# Evaluation and Interpretability: Benchmarks, Dimensions, and Attribution Tooling

October 25, 2025

# 1 Benchmarks: MMLU, GSM8K, BIG-Bench

## 1.1 Benchmark landscape

Figure ?? maps MMLU, GSM8K, and BIG-Bench along a capability spectrum. Together they cover broad knowledge, multi-step reasoning, and long-tail evaluations (including safety and creativity).

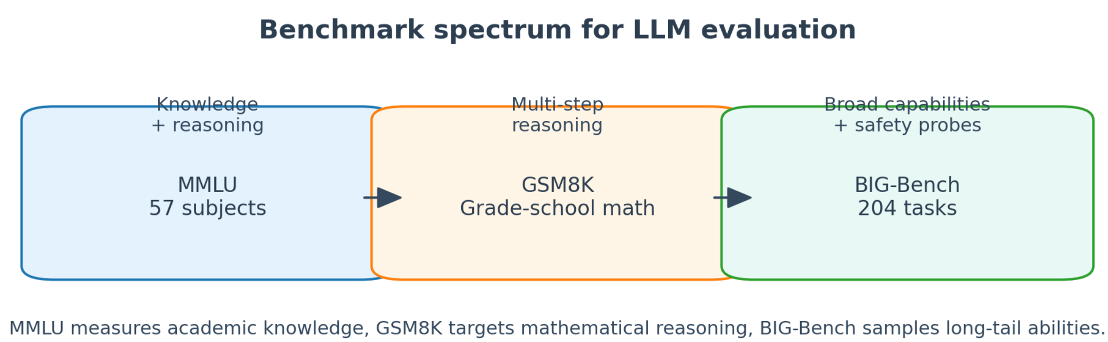


Figure 1: Benchmark spectrum across knowledge (MMLU), mathematical reasoning (GSM8K), and generalized capabilities (BIG-Bench).

## 1.2 MMLU

- 57 academic subjects with 15K multiple-choice questions spanning STEM, social sciences, humanities.
- Tests factual recall, contextual understanding, and out-of-domain transfer.
- Variants include translated versions, chain-of-thought prompting, and few-shot evaluation.

## 1.3 GSM8K

- Focuses on grade-school math word problems requiring step-by-step reasoning.
- Chain-of-thought prompting and self-consistency sampling dramatically boost accuracy.
- Extensions involve harder math sets, program-aided solving, and verification loops.

### 1.4 BIG-Bench

- 204 diverse tasks, including logic puzzles, ethics, multimodal reasoning, and adversarial challenges.
- BIG-Bench Hard isolates tasks humans solve easily but models find difficult—ideal for frontier evaluations.
- Supports crowd-sourced task contributions, enabling rapid growth of long-tail assessments.

## 2 Evaluation Dimensions: Knowledge, Reasoning, Safety, Values

### 2.1 Dimension matrix

| Dimension | Representative benchmarks | Focus areas |
|---|---|---|
| Knowledge | MMLU, TruthfulQA | Factual accuracy, specialized expertise, freshness |
| Reasoning | GSM8K, ARC-Challenge, MathBench | Multi-step deduction, symbolic manipulation, planning |
| Safety | RealToxicity, AdvBench, JailbreakBench | Harmful content detection, jailbreak resistance, policy compliance |
| Values alignment | Anthropic Helpful/Harmless, Constitutional AI evals | Moral alignment, cultural sensitivity, normative coherence |

### 2.2 Evaluation workflow

1. Maintain a unified evaluation repository (static benchmarks + custom datasets) with standardized prompts.
2. Integrate online telemetry (user feedback, refusal rates) to complement offline scores.
3. Automate reporting: trend dashboards, anomaly detection, SLA alerts.
4. Continuously red-team safety and alignment dimensions; refresh adversarial sets frequently.

### 2.3 Metrics and diagnostics

- Accuracy, macro/micro F1, exact match for classification and QA tasks.
- Chain analytics: reasoning length, error type taxonomy, tool usage counts.
- Safety metrics: refusal ratio, toxic incidence, recovery rate after unsafe prompt.
- Alignment metrics: sentiment ratio, cross-cultural consistency, human preference scores.

## 3 Attention Visualization and Attribution Analysis

### 3.1 Interpretability pipeline

Figure **??** outlines the lifecycle from instrumenting inputs to deriving insights via attention probes and attribution scores.

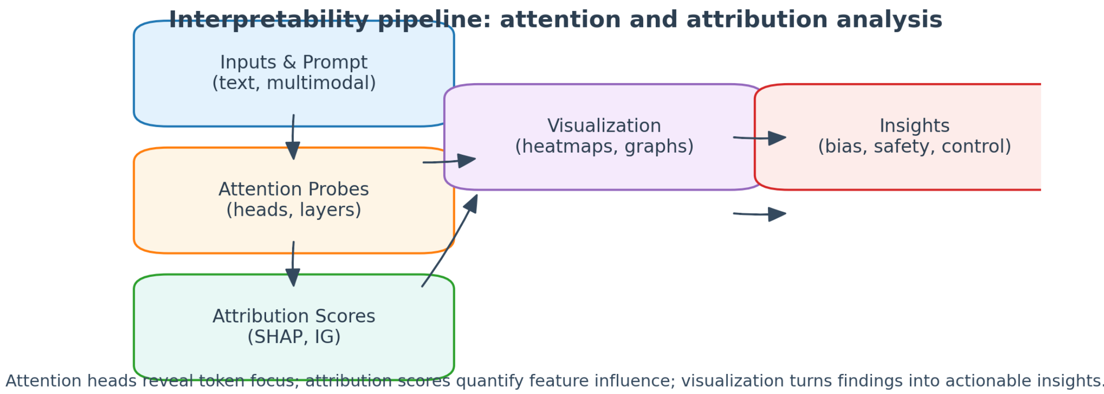


Figure 2: Interpretability pipeline combining attention probing, attribution scoring, visualization, and insights.

## 3.2 Attention probing

- **Attention rollout:** Multiplies attention matrices across layers to estimate token influence.
- **Attention flow:** Accounts for residuals/MLPs to better approximate information flow (Chefer et al.).
- **Head importance:** Gradient or masking-based head scoring reveals redundant attention heads for pruning.

## 3.3 Attribution methods

- **Integrated Gradients (IG):** Computes path-integrated gradients from a baseline to quantify feature contribution.
- **SHAP:** Game-theoretic attribution adaptable to tabular, text, and multimodal inputs.
- **Layer-wise relevance propagation (LRP):** Propagates relevance through deep networks, capturing non-linear interactions.

## 3.4 Example: IG with LLaMA

Listing 1: Integrated Gradients attribution for a LLaMA model

```python
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer
from captum.attr import IntegratedGradients

model_name = "meta-llama/Llama-2-7b-chat-hf"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name, torch_dtype=torch.float16).
    cuda()
model.eval()

prompt = "Explain the greenhouse effect."
inputs = tokenizer(prompt, return_tensors="pt").to(model.device)

def forward(inputs_ids, attention_mask):
    outputs = model(input_ids=inputs_ids, attention_mask=attention_mask)
    return outputs.logits[:, -1, :].max(dim=-1).values
```

```python

ig = IntegratedGradients(forward)
baseline = torch.zeros_like(inputs["input_ids"])

attributions, _ = ig.attribute(
    inputs["input_ids"],
    baselines=baseline,
    additional_forward_args=(inputs["attention_mask"],),
    return_convergence_delta=True,
)

tokens = tokenizer.convert_ids_to_tokens(inputs["input_ids"][0])
for token, score in zip(tokens, attributions[0].sum(dim=-1).tolist()):
    print(f"{token}: {score:.4f}")
```

### 3.5 Visualization techniques

- Token heatmaps overlay attribution scores on text; color intensity reveals focus.
- Attention graphs depict head-to-token relationships using networkx or graphviz.
- Interactive dashboards (Streamlit, Gradio) allow filtering samples, comparing models, and annotating anomalies.

## Operational recommendations

- Align evaluation dimensions with product goals; combine static benchmarks and dynamic telemetry.
- Maintain reproducible evaluation pipelines with versioned datasets and code.
- Use interpretability findings to categorize failure modes (hallucination, bias, reasoning gaps) and feed results back into training.
- Perform red-team and gray-box audits before deployment; archive evaluations for compliance reviews.

## Further reading


- Hendrycks et al. "Measuring Massive Multitask Language Understanding." ICLR, 2021.
- Cobbe et al. "Training Verifiers to Solve Math Word Problems." arXiv, 2021.
- Srivastava et al. "Beyond the Imitation Game Benchmark (BIG-bench)." arXiv, 2022.
- Chefer et al. "Transformers Interpretability Beyond Attention Visualization." CVPR, 2021.
- Mukherjee et al. "LLM Introspection: Improving Safety via Interpretability." arXiv, 2023.