Anar Labs

Anar Eval

Test before you approve.

Eval runs your model on Arabic and GCC question sets, scores each answer and ranks the models you test.

For
AI programme and procurement leads

Date
November 2026

The problem

Teams choose models with no test on record.

01

No written result

A team picks a model. Nobody can show the test that supports the choice.

02

Arabic and GCC topics

Your services are in Arabic and about GCC government. The test must use the same content.

03

No side-by-side view

Each team tests in its own way. You cannot compare the results.

What it is

One test run. One score for each model.

01

Your model

You supply the model under test. Eval reaches it through Anar Gateway.

→

02

Anar Eval

Asks each question. Scores each answer by exact match, fuzzy match or a judge model.

→

03

The leaderboard

A score for each model and each benchmark. Compare models side by side.

Anar supplies no models. You choose what to test.

The question sets

Eight benchmarks in Arabic and on GCC topics.

147

questions in 8 benchmarks. 140 of them contain Arabic text.

Count from the product source, November 2026.

- **Arabic language:** comprehension, reasoning, summarisation and translation.
- **Arabic safety:** safety and personal-data detection.
- **Government:** GCC government knowledge and government service questions.
- Anar wrote these sets. They are not published benchmarks.

Run and compare

Run a test. Read the ranking.

Run a test

Choose a model and a benchmark. Eval runs the questions in the background.

Shipped

Scoring

Exact match or a judge model, set for each question. Fuzzy match is the fallback.

Shipped

Compare and export

Compare models side by side. See the history of each model. Export results as JSON.

Shipped

Leaderboard

A ranking of models by score. After a restart, it can leave out older runs.

Partial

Shipped Partial Roadmap

Know the limits

Three limits to know before a pilot.

Judge fallback

If the judge model fails, Eval scores with fuzzy match. The result does not say so.

Partial

Model errors

A failed model call gets a score like an answer. It shows as a low score, not as an error.

Partial

Safety run

The safety run does not reach Anar Guard today. On an error, it marks the answer safe.

Partial

Your control

Your models, your judge, your questions.

Your model

Eval reaches any model that Anar Gateway serves. A list of direct endpoints is on the roadmap.

Partial

Your judge

Today you cannot choose the judge model. A judge that you choose and pin is on the roadmap.

Roadmap

Your questions

Today the question sets ship with the product. Your own question sets are on the roadmap.

Roadmap

Status today

What works now, and what does not yet.

Capability	Status	Note
Benchmarks, runs, scoring, compare, export	Shipped	8 sets, 147 questions.
Leaderboard	Partial	Can leave out older runs after a restart.
Judge and error handling	Partial	Fallbacks and model errors are not flagged.
Your own model endpoint	Partial	Through Anar Gateway only.
Safety run through Anar Guard	Partial	Does not reach Guard today.
Your own judge, your own questions	Roadmap	Planned. Not built.

Roadmap

The next steps make each result repeatable.

Now

- Eight Arabic and GCC benchmarks
- Three scoring modes
- Compare, history and export

Next

Roadmap

- Arabic text normalisation in scoring
- A list of your own model endpoints
- A record of each run that you can repeat

Later

Roadmap

- Repeated runs with confidence intervals
- A judge model that you choose
- Gulf dialect questions
- Adapters for published Arabic benchmarks

The platform

Eval uses the Anar platform for model access.

Anar Gateway

Eval can send every model call and every judge call through the gateway when you set its address.

Anar Guard

Eval can send its events to the Guard audit trail when you set the Guard address.

Test your shortlist with us.

A one-month pilot: we deploy in your environment and run it with one team.

anarlabs.ai/contact