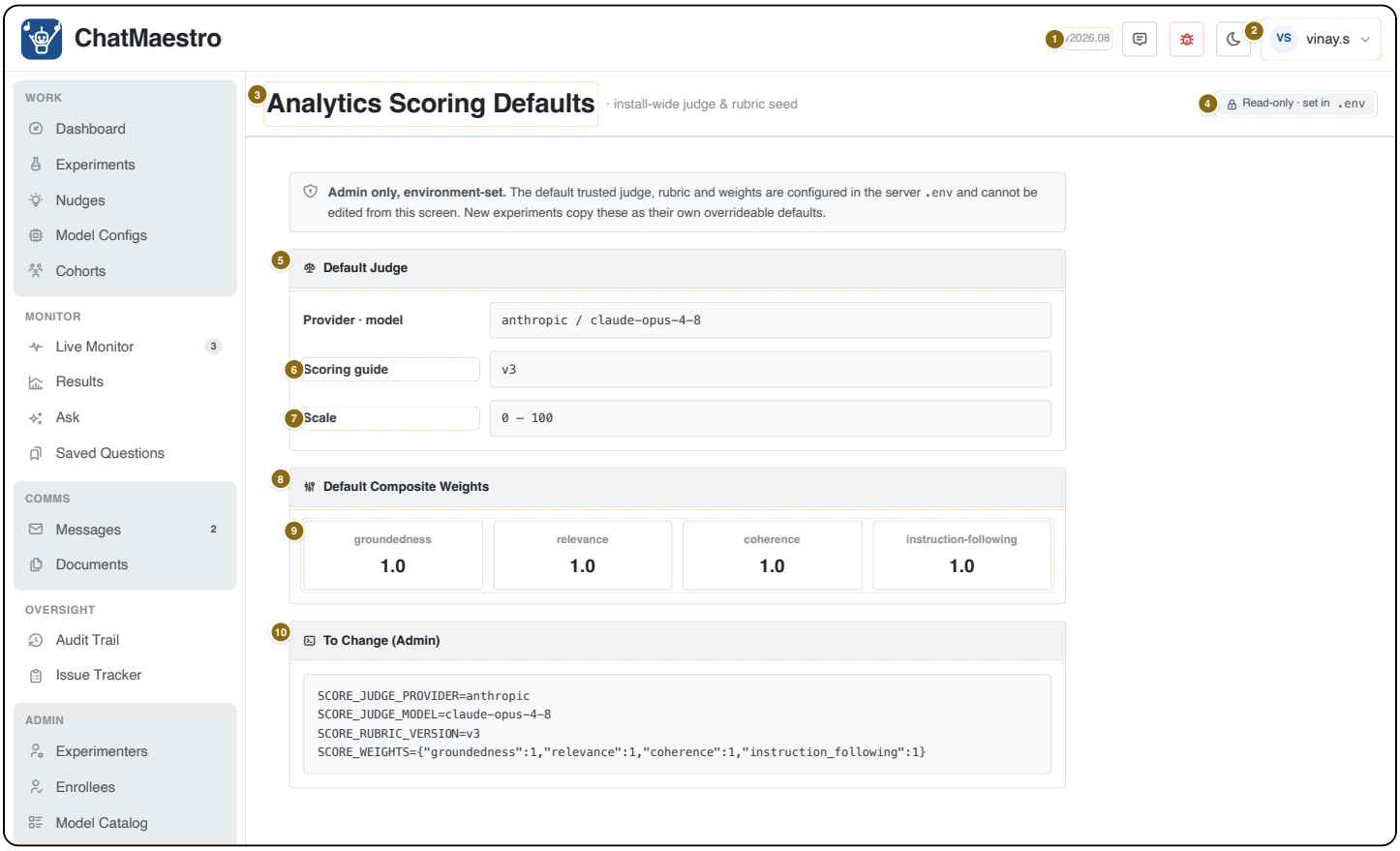
Scoring defaults

`scoring-defaults.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. **Default trusted judge** The install-wide default for the **trusted judge**, its rubric and its weights. Each experiment **copies these as its own defaults** when it is composed (Design mode → Analytics scoring) and then freezes them, so every run of that experiment is graded by one grader and its runs are comparable with each other. Changing the default here affects experiments composed afterwards, never existing ones.

4. These values are set in the server .env and are read-only here.

5. **Curated judge** When an experiment copies this seed it stores the judge as `score_judge_catalog_id` — a reference to the **model catalog**, so the judge is always a curated, priced model. Separate from any inline `judge`-role model inside a model configuration. Provider, model and rubric come from the `SCORE_*` vars in `.env`.

6. The fixed list of 4 questions the judge grades each answer on; v3 is the current version.

7. Each factor is scored 0–100; the overall score is their weighted average.

8. **Weights env var** `SCORE_WEIGHTS` — leave blank or equal ⇒ arithmetic mean. A bigger weight makes that factor count for more in the overall score.

9. **The default rubric's factors** These four are the factors of the **built-in default rubric**, and the weights here are the install-wide default an experiment starts from. The factor set is **not** fixed platform-wide: an experiment picks its rubric through `experiment.score_rubric_version` from a registry-backed drop-down, and a rubric measuring something else brings its own factors (say funniness, originality, timeliness). `experiment.score_weights` is keyed by whichever factors that rubric defines. There is still no judge panel — one trusted judge does all the scoring. A weight makes one factor count more toward `score_composite`.

10. **Seed only** Changing these updates the **seed** for *new* experiments only — existing experiments keep the copy they were composed with; a reboot applies it. There is no `app_settings` table; all install config is `.env`. Edit these settings in .env, then restart to apply.