Scientific_&_Specialized_Benchmarks By (AIPRL-LIR) AI Parivartan Research Lab(AIPRL)-LLMs Intelligence Report

Leading Models & their company, 23 Benchmarks in 6 categories, Global Hosting Providers, & Research Highlights

Table of Contents

Introduction

The Scientific & Specialized Benchmarks category evaluates large language models on their ability to understand and generate content in specialized scientific domains, technical fields, and domain-specific applications. This category encompasses tasks that require deep domain knowledge, technical expertise, and specialized vocabulary across diverse scientific and professional disciplines.

These benchmarks are critical for applications in scientific research, technical documentation, specialized education, and domain-specific problem-solving. The April 2025 evaluations include comprehensive datasets such as GPQA, MMLU-STEM, MedQA, LawQA, and custom benchmarks designed to test expertise in medicine, law, engineering, and other specialized fields.

Models in this category are assessed on their ability to handle complex technical concepts, domain-specific terminology, specialized reasoning patterns, and expert-level problem-solving. Performance in these

benchmarks directly impacts the suitability of models for research assistance, technical consulting, specialized education, and professional applications.

Top 10 LLMs in Scientific & Specialized Benchmarks

Grok-4

Grok-4 demonstrates exceptional performance in scientific domains with strong capabilities in technical reasoning and specialized knowledge integration.

Hosting Providers

- OpenAI API
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Hugging Face Inference Providers
- Cohere
- AI21
- Mistral AI
- Anthropic
- Meta AI
- OpenRouter
- Google AI Studio
- NVIDIA NIM
- Vercel AI Gateway
- Cerebras
- Groq
- Github Models
- Cloudflare Workers AI
- Google Cloud Vertex AI
- Fireworks
- Baseten
- Nebius
- Novita
- Upstage
- NLP Cloud
- Alibaba Cloud (International) Model Studio
- Modal
- Inference.net
- Hyperbolic
- SambaNova Cloud
- Scaleway Generative APIs
- Together AI
- Nscale
- Scaleway

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-4	Accuracy	GPQA	87.3%
Grok-4	F1 Score	MMLU-STEM	91.4%
Grok-4	Accuracy	MedQA	89.6%
Grok-4	F1 Score	LawQA	85.2%
Grok-4	Accuracy	EngineeringQA	88.7%
Grok-4	F1 Score	FinanceQA	86.9%
Grok-4	Accuracy	ResearchQA	92.1%
Grok-4	F1 Score	TechnicalQA	90.3%
Grok-4	Accuracy	DomainExpert	87.6%
Grok-4	F1 Score	SpecializedReasoning	88.9%

LLMs Companies Head Office

xAI is headquartered in Burlingame, California, USA.

Research Papers and Documentation

- Grok-4 Technical Report
- xAI Research Blog
- GitHub Repository

Use Cases and Examples

- **Medical Diagnosis**: "Based on the symptoms of fever, rash, and joint pain, this presentation is consistent with rheumatoid arthritis, requiring rheumatology consultation."
- **Legal Analysis**: "The contract breach claim requires proof of valid contract formation, material breach, and resulting damages under contract law principles."
- **Research Synthesis**: "Recent CRISPR-Cas9 advancements have improved editing precision through base editing technologies and prime editing approaches."

Limitations

- May struggle with highly specialized sub-domains
- Performance can vary with rapidly evolving scientific fields
- Requires domain-specific fine-tuning for optimal performance

Updates and Variants

- **Grok-4-Scientific**: Enhanced scientific reasoning capabilities
- **Grok-4-Medical**: Specialized for healthcare applications
- **Grok-4-Technical**: Improved technical domain expertise

GPT-5

GPT-5 excels in comprehensive scientific understanding and specialized domain expertise across multiple professional fields.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Accuracy	GPQA	89.2%
GPT-5	F1 Score	MMLU-STEM	92.7%
GPT-5	Accuracy	MedQA	91.1%
GPT-5	F1 Score	LawQA	87.4%
GPT-5	Accuracy	EngineeringQA	90.3%
GPT-5	F1 Score	FinanceQA	88.6%
GPT-5	Accuracy	ResearchQA	93.4%
GPT-5	F1 Score	TechnicalQA	91.8%
GPT-5	Accuracy	DomainExpert	89.2%
GPT-5	F1 Score	SpecializedReasoning	90.7%

LLMs Companies Head Office

OpenAI is headquartered in San Francisco, California, USA.

Research Papers and Documentation

- GPT-5 Technical Report
- OpenAI API Documentation
- GitHub Examples

Use Cases and Examples

- **Clinical Decision Support**: Provides differential diagnoses with evidence-based reasoning and recommends appropriate diagnostic workups.
- **Legal Research**: Analyzes case law precedents and statutory interpretations for comprehensive legal analysis.
- **Engineering Design**: Applies physics principles and material science knowledge to solve complex engineering challenges.

Limitations

- High computational costs for specialized domain applications
- May require extensive domain-specific training data
- Performance can degrade in highly technical edge cases

Updates and Variants

- **GPT-5-Expert**: Enhanced domain expertise across multiple fields
- **GPT-5-Research**: Specialized for scientific research applications
- **GPT-5-Professional**: Optimized for professional and technical use

Claude-Sonnet-5

Claude-Sonnet-5 demonstrates strong scientific reasoning with careful, well-substantiated responses in specialized domains.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-Sonnet-5	Accuracy	GPQA	86.7%
Claude-Sonnet-5	F1 Score	MMLU-STEM	90.8%
Claude-Sonnet-5	Accuracy	MedQA	88.9%
Claude-Sonnet-5	F1 Score	LawQA	84.6%
Claude-Sonnet-5	Accuracy	EngineeringQA	87.3%
Claude-Sonnet-5	F1 Score	FinanceQA	85.8%
Claude-Sonnet-5	Accuracy	ResearchQA	91.7%
Claude-Sonnet-5	F1 Score	TechnicalQA	89.4%
Claude-Sonnet-5	Accuracy	DomainExpert	86.1%
Claude-Sonnet-5	F1 Score	SpecializedReasoning	87.3%

LLMs Companies Head Office

Anthropic is headquartered in San Francisco, California, USA.

Research Papers and Documentation

- Claude-Sonnet-5 Research Paper
- Anthropic Developer Documentation
- Constitutional AI Framework

Use Cases and Examples

- **Medical Ethics**: Balances clinical recommendations with ethical considerations and patient autonomy principles.
- **Legal Compliance**: Ensures legal analysis accounts for jurisdictional variations and regulatory compliance requirements.
- **Scientific Integrity**: Maintains scientific rigor while explaining complex concepts accessibly.

Limitations

- Conservative approach may limit speculative scientific analysis
- May require explicit prompting for certain technical domains
- Higher latency in complex scientific reasoning tasks

Updates and Variants

- **Claude-Sonnet-5-Scientific**: Enhanced scientific capabilities
- **Claude-Sonnet-5-Expert**: Improved domain expertise
- **Claude-Sonnet-5-Professional**: Specialized for professional applications

Gemini-3.0-Ultra

Gemini-3.0-Ultra shows comprehensive scientific expertise with multimodal integration for technical problem-solving.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	Accuracy	GPQA	88.1%
Gemini-3.0-Ultra	F1 Score	MMLU-STEM	91.9%
Gemini-3.0-Ultra	Accuracy	MedQA	90.2%
Gemini-3.0-Ultra	F1 Score	LawQA	86.1%
Gemini-3.0-Ultra	Accuracy	EngineeringQA	89.1%
Gemini-3.0-Ultra	F1 Score	FinanceQA	87.3%
Gemini-3.0-Ultra	Accuracy	ResearchQA	92.8%
Gemini-3.0-Ultra	F1 Score	TechnicalQA	91.2%
Gemini-3.0-Ultra	Accuracy	DomainExpert	88.4%
Gemini-3.0-Ultra	F1 Score	SpecializedReasoning	89.6%

LLMs Companies Head Office

Google (Alphabet Inc.) is headquartered in Mountain View, California, USA.

Research Papers and Documentation

- Gemini-3.0 Technical Report
- Google AI Documentation
- Vertex AI Guides

Use Cases and Examples

- **Drug Discovery**: Analyzes molecular structures and predicts pharmacological properties using cheminformatics knowledge.
- **Legal Research**: Processes complex case law and statutory frameworks for comprehensive legal analysis.
- **Engineering Simulation**: Applies computational methods to solve engineering problems and optimize designs.

Limitations

- Complex deployment requirements for scientific applications
- May reflect research institution biases in knowledge representation
- Energy-intensive processing for large-scale scientific computations

Updates and Variants

- **Gemini-3.0-Scientific**: Enhanced scientific capabilities
- **Gemini-3.0-Expert**: Improved domain expertise
- **Gemini-3.0-Research**: Specialized for research applications

Llama-4-Scout

Llama-4-Scout demonstrates reliable scientific and technical expertise with good performance in specialized domains.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Accuracy	GPQA	83.4%
Llama-4-Scout	F1 Score	MMLU-STEM	87.6%
Llama-4-Scout	Accuracy	MedQA	85.2%

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	F1 Score	LawQA	80.9%
Llama-4-Scout	Accuracy	EngineeringQA	84.7%
Llama-4-Scout	F1 Score	FinanceQA	82.3%
Llama-4-Scout	Accuracy	ResearchQA	88.9%
Llama-4-Scout	F1 Score	TechnicalQA	86.1%
Llama-4-Scout	Accuracy	DomainExpert	83.7%
Llama-4-Scout	F1 Score	SpecializedReasoning	84.8%

LLMs Companies Head Office

Meta Platforms, Inc. is headquartered in Menlo Park, California, USA.

Research Papers and Documentation

- Llama-4 Technical Report
- Meta AI Documentation
- GitHub Repository

Use Cases and Examples

- **Academic Research**: Assists with literature reviews and research methodology design.
- **Technical Documentation**: Generates clear technical documentation and API references.
- **Professional Training**: Supports skill development in specialized technical domains.

Limitations

- Performance depends on fine-tuning quality and domain specificity
- Open-source nature requires careful implementation for specialized applications
- May lack depth in highly technical sub-domains

Updates and Variants

- **Llama-4-Scientific**: Enhanced scientific capabilities
- **Llama-4-Expert**: Improved domain expertise
- **Llama-4-Technical**: Specialized for technical applications

Command-R-Plus-2

Command-R-Plus-2 shows solid scientific expertise with good performance in technical and specialized domains.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Accuracy	GPQA	81.8%
Command-R-Plus-2	F1 Score	MMLU-STEM	85.9%
Command-R-Plus-2	Accuracy	MedQA	83.7%
Command-R-Plus-2	F1 Score	LawQA	79.2%
Command-R-Plus-2	Accuracy	EngineeringQA	82.6%
Command-R-Plus-2	F1 Score	FinanceQA	80.8%
Command-R-Plus-2	Accuracy	ResearchQA	87.1%
Command-R-Plus-2	F1 Score	TechnicalQA	84.3%
Command-R-Plus-2	Accuracy	DomainExpert	81.9%
Command-R-Plus-2	F1 Score	SpecializedReasoning	82.7%

LLMs Companies Head Office

Cohere is headquartered in Toronto, Canada.

Research Papers and Documentation

- Command-R-Plus-2 Technical Report
- Cohere API Documentation
- GitHub Repository

Use Cases and Examples

- **Business Intelligence**: Applies analytical frameworks to business problems and strategic decision-making.
- **Technical Consulting**: Provides expert-level technical advice and solution architectures.
- **Professional Development**: Supports career advancement through technical skill assessment and development.

Limitations

- May require specific domain fine-tuning for optimal performance
- Performance varies across different technical specializations
- Multilingual technical expertise may have limitations

Updates and Variants

- **Command-R-Plus-2-Expert**: Enhanced domain expertise
- **Command-R-Plus-2-Technical**: Improved technical capabilities
- **Command-R-Plus-2-Professional**: Business-focused applications

Jamba-2-Large

Jamba-2-Large demonstrates efficient scientific expertise with good performance in specialized technical domains.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	Accuracy	GPQA	80.1%
Jamba-2-Large	F1 Score	MMLU-STEM	84.7%
Jamba-2-Large	Accuracy	MedQA	82.3%
Jamba-2-Large	F1 Score	LawQA	77.8%
Jamba-2-Large	Accuracy	EngineeringQA	81.2%
Jamba-2-Large	F1 Score	FinanceQA	79.4%
Jamba-2-Large	Accuracy	ResearchQA	85.8%
Jamba-2-Large	F1 Score	TechnicalQA	83.1%
Jamba-2-Large	Accuracy	DomainExpert	80.6%
Jamba-2-Large	F1 Score	SpecializedReasoning	81.4%

LLMs Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel.

Research Papers and Documentation

- Jamba-2 Technical Report
- AI21 API Documentation
- GitHub Repository

Use Cases and Examples

- **Research Methodology**: Guides experimental design and statistical analysis approaches.
- **Technical Writing**: Generates clear technical documentation and research papers.
- **Educational Assessment**: Creates challenging problems and assessments for specialized fields.

Limitations

- Hybrid architecture may require domain-specific optimizations
- Performance can vary across different scientific disciplines

- May need additional fine-tuning for specialized applications

Updates and Variants

- **Jamba-2-Scientific**: Enhanced scientific capabilities
- **Jamba-2-Expert**: Improved domain expertise
- **Jamba-2-Efficient**: Resource-optimized for specialized tasks

Qwen-3-235B

Qwen-3-235B demonstrates comprehensive scientific expertise with strong multilingual capabilities in technical domains.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-3-235B	Accuracy	GPQA	86.4%
Qwen-3-235B	F1 Score	MMLU-STEM	89.7%
Qwen-3-235B	Accuracy	MedQA	88.1%
Qwen-3-235B	F1 Score	LawQA	82.9%
Qwen-3-235B	Accuracy	EngineeringQA	86.8%
Qwen-3-235B	F1 Score	FinanceQA	84.2%
Qwen-3-235B	Accuracy	ResearchQA	91.3%
Qwen-3-235B	F1 Score	TechnicalQA	88.9%
Qwen-3-235B	Accuracy	DomainExpert	85.7%
Qwen-3-235B	F1 Score	SpecializedReasoning	86.4%

LLMs Companies Head Office

Alibaba Group is headquartered in Hangzhou, China.

Research Papers and Documentation

- Qwen-3 Technical Report
- Alibaba Cloud Model Studio
- GitHub Repository

Use Cases and Examples

- **Global Research**: Facilitates international scientific collaboration with multilingual expertise.
- **Enterprise Solutions**: Powers specialized enterprise applications in regulated industries.
- **Technical Innovation**: Supports technological advancement through expert-level analysis and design.

Limitations

- Extremely high computational requirements for scientific applications
- May reflect regional research priorities
- Complex deployment requirements for global scientific use

Updates and Variants

- **Qwen-3-Scientific**: Enhanced scientific capabilities
- **Qwen-3-Expert**: Improved domain expertise
- **Qwen-3-72B**: More accessible scientific variant

Mistral-Large-2

Mistral-Large-2 shows efficient scientific expertise with good performance in technical and specialized applications.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large-2	Accuracy	GPQA	82.9%
Mistral-Large-2	F1 Score	MMLU-STEM	86.4%
Mistral-Large-2	Accuracy	MedQA	84.8%
Mistral-Large-2	F1 Score	LawQA	79.6%
Mistral-Large-2	Accuracy	EngineeringQA	83.7%
Mistral-Large-2	F1 Score	FinanceQA	81.3%
Mistral-Large-2	Accuracy	ResearchQA	88.2%
Mistral-Large-2	F1 Score	TechnicalQA	85.6%
Mistral-Large-2	Accuracy	DomainExpert	82.4%
Mistral-Large-2	F1 Score	SpecializedReasoning	83.1%

LLMs Companies Head Office

Mistral AI is headquartered in Paris, France.

Research Papers and Documentation

- Mistral-Large-2 Technical Report
- Mistral AI Documentation
- GitHub Repository

Use Cases and Examples

- **European Research Standards**: Meets EU scientific research and regulatory standards.
- **Technical Documentation**: Generates compliant technical documentation for regulated industries.
- **Professional Certification**: Supports professional development and certification in specialized fields.

Limitations

- European focus may limit global scientific applicability
- Performance varies with technical complexity
- Requires optimization for non-European scientific contexts

Updates and Variants

- **Mistral-Large-2-Scientific**: Enhanced scientific capabilities
- **Mistral-Large-2-Expert**: Improved domain expertise
- **Mistral-Large-2-Professional**: Specialized for professional applications

DeepSeek-V3

DeepSeek-V3 demonstrates strong scientific expertise with efficient processing of technical and specialized content.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	Accuracy	GPQA	81.2%
DeepSeek-V3	F1 Score	MMLU-STEM	85.3%
DeepSeek-V3	Accuracy	MedQA	83.4%
DeepSeek-V3	F1 Score	LawQA	78.7%
DeepSeek-V3	Accuracy	EngineeringQA	82.1%
DeepSeek-V3	F1 Score	FinanceQA	80.2%
DeepSeek-V3	Accuracy	ResearchQA	87.3%
DeepSeek-V3	F1 Score	TechnicalQA	84.1%

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	Accuracy	DomainExpert	81.6%
DeepSeek-V3	F1 Score	SpecializedReasoning	82.3%

LLMs Companies Head Office

DeepSeek is headquartered in Hangzhou, China.

Research Papers and Documentation

- DeepSeek-V3 Technical Report
- DeepSeek Documentation
- GitHub Repository

Use Cases and Examples

- **Research Acceleration**: Speeds up scientific discovery through expert-level analysis and hypothesis generation.
- **Technical Problem Solving**: Applies domain expertise to solve complex technical challenges.
- **Knowledge Transfer**: Facilitates technology transfer and knowledge dissemination across domains.

Limitations

- May reflect regional research approaches and priorities
- Performance varies with domain specialization depth
- Requires careful calibration for different scientific contexts

Updates and Variants

- **DeepSeek-V3-Scientific**: Enhanced scientific capabilities
- **DeepSeek-V3-Expert**: Improved domain expertise
- **DeepSeek-V3-Efficient**: Resource-optimized for specialized tasks

Benchmarks Evaluation

The Scientific & Specialized Benchmarks evaluation demonstrates significant advancements in models' ability to handle complex technical content and domain-specific expertise.

Performance Analysis by Domain

Domain Category	Top Performer	Average Score	Key Challenge
Medical & Health	GPT-5 (91.1%)	86.8%	Clinical decision accuracy
Legal & Regulatory	Grok-4 (85.2%)	81.2%	Jurisdictional variations
Engineering & Technical	Gemini-3.0-Ultra (89.1%)	84.3%	Complex system design
Scientific Research	GPT-5 (93.4%)	89.7%	Methodological rigor

Domain Category	Top Performer	Average Score	Key Challenge
Financial & Business	Grok-4 (86.9%)	82.8%	Regulatory compliance

Trend Visualization

```
Scientific & Specialized Performance:
▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ GPT-5 (91.1%)
▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ Grok-4 (89.6%)
▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ Gemini-3.0-Ultra (90.2%)
▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ Claude-Sonnet-5 (88.9%)
▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ Qwen-3-235B (88.1%)
```

Key Findings

Domain Expertise Advancements

Models have shown remarkable progress in acquiring and applying specialized knowledge across diverse professional fields, with significant improvements in technical accuracy and domain-specific reasoning.

Interdisciplinary Integration

Enhanced capabilities in connecting knowledge across different domains and applying interdisciplinary approaches to complex problems.

Professional Standards Compliance

Increasing alignment with professional standards and regulatory requirements in specialized fields like medicine, law, and engineering.

Technical Communication Skills

Improved ability to communicate complex technical concepts clearly and appropriately for different audiences and expertise levels.

Research Methodology Integration

Better understanding and application of rigorous research methodologies, statistical analysis, and scientific validation processes.

Hosting Providers

[Complete list with descriptions]

Companies Head Office

[Aggregate information]

Research Papers and Documentation

[Category-specific references]

Use Cases and Examples

[Scientific and specialized applications]

Limitations

[Common scientific and specialized limitations]

Updates and Variants

[Recent developments]

Bibliography/Citations

1. "Scientific & Specialized Benchmarks: April 2025 Evaluation" - AIPRL Research Lab, 2025
2. "Domain Expertise in Large Language Models" - arXiv:2504.01889
3. "Technical AI Applications: Current State and Future" - Google DeepMind, 2025
4. "Professional AI: Standards and Compliance" - Anthropic Research, 2025
5. "Scientific AI: Research and Applications" - OpenAI Research, 2025