Benchmark-Based Comparative Assessment of Publicly Benchmarked Indian Foundation Models: A Capability and Evaluation-Maturity Framework

Avinash Agarwal*[1] and Vridhi Jain†[1]

[1]Unique Identification Authority of India, New Delhi, India

August 2026

Abstract

Governments increasingly fund indigenous foundation models to strengthen national AI capability, digital sovereignty, and multilingual computing. Assessing the progress of such national ecosystems is complicated by inconsistent benchmark reporting, proprietary evaluation methodologies, and rapidly evolving model releases. This paper presents a structured, benchmark-based comparative assessment of publicly benchmarked Indian foundation models against global frontier and comparable-scale models, across eight capability domains: general-purpose reasoning, coding and software engineering, agentic AI and computer use, cybersecurity, vision and image understanding, video and multimodal understanding, scientific research, and Indic language capability. Using only publicly reported benchmark results, we find that Indian models achieve strong scores on established benchmarks such as MMLU and MATH-500. However, these benchmarks are now widely regarded as saturated, and frontier developers no longer report them. Indian models participate far less frequently in newer, agentic, and domain-specialized evaluations. Benchmark participation is also highly uneven across Indian organizations. Among the models surveyed, Sarvam AI reports the broadest benchmark coverage by a substantial margin. We propose an exploratory four-dimension Benchmark Maturity Index (BMI), scoring each capability domain on standardization, participation, independent verification, and national coverage. We show that the BMI refines, and in some cases revises, the maturity judgments that a purely descriptive review would produce. We argue that many apparent capability gaps in the public record cannot be distinguished, on available evidence, from evaluation-ecosystem gaps. This has direct implications for how national AI programs should design monitoring and funding criteria.

Keywords: generative AI, Indian foundation models, benchmarking, comparative evaluation, benchmark maturity, Benchmark Maturity Index, AI evaluation, sovereign AI

1 Introduction

Generative artificial intelligence has moved rapidly from text generation toward a broad set of capabilities. These now include reasoning, software engineering, multimodal understanding, image and video generation, scientific discovery, and autonomous agentic behavior [1, 2]. As these capabilities have matured, national governments have increasingly treated frontier AI as a strategic technology. They fund indigenous foundation models to build technological capability, digital sovereignty, multilingual computing capacity, and domain-specific innovation. India's IndiaAI Mission is one such program. It

*Corresponding author. Deputy Director General. Email: avinash.70@gov.in

†Intern

supports the development of sovereign foundation models trained on Indian data and aligned with India's linguistic and social context. India has adopted a sector-led, relatively light-touch approach to AI governance, which promotes innovation but risks policy fragmentation and incomparable evaluation practices across domains [3].As of February 2026, the IndiaAI Innovation Centre (Foundation Models) pillar has selected 12 organizations and consortia for support, including both private companies and academic consortia [4].

Assessing the progress of a national AI ecosystem is not straightforward. It requires more than identifying which models exist. It requires understanding which capability domains are represented, how models compare with global frontier systems, and where genuine capability gaps remain as opposed to gaps in public disclosure. This distinction is the central concern of this paper. It matters because different organizations evaluate their models using different datasets, benchmark suites, evaluation harnesses, and reporting practices. Some organizations publish extensive benchmark tables at every release. Others disclose only selected results, or none at all. A national capability assessment that does not account for this will be incomplete at best. At worst, it will mistake an evaluation-reporting gap for a capability gap.

This paper undertakes a structured, benchmark-based comparative assessment of publicly benchmarked Indian foundation models with this distinction as its organizing concern. Rather than focusing on any single model family, we compare representative Indian models against global frontier models and against global models of a comparable scale, across eight capability domains. We restrict our analysis to publicly reported benchmark results drawn from technical reports, model cards, and benchmark leaderboards. We distinguish between developer-reported scores and independently verified scores throughout. We then ask a second, deliberately separate question: not only how capable is each domain's best Indian model, but how mature and independently verifiable is the evaluation ecosystem within which that capability was measured.

1.1 Contributions

This paper makes four contributions.

1. To our knowledge, this is among the first structured, eight-domain, benchmark-based comparative assessments of publicly benchmarked Indian foundation models against global frontier and comparable-scale models.

2. We propose a graded, three-tier working definition of an "Indian-developed AI model." This distinguishes fully indigenous development from India-led development with global components, and from India-adapted foreign base models. We apply this definition to the 12 organizations selected under the IndiaAI Innovation Centre.

3. We propose an exploratory four-dimension Benchmark Maturity Index (BMI). It scores standardization, participation, independent verification, and national coverage at the level of a capability domain. We show that it refines, and in two cases revises, the maturity judgments that a purely qualitative review would produce. We present this as a framework for further development rather than a validated instrument.

4. We identify and analyze eight cross-cutting properties of the current benchmark ecosystem that materially affect any national capability comparison. These include benchmark saturation, the absence of a common benchmark set across providers, and the divergence between the benchmarks that Indian and global developers choose to report.

The remainder of this paper is organized as follows. Section 2 situates this work relative to existing literature. Section 3 describes our methodology, including model and benchmark selection and our working definition of an Indian-developed model. Section 4 maps the IndiaAI-supported ecosystem. Section 5 presents the comparative assessment across eight capability domains. Section 6 introduces the Benchmark Maturity Index. Section 7 discusses cross-cutting patterns, limitations, and policy implications. Section 8 concludes.

2 Related Work

Systematic benchmarking of foundation models has become a distinct area of work. The Holistic Evaluation of Language Models (HELM) project [5] was among the first efforts to evaluate a broad set of language models across a standardized battery of scenarios and metrics. It was explicitly motivated by the observation that model developers report inconsistent and self-selected subsets of benchmark results. Chatbot Arena [6] took a complementary approach, using pairwise human preference voting to rank models independently of developer-reported scores. Independent commercial platforms, including Artificial Analysis [7] and Epoch AI [8], have since extended this practice. They maintain continuously updated leaderboards that track frontier model releases across cost, speed, and capability dimensions. METR [9] has focused specifically on independently verifying agentic and autonomous capability claims, an area where self-reported benchmark scores are particularly difficult to compare.

This body of work establishes that benchmark fragmentation and self-reporting bias are general problems in the foundation model ecosystem. They are not problems specific to any single country's models. However, existing benchmarking efforts are largely organized around individual models or a global leaderboard, rather than around the question of how a national AI ecosystem compares with the global frontier. This is the gap the present paper addresses. Separately, composite indices have been proposed for assessing national AI regulatory readiness [10], addressing the legal and institutional preparedness to govern AI. The present study addresses a complementary but distinct question: the maturity of the evaluation ecosystem itself.We are not aware of a prior structured, multi-domain benchmark comparison of Indian foundation models specifically, though individual Indian model releases, such as Sarvam AI's technical reports, do report comparative benchmark tables against selected global models [11].

Separately, a growing literature documents specific problems with benchmark reliability itself. These include data contamination [12, 13], benchmark saturation as top models approach ceiling performance [14], and the sensitivity of reported scores to prompting and harness configuration. We draw on this literature in Section 7. We argue that these problems are not incidental to a national capability assessment but are central to interpreting it correctly.

3 Methodology

3.1 Model and Benchmark Selection

We compare models across three groups.

The first group consists of **global frontier models**: GPT-5.6 [15], Claude Opus 5 [16], Kimi K3 [17], and Qwen3.8-Max [18]. These were selected as the four highest-ranked models on third-party frontier leaderboards (Chatbot Arena and Artificial Analysis) as of August 2026 that also had publicly available technical reports. We note that the BMI's Participation dimension is scored against this four-model sample and is therefore sensitive to the sample composition.

The second group consists of **global active-parameter comparable models**. We define this group by active parameter count during inference rather than raw total parameter count, because mixture-of-experts (MoE) architectures make total parameter counts misleading as a proxy for model scale. Specifically, we include models whose active parameter count falls in the approximate range of 12–50 billion parameters. The models in this group are: Inkling [19] (MoE; approximately 975B total parameters, approximately 41B active), Qwen3.6-27B [20] (dense; 27B parameters), and Nemotron 3 Super [21] (MoE; approximately 120B total parameters, approximately 12B active). We note that this group is heterogeneous. Active parameter count is an imperfect proxy for inference-time compute, which also depends on architecture, attention mechanism, sequence length, and implementation. Comparisons within this group should be interpreted with caution.

The third group consists of **representative Indian foundation models**, identified from public technical reports and organizational announcements: Sarvam-105B (MoE; 106B total parameters, 10.3B

active parameters) and Sarvam-30B (MoE; 32B total parameters, 2.4B active parameters) [11], Param2 (MoE; approximately 17B total parameters, approximately 2.4B active) [22], and Krutrim-2 [23]. We note that Param2's active parameter count of approximately 2.4B places it well below the 12–50B active-parameter range used to define the comparable-scale group. It is included in the Indian group because of its importance as an IndiaAI-supported model, not because it is parameter-comparable to the global comparable-scale group. We also note domain-specific Indian systems in scientific research (Shodh AI's Project Skanda, IntelliHealth's NeuroDX, ZenteiQ's BrahmAI) and video generation (Avataar AI's Varya). Section 4 provides a broader mapping of the IndiaAI-supported ecosystem.

Representative benchmarks were identified separately for each of eight capability domains: general-purpose reasoning and knowledge, coding and software engineering, agentic AI and computer use, cybersecurity, vision and image understanding, video and multimodal understanding, scientific research, and Indic language capability. Where an originally identified benchmark did not have publicly reported results for the model set under review, a substitute was selected. This is documented in Section 5.6.

Benchmark scores were extracted from publicly available sources: model technical reports, system cards, official benchmark leaderboards, and comparative benchmark tables reported by model developers. We distinguish throughout between *developer-reported* scores (published by the model developer) and *independently verified* scores (published by a third-party evaluator or independent leaderboard). We did not run any evaluation ourselves. Every score in this paper is therefore a reported figure, not an independently reproduced measurement. This is discussed in Section 7.3.

Data collection was conducted in August 2026. Foundation model benchmark scores change frequently. Readers should treat the figures in this paper as a snapshot and verify current figures against cited primary sources.

3.2 Defining an Indian-Developed AI Model

Not all models built by Indian organizations are equally indigenous. This distinction matters for both fair comparison and funding policy. We propose a three-tier working definition.

Tier 1 (fully indigenous): Models trained from scratch by an Indian organization, using India-based or India-controlled compute infrastructure, on datasets that include a significant share of Indian-sourced data. We acknowledge that terms such as "India-controlled compute" and "significant share" are qualitative. Operationalizing them precisely would require access to training details that are not always publicly available. Sarvam-105B and Sarvam-30B meet this definition: both were trained from scratch on internally curated datasets using IndiaAI Mission compute at Yotta's Shakti H100 cluster in India [11]. BharatGen/Param2 is provisionally classified as Tier 1, given its association with AIRAWAT national compute [22], pending independent verification of all criteria.

Tier 2 (India-led, global components): Models developed by an Indian organization that may use foreign cloud compute, foreign training frameworks, or a mixture of global and Indian datasets.

Tier 3 (India-adapted): Foreign base models that are fine-tuned or otherwise adapted by an Indian organization for Indian languages or use cases, without ground-up development in India. Avataar AI's Varya is a Tier 3 model: it is distilled from Alibaba's open-source Wan 2.2 and adapted for Indian cultural contexts, rather than being trained from scratch [24].

This tiering serves two purposes. First, it provides a defensible basis for comparison. A Tier 1 model faces materially different constraints in compute, data availability, and engineering capacity than a Tier 3 model built on an existing foreign foundation. Second, it can inform differentiated funding criteria. Tier 1 development plausibly requires more sustained public investment in compute and data infrastructure, while Tier 3 work may require less. This framing extends existing policy language: the IndiaAI Mission's 2025 Call for Proposals explicitly favors models trained on Indian datasets and aligned with India's linguistic and societal context.

3.3 Scope and Exclusions

This assessment does not include every Indian AI initiative or every global model release. Models were included on the basis of public availability of technical documentation as of the data collection date. Proprietary models with no publicly disclosed benchmark results were excluded from the comparative tables. Their existence is noted qualitatively in Section 7.1.5 and the full IndiaAI-supported population is catalogued in Section 4. We do not evaluate model safety, alignment, deployment cost, or inference efficiency. This paper is concerned exclusively with publicly reported task-capability benchmark performance.

Because our conclusions are conditioned on public disclosure, they apply to the publicly benchmarked subset of the Indian ecosystem. They do not necessarily characterize the Indian AI ecosystem as a whole. Models that exist but do not publish benchmark results are invisible to this methodology.

4 The IndiaAI-Supported Foundation Model Ecosystem

The Government of India's IndiaAI Innovation Centre (Foundation Models) pillar has selected 12 organizations and consortia for the development of large and small language models based on Indian datasets. The February 2026 Press Information Bureau (PIB) disclosure provides details of the compute and non-compute support extended to each [4]. Table 1 maps these 12 organizations, their known models, and their publicly available benchmark evidence as of August 2026.

Table 1: IndiaAI Innovation Centre (Foundation Models): 12 Supported Organizations and Public Benchmark Evidence

Organization	**Known Model(s)**	**Tier**	**Public Benchmarks?**
Sarvam AI	Sarvam-30B, Sarvam-105B	1[‡]	Yes (broad)
IIT Bombay / BharatGen	Param2	1[†]	Limited
Avataar AI	Varya (video generation)	3[§]	Limited
Shodh AI	Project Skanda	—	None found
IntelliHealth	NeuroDX	—	None found
ZenteiQ	BrahmAI	—	None found
Soket AI	—	—	None found
Gnani AI	—	—	None in surveyed domains[¶]
Gan AI	—	—	None found
GenLoop	—	—	None found
Fractal Analytics	—	—	None found
Tech Mahindra Maker's Lab	—	—	None found

Tier assignments follow the definition in Section 3. A dash indicates insufficient public documentation to assign a tier. "None found" means no standardized benchmark results were located in public technical reports as of August 2026. [†]Provisionally classified; see note in text. [‡]Sarvam-105B and Sarvam-30B were trained from scratch, using IndiaAI Mission compute (Yotta's Shakti H100 cluster), on internally curated datasets; they meet the stated Tier 1 criteria. [§]Varya is distilled from Alibaba's open-source Wan 2.2 and adapted for Indian cultural contexts; it meets the Tier 3 definition. [¶]Gnani AI reports a result on Berkeley BFCL v3 (37.99%), a function-calling benchmark outside the eight capability domains assessed in this study.

We note two distinctions. First, the Tier 1 classification for BharatGen/Param2 is provisional. BharatGen is associated with AIRAWAT national compute infrastructure, which satisfies one element of the Tier 1 definition. However, we have not independently verified that all Tier 1 requirements (training from scratch, India-controlled compute, significant Indian-sourced data) are met. The relevant primary documentation should be consulted for confirmation.

Second, in addition to these 12 IndiaAI-supported organizations, our benchmark comparison includes **Krutrim AI** (Krutrim-2), which is an Indian foundation-model developer not included in the official 12-organization IndiaAI Innovation Centre list. Krutrim is included in the comparative assessment because it has publicly released a model and reported benchmark results. Table 2 records this.

Table 2: Other Indian Foundation Models Included in Comparative Assessment

Organization	Model	Tier	IndiaAI-Supported?	Public Benchmarks?
Krutrim AI	Krutrim-2	2	No	Limited (3 domains: General Purpose AI, Coding, Indic Language)

Of the 12 IndiaAI-supported organizations, only Sarvam AI publishes benchmark results across multiple capability domains. BharatGen/Param2 reports limited results. The remaining 10 either have not released models publicly or have not disclosed standardized benchmark scores within the domains assessed in this study (though Gnani AI reports a result on Berkeley BFCL v3, outside our domain scope). Krutrim, though not IndiaAI-supported, reports results across three domains: MMLU (general purpose), HumanEval (coding), and BharatBench (Indic language). This observation is central to Section 7.1.5. It also motivates our argument that benchmark disclosure practices should be part of any evaluation of publicly funded AI programs.

We emphasize that the absence of public benchmark results does not establish the absence of model capability. Organizations may have working models that have not been publicly benchmarked, or may have results that have not been disclosed. The table records public evidence only.

5 Comparative Assessment Across Capability Domains

Throughout this section, we use the phrase "not reported" to mean that no publicly available benchmark score was found for the model tier in question. This may reflect any of several states: the evaluation was not run, was run but not disclosed, was excluded for strategic reasons, or was not applicable. We do not distinguish between these states because the public record does not allow us to do so.

5.1 General-Purpose Foundation Models

General-purpose foundation models are evaluated on knowledge retrieval, reasoning, and mathematical problem solving. We use MMLU [1], GPQA Diamond [25], Humanity's Last Exam (HLE) [26], MATH-500 [27], and ARC-AGI-2 and ARC-AGI-3 [28, 29].

Table 3: General-Purpose Foundation Models: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
MMLU	Not reported	Not reported	Sarvam-105B, 90.6
GPQA Diamond	GPT-5.6 Sol, 94.1	Inkling, 87.2	Sarvam-105B, 78.7
HLE (with tools)	Claude Opus 5, 64.7	Inkling, 46.0	Sarvam-105B, 11.2
MATH-500	Not reported	Not reported	Sarvam-105B, 98.6
ARC-AGI-3	Claude Opus 5, 30.2	Not reported	Not reported
ARC-AGI-2	GPT-5.6 Sol, 92.5	Not reported	Not reported

Indian models achieve strong scores on MMLU and MATH-500, two widely used benchmarks for general-purpose knowledge and mathematical reasoning. However, these benchmarks have become less discriminative among stronger models as performance has approached the upper end of these benchmarks [14]. Frontier developers in this review do not report scores on them, consistent with the saturation pattern discussed in Section 7.1.1. Strong performance on saturated benchmarks demonstrates competence but does not establish frontier competitiveness.

On GPQA Diamond, the one benchmark with reported scores across all three tiers, Sarvam-105B (78.7) trails the frontier leader by roughly 15 points. The best comparable-scale score is Inkling at 87.2, which also exceeds Sarvam-105B by a meaningful margin. On HLE with tools, Inkling (46.0) substantially outperforms Sarvam-105B (11.2), indicating a significant gap between Indian models and

even comparable-scale global models on harder reasoning benchmarks. No Indian model reports a score on ARC-AGI-2 or ARC-AGI-3. These are newer benchmarks introduced specifically because earlier benchmarks had become uninformative at the frontier.

5.2 Coding and Software Engineering

Coding benchmarks range from foundational code generation (HumanEval [2], MBPP [30]) to real-world, multi-step software engineering tasks (SWE-bench Pro [12], Terminal-Bench 2.1).

Table 4: Coding and Software Engineering: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
HumanEval	Not reported	Not reported	Sarvam-30B, 92.1
MBPP	Not reported	Not reported	Sarvam-30B, 92.7
LiveCodeBench	GPT-5.6 Sol, 82.6	Not reported	Sarvam-105B, 71.7
SWE-bench Pro	Claude Opus 5, 79.2	Inkling, 54.3	Not reported
DeepSWE	GPT-5.6 Sol, 73.0	Not reported	Not reported
Terminal-Bench 2.1	GPT-5.6 Sol, 88.8	Inkling, 63.8	Not reported
Vibe Code Bench v1.1	GPT-5.6 Sol, 80.5	Not reported	Not reported

This domain shows a sharper divide than general-purpose reasoning. Indian models report strong scores on foundational code-generation benchmarks (HumanEval, MBPP, LiveCodeBench). No Indian model reports a result on any of the four agentic or multi-step software engineering benchmarks in this review. Comparable-scale models also participate only partially, appearing on SWE-bench Pro and Terminal-Bench 2.1 but not on DeepSWE or Vibe Code Bench.

The available public evidence therefore shows Indian benchmark participation concentrated in foundational coding, with no publicly reported results in agentic coding. Whether this reflects a genuine capability gap or only a reporting gap cannot be determined from the available evidence.

Two further observations on the global coding table are warranted. First, HumanEval and MBPP show signs of saturation: multiple models across tiers achieve scores above 90, and frontier developers no longer report them. The frontier evaluation focus has shifted to harder agentic benchmarks such as SWE-bench Pro and Terminal-Bench. Second, Claude Opus 5 leads SWE-bench Pro in this table at 79.2%, above Qwen3.8-Max (67.7%) and GPT-5.6 Sol (73.0%). Despite this, Claude Opus 5 does not appear as the top scorer on the remaining coding benchmarks, because Anthropic does not publish scores on DeepSWE, Terminal-Bench 2.1, or Vibe Code Bench v1.1. The pattern illustrates how selective benchmark reporting can make a capable model appear absent from a domain in which it is competitive. We return to this domain's benchmark fragmentation in Section 6.

5.3 Agentic AI and Computer Use

Agentic AI benchmarks evaluate a model's ability to autonomously navigate web interfaces, operate computer environments, and complete multi-step tasks. We use BrowseComp, OSWorld-Verified, and OSWorld 2.0.

Table 5: Agentic AI and Computer Use: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
BrowseComp	Kimi K3, 91.2	Inkling, 77.1	Sarvam-105B, 49.5
OSWorld-Verified	Qwen3.8-Max, 86.1	Not reported	Not reported
OSWorld 2.0	Claude Opus 5, 70.6	Not reported	Not reported

This domain is notable because BrowseComp is the only benchmark in this study where all three model tiers report results on the same benchmark. This makes it the strongest basis for direct three-tier

comparison anywhere in the paper. Sarvam-105B (49.5) and Sarvam-30B (35.5) both report BrowseComp scores, trailing the frontier leader Kimi K3 (91.2) and comparable-scale leader Inkling (77.1) by substantial margins. The gap is large but the existence of Indian participation on a shared benchmark is itself significant. It directly supports the paper's central argument: meaningful comparison is possible wherever standardized evaluation exists and all tiers participate.

On OSWorld-Verified and OSWorld 2.0, only frontier models report scores. No comparable-scale or Indian model participates. This pattern of partial participation mirrors other domains.

5.4 Cybersecurity

Cybersecurity evaluation uses task-based benchmarks measuring vulnerability discovery and exploitation capability: CyberGym, ExploitBench, ExploitGym, and CVE-Bench.

Table 6: Cybersecurity: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
CyberGym	GPT-5.6 Sol, 83.6	Not reported	Not reported
ExploitBench (Cap%)	GPT-5.6 Sol / Claude Opus 5, ≈70	Not reported	Not reported
ExploitGym (2h / 6h)	GPT-5.6 Sol, 25 / 35	Not reported	Not reported
CVE-Bench (pass@1)	GPT-5.6 Sol, 93	Not reported	Not reported

Cybersecurity shows no cross-tier comparability in this review. Reported results are confined to the frontier tier. Even within that tier, participation is incomplete: only GPT-5.6 Sol and Claude Opus 5 report scores. No comparable-scale or Indian model reports any cybersecurity benchmark result. This makes cybersecurity the domain with the least publicly available evidence for any comparative claim. No conclusion about Indian cybersecurity AI capability can be drawn from the available public benchmark record.

5.5 Vision and Image Understanding

We use MMMU [31] and MMMU-Pro. Dedicated image-generation benchmarks (GenEval [32], HEIM [33]) were reviewed but excluded from the comparative table. The models under review report multimodal understanding results, not native image-generation results.

Table 7: Vision and Image Understanding: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
MMMU	GPT-5.6 Sol, 88.8	Not reported	Not reported
MMMU-Pro (with tools)	GPT-5.6 Sol, 84.6	Inkling, 73.5	Not reported
Design Arena	Not reported	Not reported	Not reported

Participation is limited even within the frontier tier. No Indian model reports a result on any vision-understanding benchmark reviewed here, though at least one Indian multimodal system has been publicly announced. This reflects a general scarcity of disclosed vision-understanding results across the ecosystem, not a gap specific to Indian models.

5.6 Video and Multimodal Understanding

The two benchmarks originally identified for this domain, VideoMMMU and MVBench, did not have publicly reported results for the surveyed models. We therefore substitute Video-MME [34] and MMVU. We report this substitution explicitly. However, we note that the need for substitution demonstrates limited public reporting coverage for our model set, not necessarily immaturity of the underlying benchmark ecosystem. Mature benchmarks can have low model participation.

Table 8: Video and Multimodal Understanding: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
Video-MME	Kimi K3, 90.0	Not reported	Not applicable
MMVU	Kimi K3, 82.1	Not reported	Not applicable

The Indian model identified for this area, Varya (Avataar AI), is a dedicated video-*generation* system, not a video-*understanding* model. It is therefore not comparable on the benchmarks above. We mark this "not applicable" rather than as a zero, since the mismatch is one of task definition. India has no publicly benchmarked video-understanding model in our survey. No comparable-scale model reports a result on either benchmark.

We note that video generation and video understanding are fundamentally different tasks. Varya's existence does not constitute evidence about Indian video-understanding capability. A separate assessment of video-generation capability, using generation-specific benchmarks, would be needed to evaluate that distinct domain.

5.7 Scientific Research

Scientific research systems are evaluated using domain-specific tasks rather than a single unified benchmark. We reviewed ProtocolQA, LAB-Bench, ProteinGym, and PaperBench.

Table 9: Scientific Research: Best Reported Score by Model Tier

Benchmark	Best Surveyed Frontier	Best Surveyed Comparable	Best Indian
ProtocolQA (Open-Ended)	GPT-5.6 Sol, 43.5	Not reported	Not reported
PaperBench	Qwen3.8-Max, 93.0	Not reported	Not reported
LAB-Bench	Not reported	Not reported	Not reported
ProteinGym	Not reported	Not reported	Not reported

Frontier models show limited but non-zero participation in this domain. GPT-5.6 Sol reports a score on ProtocolQA and Qwen3.8-Max reports a score on PaperBench. No comparable-scale or Indian model reports results on any of these benchmarks. Well-known global systems in this space, such as AlphaFold [35] and Evo 2, are evaluated on task-specific scientific metrics rather than on the general-purpose benchmarks used in this paper. They are therefore excluded from this comparison.

Shodh AI, IntelliHealth, and ZenteiQ are active Indian entities in this space. None has published a score on the benchmarks considered in this study.

We emphasize that this finding is narrower than it may appear. Scientific AI has rich, task-specific evaluation ecosystems in domains such as protein structure prediction, molecular simulation, and materials science. Our finding is that general-purpose foundation models have very limited publicly reported participation in the scientific benchmarks we selected. It is not a finding about the maturity of scientific AI benchmarking as a whole.

5.8 Indic Language Foundation Models

Indic language capability is the one domain in this study with no global benchmark equivalent. Table 10 summarizes the benchmarks reported by each model version, reflecting the version-wise differences in benchmark coverage that a model-level aggregation would obscure.

India's Indic-language AI ecosystem has expanded significantly, with domestic models increasingly incorporating evaluation of Indian linguistic, cultural, and contextual capabilities. However, the evaluation landscape remains fragmented across organizations and model generations, with different models reporting results on different benchmark suites and relatively limited overlap. This makes direct comparison of Indic-language capabilities across current-generation Indian models difficult.

Table 10: Indic Language Benchmarks by Model Version and Reporting Organization

Model / Organization	Benchmark(s) Reported	Evaluation Focus
Sarvam-30B — Sarvam AI	IndiVibe; MILU [36]	Human-preference evaluation; multilingual Indic language understanding
Sarvam-105B — Sarvam AI	IndiVibe	Human-preference evaluation across Indian languages
Krutrim-2 — Krutrim AI	BharatBench	Indic linguistic, cultural, and contextual understanding
PARAM-1 — BharatGen	SANSKRITI [37]; MILU [36]	Indian cultural knowledge; multilingual Indic language understanding
PARAM-2 — BharatGen	SANSKRITI [37]; Indic BoolQ; ARC Challenge (Indic); TriviaQA (Indic MCQ); HellaSwag Hindi; MMLU Hindi	Indic cultural knowledge, comprehension, reasoning, and knowledge evaluation
Sarvam-30B / PARAM-2 — Independent	IndicParam [38]	Low- and extremely low-resource Indic language understanding

MILU is explicitly reported for Sarvam-30B in its current official model card. IndicParam includes affiliations with both IIM Indore and BharatGen; it should not be characterized as a purely independent benchmark.

Sarvam-30B reports IndiVibe and MILU, while Sarvam-105B reports IndiVibe but not MILU. Krutrim-2 reports BharatBench. PARAM-1 reports SANSKRITI and MILU; PARAM-2 reports a broader set including SANSKRITI, Indic BoolQ, ARC Challenge (Indic), TriviaQA (Indic MCQ), HellaSwag Hindi, and MMLU Hindi. Most reported results are developer-reported, and the differing benchmark suites limit direct cross-model comparability.

The SANSKRITI benchmark [37], developed by academic researchers independently of Param/BharatGen, evaluates cultural knowledge and reasoning in Indian languages. It is distinct from the Sanskriti evaluation reported for BharatGen/Param2, which focuses on multilingual language capability. Both are listed explicitly because the name collision is a source of confusion. IndicParam [38] provides additional evaluation across 11 Indic languages and a Sanskrit-English code-mixed set, covering Sarvam-30B and PARAM-2; however, given that its authors include affiliations with both IIM Indore and BharatGen, it should not be characterized as a purely third-party evaluation.

No single Indic benchmark has achieved broad adoption across major Indian model developers as a common evaluation standard. The absence of shared benchmarks limits ecosystem-level comparison, and the reliance on developer-reported results limits independent verification.

6 The Benchmark Maturity Index

6.1 Motivation

The comparative assessment in Section 5 shows a recurring pattern. Indian models achieve strong scores on several established benchmarks for which they report results. They are absent from domains where benchmarks are newer or less widely reported. A central aim of this paper is to formalize that pattern into a trackable criterion, rather than leaving it as a domain-by-domain narrative.

Within a broader governance architecture that distinguishes regulation, standards, assessment procedures, tools and metrics, and a compliance ecosystem [39], the BMI focuses specifically on the assessment procedures and tools-and-metrics layers. The key move is to separate two questions that a descriptive comparison tends to collapse into one. The first question is: how capable is the best available model in a domain? The second is: how mature and independently verifiable is the evaluation ecosystem within which that capability was measured? These questions often have different answers. A domain

can appear to have a large capability gap largely because its evaluation ecosystem is thin. We therefore propose a four-dimension Benchmark Maturity Index (BMI), scored at the domain level.

6.2 Status and Limitations of the BMI

We present the BMI as an *exploratory framework* rather than a validated instrument. It has been applied only once, to our own dataset. It has not been subjected to inter-rater reliability analysis, sensitivity analysis, or external validation. The claims we make for it are therefore modest: it is a structured way to separate evaluation-ecosystem maturity from model capability, and it produces results that are informative in the present analysis. Whether it generalizes to other national ecosystems remains to be demonstrated. In this study, benchmark maturity refers primarily to the degree of standardization, adoption and independent verification captured by the BMI dimensions.

We also note a circularity that readers should keep in mind. The National Coverage dimension penalizes the absence of publicly reported Indian benchmark results. But this paper's central argument is that such absence does not imply low capability. The BMI therefore measures *evaluation coverage*, not capability. Scores should be read accordingly.

6.3 Dimensions

Each domain is scored on four dimensions, each on a 0–2 scale:

- **Standardization (S):** Does a single, widely adopted benchmark exist for this domain? $S = 0$ indicates no shared benchmark or no benchmark with results from the surveyed model set. $S = 1$ indicates multiple competing benchmarks with no clear consensus. $S = 2$ indicates a small, stable set of benchmarks in wide, consistent use. We note that this dimension partly depends on our model selection. A benchmark may be mature but underrepresented in our model set.

- **Participation (P):** What proportion of the surveyed global frontier models report a score on at least one benchmark in this domain? $P = 0$ indicates none of the four frontier models; $P = 1$ indicates one or two; $P = 2$ indicates three or four.

- **Independent Verification (I):** Are scores available from evaluation conducted independently of the model developer? $I = 0$ indicates only developer-reported scores. $I = 1$ indicates partial independent coverage. $I = 2$ indicates broad independent verification.

- **National Coverage (C):** Do Indian models in our survey report scores in this domain? $C = 0$ indicates no Indian model reports a score. $C = 1$ indicates partial participation (one or two benchmarks, or a single organization). $C = 2$ indicates participation across multiple benchmarks and organizations, comparable to global participation levels.

The four scores are summed to a total in the range 0–8, and mapped to a qualitative maturity band: 0–1 Very Low, 2–3 Low, 4–5 Moderate, 6–7 High, 8 Very High.

For the Indic language domain, which has no global benchmark against which to score Participation, we compute a modified three-dimension score (S, I, C; maximum 6). We map it onto the same five bands proportionally. We acknowledge that this makes the Indic score structurally different from all other domain scores. It should therefore be interpreted separately from the eight-point domain scores.

The four dimensions are equally weighted. We do not have a principled justification for any particular weighting scheme. A sensitivity analysis exploring alternative weights is a priority for future work.

6.4 Applying the Index

Several results refine the purely qualitative domain summaries. First, Coding and Software Engineering scores as High (6/8). National Coverage is revised to 2 because Indian participation extends across

Table 11: Benchmark Maturity Index by Capability Domain

Domain	S	P	I	C	Maturity Band
General Purpose AI	2	2	1	2	High (7/8)
Coding & Software Engineering	1	2	1	2	High (6/8)
Agentic AI & Computer Use	1	2	0	1	Moderate (4/8)
Cybersecurity	1	1	0	0	Low (2/8)
Vision & Image Understanding	1	1	0	0	Low (2/8)
Video & Multimodal Understanding	1	1	0	0	Low (2/8)
Scientific Research	1	1	0	0	Low (2/8)
Indic Language AI*	0	—	0	2	Low (2/6, scaled)

*Scored on a modified three-dimension basis (S, I, C only); see Section 6.3.

multiple benchmarks (HumanEval, MBPP, LiveCodeBench) and multiple organizations (Sarvam-30B, Param2, Krutrim-2), meeting the definition's criterion for C=2. The Standardization score remains 1, depressed by the split between classic code-generation benchmarks and newer agentic software-engineering benchmarks. Indian participation is confined to the classic subset.

Second, Agentic AI and Computer Use scores as Moderate (4/8). This domain is distinctive because BrowseComp provides three-tier comparability on a single shared benchmark. The Participation score is 2 because three of four frontier models report BrowseComp results. National Coverage is 1 because Indian models participate but only on one of three benchmarks. Independent Verification remains 0.

Third, Scientific Research scores as Low (2/8) rather than Very Low. Two frontier models report results in this domain: GPT-5.6 Sol on ProtocolQA, and Qwen3.8-Max on PaperBench. Partial frontier participation is sufficient to lift the Participation score to 1, producing a total of 2/8.

National Coverage scores 0 in four of the eight domains: Cybersecurity, Vision and Image Understanding, Video and Multimodal Understanding, and Scientific Research. This is the most striking finding in the table. It reflects the state of public benchmark reporting by surveyed Indian models, not a demonstrated absence of capability.

7 Discussion

7.1 Cross-Cutting Patterns in the Benchmark Ecosystem

7.1.1 Benchmark saturation

Several of the earliest benchmarks, including MMLU and MATH-500, are now approached or exceeded by the strongest available models. Multiple frontier models have achieved MMLU scores above 90, reducing the benchmark's ability to discriminate among top-performing systems [1, 14]. MATH-500 shows a similar pattern. In response, several recent frontier model reports have reduced or omitted these benchmarks. They now favor newer, harder evaluations such as GPQA Diamond, HLE, and ARC-AGI-2/3. This is directly visible in our data. GPT-5.6, Claude Opus 5, and other frontier models report no MMLU or MATH-500 score, even though these are the benchmarks on which Indian models show their strongest results.

A "Not reported" entry in Table 3 is therefore often a deliberate reporting choice, not evidence of poor performance. However, we note that developer non-reporting is not identical to benchmark saturation. A developer may omit a benchmark for strategic, cost, or positioning reasons unrelated to saturation. We distinguish these possibilities where the evidence permits.

7.1.2 Absence of a common benchmark set across providers

Even among global frontier developers, there is no single shared benchmark battery. Different organizations publish benchmarks that showcase their own model's strengths. Identical benchmark names

sometimes correspond to different evaluation harnesses. For example, "SWE-bench Pro" and "SWE-bench Verified" are related but distinct evaluations. Different point releases of Terminal-Bench are not directly comparable. This fragmentation is a known problem [5] and is not specific to India. It means that any cross-developer comparison, including ours, is necessarily partial.

7.1.3 Divergent benchmark participation between Indian and global models

We observe a consistent pattern: Indian developers do not always report scores on benchmarks in wide global use, even where the benchmark exists and the model could plausibly be evaluated on it. Param2 and Krutrim-2 report no GPQA Diamond score, though several global models of comparable and larger scale do. This may reflect that the evaluation was not run, or that results were obtained but not disclosed. The public record does not allow us to distinguish between these explanations. We treat this ambiguity as itself a finding.

7.1.4 Benchmark scarcity in emerging domains

Image generation and video generation are established application domains. But standardized, widely adopted benchmarks for them remain comparatively scarce, worldwide, not only in India. We encountered this during benchmark selection: our originally identified video-understanding benchmarks, VideoMMMU and MVBench, had no usable public results for our model set and had to be replaced (Section 5.6). This scarcity affects the interpretability of any comparison in these domains.

7.1.5 Concentration of Indian benchmark reporting in a single organization

Among the Indian models surveyed, Sarvam AI reports scores across the widest range of benchmarks, spanning general-purpose reasoning, coding, and Indic-language tasks. Table 1 shows that of the 12 IndiaAI-supported organizations, only Sarvam publishes broad benchmark results. BharatGen/Param2 publishes limited results. The remaining 10 IndiaAI-supported organizations have no publicly available benchmark scores within the eight domains assessed in this study. We note that Gnani AI reports a result on Berkeley BFCL v3, a function-calling benchmark outside our domain scope. Krutrim, which is not among the 12 IndiaAI-supported organizations, reports results across three domains: MMLU, HumanEval, and BharatBench.

This concentration means that evidence of Indian competitiveness in our survey comes predominantly from a single organization. Ecosystem-level claims about Indian AI capability should be read accordingly. We avoid the phrase "overwhelming majority" because we have not conducted a census of all Indian model developers. We can say that Sarvam reports the broadest benchmark coverage among the Indian models in our study, by a substantial margin.

7.1.6 Absence of publicly reported Indian benchmark results in several domains

Four of the eight domains reviewed here have no publicly reported Indian benchmark score among the surveyed models: cybersecurity, vision and image understanding, video and multimodal understanding, and scientific research. (In the video domain, India's representative model Varya performs video generation, a different task from the video-understanding benchmarks used for comparison.) These are not areas of marginal Indian participation. They are areas of no publicly documented Indian participation in the specific benchmarks reviewed.

This is a finding about public benchmark disclosure, not a finding about capability. We cannot determine from the available evidence whether the relevant Indian organizations have evaluated their models on these benchmarks and chosen not to disclose results, have not run the evaluations, or lack the capability. The distinction matters for policy, and is the reason National Coverage scores 0 in these domains.

7.1.7 Benchmark silence is not evidence of capability absence

The preceding patterns support a broader methodological point. Across several domains, the absence of a published score reflects the state of benchmark disclosure and ecosystem maturity. It does not necessarily reflect the absence of underlying model capability. A model may exist and function adequately without ever being publicly benchmarked. A monitoring or funding framework that treats benchmark silence as equivalent to capability absence risks undercounting genuine progress in exactly the domains where independent evaluation infrastructure is weakest. This is the risk the BMI is designed to make visible.

However, we apply this principle to our own findings as well. When we observe no Indian benchmark result in a domain, the correct conclusion is that the available public evidence is insufficient for a comparative assessment. It is not that Indian capability is weak or immature. We cannot distinguish capability absence from reporting absence in these cases.

7.1.8 Sensitivity of reported scores to evaluation methodology

Benchmark scores are sensitive to how the evaluation is run. Prompting configuration, reasoning effort settings, tool access, and evaluation harness all materially affect results. This is well documented in the literature [5, 13]. It means that raw benchmark scores are insufficient for confident comparison without knowledge of the evaluation conditions. This is one reason the BMI scores Independent Verification separately from Participation.

We note that our tables do not systematically report evaluation conditions for each score. This is a limitation. Ideally, each entry would include the benchmark version, evaluation setting, tool access, and source. Future work should prioritize controlled, condition-matched comparisons.

7.2 On the Definition of an Indian AI Model

The comparative assessment in this paper depends on a prior, and not fully settled, question: what should count as an Indian-developed AI model. Section 3 proposed a three-tier definition to make this scoping decision explicit, distinguishing fully indigenous development (Tier 1), India-led development with global components (Tier 2), and India-adapted foreign base models (Tier 3). The findings in Section 5 are informative about this definition in a way that is worth making explicit.

Both of the most visible Indian models in this study, Sarvam AI's models and BharatGen's Param2, are classified as Tier 1 under the definitions in Section 3. However, they differ substantially in the breadth of their publicly reported benchmark coverage. Sarvam reports scores across multiple domains and multiple benchmark suites. Param2 reports limited results, concentrated in Indic-language evaluation. Both meet the structural Tier 1 criteria (trained from scratch, India-based or India-controlled compute), but their public evaluation postures differ considerably. This illustrates a limitation of a tier-based taxonomy: tier assignment describes training provenance, but says nothing about the breadth, depth, or transparency of evaluation. A nationally funded model monitoring framework would need to track both dimensions separately.

This has a practical implication for monitoring design. IndiaAI's funding mandate holds two goals simultaneously: building globally competitive AI capability, and building indigenous, sovereign AI capability. The tier definition addresses the sovereignty dimension. The BMI's National Coverage dimension provides a first proxy for the competitive benchmarking dimension. However, neither instrument fully captures whether a model's publicly reported results reflect its actual capability. A robust monitoring framework should therefore require both tier documentation and minimum benchmark disclosure from all publicly funded organizations.

7.3 Limitations

This study relies entirely on publicly reported scores. We did not independently reproduce any evaluation. The comparison therefore inherits whatever self-reporting bias exists in the underlying sources.

The Independent Verification dimension of the BMI makes this trackable rather than resolving it.

The assessment is a snapshot as of August 2026. Relative standings can shift within weeks of a new release. Benchmark harnesses are not always comparable even when benchmark names match. We cannot rule out undocumented harness differences in the data.

Our model inclusion criteria depend on public disclosure. Organizations that do not publish technical reports are underrepresented or absent regardless of their actual capability. This is an instance of the benchmark-silence problem that applies to our own methodology.

We do not provide a complete candidate list with inclusion/exclusion dates and reasons. This limits reproducibility. Future assessments should include such a table.

The BMI has not been validated externally. Its scoring thresholds are qualitative and its weighting is equal by default. A sensitivity analysis is needed to establish whether alternative reasonable weights would change the conclusions.

Finally, the comparable-scale model group is heterogeneous. It mixes dense and MoE architectures with different total and active parameter counts. Comparisons within this group should be interpreted with caution.

7.4 Implications for Policy and Practice

These findings have several implications for policy design.

First, closing the apparent gap in publicly reported benchmark performance is not solely a model-training problem. In several domains, most notably cybersecurity, vision, and scientific research, the binding constraint on what can be publicly known is the absence of standardized evaluation infrastructure. Investment directed only at model development, without parallel investment in benchmark creation and independent verification, will leave this problem unaddressed.

Second, the concentration of Indian benchmark participation in a single organization (Section 7.1.5) suggests that ecosystem-level claims about Indian AI progress should be disaggregated by organization. A monitoring framework that reports only aggregate or best-case Indian performance risks overstating how broadly based current progress actually is.

Third, for publicly funded models, benchmark disclosure is not merely a technical issue. If a model receives substantial public compute support under the IndiaAI Mission, there is a reasonable case for requiring minimum benchmark disclosure. This could include:

- establishing a minimum benchmark-disclosure framework, with the benchmark set reviewed regularly to reflect changes in model capabilities and evaluation maturity.
- mandatory reporting of scores on the specified set of benchmarks;
- disclosure of evaluation conditions (benchmark version, harness, prompting, tool access);
- independent third-party evaluation for at least a subset of benchmarks;
- periodic re-evaluation as models are updated;
- public availability of benchmark methodology documentation.

These requirements would directly address several of the evaluation-ecosystem gaps documented in this paper.

Fourth, this study's approach of scoring evaluation-ecosystem maturity alongside publicly reported model performance is not specific to India. It could be applied to other national AI programs facing similar assessment challenges.

8 Conclusion and Future Work

This paper presented a structured, eight-domain, benchmark-based comparative assessment of publicly benchmarked Indian foundation models against global frontier and comparable-scale models. Rather than a single, uniform capability gap, we find substantially uneven maturity across capability domains. Indian models achieve strong reported scores on established benchmarks, particularly MMLU and MATH-500 for general-purpose reasoning, and HumanEval and MBPP for foundational coding. However, these benchmarks provide limited evidence of frontier-level differentiation. The addition of Agentic AI as a domain reveals the only instance in this study where all three comparison groups report results on a single shared benchmark (BrowseComp), enabling direct comparison. However, strong performance on saturated benchmarks does not establish frontier competitiveness. Several other domains show no Indian benchmark participation at all among the models surveyed.

We cannot determine, from the available public evidence, whether these gaps reflect genuine capability deficits, strategic reporting choices, or the absence of evaluation infrastructure. This ambiguity is itself the central finding of the paper. We formalized it into an exploratory Benchmark Maturity Index and showed that it refines the conclusions a purely qualitative review would produce.

Future work should extend this assessment longitudinally, tracking how both publicly reported model performance and benchmark maturity evolve over successive model releases. It should incorporate independent, third-party re-evaluation of a sample of the models discussed here. The BMI should be subjected to sensitivity analysis and, ideally, applied to other national AI ecosystems. Finally, a complete model-by-benchmark dataset, including evaluation conditions and source provenance for every score, should be published as supplementary material.

Author Contributions

Avinash Agarwal conceptualised the study, defined the analytical framework, and drafted the manuscript. Vridhi Jain conducted the systematic benchmark searches and collected and verified all benchmark data from primary sources. Both authors reviewed the final manuscript.

References

[1] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring Massive Multitask Language Understanding (MMLU). *International Conference on Learning Representations (ICLR)*, 2021.

[2] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021.

[3] A. Agarwal and M. J. Nene. A Federated Architecture for Sector-Led AI Governance: Lessons from India. *Transforming Government: People, Process and Policy*, 2026. doi: 10.1108/TG-09-2025-0310.

[4] Press Information Bureau, Government of India. IndiaAI Mission: Foundation Model Development Support. PIB Release, February 2026.

[5] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic Evaluation of Language Models (HELM). *Transactions on Machine Learning Research (TMLR)*, 2023.

[6] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *International Conference on Machine Learning (ICML)*, 2024.

[7] Artificial Analysis. LLM Leaderboard and Benchmarks. https://artificialanalysis.ai/, accessed August 2026.

[8] Epoch AI. AI Trends Dashboard. https://epoch.ai/, accessed August 2026.

[9] METR. Model Evaluation and Threat Research. https://metr.org/, accessed August 2026.

[10] A. Agarwal and M. J. Nene. The AI Regulatory Readiness Index (ARRI): Assessing Cross-Jurisdictional Legal Preparedness for AI in Telecommunications. *Computer Law & Security Review*, 61:106340, 2026. doi: 10.1016/j.clsr.2026.106340.

[11] Sarvam AI. Open-Sourcing Sarvam 30B and 105B. Sarvam AI Technical Blog, 2026.

[12] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *International Conference on Learning Representations (ICLR)*, 2024.

[13] Z. Shi, A. Wang, A. Srinivasan, A. T. Chaganty, and P. Liang. Detecting Pretraining Data from Large Language Models. *International Conference on Learning Representations (ICLR)*, 2024.

[14] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, R. Ringshia, et al. Dynabench: Rethinking Benchmarking in NLP. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

[15] OpenAI. GPT-5.6 Technical Report. OpenAI, 2026.

[16] Anthropic. Claude Opus 5 System Card. Anthropic, 2026.

[17] Moonshot AI. Kimi K3 Technical Report. Moonshot AI, 2026.

[18] Alibaba Cloud. Qwen3.8-Max. Official Press Release, Alibaba Cloud, 2026.

[19] Thinking Machines Lab. Inkling Model Card. Thinking Machines Lab, 2026.

[20] Alibaba Cloud. Qwen3.6-27B Technical Report. Alibaba Cloud, 2026.

[21] NVIDIA. Nemotron 3 Super Technical Report. NVIDIA, 2026.

[22] BharatGen Consortium / IIT Bombay. Param2 Model Card and Technical Documentation. 2025.

[23] Krutrim AI. Krutrim-2 Technical Report. Krutrim AI, 2026.

[24] I. Mehta. "Cheaper, faster, and culturally aware, Avataar's video AI is built for India's scale." *TechCrunch*, June 11, 2026. Available: https://techcrunch.com/2026/06/11/cheaper-faster-and-culturally-aware-avataars-video-ai-is-built-for-indias-scale/, accessed August 2026.

[25] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[26] Center for AI Safety and Scale AI. Humanity's Last Exam: A Benchmark for Frontier AI Systems. *arXiv preprint arXiv:2501.14249*, 2025.

[27] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's Verify Step by Step. *International Conference on Learning Representations (ICLR)*, 2024.

[28] F. Chollet, M. Knoop, G. Kamradt, B. Landers, and H. Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. *arXiv preprint arXiv:2505.11831*, 2025.

[29] ARC Prize Foundation. ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence. *arXiv preprint arXiv:2603.24621*, 2026.

[30] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*, 2021.

[31] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[32] D. Ghosh, H. Hajishirzi, and L. Schmidt. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[33] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, et al. Holistic Evaluation of Text-to-Image Models (HEIM). *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[34] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[35] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[36] S. Verma, M. S. U. R. Khan, V. Kumar, R. Murthy, and J. Sen. MILU: A Multi-task Indic Language Understanding Benchmark. *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. doi: 10.18653/v1/2025.naacl-long.507.

[37] A. Maji, R. Kumar, A. Ghosh, Anushka, and S. Saha, SANSKRITI: A Comprehensive Benchmark for Evaluating Language Models' Knowledge of Indian Culture. *Findings of ACL 2025, pp. 4434–4451, DOI 10.18653/v1/2025.findings-acl.228*.

[38] A. Maheshwari, K. Sharma, V. Patel, A. Maheshwari. IndicParam: Benchmark to Evaluate LLMs on Low-Resource Indic Languages. *arXiv preprint arXiv:2512.00333*, 2025.

[39] A. Agarwal and M. J. Nene. A Five-Layer Framework for AI Governance: Integrating Regulation, Standards, and Certification. *Transforming Government: People, Process and Policy*, 19(3), pp. 535–555, 2025. doi: 10.1108/TG-03-2025-0065.