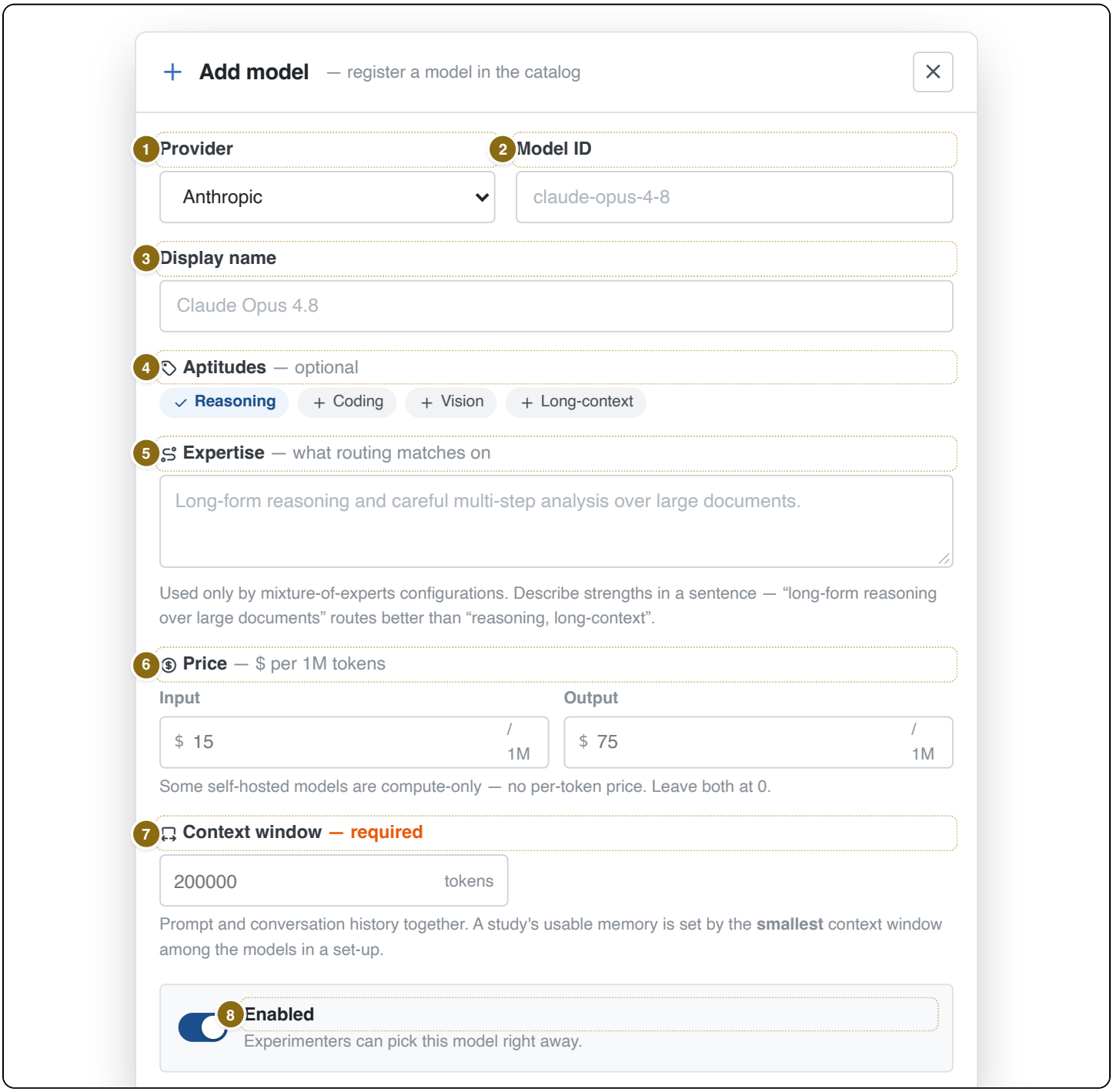
Add model

add-model.html · 760×744



Annotations

1. The company (or your own server) that hosts the model. Sign-in keys live in the server's settings — never typed in here.

2. **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.

3. The friendly name people see in menus — the exact model id above stays underneath, unchanged.

4. Optional tags shown in the picker to help experimenters choose — advisory only, they never change how the model runs.

5. **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.

6. 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.

7. **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.

8. **Retire without deleting** Turning `enabled` off later hides the model from pickers **without deleting it** — existing configurations and past runs still resolve it. On by default so experimenters can pick it right away.

9. Discard changes and close.

10. Register a model in the install catalog.