced by a presumable g factor. Haznitrama et al. (2026) report a unified general factor before showing that a neuropsychologically-grounded battery reveals cognitive gaps that it otherwise masks. Narrower still is Holm et al. (2024), where a single factor explains 95% of variance across just 8 scenarios in one language. Item response theory, recently applied to LLM benchmarks (Maia Polo et al., 2024; Zhou et al., 2025), and full-information maximum likelihood both handle incomplete data natively. IRT requires per-item responses, which published scores do not report, and FIML requires every benchmark pair to share observed models, which many pairs at our sparsity do not.

[1] We use the phrase "explanatory power" over "existence", as g is a construct of the factor model we fit.

3 METHODOLOGY

In this study we investigate the low-dimensional structure of model benchmark scores, which comprises distinct but correlated latent factors that each dominantly affects different clusters of benchmarks, and a g factor that accounts for the variances of all benchmarks. Our raw $1{,}618 \times 456$ matrix is super-sparse (~1.8% density) and missing not at random (MNAR) (Rubin, 1976), since popular models and popular benchmarks are observed far more often. We therefore apply several densification and imputation methods, each with its own biases and assumptions, and triangulate their results to find a common characteristic.

3.1 DATA COLLECTION

Scope. We limit this study to the evaluation of generative language models on a set of text-only benchmarks. We define generative language models as those that accept arbitrary prompts and produce text completions. Encoder-only classifiers, narrow task-specific systems (dedicated MT/ASR/TTS models), and undocumented community uploads are excluded. Benchmarks are included if they have at least one in-scope result row. For each row, we follow the schema from EveryEvalEver (Batzner et al., 2026) to unify the evaluation results. Full inclusion and exclusion criteria are given in Appendix C.5.

Sources. Table 1 gives the composition of the text-only corpus by source family. Appendix C lists every named source and the extraction route used for each.

Table 1. Composition of the text-only corpus by source family.

Source family	Rows	Benchmarks
Stanford HELM	4,942	138
HF Open LLM Leaderboard	4,529	12
Papers With Code	1,378	151
Other online leaderboards	971	67
Kaggle AI Benchmarks	844	26
Primary papers	587	95

Table 2. Aggregated model $\times$ benchmark matrices, text-only corpus. "Retained" is the fraction of observed cells surviving the densifier peel.

Densifier	Strategy	Shape	Density	Retained
raw	Std.	$1{,}266 \times 404$	2.2%	
raw	Aggr.	334×380	3.5%	
C	Std.	671×78	13.8%	65%
C	Aggr.	201×102	13.6%	63%
S	Std.	669×124	10.0%	75%
S	Aggr.	124×293	10.0%	81%
R	Std.	175×298	11.8%	55%
R	Aggr.	97×310	11.7%	78%

Protocol. Given the large number of fields from the EveryEvalEver schema, we follow a strict source-verification protocol, meaning each field is populated only if the verified source explicitly documented it. As an exception, some fields can be inferred from the record itself using deductive rules with small risk of error (see Appendix C.3). In particular, we put some focus on obtaining the release date field for both models and benchmarks, and only accept a release date if it is explicitly documented in the source (Appendix C.4). We also handle redundancy at two levels. Redundant duplicate rows are removed (Appendix C.6), and near-perfectly correlated benchmark identifiers (version, dialect, or language splits of one benchmark) are collapsed to one representative, which removes 47 identifiers and 2,216 rows (Appendix D).

3.2 DATA PROCESSING

Deduplication. Rows surviving duplicate removal for the same (model, benchmark) pair are averaged into a single score. To keep near-duplicate results from the same models under different conditions (e.g. reasoning effort) from breaking the IID assumption, and to further densify the data, we average rows under two collapse strategies. The **standard** strategy is variant-level. It keeps different version numbers and parameter counts, while collapsing reasoning effort, knowledge cutoff, etc. The **aggressive** strategy is family-level. Every model is collapsed to its base family token (Claude, Llama, etc.), so all sizes and generations of a family form one row.

Metric selection. 92 of our collected benchmarks were reported under several metrics. As different metrics are not comparable, we keep one metric covering the most distinct models, so the widest

comparable population survives. A per-benchmark override list handles cases where coverage alone chooses badly. Finally, benchmarks observed for only one model are dropped.

Densification. We further improve the data density by greedily peeling the matrix towards a common target density. We feature three densifiers, differing by the axis they sacrifice: **C (column-primary)** drops the least-observed benchmarks, then any model left empty, retaining famous benchmarks with wide model coverage. **R (row-primary)** drops the least-observed models, then any benchmark left empty, retaining a broad benchmark set over a small set of heavily-evaluated models, which leaves fewer models than benchmarks. **S (symmetric)** drops whichever marginal has the lowest fill-rate, privileging neither axis. After peeling, models and benchmarks that have fewer than 3 observed scores are dropped. Columns with zero variance among observed values are also dropped, since they carry no correlational signal. The algorithm is given in Appendix F. We set the target density at 10%, which is low, so that we can include as many different benchmarks as possible.

Imputation. We applied two families of missing data imputers: **full-dataset** algorithms (SoftImpute (Mazumder et al., 2010), k-NN, and missForest (Stekhoven & Bühlmann, 2012)) and **correlation recovery** which includes one-sided matrix completion (Cao et al., 2023), USVT (Chatterjee, 2015), and SoftImpute on the correlation matrix. For the correlation recovery we also use a simple filling method, where missing entries are either filled with zeros or the mean, followed by PSD smoothing. Method-level descriptions and implementation details for all of the above are given in Appendix G.

Evaluating the imputations. At each stage of the imputation, we mask ~20% of the observed cells as an evaluation set. The mask is column-stratified, so each benchmark is masked at least once and keeps at least two training observations. Held-out cells are scored by $R^2 = 1 - \mathrm{MSE}/\mathrm{MSE}_{\text{baseline}}$, where the baseline predicts each column's training mean (Appendix H). R^2 selects the hyperparameters within each method (rank, k, number of trees, etc.), and imputations whose held-out R^2 falls below 0.2 are not factored.

3.3 Factor analysis

To perform our dimension reduction, we use exploratory factor analysis (EFA) with the minimum residual estimator and the oblique promax rotation[2], followed by the Schmid-Leiman bifactor transformation (Schmid & Leiman, 1957). This runs factor analysis hierarchically, yielding one additional factor that influences the rest of the extracted factors, and is widely used in psychometric research on the g factor of intelligence (Johnson et al., 2004; 2008). The hierarchical step makes it a more principled choice than interpreting the highest-eigenvalue solution (e.g., Krakauer, 2026) as the g factor.

From the bifactor solution we use the ω_h coefficient to quantify the variance explained by the g factor.

Definition 3. Let T be a matrix of test scores that can be decomposed into independent additive components due to a **general factor** (g), **specific factors** (s)[3], and **error** (ϵ). Then ω_h is the estimand

$$\omega_h = \frac{\sigma_{\mathrm{g}}^2}{\sigma_T^2}, \qquad \sigma_T^2 = \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{s}}^2 + \sigma_\epsilon^2.$$

One additional step we do is parallel analysis (Horn, 1965) to select the number of factor analysis dimensions. To keep wall-clock time tractable we cap the number of factors extracted at 20.

3.4 Label Cohesion Analysis

We measure benchmark similarity as cosine distance between benchmarks' factor-analytic loading vectors, described further in Appendix J. Given this benchmark–benchmark distance matrix, we ask whether each label's members sit closer together than chance. For a label, let **within** be their mean pairwise distance. We compare this to the **null_mean**: benchmarks are binned into quartiles of $\log_{10}$(model coverage), and the null redraws, 2,000 times, a random same-size set matched to the label's members' coverage-quartile composition, recomputing the same mean-pairwise-distance

[2]Rotation gives an interpretable "simple structure". Oblique rotations allow the factors to correlate, while default eigendecomposition yields orthogonal solutions.

[3]Usually called "group factors". We avoid the term to prevent confusion with observation grouping.

statistic each draw. Coverage stratification matters because coverage is itself correlated with tightness in this factor space. Labels with fewer than four members, or comprising the entire benchmark set, are not scored. The reported effect size is:

$$A = 1 - \frac{\text{within}}{\text{null_mean}}$$

0 means the label's members are no tighter than a random draw, 1 means effectively identical members, and negative values mean the members are more spread out than chance.

Significance. The p-value is computed empirically from the same 2,000 permutation draws:

$$p = (1 + \#\{\text{null draws at least as tight as observed}\})/(1 + 2000)$$

Within each cell, we control the false discovery rate across the labels of one axis via Benjamini–Hochberg at q = 0.05. A label's reported significant count (e.g. "6/18") is the number of cells in which it passed this FDR-corrected threshold, out of the cells where it could be scored.

4 RESULTS

4.1 VARIANCE EXPLAINED BY g

Only 20 dataset-imputer combinations yield an R^2 above the threshold (Appendix N).

Table 3 below shows the point summaries of the factor analyses. The most important statistic here is the ω_h, which indicates the degree of indicator variances explained by the general factor. **At most, a universally causal g factor accounts for 70.8% of variance in model performance**.

The range of ω_h is wide. The best-performing imputer, SoftImpute on S Std., yielded a g factor that accounts for 25.7% of the variance. ω_h also stays in this range when we split the models by release year. Newer models score higher on the g factor, but the variance it accounts for shows no trend with release year (Appendix K). Leaving out single benchmarks shows that the change in ω_h is only weakly related to how often a benchmark is observed (mean $|r| = 0.18$, Appendix O).

Table 3. Point summaries of factor analysis results. AVE = average variance explained per factor, k = number of factors extracted, ϕ_{avg} = average inter-factor correlation.

Dataset	Imputer	k	AVE	ω_h	ϕ_{avg}	R^2
C Std.	Mean fill	14	5.5%	0.708	0.142	0.224
C Std.	missForest	4	22.1%	0.695	0.398	0.399
S Std.	SoftImpute (corr.)	5	9.3%	0.676	0.306	0.378
C Std.	Zero fill	14	5.4%	0.621	0.093	0.286
C Aggr.	missForest	4	19.3%	0.521	0.112	0.241
C Std.	k-NN	7	10.8%	0.516	0.209	0.288
C Std.	SoftImpute (corr.)	4	14.6%	0.514	0.247	0.317
R Std.	SoftImpute	20	4.6%	0.367	0.012	0.290
S Std.	SoftImpute	5	18.3%	0.257	0.094	0.504
C Std.	SoftImpute	9	10.5%	0.242	0.038	0.493
S Aggr.	SoftImpute	20	4.7%	0.225	0.031	0.282
S Std.	k-NN	11	6.9%	0.204	0.048	0.296
R Aggr.	SoftImpute	20	4.7%	0.187	0.008	0.209
C Aggr.	SoftImpute	5	17.8%	0.183	-0.001	0.337
raw Aggr.	SoftImpute	10	8.9%	0.132	0.013	0.228
C Std.	OneSidedMC	2	50.0%	0.102	0.125	0.321
raw Std.	SoftImpute	10	9.0%	0.071	-0.010	0.249
C Aggr.	OneSidedMC	2	50.0%	0.065	0.097	0.278
S Std.	missForest	4	22.5%	0.032	0.054	0.471
S Std.	OneSidedMC	2	50.0%	0.014	-0.040	0.365

4.2 BENCHMARK CLUSTERS

Figure 1 below shows an illustrative UMAP plot of benchmarks using composite distances aggregated from factor loadings, colored based on their subject matter (Appendix J). In this plot, benchmarks with a common subject only occasionally cluster together. This is supported by the low cohesion scores in Table 4, where among these labels only `code` shows a weak cohesion, while the others show no structure distinguishable from chance. Narrow labels such as `finance` are cohesive, but they have very few members (Appendix M). Across the entire figure, the spaces occupied by each flagged subject matter span across the entire plot. For coding, `livecodebench`, `swe_bench`, and `humaneval` stand far apart from each other, and the same is true for math with the benchmarks `gsm8k`, `math`, and `aime25`. In other words, **capability in one task does not always generalize to another task of the same subject**. A semantically coherent generalization is probable but not guaranteed, which makes domain abilities difficult to isolate from a purely semantic and intuitive standpoint. This phenomenon, where same-domain benchmarks lack a tendency to cluster together, is observed in nearly all of our imputations, which we discuss further in Appendix Q.

Table 4. Cohesion of subject labels. Full cohesion results for every label can be seen in Appendix M.

Label	Median A	Significant
`code`	+0.264	6/18
`math`	+0.031	0/18
`reasoning`	-0.008	0/18

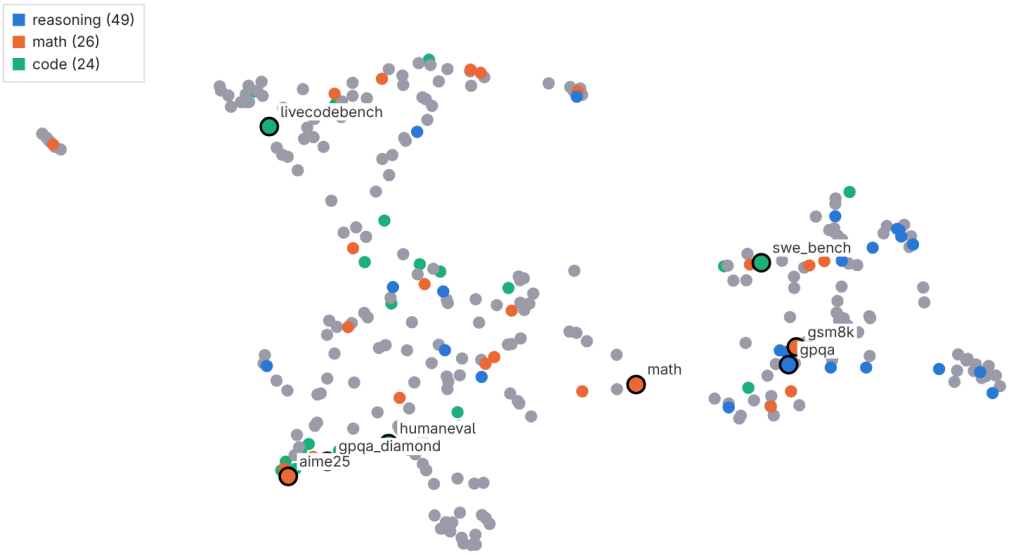

Figure 1. UMAP plot of benchmark distances from S, SoftImpute. Domain-similar benchmarks are not guaranteed to cluster together.

4.3 BENCHMARKS' g-CENTRALITY

Another point of interest for the research question is which benchmarks act as a good proxy of general intelligence, particularly as research is concerned with performance in certain specific benchmarks to quantify intelligence advancements. Table 5 answers this question by ranking the benchmarks by the normalized average rank order[4] of their loadings on the g factor (ρ). To account for confounding effects from benchmark frequency, we report the residuals of the rank order regressed on frequency (ρ_ϵ). Further details and justification are given in Appendix L.

[4]We use rank order as factor loadings vary in range, and they are normalized as different datasets have different numbers of benchmarks.

The top benchmarks are not dominated by common standard benchmarks. The top proxies include measures of creativity, legal use cases, and even emotional intelligence. There is no evidence that a g factor resembles anything like abstract reasoning. This diversity is expected on its own, since a general factor is indifferent to the content of its indicators. To add to this, standard intelligence benchmarks like `arc` and `gpqa_diamond` are placed near the middle of the rankings. **Our results are evidence that the prevailing assumption that reasoning, mathematics, and coding benchmarks are the best proxies of general intelligence does not hold**.

Table 5. Benchmarks and their average normalized rank order (ρ) of their g factor loadings. Sorted by frequency-residualized rank (ρ_ϵ). ρ ranges from 0 to 1, where 0 = ranked first, 1 = ranked last. N = number of EFA estimations with that benchmark. CI and Best/Worst refer to ρ.

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N
1	bhasa	-0.346	0.141	[-0.005, 0.287]	0.024	0.318	5
2	mtrag	-0.341	0.147	[-0.051, 0.346]	0.021	0.394	5
3	creativityprism	-0.339	0.156	[-0.019, 0.332]	0.026	0.367	5
4	eqbench	-0.332	0.156	[-0.060, 0.372]	0.017	0.451	5
5	mceval	-0.331	0.156	[0.024, 0.288]	0.051	0.333	5
6	pwc_svamp	-0.329	0.169	[-0.033, 0.371]	0.058	0.298	4
7	pwc_drop_test	-0.328	0.162	[-0.139, 0.462]	0.008	0.431	4
8	ProphetArena	-0.315	0.183	[-0.036, 0.401]	0.092	0.385	4
9	dialogbench	-0.295	0.210	[-1.351, 1.770]	0.087	0.332	2
10	pwc_piqa	-0.282	0.253	[0.162, 0.345]	0.003	0.822	20
57	gsm	-0.169	0.319	[0.162, 0.476]	0.000	0.936	20
138	arc	-0.056	0.356	[0.230, 0.481]	0.049	1.000	20
139	gpqa_diamond	-0.054	0.475	[0.355, 0.595]	0.016	0.992	20
202	gsm8k	+0.017	0.413	[0.263, 0.563]	0.000	1.000	20
327	humanitys_last_exam	+0.181	0.672	[0.162, 1.182]	0.047	0.966	5

5 Discussion

5.1 Uninterpretable hierarchical structure

Classically, cognitive sciences have approached intelligence from verbal, conceptual beginnings (Legg & Hutter, 2007). Intelligence is given a definition, then systems are developed according to the specs provided by the definition. This is not true in psychometric paradigms, which start by analyzing the covariance of performance measurements. Thus, adopting the psychometric approach for LLM benchmarks, our study has found that any substantially effective g factor cannot be coherently approximated. Adding to this, tasks with similar contents, like mathematics and coding, do not necessarily cluster together. Machine intelligence does not follow any coherent structure or definition.

Unlike theories of human psychology (Schneider & McGrew, 2018) where intelligence is explicitly assumed to possess higher-order structure, with fluid intelligence as one higher-order factor that has theoretical causal effects (van der Maas et al., 2014) on the performance of other cognitive abilities, abilities in LLMs like abstract reasoning, domain memorization, and agentic tool use, are more likely to be horizontal with respect to each other with no clear higher-order structure.

As factor analysis assumes a reflective data generating process, the next question is whether such causality can be observed and exploited in practice. One way is to follow psychological science, where training participants in one cognitive task leads to observable improvements in a different task (hence, generalization) (Simons et al., 2016). For an LLM, this means measuring a model's performance in coding, fine-tuning it to be better at mathematics, and then measuring its coding performance again. If there is an improvement between pre- and post-training performance, a causal mechanism can be empirically verified. More generally, one can fine-tune models to be better at benchmarks which are (assumed to be) proxies of general intelligence, then compare the pre- and post-training performance in a wide array of tasks.

Establishing such causal mechanisms would be practically important. If there were no causal generalization, then no blanket improvement of abilities could be gained by improving one family of

tasks. A generally-intelligent model can only be trained by including all relevant tasks in the training corpus, and there is no silver-bullet construct or ability that can be efficiently targeted in training that will causally improve aptitude in specific tasks. Such a model would require a brute-force, increasingly expansive training dataset until it covers the universe of all possible tasks.

5.2 Why do abilities correlate?

If we accept the g factor as a mixture of correlations, it remains to explain why these correlations appear in the first place. Explanations of multi-task learning suggest that cross-domain transfer occurs from a convergence to a representation shared by the trained tasks (Caruana, 1997). Contemporary paradigms like reasoning (Tang et al., 2025) suggest that tasks are decomposable into semantic primitives subject to logical/symbolic manipulation in the representational layers. However, in line with our findings, Meng et al. (2026) find that input-similarity measures identify the best transfer source for at most 2 of 8 target tasks, and that transfer is better predicted by the gradients a task induces than by its content. We therefore expect correlations between abilities to arise from highly unintuitive conditional token probabilities, which are not summarizable into human-like analogies.

5.3 Alternative to a latent factor view

Through this paper we have shown that the implicitly causal model of intelligence yields an incoherent structure. However, factor analysis by itself is an analysis of correlations, and is merely compatible with a causal model. The formative paradigm, usually associated with PCA, makes no claim about the nature of the resultant components (van der Maas et al., 2014). Even so, generalizability would be impossible without a common factor to begin with, since two tasks sharing a dominant factor decompose into a similar lower-level representation (Caruana, 1997; Menghi et al., 2025) that the network must discern. The mutualist paradigm (Borsboom & Cramer, 2013), which describes indicators as nodes in a graph of mutual causality, is harder to rule out, since a mutualist process reproduces a positive manifold without any common cause (van der Maas et al., 2014).

Mutualism (van der Maas et al., 2017) models specific abilities as having mutual causal facilitation, with no higher-order, few-dimension causal factors. For example, agentic ability could improve tool use, which in turn improves coding. The output of this analysis is a graph of partial correlations between abilities (e.g., network analysis (Borsboom & Cramer, 2013)).

This view is attractive, as it does not require a supposedly parsimonious (van der Maas & Kan, 2016) single latent variable to explain the origins of correlations (which, per our findings, are mostly unintelligible). The caveat, if this mechanism is correct, is that the g-oriented research program is untenable, as g is merely a summary scalar and not a real, targetable ability. It also leaves inter-task correlations unintuitive and not bound by domain similarity. It is a more informative model, but it does not explain the uninterpretability of the correlation structure.

6 Conclusion

We asked how the cognitive abilities of language models can be organized around a general factor. Our answer comes from factor analyses of 13,251 published evaluation scores covering 1,618 models and 456 benchmarks, triangulated so that no single treatment of the missing data carries the finding. At best, a general latent factor accounts for 70.8% of variance in model performance, but most of our solutions put it well below that. Domain-similar benchmarks are only occasionally located close together in vector space, and benchmarks with high g-loading do not share a common theme and differ from the benchmarks the field treats as standard measures of intelligence. We conclude that a hierarchical causal order of machine intelligence, as commonly but only implicitly assumed in the field, is untenable in practice given the uninterpretable factor structure. Correlations arise from unintuitive and unpredictable common features, and a non-hierarchical network of abilities is possibly a better model for machine intelligence.

References

Asad Aali, Dave Van Veen, Y. Arefeen, Jason Hom, C. Bluethgen, E. Reis, et al. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the*

American Medical Informatics Association, 32(3):470–479, 2024.

Asma Ben Abacha, Yassine Mrabet, Mark E. Sharp, Travis R. Goodwin, S. E. Shooshan, and Dina Demner-Fushman. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, volume 264 of *Studies in Health Technology and Informatics*, 2019a.

Asma Ben Abacha, Chaitanya P. Shivade, and Dina Demner-Fushman. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 370–379, 2019b.

Asma Ben Abacha, Wen wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, et al. Medec: A benchmark for medical error detection and correction in clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22539–22550, 2025.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, et al. Irokobench: A new benchmark for african languages in the age of large language models. In *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. 2732–2757, 2025.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, et al. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267, 2023.

Elif Akata, L. Schulz, Julian Coda-Forno, Seong Joon Oh, M. Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025.

Neel Alex, Eli Lifland, Lewis Tunstall, A. Thakur, Pegah Maham, C. J. Riedel, et al. Raft: A real-world few-shot text classification benchmark. In *Advances in Neural Information Processing Systems*, 2021.

Emad A. Alghamdi, Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. Aratrust: An evaluation of trustworthiness for llms in arabic. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8664–8679, 2025.

Gordon W. Allport and Henry S. Odbert. Trait-names: A psycho-lexical study. *Psychological Monographs*, 47(1):i–171, 1936.

Ebtesam Almazrouei, Ruxandra-Aimée Cojocaru, M. Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, et al. Alghafa evaluation benchmark for arabic language models. In *Proceedings of ArabicNLP 2023*, pp. 244–275, 2023.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, 2020.

Artificial Analysis. Artificial analysis intelligence index. `https://artificialanalysis.ai/`, 2025.

Thura Aung, J. Montalan, Jian Gang Ngui, and Peerat Limkonchotiwat. Burmese-san: Burmese nlp benchmark for evaluating large language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 224–245, 2026.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, H. Michalewski, David Dohan, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Don Husa, et al. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, 2024.

Jan Batzner, Sree Harsha Nelaturu, Damian Stachura, Anastassia Kornilova, Jon Crall, Tommaso Cerruti, Yanan Long, Yifan Mai, Sanchit Ahuja, Asaf Yehudai, Marek Šuppa, John P. Lalor, Oluwagbemike Olowe, Jatin Ganhotra, Brian H. Hu, Eliya Habba, Andrew M. Bean, Chang Liu, Sander Land, Steven Dillmann, Aniketh Garikaparthi, Elron Bandel, Saki Imai, James Edgell, Wm. Matthew Kennedy, Jenny Chim, Patrick Meusling, Asteria Kaeberlein, Venkata Ramachandra Karthik Chundi, Manasi Patwardhan, Martin Ku, Austin Meek, Leon Knauer, Brian Wingenroth, Srishti Yadav, Usman Gohar, Felix Friedrich, Michelle Lin, Jennifer Mickel, Arman Cohan, Stella Biderman, Irene Solaiman, Zeerak Talat, Anka Reuel, Mubashara Akhtar, Gjergji Kasneci, Avijit Ghosh, and Leshem Choshen. Every eval ever: A unifying schema and community repository for AI evaluation results. *arXiv preprint arXiv:2606.14516*, 2026.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. Peerqa: A scientific question answering dataset from peer reviews. In *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. 508–544, 2025.

S. Bedi, Hejie Cui, Miguel Fuentes, A. Unell, Michael Wornow, J. Banda, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, G. Sidorov, O. Kolesnikova, et al. Culemo: Cultural lenses on emotion - benchmarking llms for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18894–18909, 2025.

Jonathan Berant, A. Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, 2013.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. Carb: A crowdsourced benchmark for open ie. In *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. 6262–6267, 2019.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020.

Su Lin Blodgett, L. Green, and Brendan T. O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, 2016.

Ondrej Bojar, C. Buck, C. Federmann, B. Haddow, Philipp Koehn, Johannes Leveling, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, 2014.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Daniel Borkan, Lucas Dixon, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500, 2019.

Denny Borsboom and Angélique O. J. Cramer. Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, 9(1):91–121, 2013.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap Van Heerden. The theoretical status of latent variables. *Psychological Review*, 110(2):203, 2003.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap Van Heerden. The concept of validity. *Psychological Review*, 111(4):1061, 2004.

Ryan Burnell, Han Hao, Andrew R. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

Sergiu Bursuc, Theodore Ehrenborg, Shao-Wei Lin, L. Astefanoaei, Ionel Emilian Chiosa, Jure Kukovec, et al. A benchmark for vericoding: formally verified program synthesis. *Transactions on Machine Learning Research*, 2026.

Steven Cao, Percy Liang, and Gregory Valiant. One-sided matrix completion from two observations per row. In *International Conference on Machine Learning (ICML)*, 2023.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

I. Casanueva, Tadas Temcinas, D. Gerz, Matthew Henderson, and Ivan Vulic. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020.

Federico Cassano, John Gouwar, Daniel Nguyen, S. Nguyen, Luna Phipps-Costin, Donald Pinckney, et al. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023.

Raymond B. Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1):1–22, 1963.

Linzheng Chai, Shukai Liu, Jian Yang, Yuwei Yin, Ke Jin, Jiaheng Liu, et al. Mceval: Massively multilingual code evaluation. In *International Conference on Learning Representations*, 2025.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, S. Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, et al. Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*, 2025.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, T. Brants, P. Koehn, et al. One billion word benchmark for measuring progress in statistical language modeling. In *Interspeech 2014*, pp. 2635–2639, 2014.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and M. Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *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. 3563–3599, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021a.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, et al. Meddialog: A large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*, 2020a.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, SHIYANG LI, et al. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, 2020b.

Zhi-Yu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, et al. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021b.

Zhihong Chen, Maya Varma, Xiang Wan, C. Langlotz, and Jean-Benoit Delbrouck. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 469–484, 2023a.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, et al. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, 2024.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11378–11392, 2023b.

Aileen Cheng, Alon Jacovi, A. Globerson, Ben Golan, Charles Kwong, C. Alberti, et al. The facts leaderboard: A comprehensive benchmark for large language model factuality. *arXiv preprint arXiv:2512.10791*, 2025.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, S. Li, Sahithya Ravi, et al. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*, 2024.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, et al. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, 2018.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Christopher Clark, Kenton Lee, Ming-Wei Chang, T. Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

J. Clark, Eunsol Choi, Michael Collins, Dan Garrette, T. Kwiatkowski, Vitaly Nikolaev, et al. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, et al. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

K. Cobbe, V. Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. Cogbench: a large language model walks into a psychology lab. In *Proceedings of the International Conference on Machine Learning*, 2024.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, et al. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, 2018.

Christodoulos Constantinides, Dhaval Patel, Shuxin Lin, Claudio Guerrero, Sunil Dagajirao Patil, and Jayant Kalagnanam. Failuresensoriq: A multi-choice qa dataset for understanding sensor relationships and failure modes. In *Advances in Neural Information Processing Systems*, 2025.

M. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Yi Cui. Webapp1k: A practical code-generation benchmark for web app development. *arXiv preprint arXiv:2408.00019*, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. TURL: Table understanding through representation learning. *Proceedings of the VLDB Endowment*, 14(3):307–319, 2020.

Colin G. DeYoung, Lena C. Quilty, and Jordan B. Peterson. Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5):880, 2007.

J. Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, et al. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 862–872, 2021.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreyansh Goyal, M. Khapra, Anoop Kunchukuttan, et al. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12402–12426, 2023.

R. Dogan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Jinhao Duan, Renming Zhang, James Diffenderfer, B. Kailkhura, Lichao Sun, Elias Stengel-Eskin, et al. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*, 2024.

Yann Dubois, Bal'azs Galambosi, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*, 2024.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, et al. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6279–6299, 2022.

Mihail Eric, Lakshmi. Krishnan, F. Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 37–49, 2017.

FAHIM FAISAL, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, et al. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14412–14454, 2024.

Denis Federiakin. Improving LLM leaderboards with psychometrical methodology. *arXiv preprint arXiv:2501.17200*, 2025.

Zhiwei Fei, Xiaoyu Shen, D. Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, et al. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7933–7962, 2024.

Daniel Fein, S. Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7740–7755, 2026.

S. Fleming, A. Lozano, W. Haberkorn, Jenelle A. Jindal, E. Reis, Rahul Thapa, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, et al. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 96–120, 2021.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Elliott S. Glazer, Ege Erdil, T. Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

Bogdan Gliwa, Iwona Mochol, M. Biesek, and A. Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, 2019.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Y. Matias, et al. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *arXiv preprint arXiv:2502.21228*, 2025.

Linyuan Gong, Sida Wang, Mostafa Elhoushi, and Alvin Cheung. Evaluation of llms on syntax-aware code fill-in-the-middle tasks. In *Proceedings of the International Conference on Machine Learning*, 2024.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. In *Proceedings of the International Conference on Machine Learning*, 2024.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam S. Chilton, Aditya Narayana, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2023.

Lukas Haas, G. Yona, G. D'Antonio, S. Goldshtein, and D. Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

Patrick Haller, Jonas Golde, and Alan Akbik. Pecc: Problem extraction and coding challenges. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12690–12699, 2024.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5427–5444, 2020.

Michael Hardy, Anka Reuel, Lijin Zhang, Jodi M. Casabianca, Sang Truong, Yash Dave, Hansol Lee, Benjamin Domingue, and Sanmi Koyejo. AI cartography: Mapping the latent landscape of AI benchmark ecosystems. *arXiv preprint arXiv:2605.25272*, 2026.

Shreya Havaldar, Young Min Cho, Sunny Rai, and Lyle Ungar. Culturally-aware conversations: A framework & benchmark for llms. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pp. 220–229, 2025.

Faiz Ghifari Haznitrama, Faeyza Rishad Ardi, and Alice Oh. A neuropsychologically grounded evaluation of LLM cognitive abilities. *arXiv preprint arXiv:2603.02540*, 2026.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, et al. Measuring coding challenge competence with apps. In *Advances in Neural Information Processing Systems*, 2021a.

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)*, 2021b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, et al. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems*, 2021c.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, L. Espeholt, W. Kay, M. Suleyman, et al. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, 2015.

Søren Vejlgaard Holm, Lars Kai Hansen, and Martin Carsten Nielsen. Danoliteracy of generative large language models. *arXiv preprint arXiv:2410.22839*, 2024.

Shahin Honarvar, Mark van der Wilk, and Alastair Donaldson. Turbulence: Systematically and automatically testing instruction-tuned large language models for code. In *2025 IEEE Conference on Software Testing, Verification and Validation (ICST)*, pp. 80–91, 2025.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2): 179–185, 1965.

Zhaoyi Joey Hou, B. Zhang, Yining Lu, Bhiman Kumar Baghel, A. Brei, Ximing Lu, et al. Creativityprism: A cross-domain evaluation framework for large language model creativity. *Transactions on Machine Learning Research*, 2026.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, et al. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*, 2023.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, H. Foroosh, Dong Yu, et al. Sportsmetrics: Blending text and numerical data to understand information fusion in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 267–278, 2024.

Shuai Huang, Wen-Xuan Zhao, and Jun Gao. Si-bench: Benchmarking social intelligence of large language models in human-to-human conversations. *arXiv preprint arXiv:2510.23182*, 2025a.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shu-Jian Huang, et al. Benchmax: A comprehensive multilingual evaluation suite for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16751–16774, 2025b.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. In *International Conference on Learning Representations*, 2024.

Yao Huang, Yitong Sun, Yi-Chi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. Deceptionbench: A comprehensive benchmark for ai deception behaviors in real-world scenarios. In *Advances in Neural Information Processing Systems*, 2025c.

Dieuwke Hupkes and Nikolay Bogoychev. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *arXiv preprint arXiv:2504.10356*, 2025.

D. Ilica and Gilles E. Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, S. Radkani, Thomas Hikaru Clark, et al. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, 13:1245–1270, 2025.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, et al. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, et al. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025.

Arthur R. Jensen. Psychometric g: Definition and substantiation. In *The General Factor of Intelligence*, pp. 51–66. Psychology Press, 2002.

Saurabh Jha, Rohan R. Arora, Yuji Watanabe, Takumi Yanagawa, Yinfang Chen, J. Clark, et al. Itbench: Evaluating ai agents across diverse real-world it automation tasks. In *Proceedings of the International Conference on Machine Learning*, 2025.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71, 2025.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and G. Weikum. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 792–802, 2021.

Zhen Jia, Philipp Christmann, and G. Weikum. Faithful temporal question answering over heterogeneous sources. In *Proceedings of the ACM Web Conference 2024*, pp. 2052–2063, 2024.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, et al. Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shun-Yu Yao, Kexin Pei, Ofir Press, et al. Swebench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *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. 2567–2577, 2019.

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, et al. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. In *Advances in Neural Information Processing Systems*, 2024.

Sanghyun Jo, Soohyun Ryu, Sungyub Kim, Eunho Yang, and Kyungsu Kim. Ttd: Text-tag self-distillation enhancing image-text alignment in clip to alleviate single tag bias. In *European Conference on Computer Vision (ECCV)*, 2024.

Oliver P. John, Alois Angleitner, and Fritz Ostendorf. The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, 2(3):171–203, 1988.

Wendy Johnson, Thomas J. Bouchard, Robert F. Krueger, Matt McGue, and Irving I. Gottesman. Just one g: Consistent results from three test batteries. *Intelligence*, 32(1):95–107, 2004.

Wendy Johnson, Jan Te Nijenhuis, and Thomas J. Bouchard. Still just 1 g: Consistent results from five test batteries. *Intelligence*, 36(1):81–95, 2008.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Mohsinul Kabir, T. Ahmed, Md Mezbaur Rahman, Shaoxiong Ji, Hassan Alhuzali, Yuechen Jiang, et al. Xcr-bench: Benchmarking cross-cultural reasoning in llms via culture-specific items and hall's triad. *arXiv preprint arXiv:2601.14063*, 2026.

Kaggle. Kaggle benchmarks. `https://www.kaggle.com/benchmarks`, 2025.

Yannis Katsis, Sara Rosenthal, Kshitij P. Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, et al. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *arXiv preprint arXiv:2501.03468*, 2025.

Ryan Othniel Kearns. Quantifying construct validity in large language model evaluations. *arXiv preprint arXiv:2602.15532*, 2026.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, S. Dunn, Serina S. Applebaum, Zain Anwar, et al. Medcalc-bench: Evaluating large language models for medical calculations. In *Advances in Neural Information Processing Systems*, 2024.

Tomás Kociský, Jonathan Schwarz, P. Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, et al. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

S. Kolasani, Maxim Saplin, Nicholas Crispino, Kyle Montgomery, J. Davis, Matei A. Zaharia, et al. Llm chess: Benchmarking reasoning and instruction-following in llms through chess. *arXiv preprint arXiv:2512.01992*, 2025.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. Mawps: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1152–1157, 2016.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12359–12374, 2023.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, A. Sadallah, A. Alraeesi, et al. Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5622–5640, 2024a.

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*, 12:1703–1719, 2024b.

David C. Krakauer. The rise and fall of *g* in AGI. *arXiv preprint arXiv:2604.09911*, 2026.

John H. Kranzler and Arthur R. Jensen. Inspection time and intelligence: A meta-analysis. *Intelligence*, 13(4):329–347, 1989.

T. Kwiatkowski, J. Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and E. Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017.

Nicola Landro, I. Gallo, Riccardo La Grassa, and Edoardo Federici. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5):228, 2022.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, W. Shin, Seongjun Yang, et al. Ehrsql: A practical text-to-sql benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, 2022.

Kibeom Lee and Michael C. Ashton. Psychometric properties of the HEXACO-100. *Assessment*, 25 (5):543–556, 2018.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William-Chandra Tjhi. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*, 2023.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, 2005.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, et al. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11260–11285, 2024.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, et al. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Proceedings of the International Conference on Machine Learning*, 2025a.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, R. Leblond, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Yun-Han Li and Gengshen Wu. Legaleval-q: A new benchmark for the quality evaluation of llm-generated legal text. *arXiv preprint arXiv:2505.24826*, 2025.

Zhenyun Li, Kehai Chen, Yunfei Long, Xue-Feng Bai, Yaoyin Zhang, Xuchen Wei, et al. Xifbench: Evaluating large language models on multilingual instruction following. In *Advances in Neural Information Processing Systems*, 2025b.

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

H. Lightman, V. Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, et al. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Peerat Limkonchotiwat, Kanruethai Masuk, Surapon Nonesung, Chalermpun Mai-On, S. Nutanong, Wuttikorn Ponwitayarat, et al. Assessing thai dialect performance in llms with automatic benchmarks and human evaluation. *arXiv preprint arXiv:2504.05898*, 2025.

Bill Yuchen Lin, Yun-Tian Deng, K. Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, et al. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *International Conference on Learning Representations*, 2025.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Hao Liu, and Yujiu Yang. Criticbench: Benchmarking llms for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1552–1587, 2024.

Chao-Qun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, A. Luu, and L. Bing. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6134–6151, 2025.

Jiawei Liu, Chun Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023.

Yuchen Lu, Run Yang, Yichen Zhang, Shuguang Yu, Ziwei Wang, Jiayi Xiang, et al. Stateval: A comprehensive benchmark for large language models in statistics. *arXiv preprint arXiv:2510.09517*, 2025.

S. Luu, Kiet Van Nguyen, and N. Nguyen. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, Lecture Notes in Computer Science, pp. 415–426, 2021.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, A. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. Indonli: A natural language inference dataset for indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10511–10527, 2021.

Matt Mahoney. Large text compression benchmark. `http://mattmahoney.net/dc/textdata.html`, 2011.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jason T. Major, Wendy Johnson, and Thomas J. Bouchard. The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence*, 39(5): 418–433, 2011.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mathematical Association of America. American invitational mathematics examination (AIME). `https://maa.org/maa-invitational-competitions/`, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the International Conference on Machine Learning*, 2024.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

John Mendonça, A. Lavie, and Isabel Trancoso. Medal: A framework for benchmarking llms as multilingual open-domain dialogue evaluators. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2069–2097, 2026.

Yang Meng, Zhenya Liu, Zhuokai Zhao, and Yuxin Chen. Rethinking transfer in continual learning: A replay-based realisation. *arXiv preprint arXiv:2607.15587*, 2026.

Nicola Menghi, W. Jeffrey Johnston, Sara Viganò, Marc A. B. Hinrichs, Burkhard Maess, Stefano Fusi, and Christian F. Doeller. The effects of task similarity during representation learning in brains and neural networks. *Nature Communications*, 16(1):10812, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and R. Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 975–984, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Lester James Validad Miranda, Elyanah Aco, Conner Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. Filbench: Can llms understand and generate filipino? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2496–2529, 2025.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, et al. Lila: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5807–5832, 2022.

Filippo Momentè, Alessandro Suglia, Mario Giulianelli, Ambra Ferrari, Alexander Koller, Oliver Lemon, et al. Triangulating llm progress through benchmarks, games, and cognitive tests. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20051–20072, 2025.

J. Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, William-Chandra Tjhi, et al. Kalahi: A handcrafted, grassroots cultural llm evaluation suite for filipino. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pp. 497–523, 2024.

J. Montalan, J. Layacan, D. Africa, Richell Isaiah Flores, Michael T. Lopez II, Theresa Denise Magsajo, et al. Batayan: A filipino nlp benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31239–31273, 2025.

N. Moosavi, Andreas Ruckl'e, Dan Roth, and Iryna Gurevych. Learning to reason for text generation from scientific tables. *arXiv preprint arXiv:2104.08296*, 2021.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, and Shane Legg. Levels of AGI for operationalizing progress on the path to AGI. *arXiv preprint arXiv:2311.02462*, 2023.

N. Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, 2016.

Soichiro Murakami, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. Oogiri-master: Benchmarking humor understanding via oogiri. *arXiv preprint arXiv:2512.21494*, 2025.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine S.L. Dysart-Bricken, Bob Mankoff, Robert D. Nowak, et al. Which llms get the joke? probing non-stem reasoning abilities with humorbench. *arXiv preprint arXiv:2507.21476*, 2025.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018.

Kiet Van Nguyen, Duc-Vu Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and N. Nguyen. Uit-vsfc: Vietnamese students feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 19–24, 2018.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo@NIPS)*, 2016.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, J. Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020.

Yusuke Oda, Hiroyuki Fudaba, Graham Neubig, Hideaki Hata, S. Sakti, Tomoki Toda, et al. Learning to generate pseudo-code from source code using statistical machine translation (t). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 574–584, 2015.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. Afrobench: How good are large language models on african languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19048–19095, 2025.

Ola Krutrim. BharatBench: Evaluating language models on indic languages. `https://ai-labs.olakrutrim.com/`, 2024.

J. Omiye, J. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1):195, 2023.

OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, et al. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6137–6170, 2024.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, 2022.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, et al. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2858–2873, 2025.

B. Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 115–124, 2005.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, R. Bernardi, Sandro Pezzelle, et al. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, 2016.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, et al. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, 2015.

Arkil Patel, S. Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, 2021.

Antoine Peyronnet, Fabian Gloeckle, and Amaury Hayat. Lemmabench: A live, research-level benchmark to evaluate llm capabilities in mathematics. *arXiv preprint arXiv:2602.24173*, 2026.

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, et al. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146, 2026.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, et al. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*, 2023.

E. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and A. Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, 2020.

Nearchos Potamitis, Vansh Ramani, Hardika Arora, Dhairya Kuchhal, L. Klein, and Akhil Arora. Reasonbench: Benchmarking the (in)stability of llm reasoning. *arXiv preprint arXiv:2512.07795*, 2025.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, et al. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*, 2025.

P. Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, et al. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384, 2024.

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 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2373–2403, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, et al. Gpqa: A graduate-level google-proof q&a benchmark. In *Conference on Language Modeling*, 2024.

Steven P. Reise. The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5):667–696, 2012.

Steven P. Reise, Tyler M. Moore, and Mark G. Haviland. Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92(6):544–559, 2010.

William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*, 2024. R package.

Angelika Romanou, Negar Foroutan, A. Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, et al. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, 2025.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Alexander M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, A. S. Sunaryo, et al. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (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. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Apoorv Saxena, Soumen Chakrabarti, and Partha P. Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6663–6676, 2021.

John Schmid and John M. Leiman. The development of hierarchical factor solutions. *Psychometrika*, 22(1):53–61, 1957.

W. Joel Schneider and Kevin S. McGrew. The Cattell-Horn-Carroll theory of cognitive abilities. In Dawn P. Flanagan and Erin M. McDonough (eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, pp. 73–163. Guilford Press, 4th edition, 2018.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. AI Magazine, 29(3):93–106, 2008.

Moahmmadamin Shafiei and Hamidreza Saffari. Not all jokes land: Evaluating large language models understanding of workplace humor. *arXiv preprint arXiv:2506.01819*, 2025.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, E. Artemova, Valentin Malykh, et al. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4717–4726, 2020.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, S. Vosoughi, et al. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023.

Daniel J. Simons, Walter R. Boot, Neil Charness, Susan E. Gathercole, Christopher F. Chabris, David Z. Hambrick, and Elizabeth A. L. Stine-Morrow. Do "brain-training" programs work? *Psychological Science in the Public Interest*, 17(3):103–186, 2016.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11047–11073, 2024.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, E. Cardona, Dean Lee, et al. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025.

M. Sobhani, Md. Faiyaz Abdullah Sayeedi, Tasnim Mohiuddin, Md Mofijul Islam, and Swakkhar Shatabda. Mathmist: A parallel multilingual benchmark dataset for mathematical problem solving and reasoning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2524–2550, 2026.

R. Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Guijin Son, Hanwool Albert Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, et al. Kmmlu: Measuring massive multitask language understanding in korean. In *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. 4076–4104, 2025.

Alexandra Souly, Qing-Yuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, et al. A strongreject for empty jailbreaks. In *Advances in Neural Information Processing Systems*, 2024.

Charles Spearman. "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. doi: 10.2307/1412107.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Stanford CRFM. HELM arabic leaderboard. `https://crfm.stanford.edu/helm/arabic/latest/`, 2025.

Gabriel Stanovsky and Ido Dagan. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2300–2305, 2016.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Ronja Stern, Vishvaksenan Rasiah, Veton Matoshi, Srinanda Brugger Bose, M. Stürmer, Ilias Chalkidis, et al. One law, many languages: Benchmarking multilingual legal reasoning for judicial support. *arXiv preprint arXiv:2306.09237*, 2023.

Amber Stubbs, Michele Filannino, Ergin Soysal, Sam Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 2019.

L. H. Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, M. Okumura, and Hiroya Takamura. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1451–1465, 2021.

Yosephine Susanto, Adithya Venkatadri Hulagadri, J. Montalan, Jian Gang Ngui, Xian Yong, Weiqi Leong, et al. Sea-helm: Southeast asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12308–12336, 2025.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

William Tan and Kevin Zhu. Nusamt-7b: Machine translation for low-resource indonesian languages with large language models. *arXiv preprint arXiv:2410.07830*, 2024.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. *arXiv preprint arXiv:2503.11314*, 2025.

M-A-P Team, Xinrun Du, Yi-Fan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. In *Advances in Neural Information Processing Systems*, volume 38, pp. 125016–125089, 2025.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Min-Yang Tian, Luyu Gao, S. Zhang, Xinan Chen, Cun-Wei Fan, Xuefei Guo, et al. Scicode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems*, 2024.

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dien Dinh. ViGLUE: A Vietnamese general language understanding benchmark and analysis of Vietnamese language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4174–4189, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.261. URL https://aclanthology.org/2024.findings-naacl.261/.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, et al. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, 2017.

G. Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, M. Zschunke, M. Alvers, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015.

Han L. J. van der Maas and Kees-Jan Kan. Comment on "Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence" by Gilles E. Gignac (2016). *Intelligence*, 57:81–83, 2016.

Han L. J. van der Maas, Kees-Jan Kan, and Denny Borsboom. Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, 2(1):12–15, 2014.

Han L. J. van der Maas, Kees-Jan Kan, Maarten Marsman, and Claire E. Stevenson. Network models for cognitive development and intelligence. *Journal of Intelligence*, 5(2):16, 2017.

Jeyarajalingam Varsha, Menan Velayuthan, Sumirtha Karunakaran, Rasan Nivethiga, and Kengatharaiyer Sarveswaran. From phonemes to meaning: Evaluating large language models on tamil. *arXiv preprint arXiv:2511.12387*, 2025.

Vectara. Hallucination leaderboard. https://github.com/vectara/hallucination-leaderboard, 2023.

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 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10076–10132, 2025.

Bertie Vidgen, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A. Hale, et al. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

David Vilares and Carlos Gómez-Rodríguez. Head-qa: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 960–966, 2019.

VMLU Team. VMLU: Vietnamese multitask language understanding benchmark suite. https://vmlu.ai/, 2024.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, et al. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024.

Wen wai Yim, Yujuan Velvin Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, 2023.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9122–9129, 2020.

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 *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 2019.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, et al. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Lynn Waterhouse. Why multiple intelligences theory is a neuromyth. *Frontiers in Psychology*, 14: 1217288, 2023.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, et al. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeff Han, Is abella Fulford, et al. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Bosi Wen, Pei Ke, Xiao-Tao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, et al. Benchmarking complex instruction-following with multiple constraints composition. In *Advances in Neural Information Processing Systems*, 2024.

J. Weston, Antoine Bordes, S. Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, et al. Towards ai-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*, 2016.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, et al. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, 2025.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, 2018.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, et al. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 815–834, 2023.

Michael Wornow, Rahul Thapa, E. Steinberg, Jason Alan Fries, and N. Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. In *Advances in Neural Information Processing Systems*, 2023.

Xianjie Wu, Jian Yang, Lin-Zheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, et al. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Hao-Tian Xia, Zheng-Bang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, et al. Sportqa: A benchmark for sports understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5061–5081, 2024.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, et al. SORRY-Bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, 2025.

Shanli Xing, Yiyan Zhai, Alexander Jiang, Yixin Dong, Yong Wu, Zihao Ye, et al. Flashinfer-bench: Building the virtuous cycle for ai-driven llm systems. In *Proceedings of Machine Learning and Systems*, 2026.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Ao-Song Feng, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, 2025.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, et al. Berkeley function calling leaderboard. `https://gorilla.cs.berkeley.edu/leaderboard.html`, 2024.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. In *International Conference on Learning Representations*, 2026.

Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. Learning to mine aligned code and natural language pairs from stack overflow. In *Proceedings of the 15th International Conference on Mining Software Repositories*, pp. 476–486, 2018.

Pengfei Yu, Dong-Ming Shen, Silin Meng, Jaewon Lee, Wei Yin, Andrea Yaoyun Cui, et al. Rpgbench: Evaluating large language models as role-playing game engines. *arXiv preprint arXiv:2502.00595*, 2025a.

Shoubin Yu, Yue Zhang, Ziyang Wang, Jaehong Yoon, and Mohit Bansal. Mexa: Towards general multimodal reasoning with dynamic multi-expert aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22636–22652, 2025b.

Zhongkai Yu, Chenyang Zhou, Yichen Lin, Hejia Zhang, Haotian Ye, Junxia Cui, et al. Chipbench: A next-step benchmark for evaluating llm performance in ai-aided chip design. *arXiv preprint arXiv:2601.21448*, 2026.

Pardis Sadat Zahraei, Gokhan Tur, Dilek Hakkani-Tur, and Ehsaneddin Asgari. The alignment veto: How safety training suppresses cultural knowledge in llms. *arXiv preprint arXiv:2510.13154*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Jie Zhang, Cezara Petrui, Kristina Nikoli'c, and Florian Tramèr. Realmath: A continuous benchmark for evaluating language models on research-level mathematics. In *Advances in Neural Information Processing Systems*, 2025a.

Jinghao Zhang, Sihang Jiang, Shiwei Guo, Shi-Song Chen, Yanghua Xiao, Hongwei Feng, et al. Culturescope: A dimensional lens for probing cultural understanding in llms. *arXiv preprint arXiv:2509.16188*, 2025b.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9209–9219, 2021b.

Xiang Zhang, J. Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Jiahao Zhao, Yun-Jia Li, Wei Li, and Kazuyoshi Yoshii. Abc-eval: Benchmarking large language models on symbolic music understanding and instruction following. *arXiv preprint arXiv:2509.23350*, 2025.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yu Gu. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17696–17706, 2024.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, et al. Qtsumm: Query-focused summarization over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1157–1172, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, et al. Ar-lsat: Investigating analytical reasoning of text. *arXiv preprint arXiv:2104.06598*, 2021.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? Rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, et al. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *International Conference on Learning Representations*, 2025.

Tao Zou, Xinghua Zhang, Haiyang Yu, Minzheng Wang, Fei Huang, and Yongbin Li. Eifbench: Extremely complex instruction following benchmark for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20930–20953, 2025.

A LIMITATIONS

All scores in the corpus are as published, and we evaluate no model ourselves. This means we do not control the evaluation conditions behind any score, and we make no attempt to correct for differences in undocumented evaluation setup between sources. Where two sources disagree about the same evaluation, we average the two scores. Release dates enter only the exploratory analysis of Appendix K, and none of the factor analyses depends on them.

Metric direction is recorded but not applied, so a lower-is-better benchmark contributes a sign-flipped column. In a correlation-based analysis this shows up as a negative loading rather than as a bias, though orienting every column before analysis would be cleaner. Two downstream consequences follow. The cosine distance behind Figure 1 places a sign-flipped benchmark far from a same-direction benchmark measuring the same thing, and the mean fill assigns a positive correlation to pairs that should be negative.

Not every completion method we implement is carried through the full design. The results reported here cover the imputers in Table 3. The correlation-level estimators added most recently have not been run across every densifier and collapse strategy, and regularized EM-PCA is excluded entirely because its built-in cross-validation is intractable at this matrix size.

B The Use of Large Language Models

We used language models to explore the feasibility of and options for our methodological designs, but the end decisions are ultimately the authors'. For writing, AI helped us draft parts of the paper and adjust some of the prose, but every part of the text was checked, validated and further polished by the authors. AI also assisted in writing the code of the pipeline, under our full supervision and review, and we verified its numerical output against the underlying data. Lastly, we used locally-run language models to extract evaluation results and metadata from leaderboards, papers, repositories and model cards, since there was far too much to read by hand. Every extraction was checked by the authors (the release-date checks are reported in Appendix C.4), and we take full responsibility for all of the data, as well as every claim, number and citation in this paper.

C Data source and normalization

C.1 Data sources

The corpus is assembled from published evaluation records, and for every score we record which of four source tiers it comes from.

Tier 1, curated evaluation suites with a standardized harness and one evaluator across many models. These are Stanford HELM (the Classic, Lite, Safety, Reasoning, MedHELM, SEA-HELM, Arabic, ThaiExam, EWoK, TORR and Finance leaderboards) and the HuggingFace Open LLM Leaderboard (v1 and v2).

Tier 2, aggregators and result trackers. Papers With Code, Kaggle AI Benchmarks, llm-stats.com, Artificial Analysis, Vellum, LiveBench, Chatbot Arena / LMArena and pricepertoken.com.

Tier 3, benchmark-specific leaderboards, about 30 in total (e.g. BigCodeBench, CRUXEval, SWE-bench, BFCL, VMLU, SEA-LION, PubMedQA and AlpacaEval).

Tier 4, primary papers reporting original evaluations (arXiv, ACL Anthology, OpenReview, ACM Digital Library and journals). These are also our source for benchmark metadata.

The "Other online leaderboards" row of Table 1 combines Chatbot Arena / LMArena (202 rows), llm-stats.com (121), Vellum (96), Artificial Analysis (77), LiveBench (55) and the remaining named leaderboards (420). For 295 rows (2.2%) the source organization was not recorded, and we assigned these by source name and URL host.

C.2 Extraction routes

- **Papers With Code.** The public API is defunct (the domain redirects to HuggingFace). Evaluation tables are instead read from a daily-published archive of the evaluation tables, in four shards. Each row is one *task* with nested datasets, each carrying its own leaderboard. Extraction flattens task → dataset → leaderboard row into result rows, generating benchmark identifiers under their own namespace and applying the scope filter (Appendix C.5.2) to exclude non-LLM tasks.
- **Kaggle AI Benchmarks.** Community-maintained leaderboard datasets, extracted per dataset owner and namespaced by owner as well as by benchmark. We retain the owner namespace

precisely so that cross-source duplicates remain detectable, and several were subsequently detected (Appendix D).
- **Stanford HELM.** Extracted per sub-project into staging files, then merged. We stage before merging so that a partial or malformed extraction can be discarded without touching the canonical tables.
- **Papers and repositories.** ArXiv PDFs and abstracts converted to HTML for methodology extraction. GitHub READMEs and evaluation scripts read directly from the raw file host, under both default-branch names, and HuggingFace dataset cards through the Hub API.

C.3 Inferred fields

Two fields are filled from the record itself when the source leaves them blank.

1. **Inference platform of closed models.** A closed-weights model cannot have been run locally, so we set its inference platform to a vendor API.
2. **Temperature under the evaluation harness.** Generative tasks in EleutherAI's evaluation harness default to greedy decoding, so for benchmarks documented to use the harness we record a sampling temperature of zero.

C.4 Release dates

We record the release date of every model and benchmark to the month, together with the kind of evidence behind it. For models, the direct sources are the creation date of a HuggingFace repository named after the model (801 models), a web page that we fetched and checked to name the model and give the date (434), the model's own paper (25) and a date written in the model name (1). These date 1,261 of the 1,618 models (78%). Most of the rest come from a cited page that we could not re-read (210) or from a single language-model lookup (89), and 5 models are undated. Benchmark dates are weaker. Only one benchmark is dated from a direct source (its arXiv identifier), and most carry either a date from an earlier pass whose origin was not recorded (223) or one that two lookups agree on only to the year (136).

Every date that came with a citation was checked by fetching the cited page and asking whether it names the model and gives the date. Of the 565 such dates, 404 were confirmed and 96 could not be checked, since the page was blocked or paywalled. The other 65 were re-checked by hand against other sources, and 61 of them were resolved this way.

Two caveats remain. A repository is often created some time before the model is made public, so its creation date is a lower bound on the release. Also, dates from language-model lookups tend to be too early for models released after the training cutoff of the lookup model. Correcting the weaker tiers moved 119 model dates, 85 of them to a later month, but the dates still resting on a single lookup may run early.

C.5 Inclusion and exclusion criteria

C.5.1 Models

Included. We include general-purpose generative LLMs, domain- or task-adapted models (code, medical, legal) that still accept arbitrary prompts, and multimodal models built by adding an encoder to an LLM backbone, provided the backbone still handles arbitrary text prompts.

Excluded. We exclude encoder-only or classification-only models, narrow single-purpose systems that cannot be prompted generally (such as dedicated translation or speech systems), bare embedding or vision encoders, non-deployable research systems, evaluation metrics misfiled as models, and undocumented community uploads whose source cannot be traced.

Not a separate model. A different setup of the same model, such as a context-length variant, a reasoning mode, an effort level or a prompting scheme, is recorded on the result row instead.

Borderline cases, such as encoder-decoder models, are reviewed one by one against the same criterion, namely whether the model can follow an arbitrary prompt and generate text.

C.5.2 BENCHMARKS

We keep only benchmarks that do not require image, audio, video or speech understanding, and remove benchmarks that are left with no results once out-of-scope models are removed. The content of a benchmark is not otherwise filtered.

C.5.3 BENCHMARK TRANSLATION DUPLICATES

We remove benchmarks that are literal translations of another benchmark in the corpus, and keep those whose content is sourced separately for each language, checking each case against the benchmark's paper. The removed ones are the per-language variants of MGSM and Global MMLU Lite, a translated Arabic MMLU, the Multilingual MMLU and Global MMLU aggregates, IndicXNLI and HumanEval-XL. Translated benchmarks with no original in the corpus (the language splits of XCOPA, XNLI and XQuAD) are merged into one benchmark each. Natively multilingual benchmarks such as MultiLoKo, ArabicMMLU, FLORES, LINDSEA and the Thai national exams are kept. We also keep cross-language aggregates (MGSM, Belebele and MultiLoKo), since none of them duplicates a column we hold (MGSM shares no models with GSM8K, for instance) and they also measure transfer across languages.

C.6 DUPLICATE DETECTION AND INTEGRITY CHECKS

Two rows are duplicates if they agree on model, benchmark, metric, setup, source, model identifier and language, and redundant copies are removed after review. Rows that differ in any of these are separate evaluations and are kept, and all rows of one model-benchmark pair are averaged at aggregation (87 model-benchmark-metric triples have scores from more than one source). After every write we check that every result row points to an existing model and benchmark and that no model or benchmark is left without results, and we flag benchmarks with fewer than five rows for review.

C.7 SCORE NORMALIZATION

Scores are normalized to a 0 to 100 scale. Papers With Code and Kaggle report scores in mixed formats. For these two sources, a raw value in $[0, 1]$ is multiplied by 100 and a value above 1 is kept as it is. The other sources report on one scale per leaderboard and are converted as a whole. A few columns remain on a 0 to 1 scale, which column standardization absorbs. Results are capped at 100 to absorb floating-point noise. Exempt metrics, kept on their native scale, are perplexity, bits-per-byte, BLEURT, BERTScore, Elo, and count-type metrics ("# eval").

C.8 CANONICAL METRIC SELECTION

When a benchmark is reported under more than one metric, we keep one. Metric names are first normalized for case and whitespace, and spelling variants of the same measurement (such as written-out and abbreviated forms of accuracy or bits per byte) are merged by a hand-curated alias map. We then keep the metric that covers the most models, unless a per-benchmark override applies, and break ties by row count and then by name. This affects 92 benchmarks and drops 1,420 rows (706 model-benchmark cells), and about half of these benchmarks lose no model.

Accuracy and exact match are kept apart, even though they look like the same measure on multiple-choice tasks. On the sixteen benchmarks that carry both, exact match comes from HELM and accuracy mostly from the Open LLM Leaderboard and papers, and no model is scored both ways, so the offset between the two cannot be estimated. Merging them would put two evaluation regimes into one column. The coverage rule picks accuracy on four of these benchmarks (MMLU, TruthfulQA, HellaSwag and PubMedQA) and exact match on the other twelve.

C.9 REMAINING COLUMN DEFECTS

Two columns need fixing after metric selection. On GPQA, 447 of 454 rows are the normalized accuracy of the Open LLM Leaderboard v2 (chance mapped to zero, with negative values clamped to zero), and the other 7 are raw accuracy from papers and llm-stats.com. The two ranges do not overlap, so we drop the 7 raw rows. Since 56 of the remaining rows sit at the clamp, GPQA separates

weak models poorly. On ELEPHANT the metric field holds model configurations instead of metrics, so we remove the benchmark (9 models).

Three other benchmarks (WildBench, SEA-Exam and MultiPL-E) have sources with non-overlapping score ranges. We leave them as they are, since trackers of frontier models evaluate stronger models and a gap alone does not show a scale conflict.

D SCORE-REDUNDANCY PRUNING

This pass is applied to the text-only copy, after the scope filter of Appendix C.5.2. Each family was audited by computing the full pairwise Pearson correlation among its columns over the models evaluated on both columns. The removal decision was taken per family, on the evidence, and we verified that the resulting cascades orphan no models.

Table A1. Benchmark families audited for score redundancy, with the correlation evidence and decision for each.

Family	Correlation evidence	Decision	Rows removed
LiveCodeBench release windows v1–v6 (Kaggle)	mean pairwise $r = 0.995$, worst pair 0.987, over 45 shared models	Keep the aggregate, drop 6 per-version identifiers	270
TwitterAAE dialect splits (African-American and White English)	$r = 0.993$–0.999 with each other and the parent, over 32 shared models	Keep the parent, drop both dialect splits	64
GPQA variants (few/zero-shot $\times$ diamond/main, Kaggle)	mean $r = 0.944$, worst pair 0.915, over 46–47 shared models, with the better-populated canonical GPQA and GPQA Diamond already present	Drop all 4 Kaggle variants	185
MultiLoKo per-language splits (31 languages, Kaggle)	mean pairwise $r = 0.82$, near-duplicate for well-resourced pairs (Simplified/Traditional Mandarin 0.989, Italian/Swedish 0.983) and noisy for low-resource pairs on small overlap	Keep the paper-sourced MultiLoKo across-language aggregate, drop all 31 per-language identifiers	1,523
A Kaggle re-import of MMLU	Cross-source re-import (44 rows) of the canonical MMLU column (468 rows)	Drop the re-import	44
A Kaggle re-import of SciCode	Not a third metric. Per-model value matching shows it splices main-problem scores for 30 of 46 models and sub-problem scores for the other 13, which is a scraping artifact	Drop as a data-integrity fix. The 4 explicit split variants ($r = 0.71$–0.96) are **kept**, as their correlations are not uniform enough to treat as duplicates	46

Table A1. (continued)

Family	Correlation evidence	Decision	Rows removed
Stanford HELM ThaiExam sub-splits	Two clusters, not uniform redundancy: {ONET, IC, A-Level} at $r = 0.92$–0.95, while {TGAT, TPAT1} correlate weakly with that cluster ($r = 0.70$–0.88). The TGAT/A-Level gap was verified as systematic, not noise (several multilingual models score 35–45 points higher on TGAT)	Drop ONET and IC, and **keep** A-Level as the knowledge-cluster representative and keep TGAT and TPAT1, which carry distinct variance	84
		Total	**2,216**

Two of the seven audited families were thus deliberately left partially or fully intact, which is the point of auditing by correlation rather than by name. This pass leaves 456 benchmarks. The canonical-metric filter (Appendix C.8), the source-scale fix and the single-row anomaly removal then drop a further 1,463 result rows, for a final corpus of **456 benchmarks, 1,618 models, 13,251 result rows**.

E MODEL-IDENTITY COLLAPSE

Both strategies operate on the source-specific model identifier, which each result row carries exactly as its source spelled it, alongside the canonical model it resolves to. Keeping both is what makes cross-source duplicate detection possible after canonicalization. Each identifier's family and parameter-count metadata is first canonicalized to the first non-null value observed for it. Multiple evaluations of the same (identifier, benchmark) are averaged before collapsing, and rows sharing a collapse key are averaged again per benchmark.

E.1 STANDARD (VARIANT-LEVEL)

We apply the following steps in order to each identifier.

1. Strip the organization prefix, meaning everything before the first slash.
2. Reduce parenthesized content to a parameter count where one is present, and drop it otherwise.
3. Strip trailing separators and any text after them, reasoning-effort phrases such as "medium effort" or "high reasoning", every date format we encountered, and any bare four-digit number, which in this corpus is a release year, a checkpoint stamp, or a context length.
4. Normalize version numbers written with hyphens so that they read as decimals.
5. Where a model family is recorded and is not itself numeric, take it as the stem and tokenize the remaining suffix. Otherwise tokenize the whole cleaned identifier.
6. Then, token by token, drop generic training and serving tokens (instruction tuning, chat, base, the preference-optimization family, thinking and reasoning markers, greedy decoding, and API markers), language and region tokens, legacy engine names, and month names. Preserve named tier tokens, since Opus, Sonnet, Haiku, Pro, Mini, Flash, Maverick and Scout each denote a distinct released model rather than a serving option. Drop context-length tokens. Recognize a parameter count either from a billions suffix or by matching the identifier's own recorded size, falling back to a whitelist of common sizes. Merge a version number into the family stem when one extends the other.
7. Emit the family stem, the preserved tokens, and the parameter count.

The common-size whitelist is guarded against version-number collisions, since a bare 3 or 4 in a Claude or GPT identifier is a version rather than a parameter count.

E.2 AGGRESSIVE (FAMILY-LEVEL)

Take the first alphabetic token of the recorded model family where it is not numeric. Otherwise strip the organization prefix, the parentheses and the dates from the identifier, and take its first alphabetic token. Everything else is discarded.

E.3 POST-COLLAPSE FILTERING

Benchmarks observed for only one collapse key are dropped, then collapse keys with no remaining benchmarks are dropped. This yields the two raw matrices of Table 2, 1,266 × 404 at 2.2% and 334 × 380 at 3.5%, from 2,183 distinct source-level model identifiers.

F DENSIFICATION ALGORITHM

Target density $\tau = 0.10$. Let $\mathbf{M}$ be the boolean observation mask.

C (column-primary peel). While density $< \tau$: drop the benchmark (column) with the fewest observations across currently-kept rows, then drop any model left with zero observations.

R (row-primary peel). While density $< \tau$: drop the model with the fewest observations among currently-kept columns, then drop any benchmark left with zero observations.

S (symmetric peel). While density $< \tau$: compute each kept column's and each kept row's *fill rate* (observations ÷ current opposite-axis size) and drop whichever single marginal has the lowest rate, then clear emptied rows and columns on both axes.

Minimum-observation floor. After peeling, iterate to a fixed point, dropping any kept row or column with fewer observations than the floor *within the currently kept submatrix*. The loop is required because dropping a sparse row can starve a column and vice versa.

Degenerate-column guard. Finally drop any column with fewer than 2 observed values or zero variance among its observed values, matching exactly what the downstream estimators would drop at runtime. This parity is deliberate: it keeps the reported matrix shape equal to the shape actually factored.

The tables analyzed in this paper were generated with the floor set to 3 observations.

G IMPUTATION METHODS

Every method implemented in R shares one interface. It takes the sparse matrix in, and returns the completed matrix, the swept-parameter grid, and the held-out RMSE and R^2 at each parameter value. None of them factor. This is what allows the factoring stage to be identical across methods. Table A2 shows the list of full-dataset missing data estimators, while Table A3 shows the list for the correlation-level imputers.

Table A2. Estimators of the missing dataset entries.

Method	Description	Package	Configuration
SoftImpute (Mazumder et al., 2010)	Nuclear-norm-penalized low-rank completion by iterative soft-thresholded SVD, assuming a low-rank signal plus noise. Primary cell-level method.	softImpute	sweeps rank: $1\dots10$ (capped at $\min(n,p)-1$), and at each rank a 30-point geometric λ grid from λ_0 down to $\lambda_0/100$, ALS with warm starts
k-NN	Each missing cell filled from the k most similar models, an assumption-light baseline with no low-rank, linearity, or normality assumption.	VIM	sweeps k: $1\dots10$ (capped below n), Gower distance over benchmarks, weighted-mean aggregation
missForest (Stekhoven & Bühlmann, 2012)	Iterative random-forest imputation, nonparametric, able to capture nonlinear dependence the low-rank methods cannot represent.	missForest	sweeps number of trees: {50, 100, 200, 400}, at most 10 iterations

Table A3. Estimators of the missing correlation entries.

Estimator	Description	Package	Configuration
One-sided matrix completion (Cao et al., 2023)	Recovers the right-singular vectors of a reduced matrix of a large, super-sparse dataset. Originally demonstrated to work with simply 2 observations per row.	Custom Julia code	Sweeps rank $1\dots10$, selecting the rank with the best R^2.
SoftImpute (Mazumder et al., 2010)	Applies SoftImpute's low-rank completion to the observed pairwise correlation matrix rather than the data matrix, whose missing entries are exactly the benchmark pairs never co-observed.	softImpute	sweeps rank $1\dots10$ with the same nested λ grid as the cell-level methods above
USVT (Chatterjee, 2015)	Universal singular value thresholding: completes the correlation matrix by hard-thresholding its singular values.	filling	fixed singular-value threshold $\eta = 0.01$, no sweep

G.1 One-sided matrix completion

We use a custom implementation of OSMC (Cao et al., 2023) written in Julia. The premise is that when observations are too sparse to complete cells, the right singular vectors (the benchmark-space factors) may still be recoverable. The estimator targets $\Theta = \frac{1}{n} Z^\top Z$ over the n models. Each product $z_{ij} z_{ij'}$ of two observed standardized scores in one row is an estimate of $\Theta_{jj'}$, and we fit $\hat{\Theta} = \hat{V}\hat{V}^\top$, with $\hat{V} \in \mathbb{R}^{p \times r}$, to all such products by squared loss using Adam. Off-diagonal and diagonal terms are each averaged over their total count across rows.

Its native error is defined on pairwise products, which is not comparable to the other methods, so each held-out cell is additionally predicted from the recovered covariance by the conditional-Gaussian (best linear) predictor $\hat{z}_j = V_j^\top V_S^+ z_S$, solved in the r-dimensional factor space rather than by inverting

the rank-deficient $|S| \times |S|$ covariance block, which is numerically unstable on richly-observed rows. For the hold-out column stratification, the masking also ensures each row has at least 2 observations.

H IMPUTATION EVALUATION

H.1 METRICS

Within each benchmark j, sample $n_{\text{hold}} = \max\left(1,\ \min(\lfloor 0.2\, n_{\text{obs}(j)} \rfloor,\ n_{\text{obs}(j)} - 2)\right)$ observed cells for $n_{\text{obs}(j)} > 2$, so at least 2 training observations remain in every column. Every column with more than 2 observations contributes at least one held-out cell, so no benchmark is unrepresented in the evaluation set.

At each stage of the imputation, we mask ~20% of the observed cells as an evaluation set. To prevent high-observation benchmarks from inflating the score, we use a column-stratified mask, such that each benchmark is masked at least once and keeps at least two training observations in every column. Columns are standardized using training-cell moments only.

Let MSE_j and base_j be the mean squared error and mean baseline within column j. Then

$$\text{RMSE} = \frac{1}{p}\sum_j \sqrt{\text{MSE}_j}, \qquad R^2 = 1 - \frac{\frac{1}{p}\sum_j \text{MSE}_j}{\frac{1}{p}\sum_j \text{base}_j}$$

To filter out likely untrustworthy imputations, we skip running factor analysis on results where the held-out $R^2 < 0.2$. R^2 is also used for hyperparameter selection within each method (rank, number of trees, etc.).

H.2 CORRELATION-LEVEL IMPUTERS

Since correlation-level imputers cannot recover individual cells, we held out a fraction of observed cells, then predict each held-out cell from the row's remaining observed cells using the conditional-Gaussian (best linear) predictor $\hat{z}_j = \Theta_{jS}\,\Theta_{SS}^{+}\,z_S$, where S indexes the surviving cells of that row and Θ is the imputed (or fitted) covariance structure. The best guess for a missing cell is a weighted combination of the row's other cells, with weights fixed by the recovered correlations. The correlation-level imputers are listed in Table A3 above.

I FACTOR ANALYSIS DETAILS

I.1 ESTIMATOR

We fit the exploratory factor analysis with the minimum-residual estimator, applying a promax rotation whenever more than one factor is extracted and no rotation otherwise. Where the default squared-multiple-correlation start for the communalities errors on a singular correlation matrix, we retry the fit from a unity diagonal. Only hard errors trigger that fallback, since benign warnings still return a usable fit. The factor analysis used the R package `psych` (Revelle, 2024).

I.2 FACTOR COUNT

We use Horn's parallel analysis (Horn, 1965) in its PC flavor. The observed eigenvalues of the correlation matrix are compared position-by-position against the 95th percentile of eigenvalues from 100 random $n \times p$ standard-normal matrices' correlation matrices, and

$$n_f = \#\{i : \lambda_i^{\text{obs}} > \lambda_i^{\text{cut}}\}, \quad n_f \geq 2.$$

The count is then capped at 20, beyond which the bifactor fits become prohibitively slow, and at $\min(p-1,\ n-1,\ \text{rank}(R)-1)$, since the completed and surrogate matrices are frequently rank-deficient or have $p \gg n$ and the estimator would otherwise error. If the fit still fails, we decrement the count until it succeeds, and record the count actually used.

I.3 BIFACTOR DECOMPOSITION

The bifactor step uses the Schmid-Leiman (Schmid & Leiman, 1957; Reise et al., 2010; Reise, 2012) transformation for EFA without sign flipping. We record per cell the full Schmid-Leiman loading matrix (the general factor plus the domain factors, per benchmark) and ω_h. From the first-order solution we also record cumulative variance explained, per-factor proportions, and the inter-factor correlation matrix Φ with its mean off-diagonal.

J BENCHMARK EMBEDDING

Figure 1 places each benchmark in two dimensions using a composite distance computed from the factor loadings rather than from the score matrix. This section describes how that distance is built and how the figure is colored.

J.1 COMPOSITE DISTANCING

Every bifactor solution assigns each benchmark a row of loadings, consisting of its loading on the general factor followed by its loadings on each specific factor. We take that row as the benchmark's vector within that solution. A benchmark absent from a given solution does not contribute to it.

Within a single solution we compute the cosine distance between every pair of benchmark vectors, which is one minus their cosine similarity, clipped to the range 0 to 2. Cosine distance is invariant to the rotation and to the sign of the factors, so per-solution distances remain comparable even though the factors themselves are not. We then average each pair's distance across every solution in which both benchmarks appear. This average is the composite distance, and it is what the figure embeds.

The composite distance matrix is passed to UMAP (McInnes et al., 2018) as a precomputed metric, with two components, ten neighbors, a minimum distance of 0.15, and a fixed random seed. Nothing is re-standardized at this stage, since the distances already share a common scale.

J.2 UNCOMPUTABLE DISTANCE HANDLING

Two benchmarks that never appear together in any solution have no measured distance between them. We fill each such entry with the mean of the distances that one of the two benchmarks does have, falling back to the global mean when neither has any. Luckily, missing entries are only observed in the R dataset (which also has only one valid imputation, SoftImpute), with a very small missing proportion. The R dataset has 318 benchmarks, and 160 out of 50,403 distances (0.32%) were uncomputable. The statistics are identical for each release year from 2020 to 2025 and for the aggregates. Uncomputable distances were not observed in any other dataset, imputation, or release-year combination.

J.3 LABELS

All benchmarks were labeled along three independent, non-exclusive axes: subject (content domain), task (administration format), and language. A benchmark can carry several subject tags, and some of those tags are nested within a broader parent label (e.g. medical under specialized_domain). Labels were curated from each benchmark's documentation, independent of the distance-geometry and cohesion analyses.

J.4 UMAP LIMITATIONS

UMAP preserves neither density nor global distance. Groups that appear tight or far apart in two dimensions are partly an artifact of the embedding. We therefore read the figure as a visual summary, and any claim we make about clustering rests on the composite distance matrix itself.

K RELEASE-DATE ANALYSIS

Here we check whether the factor structure changes with model release date, and whether a benchmark's release date relates to its place in that structure. Both analyses are exploratory and use the matrices of the main analysis at the 10% target density.

K.1 SETUP

We group models by release year into four groups (2022 or earlier, 2023, 2024, and 2025 or later), since the earlier years hold too few models to stand alone. Release dates are known for 99.8% to 99.9% of the models in each matrix. We take the completed C and S matrices (standard collapse) of the three imputers whose output rows are the models themselves (SoftImpute, missForest and k-NN), split their rows by year group, and run each group through the same parallel analysis, EFA and Schmid-Leiman pipeline as the pooled data (Appendix I). The correlation-level estimators are left out, since their completed matrix is a synthesized surrogate whose rows are not the models (Appendix H.2).

A smaller sample moves ω_h on its own. For each year group we therefore also factor 50 random subsets of the same size, drawn from all dated models regardless of year, and only count a group as different when its ω_h falls outside the 5th to 95th percentile of its subsets. We also compare the general factor of each fit with the pooled one by Tucker's congruence.

It must be stressed, however, that each matrix is imputed once over all models before the split. The missing cells of one year group are filled partly from the other groups, and a benchmark that no model in the group took is filled entirely from them. This happens often (Table A4), up to 42 of the 124 S benchmarks in the oldest group. Imputing each group on its own does not work either, since a benchmark with no observation in a group cannot be imputed from inside it. If the correlations between benchmarks are the same in every year and only the level of performance shifts, the shared imputation does no harm. If they are not, it pulls every group toward the pooled structure, and the differences we report below are smaller than the true ones.

Table A4. Size and coverage of each release-year group. Each cell gives the number of models, the share of observed cells, and the number of benchmarks with no observed score in that group.

Densifier	Benchmarks	≤ 2022	2023	2024	≥ 2025
C	78	87/20%/10	196/14%/6	294/12%/21	93/12%/10
S	124	86/14%/42	196/9%/30	292/9%/26	94/10%/16

K.2 MODELS BY RELEASE YEAR

Figure A1 plots the ω_h of each year group against the band of its random subsets. There is no trend with release year. The imputers do not even agree on which year has the strongest general factor. Still, the year groups differ from random subsets more often than chance allows. Of the 24 fits, 8 fall outside the 90% band (6 above and 2 below) where 2 or 3 are expected, and the median congruence of their general factor with the pooled one is 0.82, against 0.97 for the random subsets. In other words, the structure does shift from one year to the next, but in no consistent direction.

Newer models do score higher on the general factor (Table A5). We score each model as the loading-weighted sum of its standardized completed scores on the pooled general factor, so the mean over all models is zero. In four of the six solutions the mean score goes from between -1.5 and -1.0 for models of 2022 or earlier to between 0.0 and 1.2 for models of 2025 or later. In the other two (missForest and k-NN on S) the newest models still score highest. So newer models are better on average, but their general factor does not explain more of the variance. We find no sign that later model generations are converging on a more general intelligence.

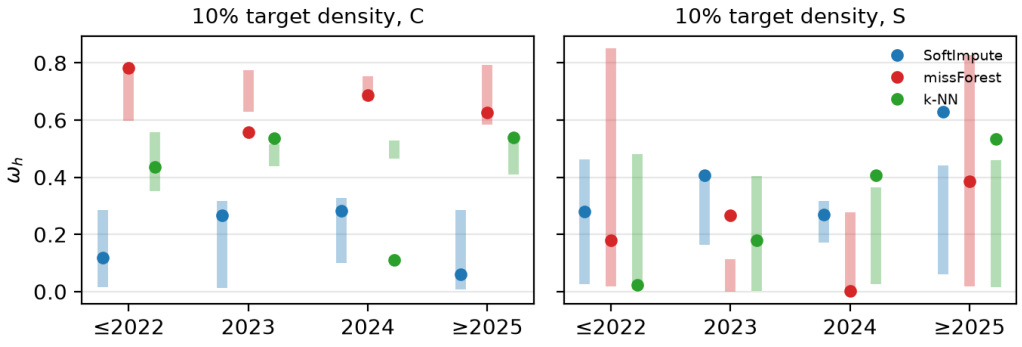

Figure A1. ω_h of each release-year group (dots) against the 5th to 95th percentile of 50 random subsets of the same size (bars), per imputer, on the C (left) and S (right) matrices.

Table A5. Mean general-factor score of each release-year group, with mean zero over all models.

Densifier	Imputer	≤ 2022	2023	2024	≥ 2025
C	SoftImpute	-1.33	-0.29	0.59	0.00
C	missForest	-1.02	-0.40	0.22	1.11
C	k-NN	-1.36	-0.34	0.25	1.22
S	SoftImpute	-1.47	-0.21	0.37	0.67
S	missForest	0.43	-0.32	-0.19	0.86
S	k-NN	-0.52	0.06	-0.27	1.20

K.3 Without imputation

Only six benchmarks are observed in every year group (BBH, GPQA, IFEval, MATH, MMLU-Pro and MuSR, all from the Open LLM Leaderboard v2). For the 330 models that have all six, we compute the correlation matrix of their observed scores within each year group and take the share of variance of its first eigenvalue. This share is 0.43 for 2022 or earlier, 0.71 for 2023, 0.72 for 2024 and 0.57 for 2025 or later. The oldest models sit near the floor of most of these benchmarks (a mean MATH score of 1 out of 100, for instance), which lowers their correlations, and the newest group holds only 20 models, with a bootstrap interval ([0.50, 0.69]) that overlaps that of 2023 ([0.66, 0.81]). We read no trend into this check either.

K.4 Benchmark release date

We also check whether the release year of a benchmark relates to its place in the pooled solutions. Benchmark dates are less reliable than model dates (Appendix C.4), so we only use the year. Since only the pooled loadings are needed here, all five imputers enter (Table A6).

First off, release year does not predict the general-factor loading. The Spearman correlation between year and absolute loading is significant in only 3 of 10 solutions, twice positive and once negative.

Benchmarks from the same era do sit closer together in loading space. We group benchmarks into four eras (2019 or earlier, 2020 to 2021, 2022 to 2023, and 2024 or later), and the mean within-era Euclidean distance between their loading vectors is smaller than under 2,000 permutations of the era labels in 9 of 10 solutions ($z < -1.96$).

Most of this closeness, however, comes from which models took each benchmark. Newer benchmarks are mostly taken by newer models (the Spearman correlation between a benchmark's year and the mean year of its models is 0.67 on S and 0.68 on C), and two benchmarks taken by the same models are filled in alike by the imputers. We regress the rank of each pair's loading distance on the ranks of its year gap and of the Jaccard distance between the two sets of observed models. The co-observation

term is the larger one in all 10 solutions (0.20 to 0.54, against 0.01 to 0.20 for the year gap), although the year gap keeps a significant effect of its own in 6 of 10 (permutation test with 500 relabelings).

The observed scores alone show a small era effect on S. Among benchmark pairs observed together on at least 30 models, pairs released four or more years apart correlate at 0.36, against 0.50 for pairs released within a year of each other. On C there is no gap (0.51 against 0.51). Succinctly, most of the era effect comes from the missingness pattern, and the small part left matches the year-to-year shift we find for the models.

Table A6. Benchmark release year against the pooled solutions. ρ is the Spearman correlation between release year and absolute general-factor loading. Era z compares within-era loading distance with permuted eras (negative means same-era benchmarks sit closer). The last two columns are standardized rank-regression coefficients of a pair's loading distance on its release-year gap (with permutation p) and on the Jaccard distance between its sets of observed models (co-obs.).

Densifier	Imputer	ρ	Era z	Year gap (p)	Co-obs.
C	SoftImpute	-0.16	-6.1	0.07 (0.022)	0.39
C	missForest	0.21	-4.2	0.10 (0.054)	0.31
C	k-NN	-0.34	-8.2	0.06 (0.082)	0.43
C	OneSidedMC	0.10	-12.6	0.15 (0.000)	0.54
C	SoftImpute (corr.)	-0.06	-6.1	0.11 (0.018)	0.25
S	SoftImpute	-0.03	-6.9	0.09 (0.004)	0.20
S	missForest	0.22	-14.2	0.20 (0.000)	0.47
S	k-NN	0.26	-6.0	0.01 (0.314)	0.24
S	OneSidedMC	0.13	-9.9	0.10 (0.000)	0.45
S	SoftImpute (corr.)	-0.02	-1.7	0.04 (0.158)	0.20

L BENCHMARK g-RANKINGS

L.1 FREQUENCY AND g-RANKING CORRELATIONS

A pressing concern with regard to our findings on the top g benchmarks in Table 5 of the main content is to what degree a benchmark's g factor loading is correlated with its non-missing frequency. This is important to know, as our datasets can possess higher or lower correlations introduced as artifacts by our imputation methods.

Naively, as shown in Table A7 below, benchmark frequency *is* correlated with g loadings. Unsigned averages suggest this correlation is modest ($r = 0.3432$), but the correlations are quite dispersed, and some dataset-imputation combinations can be as high as ~0.5. Thus, identifying benchmarks which proxy a supposed latent g factor well requires adjusting the statistics with respect to their frequency.

Table A7. Correlations between benchmark frequency and their g loadings. k = number of factors extracted.

Method	Dataset	k	r	N
Mean fill	C Std.	2	+0.4192	78
Mean fill	C Std.	14	+0.4192	78
k-NN	C Std.	2	+0.4065	78
k-NN	C Std.	7	+0.4065	78
k-NN	S Std.	2	+0.2335	124
k-NN	S Std.	11	+0.2335	124
missForest	C Aggr.	2	+0.3251	102
missForest	C Aggr.	4	+0.3251	102
missForest	C Std.	2	+0.2203	78
missForest	C Std.	4	+0.2203	78
missForest	S Std.	2	-0.2462	124
missForest	S Std.	4	-0.2462	124

Table A7. (continued)

Method	Dataset	k	r	N
OneSidedMC	C Aggr.	2	-0.4090	102
OneSidedMC	C Std.	2	+0.4786	78
OneSidedMC	S Std.	2	+0.3609	124
SoftImpute	C Aggr.	2	+0.5763	102
SoftImpute	C Aggr.	5	+0.5763	102
SoftImpute	C Std.	2	+0.5513	78
SoftImpute	C Std.	9	+0.5513	78
SoftImpute	R Aggr.	2	-0.3041	310
SoftImpute	R Aggr.	20	-0.3041	310
SoftImpute	R Std.	2	-0.2140	298
SoftImpute	R Std.	20	-0.2140	298
SoftImpute	S Aggr.	2	-0.3161	293
SoftImpute	S Aggr.	20	-0.3161	293
SoftImpute	S Std.	2	+0.1885	124
SoftImpute	S Std.	5	+0.1885	124
SoftImpute	raw Aggr.	2	-0.2404	380
SoftImpute	raw Aggr.	10	-0.2404	380
SoftImpute	raw Std.	2	+0.1815	404
SoftImpute	raw Std.	10	+0.1815	404
SoftImpute (corr.)	C Std.	2	+0.3788	78
SoftImpute (corr.)	C Std.	4	+0.3788	78
SoftImpute (corr.)	S Std.	2	+0.3890	124
SoftImpute (corr.)	S Std.	5	+0.3890	124
Zero fill	C Std.	2	+0.5343	78
Zero fill	C Std.	14	+0.5343	78
Average			**+0.1783**	37 (groups)
Average $\lvert r\rvert$			**0.3432**	

L.2 Frequency adjustment

Let C denote a factor cell: a method and dataset combination for which factor loadings are available. Within each cell C, benchmarks are ranked by absolute loading $|g_i|$ (i indexes benchmarks). The rank r_i is normalized within the cell,

$$a_i = \frac{r_i - 1}{n - 1}, \qquad r_i \in \{1, \dots, n\},$$

so $a_i = 0$ at the top of the cell and $a_i = 1$ at the bottom. This makes cells with different numbers of benchmarks n comparable. Each benchmark's raw score is the mean of its a_i over the cells in which it appears. Benchmarks appearing in fewer than two cells are excluded.

For a dataset with m models, the frequency of benchmark i is the proportion of models with an observed score,

$$f_i = \frac{\#\{\text{models with non-missing score for } i\}}{m}.$$

Columns with fewer than two observations or zero variance are excluded before computing f_i. Where several datasets contribute, the benchmark's frequency is the mean of its per-dataset f_i.

As shown in Table A8 below, within most cells, frequently measured benchmarks obtain higher $|g_i|$ and lower (better) a_i. A benchmark may therefore rank highly partly because it is measured often rather than because it is a strong indicator of g.

The adjustment is applied within each cell rather than to the pooled averages. Within cell C, the normalized ranks are regressed on the cell's own frequencies by ordinary least squares,

$$a_i = \alpha + \beta f_i + \varepsilon_i, \qquad \hat{\varepsilon}_i = (a_i - \bar{a}) - \hat{\beta}\,(f_i - \bar{f}),$$

and the residual $\hat{\varepsilon}_i$ replaces a_i. By construction the residuals are uncorrelated with f_i within the cell, so $\hat{\beta}$ removes exactly the linear component of the within-cell rank–frequency association. A cell with fewer than three frequency–rank pairs, or with zero frequency variance, cannot be fitted. Its benchmarks receive centered raw ranks $(a_i - \bar{a})$, and such cells are counted and reported.

The adjusted score of benchmark i is the mean of its residuals over the cells in which it appears. Rankings are reported on this scale, with the raw average rank retained for comparison.

L.3 Diagnostics

We present two diagnostics that justify the method of adjustment. Table A8 below reports, within each cell, the Pearson correlation $r(f_i, a_i)$, with the mean r, the mean $|r|$, and the counts of negative and positive cells. Negative r indicates that more frequently measured benchmarks load higher on g.

Table A8. Correlations between benchmark frequency and their g rankings. Mean $r = -0.173$, mean $|r| = 0.273$; 13 negative and 7 positive cells.

Method	Dataset	N	r
Mean fill	C Std.	78	-0.429
k-NN	C Std.	78	-0.429
k-NN	S Std.	124	+0.020
missForest	C Aggr.	102	-0.342
missForest	C Std.	78	-0.106
missForest	S Std.	124	+0.261
OneSidedMC	C Aggr.	102	+0.383
OneSidedMC	C Std.	78	-0.242
OneSidedMC	S Std.	124	-0.271
SoftImpute	C Aggr.	102	-0.470
SoftImpute	C Std.	78	-0.497
SoftImpute	R Aggr.	310	+0.006
SoftImpute	R Std.	298	+0.222
SoftImpute	S Aggr.	293	+0.094
SoftImpute	S Std.	124	-0.162
SoftImpute	raw Aggr.	380	+0.012
SoftImpute	raw Std.	404	-0.091
SoftImpute (corr.)	C Std.	78	-0.425
SoftImpute (corr.)	S Std.	124	-0.431
Zero fill	C Std.	78	-0.572

Second, Table A9 reports the pooled correlation between mean frequency and mean rank before and after the adjustment, stratified by the number of cells k in which a benchmark appears. The within-cell association does not survive pooling with a consistent sign: cells disagree in direction, so the pooled raw correlation is small even though the within-cell correlations are not.

Table A9. Pooled correlation between mean frequency and mean rank.

N cells	N benchmarks	r raw	r adjusted
2	54	-0.023	-0.021
3	26	-0.024	-0.049
4	24	-0.214	-0.232
5	141	+0.061	+0.046
8	10	+0.040	+0.038
10	33	+0.518	+0.516

Table A9. (continued)

N cells	N benchmarks	r raw	r adjusted
13	13	-0.538	-0.539
20	78	-0.382	+0.073
Pooled	380	-0.128	-0.039

For this reason, we report the cellwise-frequency-adjusted average normalized rank as the primary ordering of benchmarks (Table 5 in the main text). This ordering removes the linear component of the within-cell rank–frequency association, which reaches $|r| \approx 0.3$ in the average cell, before averaging across cells. The raw average normalized rank is retained alongside it for comparison, as the two orderings agree closely. The adjustment mainly matters for benchmarks measured across many cells, where the within-cell association is strongest and where the adjusted ranking reorders the raw ranking.

The full ranking of all 380 benchmarks is given in Appendix P.

M LABEL COHESION RESULTS

We present the full label cohesion results for all labels here. From the subject category, only `finance` and `professional_writing` show significantly strong cohesion, both of which are very specific labels with low median n. The next few labels show moderate to weak cohesion at best, a lot of which do not have enough significance to distinguish from chance. Labeling by language surprisingly shows many weakly cohesive structures, especially for `monolingual_non_english`, which is an aggregate label. Individual labels for each language are not used due to low n. All of the results support our claim that benchmark clusters are weakly cohesive.

Table A10. Cohesion by task-format label, sorted by descending median A.

Label	Median A	Significant	Median n
`conversation`	+0.283	3/6	7
`likelihood_probe`	+0.153	6/10	19
`short_qa`	+0.098	2/18	15
`classification`	+0.095	2/18	15
`interactive`	+0.078	1/10	16
`long_reasoning`	+0.066	0/18	9
`multiple_choice`	+0.059	8/18	37
`long_form_generation`	+0.023	4/18	32
`extraction`	-0.000	0/14	5
`ranking`	-0.072	0/2	10
`sentence_completion`	-0.101	0/18	6

Table A11. Cohesion by language-coverage label, sorted by descending median A.

Label	Median A	Significant	Median n
`monolingual_non_english`	+0.291	16/18	19
`multilingual`	+0.204	6/14	9
`crosslingual`	+0.137	1/10	12
`english`	-0.034	0/18	85

Table A12. Cohesion by subject-matter label, sorted by descending median A.

Label	Median A	Significant	Median n
finance	+0.870	10/10	4
professional_writing	+0.632	4/6	5
fact_verification	+0.300	0/4	5
commonsense	+0.287	5/18	11
code	+0.264	6/18	11
translation	+0.249	6/10	16
language_understanding	+0.247	4/6	10
linguistic_competence	+0.226	0/10	6
social_media	+0.226	2/10	6
sentiment	+0.217	2/10	6
dialogue	+0.203	0/6	5
industrial	+0.186	0/2	4
cognitive	+0.178	0/6	12
society_culture	+0.172	2/6	14
agentic	+0.171	0/10	10
recall	+0.169	13/18	21
factuality	+0.154	1/14	7
structured_data	+0.111	0/6	15
language_modelling	+0.106	0/6	5
encyclopedic	+0.104	1/18	23
world_knowledge	+0.092	0/18	7
generation	+0.091	1/10	37
legal	+0.069	0/4	4
multi_subject	+0.068	6/18	28
logical_reasoning	+0.068	2/18	6
medical	+0.066	1/6	40
instruction_following	+0.044	0/10	16
creativity	+0.042	0/4	4
safety	+0.042	0/18	8
math	+0.031	0/18	10
science	+0.029	0/18	11
specialized_domain	+0.029	1/18	13
games	+0.022	0/6	6
natural_language_inference	+0.018	0/6	9
source_genre	+0.007	0/18	39
text_classification	-0.007	0/6	12
reasoning	-0.008	0/18	19
summarization	-0.008	0/6	13
reading_comprehension	-0.013	0/18	9
retrieval	-0.014	0/6	6
language_processing	-0.016	0/18	26
alignment	-0.016	0/18	26
temporal_reasoning	-0.024	0/2	7
news	-0.042	0/18	6
toxicity	-0.056	1/14	6
bias_fairness	-0.069	0/6	4
information_extraction	-0.120	0/2	4
fiction	-0.232	0/6	4

N IMPUTATION RESULTS

The table below lists, for every dataset-imputer combination, the held-out RMSE, R^2, and the selected configuration. Rows are sorted by R^2. As described in the results, only 20 of the combinations pass

the $R^2 \geq 0.2$ gate. Of all the methods presented, only USVT yielded no valid solutions or crashes mid-estimation.

Table A13. Results of all imputation runs, sorted by R^2.

Dataset	Imputer	RMSE	R^2	Configuration
S Std.	SoftImpute	0.6338	0.504	rank=5 (swept 1..10)
C Std.	SoftImpute	0.6575	0.493	rank=9 (swept 1..10)
S Std.	missForest	0.6798	0.471	ntree=400 (swept [50,100,200,400])
C Std.	missForest	0.7321	0.399	ntree=50 (swept [50,100,200,400])
S Std.	SoftImpute (corr.)	0.7637	0.378	rank=6 (swept 1..10)
S Std.	OneSidedMC	0.7554	0.365	r=2 (swept 1..10)
C Aggr.	SoftImpute	0.6739	0.337	rank=5 (swept 1..10)
C Std.	OneSidedMC	0.7261	0.321	r=2 (swept 1..10)
C Std.	SoftImpute (corr.)	0.8126	0.317	rank=5 (swept 1..7)
S Std.	k-NN	0.8104	0.296	k=5 (swept 1..10)
R Std.	SoftImpute	0.7245	0.290	rank=4 (swept 1..10)
C Std.	k-NN	0.8232	0.288	k=5 (swept 1..10)
C Std.	Zero fill	0.8910	0.286	fill=zero
S Aggr.	SoftImpute	0.7068	0.282	rank=10 (swept 1..10)
C Aggr.	OneSidedMC	0.8815	0.278	r=2 (swept 1..10)
raw Std.	SoftImpute	0.7401	0.249	rank=9 (swept 1..10)
C Aggr.	missForest	0.8275	0.241	ntree=50 (swept [50,100,200,400])
raw Aggr.	SoftImpute	0.7457	0.228	rank=10 (swept 1..10)
C Std.	Mean fill	0.9268	0.224	fill=mean
R Aggr.	SoftImpute	0.7370	0.209	rank=10 (swept 1..10)
S Std.	Zero fill	0.8926	0.180	fill=zero
S Std.	Mean fill	0.8923	0.179	fill=mean
S Aggr.	OneSidedMC	0.9311	0.160	r=2 (swept 1..10)
C Aggr.	k-NN	0.9253	0.085	k=6 (swept 1..10)
S Aggr.	SoftImpute (corr.)	1.0514	0.076	rank=1
raw Aggr.	SoftImpute (corr.)	1.1797	0.075	rank=3 (swept 1..3)
S Aggr.	k-NN	1.0841	0.069	k=4 (swept 1..10)
R Std.	SoftImpute (corr.)	1.2346	0.047	rank=5 (swept 1..10)
raw Std.	OneSidedMC	1.1045	0.044	r=2 (swept 1..10)
S Aggr.	missForest	1.0635	0.041	ntree=100 (swept [50,100,200,400])
R Std.	missForest	1.2253	0.036	ntree=200 (swept [50,100,200,400])
R Std.	OneSidedMC	1.3507	0.035	r=2 (swept 1..10)
raw Aggr.	missForest	1.2200	0.026	ntree=400 (swept [50,100,200,400])
raw Aggr.	k-NN	1.2247	0.021	k=3 (swept 1..10)
R Aggr.	SoftImpute (corr.)	1.3517	0.019	rank=1
raw Std.	SoftImpute (corr.)	1.2551	0.015	rank=2 (swept 1..10)
raw Aggr.	OneSidedMC	1.5050	0.015	r=2 (swept 1..10)
raw Std.	missForest	1.2441	0.015	ntree=200 (swept [50,100,200,400])
R Aggr.	k-NN	1.4189	0.013	k=2 (swept 1..10)
R Std.	k-NN	1.2854	0.011	k=3 (swept 1..10)
R Aggr.	OneSidedMC	2.0522	0.007	r=2 (swept 1..10)
R Aggr.	missForest	1.3915	0.006	ntree=200 (swept [50,100,200,400])
raw Std.	Mean fill	1.3404	0.000	fill=mean
raw Std.	k-NN	1.3065	-0.004	k=5 (swept 1..10)
R Std.	Mean fill	1.3990	-0.049	fill=mean
S Aggr.	Mean fill	1.2040	-0.087	fill=mean
S Aggr.	Zero fill	1.2298	-0.139	fill=zero
R Std.	USVT	2.5911	-1.635	eta=0.01
S Std.	USVT	1.7823	-2.965	eta=0.01
raw Std.	USVT	2.9244	-5.031	eta=0.01
S Aggr.	USVT	2.9488	-12.278	eta=0.01

O OMEGA SENSITIVITY

To quantify to what degree missing observations affect our results, we run leave-one-covariate-out (LOCO) factor analyses for each valid dataset. For each benchmark in the dataset, we run factor analysis with the benchmark left out, and store the difference in ω_h as a measure of sensitivity. Table A14 shows the correlations between benchmark frequency (normalized within the dataset) and the deltas. Signed averages of r show negligible correlation at $r = 0.054$, but this may simply be because signed correlations cancel out to 0. Average of unsigned, absolute r yielded a larger but still modest correlation of $r = 0.184$.

Table A14. Correlations between benchmark observation and their $\Delta\omega_h$. k = number of factors extracted.

Method	Dataset	k	r	N
Mean fill	C Std.	2	+0.0701	78
Mean fill	C Std.	14	+0.1316	78
k-NN	C Std.	2	+0.0268	78
k-NN	C Std.	7	+0.3205	78
k-NN	S Std.	2	+0.1765	124
k-NN	S Std.	11	+0.2069	124
missForest	C Aggr.	2	+0.0173	102
missForest	C Aggr.	4	+0.2315	102
missForest	C Std.	2	+0.0318	78
missForest	C Std.	4	+0.0100	78
missForest	S Std.	2	-0.3363	124
missForest	S Std.	4	-0.2933	124
OneSidedMC	C Aggr.	2	-0.4256	102
OneSidedMC	C Std.	2	+0.3207	78
OneSidedMC	S Std.	2	+0.3462	124
SoftImpute	C Aggr.	2	-0.0747	102
SoftImpute	C Aggr.	5	+0.1804	102
SoftImpute	C Std.	2	+0.4196	78
SoftImpute	C Std.	9	+0.3245	78
SoftImpute	R Aggr.	2	+0.1227	310
SoftImpute	R Aggr.	20	-0.0677	310
SoftImpute	R Std.	2	+0.0164	298
SoftImpute	R Std.	20	-0.1719	298
SoftImpute	S Aggr.	2	-0.2156	293
SoftImpute	S Aggr.	20	-0.0837	293
SoftImpute	S Std.	2	+0.2324	124
SoftImpute	S Std.	5	+0.0361	124
SoftImpute	raw Aggr.	2	-0.2083	380
SoftImpute	raw Aggr.	10	-0.3003	380
SoftImpute	raw Std.	2	+0.2880	404
SoftImpute	raw Std.	10	+0.0528	404
SoftImpute (corr.)	C Std.	2	-0.0565	78
SoftImpute (corr.)	C Std.	4	+0.0917	78
SoftImpute (corr.)	S Std.	2	-0.1664	124
SoftImpute (corr.)	S Std.	5	+0.2068	124
Zero fill	C Std.	2	+0.2965	78
Zero fill	C Std.	14	+0.2493	78
Average r			**+0.0542**	37 (groups)
Average $\lvert r \rvert$			**0.1840**	

P FULL g-RANKINGS

Table A15 lists all 380 benchmarks used across our analyses, sorted by their rank order as in Table 5 of the main text. The ranking method and its diagnostics are described in Appendix L.

Table A15. All used 380 benchmarks, sorted by their average normalized rank order (ANR) of their g factor loadings, residualized (RANR) against their frequency. The normalized rank order ranges from 0 to 1. 0 = ranked first, 1 = ranked last. N cells = number of EFA solutions with that benchmark. CI and Best/Worst refers to ANR. Leading zeros are omitted. Reference is the paper introducing the benchmark; † marks benchmarks without one, cited by the source of their scores.

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
1	bhasa	-.346	.141	[-.005, .287]	.024	.318	5	Leong et al., 2023
2	mtrag	-.341	.147	[-.051, .346]	.021	.394	5	Katsis et al., 2025
3	creativityprism	-.339	.156	[-.019, .332]	.026	.367	5	Hou et al., 2026
4	eqbench	-.332	.156	[-.060, .372]	.017	.451	5	Paech, 2023
5	mceval	-.331	.156	[.024, .288]	.051	.333	5	Chai et al., 2025
6	pwc_svamp	-.329	.169	[-.033, .371]	.058	.298	4	Patel et al., 2021
7	pwc_drop_test	-.328	.162	[-.139, .462]	.008	.431	4	Dua et al., 2019
8	ProphetArena	-.315	.183	[-.036, .401]	.092	.385	4	Yang et al., 2026
9	dialogbench	-.295	.210	[-1.351, 1.770]	.087	.332	2	Ou et al., 2024
10	pwc_piqa	-.282	.253	[.162, .345]	.003	.822	20	Bisk et al., 2020
11	pwc_timequestions	-.275	.230	[.173, .288]	.226	.235	2	Jia et al., 2021
12	tablebench_data_analysis	-.272	.223	[-.185, .631]	.000	.787	5	Wu et al., 2025
13	pwc_multinli	-.271	.233	[-.459, .925]	.179	.288	2	Williams et al., 2018
14	lawbench	-.264	.222	[-.288, .731]	.057	.452	3	Fei et al., 2024
15	pwc_arc_challenge	-.260	.277	[.149, .405]	.000	.808	20	Clark et al., 2018
16	tablebench_fact_checking	-.260	.235	[-.165, .635]	.020	.784	5	Wu et al., 2025
17	sea_helm	-.257	.235	[-.069, .539]	.045	.603	5	Susanto et al., 2025
18	pinocchio	-.256	.249	[-.564, 1.062]	.185	.313	2	Hu et al., 2023
19	tombench	-.255	.234	[.056, .413]	.062	.446	5	Chen et al., 2024
20	indicgenbench	-.252	.253	[-.347, .853]	.206	.300	2	Singh et al., 2024
21	dischargeme	-.250	.245	[.093, .396]	.087	.402	5	Bedi et al., 2025†
22	pwc_gem_xsum	-.246	.259	[-.508, 1.026]	.199	.319	2	Gehrmann et al., 2021
23	milu	-.241	.248	[.092, .404]	.044	.350	5	Verma et al., 2025
24	dialectbench	-.238	.267	[-.357, .892]	.218	.317	2	FAISAL et al., 2024
25	mena_bench	-.232	.256	[-.163, .675]	.082	.859	5	Zahraei et al., 2025†
26	ReasonBENCH	-.232	.266	[.044, .489]	.075	.380	4	Potamitis et al., 2025
27	sea_exam	-.230	.261	[-.007, .530]	.055	.559	5	Liu et al., 2025
28	evalplus	-.230	.269	[.020, .517]	.102	.479	4	Liu et al., 2023
29	sportsmetrics	-.226	.272	[-.132, .676]	.124	.653	4	Hu et al., 2024
30	kalahi	-.224	.264	[.146, .382]	.148	.362	5	Montalan et al., 2024
31	batayan	-.222	.268	[.053, .482]	.071	.456	5	Montalan et al., 2025
32	bharatbench	-.221	.268	[-.141, .676]	.111	.856	5	Ola Krutrim, 2024
33	tablebench_numerical_reasoning	-.217	.278	[-.146, .702]	.021	.851	5	Wu et al., 2025
34	pwc_pecc	-.216	.273	[.056, .490]	.079	.516	5	Haller et al., 2024
35	race_based_med	-.213	.281	[.042, .520]	.108	.579	5	Omiye et al., 2023
36	shc_conf_med	-.211	.283	[-.036, .602]	.098	.722	5	Bedi et al., 2025†
37	helm	-.209	.282	[-.066, .630]	.048	.764	5	Liang et al., 2023
38	financial_scenarios	-.204	.304	[.121, .488]	.057	.764	10	Liang et al., 2023†
39	kaggle_olamysiak_eclektic	-.203	.310	[.121, .498]	.006	.821	13	Goldman et al., 2025
40	aci_bench	-.200	.294	[.109, .480]	.106	.465	5	wai Yim et al., 2023
41	turl_col_type	-.193	.301	[-.167, .769]	.061	.958	5	Deng et al., 2020

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
42	financebench	-.192	.317	[.174, .460]	.061	.707	10	Islam et al., 2023
43	cmmlu	-.191	.294	[-.276, .864]	.104	.549	3	Li et al., 2024
44	pwc_race	-.191	.299	[-.149, .747]	.032	.886	5	Lai et al., 2017
45	pwc_turbulence	-.191	.294	[.149, .439]	.246	.359	3	Honarvar et al., 2025
46	pwc_big_bench_ reasoning_about_ colored_objects	-.191	.299	[-.196, .793]	.010	.929	5	Srivastava et al., 2023
47	fin_qa	-.190	.319	[.130, .508]	.029	.756	10	Chen et al., 2021b
48	banking77	-.187	.322	[.104, .540]	.013	.886	10	Casanueva et al., 2020
49	wikitq	-.184	.311	[-.126, .748]	.007	.794	5	Pasupat & Liang, 2015
50	cac	-.183	.305	[-.156, .765]	.099	.965	5	Havaldar et al., 2025
51	pwc_big_bench_ winowhy	-.181	.322	[-.870, 1.514]	.042	.876	3	Srivastava et al., 2023
52	pwc_big_bench_ disambiguation_qa	-.180	.310	[-.178, .798]	.013	.902	5	Srivastava et al., 2023
53	synthetic_reasoning	-.179	.343	[.250, .435]	.065	.748	20	Liang et al., 2023
54	afrobench	-.175	.315	[.074, .555]	.000	.534	5	Ojo et al., 2025
55	mt_bench	-.173	.316	[-.037, .669]	.120	.798	5	Zheng et al., 2023
56	followbench	-.170	.319	[.065, .574]	.185	.551	4	Jiang et al., 2024
57	gsm	-.169	.319	[.162, .476]	.000	.936	20	Cobbe et al., 2021
58	pwc_arc_easy	-.169	.341	[.184, .499]	.041	.970	13	Clark et al., 2018
59	xifbench	-.169	.320	[-.089, .728]	.034	.849	5	Li et al., 2025b
60	openbookqa	-.168	.330	[.180, .480]	.013	.931	20	Mihaylov et al., 2018
61	pwc_big_bench_ date_understanding	-.166	.323	[-.156, .803]	.003	.919	5	Srivastava et al., 2023
62	belebele	-.165	.323	[-.120, .765]	.041	.946	5	Bandarkar et al., 2024
63	americasnli	-.162	.343	[-2.033, 2.719]	.156	.530	2	Ebrahimi et al., 2022
64	natural_qa_ closedbook	-.162	.327	[.208, .445]	.052	.909	20	Kwiatkowski et al., 2019
65	medcalc_bench	-.160	.335	[-.029, .698]	.074	.834	5	Khandekar et al., 2024
66	starr_patient_ instructions	-.159	.335	[.017, .653]	.175	.769	5	Bedi et al., 2025[†]
67	pubmedqa	-.156	.348	[.134, .562]	.050	.812	8	Jin et al., 2019
68	pwc_tiq	-.155	.350	[-1.070, 1.770]	.238	.462	2	Jia et al., 2024
69	pwc_big_bench_ penguins_in_a_table	-.155	.334	[-.220, .888]	.010	.933	5	Srivastava et al., 2023
70	pwc_big_bench_ strategyqa	-.155	.348	[-.969, 1.665]	.039	.960	3	Suzgun et al., 2023
71	akata_games_2023	-.154	.334	[-.360, 1.028]	.017	.975	4	Akata et al., 2025
72	thaiexam	-.154	.374	[.263, .486]	.016	.980	20	Pipatanakul et al., 2023
73	emobench	-.154	.334	[.125, .544]	.159	.586	5	Sabour et al., 2024
74	synthetic_ reasoning_natural	-.153	.369	[.274, .464]	.143	.724	20	Liang et al., 2023
75	pwc_big_bench_ sports_ understanding	-.150	.338	[-.216, .892]	.000	.898	5	Srivastava et al., 2023
76	pwc_record	-.149	.361	[.116, .606]	.140	.881	7	Wang et al., 2019
77	pwc_big_bench_ causal_judgment	-.148	.341	[-.167, .850]	.003	.836	5	Srivastava et al., 2023
78	pwc_multirc	-.147	.357	[.208, .507]	.069	.594	8	Wang et al., 2019
79	pwc_ncbi_disease	-.146	.359	[-1.081, 1.799]	.246	.472	2	Dogan et al., 2014
80	pwc_rucos	-.146	.353	[.105, .600]	.237	.411	3	Shavrina et al., 2020
81	agentif	-.146	.346	[.041, .652]	.074	.732	5	Qi et al., 2025
82	medqa	-.144	.362	[.214, .511]	.000	.992	20	Jin et al., 2021
83	thai_exam_a_level	-.144	.384	[.277, .491]	.008	.772	20	Pipatanakul et al., 2023

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
84	flores_200	-.142	.363	[-1.672, 2.399]	.203	.524	2	Costa-jussà et al., 2022
85	litbench	-.140	.352	[-.082, .786]	.024	.933	5	Fein et al., 2026
86	pwc_parus	-.139	.360	[-.430, 1.151]	.146	.726	3	Shavrina et al., 2020
87	shc_ptbm_med	-.137	.357	[.045, .669]	.094	.680	5	Bedi et al., 2025†
88	bigcodebench	-.131	.379	[.220, .538]	.059	.846	13	Zhuo et al., 2025
89	thai_exam_tpat1	-.131	.397	[.296, .498]	.013	.676	20	Pipatanakul et al., 2023
90	mgsm	-.126	.370	[-.070, .810]	.108	.993	5	Shi et al., 2023
91	nusamt	-.122	.383	[-.855, 1.621]	.285	.480	2	Tan & Zhu, 2024
92	math_chain_of_thought	-.120	.367	[.230, .505]	.037	.935	20	Hendrycks et al., 2021c
93	legalbench	-.117	.392	[.243, .542]	.045	.963	20	Guha et al., 2023
94	pwc_big_bench_temporal_sequences	-.113	.376	[-.190, .942]	.005	.980	5	Srivastava et al., 2023
95	neuro_eval	-.113	.386	[.205, .566]	.303	.553	4	Haznitrama et al., 2026
96	thaih6	-.109	.380	[-.021, .780]	.154	.921	5	Limkonchotiwat et al., 2025†
97	ewok_spatial_relations	-.108	.387	[.015, .758]	.010	.839	5	Ivanova et al., 2025
98	tab_fact	-.107	.388	[.024, .751]	.161	.861	5	Chen et al., 2020b
99	mtsamples_replicate	-.106	.388	[.100, .677]	.198	.680	5	Bedi et al., 2025†
100	pwc_cc3m_tagmask	-.096	.409	[-4.122, 4.939]	.052	.765	2	Jo et al., 2024†
101	quac	-.095	.429	[.304, .553]	.020	.878	20	Choi et al., 2018
102	kaggle_andrewmingwang_scicode_subproblem_standard	-.095	.435	[.298, .572]	.000	.935	20	Tian et al., 2024
103	wmt_14	-.092	.417	[.260, .575]	.000	.971	20	Bojar et al., 2014
104	thai_exam_tgat	-.092	.436	[.323, .549]	.003	.901	20	Pipatanakul et al., 2023
105	swiss_legal_bench	-.092	.414	[-3.097, 3.924]	.137	.690	2	Stern et al., 2023†
106	pwc_webapp1k_react	-.091	.398	[.083, .712]	.146	.825	5	Cui, 2024
107	ewok_social_interactions	-.090	.404	[.047, .762]	.088	.866	5	Ivanova et al., 2025
108	ChipBench	-.090	.416	[-3.354, 4.185]	.119	.712	2	Yu et al., 2026
109	kaggle_yulongt_facts_parametric	-.090	.419	[.191, .646]	.024	.951	10	Kaggle, 2025†
110	ewok_physical_interactions	-.088	.406	[-.020, .832]	.047	.970	5	Ivanova et al., 2025
111	pwc_copa	-.088	.410	[.242, .579]	.129	.683	8	Wang et al., 2019
112	shc_bmt_med	-.088	.406	[.149, .663]	.084	.625	5	Bedi et al., 2025†
113	gtbench	-.088	.403	[.241, .565]	.301	.579	5	Duan et al., 2024
114	multiloko	-.087	.439	[.326, .552]	.065	.909	20	Hupkes & Bogoychev, 2025
115	mimic_rrs	-.087	.407	[.142, .672]	.129	.704	5	Chen et al., 2023a
116	kaggle_aminmohamedmohami_browsecomp	-.084	.446	[.297, .595]	.016	.961	20	Wei et al., 2025
117	include	-.084	.405	[-.002, .813]	.094	.696	4	Romanou et al., 2025
118	pwc_commitmentbank	-.083	.406	[.184, .629]	.195	.605	5	Wang et al., 2019
119	mbpp	-.082	.457	[.337, .577]	.030	.905	20	Austin et al., 2021
120	aime25	-.080	.449	[.277, .622]	.008	.987	20	Mathematical Association of America, 2025

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
121	kaggle_sjmikler_livecodebench	-.079	.449	[.299, .600]	.029	1.000	20	Jain et al., 2025
122	filbench	-.077	.412	[.125, .699]	.067	.715	5	Miranda et al., 2025
123	shc_ent_med	-.077	.417	[.002, .833]	.127	.912	5	Bedi et al., 2025†
124	mmlu_prox	-.073	.455	[.321, .589]	.036	.948	20	Xuan et al., 2025
125	mmlu_pro	-.073	.280	[.139, .421]	.000	.927	20	Wang et al., 2024
126	shc_sequoia_med	-.071	.423	[-.029, .875]	.107	.943	5	Bedi et al., 2025†
127	kaggle_andrewmingwang_scicode_main_with_background	-.070	.442	[.273, .611]	.089	.984	13	Tian et al., 2024
128	ilakkanam	-.070	.420	[.139, .700]	.099	.638	5	Varsha et al., 2025
129	mmlu	-.069	.340	[.194, .486]	.000	1.000	20	Hendrycks et al., 2021b
130	madinah_qa	-.069	.460	[.332, .588]	.026	.976	20	Stanford CRFM, 2025†
131	burmesesan	-.067	.421	[-.037, .880]	.089	.970	5	Aung et al., 2026
132	raft	-.065	.459	[.327, .591]	.010	.902	20	Alex et al., 2021
133	humorbench	-.063	.431	[.157, .706]	.057	.637	5	Narad et al., 2025
134	global_piqa	-.063	.440	[-.055, .935]	.401	.479	2	Chang et al., 2025
135	arabicmmlu	-.062	.467	[.343, .592]	.079	.973	20	Koto et al., 2024a
136	medbullets	-.061	.433	[.112, .754]	.162	.809	5	Chen et al., 2025
137	sotopia	-.057	.433	[-.034, .899]	.072	.790	4	Zhou et al., 2024
138	arc	-.056	.356	[.230, .481]	.049	1.000	20	Clark et al., 2018
139	gpqa_diamond	-.054	.475	[.355, .595]	.016	.992	20	Rein et al., 2024
140	wikifact	-.052	.471	[.365, .577]	.000	.851	20	Liang et al., 2023
141	ewok_physical_relations	-.050	.445	[.076, .813]	.040	.871	5	Ivanova et al., 2025
142	pwc_commonsenseqa	-.049	.456	[.106, .805]	.067	.990	8	Talmor et al., 2019
143	mmedbench	-.048	.442	[-.101, .985]	.116	.771	4	Qiu et al., 2024
144	freshqa	-.048	.440	[.026, .854]	.027	.931	5	Vu et al., 2024
145	criticbench	-.047	.443	[.148, .739]	.107	.717	5	Lin et al., 2024
146	facts_search	-.044	.463	[.270, .656]	.000	.813	10	Cheng et al., 2025
147	cogbench	-.042	.470	[.290, .650]	.081	.884	13	Coda-Forno et al., 2024
148	swe_bench	-.042	.466	[.290, .643]	.081	.780	10	Jimenez et al., 2024
149	arabic_exams	-.042	.487	[.359, .615]	.065	.980	20	Hardalov et al., 2020
150	medmcqa	-.041	.453	[.188, .718]	.177	.729	5	Pal et al., 2022
151	pwc_wnli	-.041	.464	[-2.057, 2.985]	.266	.662	2	Wang et al., 2018
152	alrage	-.040	.489	[.361, .617]	.041	.846	20	Stanford CRFM, 2025†
153	truthfulqa	-.038	.366	[.258, .475]	.016	.789	20	Lin et al., 2022
154	pwc_mawps	-.036	.453	[.095, .810]	.137	.785	5	Koncel-Kedziorski et al., 2016
155	sportqa	-.035	.453	[.261, .645]	.276	.541	4	Xia et al., 2024
156	arena_hard_auto	-.034	.497	[.362, .632]	.078	.922	20	Li et al., 2025a
157	vietnamese_glue	-.033	.457	[-.098, 1.013]	.015	.818	4	Tran et al., 2024
158	numeric_nlg	-.031	.463	[.027, .899]	.077	.985	5	Suadaa et al., 2021
159	n2c2_ct_matching	-.031	.463	[.242, .684]	.217	.690	5	Stubbs et al., 2019
160	pwc_big_bench_formal_fallacies_syllogisms_negation	-.031	.459	[.079, .839]	.058	.818	5	Srivastava et al., 2023
161	pwc_big_bench_logic_grid_puzzle	-.031	.472	[.074, .871]	.377	.658	3	Srivastava et al., 2023
162	ewok_material_dynamics	-.030	.464	[.145, .783]	.074	.779	5	Ivanova et al., 2025
163	multichallenge	-.030	.460	[.212, .709]	.285	.774	5	Sirdeshmukh et al., 2025
164	ewok	-.028	.466	[.088, .844]	.054	.906	5	Ivanova et al., 2025

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
165	bbh	-.028	.323	[.182, .464]	.000	.959	20	Suzgun et al., 2023
166	ewok_agent_properties	-.028	.467	[.028, .905]	.007	.995	5	Ivanova et al., 2025
167	pwc_danetqa	-.026	.472	[-.052, .997]	.241	.654	3	Shavrina et al., 2020
168	msmarco_regular	-.026	.484	[.332, .636]	.244	.919	10	Nguyen et al., 2016
169	simpleqa	-.026	.486	[.383, .590]	.171	.692	13	Wei et al., 2024
170	alghafa	-.022	.507	[.377, .637]	.039	1.000	20	Almazrouei et al., 2023
171	kaggle_andrewmingwang_scicode_subproblem_with_background	-.022	.508	[.368, .648]	.049	.935	20	Tian et al., 2024
172	pwc_frontiermath	-.019	.470	[.091, .850]	.069	.849	5	Glazer et al., 2024
173	FlashInfer-Bench	-.018	.480	[-.038, .998]	.062	.847	4	Xing et al., 2026
174	math500	-.017	.516	[.352, .679]	.013	1.000	20	Lightman et al., 2024
175	indoculture	-.017	.473	[-.210, 1.155]	.037	.897	4	Koto et al., 2024b
176	mtsamples_procedures	-.016	.479	[.158, .799]	.272	.926	5	Bedi et al., 2025†
177	gpqa	-.015	.338	[.219, .456]	.040	.919	20	Rein et al., 2024
178	indicqa	-.014	.496	[.322, .670]	.104	.854	10	Doddapaneni et al., 2023
179	pwc_codecontests	-.013	.492	[-5.322, 6.306]	.035	.950	2	Li et al., 2022
180	benchmax	-.011	.476	[.393, .559]	.384	.558	5	Huang et al., 2025b
181	msmarco_trec	-.010	.500	[.368, .631]	.236	.854	10	Nguyen et al., 2016
182	pwc_gigaword	-.010	.495	[-2.480, 3.471]	.261	.730	2	Rush et al., 2015
183	cruxeval	-.010	.496	[.280, .712]	.192	.970	8	Gu et al., 2024
184	livebench	-.008	.484	[.103, .864]	.040	.751	5	White et al., 2025
185	kmmlu	-.006	.482	[.063, .901]	.114	.877	5	Son et al., 2025
186	summarization_xsum	-.006	.525	[.435, .616]	.130	.801	20	Narayan et al., 2018
187	winogrande	-.006	.392	[.274, .510]	.117	.950	20	Sakaguchi et al., 2020
188	bfcl	-.005	.527	[.390, .663]	.106	.926	20	Yan et al., 2024
189	DeceptionBench	-.002	.504	[-1.136, 2.143]	.375	.633	2	Huang et al., 2025c
190	culturescope	-.001	.502	[-.372, 1.375]	.151	.855	3	Zhang et al., 2025b
191	pwc_obqa	-.001	.487	[.239, .735]	.169	.650	5	Mihaylov et al., 2018
192	chatbot_arena	+.001	.499	[.356, .641]	.057	.992	20	Zheng et al., 2023
193	flores_en_id	+.002	.511	[.338, .685]	.051	.792	10	Costa-jussà et al., 2022
194	natural_qa_openbook_longans	+.003	.493	[.390, .596]	.099	.798	20	Kwiatkowski et al., 2019
195	ewok_social_properties	+.005	.500	[.199, .800]	.125	.797	5	Ivanova et al., 2025
196	artificial_analysis_intelligence	+.006	.542	[.440, .644]	.179	1.000	20	Artificial Analysis, 2025
197	pwc_lambada	+.007	.500	[.176, .823]	.243	.868	5	Paperno et al., 2016
198	ewok_quantitative_properties	+.009	.503	[.179, .828]	.091	.799	5	Ivanova et al., 2025
199	sorry_bench	+.012	.502	[.158, .846]	.322	.963	5	Xie et al., 2025
200	ewok_physical_dynamics	+.014	.508	[.233, .782]	.172	.767	5	Ivanova et al., 2025
201	pwc_strategyqa	+.017	.522	[-5.303, 6.346]	.063	.980	2	Geva et al., 2021
202	gsm8k	+.017	.413	[.263, .563]	.000	1.000	20	Cobbe et al., 2021
203	kaggle_vijitsingh1_mgsm_english	+.017	.546	[.436, .656]	.136	.911	20	Shi et al., 2023
204	bbq	+.020	.531	[.390, .671]	.000	.976	20	Parrish et al., 2022
205	Vericoding	+.021	.526	[.305, .748]	.509	.544	2	Bursuc et al., 2026
206	boolq	+.022	.545	[.441, .649]	.040	.919	20	Clark et al., 2019

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
207	pwc_anli_test	+.023	.511	[.221, .802]	.247	.819	5	Nie et al., 2020
208	ewok_social_relations	+.025	.519	[.145, .894]	.064	.911	5	Ivanova et al., 2025
209	moralbench	+.027	.512	[-.424, 1.448]	.132	.886	3	Ji et al., 2025
210	ehr_sql	+.027	.521	[.145, .897]	.204	.865	5	Lee et al., 2022
211	medec	+.031	.525	[.140, .911]	.133	.845	5	Abacha et al., 2025
212	babi_qa	+.033	.554	[.442, .667]	.050	.921	20	Weston et al., 2016
213	vectara	+.034	.522	[.412, .631]	.435	.660	5	Vectara, 2023
214	ewok_material_properties	+.035	.529	[.124, .934]	.030	.940	5	Ivanova et al., 2025
215	pwc_webapp1k_duo_react	+.036	.524	[.305, .743]	.293	.723	5	Cui, 2024
216	kaggle_andrewmingwang_scicode_main_standard	+.036	.550	[.361, .738]	.008	.911	13	Tian et al., 2024
217	kaggle_andrewmingwang_dsqa	+.036	.526	[.309, .743]	.364	.722	5	Kaggle, 2025†
218	scigen	+.037	.531	[.137, .925]	.118	.955	5	Moosavi et al., 2021
219	pwc_terra	+.037	.536	[-.367, 1.438]	.236	.939	3	Shavrina et al., 2020
220	math_regular	+.038	.575	[.463, .687]	.077	1.000	20	Hendrycks et al., 2021c
221	culemo	+.038	.544	[.510, .577]	.541	.546	2	Belay et al., 2025
222	eifbench	+.042	.534	[.418, .649]	.462	.665	5	Zou et al., 2025
223	pwc_aime24	+.044	.532	[.260, .804]	.290	.683	4	Mathematical Association of America, 2025
224	aratrust	+.047	.576	[.441, .710]	.052	.974	20	Alghamdi et al., 2025
225	livecodebench	+.049	.540	[.194, .887]	.178	.914	5	Jain et al., 2025
226	pwc_storycloze	+.050	.550	[.309, .791]	.198	.980	8	Mostafazadeh et al., 2016
227	pwc_siqa	+.052	.539	[-.130, 1.208]	.025	.887	4	Sap et al., 2019
228	hellaswag	+.052	.463	[.340, .586]	.126	.983	20	Zellers et al., 2019
229	indicsentiment	+.052	.562	[.345, .780]	.065	.951	10	Doddapaneni et al., 2023
230	entity_data_imputation	+.055	.578	[.449, .707]	.030	.990	20	Liang et al., 2023
231	pwc_apps	+.060	.565	[-2.515, 3.645]	.323	.807	2	Hendrycks et al., 2021a
232	lindsea_pragmatics_presuppositions_id	+.061	.571	[.373, .770]	.049	.896	10	Leong et al., 2023
233	LemmaBench	+.063	.566	[.390, .743]	.494	.636	3	Peyronnet et al., 2026
234	mental_health	+.063	.558	[.225, .890]	.252	.846	5	Bedi et al., 2025†
235	aime_2025	+.065	.559	[.174, .943]	.189	.893	5	Mathematical Association of America, 2025
236	culturalbench	+.065	.554	[.295, .813]	.393	.882	5	Chiu et al., 2024
237	pwc_peerqa	+.066	.554	[.322, .786]	.330	.816	5	Baumgärtner et al., 2025
238	pwc_rwsd	+.068	.567	[-.369, 1.503]	.315	1.000	3	Shavrina et al., 2020
239	kaggle_aminmohamedmohami_indic_gen_bench	+.068	.562	[.353, .772]	.421	.844	5	Singh et al., 2024
240	StatEval	+.071	.577	[-2.021, 3.174]	.372	.781	2	Lu et al., 2025
241	pwc_carb	+.073	.558	[-.468, 1.584]	.135	.960	3	Bhardwaj et al., 2019
242	facts_grounding	+.076	.608	[.481, .736]	.223	.987	20	Cheng et al., 2025
243	pwc_lidirus	+.077	.576	[-.247, 1.398]	.248	.910	3	Shavrina et al., 2020
244	flores_id_en	+.077	.587	[.365, .809]	.111	.984	10	Costa-jussà et al., 2022

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
245	naturalquestions	+.079	.568	[.166, .970]	.032	.836	5	Kwiatkowski et al., 2019
246	nusax	+.080	.589	[.451, .728]	.229	.911	10	Winata et al., 2023
247	bold	+.082	.614	[.494, .734]	.143	.984	20	Dhamala et al., 2021
248	harmbench	+.082	.593	[.474, .712]	.244	1.000	20	Mazeika et al., 2024
249	entity_matching	+.082	.605	[.461, .750]	.069	.961	20	Liang et al., 2023
250	wisesight	+.083	.597	[.458, .736]	.135	.921	13	Susanto et al., 2025†
251	pwc_oie2016	+.085	.571	[-.494, 1.635]	.140	.997	3	Stanovsky & Dagan, 2016
252	irokobench	+.086	.578	[.130, 1.026]	.144	.953	5	Adelani et al., 2025
253	pwc_newsqa	+.092	.584	[.177, .991]	.079	.949	5	Trischler et al., 2017
254	xstest	+.095	.606	[.481, .731]	.179	.987	20	Röttger et al., 2024
255	kaggle_andrewmingwang_simpleqa_verified	+.097	.629	[.520, .738]	.268	1.000	20	Haas et al., 2025
256	summarization_cnndm	+.098	.630	[.515, .745]	.188	.987	20	Hermann et al., 2015
257	shc_privacy_med	+.100	.594	[.365, .823]	.343	.760	5	Bedi et al., 2025†
258	xnli	+.101	.615	[.496, .734]	.249	.871	13	Conneau et al., 2018
259	civil_comments	+.101	.624	[.518, .731]	.277	.909	20	Borkan et al., 2019
260	narrative_qa	+.102	.591	[.482, .700]	.059	.870	20	Kociský et al., 2018
261	LegalEval-Q	+.104	.608	[.063, 1.153]	.355	.749	3	Li & Wu, 2025
262	musr	+.105	.458	[.331, .584]	.109	.971	20	Sprague et al., 2024
263	mimiciv_billing_code	+.105	.600	[.259, .940]	.196	.862	5	Bedi et al., 2025†
264	flores_en_vi	+.105	.615	[.411, .820]	.128	.953	10	Costa-jussà et al., 2022
265	simple_safety_tests	+.106	.616	[.488, .744]	.033	1.000	20	Vidgen et al., 2023
266	disinformation_wedging	+.106	.638	[.506, .770]	.129	1.000	20	Liang et al., 2023
267	flores_th_en	+.107	.617	[.424, .809]	.163	.943	10	Costa-jussà et al., 2022
268	math	+.107	.454	[.305, .603]	.024	.987	20	Sobhani et al., 2026
269	flores_vi_en	+.107	.617	[.404, .830]	.084	.976	10	Costa-jussà et al., 2022
270	shc_proxy_med	+.109	.603	[.222, .984]	.117	.939	5	Bedi et al., 2025†
271	real_toxicity_prompts	+.109	.641	[.513, .769]	.065	.950	20	Gehman et al., 2020
272	shc_gip_med	+.111	.605	[.226, .984]	.154	.877	5	Bedi et al., 2025†
273	shc_sei_med	+.112	.606	[.365, .847]	.316	.860	5	Bedi et al., 2025†
274	worldvaluesbench	+.112	.602	[.227, .977]	.082	.829	5	Zhao et al., 2024
275	pwc_rcb	+.113	.612	[.196, 1.028]	.454	.788	3	Shavrina et al., 2020
276	pwc_penn_treebank_word_level	+.113	.618	[-1.204, 2.441]	.475	.762	2	Marcus et al., 1993
277	RealMath	+.114	.602	[.178, 1.026]	.065	.879	5	Zhang et al., 2025a
278	aime	+.115	.619	[.600, .638]	.617	.620	2	Mathematical Association of America, 2025
279	kaggle_sripalthilakraj_global_mmlu_lite_english	+.115	.627	[.463, .791]	.057	.960	13	Singh et al., 2025
280	pwc_samsum	+.118	.623	[-3.893, 5.140]	.268	.979	2	Gliwa et al., 2019
281	maliciousinstruct	+.118	.607	[.225, .989]	.136	.990	5	Huang et al., 2024
282	lindsea_syntax_minimal_pairs_id	+.119	.629	[.433, .824]	.269	.978	10	Leong et al., 2023
283	mega	+.121	.610	[-.073, 1.293]	.139	.992	4	Ahuja et al., 2023
284	indommlu	+.121	.613	[.134, 1.092]	.134	.962	5	Koto et al., 2023

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
285	xcopa	+.128	.642	[.483, .801]	.187	1.000	13	Ponti et al., 2020
286	ifeval	+.129	.482	[.360, .605]	.030	.984	20	Zhou et al., 2023
287	indonli	+.131	.641	[.515, .767]	.447	.942	10	Mahendra et al., 2021
288	flores_en_ta	+.131	.641	[.499, .783]	.350	.974	10	Costa-jussà et al., 2022
289	lsat_qa	+.133	.655	[.542, .767]	.221	.948	20	Zhong et al., 2021
290	twitter_aae	+.133	.643	[.466, .819]	.261	.967	10	Blodgett et al., 2016
291	kaggle_ andrewmingwang_ facts	+.133	.641	[.401, .880]	.081	1.000	10	Jacovi et al., 2025
292	kaggle_ andrewmingwang_ asset_ops_bench	+.134	.643	[.383, .904]	.033	.992	10	Constantinides et al., 2025
293	medal	+.135	.640	[-3.237, 4.518]	.335	.945	2	Mendonça et al., 2026
294	kaggle_jonlipovetz_ game_arena	+.136	.626	[.349, .903]	.310	.815	5	Kaggle, 2025†
295	pwc_asdiv_a	+.140	.645	[-2.125, 3.415]	.427	.863	2	Miao et al., 2020
296	anthropic_red_team	+.142	.652	[.538, .767]	.041	.951	20	Ganguli et al., 2022
297	uitvsfc	+.142	.652	[.501, .803]	.268	.862	10	Nguyen et al., 2018
298	the_pile	+.142	.674	[.568, .780]	.333	1.000	20	Gao et al., 2021
299	disinformation_ reiteration	+.143	.675	[.545, .805]	.000	.948	20	Liang et al., 2023
300	ttcw	+.144	.649	[-1.535, 2.833]	.478	.821	2	Chakrabarty et al., 2024
301	legal_support	+.145	.667	[.548, .785]	.114	.909	20	Liang et al., 2023
302	tydiqa	+.145	.655	[.432, .878]	.203	.997	10	Clark et al., 2020
303	pwc_muserc	+.147	.646	[.460, .832]	.565	.712	3	Shavrina et al., 2020
304	mexa	+.147	.635	[.276, .994]	.191	.949	5	Yu et al., 2025b
305	opencompass	+.147	.653	[.496, .810]	.640	.665	2	OpenCompass Contributors, 2023
306	SuperGPQA	+.151	.640	[.251, 1.029]	.131	.918	5	Team et al., 2025
307	oogiri	+.152	.641	[.148, 1.133]	.205	.932	4	Murakami et al., 2025
308	mimic_bhc	+.154	.648	[.361, .936]	.308	.956	5	Aali et al., 2024
309	lindsea_pragmatics_ scalar_implicatures_ id	+.154	.664	[.494, .834]	.106	.922	10	Leong et al., 2023
310	humaneval	+.156	.661	[.398, .924]	.256	.993	8	Chen et al., 2021a
311	medhallu	+.160	.655	[.336, .973]	.233	.900	5	Pandit et al., 2025
312	flores_ta_en	+.161	.671	[.487, .855]	.242	.984	10	Costa-jussà et al., 2022
313	med_dialog	+.162	.656	[.393, .918]	.356	.859	5	Chen et al., 2020a
314	pwc_wikitext_2	+.163	.668	[-2.208, 3.545]	.442	.894	2	Merity et al., 2017
315	fever	+.164	.650	[.257, 1.043]	.468	.754	3	Thorne et al., 2018
316	flores_en_th	+.167	.677	[.536, .818]	.317	.901	10	Costa-jussà et al., 2022
317	pwc_wikitext_103	+.168	.672	[.349, .995]	.646	.697	2	Merity et al., 2017
318	pwc_safim	+.170	.667	[.003, 1.330]	.050	.951	4	Gong et al., 2024
319	pwc_abstractive_ text_ summarization_ from_il_post	+.173	.678	[-1.634, 2.990]	.496	.860	2	Landro et al., 2022
320	vihsd	+.175	.689	[.560, .818]	.220	.990	13	Luu et al., 2021
321	blimp	+.176	.686	[.590, .782]	.536	.967	10	Warstadt et al., 2020
322	thaitoxicitytweets	+.177	.687	[.508, .867]	.114	.987	10	Susanto et al., 2025†
323	mlhsd	+.179	.689	[.536, .843]	.228	.942	10	Susanto et al., 2025†
324	xquad	+.179	.693	[.576, .810]	.407	.926	13	Artetxe et al., 2020
325	kaggle_nanliao7_ itbench	+.179	.673	[.325, 1.020]	.186	.875	5	Jha et al., 2025
326	pwc_text8	+.180	.685	[-2.054, 3.424]	.470	.901	2	Mahoney, 2011

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
327	humanitys_last_exam	+.181	.672	[.162, 1.182]	.047	.966	5	Phan et al., 2026
328	copyright_text	+.187	.719	[.616, .822]	.228	.982	20	Liang et al., 2023
329	kaggle_jonlipovetz_chess_suite	+.188	.678	[.267, 1.090]	.270	.993	5	Kaggle, 2025†
330	qtsumm	+.190	.684	[.366, 1.003]	.401	.966	5	Zhao et al., 2023
331	llmchess	+.194	.689	[.512, .867]	.439	.792	5	Kolasani et al., 2025
332	pwc_sst_5_fine_grained_classification	+.196	.699	[.104, 1.295]	.501	.966	3	Socher et al., 2013
333	vmlu	+.201	.707	[.434, .980]	.017	.970	8	VMLU Team, 2024
334	pwc_openwebtext	+.201	.705	[-2.692, 4.102]	.438	.973	2	Gokaslan & Cohen, 2019
335	kaggle_nanliao7_enterprise_ops	+.204	.697	[.354, 1.040]	.216	.918	5	Kaggle, 2025†
336	triangulating	+.209	.706	[.080, 1.333]	.127	.958	4	Momentè et al., 2025
337	NormAd	+.209	.714	[.554, .875]	.702	.727	2	Rao et al., 2025
338	StrongREJECT	+.215	.720	[-.422, 1.862]	.630	.810	2	Souly et al., 2024
339	imdb	+.217	.740	[.630, .850]	.089	.974	20	Maas et al., 2011
340	pwc_bioasq	+.219	.722	[.055, 1.389]	.417	.923	3	Tsatsaronis et al., 2015
341	medi_qa	+.222	.717	[.342, 1.091]	.212	.961	5	Abacha et al., 2019b
342	pwc_asqp	+.229	.734	[-.247, 1.715]	.657	.811	2	Zhang et al., 2021b
343	shc_cdi_med	+.230	.724	[.422, 1.026]	.391	.989	5	Bedi et al., 2025†
344	abceval	+.231	.720	[.238, 1.202]	.273	.913	4	Zhao et al., 2025
345	alpacaeval	+.231	.718	[.629, .808]	.620	.793	5	Dubois et al., 2024
346	pwc_tasd	+.231	.737	[-.243, 1.717]	.660	.814	2	Wan et al., 2020
347	triviaqa	+.233	.738	[.526, .949]	.356	.990	8	Joshi et al., 2017
348	sibench	+.235	.723	[.432, 1.013]	.464	.994	5	Huang et al., 2025a
349	complexbench	+.236	.724	[.569, .878]	.562	.887	5	Wen et al., 2024
350	wildbench	+.239	.747	[.571, .923]	.437	.986	8	Lin et al., 2025
351	xcr_bench	+.239	.728	[.286, 1.170]	.141	.997	5	Kabir et al., 2026
352	medalign	+.241	.735	[.510, .960]	.514	.923	5	Fleming et al., 2024
353	pwc_multitq	+.244	.750	[-1.428, 2.927]	.578	.921	2	Chen et al., 2023b
354	dyck_language	+.247	.770	[.670, .869]	.238	.984	20	Liang et al., 2023
355	pwc_one_billion_word	+.248	.753	[-1.472, 2.978]	.578	.928	2	Chelba et al., 2014
356	ice	+.255	.765	[.584, .946]	.280	.990	10	Liang et al., 2023†
357	facts_parametric	+.260	.753	[.462, 1.045]	.434	.979	5	Cheng et al., 2025
358	clear	+.263	.757	[.434, 1.081]	.303	.955	5	Bedi et al., 2025†
359	multipl_e	+.269	.756	[.338, 1.173]	.164	.983	5	Cassano et al., 2023
360	pwc_pubmed	+.276	.781	[.282, 1.281]	.742	.821	2	Sen et al., 2008
361	ShoppingMMLU	+.278	.781	[.562, 1.000]	.715	.881	3	Jin et al., 2024
362	pwc_arxiv_hep_th_citation_graph	+.278	.784	[.283, 1.284]	.744	.823	2	Leskovec et al., 2005
363	workplacehumor	+.282	.771	[.530, 1.011]	.469	.995	5	Shafiei & Saffari, 2025
364	rpgbench	+.285	.776	[.529, 1.023]	.441	.948	5	Yu et al., 2025a
365	ehrshot	+.292	.786	[.528, 1.044]	.458	.974	5	Wornow et al., 2023
366	pwc_webquestions	+.295	.781	[.547, 1.014]	.720	.889	3	Berant et al., 2013
367	chw_care_plan	+.308	.802	[.641, .964]	.646	.935	5	Bedi et al., 2025†
368	head_qa	+.312	.806	[.633, .979]	.615	.976	5	Vilares & Gómez-Rodríguez, 2019
369	medication_qa	+.322	.816	[.463, 1.170]	.310	.990	5	Abacha et al., 2019a
370	pwc_lila_ood	+.338	.843	[.319, 1.366]	.801	.884	2	Mishra et al., 2022
371	pwc_sst_2_binary_classification	+.342	.847	[-.588, 2.283]	.734	.960	2	Socher et al., 2013
372	multi_if	+.362	.852	[.678, 1.026]	.710	.977	4	He et al., 2024

Table A15. (continued)

No	Benchmark	ρ_ϵ	ρ	95% CI	Best	Worst	N	Reference
373	pwc_lila_iid	+.369	.874	[.452, 1.297]	.841	.908	2	Mishra et al., 2022
374	pwc_ag_news	+.371	.874	[.520, 1.228]	.710	.964	3	Zhang et al., 2015
375	pwc_rte	+.373	.879	[.244, 1.514]	.829	.929	2	Wang et al., 2019
376	pwc_kvret	+.397	.902	[.729, 1.075]	.888	.916	2	Eric et al., 2017
377	pwc_cronquestions	+.419	.924	[.531, 1.317]	.893	.955	2	Saxena et al., 2021
378	pwc_mr	+.441	.946	[.259, 1.633]	.892	1.000	2	Pang & Lee, 2005
379	pwc_django	+.478	.984	[.779, 1.189]	.968	1.000	2	Oda et al., 2015
380	pwc_conala	+.492	.997	[.996, .998]	.997	.998	2	Yin et al., 2018

Q COMMON-SUBJECT UMAP PLOTS

This section presents the UMAP plots for the benchmarks' composite distance, with each figure showing the plots of one imputation method. Each plot is composited over the standard and aggressive collapse strategies. The aggregate plots are composited over all collapse strategies and all imputation methods. Since the R and raw datasets only have 1 valid imputation each, they do not have a dedicated aggregate figure. This section has 16 figures in total.

Notably, through the different imputation strategies, all of the visualizations show the same pattern, in which shared-domain benchmarks tend to be dispersed across the vector space. Highlighting the same benchmarks as in the results (with some benchmarks discarded by each densifier), we see the same pattern that, for example, `aime25` and `gsm8k` are distanced quite far from each other. Through most of the different possible solutions, our core claim that same-domain benchmarks are not guaranteed to cluster together is robust.

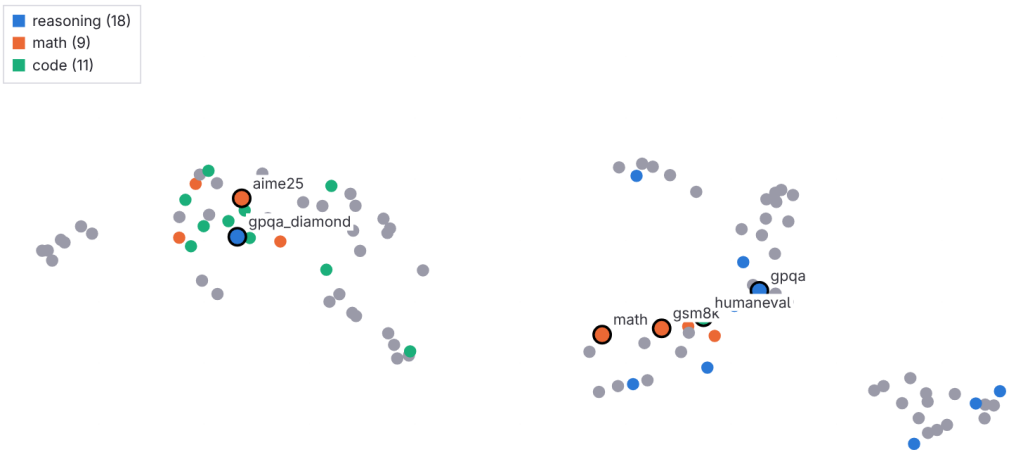

Figure A2. C dataset, aggregated across 7 imputations.

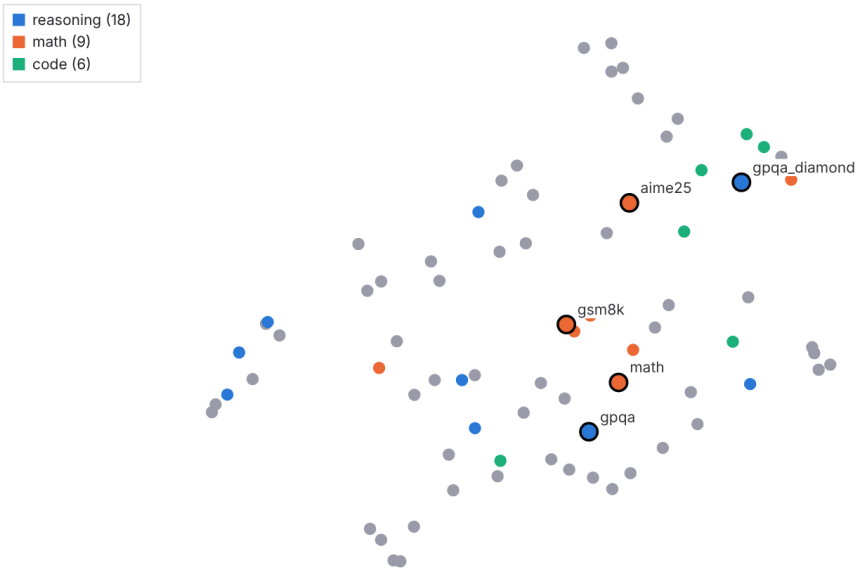

Figure A3. C dataset, imputed with mean fill.

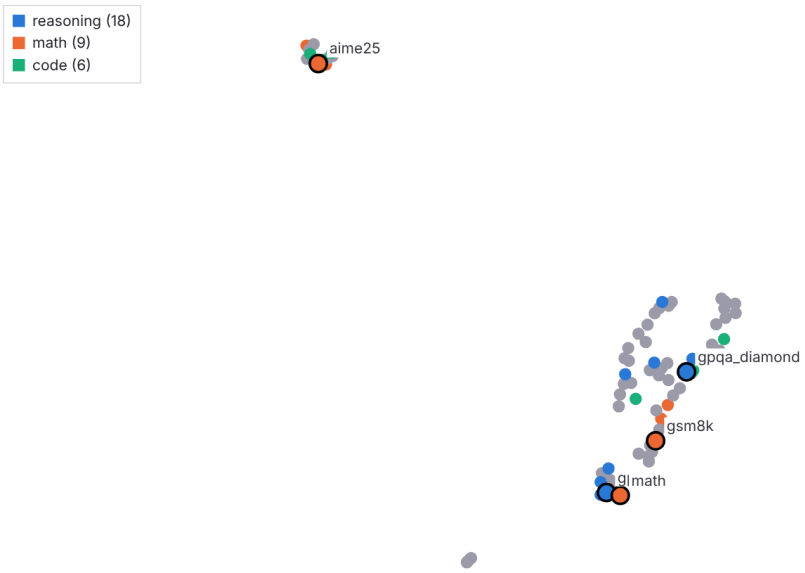

Figure A4. C dataset, imputed with k-NN.

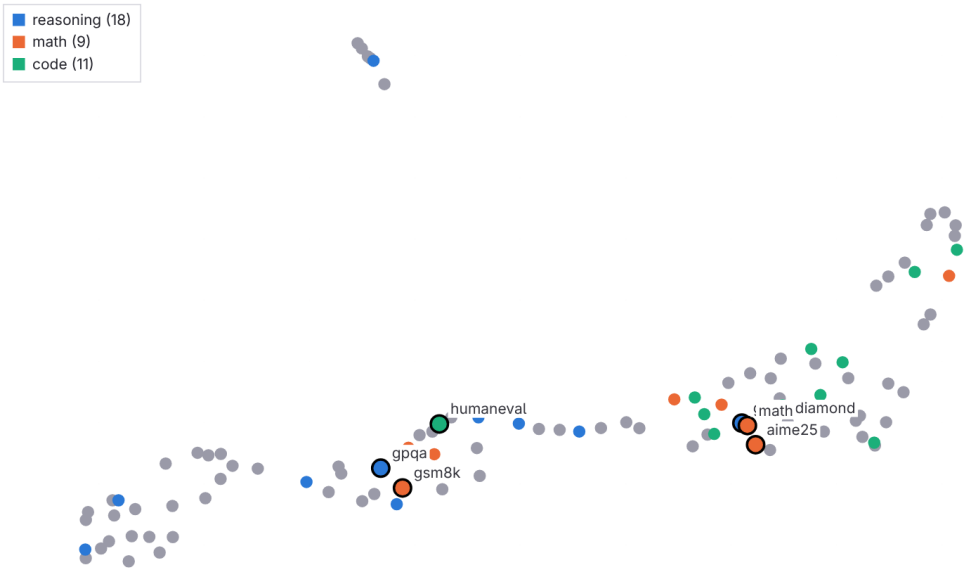

Figure A5. C dataset, imputed with missForest.

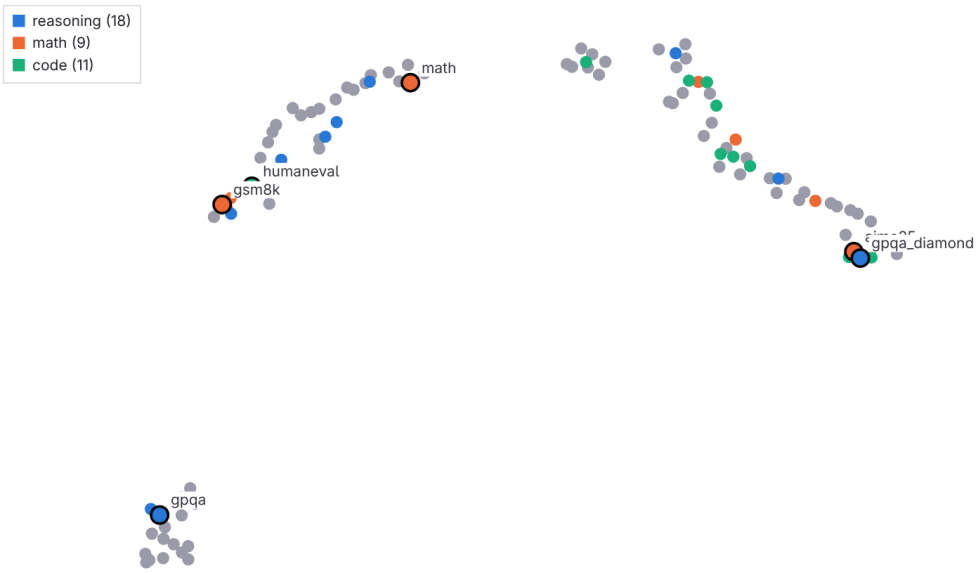

Figure A6. C dataset, imputed with OneSidedMC.



Figure A7. C dataset, imputed with SoftImpute.

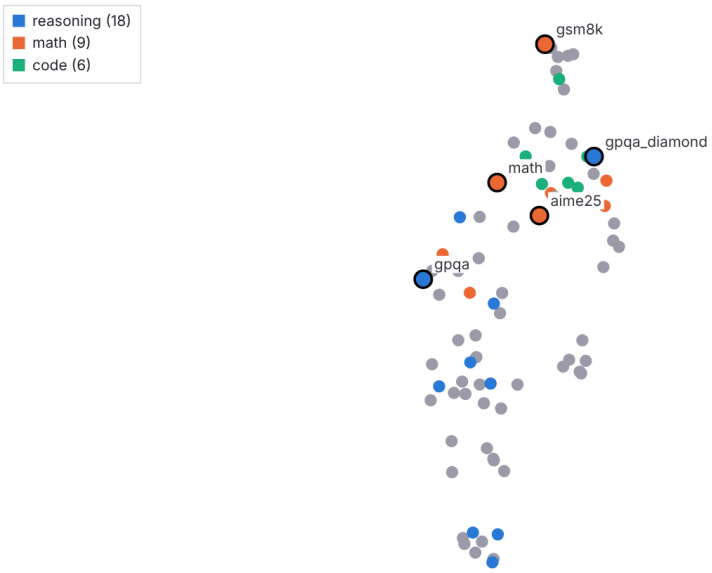

Figure A8. C dataset, imputed with SoftImpute (corr.).

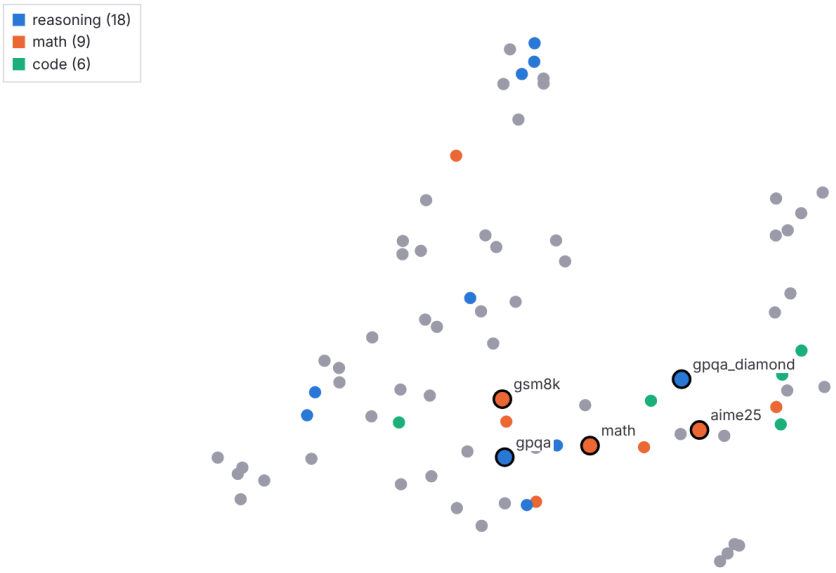

Figure A9. C dataset, imputed with zero fill.

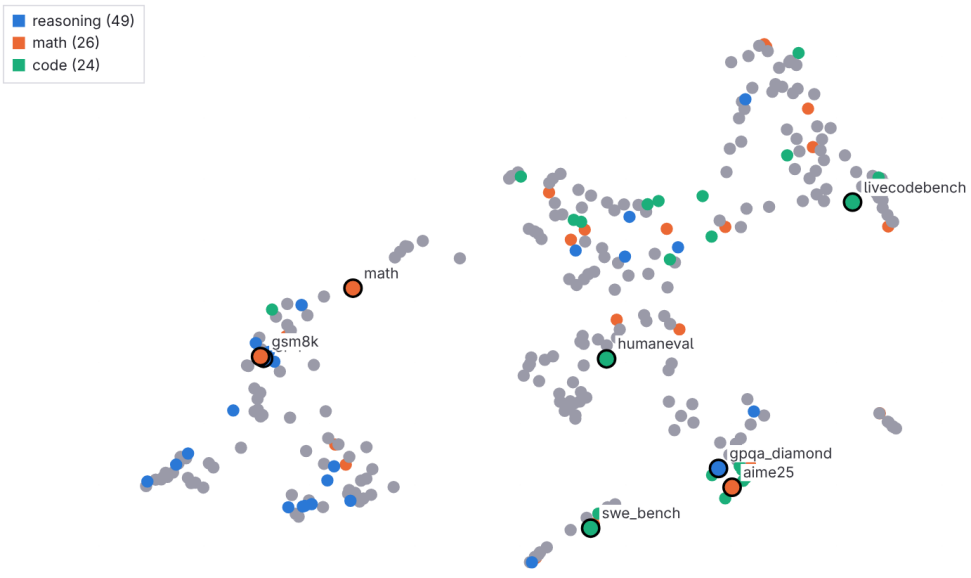

Figure A10. S dataset, aggregated across 5 imputations.

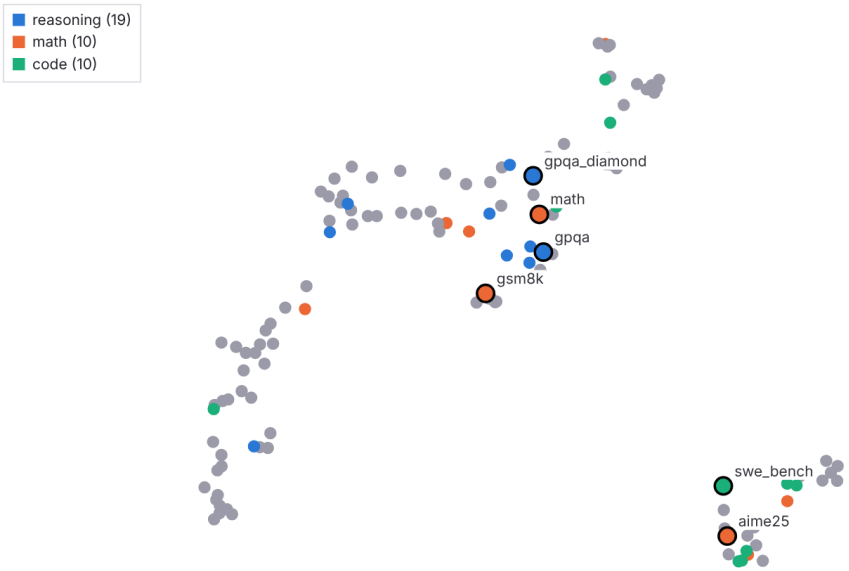

Figure A11. S dataset, imputed with k-NN.

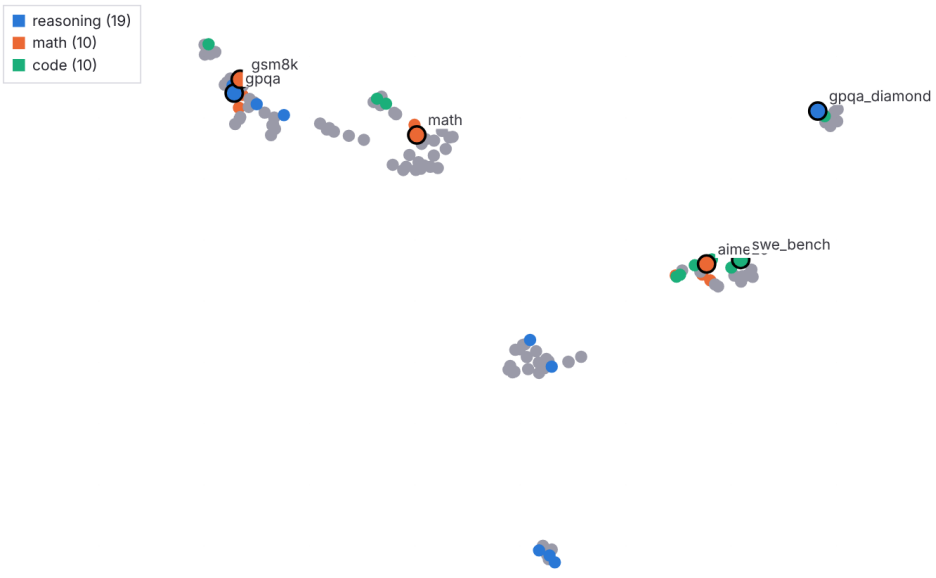

Figure A12. S dataset, imputed with missForest.

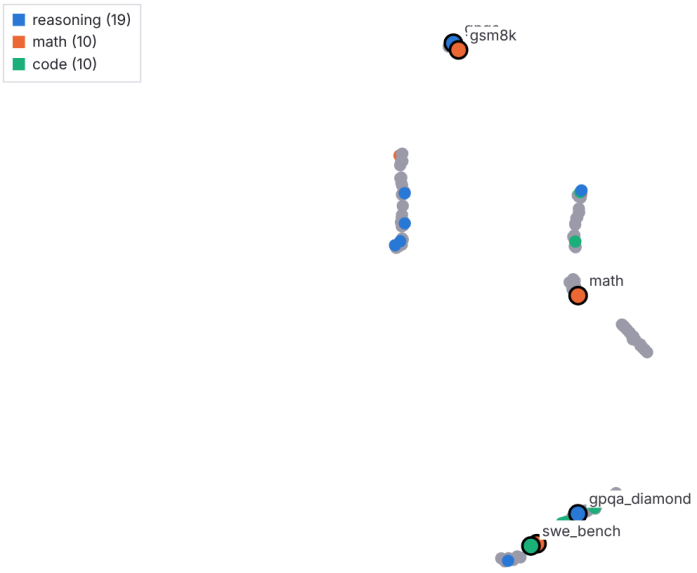

Figure A13. S dataset, imputed with OneSidedMC.

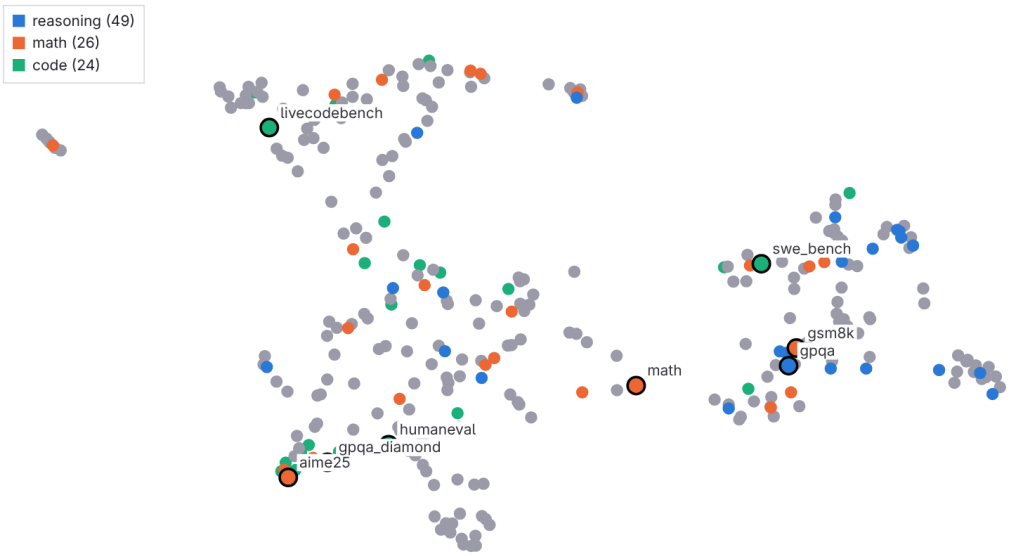

Figure A14. S dataset, imputed with SoftImpute.

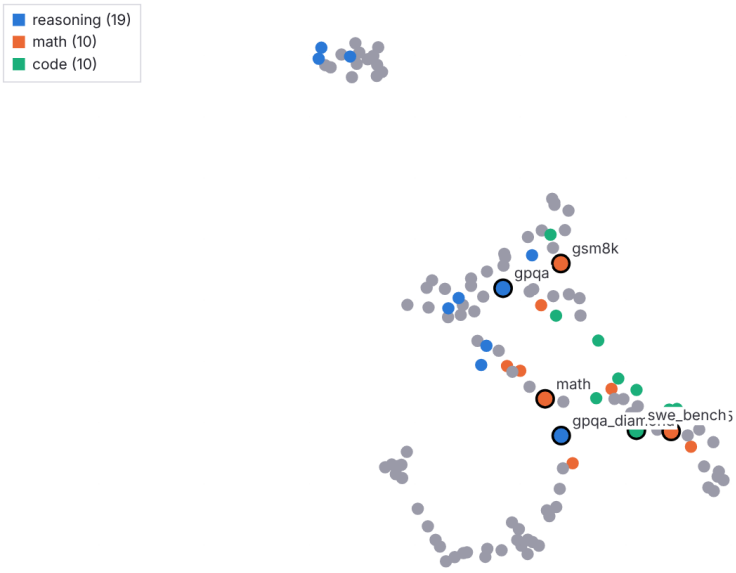

Figure A15. S dataset, imputed with SoftImpute (corr.).

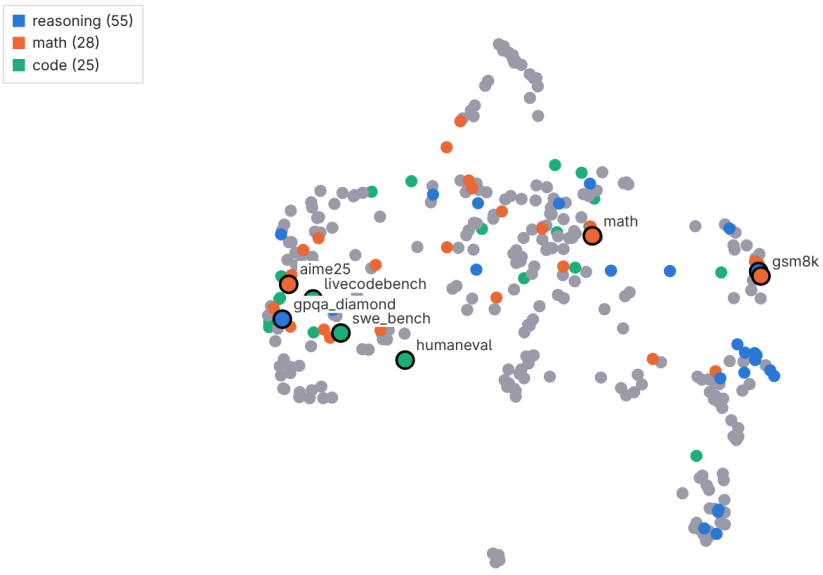

Figure A16. R dataset, imputed with SoftImpute.

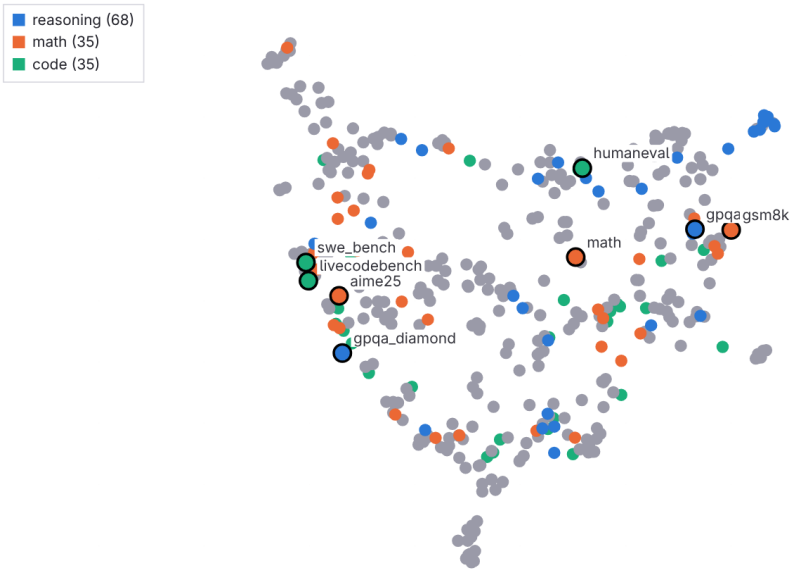

Figure A17. Raw dataset, imputed with SoftImpute.