How to Choose the Right Model

Amplify AI provides access to seven AI models. Each model is designed for different types of work. The best model for your task depends on the complexity of the task, the type and amount of information you are using, and the level of reasoning you need. Review the information below to see what each model does best and examples of when to use it.

ICON GUIDE

- **Resource-efficient model** — All models currently available to Mercer users are in this tier.
- **Higher-cost model** — Less resource-efficient. Use only when needed.
- Generates simple images
- Can analyze images (upload and discuss)

MODEL	COST	CAPABILITIES	WHAT IT DOES (IN PLAIN LANGUAGE)	WHAT IT'S BEST FOR	EXAMPLE USE CASES
Claude Haiku 5.5	$		The fastest and most cost-efficient model in the Claude 5.5 family.	Everyday tasks that benefit from speed, clear analysis, reliable instruction-following, and solid coding help.	Review and revise documents, compare information, summarize reports, analyze images, draft or troubleshoot code, brainstorm ideas, and answer moderately complex questions efficiently.
Gemini 2.0 Flash	$		A fast, multimodal model built for large-context information processing. Works with text, images, video, and audio and can handle very large amounts of source material.	Tasks with large amounts of information, long documents, multiple media types, or rapid summarization.	Analyze long documents or multiple files, summarize extensive material, extract information from graphics, work with audio or video, compare large tables, and process content quickly.
Gemini 3.0 Flash	$		A newer Flash model that combines fast responses with stronger reasoning, coding, tool use, and multimodal understanding across text, images, video, audio, and PDFs.	Fast multimodal analysis, coding assistance, agent-style tasks, and complex work where speed still matters.	Analyze PDFs and screenshots, reason across mixed media, troubleshoot code, compare sources, create structured outputs, and experiment with advanced Gemini workflows.
GPT-6.0 Luna (Default)	$		The fastest, most cost-effective default model in the GPT family. A versatile, general-purpose model optimized for cost-sensitive, high-volume work. It combines fast responses, adjustable reasoning, and a very large context window.	Writing, analysis, planning, large-document work, routine problem solving, and general productivity. A strong default choice for most tasks.	Draft or revise documents, summarize and analyze information, brainstorm ideas, work with long files, create structured content, study, plan projects, and get coding help.
GPT-4.1 Mini	$		A smaller, faster GPT-4.1 model with strong instruction-following, tool calling, coding, and long-context comprehension. It responds without an extended reasoning step.	Long documents, coding, structured instructions, tool use, and tasks that need precision with low latency.	Summarize or compare long documents, follow detailed formatting rules, revise code, extract structured data, analyze images, and process large prompts efficiently.
GPT-4o Mini	$		A fast, capable model for everyday tasks. Provides solid responses with good language understanding and multimodal reasoning at very low cost.	Quick drafts, summaries, image-aware tasks, and straightforward analysis.	Write, summarize, or edit text; create outlines; simplify complex information; analyze images; work with simple tables or spreadsheet data; and answer questions on a variety of topics.
GPT-5 Nano	$		OpenAI's smallest, fastest, and most cost-efficient GPT-5 model. Best for simple, well-defined tasks that do not require deep reasoning.	Quick, routine tasks where speed, efficiency, and high-volume processing are most important.	Summarize information, classify or organize content, extract key points, reformat text, generate simple lists, handle repetitive work, and process straightforward requests quickly.
GPT-5 Mini	$		A more capable small model with stronger reasoning and a larger context window than Nano. Good for tasks that need more precision, structure, or detailed outputs.	Structured work, document and data analysis, coding assistance, content creation, and tasks requiring more reasoning than Nano.	Create detailed documents, rubrics or templates; analyze spreadsheets and data; develop or troubleshoot formulas; write or revise code; analyze documents and images; and follow complex instructions.