

Benchmark Profiling: Mechanistic Diagnosis of LLM Benchmarks

Anonymous ACL submission

Abstract

Large Language Models are commonly judged by their scores on standard benchmarks, yet such scores often overstate real capability since they mask the mix of skills a task actually demands. For example, ARC is assumed to test reasoning, while HellaSwag is designed to evaluate commonsense. However, we lack a systematic way to verify if these benchmarks actually measure these labels. We introduce BENCHMARK PROFILING, a diagnostic framework that decomposes benchmark performance into ten cognitively grounded abilities. The method combines gradient-based importance scoring with targeted parameter ablation to compute an *Ability Impact Score* (AIS) that quantifies how much each ability contributes to a model’s success on a given benchmark. Profiling three instruction-tuned models across ten widely used benchmarks yields four key findings: (i) most benchmarks draw on several abilities rather than one, (ii) datasets with similar labels rely on distinct ability mixtures, (iii) code-generation benchmarks reward broad, multi-skill improvement and thus show only modest gains from narrow domain-specific fine-tuning, and (iv) abilities irrelevant to the task could negatively affect performance. BENCHMARK PROFILING therefore explains why performance gains do not always translate into user-perceived competence and offer a transparent tool for benchmark audit and model interpretability. Code and datasets are publicly available.¹

1 Introduction

Modern evaluations of Large Language Models (LLMs) depend heavily on standardized benchmarks designed to test capabilities like reasoning, commonsense, and knowledge (Liang et al., 2022; Cobbe et al., 2021; Zellers et al., 2019). While these benchmarks provide quantitative measures

¹Due to anonymity, links will be released upon acceptance.

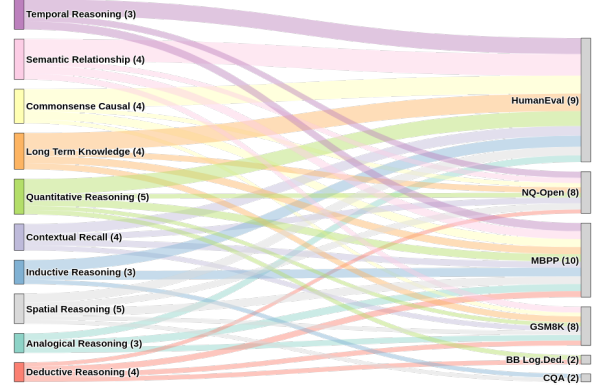


Figure 1: Top ability–benchmark links for Llama-3.1-8B-Instruct derived from its Benchmark Profile (ribbons shown only where AIS > 0.05; ribbon width \propto impact).

of performance, a growing body of evidence suggests a discrepancy between high scores on automated metrics and the qualities humans value in LLM interaction. For instance, models optimized for benchmarks can sometimes produce outputs that are misaligned with human preferences, as evidenced by the mismatched rankings between platforms like the Open LLM Leaderboard (Fourrier et al., 2024) and the Chatbot Arena LLM Leaderboard (Chiang et al., 2024). This misalignment raises a critical concern: **current benchmarks may not accurately measure the abilities they claim to assess**, undermining their reliability as indicators of true model competence.

The core issue lies in the ambiguity of benchmark design. High accuracy scores on benchmarks, are often taken as direct proof that a model possesses the high-level ability suggested by the benchmark’s label (e.g., *math* or *commonsense*), despite a lack of rigorous verification (Eriksson et al., 2025). In reality, models might exploit dataset artifacts or memorize patterns to achieve high scores without genuine understanding (McCoy et al., 2019; Geva et al., 2021). Without knowing what benchmarks truly measure, we cannot reliably improve

models or design evaluations that reflect real-world requirements (Bowman and Dahl, 2021).

To address this, we introduce BENCHMARK PROFILING, a methodology that systematically diagnoses the functional abilities required by LLM benchmarks. By defining 10 operationalized abilities (e.g., Deductive Reasoning, Contextual Recall) derived from established models of human intelligence (Carroll, 1993), we create measurement criteria that reflect both computational performance and the cognitive dimensions humans intuitively value in real-world interactions. This approach directly tackles the *Performance-Perception Paradox*, where models dominate benchmarks yet underwhelm users, by ensuring evaluations test the same competencies people assess when judging capability (Kyllonen, 2021). Bridging this gap, our profiles reveal whether "high-scoring" models truly exhibit the abilities users expect from labels like *math* or *commonsense*. The BENCHMARK PROFILING framework measures how much each ability actually contributes to a model's success on each benchmark, using targeted parameter ablation and our proposed Ability Impact Score (AIS). This approach produces diagnostic profiles that reveal the true combination of abilities required for high performance on every benchmark.

2 Related Work

The Benchmarking paradigm in LLM Evaluation

Large Language Models (LLMs) are predominantly evaluated through standardized benchmarks like MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), and TruthfulQA (Lin et al., 2021), which quantify performance on tasks such as commonsense reasoning, mathematical problem-solving, and factual accuracy. While these benchmarks have driven rapid progress via leaderboards, their limitations in capturing real-world competence and human-aligned abilities have become increasingly apparent. Critiques highlight issues such as dataset contamination, prompt sensitivity, and the prevalence of shortcut learning, where models exploit artifacts rather than demonstrating genuine understanding (McCoy et al., 2019; Geva et al., 2021; Bowman and Dahl, 2021).

A growing body of research reveals systemic flaws in the benchmarking paradigm. Studies demonstrate that models often achieve high scores through memorization or spurious correlations,

failing to exhibit robust reasoning or adaptability (Banerjee et al., 2024; Oh et al., 2024). For instance, inherent limitations, such as overfitting to narrow metrics, and a lack of theoretical justification for real-world relevance, are shown in benchmarks like MMLU containing mislabeled or irrelevant questions (Fodor, 2025). These critiques align with observations of the *performance-perception paradox*, a term we introduce to describe the disconnect between benchmark-driven progress and the abilities users intuitively expect from LLMs in practical applications. Further analyses reveal that benchmarks often inadequately measure reasoning depth, exhibit cultural biases, and struggle with adversarial robustness (McIntosh et al., 2025), underscoring the need for evaluations grounded in human-aligned competencies.

Mechanistic Interpretability in Language Models

Mechanistic interpretability seeks to reverse engineer neural networks by mapping their internal computations to human-understandable algorithms and concepts, aiming for a granular, causal understanding of model behavior (Bereska and Gavves, 2024). This approach distinguishes itself through its ambition to *completely specify a neural network's computation*, enabling researchers to diagnose how models encode knowledge and execute task. In language models, mechanistic studies have uncovered computational mechanisms for syntactic processing (Hewitt and Manning, 2019) and factual recall (Elhage et al., 2022), often through techniques like probing and ablation. Central to this effort are gradient-based importance scores, which quantify parameter contributions to task performance by analyzing the sensitivity of the loss function to perturbations (Molchanov et al., 2019; Michel et al., 2019). These methods, validated in prior work, provide a practical means to identify critical parameters for specific abilities, bridging the gap between reverse engineering and actionable diagnostics.

Recent advances in mechanistic interpretability provide a foundation for critically assessing benchmark validity. While earlier studies focused on model behaviors (Yu and Ananiadou, 2024; Zhang et al., 2024b; Stolfo et al., 2023) (e.g., shortcut learning in arithmetic tasks (Geva et al., 2021)), our work introduces a novel application of these insights to evaluate whether benchmarks genuinely measure the abilities they claim to assess. By oper-

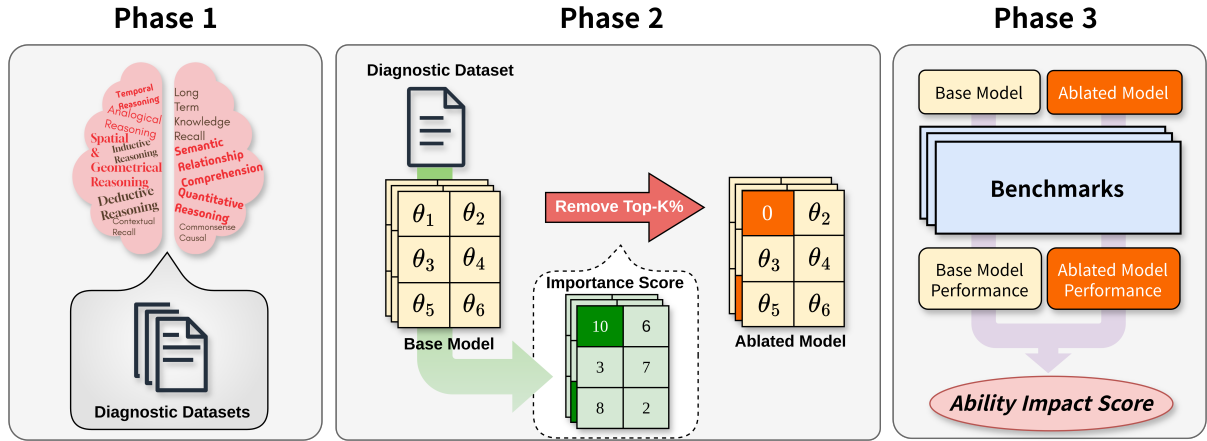


Figure 2: Three step pipeline of BENCHMARK PROFILING. **Left:** We define ten cognitively motivated abilities and create a dedicated diagnostic dataset for each one. **Middle:** Using the diagnostic dataset, we rank the base model’s parameters by gradient-based importance, and zero out (orange) the top k percent associated with that ability. **Right:** We run the original and ability-ablated models on downstream benchmarks. Their task accuracies yield the *Ability Impact Score* (AIS), which quantifies how strongly the benchmark depends on the ablated ability.

ationalizing cognitive dimensions and quantifying their impact through targeted parameter ablation, we demonstrate how mechanistic tools can diagnose mismatches between benchmark requirements and human-aligned competencies. This approach addresses a key gap in prior critiques, which identified limitations but lacked methodologies to systematically evaluate benchmark validity (Fodor, 2025; McIntosh et al., 2025), positioning mechanistic interpretability as a critical tool for advancing evaluation frameworks that reflect real-world capabilities.

3 Methodology

BENCHMARK PROFILING is a systematic methodology designed to diagnose the ability composition of LLM evaluation benchmarks. It quantifies the dependence of benchmarks on a predefined set of fundamental operationalized abilities by measuring the impact of selectively ablating ability-specific parameters within an LLM. The methodology comprises three main phases:

Phase 1: Defining Abilities

A cornerstone of BENCHMARK PROFILING is establishing a set of well-defined, fundamental abilities that serve as the diagnostic criteria. To address the *Performance-Perception Paradox* where models excel on benchmarks yet underperform in human-aligned contexts, we ground these criteria in established cognitive science frameworks (Laura, 2021; Beinborn and Hollenstein, 2024). By building on taxonomies like Cattell-Horn-Carroll (CHC)

theory (Carroll, 1993), which describes human cognitive abilities such as fluid reasoning and working memory, we ensure our operationalized abilities reflect dimensions humans intuitively recognize as markers of intelligence. This human-centric foundation bridges the gap between benchmark scores and the competencies users expect LLMs to exhibit in real-world interactions. In designing these 10 abilities found in Table 1, we balance theoretical robustness with practical relevance by adapting cognitive science principles to the context of LLM evaluation tasks, ensuring that each ability is both grounded in human cognition and directly applicable to benchmarking modern language models. While inspired by human cognition, these terms refer to specific, operationalized functional capacities within the LLM architecture. Detailed definitions are in Appendix A.

For each defined ability a , a diagnostic dataset D_a is created (2000 MCQs per ability in this work) which is designed to specifically measure that ability. Crucially, these datasets are validated in Section 6.1, and creation details are in Appendix B.

Phase 2: Identifying Abilities

This phase identifies specific components within the LLM, which are responsible for each defined abilities.

Within a chosen LLM (Θ), parameters critical for each dataset D_a are identified.

Importance Scoring We compute gradient-based importance scores $I_j^a(\theta)$ for each param-

Abbr.	Ability	Operationalization in Diagnostic Dataset
<i>Ana</i>	Analogical Reasoning	Present an analogy or proportional pair (e.g. $A:B :: C:?$) and ask which option best completes the relationship. Distractors ensure success requires mapping the underlying relation rather than surface word similarity.
<i>Com</i>	Commonsense & Causal Reasoning	Give a short everyday vignette and ask for the most plausible cause, effect, or next event; items hinge on everyday causal plausibility, not memorised facts.
<i>Cxt</i>	Contextual Recall	Provide a brief passage, then ask for verbatim details or their conjunction without new inference, isolating short-term textual memory.
<i>Ded</i>	Deductive Reasoning	Present premises that logically entail one conclusion; decoy options violate at least one logical step, forcing rule-based inference.
<i>Ind</i>	Inductive Reasoning	Show a short pattern or sequence and ask the model to infer the governing rule and extrapolate, so only rule discovery generalizes.
<i>LTK</i>	Long-Term Knowledge Recall	Ask about stored factual knowledge (history, science, geography) absent from the prompt; items use low-frequency facts to reduce chance memorization from local context.
<i>Quant</i>	Quantitative Reasoning	Pose a word problem with numerical data requiring arithmetic or counting; multi-step reasoning and distractor numbers discourage pattern matching.
<i>Sem</i>	Semantic Relationship Comprehension	Give a passage with several entities and ask about their roles or relations (e.g. part-whole, managerial hierarchy); questions test explicit and implicit links, not mere co-occurrence.
<i>Spat</i>	Spatial & Geometrical Reasoning	Describe spatial layouts or geometric facts, then ask about positions, directions, shapes, or distances; requires constructing a mental map or performing shape-based deductions.
<i>Temp</i>	Temporal Reasoning	Present events with time markers (dates, times, order words) and ask about sequence, simultaneity, or duration; items mix explicit and implicit cues to test chronology.

Table 1: Operationalized abilities and their abbreviations used in the BENCHMARK PROFILING framework.

ter θ_j using a first-order Taylor approximation of the loss $L(D_a, \theta)$ on dataset D_a (Molchanov et al., 2019; Michel et al., 2019).

$$I_j^a(\theta) \approx \left| \frac{\partial L(D_a, \theta)}{\partial \theta_j} \cdot \theta_j \right| \quad (1)$$

Gradients are obtained via fine-tuning on D_a . This fine-tuning is performed solely to facilitate accurate gradient estimation. The resulting model state is **discarded**.

Parameter Selection MLP layer parameters are ranked by $I_j^a(\theta)$, and the top-k% are selected as the parameter subset associated with ability a . For each ability a , an ablated model Θ^a is created by taking the original model Θ and setting the value of identified top-k% MLP parameters for ability a to zero. Preliminary experiments revealed that restricting ablations to MLP weights yields the clearest ability-specific signal with minimal collateral damage; see Section 6.2 for details.

Phase 3: Benchmark Profiling

This phase involves evaluating baseline and ablated models on target benchmarks, calculating the Ability

Impact Score (AIS) to normalize performance changes, and constructing the Benchmark Profile from these AIS values.

The original model Θ and each ablated model Θ^a are evaluated on target benchmarks b . Let baseline performance be $P_b(\Theta)$ and ablated performance be $P_b(\Theta^a)$.

To quantify benchmark reliance on each ability, we define the Ability Impact Score (AIS) for ability a on benchmark b , measuring the proportion of performance loss relative to the model’s baseline improvement over chance:

$$\text{AIS}_b^a = \frac{P_b(\Theta) - P_b(\Theta^a)}{P_b(\Theta) - P_b^{\text{chance}}} \quad (2)$$

where P_b^{chance} is chance-level performance for benchmark b . An AIS near 1 indicates strong dependence, while an AIS near 0 suggests little or no reliance. A negative AIS means that performance actually improves after the ability is ablated, signaling that the ability can be detrimental for that benchmark.

The calculated AIS values (AIS_b^a) are organized into the **Benchmark Profile**, providing a quan-

tative summary of each benchmark’s measured reliance on the defined operationalized abilities.

4 Experiments

This section details how BENCHMARK PROFILING is applied to derive Ability Impact Scores (AIS) for a suite of standard benchmarks. Section 4.1 explains the procedure for pinpointing ability-specific parameters and computing AIS values. Section 4.2 describes the experimental setup, including the language models, diagnostic datasets, and target benchmarks. The Benchmark Profiles produced by these experiments are interpreted in Section 5.

4.1 Experiment Design

Our goal is to quantify how strongly each benchmark in the curated suite depends on each of the ten operationalized abilities. For every ability we first rank model weights by gradient-based importance, then ablate the *top* 1.024 % of MLP parameters associated with that ability. A preliminary sweep across smaller and larger k values ranging from 0.001 % to 4.096 % showed that 1.024 % is the smallest budget that produces a clear, ability-specific signal without inflicting unnecessary collateral damage on unrelated capabilities. Applying this threshold yields ten ability-ablated models, each of which selectively disrupts one functional component while leaving the rest of the network, and its fluency, largely intact.

The core aim of our experiment design is to systematically measure how much each benchmark in our curated suite depends on each of the 10 operationalized abilities. For each ability, we identify the most critical model parameters using gradient-based importance scores, then create an ablated model by zeroing out the top 1.024% of MLP parameters associated with that ability. This process yields 10 ability-ablated models, each designed to selectively disrupt one functional component while leaving the rest of the model intact.

We systematically evaluated the baseline and ability-ablated models on our curated suite of 10 benchmarks, applying each benchmark’s standard evaluation metric. For every ability-benchmark pair, we computed the AIS as the normalized performance drop relative to the model’s improvement over chance, as formalized in Equation 2. This yields the Benchmark Profile, which quantifies the functional dependence of each benchmark on each operationalized ability. The Benchmark Profile

serves as the foundation for all subsequent analyses, providing a quantitative map of dependencies that we interpret in the following sections.

4.2 Experimental Setup

This section outlines the core components used in our experiments: the LLMs subjected to profiling, the diagnostic datasets developed to isolate specific abilities, and the suite of benchmarks selected for analysis, including details on their evaluation.

Models Our primary experiments leverage Llama-3.1-8B-Instruct (Meta, 2024), a widely recognized instruction-tuned model. To assess the generalizability of our findings, robustness checks replicate key analyses on two additional models: Qwen/Qwen2.5-7B-Instruct (Team, 2024) and mistralai/Mistral-7B-Instruct-v0.3 (Jiang et al., 2024). All models are used in their base precision (BF16) and evaluated using greedy decoding, consistent with common practices, unless otherwise specified by a benchmark’s standard protocol.

Diagnostic Datasets We employ the 10 diagnostic datasets designed to target the operationalized abilities defined in Section 3 and listed in Table 1. Each dataset consists of 2000 4-choice Multiple-Choice Questions (MCQs), totaling 20,000 examples. These datasets were synthetically generated using the o4-mini-2025-04-16 API via carefully crafted few-shot prompting strategies tailored to each ability. Detailed descriptions of the generation prompts and examples for each ability dataset are provided in Appendix B. The validation of these datasets is presented in Section 6.1.

Benchmark Details For the application of BENCHMARK PROFILING, we selected the curated suite of 10 standard LLM evaluation tasks. This suite was chosen to encompass a variety of task formats and evaluation paradigms common in LLM assessment.

The selected benchmarks include several multiple-choice question-answering tasks: ARC-Challenge (Clark et al., 2018) which uses a 4-choice format; HellaSwag (Zellers et al., 2019), also 4-choice, requiring sentence completion; WinoGrande (Sakaguchi et al., 2021), a 2-choice pronoun resolution task; CommonsenseQA (Talmor et al., 2019), a 5-choice QA task; LogiQA (Liu et al., 2020), a 4-choice QA over logical passages; and BIG-Bench Logical Deduction (Srivastava et al., 2022), a 5-choice task.

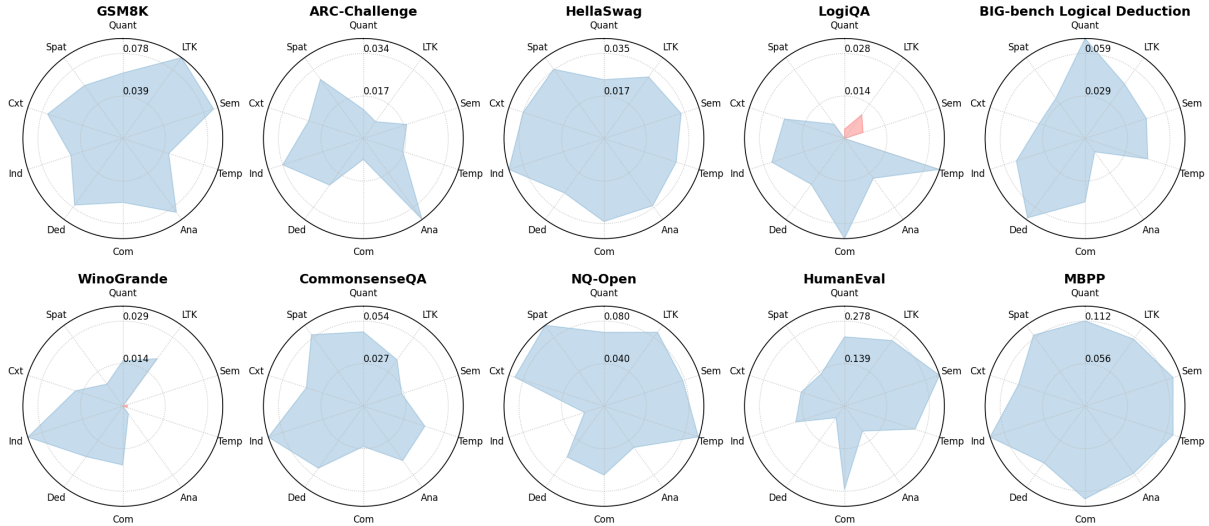


Figure 3: Ability Impact Score radar plots for the ten benchmarks profiled on Llama-3.1-8B-Instruct. Axes are labeled with the following abbreviated abilities. Blue and red shading indicates positive and negative AIS values.

The suite also incorporates generation tasks. GSM8K (Cobbe et al., 2021) requires generating a chain-of-thought leading to a final numerical answer, which is then matched for evaluation. Natural Questions Open (NQ-Open) (Kwiatkowski et al., 2019) is an open-domain QA task where short generated answers are evaluated by exact match. For coding, HumanEval (Chen et al., 2021) and MBPP (Mostly Basic Python Problems) (Austin et al., 2021) require the model to generate Python code, which is then evaluated for functional correctness using a pass@1 metric.

To ensure consistency and facilitate reproducible evaluations across this diverse suite, we utilized the EleutherAI Language Model Evaluation Harness (Gao et al., 2024) for executing the benchmark tasks and collecting performance metrics. For each benchmark, we adhere to its standard evaluation protocol and primary metric. These performance scores are subsequently used to calculate the AIS as defined in Section 3 Phase 2. The chance-level performance P_b^{chance} for each benchmark, critical for the AIS calculation, is determined by its specific format (e.g., 0.25 for 4-choice MCQs, 0.5 for 2-choice, 0.2 for 5-choice, and 0 for generation tasks). This diverse set of task formats and evaluation approaches allows us to investigate how ability dependencies manifest across different interaction and assessment modalities.

5 Main Results

This section interprets the benchmark profile of Llama-3.1-8B-Instruct. We visualize AIS the pat-

tern with radar plots and compare cross-model similarity with Jensen–Shannon statistics. We then highlight four empirical observations, showing (i) that popular benchmarks exercise multiple abilities rather than a single labeled skill, (ii) that seemingly related datasets often reward very different mixtures of abilities, (iii) that code-generation tasks demand the broadest spectrum of abilities and therefore penalize narrow fine-tuning, and (iv) that certain abilities can act as distractors on tightly constrained reasoning tasks. The remainder of the section presents the visual evidence and discusses each *Key Finding* in detail. Detailed AIS matrix can be found in Table 3.

Key Finding 1: Benchmarks Combine Multiple Abilities Figure 3 reveals that every benchmark draws on a rich mixture of skills. HellaSwag and MBPP shows a broad footprint, while WinoGrande, the narrowest profile, still includes on more than one competency. **GSM8K** peaks in *Long-Term Knowledge Recall* and *Semantic Relationship*, while *Quantitative Reasoning* is only moderate. **ARC-Challenge** centers on *Analogical* and *Inductive Reasoning* with minimal *Long Term Knowledge Recall*. **LogiQA**, marketed as a logical reasoning benchmark, in fact leans most on *Temporal Reasoning* and *Commonsense Causal Reasoning*, with *Deductive Reasoning* contributing only modestly. These composite patterns confirm that task labels such as *math* or *logic* under-specify what is really being measured.

Key Finding 2: Benchmarks with Similar Labels

Test Different Abilities Figure 3 compares two question-answering datasets that are often grouped under *knowledge QA* yet rely on markedly different skill mixes. **CommonsenseQA**, a 5-choice multiple-choice benchmark, peaks in *Inductive Reasoning* and draws secondary support from *Deductive Reasoning*, *Spatial Reasoning*, *Quantitative Reasoning*, and *Analogical Reasoning*. In contrast, **Natural Questions Open** (NQ-Open), an open-ended retrieval task, scores highest on *Temporal Reasoning*, *Spatial Reasoning*, *Semantic Relationship*, *Long-Term Knowledge*, and *Contextual Recall*.

These divergent ability footprints translate into sharply different accuracies: Llama-3.1-8B-Instruct answers 77.1% of CommonsenseQA items correctly yet attains only an 17.9% exact-match rate on NQ-Open; Qwen2.5-7B-Instruct shows a similar contrast (82.7% vs. 4.7%). Even after accounting for the easier multiple-choice format of CommonsenseQA, the gap remains large. Such crossed scores illustrate how a model can excel on one *knowledge QA* benchmark while struggling on another that depends on a different blend of abilities, underscoring the diagnostic value of BENCHMARK PROFILING.

Key Finding 3: Code Benchmarks Demand

Broad Skill Sets The two bottom right panels of Figure 3 show that **HumanEval** and **MBPP** produce the largest AIS values, indicating that success depends on many abilities at once. HumanEval is driven most by *Semantic Relationship*, aligning with the need to interpret function specifications precisely, whereas MBPP lights up almost every axis forming an almost complete disk.

The wide spread of AIS values for MBPP aligns with findings that coding datasets inherently correlate with multiple reasoning abilities due to their structured, logic-driven nature (Zhang et al., 2024a). This mutual reinforcement has been shown evident in training dynamics: models exposed to code data not only excel at programming tasks but also exhibit enhanced performance on mathematical and logical reasoning benchmarks (Ma et al., 2023; Tao et al., 2024). This correlation between code and broad spectrum of reasoning abilities explains why MBPP’s profile lights up nearly every ability axis. The structured syntax and semantic precision required in coding tasks act as a scaffold for multitask learning, reinforcing skills like

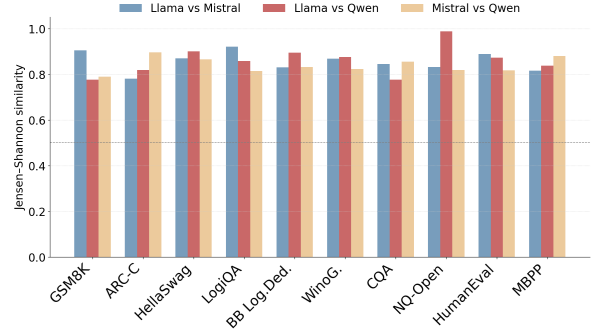


Figure 4: Jensen–Shannon Similarity after min-max normalization. Each bar compares two models on a single benchmark.

deductive reasoning and contextual recall that are critical for both programming and general problem-solving.

Key Finding 4: Irrelevant Abilities Can Hurt

Performance Figure 3 exposes a small but consistent pocket of *negative AIS* values shown in red: ablating *Long-Term Knowledge*, *Semantic Relationship*, or *Quantitative Reasoning* increases LogiQA accuracy by 1–2 percentage points, and removing *Temporal Reasoning* or *Semantic Relationship* yields a similar boost on WinoGrande. This pattern is absent from the other eight benchmarks, indicating that negative transfer emerges only when the dataset contains spurious cues that conflict with its core reasoning chain. LogiQA is explicitly constructed so that the correct answer is derivable *only* from the supplied premises; introducing external factual recall or numerical heuristics therefore lures a model toward plausible-but-invalid distractors (Liu et al., 2020). WinoGrande was adversarially re-balanced to neutralize superficial lexical biases, forcing systems to rely on fine-grained syntactic cues; augmenting the model with world knowledge or event-ordering heuristics re-introduces precisely the shortcut signals the benchmark was designed to suppress (Sakaguchi et al., 2021; McCoy et al., 2019). More broadly, multi-task learning research shows that adding tasks or features weakly correlated with the gold decision boundary can hurt generalization, a phenomenon known as *negative transfer* (Zhang et al., 2022). Recent mechanistic and robustness studies echo this observation, demonstrating that *adding* external knowledge or auxiliary data can introduce spurious correlations that *degrade* downstream logical-reasoning accuracy (Schuff et al., 2021; Compton et al., 2023). Because the remaining benchmarks either reward

Ablated Ability	GSM8K			NQ-Open		
	Base	MLP	All	Base	MLP	All
Contextual Recall	0.773	0.7354	0.1024	0.1789	0.1202	0.0374
Quantitative Reasoning		0.7422	0.0902		0.1357	0.0163

Table 2: Top-1.024 % ablation applied either to MLP weights only (*MLP*) or to all weights (*All*). Results are reported as exact-match accuracy.

those auxiliary abilities or embed them in ways that align with the task objective, ablating them provides no systematic benefit, hence negative AIS values appear only for LogiQA and WinoGrande.

Robustness Across Models After min-max-normalizing each benchmark column of the AIS matrix, we measure agreement with *Jensen–Shannon Similarity (JSS)*, which is derived from the Jensen-Shannon Divergence. For two discrete distributions p and q , it is expressed as:

$$\text{JSS}(p, q) = 1 - \frac{1}{2}(\text{D}_{\text{KL}}(p \parallel m) + \text{D}_{\text{KL}}(q \parallel m))$$

where $m = \frac{1}{2}(p + q)$ and D_{KL} is Kullback–Leibler divergence. By construction, $\text{JSS}(p, q) \in [0, 1]$, with higher values indicating greater similarity.

Figure 4 plots the pairwise similarities for all ten benchmarks. Every bar is above the gray 0.5 reference line (range 0.53–0.89, mean 0.64), indicating that Llama-3.1-8B, Mistral-7B, and Qwen-2.5 share broadly consistent ability footprints despite architectural differences.

6 Validation of Methodology Components

We validate our method in two ways, first by having human experts confirm that each diagnostic dataset genuinely targets its stated ability, and second by demonstrating that ablating only MLP parameters weakens the intended skills while preserving overall model fluency better than ablating all layers.

6.1 Expert Evaluation of the Diagnostic Datasets

To confirm that each prompt truly targeted its intended ability, we asked ten independent domain specialists to review a stratified sample of items from every ability category (see Appendix C for annotator demographics and instructions). Each expert saw the *context*, *question*, and *answer* for every item, then (i) selected which of the ten ability labels best described the required skill and (ii) judged whether the item fit that definition. Experts

selected the correct label in **92.2 %** of cases (individual ability accuracies: 74, 88, 92, 92, 92, 94, 94, 98, 98, 100), confirming that the items faithfully captured their intended skills.

6.2 MLP-Only Ablation

Zeroing attention weights as well as MLP weights quickly dismantles the self-attention pathways that bind tokens into a coherent context that supports coherent text. Table 2 contrasts the two ablation regimes on Llama-3.1-8B-Instruct for the *Contextual Recall* and *Quantitative Reasoning* abilities. The *MLP-only* variant yields only modest accuracy drops, whereas the *all-layer* variant slashes performance on every setting in the table, confirming that attention layer damage wipes out far more capability than is needed for diagnostic purposes. Appendix E shows generation examples that match these numbers: the MLP-only model stays fluent, whereas the all-layer model lapses into repetitive, incoherent text.

7 Conclusion

Current benchmark tasks often obscure which skills a language model actually employs during evaluation, making it difficult to know when a reported gain reflects robust problem-solving ability or an exploitable shortcut. We introduce BENCHMARK PROFILING, a systematic framework that decomposes benchmark performance into ten operationalized abilities grounded in cognitive science. By combining gradient-based importance scoring, targeted parameter ablations, and the Ability Impact Score, our method delivers an interpretable ability fingerprint for every benchmark–model pair.

Experiments applying BENCHMARK PROFILING to widely used models and benchmarks uncovered patterns indicating that most benchmarks tap several underlying abilities, tasks with the same label often depend on different ability blends, code-generation benchmarks reward broad multi-skill competence rather than narrow domain tuning, and adding abilities a task does not truly demand can even reduce performance. These insights clarify why leaderboard gains sometimes fail to translate into practical capability.

BENCHMARK PROFILING thus provides researchers and practitioners with transparent diagnostics, enabling better-aligned model evaluations, targeted improvements in model design, and more accurate interpretations of benchmark results.

Limitations

Synthetic diagnostics All probing datasets are synthetic; their generation templates and few-shot examples are listed in Appendix B, and domain experts confirmed their *face validity* in Appendix C.

Model scale and compute All experiments use three open models: Llama-3.1-8B, Qwen-2.5-7B, and Mistral-7B. For each ability we distributed the gradient-importance computation across eight NVIDIA A100-80 GB GPUs, which completed in about 25 minutes. The subsequent weight-zeroing step ran on a single A100-80 GB GPU and finished in roughly 5 minutes. Thus profiling one model over ten abilities plus downstream benchmark evaluation fits comfortably within a few GPU-hours. Profiling models beyond the 7–8 B range may still require memory-efficient techniques such as gradient check-pointing.

Licensing and intended use The diagnostic datasets and code will be released under the MIT License for research and non-commercial use. They are not intended for high-stakes deployment or for ranking commercial systems without additional validation.

Documentation We provide full data statistics, generation templates, and class labels in Appendix B. A README.md with installation and reproduction scripts will accompany the code repository.

Ethics Statement

Data privacy and content All diagnostic items are generated from templated prompts and contain no personal or identifying details. Volunteers manually screened a random sample and reported no offensive content (Appendix C).

Benchmark licenses We rely only on benchmarks released under permissive licenses: ARC-Challenge, CommonsenseQA, GSM8K, HellaSwag, HumanEval, LogiQA, MBPP, Natural Questions Open, WinoGrande, and BIG-Bench Logical Deduction. Our use remains within each dataset’s original research intent.

Synthetic artifact release To maintain anonymity during review, the diagnostic datasets, generation scripts, and validation labels will be placed in a public GitHub repository once the paper is accepted. They will be distributed under the CC-BY-SA-4.0 license; accompanying code

will use the MIT license. The README file will specify intended research use and disclaim commercial deployment without additional validation.

Potential misuse Knowing how benchmarks decompose into abilities could, in theory, help actors craft adversarial tests or game leaderboard metrics. We consider this risk low because reproducing our pipeline requires non-trivial compute, and transparency ultimately benefits the community by exposing hidden shortcuts.

Human subjects Ten adult volunteers participated in item validation. No personal data were collected or stored beyond coarse demographics. Details are in Appendix C.

References

- John R Anderson. 2013. *The architecture of cognition*. Psychology Press.
- John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An integrated theory of the mind. *Psychological review*, 111(4):1036.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. [The vulnerability of language model benchmarks: Do they accurately reflect true llm performance?](#) *Preprint*, arXiv:2412.03597.
- Lisa Beinborn and Nora Hollenstein. 2024. *Cognitive plausibility in natural language processing*. Springer.
- Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety—a review. *arXiv preprint arXiv:2404.14082*.
- Samuel R. Bowman and George E. Dahl. 2021. [What will it take to fix benchmarking in NLP?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2473–2487, Online. Association for Computational Linguistics.
- John Bissell Carroll. 1993. *Human cognitive abilities: A survey of factor-analytic studies*. 1. Cambridge university press.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg

698	Brockman, and 1 others. 2021. Evaluating large	Clémentine Fourrier, Nathan Habib, Alina Lozovskaya,	746
699	language models trained on code. <i>arXiv preprint</i>	Konrad Szafer, and Thomas Wolf. 2024. Open llm	747
700	<i>arXiv:2107.03374</i> .	leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard .	748
701	Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anasta-		749
702	sios Nikolas Angelopoulos, Tianle Li, Dacheng Li,	Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman,	750
703	Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E	Sid Black, Anthony DiPofi, Charles Foster, Laurence	751
704	Gonzalez, and 1 others. 2024. Chatbot arena: An	Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li,	752
705	open platform for evaluating llms by human pref-	Kyle McDonell, Niklas Muennighoff, Chris Ociepa,	753
706	erence. In <i>Forty-first International Conference on</i>	Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya	754
707	<i>Machine Learning</i> .	Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness .	755
708	Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,		756
709	Ashish Sabharwal, Carissa Schoenick, and Oyvind	Dedre Gentner. 1983. Structure-mapping: A theoretical	757
710	Tafjord. 2018. Think you have solved AI? try the	framework for analogy. <i>Cognitive science</i> , 7(2):155–	758
711	ARC AI reasoning challenge . In <i>AAAI Workshop on</i>	170.	759
712	<i>Reasoning for Complex Question Answering</i> .		
713	Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,	Mor Geva, Daniel Khashabi, Elad Segal, Jonathan Berant,	760
714	Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias	and Ido Dagan. 2021. Aristotle: A dataset for focused	761
715	Plappert, Jerry Tworek, Jacob Hilton, Reiichiro	logical reasoning . In <i>Findings of the Association for</i>	762
716	Nakano, Christopher Hesse, and John Schulman.	<i>Computational Linguistics: EMNLP 2021</i> , pages 1–14,	763
717	2021. Training verifiers to solve math word prob-	Punta Cana, Dominican Republic. Association for Com-	764
718	lems . In <i>Proceedings of the International Conference</i>	putational Linguistics.	765
719	<i>on Learning Representations (ICLR)</i> .		
720	Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,	766
721	Ranganath. 2023. When more is less: Incorporating	Mantas Mazeika, Dawn Song, and Jacob Steinhardt.	767
722	additional datasets can hurt performance by introduc-	2021. Measuring massive multitask language under-	768
723	ing spurious correlations. In <i>Machine Learning for</i>	standing. In <i>Proceedings of the International Confer-</i>	769
724	<i>Healthcare Conference</i> , pages 110–127. PMLR.	<i>ence on Learning Representations (ICLR)</i> .	770
725	Joseph Cummings and Jason Wilson. 2019. Clark at	John Hewitt and Christopher D. Manning. 2019. A struc-	771
726	semeval-2019 task 3: exploring the role of context to	tural probe for finding syntax in word representations .	772
727	identify emotion in a short conversation. In <i>Proceed-</i>	In <i>Proceedings of the 2019 Conference of the North</i>	773
728	<i>ings of the 13th International Workshop on Semantic</i>	<i>American Chapter of the Association for Computational</i>	774
729	<i>Evaluation</i> , pages 159–163.	<i>Linguistics: Human Language Technologies, Volume 1</i>	775
730	Meredyth Daneman and Patricia A Carpenter. 1980.	(<i>Long and Short Papers</i>), pages 4129–4138, Minneapo-	776
731	Individual differences in working memory and read-	lis, Minnesota. Association for Computational Linguis-	777
732	ing. <i>Journal of verbal learning and verbal behavior</i> ,	tics.	778
733	19(4):450–466.	John H Holland. 1986. <i>Induction: Processes of inference,</i>	779
	Nelson Elhage, Tristan Hume, Catherine Olsson,	<i>learning, and discovery</i> . MIT press.	780
	Nicholas Schiefer, Tom Henighan, Shauna Kravec,	Keith J Holyoak. 2012. 13 analogy and relational reason-	781
	Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain,	ing. <i>The Oxford handbook of thinking and reasoning</i> ,	782
	Carol Chen, Roger Grosse, Sam McCandlish, Jared	page 234.	783
	Kaplan, Dario Amodei, Martin Wattenberg, and	Albert Jiang, Alexandre Sablayrolles, Arthur Mensch,	784
	Christopher Olah. 2022. Toy models of superposition.	Chris Bamford, Devendra Singh Chaplot, Diego de las	785
	<i>Transformer Circuits Thread</i> . https://transformer-	Casas, Emma Rolland, Gergely Lengyel, Guillaume	786
	<i>circuits.pub/2022/toy_model/index.html</i> .	Lample, Jia Han, Jian Zhu, Thibaut Lavril, Thomas	787
734	Randall W Engle and Michael J Kane. 2004. Executive	Lacroix, Louis Martin, Ragheb El Sayed, Edouard	788
735	attention, working memory capacity, and a two-factor	Grave, Ahmed Sultan, Antoinette Savary, Bilal Ayed,	789
736	theory of. <i>The psychology of learning and motivation:</i>	and 18 others. 2024. Mistral large. <i>arXiv preprint</i>	790
737	<i>Advances in research and theory</i> , 44:145.	<i>arXiv:2402.19455</i> .	791
738	Maria Eriksson, Erasmo Purificato, Arman Noroozian,	Philip N Johnson-Laird. 2001. Mental models and deduc-	792
739	Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and	tion. <i>Trends in cognitive sciences</i> , 5(10):434–442.	793
740	David Fernandez-Llorca. 2025. Can we trust ai bench-		
741	marks? an interdisciplinary review of current issues in	Michael J Kane and Randall W Engle. 2002. The role	794
742	ai evaluation. <i>arXiv preprint arXiv:2502.06559</i> .	of prefrontal cortex in working-memory capacity, ex-	795
743	James Fodor. 2025. Line goes up? inherent limitations	ecutive attention, and general fluid intelligence: An	796
744	of benchmarks for evaluating large language models .	individual-differences perspective. <i>Psychonomic bul-</i>	797
745	<i>Preprint</i> , arXiv:2502.14318.	<i>letin & review</i> , 9(4):637–671.	798

799	Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield,	Paul Michel, Omer Levy, and Graham Neubig. 2019. Are	851
800	Michael Collins, Ankur Parikh, Chris Alberti, Danielle	sixteen heads really better than one? In <i>Advances in</i>	852
801	Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee,	<i>Neural Information Processing Systems</i> .	853
802	and 1 others. 2019. Natural questions: a benchmark		
803	for question answering research. <i>Transactions of the</i>	George A Miller. 1995. Wordnet: a lexical database for	854
804	<i>Association for Computational Linguistics</i> , 7:453–466.	english. <i>Communications of the ACM</i> , 38(11):39–41.	855
805	Patrick C Kyllonen. 2021. Taxonomy of cognitive abil-	Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio,	856
806	ities and measures for assessing artificial intelligence	and Jan Kautz. 2019. Importance estimation for neural	857
807	and robotics capabilities. <i>AI and the Future of Skills,</i>	network pruning. <i>Proceedings of the IEEE/CVF Con-</i>	858
808	<i>Volume</i> , page 50.	<i>ference on Computer Vision and Pattern Recognition</i>	859
		(CVPR).	860
809	Brenden Lake and Marco Baroni. 2018. Generalization	Klaus Oberauer and Stephan Lewandowsky. 2016. Con-	861
810	without systematicity: On the compositional skills of	trol of information in working memory: Encoding and	862
811	sequence-to-sequence recurrent networks. In <i>Interna-</i>	removal of distractors in the complex-span paradigm.	863
812	<i>tional conference on machine learning</i> , pages 2873–	<i>Cognition</i> , 156:106–128.	864
813	2882. PMLR.		
814	DOCKINGS Laura. 2021. Guidance on key considerations	Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. 2024.	865
815	for the identification and selection of safer chemical	The generative AI paradox in evaluation: “what it can	866
816	alternatives.	solve, it may not evaluate” . In <i>Proceedings of the 18th</i>	867
		<i>Conference of the European Chapter of the Association</i>	868
817	Aitor Lewkowycz, Anders Andreassen, David Dohan,	<i>for Computational Linguistics: Student Research Work-</i>	869
818	Ethan Dyer, Henryk Michalewski, Vinay Ramasesh,	<i>shop</i> , pages 248–257, St. Julian’s, Malta. Association	870
819	Ambrose Slone, Cem Anil, Imanol Schlag, Theo	for Computational Linguistics.	871
820	Gutman-Solo, and 1 others. 2022. Solving quantitative		
821	reasoning problems with language models. <i>Advances in</i>	Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton	872
822	<i>Neural Information Processing Systems</i> , 35:3843–3857.	Bakhtin, Yuxiang Wu, Alexander H Miller, and Se-	873
		bastian Riedel. 2019. Language models as knowledge	874
		bases? <i>arXiv preprint arXiv:1909.01066</i> .	875
823	Percy Liang and 1 others. 2022. Holistic evaluation of	J Raven. 1939. Progressive matrices: A perceptual test of	876
824	language models. In <i>arXiv preprint arXiv:2211.09110</i> .	intelligence, hk lewis, london.	877
825	Stephanie Lin, Jacob Hilton, and Owain Evans. 2021.	Adam Roberts, Colin Raffel, and Noam Shazeer. 2020.	878
826	Truthfulqa: Measuring how models mimic human false-	How much knowledge can you pack into the parameters	879
827	hoods. <i>arXiv preprint arXiv:2109.07958</i> .	of a language model? <i>arXiv preprint arXiv:2002.08910</i> .	880
828	Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile	Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula,	881
829	Wang, and Yue Zhang. 2020. Logiqa: A challenge	and Yejin Choi. 2021. Winogrande: An adversarial	882
830	dataset for machine reading comprehension with logical	winograd schema challenge at scale. <i>Communications</i>	883
831	reasoning. <i>arXiv preprint arXiv:2007.08124</i> .	<i>of the ACM</i> , 64(9):99–106.	884
832	Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang,	Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi,	885
833	Changjian Wang, and Shanshan Li. 2023. At which	and Dan Roth. 2020. Commonsense reasoning for nat-	886
834	training stage does code data help llms reasoning?	ural language processing. In <i>Proceedings of the 58th</i>	887
835	<i>Preprint</i> , arXiv:2309.16298.	<i>annual meeting of the association for computational</i>	888
		<i>linguistics: Tutorial abstracts</i> , pages 27–33.	889
836	Kannan Mani and Philip N Johnson-Laird. 1982. The	W Joel Schneider and Kevin S McGrew. 2012. The cattell-	890
837	mental representation of spatial descriptions. <i>Memory</i>	horn-carroll model of intelligence.	891
838	<i>& Cognition</i> , 10(2):181–187.		
839	R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019.	W Joel Schneider and Kevin S McGrew. 2018. The cattell-	892
840	Right for the wrong reasons: Diagnosing syntactic	horn-carroll theory of cognitive abilities. <i>Contempo-</i>	893
841	heuristics in natural language inference. <i>arXiv preprint</i>	<i>rary intellectual assessment: Theories, tests, and issues</i> ,	894
842	<i>arXiv:1902.01007</i> .	733:163.	895
843	Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage,	Hendrik Schuff, Hsiu-Yu Yang, Heike Adel, and	896
844	Tong Liu, Dan Xu, Paul Watters, and Malka N Hal-	Ngoc Thang Vu. 2021. Does external knowledge	897
845	gamuge. 2025. Inadequacies of large language model	help explainable natural language inference? auto-	898
846	benchmarks in the era of generative artificial intelli-	matic evaluation vs. human ratings. <i>arXiv preprint</i>	899
847	gence . <i>IEEE Transactions on Artificial Intelligence</i> ,	<i>arXiv:2109.07833</i> .	900
848	page 1–18.	Bibhya Nand Sharma, Aluwesi Volau Fonolahi, Akeshnil	901
849	AI @ Meta. 2024. Introducing llama 3.1: Our most capa-	Bali, and Swasti Shubha Narayan. 2019. The online	902
850	ble models to date. Accessed: April 25, 2025.	mathematics diagnostic tool for transformative learning	903
		in the pacific. In <i>Cases on smart learning environments</i> ,	904
		pages 63–80. IGI Global Scientific Publishing.	905

906
907
908

909
910
911

912
913
914
915

916
917
918
919
920
921
922
923
924

925
926
927
928
929

930
931

932
933
934

935
936
937
938

939
940
941
942
943

944
945
946

947
948
949
950
951

952
953
954

955
956

957
958

Steven Sloman and Steven A Sloman. 2009. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Aarohi Srivastava and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In *arXiv preprint arXiv:2206.04615*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. [A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis](#). *Preprint*, arXiv:2305.15054.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianhua Tao, Junbo Li, Bowen Tan, Hongyi Wang, William Marshall, Bhargav M Kanakiya, Joel Hestness, Natalia Vassilieva, Zhiqiang Shen, Eric P. Xing, and Zhengzhong Liu. 2024. [Crystal: Illuminating llm abilities on language and code](#). *Preprint*, arXiv:2411.04156.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2406.04721*.

Eva Tideman and Jan-Eric Gustafsson. 2004. Age-related differentiation of cognitive abilities in ages 3–7. *Personality and Individual Differences*, 36(8):1965–1974.

Zeping Yu and Sophia Ananiadou. 2024. [Interpreting arithmetic mechanism in large language models through comparative neuron analysis](#). *Preprint*, arXiv:2409.14144.

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

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. 2022. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329.

Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. 2024a. [Unveiling the impact of coding data instruction fine-tuning on large language models reasoning](#). *Preprint*, arXiv:2405.20535.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. [Unveiling linguistic regions in large language models](#). *Preprint*, arXiv:2402.14700.

A Operationalized Ability Definitions and Diagnostic Task Principles

This appendix justifies the ten abilities used in BENCHMARK PROFILING, situates each one within

the Cattell–Horn–Carroll (CHC) model of intelligence (Carroll, 1993; Schneider and McGrew, 2012), and explains how the corresponding synthetic diagnostic dataset was constructed.

Human cognition is *distributed*: higher level skills co-recruit multiple lower level processes, and narrow processes are re-used across domains (Anderson, 2013; Oberauer and Lewandowsky, 2016). CHC therefore models abilities as *correlated but separable* factors rather than mutually exclusive boxes. In the same spirit, our ten abilities were by design, chosen to be *distinct enough* to yield interpretable weight profiles yet *not so orthogonal* that they ignore real cognitive interactions. Mild overlap is expected and even desirable: it lets our ablation analysis reveal which *mixtures* of skills a benchmark rewards. What matters empirically is that each diagnostic dataset is *maximally diagnostic* for its target ability so that the gradient-importance procedure reliably tags a concentrated slice of parameters. The robustness of the Ability-Impact profiles across three models (Section 5) supports this assumption.

Below, each ability entry follows the same template: (i) cognitive-science grounding and CHC slot, (ii) a motivating example, and (iii) how the dataset was generated to isolate that skill.

Analogical Reasoning (CHC: *Gf*–Induction). A proportional analogy such as *bird, : nest, :: bee, : ?* demands mapping a relational schema rather than surface similarity; Raven’s Progressive Matrices and related tasks tap the same faculty (Raven, 1939; Gentner, 1983; Holyoak, 2012).

We authored four prompt templates that supply an *A:B :: C:?* stem and four distractors. Distractors are chosen by perturbing either A or B to share lexical or semantic features without preserving the relation (e.g., *hive*, (*correct*) vs. *honey*, *sting*, *wasp*). This forces the model to attend to the latent mapping.

Commonsense & Causal Reasoning (CHC: *Gf* + script knowledge). Inferring that a neglected plant will wilt integrates causal schemas learned from everyday experience (Sloman and Sloman, 2009; Sap et al., 2020).

Each question describes a three-to-five sentence vignette drawn from diverse domains (kitchen accidents, school routines, etc.). We then ask for the most plausible cause or effect, sampling distractors from unrelated but thematically similar events to eliminate superficial cueing. Scenarios were gen-

erated by large-model completion and manually filtered for obvious lexical shortcuts.

Contextual Recall (CHC: *Gsm*). Working-memory span underpins reading comprehension (Daneman and Carpenter, 1980; Kane and Engle, 2002).

Two template families were used: (i)*single-fact* passages of 2–3 sentences followed by a verbatim retrieval question, and (ii)*multi-hop* passages of 4–6 sentences where the queried detail is the conjunction of two facts stated far apart. All answers are extractive so no external knowledge is useful.

Deductive Reasoning (CHC: *Gf*–Sequential Reasoning). Classical syllogisms illustrate rule-based deduction; accuracy correlates with measures of logical capacity (Johnson-Laird, 2001).

Premises are generated by a symbolic template engine that instantiates first-order logic patterns (e.g., *All S are P; No P are R; therefore ?*). Distractors violate exactly one rule to ensure that only a valid derivation succeeds.

Inductive Reasoning (CHC: *Gf*–Induction). Discovering hidden regularities in sequences is central to hypothesis formation (Holland, 1986; Lake and Baroni, 2018).

We mine integer, geometric, and lexical pattern families (arithmetic progression, polygon naming, etc.). For each instance we sample five in-context elements and ask for the sixth. Distractors follow decoy rules (e.g., additive offset vs. multiplicative) to penalise surface heuristics.

Long-Term Knowledge Recall (CHC: *Glr*). Retrieving stored facts such as *Canberra is Australia’s capital* maps to *Glr* in CHC and has been probed extensively in LLMs (Petroni et al., 2019; Roberts et al., 2020).

We queried Wikidata for low-frequency entities, then generated four-choice trivia questions via a templating script. We discard items whose answer string appears verbatim in the question to curb lexical leakage.

Quantitative Reasoning (CHC: *Gq* + numeric *Gf*). Multi-step word problems activate both quantitative knowledge and fluid reasoning (Cobbe et al., 2021; Lewkowycz et al., 2022).

Templates embed 3–5 numbers, at least one of which is irrelevant, and require two operations (addition then division, etc.). Distractor an-

swers are produced by common student errors (off-by-one, wrong operator) as recommended by math-education literature (Sharma et al., 2019).

Semantic Relationship Comprehension (CHC: *Gc*). Understanding taxonomical and role relations underlies lexical semantics (Miller, 1995; Cummings and Wilson, 2019).

Each passage introduces 3–4 named entities in a mini-scenario (corporate hierarchy, biological taxonomy). We ask about an implicit relationship (*Who is Alice to Charlie?*) while distractors share topical words but break the relation type.

Spatial & Geometrical Reasoning (CHC: *Gv*). Textual spatial reasoning engages mental imagery and visuospatial sketchpad resources (Mani and Johnson-Laird, 1982).

We generate short descriptions of object layouts on a 3×3 grid and ask queries like *Which object is directly below the circle?* Distractors include objects that are correct under mirror-flipped or rotated interpretations, so success requires consistent coordinate mapping.

Temporal Reasoning (CHC: sequencing facet of *Gf*). Temporal sequencing develops early and is essential for narrative comprehension (Anderson et al., 2004).

Templates mention explicit times, durations, or adverbial order cues; questions ask which event came first, lasted longer, or overlapped. Distractors are derived by permuting the true order.

Potential Overlaps and Taxonomy Limits

Because CHC factors are *correlated* (e.g., *Gf* tasks share variance with *Gsm*), perfect separation is neither feasible nor theoretically warranted (Tideman and Gustafsson, 2004; Engle and Kane, 2004). We therefore aim for *dominant association*: each diagnostic set should be *most strongly* but not exclusively linked to its labeled ability. Empirically, the distinct weight clusters uncovered by gradient analysis and the reproducible Benchmark Profiles across models (Figure 4) confirm that residual overlap does not hinder interpretability. Indeed, partial overlap lets the Benchmark Profile capture real-world skill synergies, aligning with evidence that complex tasks draw on multiple CHC factors simultaneously (Schneider and McGrew, 2018).

Llama-3.1-8B-Instruct (k=1.024% MLP Ablation)										
Ablated Ability	GSM8K	ARC-C	HellaSwag	LogiQA	BB Log.Ded.	WinoG.	CQA	NQ-Open	HumanEval	MBPP
Analogical Reasoning	0.0833	0.0398	0.0337	0.0163	0.0114	0.0032	0.0426	0.0480	0.1006	0.1090
Commonsense Causal	0.0583	0.0083	0.0337	0.0332	0.0439	0.0198	0.0256	0.0648	0.2730	0.1220
Contextual Recall	0.0723	0.0229	0.0345	0.0210	0.0334	0.0168	0.0381	0.0884	0.1494	0.0922
Deductive Reasoning	0.0750	0.0229	0.0273	0.0188	0.0678	0.0210	0.0486	0.0591	0.0469	0.0922
Inductive Reasoning	0.0499	0.0338	0.0406	0.0254	0.0500	0.0337	0.0635	0.0193	0.1678	0.1316
Long Term Knowledge	0.0913	0.0083	0.0309	-0.0098	0.0466	0.0198	0.0364	0.0861	0.2657	0.1090
Quantitative Reasoning	0.0598	0.0116	0.0239	-0.0031	0.0692	0.0153	0.0472	0.0696	0.2272	0.1123
Semantic Relationship	0.0872	0.0182	0.0330	-0.0065	0.0447	-0.0016	0.0256	0.0783	0.3275	0.1220
Spatial Reasoning	0.0598	0.0291	0.0348	0.0059	0.0344	0.0092	0.0561	0.0942	0.1304	0.1156
Temporal Reasoning	0.0441	0.0165	0.0309	0.0332	0.0457	-0.0016	0.0411	0.0936	0.2430	0.1220

Table 3: Ability Impact Score (AIS) matrix for Llama-3.1-8B-Instruct across a curated suite of 10 benchmarks. Higher AIS values indicate greater performance loss upon ability ablation relative to the baseline’s improvement over chance, suggesting higher dependence of the benchmark on that ability.

B Dataset Creation Prompts and Examples

The diagnostic datasets were generated using the o4-mini-2025-04-16 API, guided by specific few-shot examples and instructions for each operationalized ability and task type. Table 5 provides these details, used to generate 2000 examples per ability (split across task types for Contextual Recall). Output was requested in JSON format with keys: context, question, options, answer, answer_index.

Table 5 lists one representative template per ability along with the natural-language instructions fed to the API. Each accepted completion was stored in a JSON schema with keys context, question, options (an array of four or five strings), answer (the correct option token), and answer_index (zero-based integer).

C Human Evaluation of Diagnostic Items

To verify that each synthetic question truly targets its intended skill, we invited ten independent volunteers to label a stratified sample of items drawn from the ten diagnostic datasets.

Table 1 was provided to the experts for reference. Each volunteer saw fifty items (five from every ability) presented one at a time, as illustrated in Figure 5. For every item they selected the single ability that best matched the question and flagged any unclear or sensitive content.

All annotators held at least a bachelor’s degree and were either postgraduate students or early-career researchers who responded to an internal mailing list. Participation was voluntary and unpaid. Three identified as women and seven as men, with ages ranging from 21 to 29. Annotators were

drawn from institutions in Asia and North America.

D Detailed AIS Results and Raw Accuracies

Table 3 provides the detailed AIS matrix and raw accuracies for the baseline model Θ and all 10 ability-ablated models Θ^a at k=1.024% across all evaluated benchmarks. These scores form the basis for creating the visualizations in Section 5.

E Qualitative Impact of Ablations

Table 4 contrasts the answers that the **Base**, **MLP-only**, and **All-layer** ablation versions of Llama-3.1-8B-Instruct give to the same GSM8K test question. The base model returns the correct total of **200 minutes**. The MLP-only model remains fluent but over-counts, replying with **12,000 minutes**. In the all-layer variant the response collapses into a repetitive numeric loop and never produces an answer. These qualitative differences align with the accuracy drops in Table 2 and underline why our study restricts ablation to MLP weights: they weaken targeted reasoning without destroying overall generation.

Please select the ability that best fits the data for a benchmark dataset. Refer to the provided table for ability definitions.
 [#####.....] 14/50
 Data:
 Context: Sequence of letters: B, D, G, K, P, ? Question: What is the next letter in the sequence? Answer: V

1. Analogical Reasoning
2. Commonsense & Causal Reasoning
3. Contextual Recall
4. Deductive Reasoning
5. Inductive Reasoning
6. Long-Term Knowledge Recall
7. Quantitative Reasoning
8. Semantic Relationship Comprehension
9. Spatial & Geometrical Reasoning
10. Temporal Reasoning

Your choice (1-10):

Figure 5: Interface shown to volunteer experts during item validation. Progress is indicated by a bar at the top. Annotators read the prompt, inspect the ten ability options, and enter a numeric choice.

Base Model (no ablation)	
User	<i>Charisma works for 8 hours every day. She has a timer to remind her to get up and walk for 5 minutes every hour she's at work. After 5 days at the office, how many minutes has she walked?</i>
Assistant	To find out how many minutes Charisma walked, we need to calculate the total number of minutes she walked in 5 days. Charisma walks 5 minutes every hour. Since she works 8 hours a day, she walks $5 * 8 = 40$ minutes per day. In 5 days, she will walk $40 * 5 = 200$ minutes. The answer is 200 minutes.
MLP-Only Ablation (fluent but weakened)	
User	<i>Same GSM8K question.</i>
Assistant	To find the total number of minutes Charisma walked, we need to calculate the total number of minutes she walked in 5 days. Charisma works 8 hours a day and walks 5 minutes every hour. First, we need to find the total number of minutes she works in a day. $8 \text{ hours} * 60 \text{ minutes per hour} = 480$ minutes per day Since she walks 5 minutes every hour, we multiply the total minutes she works in a day by 5. $480 \text{ minutes per day} * 5 \text{ minutes per hour} = 2400$ minutes per day Now, we need to find the total number of minutes she walked in 5 days. $2400 \text{ minutes per day} * 5 \text{ days} = 12000$ minutes Therefore, Charisma walked 12000 minutes in 5 days.
All-Layer Ablation (degraded)	
User	<i>Same GSM8K question.</i>
Assistant	Charisma works 8 hours a day, 5 days a week, so $5 * 8 = 40$ hours. $40 * 60 = 2400$ minutes. $2400 / 5 = 4800$ minutes in 5 days. $8 * 60 = 4800$ minutes in 5 days. $4800 / 5 = 9600 / 5 = 9600 / 5 = 9600 / 5 = 1920$ 5-minute walks, $5 \text{ minutes} * 5 = 25$ minutes per walk, $5 * 5 = 25$. $25 * 5 = 1250$, $1250 / 5 = 2500$, $2500 / 5 = 500$, $500 / 5 = 100$, $100 / 5 = 20$. $20 * 5 = 1000$, $1000 / 5 = 200$, $200 / 5 = 40$, $40 / 5 = 8$. $8 * 5 = 40$, $40 / 5 = 8$, $8 / 5 = 1.6$, $1.6 * 5 = 8$, $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$. $8 / 5 = 1.6$, $1.6 * 5 = 8$.

Table 4: GSM8K test prompt evaluated by three model variants. The base model answers correctly, the MLP-only model remains fluent but over-counts, and the all-layer model degenerates into repetitive incoherence, illustrating why attention weights are left intact in our study.

Table 5: Few-shot Examples and Instructions for Diagnostic Dataset Generation.

Ability	Task Type	Representative Few-shot Example	Instruction
Analogical Reasoning	analogy	Context: 'Light : Dark :: Truth : ?' Question: 'Which option best completes the analogy?' Options: ['Lie', 'Fact', 'Shadow', 'Wisdom'] Answer: 'Lie'.	Create a new analogy question. Use 'A : B :: C : ?' style or a similar analogical relationship in context, and ask which option completes it. Output as JSON with the required fields.
Commonsense & Causal Reasoning	everyday_cause_effect	Context: 'Tom left his ice cream in the sun on a hot day.' Question: 'What likely happened to the ice cream?' Options: ['It melted', 'It froze', 'It caught fire', 'It grew larger'] Answer: 'It melted'.	Now write a new commonsense cause-and-effect question. The context should be a simple scenario, and the question asks for a logical outcome or reason. Ensure the answer is based on everyday common sense. Output in JSON format.
Contextual Recall	single_fact_recall	Passage: 'In an ancient manuscript... Nile flows through Egypt while the Tiber traverses Rome.' Question: 'Based on the passage, which river is associated with Rome?' Options: ['Nile', 'Tiber', 'Danube', 'Ganges'] Answer: 'Tiber'.	Using the context... generate one new, diverse and non-redundant MCQ. Output valid JSON object with keys: context, question, options, answer, answer_index. Use varied domains.
	multi_hop_recall	Passage: 'At a Geneva symposium... 1969 lunar landing spurred robotics. Later in Tokyo... noted automation...' Question: 'Which event... linked to inspiration for robotics?' Options: ['Geneva...', '1969 lunar landing', 'Tokyo...', 'Automation...'] Answer: '1969 lunar landing'.	Now generate a new multi-hop question. Passage should blend facts, question require combination. Present JSON ensuring multi-hop reasoning.
Deductive Reasoning	logic_puzzle	Passage: 'If every cat... black animals are calm... Whiskers is in the room.' Question: 'Based on the passage, what can we deduce about Whiskers?' Options: ['Whiskers is calm.', 'Whiskers is not black.', ...] Answer: 'Whiskers is calm.'.	Now create a new deductive reasoning question. Provide context with premises/clues, question requires deducing answer. Output JSON.
Inductive Reasoning	pattern_completion	Sequence: 'A, C, E, G, ?' Question: 'What is the next letter...?' Options: ['H', 'I', 'J', 'K'] Answer: 'I'.	Generate new inductive reasoning question based on pattern. Ensure question asks for next element/rule. Provide answer/answer_index JSON.
Long Term Knowledge Recall	world_fact	Context: 'This question is about world geography.' Question: 'What is the capital city of Australia?' Options: ['Sydney', 'Canberra', 'Melbourne', 'Perth'] Answer: 'Canberra'.	Create new world-knowledge question. Provide brief context if needed, question must be answered from general knowledge. Ensure JSON format.
Quantitative Reasoning	arithmetic_word_problem	Context: 'Alice had 5 apples. She gave 2 to Bob and then bought 3 more.' Question: 'How many apples does Alice have now?' Options: ['6', '5', '8', '10'] Answer: '6'.	Generate new math word problem/quantitative question. Context provides numbers/scenario, question asks for result. Provide answer/answer_index JSON.
Semantic Relationship	roles_and_relations	Passage: 'Alice is Bob's mother. Bob is Charlie's teacher.' Question: 'Who is Alice to Charlie?' Options: ['His mother', 'His teacher', 'His grandmother', 'Not related'] Answer: 'His grandmother'.	Generate new passage and question about relationships/roles. Passage contains ≥ 2 entities with relationship. Ask question testing understanding. Output JSON.
Spatial Reasoning	spatial_relation	Context: 'There is a triangle to the left of a square, and a circle above the triangle.' Question: 'Which shape is directly below the circle?' Options: ['Triangle', 'Square', 'Circle', 'None'] Answer: 'Triangle'.	Generate new spatial reasoning question. Context: description of locations or simple geometry. Ask about relative position, direction, or basic inference. Provide JSON output.
Temporal Reasoning	temporal_order	Context: 'John's meeting started at 9:00 AM... lasted 2 hours. Mary's meeting started at 10:30 AM...' Question: 'Whose meeting ended later?' Options: ['John', 'Mary', 'Same time', 'Not enough info'] Answer: 'John'.	Now create new temporal reasoning question. Context with ≥ 2 events/time points. Ask about order/timing (e.g., first, duration). Output JSON.