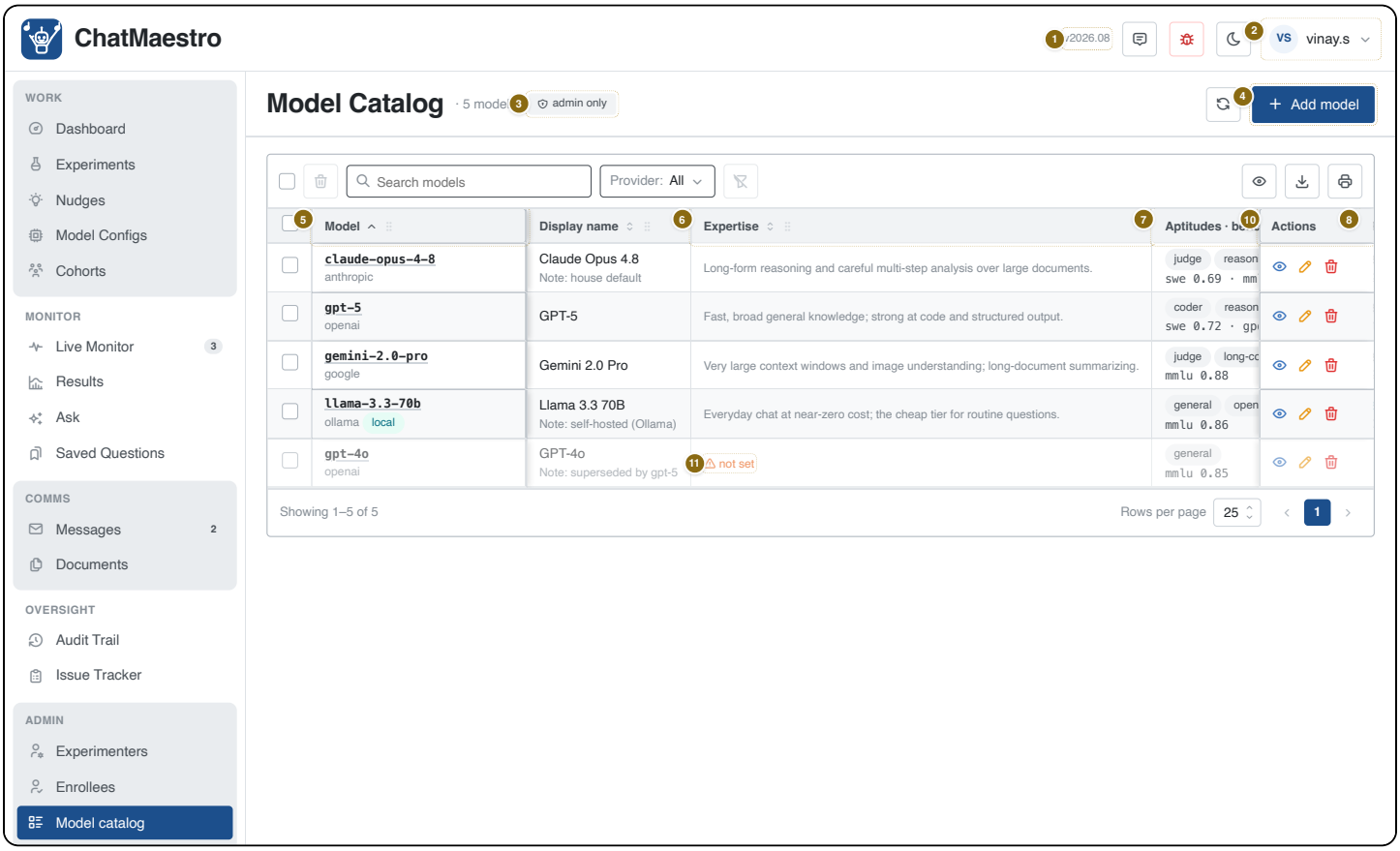
Model catalog

`model-catalog.html · 1360×820`



Annotations

1. **Version pill (right-aligned)** Dimmed build pill sits on the RIGHT, beside the account menu. Hover reveals the build sha; on phones it folds into the account menu.

2. **Account menu** Avatar + name + chevron open the account menu (profile, settings, sign out). On narrow screens the feedback / bug / dark-mode icons and version pill collapse into here.

3. **The model menu** The install's menu of models (`model_catalog`) that model configurations pick from. DB-backed, so you curate it live — **no code deploy**. Holds **no secrets**: API keys are per-provider in `.env`.

4. Register a model in the install catalog.

5. **Unique key** Keys are per-provider, so `provider` + `model` is unique. Each model belongs to a provider; the provider and model id together name the row.

6. **The one field routing reads** `expertise` — unlike aptitudes and benchmarks, which only inform a human, this sentence **is** read by the system: a `moe` configuration embeds it and routes each prompt to the nearest match. Leave it empty and every prompt reads as a tie, so the experiment's judge has to break it — cheap to fix, expensive to ignore. One sentence saying what this model is good at — the text a mixture-of-experts set-up matches a prompt against when it routes.

7. **Advisory only** Aptitudes and benchmarks are *advisory* — they inform the picker; the system never makes a correctness decision from them. Routing reads `expertise`, not these. What each model is good at, plus test scores — a guide to help you pick, not a hard rule.

8. **Frozen at run time** Price (`input_price` · `output_price`, US$ per 1M tokens, input and output priced separately) is read at run time to **freeze each round's dollar cost**, so later edits never rewrite historical spend. A self-hosted model is priced at 0. Cost to run the model, in US dollars per million tokens (chunks of text) — input, then output. A local model is $0.

9. **Retire without deleting** Toggle `enabled` off to retire a model: it drops from pickers but still resolves for existing configs and history. Turn a model off to hide it from menus without deleting it — past experiments still work.

10. **Concurrent-edit safety (OCC)** Editing a row (the pencil) or toggling `enabled` is a **versioned write** on that catalog entry — `UPDATE model_catalog SET …, version = version + 1 WHERE id = … AND version = (the value loaded)`; if someone changed it first, the save is refused with **"This record changed since you opened it — reload."**

11. No expertise sentence — a routed configuration cannot match a prompt to this model, so every round would fall to the judge tie-break.