

Edit catalog model

edit-catalog-model.html · 760×744


1 **claude-opus-4-8**

EDITING — edit this catalog entry

✕

2 **Provider**

Anthropic

▼

3 **Model ID**

claude-opus-4-8

4 **Display name**

Claude Opus 4.8

5  **Aptitudes** — optional

✓ Judge

✓ Reasoning

✓ General

+ Coding

+ Vision

+ Long-context

6  **Expertise** — what routing matches on

Long-form reasoning and careful multi-step analysis over large documents.

Used only by mixture-of-experts configurations. Describe strengths in a sentence — “long-form reasoning over large documents” routes better than “reasoning, long-context”.

7  **Price** — \$ per 1M tokens

Input	Output
\$ 15 / 1M	\$ 75 / 1M

Some self-hosted models are compute-only — no per-token price. Leave both at 0.

8  **Context window** — required

200000

tokens

Prompt and conversation history together. A study's usable memory is set by the **smallest** context window among the models in a set-up.

9  **Enabled**

Experimenters can pick this model.

Annotations

- 1 Editing an existing catalog model** Opened from the edit pencil, the View (eye) icon, or the row identity — the editor shows the **actual catalog entry** (its model id in the title), not a blank “Add model” form.
- 2** The company (or your own server) that hosts the model. Sign-in keys live in the server's settings — never typed in here.

- 3 **Unique key** Stored in `model_catalog` — the menu of models that configurations pick from. Keys are per-provider, so `provider + model` must be unique. Paste the provider's exact model id — this is the string the system sends when it calls the model, so a typo means the call fails.
- 4 The friendly name people see in menus — the exact model id above stays underneath, unchanged.
- 5 Optional tags shown in the picker to help experimenters choose — advisory only, they never change how the model runs.
- 6 **The one field routing reads** expertise — aptitudes above are advisory tags for filtering the picker; this sentence is embedded and matched against each incoming prompt by a moe configuration. Empty means every prompt reads as a tie and the experiment's judge has to break it. One sentence in plain prose describing what this model is good at. A routed (mixture-of-experts) configuration matches each prompt against this text, so write it as a description, not as tags.
- 7 Cost in US dollars per million tokens (chunks of text), input and output priced separately. Self-hosted models are compute-only — leave both at 0.
- 8 **Required, and it governs the whole set-up** `model_catalog.context_window` is **mandatory**: nothing else in the system knows how much a model can read, and a wrong value is not recoverable from a provider's reply. It bounds the chat history replayed to a study's models — and because every member of a set-up must receive identical input, the budget comes from the **smallest** window among the members, not the largest. Overstating one model therefore silently over-fills every model it is paired with. Room for the system prompt, retrieved context, the current prompt and a reserve for the answer all come off it first (see ALGORITHMS §23). How much text this model can read in one call, counted in tokens (chunks of text). Required — nothing else in the system knows it.
- 9 **Retire without deleting** Turning enabled off hides the model from pickers **without deleting it** — existing configurations and past runs still resolve it.
- 10 **Concurrent-edit safety (OCC)** Save issues a **versioned write** — `UPDATE model_catalog SET ..., version = version + 1 WHERE id = ... AND version = (loaded value)`. If someone saved first, the save is refused with “This record changed since you opened it — reload”.
- 11 Discard changes and close.
- 12 Save edits to this catalog model.