

Model configs

model-configurations.html · 1360×820

ChatMaestro 1/2026.08 2 PN priya.n

Model Configs · 8 configurations

+ New model config

Search model configs Combine: All Self-improve: All

Name	Combine	Models	Self-improve	Notes	Actions
GPT-5 + Claude ensemble	synthesize	11	12 YES 3x	Two strong generalists merged; headline runs.	13
Single · claude-opus-4-8	single	14	15 NO	Fast single-model baseline.	
Judge-best trio	judge_best	16	3 NO	Pick the best of three per turn.	
MoE router · 3 experts	moe	18	3 NO	Cheap / strong / long-context experts; one answer per row.	
Vote · 3-model	vote	3	3 NO	Weighted majority vote.	
Haiku speed-single	single	1	NO	Cheapest single model; smoke tests.	
Opus + Sonnet synth	synthesize	2	2 YES 2x	Anthropic pair; GPT-5 as a third generator.	
Long-context judge panel	judge_best	4	NO	Long docs; 4 candidates scored.	

Showing 1–8 of 8 Rows per page 25 1

Annotations

- 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.
- 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.
- Reusable model** One logical model to answer with — its constituent models + how they combine. Authored **independently** of any experiment and reused across runs (a run picks one at launch, like it picks a cohort). The common case is a single model with Single, which short-circuits the ensemble machinery. Edits apply to future runs; a run **freezes** its config at launch.
- Create a new model configuration.
- Confined horizontal scroll** The scrollbar belongs to THIS viewport — only the middle columns scroll. The Select column (left) and Actions column (right) stay pinned; the Name column is pinned-left too, so the row's identity never scrolls away.
- Name** name — unique & reusable across runs.
- Combine method** combine_method = single / synthesize / vote / consensus / judge_best / moe — how the constituents become one output. The first five **fan out then reduce**; moe (mixture of experts) **routes then calls one**, matching the prompt against each model's expertise and escalating a tie to the experiment's trusted judge.

- 8 **Constituent models** Every member is a **generator** identified by its `catalog_id`, with a temperature and a weight. There is **no role or lead flag** — the highest-weight member is the aggregator under Synthesize. No judge member: any scoring a method needs is done by the experiment's **single trusted judge**.

- 9 **Self-improvement loop** When enabled, the config loops **improve** → **score** → **refine** (feed the best/aggregate back to the generators). The *inline* score that drives selection comes from the experiment's **single trusted judge** (or a rubric peer-vote) — **separate** from the uniform analytics use of that judge. Each pass is saved as a candidate.

- 10 **Notes** notes — free-text, searchable.

- 11 **Per-model overrides** temperature is per-use; weight is relative, normalized at combine time. The **Prompt** column sets each model's override mode — inherit / append.

- 12 **Loop controls** `max_iterations` (blank = single pass) · `score_target` (0–100, default 90) · `min_score_gain` (stop under this gain) · `underperform_threshold` (prune, default 50).

- 13 **Concurrent-edit safety (OCC)** Save issues a **versioned write** on the configuration — `UPDATE model_config SET ..., version = version + 1 WHERE id = ... AND version = (the value this form loaded)`. Editing its constituent models bumps the *same* version, so the config is guarded as one unit; a stale save is refused with “**This record changed since you opened it — reload.**”

- 14 Higher values make answers more varied and creative; lower values make them more focused.

- 15 **Inline weights** `score_weights`; null = equal. This is the config's *inline* weighting (drives which candidate wins) — distinct from the experiment's **analytics** weights used for cross-run comparison.

- 16 How much each model counts when the answers are combined.

- 17 Mixture of experts — a router picks ONE model to answer; the others are never called.

- 18 **Routed, not fanned out** Three members, but only ONE answers each round: the prompt is matched by meaning against each model's `model_catalog.expertise`, and a tie escalates to the experiment's trusted judge at temperature 0. So generator cost is one call per round however many experts there are, and the per-round decision is recorded in `chat_round.moe_routing`.