



The Right AI for What?

November 2025 • The right AI LLM model for every use case



23

Different Models



11

AI Providers



Your AI

Memory Aid



11.2025

Current Status

Created by Matthias Herbert (<https://www.linkedin.com/in/matthiasherbert/>)

Decision Matrix

Quick guide for choosing the right model

Highest Intelligence & Analysis

Gemini 3 Pro

Complex Agents/Computer Control

Claude Sonnet 4.5

Coding & Refactoring

Claude Sonnet 4.5

Qwen 2.5 Coder (Local)

Science & Mathematics

o3 / o3-mini

DeepSeek R1

Processing Millions of Documents

Gemini 3 Pro (Quality)

Llama 4 Scout (Cost)

Ultra Budget

DeepSeek V3.2-Exp

1. General Purpose / Conversational AI

Google Gemini 3 Pro

OpenAI GPT-5.1

Edit with Lovable

^ More Details

Currently leading LLM (Release: Nov 18, 2025, Source: Google DeepMind). GPQA Diamond: 91.9%, Humanity's Last Exam: 37.5%, ARC-AGI-2: 31.1%, MMMU-Pro: 81%, Video-MMMU: 87.6%. Context window: up to 1M Tokens. Multimodal: Text, Image, Video, Audio, Code. Pricing: not fully publicly available

Currently leading LLM (Release: Nov 18, 2025, Source: Google DeepMind). GPQA Diamond: 91.9%, Humanity's Last Exam: 37.5%, ARC-AGI-2: 31.1%, MMMU-Pro: 81%, Video-MMMU: 87.6%. Context window: up to 1M Tokens. Multimodal: Text, Image, Video, Audio, Code. Pricing: not fully publicly available

^ More Details

New flagship version (Release: Nov 12, 2025, Source: OpenAI) with GPT-5.1 Instant & GPT-5.1 Thinking. GPQA Diamond: 88.1%, SWE Bench: 76.3%. Improved conversational quality. Pricing: not fully publicly available

New flagship version (Release: Nov 12, 2025, Source: OpenAI) with GPT-5.1 Instant & GPT-5.1 Thinking. GPQA Diamond: 88.1%, SWE Bench: 76.3%. Improved conversational quality. Pricing: not fully publicly available

^ More Models (2)

Google Gemini 2.5 Pro

^ More Details

Robust multimodal capabilities (Source: Google DeepMind). GPQA Diamond: 86.4%, MMMLU: 89.2%. Strong performance in creative and analytical tasks. Pricing: not fully publicly available

Robust multimodal capabilities (Source: Google DeepMind). GPQA Diamond: 86.4%, MMMLU: 89.2%. Strong performance in creative and analytical tasks. Pricing: not fully publicly available

Claude Sonnet 4

^ More Details

Balanced performance for production use (Source: Anthropic). Excellent instruction following. Release: 2024. Pricing: ~\$3/\$15 per 1M Tokens

Balanced performance for production use (Source: Anthropic). Excellent instruction following. Release: 2024. Pricing: ~\$3/\$15 per 1M Tokens

Google Gemini 2.5 Pro

^ More Details

Robust multimodal capabilities (Source: Google DeepMind). GPQA Diamond: 86.4%, MMMLU: 89.2%. Strong performance in creative and analytical tasks. Pricing: not fully publicly available

Claude Sonnet 4

^ More Details

Balanced performance for production use (Source: Anthropic). Excellent instruction following. Release: 2024. Pricing: ~\$3/\$15 per 1M Tokens

2. Advanced Reasoning & Mathematics

Google Gemini 3 Pro

^ More Details

GPQA Diamond: 91.9% (top scorer, Source: Google DeepMind). Humanity's Last Exam: 37.5%. ARC-

OpenAI o3

^ More Details

Specialized reasoning model (Source: OpenAI). AIME 2025: 98.4% (Source: O

Edit with Lovable

AGI-2: 31.1% (45.1% with Deep Think). Leading in scientific reasoning. Context window: up to 1M Tokens. Pricing: not fully publicly available
GPQA Diamond: 91.9% (top scorer, Source: Google DeepMind). Humanity's Last Exam: 37.5%. ARC-AGI-2: 31.1% (45.1% with Deep Think). Leading in scientific reasoning. Context window: up to 1M Tokens. Pricing: not fully publicly available

analytical thinking. Pricing: not fully publicly available
Specialized reasoning model (Release: Dec 2024). AIME 2025: 98.4% (Source: OpenAI). Deep analytical thinking. Pricing: not fully publicly available

^ More Models (3)

Moonshot AI Kimi K2 Thinking

^ More Details

AIME 2025: 99.1%, Humanity's Last Exam: 44.9% (Source: Moonshot AI). Strong reasoning performance. Pricing: not fully publicly available
AIME 2025: 99.1%, Humanity's Last Exam: 44.9% (Source: Moonshot AI). Strong reasoning performance. Pricing: not fully publicly available

xAI Grok 4

^ More Details

GPQA Diamond: 87.5% (Source: xAI), specialized in scientific research and complex reasoning. Release: Nov 2025. Pricing: not fully publicly available
GPQA Diamond: 87.5% (Source: xAI), specialized in scientific research and complex reasoning. Release: Nov 2025. Pricing: not fully publicly available

Google Gemini 2.5 Deep Think (Preview)

^ More Details

Status: Experimental/Preview. Specialized in deep analytical thinking with extended reasoning capabilities. Benchmark data: not fully publicly available
Status: Experimental/Preview. Specialized in deep analytical thinking with extended reasoning capabilities. Benchmark data: not fully publicly available

Moonshot AI Kimi K2 Thinking

^ More Details

AIME 2025: 99.1%, Humanity's Last Exam: 44.9% (Source: Moonshot AI). Strong reasoning performance. Pricing: not fully publicly available

xAI Grok 4

^ More Details

GPQA Diamond: 87.5% (Source: xAI), specialized in scientific research and complex reasoning. Release: Nov 2025. Pricing: not fully publicly available

Edit with **Lovable**

Google Gemini 2.5 Deep Think (Preview)

More Details

Status: Experimental/Preview. Specialized in deep analytical thinking with extended reasoning capabilities.
Benchmark data: not fully publicly available

3. Software Engineering & Coding

Claude Sonnet 4.5

More Details

SOTA on SWE-bench: 82% (Source: Anthropic).
Leading in Agentic Capabilities (SWE-bench Verified 61.4%) and Computer Use (OSWorld 61.4%). Release: Oct 2024. Pricing: \$3/\$15 per 1M Tokens (Input/Output)

SOTA on SWE-bench: 82% (Source: Anthropic).
Leading in Agentic Capabilities (SWE-bench Verified 61.4%) and Computer Use (OSWorld 61.4%). Release: Oct 2024. Pricing: \$3/\$15 per 1M Tokens (Input/Output)

OpenAI GPT-5.1

More Details

Robust coding model (Release: Nov 12, 2025).
SWE Bench: 76.3% (Source: OpenAI). Optimized for repository context. Pricing: not fully publicly available

Robust coding model (Release: Nov 12, 2025).
SWE Bench: 76.3% (Source: OpenAI). Optimized for repository context. Pricing: not fully publicly available

More Models (4)

Google Gemini 3 Pro

More Details

SWE Bench: 76.2% (Source: Google DeepMind).
State-of-the-art in multimodal coding tasks (screenshots → code). Release: Nov 18, 2025.
Pricing: not fully publicly available

SWE Bench: 76.2% (Source: Google DeepMind).
State-of-the-art in multimodal coding tasks (screenshots → code). Release: Nov 18, 2025.
Pricing: not fully publicly available

xAI Grok 4

More Details

SWE Bench: 75% (Source: xAI). Strong coding performance with focus on efficiency. Release: Nov 2025. Pricing: not fully publicly available

SWE Bench: 75% (Source: xAI). Strong coding performance with focus on efficiency. Release: Nov 2025. Pricing: not fully publicly available

Alibaba Qwen3-Coder

More Details

Open-source option with strong coding performance (Source: Alibaba). Self-hostable without vendor lock-in. Release: 2025. Free (Self-Hosted)

Open-source option with strong coding performance (Source: Alibaba). Self-hostable without vendor lock-in. Release: 2025. Free (Self-Hosted)

DeepSeek R1

More Details

Robust open-source alternative with transparent reasoning (Source: DeepSeek). Release: Jan 2025. Free (Self-Hosted) or \$0.14/\$0.28 per 1M Tokens via API

Robust open-source alternative with transparent reasoning (Source: DeepSeek). Release: Jan 2025. Free (Self-Hosted) or \$0.14/\$0.28 per 1M Tokens via API

Google Gemini 3 Pro

More Details

SWE Bench: 76.2% (Source: Google DeepMind). State-of-the-art in multimodal coding tasks (screenshots → code). Release: Nov 18, 2025. Pricing: not fully publicly available

xAI Grok 4

More Details

SWE Bench: 75% (Source: xAI). Strong coding performance with focus on efficiency. Release: Nov 2025. Pricing: not fully publicly available

Alibaba Qwen3-Coder

More Details

Open-source option with strong coding performance (Source: Alibaba). Self-hostable without vendor lock-in. Release: 2025. Free (Self-Hosted)

DeepSeek R1

More Details

Robust open-source alternative with transparent reasoning (Source: DeepSeek). Release: Jan 2025. Free (Self-Hosted) or \$0.14/\$0.28 per 1M Tokens via API

4. Agentic AI & Computer Use

Claude Sonnet 4.5

More Details

Absolute SOTA leader in Computer Use: OSWorld 61.4% (Source: Anthropic). Best choice for multi-step autonomous agent workflows. Release: Oct 2024. Pricing: \$3/\$15 per 1M Tokens

Absolute SOTA leader in Computer Use: OSWorld 61.4% (Source: Anthropic). Best choice for multi-step autonomous agent workflows. Release: Oct 2024. Pricing: \$3/\$15 per 1M Tokens

OpenAI GPT-5.1

More Details

Agentic coding with real-time interaction (Release: Nov 12, 2025, Source: OpenAI). Optimized for autonomous workflows. Pricing: not fully publicly available

Agentic coding with real-time interaction (Release: Nov 12, 2025, Source: OpenAI). Optimized for autonomous workflows. Pricing: not fully publicly available

More Models (1)

Alibaba Qwen3-Max

More Details

1T+ parameters with Production-Ready Thinking Mode (Source: Alibaba). Strong Agentic Bench

Edit with Lovable

performance. Release: 2025. Pricing: not fully publicly available
1T+ parameters with Production-Ready Thinking Mode (Source: Alibaba). Strong Agentic Bench performance. Release: 2025. Pricing: not fully publicly available

Alibaba Qwen3-Max

∨ More Details

1T+ parameters with Production-Ready Thinking Mode (Source: Alibaba). Strong Agentic Bench performance. Release: 2025. Pricing: not fully publicly available

5. Long Context & Document Analysis

Google Gemini 3 Pro

^ More Details

Up to 1M Tokens context window (Source: Google DeepMind). Leading in document understanding. Release: Nov 18, 2025. Pricing: not fully publicly available

Up to 1M Tokens context window (Source: Google DeepMind). Leading in document understanding. Release: Nov 18, 2025. Pricing: not fully publicly available

DeepSeek V3.2-Exp

^ More Details

164K context window with DeepSeek Sparse Attention (DSA, Source: DeepSeek). Release: Sept 2025. Pricing: 50%+ cheaper through Sparse Attention (~\$0.14/\$0.28 per 1M Tokens)

164K context window with DeepSeek Sparse Attention (DSA, Source: DeepSeek). Release: Sept 2025. Pricing: 50%+ cheaper through Sparse Attention (~\$0.14/\$0.28 per 1M Tokens)

∨ More Models (1)

OpenAI GPT-5.1

∨ More Details

Extended context window with optimized processing (Source: OpenAI). Release: Nov 12, 2025. Pricing: not fully publicly available

6. Ultra-Budget & High-Throughput

DeepSeek V3.2-Exp

^ More Details

50%+ API price reduction through Sparse Attention (Source: DeepSeek). Release: Sept 29, 2025. Pricing: ~\$0.14/\$0.28 per 1M Tokens (Input/Output) - New market standard for cost efficiency

50%+ API price reduction through Sparse Attention (Source: DeepSeek). Release: Sept 29, 2025. Pricing: ~\$0.14/\$0.28 per 1M Tokens

Claude 3.5 Haiku

^ More Details

Fastest Claude model (Source: Anthropic). Release: Nov 2024. Pricing: \$0.80/\$4 per 1M Tokens - optimal for high user volumes

Fastest Claude model (Source: Anthropic). Release: Nov 2024. Pricing: \$0.80/\$4 per 1M Tokens - optimal for high user volumes

Edit with Lovable

(Input/Output) - New market standard for cost efficiency

^ More Models (2)

Meta Llama 4 Scout

^ More Details

Fastest model: 2600 Tokens/second (Source: Meta). Release: 2025. Open-weight - free with self-hosting

Fastest model: 2600 Tokens/second (Source: Meta). Release: 2025. Open-weight - free with self-hosting

Alibaba Qwen3-Flash

^ More Details

Cost-effective open-weight option (Source: Alibaba). Release: 2025. Free (Self-Hosted) or low API pricing

Cost-effective open-weight option (Source: Alibaba). Release: 2025. Free (Self-Hosted) or low API pricing

Meta Llama 4 Scout

^ More Details

Fastest model: 2600 Tokens/second (Source: Meta). Release: 2025. Open-weight - free with self-hosting

Alibaba Qwen3-Flash

^ More Details

Cost-effective open-weight option (Source: Alibaba). Release: 2025. Free (Self-Hosted) or low API pricing

7. Multimodal Understanding (Vision + Audio)

Google Gemini 3 Pro

^ More Details

Leading in Visual Reasoning: ARC-AGI 2: 31% (45.1% with Deep Think, Source: Google DeepMind). MMMU-Pro: ~81%, Video-MMMU: ~87.6%. Release: Nov 18, 2025. Pricing: not fully publicly available

Leading in Visual Reasoning: ARC-AGI 2: 31% (45.1% with Deep Think, Source: Google DeepMind). MMMU-Pro: ~81%, Video-MMMU: ~87.6%. Release: Nov 18, 2025. Pricing: not fully publicly available

OpenAI GPT-5.1

^ More Details

ARC-AGI 2: 18% (Source: OpenAI). Strong vision capabilities. Release: Nov 12, 2025. Pricing: not fully publicly available

ARC-AGI 2: 18% (Source: OpenAI). Strong vision capabilities. Release: Nov 12, 2025. Pricing: not fully publicly available

^ More Models (1)

Google Gemini 2.5 Deep Think (Preview)

Edit with Lovable

^ More Details

Status: Experimental/Preview. Multimodal capabilities with deep analysis (Source: Google DeepMind). Benchmark data: not fully publicly available

Status: Experimental/Preview. Multimodal capabilities with deep analysis (Source: Google DeepMind). Benchmark data: not fully publicly available

Google Gemini 2.5 Deep Think (Preview)

^ More Details

Status: Experimental/Preview. Multimodal capabilities with deep analysis (Source: Google DeepMind). Benchmark data: not fully publicly available

8. Image Generation & Editing

Google Nano Banana Pro

^ More Details

New flagship for image generation (Release: Nov 20, 2025, Source: Google DeepMind). Focus on AI Image Editor. Can merge multiple images and perform complex edits via natural language. Pricing: not fully publicly available

New flagship for image generation (Release: Nov 20, 2025, Source: Google DeepMind). Focus on AI Image Editor. Can merge multiple images and perform complex edits via natural language. Pricing: not fully publicly available

Midjourney V7

^ More Details

Leading in artistic and aesthetic images (Source: Midjourney). Release: 2024. Pricing: Subscription-based from \$10/month

Leading in artistic and aesthetic images (Source: Midjourney). Release: 2024. Pricing: Subscription-based from \$10/month

^ More Models (2)

Ideogram v2.0

^ More Details

Leading in text-in-images (Source: Ideogram). Release: 2024. Pricing: Free with limits, Pro from \$8/month

Leading in text-in-images (Source: Ideogram). Release: 2024. Pricing: Free with limits, Pro from \$8/month

Recraft V3

^ More Details

Optimized for precise style control (Source: Recraft). Release: 2024. Pricing: Free with limits, Pro plans available

Optimized for precise style control (Source: Recraft). Release: 2024. Pricing: Free with limits, Pro plans available

Ideogram v2.0

^ More Details

Edit with Lovable

Leading in text-in-images (Source: Ideogram). Release: 2024. Pricing: Free with limits, Pro from \$8/month

Recraft V3

More Details

Optimized for precise style control (Source: Recraft). Release: 2024. Pricing: Free with limits, Pro plans available

9. Video & Audio Generation

Google Veo 3.1

More Details

Video generation with 'Ingredients to Video' feature (Source: Google DeepMind). Release: Nov 2025. Pricing: not fully publicly available

Video generation with 'Ingredients to Video' feature (Source: Google DeepMind). Release: Nov 2025. Pricing: not fully publicly available

OpenAI Sora

More Details

State-of-the-art video generation with realistic scenes (Source: OpenAI). Release: Dec 2024. Pricing: Subscription-based, part of ChatGPT Plus/Pro

State-of-the-art video generation with realistic scenes (Source: OpenAI). Release: Dec 2024. Pricing: Subscription-based, part of ChatGPT Plus/Pro

10. Open-Source & Self-Hosted

DeepSeek R1

More Details

Powerful open-source alternative with transparent reasoning (Source: DeepSeek). Release: Jan 2025. Free (Self-Hosted) or \$0.14/\$0.28 per 1M Tokens via API

Powerful open-source alternative with transparent reasoning (Source: DeepSeek). Release: Jan 2025. Free (Self-Hosted) or \$0.14/\$0.28 per 1M Tokens via API

Alibaba Qwen3-Max

More Details

1T+ parameters, self-hostable without vendor lock-in (Source: Alibaba). Production-Ready Thinking Mode. Release: 2025. Free (Self-Hosted)

1T+ parameters, self-hostable without vendor lock-in (Source: Alibaba). Production-Ready Thinking Mode. Release: 2025. Free (Self-Hosted)

More Models (2)

Meta Llama 4 Scout

More Details

Open-weight model with 2600 Tokens/second (Source: Meta). Release: 2025. Free under open-weight license

Alibaba Qwen3-Coder

More Details

Robust open-source option specialized in software development (Source: Alibaba). Release: 2025. Free (Self-Hosted)

Edit with Lovable

Open-weight model with 2600 Tokens/second (Source: Meta). Release: 2025. Free under open-weight license

Robust open-source option specialized in software development (Source: Alibaba). Release: 2025. Free (Self-Hosted)

Meta Llama 4 Scout

✓ More Details

Open-weight model with 2600 Tokens/second (Source: Meta). Release: 2025. Free under open-weight license

Alibaba Qwen3-Coder

✓ More Details

Robust open-source option specialized in software development (Source: Alibaba). Release: 2025. Free (Self-Hosted)

Important Trends (As of November 2025)

Current developments in the AI landscape

With Gemini 3 Pro and o3, "Reasoning" (extended thinking) is no longer a niche, but standard for complex tasks.

Models like Gemini 3 Pro process video and audio natively ("token in, token out") instead of through workarounds, massively reducing latency.

The shift from "chatbot" to "agent" (operating software) is complete. Models are now measured by their ability to perform real work steps in operating systems (see OSWorld Benchmark).



Our Forward-Looking Training Programs [↗](#)

(<https://lom.link/ai-offer>)

Systematic AI competency development for various roles and experience levels

Edit with **Lovable**

Management Matthias Herbert

Consultant & AI Strategy Expert

✉ matthias.herbert@obviousworks.ch