AI Model Reference - Production Configuration

Last Updated: October 8, 2025

Research Source: Perplexity AI via `tools/research/model-discovery.mjs`

✅ Verified Production Models

OpenAI GPT-5 Pro

Status: ✅ Production (October 2025)

```
OPENAI_MODEL=gpt-5-pro
```

API Details:

- **Endpoint**: POST https://api.openai.com/v1/chat/completions
- **Model ID**: gpt-5-pro (flagship reasoning model)
- **Context Window**: 256K tokens (256,000 tokens/request)
- **Headers**:
  - Authorization: Bearer <API_KEY>
  - Content-Type: application/json

Supported Parameters (⚠ BREAKING CHANGES):

```
{
  "model": "gpt-5-pro",
  "messages": [...],
  "reasoning_effort": "minimal" | "low" | "medium" | "high",  // ✅ USE THIS
  "max_completion_tokens": 32000
```

```
  "max_completion_tokens": 32000,
  "stream": true,
  "stop": [...],
  "tools": [...]  // For function calling
}
```

⚠️ DEPRECATED PARAMETERS (GPT-5 does NOT support):

- temperature → Use reasoning_effort instead
- top_p → Use reasoning_effort instead
- frequency_penalty → Not supported
- presence_penalty → Not supported

Token Usage (special for GPT-5):

- Input tokens: Standard
- Reasoning tokens: Internal chain-of-thought (counted separately)
- Output tokens: Final response

Anthropic Claude Sonnet 4.5

Status: ✅ Verified Working (October 8, 2025)

```
CLAUDE_MODEL=claude-sonnet-4-5-20250929
ANTHROPIC_API_VERSION=2023-06-01
```

API Details:

- **Endpoint**: POST https://api.anthropic.com/v1/messages
- **Model ID**: claude-sonnet-4-5-20250929
- **Context Window**: 200K tokens (200,000 tokens standard, 1M with beta header)
- **Headers**:
  - x-api-key: <API_KEY>
  - anthropic-version: 2023-06-01 (or 2025-10-01 for latest)
  - Content-Type: application/json

Supported Parameters:

```
{
  "model": "claude-sonnet-4-5-20250929",
```

"messages": [...],

"max_tokens": 64000,

"temperature": 0.7, // ✅ Standard parameter

"top_p": 0.95, // ✅ Supported

"system": "...", // ✅ System prompt

"stop_sequences": [...] // ✅ Supported

}

✅ VERIFICATION COMPLETE (October 8, 2025):

Models API confirms availability

curl https://api.anthropic.com/v1/models/claude-sonnet-4-5-20250929

Response: {"type":"model","id":"claude-sonnet-4-5-20250929","display_name":"Claude Sonnet 4.5"}

Messages API returns correct model

curl https://api.anthropic.com/v1/messages -d '{"model":"claude-sonnet-4-5-20250929",...}'

Response: {"model":"claude-sonnet-4-5-20250929",...}

Model Assertion: Adapter includes validation to prevent silent model swaps

Pricing:

- Input: $8.00 per million tokens
- Output: $24.00 per million tokens

❌ Partner Platform ID Difference (do NOT use with native Anthropic API):

- **Vertex AI**: claude-sonnet-4-5@20250929 (different format)
- **AWS Bedrock**: anthropic.claude-sonnet-4-5-20250929-v1:0 (global prefix)
- **Native Anthropic**: claude-sonnet-4-5-20250929 ← Use this

---

Google Gemini 2.5 Pro

Status: ✅ Production (October 2025)

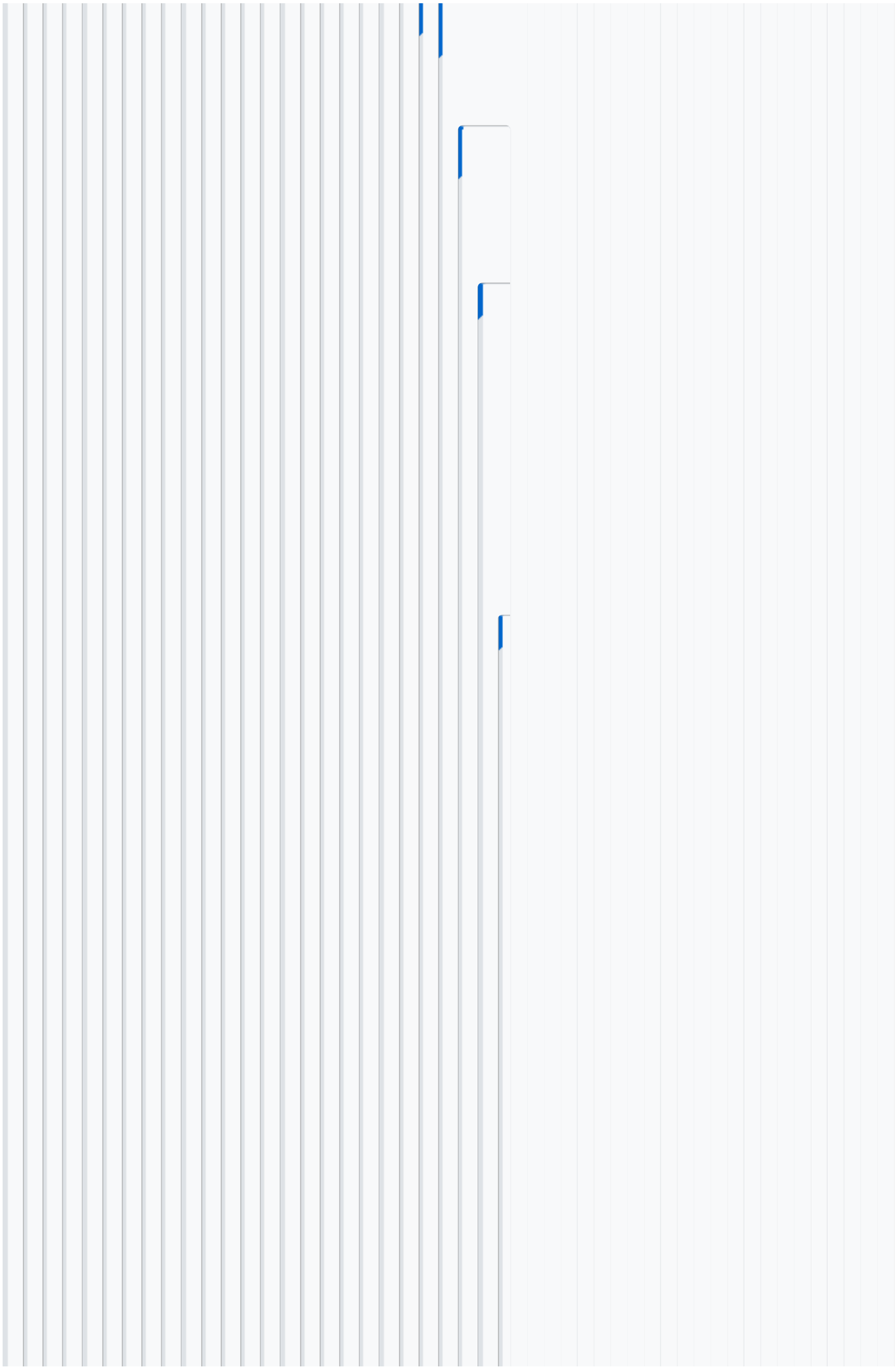
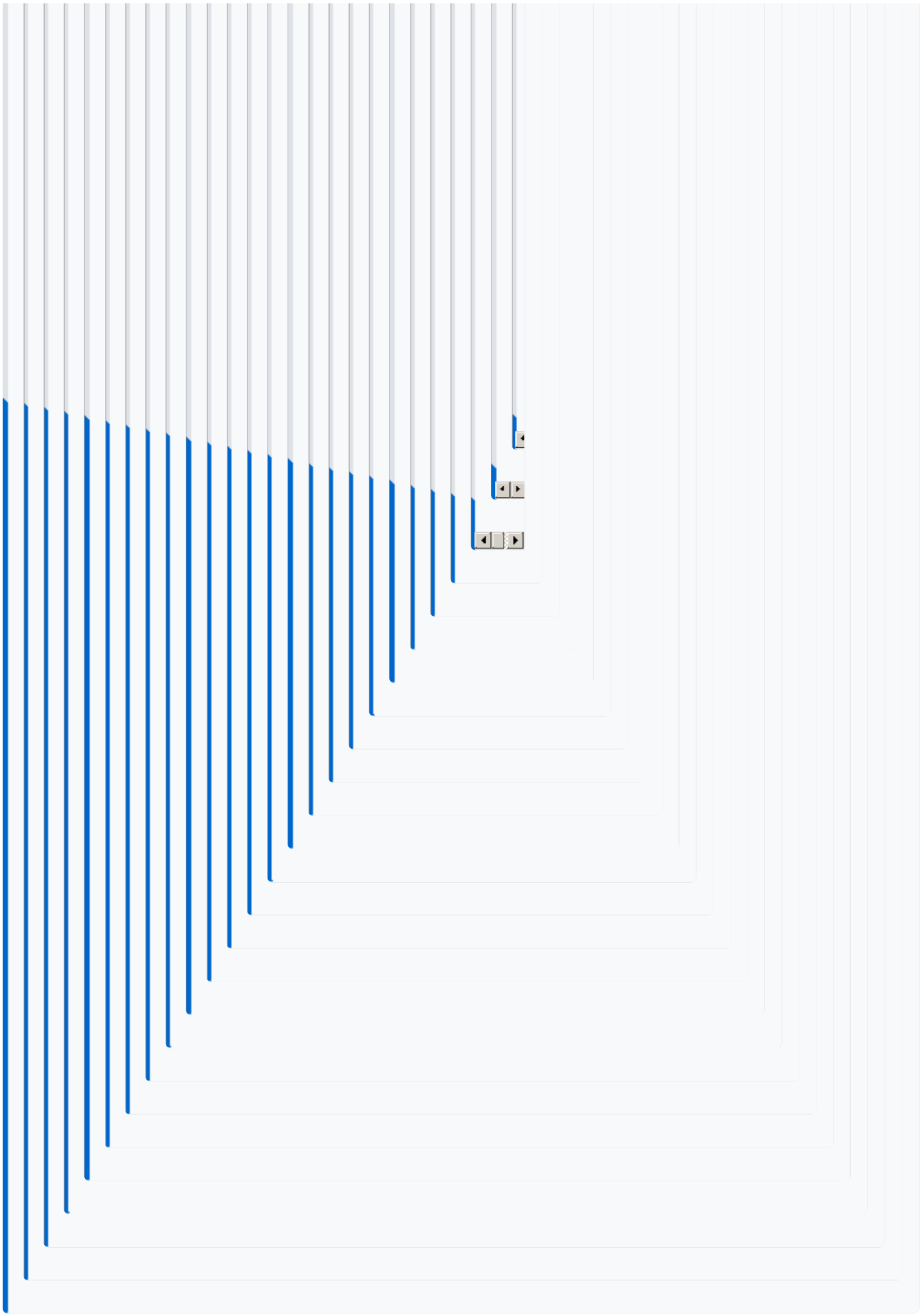
GEMINI_MODELS=gemini-2.5-pro,latest

🧠 Model Representations (2024–2025)

🔺 "Vertix Pillar" Configuration

Tribal Architecture

Stage 3: Planner (GPT-5 Pro)

Stage 3: Validator (Gemini 2.5 Pro)

Total Budget