Apriel-1.5-15B-Thinker: Mid-training is all you need

SLAM Lab
ServiceNow *

Abstract

We present Apriel-1.5-15B-Thinker, a 15-billion parameter open-weights multimodal reasoning model that achieves frontier-level performance through training design rather than sheer scale. Starting from Pixtral-12B, we apply a progressive three-stage methodology: (1) depth upscaling to expand reasoning capacity without pretraining from scratch, (2) staged continual pre-training that first develops foundational text and vision understanding, then enhances visual reasoning through targeted synthetic data generation addressing spatial structure, compositional understanding, and fine-grained perception, and (3) high-quality text-only supervised fine-tuning on curated instruction-response pairs with explicit reasoning traces spanning mathematics, coding, science, and tool use. Notably, our model achieves competitive results without reinforcement learning or preference optimization, isolating the contribution of our data-centric continual pre-training approach. On the Artificial Analysis Intelligence Index, Apriel-1.5-15B-Thinker attains a score of 52, matching DeepSeek-R1-0528 despite requiring significantly fewer computational resources. Across ten image benchmarks, its performance is on average within five points of Gemini-2.5-Flash and Claude Sonnet-3.7, a key achievement for a model operating within single-GPU deployment constraints. Our results demonstrate that thoughtful mid-training [2] design can close substantial capability gaps without massive scale, making frontier-level multimodal reasoning accessible to organizations with limited infrastructure. We release the model checkpoint, all training recipes, and evaluation protocols under the MIT license to to advance open-source research.

1 Introduction

Large language models (LLMs) continue to advance rapidly across general capability, long-context reasoning, and multimodal understanding. Open-weight families such as Qwen [1, 2] and Llama [3, 4] have demonstrated strong, scalable baselines, while proprietary systems like Gemini [5, 6] and Claude [7] have pushed frontier performance across complex reasoning and multimodal tasks. Recent "reasoning-first" training approaches, exemplified by DeepSeek-R1 [8], reveal that careful data curation and training strategy can unlock sophisticated chain-of-thought competence without relying solely on extreme scale. Yet despite these advances, a fundamental tension persists between

*Contributors are listed in Section 8

[2]We define mid-training as a combination of the continual pretraining and SFT stages

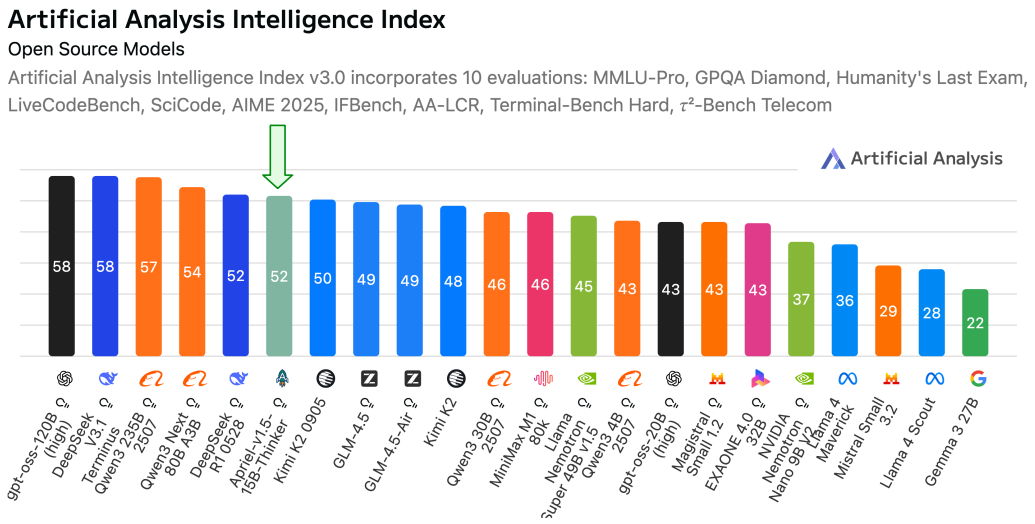


Figure 1: Apriel-1.5-15B-Thinker compared to the best open source LLMs on the Artificial Analysis Intelligence Index.

capability and accessibility, as these insights have not fully addressed the challenges facing real-world applications.

Two critical barriers remain for widespread adoption. First, organizations requiring on-premises or air-gapped deployments for privacy and compliance need compact models with predictable resource footprints that can operate within strict infrastructure constraints. Second, the cost profile spanning both training and inference often becomes the decisive factor in whether frontier-level capability can be deployed at production scale. These practical considerations raise a fundamental question: *Can a compact, open, multimodal model achieve frontier-level reasoning while remaining economical to train and deploy?*

This work introduces **Apriel-1.5-15B-Thinker**, a 15B-parameter open-weights multimodal reasoning model designed with that guiding question in mind. Our approach centers on the mid-training/continual pretraining phase, where both data selection and staged presentation exert a strong influence on downstream reasoning. Concretely, the training corpus spans curated pretraining-style corpora, diverse web-style text and images, reasoning-rich samples, and a mix of verified and unverified synthetic data, all introduced through a staged curriculum. Our core innovation lies in a progressive, cost-effective multimodal training pipeline that effectively scales reasoning capabilities across text and vision through three carefully orchestrated stages:

(1) Integrated Multimodal Architecture: Beginning with Pixtral-12B [9] as our foundation, we expand it to a model size capable of advanced reasoning across modalities, without requiring pretraining from scratch.

(2) Staged Multimodal Continual Pretraining (CPT): We adopt a two-phase CPT strategy. The first phase develops foundational text reasoning and broad multimodal capabilities, while the second enhances visual reasoning through synthetic data targeting spatial structure, compositional understanding, and fine-grained perception. This staged progression enables balanced strengthening of both modalities and provides a stable foundation for subsequent training stages, even when later stages emphasize a narrower set of modalities.

(3) High-Quality Supervised Fine-Tuning (SFT): We curate a diverse, high-quality, and high-signal set of samples for supervised fine-tuning. Each response includes explicit reasoning traces, enabling the model to learn transparent thought processes. Coupled with the strong base model, this yields frontier-level performance across a broad range of reasoning benchmarks without requiring additional post-training.

Given computational constraints, the current release focuses on maximizing the potential of the base model through mid-training, without employing reinforcement learning or preference optimization.

This design choice also allows for a clearer assessment of the contribution of the mid-training recipe itself to the overall performance of the model.

The result is a compact model tailored to enterprise-friendly deployment constraints (memory, latency, throughput) while still achieving frontier-level reasoning. Apriel-1.5-15B-Thinker attains a score of 52 on the Artificial Analysis Intelligence Index, matching DeepSeek-R1-0528 [10] despite requiring significantly fewer computational resources. Across ten multimodal benchmarks, the model demonstrates competitive performance, averaging only 5 points behind Gemini-2.5-Flash and Claude Sonnet-3.7 [6, 7], a remarkable achievement for a 15B parameter model operating within single-GPU deployment constraints. These empirical results provide compelling evidence that thoughtful continual pre-training with heterogeneous synthetic signals, applied to a compact architecture, can close substantial capability gaps without massive scale or expensive RL pipelines. By releasing this open, compact, multimodal reasoning model that approaches the frontier, we aim to catalyze research on mid-training curricula and lower operational barriers for privacy-preserving, cost-aware deployments across diverse organizational contexts.

Summary of Contributions. This work advances the state of efficient multimodal reasoning through four interconnected contributions that challenge conventional assumptions about scale, cost, and accessibility:

- **Compact and Deployable Frontier-level Models:** We show that relatively small models can achieve frontier-level performance, narrowing the gap to leading proprietary systems through training design rather than sheer parameter count. Their modest compute and memory footprint further makes them practical for on-premises deployment in constrained environments.
- **Efficient and Democratized Scaling:** Our staged continual pretraining recipe strengthens both textual and visual reasoning while remaining feasible under realistic budgets. Techniques such as depth upscaling (capacity expansion without pretraining from scratch), selective loss computation, and checkpoint averaging improve efficiency and stability. This maximizes the potential of the base model and demonstrates that approaching state-of-the-art performance is not restricted to organizations with tens of thousands of GPUs.
- **Comprehensive Cross-Modal Evaluation:** We demonstrate strong results across text reasoning benchmarks (e.g. AIME'25: 88%, IFBench: 62%, τ^2-Bench Telecom: 68%) and multimodal tasks (e.g. MMMU: 70.2%, MathVista: 75.5%, CharXiv: 88.2%). These results underscore broad reasoning competence across domains, supported by both internal evaluation and third-party validation.
- **Open-Source Compact Multimodal Foundation Model:** To our knowledge, this is the first openly released compact multimodal reasoning model operating at frontier level. We provide weights, training recipes, and evaluation artifacts under a permissive license, democratizing access and enabling reproducibility and further study

We structure the report as follows: Section 2 introduces the multimodal architecture integration, including depth upscaling and cross-modal training procedures. Section 3 presents our staged continual pretraining, detailing the data and training methodology across the foundational reasoning, image understanding phase and the visual reasoning phase. Section 4 describes our high-quality data curation and training for supervised fine-tuning. Section 5 describes our evaluation methodology for text and image benchmarks, incorporating internal validation and third-party assessment. Section 6 reports comprehensive evaluations across text and multimodal benchmarks, along with additional analysis. Finally, Section 7 concludes the report with possible directions for future work.

2 Architecture and Model Upscaling

Base Model To enable multimodal capabilities in a compute efficient manner, we build on Pixtral-12B-Base-2409 [9][3]. Pixtral follows the LLaVA architecture [11], consisting of a vision encoder connected to a multimodal decoder through a two-layer fully connected projection network.

Depth Upscaling Following the approach adopted in Apriel-Nemotron-15B-Thinker [12], we first upscale the base model via depth upscaling to balance compute, latency, and performance, while maintaining deployability on a single high-end GPU. To upscale the multimodal model, we

[3]https://huggingface.co/mistralai/Pixtral-12B-Base-2409 . We used a version from Unsloth, which is no longer available at https://huggingface.co/unsloth as of this writing.

first upscale the decoder by increasing the number of hidden layers from 40 to 48, training on a large corpus of text tokens. Half of these tokens serve as replay data, and the rest are drawn from diverse domains including high-quality web content, technical literature, mathematical problem sets, programming code, and StackExchange discussions.

Projection network realignment Next, the projection network is realigned by training on data from image captioning datasets, multimodal instruction-response pairs, and document understanding scenarios. During this stage, the vision encoder and the decoder remain frozen.

Training Setup Both depth upscaling and projection network realignment were trained with a sequence length of 8192 (with sequence packing) and a learning rate of 5e-5 with linear decay. The weights of six equispaced intermediate checkpoints from the depth upscaling stage were averaged in equal proportions before projection network realignment. The final checkpoint obtained from the projection network realignment stage was used for subsequent stages of training.

3 Continual Pretraining (CPT)

To strengthen the foundational capabilities of the base model, after scaling up the model we further enhance its textual and visual reasoning capabilities with multimodal continual pretraining (CPT). The CPT process is divided into two stages: the first focuses on enhancing the model's textual reasoning and image understanding capabilities, while the second aims at further improving its visual reasoning capabilities. The two stages are described in detail below.

3.1 CPT Stage 1

Foundational Reasoning and Multimodal Data The first stage involves training on a dataset that comprises of 50% text-only tokens covering mathematical and scientific reasoning, coding tasks, and general knowledge; 20% tokens replayed from the decoder upscaling stage; and 30% multimodal tokens drawn from data on document understanding, chart understanding and reasoning, image captioning, long-form image descriptions, OCR-related tasks, and reasoning over mathematical and logical problems in visual contexts.

Training Setup Since this stage involved data addressing foundational vision capabilities of the model, the vision encoder, projection network, and decoder were kept unfrozen. The training was performed at a sequence length of 32768 (with sequence packing) and a learning rate of 5e-5 with cosine decay and 10% warmup. Loss was computed on all the tokens in the sequence. The weights of three equispaced intermediate checkpoints were averaged in equal proportions to form the final checkpoint from this stage.

3.2 CPT Stage 2

Targeted Visual Reasoning Data via Synthetic Augmentation To further strengthen visual reasoning after the first stage, we construct a targeted multimodal dataset by employing a synthetic data generation pipeline to large collections of raw images. The pipeline transforms each image into one or more task-centric training samples. This shifts the original image distribution to a custom curriculum that encourages the model to learn spatial structure, compositionality, and fine-grained perception that transfer to more complex visual reasoning. The following are the primary categories we target:

- **Image Reconstruction**: Learn holistic scene priors and part–whole reasoning by masking image regions.
- **Visual Matching**: Improve correspondence, retrieval, and fine-grained discrimination by matching cropped or augmented anchors to candidates across views or images.
- **Object Detection**: Strengthen grounding and localization by identifying object presence and approximate location.
- **Counting**: Enhance the ability to count and distinguish specific visual elements by querying total or category-specific counts.

Data Hygiene and Difficulty Control For each task, we modulate difficulty through controlled augmentation depending on the task. This helps the model learn a more robust spatial reasoning, compositional understanding, and precise grounding, while remaining broadly applicable across diverse visual domains.

Training Setup In this stage, the vision encoder was frozen, with just the projection network and decoder updated during training. The training was performed at a sequence length of 16384 (with sequence packing), and learning rate of 1e-5 with cosine decay and 10% warmup. For samples having an instruction-response format, we compute loss only on the reponses in this stage. The final checkpoint from this stage was considered as the base model for future stages.

Evaluating effectiveness of Stage-2 To evaluate the effectiveness of the second CPT stage, we conducted two small-scale SFT experiments, initialized from the final checkpoints of CPT Stage 1 and Stage 2, using 17k text-based reasoning samples designed to mimic our full SFT setup. Table 1 presents a comparative evaluation of SFT after the two CPT stages across a range of multimodal and math-focused vision benchmarks (see 5.2). Stage 2 consistently improves performance over Stage 1, with notable gains on tasks such as MathVerse (Vision Dominant: +9.65 points), CharXiv (Descriptive: +5.98 points), and AI2D Test (+3.7 points). These results demonstrate that CPT Stage 2 provides substantial benefits for visual reasoning tasks.

Benchmark	SFT on CPT Stage 1	SFT on CPT Stage 2
MMMU (Val)	64.11	69.10
MathVision	44.4	47.36
MathVista	71.8	74.10
MathVerse (Vision Dominant)	53.04	62.69
MathVerse (Text Dominant)	70.81	78.42
MMStar	61.80	66.30
CharXiv (descriptive)	80.22	86.20
CharXiv (reasoning)	43.5	48.00
AI2D Test	78.1	81.8

Table 1: Evaluating effectiveness of CPT Stage 2 across multiple vision benchmarks.

4 Supervised Fine Tuning (SFT)

Following the upscaling and continual pretraining stages, which yielded a base model with strong reasoning capabilities, we performed Supervised Fine-Tuning (SFT) to develop the model into a full-fledged reasoner.

Data Curation Given compute constraints that preclude training a larger annotator model or scaling post-training runs from a cold-start SFT, we emphasize curating and synthesizing high-quality, high-signal prompts and employ open-source models as annotators. We curate and synthesize a diverse set of prompts[13]. Small-scale ablations using DeepSeek-R1-0528 [8] and gpt-oss-120b [14] presented in Table 2 show minimal performance differences between annotators for our base model. We therefore adopt gpt-oss-120b as our annotator model due to its greater compute efficiency. For verifiable domains, such as Math, Coding and Science, we follow the synthetic data generation methodology in [15] to synthesize high quality, execution verifiable data samples across domains starting from a seed taxonomy and samples, and evolving iteratively toward more complex scenarios [16]. That said, this release prioritized performance, with safety mitigations included but not pursued to the same depth.

To ensure the highest data quality and maximize sample efficiency, we invested significantly in a comprehensive data processing pipeline. We followed a multi-step filtering process that included rigorous de-duplication to enhance data diversity, content filtering to remove unsafe or inappropriate material, and heuristic filtering to remove low-quality samples. Following this initial cleaning, we verified the data's correctness using LLM-as-Judge and execution-based verification where applicable, implementing rejection sampling to discard incorrect or low-quality instruction-response pairs. This verification stage also included format-based checks to ensure structural correctness for samples

where a specific output like JSON or XML was expected. Finally, all samples were processed with consistent formatting using our custom chat template, and a decontamination stage to remove any samples overlapping with the benchmarks.

Benchmark	Annotator: DeepSeek-R1-0528	Annotator: gpt-oss-120b
GPQA Diamond	67.67	65.82
AIME'24	80.66	80.67
AIME'25	75.33	74

Table 2: Performance on benchmarks relevant to a small scale SFT set, annotated with DeepSeek-R1-0528 and gpt-oss-120b. Overall, we find the benchmark performance to be similar for both annotator models.

Data Composition We use a large and diverse dataset containing millions of high-quality instruction–response pairs. Each response contained explicit reasoning steps leading to the final response, followed by the final response itself. The final dataset comprised samples from domains including mathematical reasoning, coding, scientific reasoning, tool calling, generic reasoning and knowledge-intensive samples, conversations, instruction-following, security, content moderation, and robustness. This ensures that the model is both capable and reliable across diverse scenarios.

Training We first performed an initial SFT for 4 epochs at a sequence length of 32768 (with sequence packing) and a learning rate of 1e-5 with cosine decay. To further improve performance, we conducted two smaller SFT runs on top of the large-scale SFT: (1) trained with a stratified 25% subset of the full dataset for 4 more epochs at the same sequence length and (2) a longer-sequence run at 49,152 sequence length, using 25k samples between 32768 and 49152 tokens and 100k samples $\leq$ 32768 tokens, randomly drawn from the original mix. The models from these two smaller runs were merged by averaging their weights in equal proportions to produce the final APRIEL-1.5-15B-THINKER checkpoint. These smaller runs provided inexpensive gains in overall and long-context performance, and the merge balanced the benefits of both. As this phase consisted entirely of text data, only the decoder was updated. In all SFT runs, loss was computed only on response, and the chat template was applied to all samples.

5 Evaluation Methodology

5.1 Text evaluation

To report the evaluation results for the Apriel-1.5-15B-Thinker, we relied on the **Artificial Analysis Intelligence Index**, an independent combination metric for measuring general intelligence in large language models (LLMs). Using this external source ensures that results are unbiased and comparable across organizations, as the scoring is not influenced by in-house test sets or proprietary metrics. Although our internal evaluation also show very similar metrics as reported by Artificial Intelligence.

The Artificial Analysis Intelligence Index is notable for its *breadth and methodological rigor*. It aggregates results from ten heterogeneous benchmarks, with each benchmark targeting a distinct dimension of model capability:

- **MMLU-Pro** – advanced multi-domain knowledge and reasoning
- **GPQA Diamond** – graduate-level problem solving in science/engineering
- **Humanity's Last Exam** – multi-disciplinary high-difficulty reasoning
- **LiveCodeBench** – functional correctness in code generation
- **SciCode** – scientific computing and reasoning tasks
- **AIME 2025** – competition-level mathematics
- **IFBench** – instruction following and compliance
- **AA-LCR** – long-context reasoning
- **Terminal-Bench Hard** – real-world Linux shell execution and system tool use in end-to-end tasks

- τ^2**-Bench Telecom** – specialized domain evaluation in applied tasks

By normalizing across domains, evaluation difficulty, and inter-benchmark variance, the Index provides a *holistic measure of intelligence* rather than domain-specific performance. This methodology makes it a well-respected yardstick across academia and industry[4].

5.2 Vision evaluation

For the vision component, we focus on image evaluations since our training has been conducted primarily with images. We evaluate vision capabilities using the VLMEvalKit[17] toolkit, which standardizes data loading, prompting, post-processing, and scoring for reproducible comparisons across diverse tasks. Our benchmark suite spans the following areas:

- **General Multi-modal Reasoning**
  - MMMU[18]: Multi-modal understanding benchmark focusing on evaluating visual knowledge and reasoning.
  - MMMU-Pro[19]: Enhanced Multi-modal understanding benchmark focusing on evaluating visual knowledge and reasoning.
  - MMStar[20]: Vision-indispensable benchmark focusing on tasks that cannot be solved with only knowledge or without using the image.
- **Visual Logic**
  - LogicVista[21]: Multi-modal logical reasoning benchmark targeting different reasoning skill types in visual contexts.
- **Mathematical Vision and Quantitative Reasoning**
  - MathVision[22]: Mathematical reasoning within visual contexts.
  - MathVista[23]: Benchmark combining challenges from various visual and mathematical tasks.
  - MathVerse[24]: Mathematical benchmark measuring model performance across different levels of information content across multiple modalities.
- **Document/Diagram Understanding**
  - CharXiv[25]: Benchmark measuring descriptive and reasoning question answering capabilities across basic and complex chart elements respectively.
  - AI2D[26]: Diagram understanding benchmark.
- **Open-domain Vision–Language Reasoning**
  - BLINK[27]: Benchmark measuring performance on various visual perception tasks.

For each dataset, we adhere to official or community-standard protocols as implemented in VLMEvalKit and adopt consistent prompts and inference settings across models to ensure fair comparisons.

6 Results and Observations

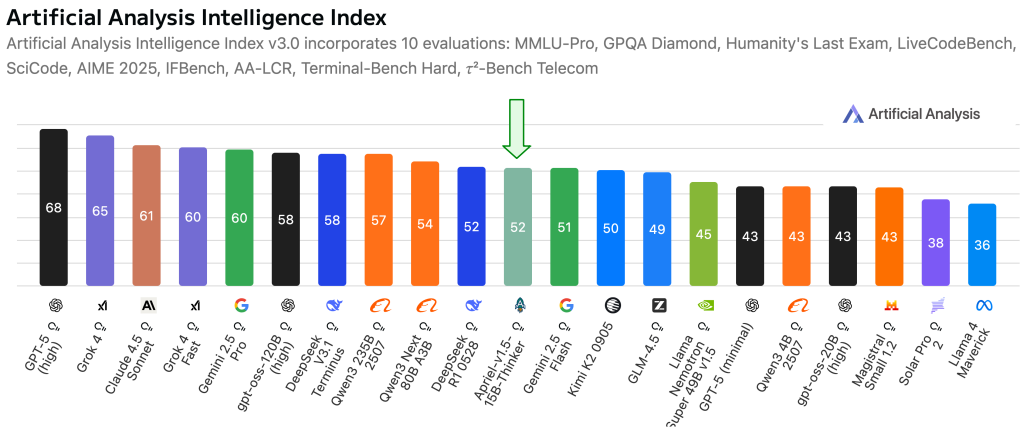
6.1 Text Benchmarks

Figure 2 shows Apriel-1.5-15B-Thinker achieves a score of **52**. It surpasses larger open-weight systems such as **Llama Nemotron Super 49B v1.5 (45)** and **gpt-oss-20B (43)**, while performing comparably to models such as **DeepSeek-R1-0528** and **Gemini-2.5-Flash**.

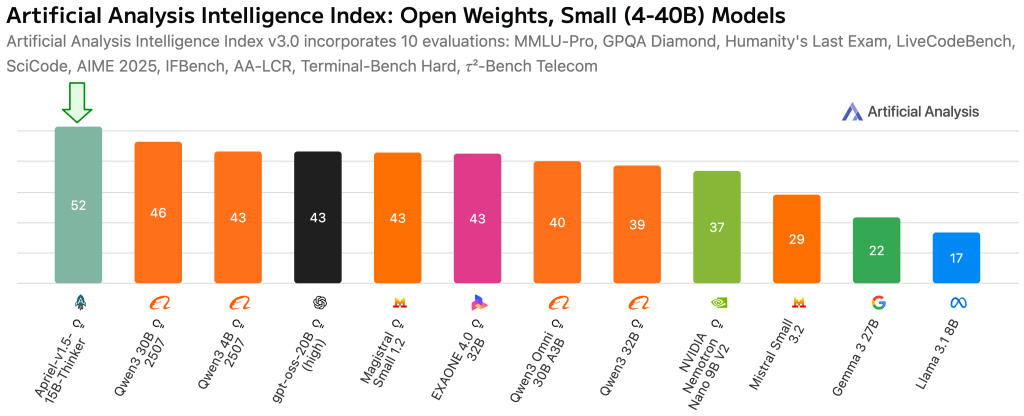
The aggregated results across the AA intelligence index offer a consolidated view of the model's reasoning performance. The model demonstrates state-of-the-art accuracy on several challenging evaluations, achieving 87% on AIME2025, 62% on IF-Bench, and 68% on τ^2Bench (Telecom). These results highlight its strong mathematical reasoning, robust instruction-following, and domain-specific problem-solving capabilities, where it consistently outperforms significantly larger open-source baselines.

On TerminalBench-Hard, the model achieves a score of 10%. This benchmark evaluates performance in terminal-based environments across a wide spectrum of technical tasks. Despite its smaller

[4]https://artificialanalysis.ai/methodology/intelligence-benchmarking



Apriel-1.5-15B-Thinker compared with state-of-the-art LLMs.



Apriel-1.5-15B-Thinker compared with SOTA open-source models.

Figure 2: Apriel-1.5-15B-Thinker ranks first in Artificial Analysis Intelligence index among the SOTA small open-source models and delivers performance competitive to larger open-source and proprietary models (as of September 26th, 2025).

parameter count, our model performs competitively with much larger proprietary systems such as GPT-4.1 and Gemini 2.5 Flash (both at 13%) and Qwen3-250B (13%). Notably, it surpasses strong open-source peers of comparable size, including `gpt-oss-20b`, which scores 6%.

These results underscore the efficiency and competitiveness of the model, offering strong reasoning and agentic capabilities without the overhead of massive parameter counts.

The detailed breakdown in Table 3 reports the scores provided by *Artificial Analysis*, with the exception of Apriel-1.5-15B-Thinker (SELF-REPORTED), where evaluations were conducted internally.

The divergence in internal evaluation arises primarily from differences in benchmarking conditions. In particular, the judging models for AA-LCR and τ^2Bench differ (`GPT-4.1` versus `Qwen3-235B-A22B-2507`). For `AIME2025`, no language model equality checker was employed. Moreover, `TerminalBench` was executed under a shorter timeout constraint, further contributing to lower scores.

Figure 3 demonstrates performance relative to scale, and falls within the *"most attractive quadrant"* – the region where models combine moderate scale with disproportionately high performance. This placement underscores Apriel-1.5-15B-Thinker's *superior cost-to-intelligence trade-off*, offering reduced compute requirements and faster inference while maintaining robust general capabilities.

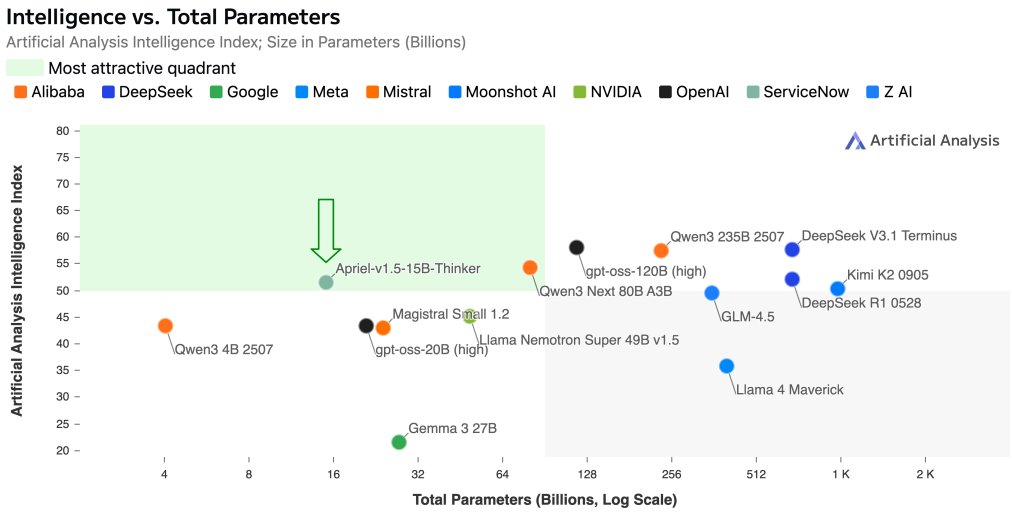


Figure 3: Artificial Analysis Intelligence Index vs. Total Parameters (log scale). Apriel-1.5-15B-Thinker lies in the "most attractive quadrant."

Benchmark	Artificial Analysis Intelligence Index	MMLU-Pro	GPQA Diamond	HLE	LiveCodeBench	SciCode	AIME2025	IF-Bench	AA-LCR	TerminalBench Hard	τ^2-Bench Telecom
Proprietary Models											
GPT-5 (High)	68.47	87.1	85.4	26.5	84.6	42.9	94.3	73.1	75.6	30.5	84.8
Grok 4	65.26	86.6	87.7	23.9	81.9	45.7	92.7	53.7	68	37.6	74.9
Claude 4.5 Sonnet	61.29	87.5	83.4	17.3	57.7	44.7	88	57.3	65.7	33.3	78.1
Grok 4 Fast	60.25	85	84.7	17	83.2	44.2	89.7	50.5	64.7	17.7	65.8
Gemini 2.5 Pro	59.59	86.2	84.4	21.1	80.1	42.8	87.7	48.7	66	24.8	54.1
Claude 4.1 Opus	59.27	88	80.9	11.9	65.4	40.9	80.3	55.4	66.3	32.1	71.4
Claude 4 Sonnet	56.51	84.2	77.7	9.6	65.5	40	74.3	54.7	64.7	29.8	64.6
Magistral Medium 1.2	52.05	81.5	73.9	9.6	75	39.2	82	43	51.3	12.8	52
Gemini 2.5 Flash	51.18	83.2	79	11.1	69.5	39.4	73.3	50.3	61.7	12.8	31.6
GPT-5 (Minimal)	43.48	80.6	67.3	5.4	55.8	38.8	31.7	45.6	25	17.7	67
Large Open Weight Models											
gpt-oss-120B (High)	57.98	80.8	78.2	18.5	65.3	36.2	93.4	69	50.7	22	65.8
DeepSeek v3.1 Terminus	57.71	85.1	79.2	15.2	79.8	40.6	89.7	57	65	28.4	37.1
Qwen3 235B 2507	57.47	84.3	79	15	78.8	42.4	91	51.2	67	12.8	53.2
DeepSeekR1 0528	52.01	84.9	81.3	14.9	77	40.3	76	39.6	54.7	14.9	36.5
Kimi K2 0905	50.4	81.9	76.7	6.3	61	30.7	57.3	41.7	52.3	22.7	73.4
GLM 4.5	49.44	83.5	78.2	12.2	73.8	34.8	73.7	44.1	48.3	21.3	24.6
GLM 4.5 Air	48.81	81.5	73.3	6.8	68.4	30.6	80.7	37.6	43.7	19.1	46.5
Llama 4 Maverick	35.8	80.9	67.1	4.8	39.7	33.1	19.3	43	46	6.4	17.8
Small Open Weight Models											
Qwen3 Next 80B A3B	54.32	82.4	75.9	11.7	78.4	38.8	84.3	60.7	60.3	9.2	41.5
gpt-oss-20B (High)	43.27	73.6	61.7	8.5	57.2	35.4	61.7	60.5	18.7	5.7	49.7
Llama Nemotron Super 49B v1.5	45.22	81.4	74.8	6.8	73.7	34.8	76.7	37	34	5	28.1
Qwen3 4B 2507	43.36	74.3	66.7	5.9	64.1	25.6	82.7	49.8	37.7	1.4	25.4
Magistral Small 1.2	42.97	76.8	66.3	6.1	72.3	35.2	80.3	44.4	16.3	4.3	27.8
exaone 4.0 32B	42.64	81.8	73.9	10.5	74.7	34.4	80	36.3	14	3.5	17.3
Apriel-1.5-15B-Thinker	51.57	77.3	71.3	12	72.8	34.8	87.5	61.7	20	9.9	68.4
Apriel-1.5-15B-Thinker (SELF-REPORTED)	50	76.48	70.61	11	71.6	36.46	83.67	60.45	26.33	5.7	57.8

Table 3: Evaluation (pass@1 or accuracy) on benchmarks with **maximum reasoning**, as applicable: MMLU-Pro, GPQA Diamond, HLE, LiveCodeBench, SciCode, AIME2025, IF-Bench, AA-LCR, TerminalBench-Hard, and τ^2-2Bench. Orange = proprietary models, blue = >50B open-weight models, green = <50B open-weight models.

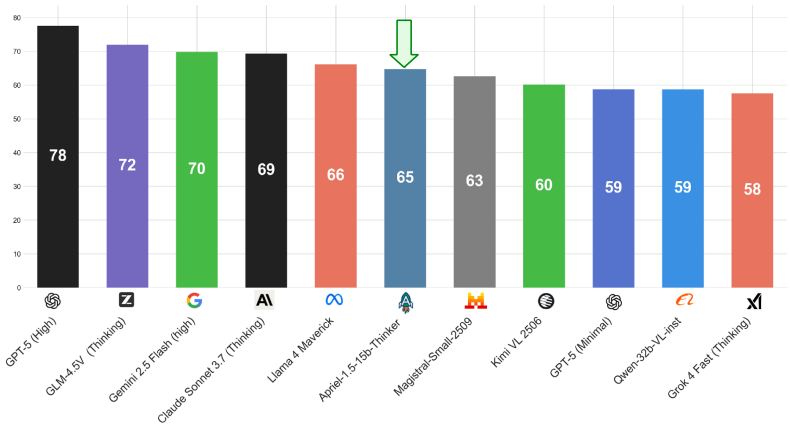


Figure 4: Average performance across the benchmark suite (higher is better). The chart aggregates scores from MMMU [18], MMMU-Pro [28], LogicVista [21], MathVision [22], MathVista [23], MathVerse [24], MMStar [20], CharXiv [25], AI2D [26], and BLINK [27].

The findings support the broader trend that *smaller, efficiently trained models can close the gap with frontier models*.

6.2 Vision Benchmarks

Figure 4 summarizes model-level averages over the full image benchmark suite, highlighting overall capability and robustness. Detailed per-benchmark, per-model scores are reported in Table 4, where higher values indicate better performance. Together, these views provide both a high-level comparison and fine-grained insight into strengths and weaknesses across categories.

Benchmark	MMMU	MMMU-PRO		LogicVista	MathVision	MathVista	MathVerse		MMStar	CharXiv		AI2D	BLINK	Average
	Val	10 C	vision				Vision-dom	Text-dom		Des	R	test		
Proprietary Models														
GPT-5 (High)	81.33	74.73	66.93	69.35	67.10	83.30	79.82	84.64	77.74	91.25	71.50	90.05	70.22	77.54
GPT-5 (Minimal)	66.66	66.06	57.68	44.51	35.52	61.20	39.84	43.78	63.60	82.45	52.80	85.16	64.59	58.76
Grok 4 Fast (Thinking)	70.11	61.61	22.94	47.42	48.35	68.20	54.69	72.20	64.80	68.15	33.50	81.86	54.39	57.56
Claude Sonnet 3.7 (Thinking)	73.66	64.5	60.11	69.12	50.32	74.60	56.09	69.28	70.00	93.27	70.90	84.19	64.49	69.27
Gemini 2.5 Flash (High)	70.66	67.86	56.76	63.75	59.21	78.50	70.68	78.80	73.86	83.60	56.50	82.09	65.64	69.84
Large Open Weight Models														
Llama 4 Maverick	72.22	63.41	54.45	58.38	43.09	72.60	63.32	70.30	71.00	87.70	55.00	85.78	62.28	66.12
GLM-4.5V (Thinking)	74.33	64.16	61.50	63.53	59.53	83.60	68.65	77.41	74.46	90.80	63.00	87.75	66.59	71.95
Small Open Weight Models														
Qwen 2.5 VL 32B-Instruct	64.40	51.90	45.02	52.15	38.15	75.90	51.64	62.69	66.30	72.50	39.10	81.60	61.90	58.71
Magistral Small-2509	70.00	55.72	46.06	54.80	55.92	73.40	53.68	67.76	65.13	82.27	52.90	82.18	54.48	62.64
Kimi VL 2506	61.44	48.09	42.89	46.97	50.00	79.80	52.91	67.13	68.93	78.70	45.60	82.70	56.33	60.11
Apriel-1.5-15B-Thinker	70.22	55.38	48.21	58.39	50.99	75.50	58.38	76.40	67.73	88.20	50.10	82.87	58.71	64.70

Table 4: Evaluation (pass@1 or accuracy, as applicable) on multimodal benchmarks covering general reasoning (MMMU, MMStar), visual logic (LogicVista), mathematical vision tasks (MathVision, MathVista, MathVerse), document-level understanding (CharXiv), diagram understanding (AI2D), and open-domain vision-language reasoning (BLINK). orange = proprietary models, blue = >100B open weight models, green = <50B open weight models.

As shown in Figure 4, the strongest overall performance is achieved by larger, advanced reasoning models, with notable leads on broad multimodal and STEM-oriented tasks. *Apriel-1.5-15B-Thinker* attains a solid position among the evaluated models, beating most similarly-sized and even larger open-weights vision-language models such as Kimi-VL-2506[29] and Qwen-2.5-VL-3B-Instruct[30]. Despite its smaller size (15B parameters), Apriel closely tracks much larger models such as Llama 4 Maverick[4] (400B parameters) and outperforms several larger proprietary baselines (e.g., GPT-5 Minimal [31], Grok 4 Fast [32]) in the overall average score.

The detailed breakdown in Table 4 indicates strong results on document-centric and diagram understanding benchmarks (e.g., CharXiv, AI2D), competitive performance on general multimodal reasoning (MMStar), and visual mathematical skills (MathVista). Apriel demonstrates particularly strong performance in document understanding tasks, achieving 88.20% on CharXiv descriptive tasks, which is the third-highest score after Claude and GPT-5 (High). Similarly, on MathVerse (Text

Dominant), Apriel scores 76.40%, outperforming several larger models including Claude Sonnet, Magistral, and Llama 4 Maverick[4]. The results suggest a pattern where Apriel performs better on tasks that combine visual inputs with substantial textual reasoning components, while showing moderate performance on purely visual reasoning tasks. For instance, on MMMU (70.22%), Apriel demonstrates competitive performance, whereas on vision-dominant tasks like MMMU-PRO (Vision) at 48.21%, it shows room for improvement. Most models, including Apriel, show stronger performance on structural understanding tasks (AI2D) and descriptive document tasks (CharXiv descriptive) compared to more complex reasoning tasks (CharXiv reasoning, LogicVista). The Apriel model demonstrates this pattern as well, with a notable 38.1 percentage point difference between its performance on CharXiv descriptive (88.20%) and CharXiv reasoning (50.10%) tasks, highlighting a gap between surface-level document comprehension and deeper contextual reasoning. Performance on the most demanding STEM-centric and visual logic tasks remains a key opportunity for further improvement.

7 Conclusion and Future Work

Our work shows that a 15-billion-parameter model can reach frontier-level reasoning by prioritizing data quality and a deliberately structured **mid-training** pipeline—staged continual pretraining (CPT) followed by large-scale, high-signal SFT without reinforcement learning or preference optimization. This data-centric recipe yields measurable gains during CPT (e.g., **+9.65** on MathVerse Vision-Dominant) and culminates in strong text-reasoning results on **AIME** and **GPQA**, while remaining competitive on multimodal benchmarks. Crucially, the final model operates on a **single-GPU**, delivering a favorable performance–efficiency trade-off that makes frontier-level reasoning accessible to organizations with limited computational infrastructure.

While this work focused primarily on text-based reasoning, the model's multimodal results offer a solid foundation for future development. Our next steps will extend multimodal capabilities more comprehensively and strengthen agentic abilities to support interactive workflows, with targeted alignment techniques where appropriate. Future development will continue to be guided by the core principles demonstrated here: strategic mid-training design, efficient architectural scaling and a continued focus on high-quality, targeted data.

References

[1] An Yang et al. "Qwen2 Technical Report". In: *arXiv preprint arXiv:2407.10671* (2024). URL: https://arxiv.org/abs/2407.10671.

[2] An Yang et al. "Qwen2.5 Technical Report". In: *arXiv preprint arXiv:2412.15115* (2024). URL: https://arxiv.org/abs/2412.15115.

[3] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[4] Meta AI. *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*. Apr. 2025. URL: https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[5] Gemini Team, Google. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: *arXiv preprint arXiv:2403.05530* (2024). URL: https://arxiv.org/abs/2403.05530.

[6] Gemini Team, Google DeepMind. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context*. Tech. rep. Google DeepMind, 2025. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.

[7] Anthropic. *Introducing Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed 2025-09-29. 2024.

[8] DeepSeek-AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.

[9] Pravesh Agrawal et al. *Pixtral 12B*. 2024. arXiv: 2410.07073 [cs.CV]. URL: https://arxiv.org/abs/2410.07073.

[10] Zehui Ren et al. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *arXiv preprint arXiv:2501.12948* (2025). URL: https://arxiv.org/abs/2501.12948.

[11] Haotian Liu et al. “Visual Instruction Tuning”. In: *NeurIPS*. 2023.

[12] Shruthan Radhakrishna et al. *Apriel-Nemotron-15B-Thinker*. 2025. arXiv: 2508.10948 [cs.LG]. URL: https://arxiv.org/abs/2508.10948.

[13] Sathwik Tejaswi Madhusudhan et al. *Millions scale dataset distilled from R1-32b*. https://huggingface.co/datasets/ServiceNow-AI/R1-Distill-SFT. 2025.

[14] OpenAI. *gpt-oss-120b & gpt-oss-20b Model Card*. 2025. arXiv: 2508.10925 [cs.CL]. URL: https://arxiv.org/abs/2508.10925.

[15] Aman Tiwari et al. “Auto-Cypher: Improving LLMs on Cypher generation via LLM-supervised generation-verification framework”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 623–640. ISBN: 979-8-89176-190-2. DOI: 10.18653/v1/2025.naacl-short.53. URL: https://aclanthology.org/2025.naacl-short.53/.

[16] Can Xu et al. *WizardLM: Empowering large pre-trained language models to follow complex instructions*. 2025. arXiv: 2304.12244 [cs.CL]. URL: https://arxiv.org/abs/2304.12244.

[17] Haodong Duan et al. *VLMEvalKit: An open-source toolkit for evaluating large multi-modality models*. 2024. arXiv: 2407.11691 [cs.CL]. URL: https://arxiv.org/abs/2407.11691.

[18] Xiang Yue et al. “MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI”. In: *arXiv preprint arXiv:2311.16502* (2023). URL: https://arxiv.org/abs/2311.16502.

[19] Xiang Yue et al. “MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark”. In: *arXiv preprint arXiv:2409.02813* (2024).

[20] Lin Chen et al. *Are We on the Right Way for Evaluating Large Vision-Language Models?* 2024. URL: https://arxiv.org/abs/2403.20330.

[21] Yijia Xiao et al. *LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts*. 2024. URL: https://arxiv.org/abs/2407.04973.

[22] Ke Wang et al. *Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset*. 2024. URL: https://arxiv.org/abs/2402.14804.

[23] Pan Lu et al. “MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts”. In: *arXiv preprint arXiv:2310.02255* (2023). URL: https://arxiv.org/abs/2310.02255.

[24] Renrui Zhang et al. *MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?* 2024. URL: https://arxiv.org/abs/2403.14624.

[25] Zirui Wang et al. “CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal Large Language Models”. In: *arXiv preprint arXiv:2406.18521* (2024). URL: https://arxiv.org/abs/2406.18521.

[26] Aniruddha Kembhavi et al. *A Diagram Is Worth A Dozen Images*. 2016. URL: https://arxiv.org/abs/1603.07396.

[27] Xingyu Fu et al. *BLINK: Multimodal Large Language Models Can See but Not Perceive*. 2024. URL: https://arxiv.org/abs/2404.12390.

[28] Xiang Yue et al. *MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark*. 2024. URL: https://arxiv.org/abs/2409.02813.

[29] Kimi Team et al. *Kimi-VL Technical Report*. 2025. arXiv: 2504.07491 [cs.CV]. URL: https://arxiv.org/abs/2504.07491.

[30] Shuai Bai et al. *Qwen2.5-VL Technical Report*. 2025. URL: https://arxiv.org/abs/2502.13923.

[31] OpenAI. *GPT-5 System Card*. White paper. 2025. URL: https://cdn.openai.com/gpt-5-system-card.pdf.

[32] X.AI Corp. *Grok-4 Fast System Card*. White paper. 2025. URL: https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf.

8 Contributions and Acknowledgments

Core Contributors

Shruthan Radhakrishna, Aman Tiwari, Aanjaneya Shukla, Masoud Hashemi, Rishabh Maheshwary, Shiva Krishna Reddy Malay, Jash Mehta, Pulkit Pattnaik, Saloni Mittal, Khalil Slimi, Kelechi Ogueji, Akintunde Oladipo, Soham Parikh, Oluwanifemi Bamgbose

Secondary Contributors

Toby Liang, Ahmed Masry, Khyati Mahajan, Sai Rajeswar Mudumba, Vikas Yadav

Leads & Management

Sathwik Tejaswi Madhusudhan	Technical co-lead (Mid-Training and Post-Training)
Torsten Scholak	Technical co-lead (Pre-Training and Architecture)
Sagar Davasam	Applied Research Manager
Srinivas Sunkara	VP, Applied Research
Nicholas Chapados	VP, AI Research

Upstream Contributors

We gratefully acknowledge the contributors whose work on Apriel-Nemotron-15B-Thinker and related projects was reused as part of the current work:

Gopal Sarda, Anil Turkkan, Shashank Maiya, Dhruv Jhamb, Jishnu S Nair, Akshay Kalkunte, Bidyapati Pradhan, Surajit Dasgupta, Jaykumar Kasundra, Anjaneya Praharaj, Sourabh Surana, Lakshmi Sirisha Chodisetty, Abhigya Verma, Abhishek Bharadwaj, Nandhakumar Kandasamy, Naman Gupta, Segan Subramanian

We also thank Anil Madamala for leading evaluations and benchmarking, and Segan Subramanian and Vipul Mittal for leading data infrastructure.