

Your first plan, one step at a time.

A visual guide to choosing model prices, understanding your estimate, and keeping a plan you can revisit.

Updated 30 October 2026

Online: modelfitopenfire.vercel.app/guide

The walkthrough uses a support chatbot with 1,000 tasks per month. Usage is illustrative, not measured traffic. Screenshots show the real website and catalog prices captured on 30 October 2026; prices and available models may change.

Follow the written steps, enlarge a screenshot, or play a short walkthrough. The HTML and GIFs work offline; links to the actual planner require internet access.

This PDF uses still screenshots. Open the online guide or the included HTML for the animated walkthroughs.

01. Start with your workload

02. Choose a model and its price

03. Describe the calls behind one task

04. Read the estimate without overreading it

05. Save without getting stuck

06. Export a copy you can share

07. Revisit prices without losing your baseline

08. Use an account only when you need one

Start with your workload

One task means one useful outcome, such as answering a customer question.

1. Open Plan. You can use the planner without an account or an inference API key.
2. Under Work type, choose Support chatbot for an editable starting point, or keep the blank plan.
3. Name your project, choose Monthly or One-time project, and enter Tasks per period. You can also calculate volume with Users × tasks per user.
4. Replace the example's 1,000 monthly tasks with your expected usage. A budget is optional.

An example supplies usage assumptions, not provider prices or proof of model quality. Choosing another example can replace your draft; save important edits first.

• Required to save. Other fields are optional.

What are you building?

One task could be one customer answer or one summary.

Project name *

My support chatbot

Work type ⓘ **Planning period ***

Choose an example workload ▼ Monthly ▼

Estimate your volume **Tasks per period *** ⓘ

I know the task count ▼ 1000

Budget in USD (optional) ⓘ

Calculated in your browser. No API calls.

Ready to save. Add model prices when you want a cost estimate.

Back Choose model prices

The first step defines what you are budgeting for.

Choose a model and its price

A model and the provider serving it are a pair. Compare the price of the exact pair you intend to use.

1. Open Models & prices. Search in Find a model or provider, then select a Sourced model rate.
2. Check both Input price / 1M and Output price / 1M. They are USD per one million tokens, not per request.
3. Use Browse sourced prices to compare offers and follow a provider link. You may enter verified rates manually instead.

“Copied snapshot; not silently refreshed” is an information notice. The plan keeps the selected prices until you choose to update them. “Observed” records when ModelFare fetched the source catalog, not when the provider changed its prices. Models.dev is a community source; confirm important budgets against the provider.

* Required to save. Other fields are optional.

Choose your model rates.

Select a sourced rate or enter your own, USD per million tokens.

[Browse sourced prices](#)

Model 1

Choose a sourced rate

Find a model or provider

qwen3.8

Sourced model rate

Select a model and provider

Model name / ID *

my-model

Input price / 1M Output price / 1M

0.99 1.49

Model: qwen-3.8-27b at Cerebras. Source: Models.dev - observed 29/10/2026. Copied snapshot; not silently refreshed.

[+ Add model](#)

Ready to save.

[Back](#) [Map your workflow](#)

Historical capture of a sourced offer, not a current provider quote or a model recommendation.

Describe the calls behind one task

A task may need one model call or several stages. Include the work that actually happens.

1. Open Workflow. Give each stage a name, choose its model, and set Calls per task.
2. Enter Input tokens / call and Output tokens / call. Input includes instructions, history, documents and tool results sent with that call.
3. Use Add stage when another step, such as a separate review, makes another model call.
4. Help me estimate input tokens can turn a representative text sample into a rough local assumption. Select its target stage and click Use this assumption.

The text helper estimates characters ÷ 4. It fills input tokens only. Set output tokens separately, and include any context missing from your sample. It is approximate across languages and tokenizers; pasted text is processed locally.

• Required to save. Other fields are optional.

Map your workflow.

Add a stage for each model call in your workflow.

Stage 1

Stage name *

Answer a customer question

Model *

my-model

Calls per task *

1

Input tokens / call *

1000

Output tokens / call *

500

Compare models for this stage

+

Scenarios, retries & fallback

+

+ Add stage

Help me estimate input tokens

+

Ready to save.

Back

Save on this device

Start with one stage, then add only the extra calls your feature needs.

Read the estimate without overreading it

The number is a forecast from your inputs. It is not a bill, a benchmark, or a savings guarantee.

1. Your estimate updates as you edit. Low, Base and High select different token assumptions.
2. Open Scenarios, retries & fallback in a stage to edit those assumptions. Blank low/high token values use the base value.
3. For retries, 0.1 means a 10% retry probability. Maximum attempts includes the first call. A fallback represents one additional call at its escalation probability.
4. Read Where the money goes for primary calls, retries and fallback costs. Review What this estimate includes and Prices behind this estimate.

Low/base/high are scenarios, not confidence intervals. The estimate covers supported text-model API charges; hosting, storage, taxes, external tools and unsupported pricing rules are excluded. Unknown prices never mean free.

Your estimate

Example usage, not measured spend. Verify prices before spending.

Low

Base

High

\$1.74

base scenario - per month

Per task \$0.001735

Low-high range \$0.89-\$2.97

Where the money goes

Answer a customer question	\$1.74
Primary calls	\$1.74
Retries	\$0.00
Fallback	\$0.00

Answer a customer question is your biggest cost driver.

What changes between scenarios? +

What this estimate includes +

Prices behind this estimate +

Export plan +

USD text API estimate. Hosting, taxes, and unsupported charges excluded.

These totals belong to the captured example and its historical prices. Your workload produces different totals.

Save without getting stuck

An asterisk marks required fields. The checklist tells you exactly what is still missing.

1. Look above the save controls for Complete ... required fields to save.
2. Click a missing-field button to open the right step and focus that input. Fill every remaining required field.
3. At the Workflow step, select Save on this device. Look for Saved on this device. Nothing was uploaded.
4. Return to Saved in the navigation to reopen your plan. Save again after making changes.

A blank price can be saved as unknown, but cannot produce a complete estimate. Invalid numbers still block saving and show a reason. Device saves stay in that browser; clearing site data or using a different device will not preserve them. Export a copy for portability.

Complete 2 required fields to save.

Task volume

Stage 1: output tokens per call

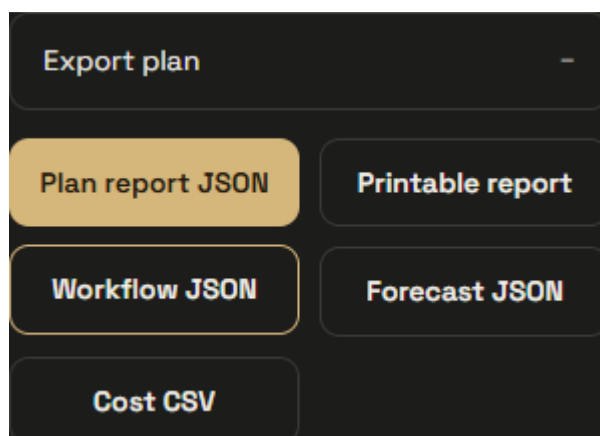
Applying the input-token helper alone does not fill task volume or output tokens.

Export a copy you can share

Use an editable file for your next planning session, or a readable report for a review.

1. Open Export plan beneath the estimate.
2. Choose Workflow JSON for the editable inputs. Import it later using Import at the top of the planner.
3. Choose Plan report JSON for the forecast and documented assumptions, Printable report for a readable HTML report, or Cost CSV for a spreadsheet breakdown.
4. Forecast JSON contains calculated results. A complete, valid estimate is needed for forecast and report exports; Workflow JSON can retain unknown prices.

Downloads contain your plan information. Review project names and assumptions before sharing them. Exporting does not publish a public post or upload your plan to another user.



Choose the format that suits the person reviewing your plan.

Revisit prices without losing your baseline

The catalog can change while a saved plan keeps its original assumptions.

1. Open Versions & updated prices beneath the planner.
2. Click Check updated prices. Compare Current base estimate with With available updates and read the per-model notes.
3. When an eligible update is available and the estimate is complete, choose Apply as a new version. This saves a separate device project.
4. No newer source snapshot available simply means there is nothing newer in the published catalog. Manually edited rates are retained.

This checks ModelFare's latest published catalog, not a live provider quote. Refreshes are scheduled every six hours and can be delayed. A newer retrieval can confirm unchanged rates. Stale, missing or conditional offers are not silently substituted.

Has the price changed?

Compare with the latest available catalog snapshot. Your original plan stays intact.

Check updated prices

Current base estimate

\$1.74

With available updates

\$1.74

my-model: No newer source snapshot available.

Apply as a new version

This saves a separate device project, including any current edits. No automatic price substitutions.

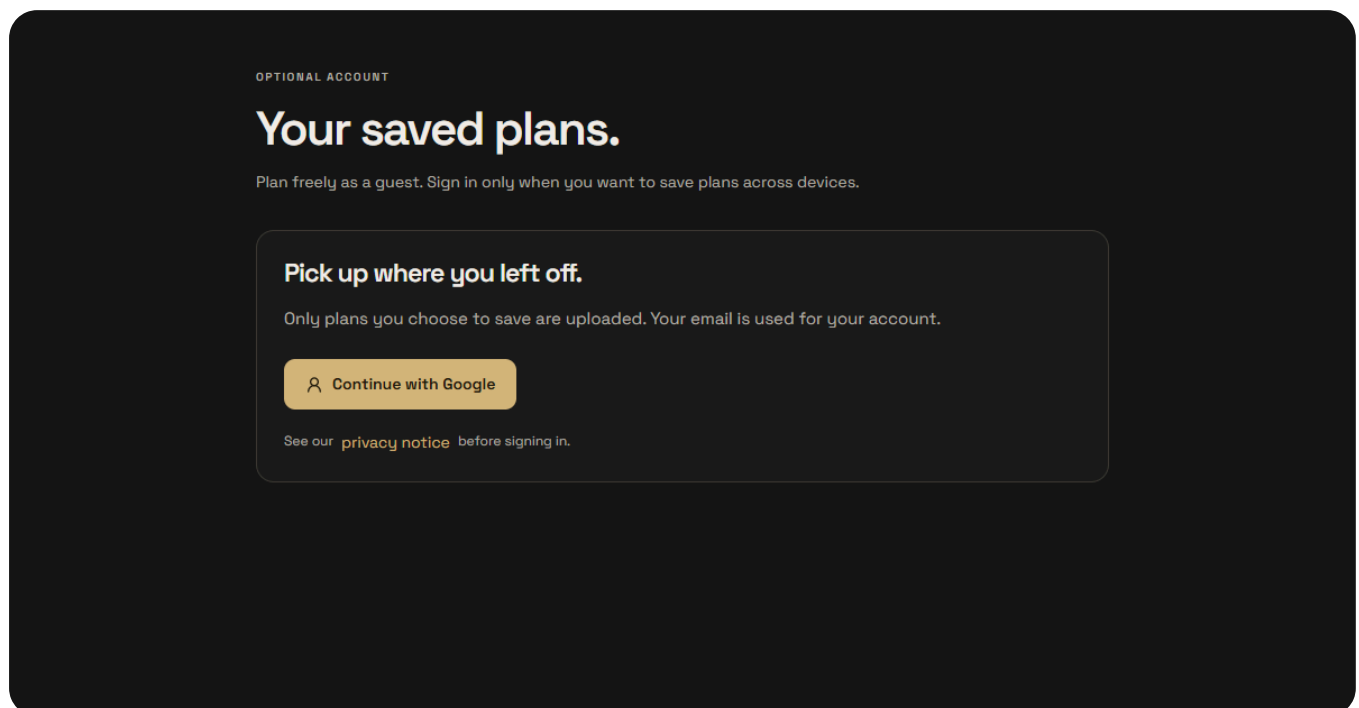
Keeping versions makes a price change explainable instead of surprising.

Use an account only when you need one

Guest planning and device saving remain available without signing in.

1. For private cloud saving, open Sign in and choose Continue with Google. Complete Google's screen and return to ModelFare.
2. Review and accept the current terms when asked, then use Save privately to account for a selected plan.
3. To leave, open Your account and use Delete account and cloud data. Follow the fresh-sign-in and DELETE confirmation prompts.
4. Device copies are separate. Select the device-data deletion option if you also want to remove those copies from the current browser.

Google sign-in is the enabled login method at publication. Email/password registration is not available yet. Accounts process authentication details and explicitly saved plans; this is not a zero-data-collection service. Deletion removes active account data, while provider backups/logs and copies you downloaded have separate retention.



Captured while signed out. No personal account details appear in this guide.

A few words, explained.

Task

One outcome for a user, such as a customer answer. A task can contain multiple model calls.

Token

A unit of text a model processes. Tokens are not the same as words; counts vary by model and language.

Input / output

Input is what the model receives. Output is what it produces. Providers can charge different rates for each.

Retry / fallback

A retry repeats an attempt. A fallback adds a call to another configured model when escalation occurs.

Snapshot

A recorded set of prices and its source timestamp, kept with a plan so it can be reproduced.

Cost per success

Measured evaluation cost divided by successful tasks. The website forecast does not measure task success; use the local CLI for that.

Common questions.

Why is Save disabled?

Check the required-field list first. Task volume, input tokens and output tokens must be filled; other prefilled required fields must remain valid. If nothing is missing, read the validation message. Applying the token helper fills only input tokens.

Why does a selected price say copied snapshot?

The selection worked. That notice explains where the price came from and why it will stay fixed in your plan. Use Check updated prices when you want to compare a newer published snapshot.

Does choosing a model make a paid API call?

No. Website planning uses prices and your assumptions. It does not run inference or charge your model-provider account. Real inference happens only if you separately configure and run the local evaluation CLI.

Does the cheapest model work best for my task?

The forecast cannot tell you that. Test candidate models against your own meaningful checks, then compare task quality and cost per successful result.

Why do my numbers differ from this guide?

These captures use example traffic and the prices available at capture time. Your model, provider, traffic, token assumptions, retries, fallback and later catalog updates can all change the result.

What should I send when asking for help?

Describe the step, the result you expected and the error text. Share a redacted screenshot or export only if needed. Never include provider keys, passwords, login codes, private prompts or another person's data.

Test beyond the forecast.

The website forecasts costs from assumptions. It does not run models or measure their quality. The local ModelFit CLI can evaluate real tasks with meaningful checks and compare pass rate, token usage and cost per success.

Export Workflow JSON to reproduce a forecast without inference. Real model evaluations require configuring a provider, credentials, prices and tasks, and consume your provider credits. Mock defaults are for development, not quality evidence. Small samples do not establish a reliable model ranking.

Agent tasks execute trusted commands in workspace copies. That sandbox is not a security boundary or network isolation.

Read the CLI instructions at github.com/Ujjansh05/ModelFit.

Share your experience.

Try one of your own tasks. Was any input unclear? Could you explain the estimate? Could you save and reopen the plan?

Describe the step, expected result and actual behavior. Share only redacted screenshots or exports. Never include passwords, API keys, private prompts or someone else's data.

[Privacy and contact](#) · [Report an issue](#)

You may redistribute this guide under the included MIT license. The screenshots reflect the product at capture time. Model/provider names identify the selected offers; no endorsement or guaranteed savings are implied.