| **Model** | **Purpose** | **Strengths** | **Weaknesses** | **Best For** |
| --- | --- | --- | --- | --- |
| **Claude 2** | Conversational AI | Great for casual chats, simple automation, and creativity | Limited in handling complex tasks | Everyday users for scheduling, writing, brainstorming |
| **Claude 2.1** | Enhanced reasoning & conversations | Better at following instructions and long-form discussions | Not optimized for large datasets | Users managing reports, projects, detailed queries |
| **Claude 3 Haiku** | Creative & concise writing | Great for poetry, storytelling, expressive writing | Weak in logic-heavy tasks | Writers, marketers, creatives |
| **Claude 3 Opus** | Structured & professional writing | Produces formal reports, long documents | Not ideal for quick chats | Professionals drafting reports, articles |
| **Claude 3 Sonnet** | Poetic & formal writing | Great for structured storytelling and speeches | Limited outside of creative writing | Poets, speechwriters, authors |
| **Claude 3.5 Sonnet** | Advanced structured writing | Higher-quality, polished output | Slower with large data | Users needing high-quality language generation |
| **Claude 3.7 Sonnet** | Fluency & reasoning boost | Handles complex, longer content well | Not ideal for highly technical tasks | Writers, marketers, professionals |
| **Claude Instant** | Fast, lightweight AI | Quick responses for conversations & simple tasks | Struggles with complex instructions | Users needing real-time chat & quick writing |
| **Gemini Flash 1.5** | Productivity & analytics | Handles business tasks, summaries, decision-making | Lacks deep analytical capabilities | Business users for data handling & reports |
| **Gemini Pro 1.5** | Long-form & data processing | Excellent for reports, technical writing | Slower with simple tasks | Business & academic users needing detailed reports |
| **Gemini 2 Flash** | Fast task completion | Quick summaries, answering questions | Not suited for deep analysis | Professionals needing fast AI assistance |
| **Gemini 2 Flash Lite** | Instant responses | Ultra-fast outputs | Limited complexity & depth | Chatbots, mobile assistants, time-sensitive tasks |
| **Gemini 2 Flash Pro** | Business intelligence | Balances speed with deeper task understanding | Slower than Lite for quick tasks | Teams needing decision-making support |
| **GPT-3.5 (16K)** | Long conversation memory | Retains context in extended chats | Less advanced than GPT-4 | Customer support, detailed writing |
| **GPT-4** | Advanced AI for various tasks | Great for writing, coding, data analysis | Slower than lighter models | Professionals, creatives, developers |
| **GPT-4 (32K)** | Long document processing | Handles complex, in-depth content | More resource-intensive | Legal, research, long-form reports |
| **GPT-4 (Turbo)** | Fast, high-quality AI | Combines GPT-4 intelligence with speed | Not ideal for massive data tasks | Real-time support, customer service |
| **GPT-4o** | Everyday AI tasks | Balanced performance & efficiency | Weaker in highly complex tasks | Daily tasks like emails, summaries |
| **GPT-4o Mini** | Quick, simple AI | Fast and responsive | Not for detailed work | Answering questions, short content |
| **Llama3 40 5b** | Data-driven AI | Great for research & analytics | Requires technical knowledge | Data scientists, researchers |
| **Llama3 70b** | Heavy data processing | Unmatched power for AI tasks | Too complex for everyday use | Machine learning & AI experts |
| **Mistral 8 7b** | Document & data tasks | Good balance of speed and accuracy | Not for large-scale analytics | Business users for data processing |
| **Mistral Large** | Enterprise AI | Handles big automation & analytics | Requires more resources | Advanced business intelligence |
| **Nova Pro** | Advanced AI for precision tasks | High accuracy & computational power | More expensive & resource-heavy | Enterprises, AI researchers |
| **Nova Lite** | Mid-level AI | Fast, cost-effective AI for medium complexity | Less capable for deep analysis | Mid-sized businesses, recommendation systems |
| **Nova Micro** | Lightweight AI | Fast, low-resource tasks | Limited complexity & depth | Startups, chatbots, simple automation |
| **O1** | General-purpose AI | Handles various tasks well | Not great for specialized technical work | Users needing an all-around assistant |
| **O1 Mini** | Quick AI for simple tasks | Fast responses | Limited depth | Email drafting, casual writing |
| **O3 Mini** | Conversational AI | Smooth & reliable for daily tasks | Not for deep analysis | Writing help, summaries, chat support |

Short Summary

### **Reasoning-Focused Models**

* **Claude 3 Haiku, Claude 3 Opus, Claude 3 Sonnet, Claude 3.5 Sonnet, GPT-4**These models are generally well-suited for tasks involving natural language understanding, critical reasoning, and complex context analysis. They’re adept at generating coherent, contextually relevant text, often used in reasoning-heavy tasks like summarization, dialogue, and customer support.

### **2. Math-Focused Models**

* **GPT-4 (32K), Gemini Pro, Llama3 70b**These models perform well in mathematical reasoning and complex calculations due to their larger token limit and advanced training on mathematical datasets. They're suitable for applications requiring high accuracy in mathematical problem-solving, such as finance or quantitative analysis.

### **3. Coding-Focused Models**

* **GPT-4 (Turbo), Claude Instant, Command, Command Light**These models are optimized for tasks that involve coding and software development, with strengths in understanding programming languages, debugging, and generating code snippets. They’re used for code generation, code completion, and other software development tasks.

### **4. General Purpose / High Similarity Across Domains**

* **GPT-3.5 (16K), GPT-4o, GPT-4o Mini, Titan Express (8K), Titan Lite (4K)**These models are versatile and can perform reasonably well across reasoning, math, and coding. They provide a balanced performance across tasks, making them suitable for general-purpose applications where a broad range of abilities is needed.

### **Creative Writing**

* **Claude 3 Opus, GPT-4, GPT-4 (32K)**These models are strong in generating expressive and engaging content, ideal for tasks requiring a creative touch like storytelling, article writing, or creating persuasive text.

### **2. Mathematics**

* **GPT-4 (32K), Gemini Pro, Llama3 70b**These models excel at handling complex math tasks, offering accuracy for scientific, quantitative, and problem-solving applications.

### **3. Summarization**

* **Claude 3 Haiku, GPT-4 Turbo, Titan Express (8K)**These models are efficient and precise in summarizing long texts, making them suitable for creating concise overviews of reports, articles, or documents.

### **4. Cost-Efficiency**

* **GPT-3.5 (16K), Titan Lite (4K), Llama3 40 5b,** Mistral 3B These models offer solid performance at a lower cost, making them ideal for high-usage applications that require a balance of quality and affordability.

### **5. Quality of Response**

* **Claude 3.5 Sonnet, GPT-4 (32K), Gemini Pro 1.5**These models deliver highly accurate, contextually aware responses, suitable for tasks where quality and reliability of response are critical, such as customer service and knowledge-based queries.

### **6. Reasoning**

* **Claude 3 Sonnet, GPT-4, Llama3 70b**Known for strong logical and contextual reasoning, these models are well-suited for complex problem-solving and understanding nuanced, layered queries.

### **7. Instruction Following**

* **GPT-4 Turbo, Claude Instant, Command Light**These models excel at following instructions precisely, making them ideal for technical support, step-by-step processes, and other guided procedural tasks.

### **8. Mathematical Ability**

* **GPT-4 (32K), Claude 3 Haiku, Llama3 70b**With high mathematical accuracy, these models are effective for academic or technical tasks involving rigorous math, from calculus to algebra.

### **9. Coding**

* **GPT-4 Turbo, Claude Instant, Command, Command Light**These models are optimized for coding-related tasks, including code generation, debugging, and providing step-by-step code instructions, making them highly suitable for software development, code assistance, and programming education.

Deep Reasoning : O1 Models

Multilingual: Gemini and Llama and GPT models are multilingual

Inspo:



