AI Providers Models Research - Post June 2024

Research compiled on July 24, 2025

This document provides comprehensive information about AI models from major providers that were released after June 2024, including their specifications, capabilities, pricing, and implementation details.

Table of Contents

Google Gemini

Google has released several advanced models in their Gemini 2.5 and 2.0 series, along with specialized models for image and video generation.

Gemini 2.5 Pro

Release: After June 2024
Model Code: `gemini-2.5-pro`

Specifications:
- **Context Length**: 1,048,576 tokens (input) / 65,536 tokens (output)
- **Modalities**: Audio, images, videos, text, PDF → Text output
- **Optimized For**: Enhanced thinking, reasoning, multimodal understanding, advanced coding

Key Capabilities:
- Structured outputs
- Context caching
- Function calling
- Code execution
- Search grounding
- Advanced “thinking” model for complex reasoning

Pricing (per 1M tokens):
- Input: $1.25 (≤200k tokens) / $2.50 (>200k tokens)
- Output (including thinking tokens): $10.00 (≤200k) / $15.00 (>200k)
- Context caching: $0.31/$0.625 + $4.50/1M tokens/hour storage
- Grounding: 1,500 RPD free, then $35/1,000 requests

Use Cases: Complex reasoning tasks, advanced coding, long-context analysis, multimodal understanding

Gemini 2.5 Flash

Release: After June 2024
Model Code: `gemini-2.5-flash`

Specifications:
- **Context Length**: 1,048,576 tokens (input) / 65,536 tokens (output)
- **Modalities**: Audio, images, videos, text → Text output
- **Optimized For**: Price-performance balance, adaptive thinking

Key Capabilities:
- Function calling
- Code execution
- Search grounding
- Adaptive thinking (enabled by default, can be disabled)
- Live API support

Pricing (per 1M tokens):
- Input: $0.30 (text/image/video) / $1.00 (audio)
- Output: $2.50
- Context caching: $0.075/$0.25 + $1.00/1M tokens/hour storage
- Live API: Input $0.50 (text) / $3.00 (audio/image/video); Output $2.00 (text) / $12.00 (audio)

Use Cases: High-throughput tasks, real-time applications, cost-effective multimodal processing

Gemini 2.5 Flash-Lite

Release: After June 2024
Model Code: `gemini-2.5-flash-lite`

Specifications:
- **Context Length**: 1,048,576 tokens (input) / 65,536 tokens (output)
- **Modalities**: Text, image, video, audio → Text output
- **Optimized For**: Cost efficiency, high throughput

Key Capabilities:
- Structured outputs
- Context caching
- Function calling
- Most cost-efficient option

Pricing (per 1M tokens):
- Input: $0.10 (text/image/video) / $0.30 (audio)
- Output: $0.40
- Context caching: $0.025/$0.125 + $1.00/1M tokens/hour storage

Use Cases: High-volume processing, cost-sensitive applications, basic multimodal tasks

Gemini 2.5 Flash Native Audio

Release: After June 2024
Model Codes:
- `gemini-2.5-flash-preview-native-audio-dialog`
- `gemini-2.5-flash-exp-native-audio-thinking-dialog`

Specifications:
- **Context Length**: 128,000 tokens (input) / 8,000 tokens (output)
- **Modalities**: Audio, video, text → Text and audio output
- **Optimized For**: Conversational audio outputs

Key Capabilities:
- High-quality conversational audio
- Internal “thinking” steps (experimental version)
- Native audio processing

Pricing (per 1M tokens):
- Input: $0.50 (text) / $3.00 (audio/video)
- Output: $2.00 (text) / $12.00 (audio)

Use Cases: Voice assistants, conversational AI, audio-first applications

Gemini 2.5 Text-to-Speech Models

Release: After June 2024

Flash Preview TTS

- **Model Code**: `gemini-2.5-flash-preview-tts`
- **Context Length**: 8,000 tokens (input) / 16,000 tokens (output)
- **Pricing**: $0.50 input / $10.00 output (per 1M tokens)
- **Use Case**: Low-latency text-to-speech

Pro Preview TTS

- **Model Code**: `gemini-2.5-pro-preview-tts`
- **Context Length**: 8,000 tokens (input) / 16,000 tokens (output)
- **Pricing**: $1.00 input / $20.00 output (per 1M tokens)
- **Use Case**: High-quality text-to-speech

Imagen 4 (Image Generation)

Release: After June 2024
Model Codes:
- `imagen-4.0-generate-preview-06-06` (Standard)
- `imagen-4.0-ultra-generate-preview-06-06` (Ultra)

Specifications:
- **Context Length**: 480 tokens (input) / 1-4 images (output)
- **Modalities**: Text → Images
- **Optimized For**: Better text rendering and image quality

Pricing (per image):
- Standard: $0.04
- Ultra: $0.06

Use Cases: High-quality image generation, text rendering in images, creative content

Veo 3 Preview (Video Generation)

Release: After June 2024
Model Code: `veo-3.0-generate-preview`

Specifications:
- **Context Length**: 1,024 tokens (input) / 1 video (output)
- **Modalities**: Text → Video with audio
- **Optimized For**: Video generation

Pricing (per second):
- Video with audio: $0.75
- Video without audio: $0.50

Use Cases: Video content creation, multimedia presentations, creative video generation

Implementation Examples

Basic Text Generation

```
import google.generativeai as genai

genai.configure(api_key="YOUR_API_KEY")
model = genai.GenerativeModel('gemini-2.5-flash')

response = model.generate_content("Explain quantum computing")
print(response.text)
```

Multimodal Processing

```
import PIL.Image

model = genai.GenerativeModel('gemini-2.5-pro')
image = PIL.Image.open('image.jpg')

response = model.generate_content([
    "Describe this image in detail",
    image
])
print(response.text)
```

Audio Processing

```
model = genai.GenerativeModel('gemini-2.5-flash')

# Upload audio file
audio_file = genai.upload_file('audio.mp3')

response = model.generate_content([
    "Transcribe and summarize this audio",
    audio_file
])
print(response.text)
```

---

xAI (Grok)

xAI has released Grok 4, their most advanced reasoning model with enhanced capabilities.

Grok 4 (grok-4-0709)

Release: July 2024
Model Code: `grok-4-0709`

Specifications:
- **Context Length**: 256,000 tokens
- **Modalities**: Text with vision (image generation and other capabilities coming soon)
- **Model Type**: Reasoning model (no non-reasoning mode available)

Key Capabilities:
- Function calling
- Structured outputs
- Advanced reasoning
- Vision capabilities
- No support for `presencePenalty`, `frequencyPenalty`, `stop` parameters
- No `reasoning_effort` parameter support

Pricing (per million tokens):
- Input: $3.00
- Output: $15.00

Rate Limits:
- 2 million tokens per minute (2Mtpm)
- 480 requests per minute (480rpm)

Use Cases: Complex reasoning tasks, mathematical problem solving, advanced analysis, vision-based reasoning

Implementation Examples

Basic Reasoning

```
import openai

client = openai.OpenAI(
    api_key="YOUR_XAI_API_KEY",
    base_url="https://api.x.ai/v1"
)

response = client.chat.completions.create(
    model="grok-4-0709",
    messages=[
        {"role": "user", "content": "Solve this complex math problem step by
step: ..."}
    ]
)

print(response.choices[0].message.content)
```

Function Calling

```
tools = [
    {
        "type": "function",
        "function": {
            "name": "calculate",
            "description": "Perform mathematical calculations",
            "parameters": {
                "type": "object",
                "properties": {
                    "expression": {"type": "string"}
                }
            }
        }
    }
]

response = client.chat.completions.create(
    model="grok-4-0709",
    messages=[{"role": "user", "content": "Calculate 15% of 2847"}],
    tools=tools
)
```

Groq

Groq has released numerous models after June 2024, focusing on high-performance inference across various model families.

Speech-to-Text Models

Whisper Large V3

Release: June 24, 2024
Model Code: `whisper-large-v3`

Specifications:
- **Developer**: OpenAI (hosted on Groq)
- **Max File Size**: 100 MB
- **Modalities**: Audio → Text
- **Special Features**: Word-level timestamps support

Rate Limits (Free Developer tier):
- RPM: 20 requests per minute
- RPD: 2000 requests per day
- ASH: 7200 audio seconds per hour
- ASD: 28800 audio seconds per day

Whisper Large V3 Turbo

Release: October 9, 2024
Model Code: `whisper-large-v3-turbo`

Specifications:
- **Developer**: OpenAI (hosted on Groq)
- **Optimized For**: Faster inference than standard Whisper Large V3

Distil-Whisper Large V3 EN

Release: August 20, 2024
Model Code: `distil-whisper-large-v3-en`

Specifications:
- **Developer**: Hugging Face (hosted on Groq)
- **Optimized For**: English-only, faster processing

Language Models

Llama 3.1 Series

Release: July 23, 2024

Llama 3.1 8B Instant
- **Model Code**: `llama-3.1-8b-instant`
- **Optimized For**: Fast inference, lightweight tasks

Llama 3.1 70B Versatile
- **Model Code**: `llama-3.1-70b-versatile`
- **Optimized For**: Balanced performance and capability

Llama 3.1 405B Reasoning (Later moved offline)
- **Model Code**: `llama-3.1-405b-reasoning`
- **Optimized For**: Complex reasoning tasks

Llama 3.2 Series

Release: September 25, 2024

Llama 3.2 1B Preview
- **Model Code**: `llama-3.2-1b-preview`
- **Optimized For**: Ultra-lightweight applications

Llama 3.2 3B Preview
- **Model Code**: `llama-3.2-3b-preview`
- **Optimized For**: Edge deployment, mobile applications

Llama 3.2 90B Text Preview
- **Model Code**: `llama-3.2-90b-text-preview`
- **Optimized For**: High-performance text processing

Llama 3.2 Vision Models

Release: September 27, 2024 & October 9, 2024

Llama 3.2 11B Vision Preview
- **Model Code**: `llama-3.2-11b-vision-preview`
- **Modalities**: Text + Vision
- **Optimized For**: Multimodal understanding

Llama 3.2 90B Vision Preview
- **Model Code**: `llama-3.2-90b-vision-preview`

- **Modalities**: Text + Vision
- **Optimized For**: Advanced multimodal reasoning

Llama 3.3 Series

Release: December 6, 2024

Llama 3.3 70B Versatile
- **Model Code**: `llama-3.3-70b-versatile`
- **Optimized For**: General-purpose tasks

Llama 3.3 70B SpecDec
- **Model Code**: `llama-3.3-70b-specdec`
- **Optimized For**: Speculative decoding for faster inference

Tool Use Models

Release: July 16, 2024

Llama3 Groq 70B Tool Use
- **Model Code**: `Llama3-groq-70b-tool-use`
- **Capabilities**: Function calling, tool integration

Llama3 Groq 8B Tool Use
- **Model Code**: `Llama3-groq-8b-tool-use`
- **Capabilities**: Lightweight tool use

Recent Advanced Models (2025)

DeepSeek R1 Distilled Models

Release: January 26, 2025 & February 3, 2025

DeepSeek R1 Distill Llama 70B
- **Model Code**: `deepseek-r1-distill-llama-70b`
- **Optimized For**: Reasoning capabilities

DeepSeek R1 Distill Llama 70B SpecDec
- **Model Code**: `deepseek-r1-distill-llama-70b-specdec`
- **Optimized For**: Fast reasoning inference

Qwen Series

Release: February 10, 2025 & February 13, 2025

Qwen 2.5 32B
- **Model Code**: `qwen-2.5-32b`
- **Optimized For**: General-purpose tasks

Qwen 2.5 Coder 32B
- **Model Code**: `qwen-2.5-coder-32b`
- **Optimized For**: Code generation and programming

Qwen QwQ 32B
- **Model Code**: `qwen-qwq-32b`
- **Release**: March 5, 2025
- **Optimized For**: Question-answering tasks

Text-to-Speech Models

Release: March 26, 2025

PlayAI TTS
- **Model Code**: `playai-tts`
- **Modalities**: Text → Audio

PlayAI TTS Arabic
- **Model Code**: `playai-tts-arabic`
- **Modalities**: Text → Audio (Arabic language)

Meta Llama 4 Series

Release: April 5, 2025

Llama 4 Maverick 17B
- **Model Code**: `meta-llama/llama-4-maverick-17b-128e-instruct`
- **Optimized For**: Advanced instruction following

Llama 4 Scout 17B
- **Model Code**: `meta-llama/llama-4-scout-17b-16e-instruct`
- **Optimized For**: Exploration and discovery tasks

Implementation Examples

Speech-to-Text

```
from groq import Groq

client = Groq(api_key="YOUR_GROQ_API_KEY")

with open("audio.mp3", "rb") as file:
    transcription = client.audio.transcriptions.create(
        file=file,
        model="whisper-large-v3",
        response_format="json",
        timestamp_granularities=["word"]
    )

print(transcription.text)
```

Text Generation

```
completion = client.chat.completions.create(
    model="llama-3.3-70b-versatile",
    messages=[
        {"role": "user", "content": "Explain machine learning"}
    ],
    temperature=0.7,
    max_tokens=1024
)

print(completion.choices[0].message.content)
```

Vision Model

```
import base64

def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode('utf-8')

base64_image = encode_image("image.jpg")

completion = client.chat.completions.create(
    model="llava-v1.5-7b-4096-preview",
    messages=[
        {
            "role": "user",
            "content": [
                {"type": "text", "text": "Describe this image"},
                {
                    "type": "image_url",
                    "image_url": {"url": f"{base64_image}"}
                }
            ]
        }
    ]
)
```

Anthropic Claude

Anthropic has released several advanced Claude models with enhanced reasoning capabilities and computer use features.

Claude 3.5 Sonnet (October 2024 Upgrade)

Release: October 22, 2024
Model Code: `claude-3-5-sonnet-20241022` (alias: `claude-3-5-sonnet-latest`)

Specifications:
- **Context Length**: 200,000 tokens
- **Max Output**: 8,192 tokens
- **Training Data Cut-off**: April 2024
- **Modalities**: Text, Vision
- **Capabilities**: High intelligence, multilingual, vision, priority tier

Key Features:
- Computer use tool support (beta)
- Advanced reasoning capabilities
- Vision processing
- Desktop automation through screenshots and controls

Pricing (per million tokens):
- Base Input: $3.00
- 5m Cache Writes: $3.75
- 1h Cache Writes: $6.00

- Cache Hits & Refreshes: $0.30
- Output: $15.00

Rate Limits (Tier 1):
- RPM: 50 requests per minute
- ITPM: 30,000 input tokens per minute
- OTPM: 8,000 output tokens per minute

Computer Use Tool:
- **Tool Version**: `computer_20241022`
- **Beta Flag**: `"computer-use-2024-10-22"`
- **System Prompt Overhead**: 466-499 tokens
- **Tool Token Usage**: 683 tokens
- **Actions**: `screenshot`, `left_click`, `type`, `key`, `mouse_move`

Claude Haiku 3.5

Release: October 22, 2024
Model Code: `claude-3-haiku-20241022`

Specifications:
- **Context Length**: 200,000 tokens
- **Max Output**: 8,192 tokens
- **Optimized For**: Speed and efficiency
- **Capabilities**: Fastest model in the Claude family

Claude Sonnet 3.7

Release: February 19, 2025
Model Code: `claude-3-sonnet-20250219`

Specifications:
- **Context Length**: 200,000 tokens
- **Max Output**: 8,192 tokens
- **Key Features**: Extended thinking capabilities
- **Added to claude.ai**: February 24, 2025

Claude Sonnet 4

Release: May 14, 2025
Model Code: `claude-4-sonnet-20250514`

Specifications:
- **Context Length**: 200,000 tokens
- **Max Output**: 8,192 tokens
- **Key Features**: High-performance model with exceptional reasoning and efficiency
- **Added to claude.ai**: May 22, 2025

Claude Opus 4

Release: May 14, 2025
Model Code: `claude-4-opus-20250514`

Specifications:
- **Context Length**: 200,000 tokens
- **Max Output**: 8,192 tokens

- **Key Features**: Anthropic's most capable and intelligent model, setting new standards in complex reasoning and advanced coding
- **Added to claude.ai**: May 22, 2025

Implementation Examples

Basic Text Generation

```python
import anthropic

client = anthropic.Anthropic(api_key="YOUR_ANTHROPIC_API_KEY")

message = client.messages.create(
    model="claude-3-5-sonnet-20241022",
    max_tokens=1024,
    messages=[
        {"role": "user", "content": "Explain quantum computing"}
    ]
)

print(message.content[0].text)
```

Vision Processing

```python
import base64

def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode('utf-8')

base64_image = encode_image("image.jpg")

message = client.messages.create(
    model="claude-3-5-sonnet-20241022",
    max_tokens=1024,
    messages=[
        {
            "role": "user",
            "content": [
                {
                    "type": "image",
                    "source": {
                        "type": "base64",
                        "media_type": "image/jpeg",
                        "data": base64_image
                    }
                },
                {
                    "type": "text",
                    "text": "Describe this image in detail"
                }
            ]
        }
    ]
)
```

Computer Use Tool

```
message = client.messages.create(
    model="claude-3-5-sonnet-20241022",
    max_tokens=1024,
    tools=[
        {
            "type": "computer_20241022",
            "name": "computer",
            "display_width_px": 1024,
            "display_height_px": 768,
            "display_number": 1
        }
    ],
    betas=["computer-use-2024-10-22"],
    messages=[
        {
            "role": "user",
            "content": "Take a screenshot and describe what you see"
        }
    ]
)
```

Function Calling

```
tools = [
    {
        "name": "get_weather",
        "description": "Get weather information for a location",
        "input_schema": {
            "type": "object",
            "properties": {
                "location": {"type": "string", "description": "City name"}
            },
            "required": ["location"]
        }
    }
]

message = client.messages.create(
    model="claude-3-5-sonnet-20241022",
    max_tokens=1024,
    tools=tools,
    messages=[
        {"role": "user", "content": "What's the weather like in San Francisco?"}
    ]
)
```

Summary and Recommendations

Key Trends Post-June 2024

1. **Multimodal Capabilities**: All providers have enhanced their multimodal offerings, with Google leading in audio/video generation, and Claude introducing computer use capabilities.

2. **Reasoning Models**: xAI's Grok 4 and various reasoning-focused models from other providers show the industry's focus on advanced reasoning capabilities.
3. **Efficiency Optimizations**: Multiple model variants optimized for different use cases (speed vs. capability vs. cost).
4. **Specialized Tools**: Computer use (Claude), native audio processing (Gemini), and advanced vision capabilities across all providers.

Best Practices

1. **Choose the Right Model**: Match model capabilities to your specific use case requirements.
2. **Implement Caching**: Use context caching where available to reduce costs for repeated contexts.
3. **Monitor Rate Limits**: Implement proper rate limiting and error handling in your applications.
4. **Security Considerations**: Be cautious with computer use tools and implement proper sandboxing.
5. **Cost Optimization**: Use lite/efficient models for high-volume, simple tasks and reserve premium models for complex reasoning.

Provider Strengths

- **Google Gemini**: Best for multimodal applications, especially audio/video processing
- **xAI Grok**: Excellent for complex reasoning and mathematical problem-solving
- **Groq**: Superior inference speed and variety of model options
- **Anthropic Claude**: Leading in safety, computer use capabilities, and advanced reasoning

This research provides a comprehensive foundation for integrating these advanced AI capabilities into applications and choosing the most appropriate models for specific use cases.