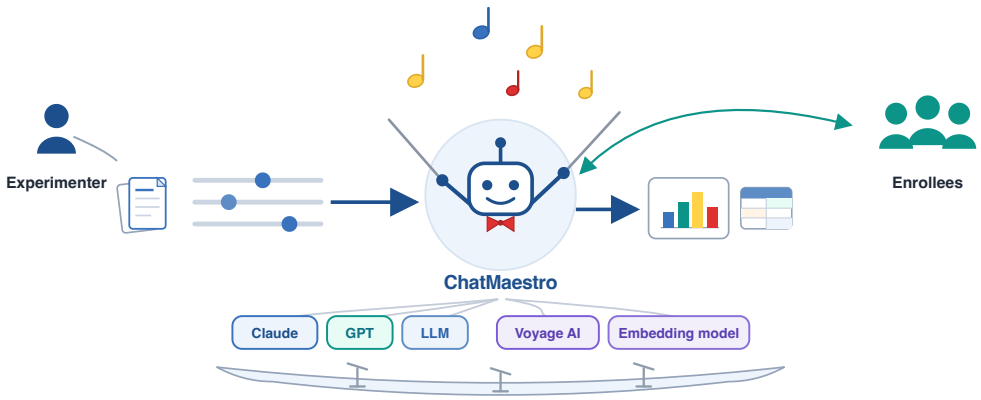
ChatMaestro

Controlled experiments for large language model (LLM) answer quality.

DATA SHEET
OVERVIEW

Revision: 2026-08-30



ChatMaestro runs controlled experiments on how the way you configure or talk to a large language model changes the quality of its answers. An experiment fixes a shared set-up and varies one parameter at a time across its runs — each a session with a cohort of real people, all scored the same way by one trusted judge. Runs are comparable within their experiment, and each stays reproducible for audit.

THE PROBLEM

You can't compare two LLM set-ups fairly by reading a few chats. Answer quality shifts with the system prompt, the guidance you give, the model or ensemble of models behind it, the opening message, the framing, and even the audience. Change several at once, and the result tells you nothing. Score each set-up with a different judge, and the numbers don't line up. A fair comparison changes one factor at a time, holds everything else constant, and scores every run with the same judge. That is exactly what ChatMaestro is built to do.

HOW IT WORKS

1

Compose an experiment

Fix the shared set-up every run holds constant — prompts, scenario, opening message, context documents, scoring, one judge — and let the nudge, model configuration, and cohort vary.

2

Launch a group of runs

Launch several runs that change just one of those parameters and hold the others fixed, each against its own cohort. Every set-up is frozen at launch, and enrollees chat in real time.

3

Score every run

The one trusted judge scores every answer on the same rubric, and each run rolls up into one scorecard — comparable to the other runs in its group.

4

Compare the run group

Run analytics across the scorecards of a group of runs. Because the runs differ in just one parameter, any difference in score points at it. Comparisons stay within the experiment.

WHAT YOU GET

Controlled comparison

Across an experiment's runs you vary exactly one parameter — the nudge, the model configuration, or the cohort — and hold everything else fixed. Any difference in the scores then points at the one thing you changed; runs stay comparable within their experiment.

One comparable scorecard

Each run produces one weighted 0–100 score across the rubric's quality factors — four by default — ready to rank the experiment's runs side by side.

Bring your own documents

Upload or link files; ChatMaestro extracts the text (including scanned PDFs) and retrieves it into a run's chats.

Single models, ensembles, or routing

Run a single model, or combine several by synthesizing them, voting among them, finding their consensus, or letting a judge pick the best. Or route: a mixture-of-experts set-up sends each question to the model best suited to it, so you pay for one answer instead of several — and the platform measures whether that pays off. Optional self-improvement passes refine an answer before it's shown, and an optional prompt enhancer can sharpen the user's request before the models answer.

Ask your data in plain language

Ask analytical and operational questions and get a short narrative plus downloadable tables and/or charts. Common questions are one-click saved buttons, and run results can always be exported to a third-party analytics tool.

Full provenance

Every run's set-up, cost, and history stay recorded and queryable, and a finished run's results never drift because the set-up it used is locked while any run references it.

WHY THE RESULTS ARE TRUSTWORTHY

- **Frozen set-ups**
  Each run's complete set-up is locked in when it launches. Editing a study later never changes what a past run used, so a finished run's results stay valid over time.
- **One trusted judge**
  A single automated judge — an industry-standard, reference-free "LLM-as-a-judge" — is fixed for the whole experiment, so it scores every answer of every run the same way, on the same rubric, with no reference answer needed. That one constant grader is what makes an experiment's runs comparable rather than a matter of opinion. To check it against people, export the conversations and their scores and compare offline.
- **Precise cost accounting**
  Cost is tracked for every exchange and split between generating answers and judging them, so a constant measurement overhead can never distort a comparison.

WHO IT'S FOR

Teams choosing between models, prompts, or prompting strategies need evidence, not opinion.

Researchers study how guidance and framing change the way a model behaves with real people.

Anyone who needs defensible, comparable results can hand them to a stakeholder and re-run the exact same setup next quarter.

BUILT FOR INTEGRITY

Blinded by default. Enrollees are pseudonymous, appearing under system-assigned handles rather than their real names, and they don't see which set-up they're in.

Runs stay isolated. An enrollee can't be pulled into two live runs at once, so runs don't bleed into each other.

Ask, safely. Every question runs read-only and only ever sees what the person asking is already allowed to see.

AT A GLANCE

Roles	Admin · experimenter · enrollee (enrollees pseudonymous and blinded by default)
Scoring	A reference-free LLM-as-a-judge rubric of quality factors, chosen from a built-in list — the default four are groundedness, relevance, coherence, instruction-following — combined into a weighted 0–100 composite by one trusted judge, applied the same way across every run of an experiment
Models	Any hosted or local model; single models or ensembles, with optional self-improvement and Socratic questioning
Documents	Upload or link → extract the text, using optical character recognition (OCR) for scanned files → retrieved into a run's chats
Ask	Plain-language questions — analytical and operational — answered as a narrative plus downloadable tables and/or charts; run results also exportable to a third-party analytics tool
Access	Everything you do runs under your own permissions