```
# ChatGPT Models (OpenAI)

## GPT-4o ("omni")

### Strengths
- **GPT-4 level intelligence with optimized speed and cost**: Flagship model for 2024-2025. As performant or superior to GPT-4 Turbo, with latency comparable to GPT-3.5 Turbo, and ~50% less expensive than GPT-4 Turbo via API.
- **Natively multimodal (text, image, audio, video)**: Designed from the ground up to process different input and output formats, with advanced voice capabilities.
- **Significant reduction in multimodal latency**: Responds to voice inputs in ~300 ms.
- **Improvements in vision and audio**: Better understanding of visual scenes, intonation, and multi-voice speech generation.
- **Better multilingual support**: Notable improvements in European and Asian languages.
- **Context window: 128k tokens.**
- **Accessible for free via ChatGPT (with limits) and via the API.**

### Weaknesses
- **Still more expensive than GPT-3.5 Turbo for simple tasks.**
- **Some advanced multimodal features are still being rolled out.**
- **Best practices for use are still being explored.**

### Recommended Use Cases
- Intelligent voice assistants.
- Joint text/image/audio analysis.
- Fast and contextual multilingual translation.
- Multimodal collaborative tools.
- Applications with visual or sound understanding.

---

## GPT-4 Turbo (and variants)

### Strengths
- **Very good at complex reasoning and code generation.**
- **Context window: 128k tokens.**
- **Vision available (GPT-4 Turbo with Vision, e.g., `gpt-4-vision-preview`).**
- **Up-to-date knowledge (April or December 2023 depending on the version).**
```

```
### Weaknesses
- **More expensive and slower than GPT-4o.**
- **Not natively multimodal (audio/visual capabilities are less integrated).**
- **Generally surpassed by GPT-4o.**

### Recommended Use Cases
- Advanced code development.
- High-quality technical writing or long-form content.
- Analysis of large documents.
- Personalized tutorials on complex subjects.

---

## GPT-3.5 Turbo

### Strengths
- **Excellent speed/cost ratio.**
- **Very good for classic tasks: summarization, classification, simple
writing, chatbot.**
- **Context: 4K or 16K tokens available.**
- **Stable and widely documented model.**

### Weaknesses
- **Less performant for complex tasks.**
- **No vision or audio.**
- **Knowledge often limited to 2021 (depending on version).**

### Recommended Use Cases
- Customer chatbots with frequent requests.
- Summaries, product descriptions, emails.
- Latency or cost-sensitive applications.
- Rapid prototyping.

---

## "Mini" or Specialized Models

### GPT-3.5 Turbo ("light" version)
- **Designed for simple tasks, with reduced cost and latency.**
- **Less performant than GPT-4 and GPT-4o models.**
```

```
### Embedding Models (`text-embedding-3-small`, `text-embedding-ada-002`)
- **`text-embedding-3-small`**:
  - Very good quality/price ratio.
  - Ideal for semantic search, recommendations, clustering.
- **`text-embedding-ada-002`**:
  - Very economical, still used in many systems.

### Fine-tuned Models (on GPT-3.5 Turbo)
- **Optimized for specific tasks.**
- **Can surpass GPT-4 in a well-trained narrow domain.**
- **Useful for strict styles, tones, or formats.**

---

## Older Models (`davinci`, `curie`, `babbage`, `ada`)

### Strengths (historical)
- Excellent performance at their release (notably `text-davinci-003`).

### Weaknesses
- Technically and economically outdated models.
- Limited context windows.
- Not recommended for current projects.

### Use Cases
- Maintenance of existing systems.
- Rare cases of fine-tuning on `davinci`.

---

> **Last updated: May 2025**
> For official and up-to-date information:
[https://platform.openai.com/docs/models](https://platform.openai.com/docs/models)
```