Fube

GUIDE

The AI Agent budget: how much you can use it and what to do if it runs out



The AI Agent budget: how much you can use it and what to do if it runs out

> There's a conversation almost everyone has with the AI Agent halfway through the second month: you write to it, and the text box won't let you send. Nothing is broken. You've used up the monthly allowance your plan includes for talking to it.

This guide explains where that allowance comes from, what drains it faster than you think, when it fills up again and what you can do in the meantime.

Before you start

The AI Agent is included in **all** Fube plans, the free one included. What changes from one plan to another isn't whether you have it or not: it's **how much you can use it each month**.

And one distinction that saves confusion from day one: **the agent budget and scanning credits are two different things**. Credits are spent when you photograph a receipt so that Fube reads the amount; the budget is spent when you chat. Using the agent for anything else doesn't touch your scanning credits, and scanning a receipt doesn't touch your agent budget.

Step 1 — Check how much you've used

There are two places, and they serve different purposes.

The indicator in the chat itself, next to the message box, is the everyday one: it tells you the percentage of the cycle you've used so far. Hover over it and the breakdown in tokens appears, plus a line reminding you when it renews.

The Settings > Subscription > AI Agent screen is the complete one: the cycle budget, amount used, amount available, the analytics for the period and the interaction-by-interaction detail. It's the only place where you see everything together, and the one you need when your team's usage shoots up and you want to know where it's coming from.

Step 2 — Understand what's being measured

When you expand the indicator you'll see four figures. They aren't four independent counters: they are the four ways a conversation spends.

Metric	What it is
Input	What reaches the agent: your message, the conversation history and any context you've attached
Output	What the agent writes back
Cache	The part of the context that is reused between consecutive messages, and which costs less
Reasoning	What the agent spends thinking before it replies

What you really need to keep an eye on day to day is the **percentage**. The breakdown helps you understand why a thirty-message conversation costs far more than thirty separate questions: in a long conversation, **the entire previous history travels as input with every new message**. That is the mechanism behind almost every nasty surprise.

The indicator changes colour according to the percentage used: green below 50%, amber between 50% and 80%, and red from there on. Once you pass **90%**, the chat indicator turns red and explicitly warns you that less than 10% remains.

Step 3 — Know when it renews (it's not the 1st)

The budget resets on the **monthly anniversary of your sign-up**, not on the first of each month. If you subscribed on 15 January, your allowance fills up again on the 15th of every month.

And it doesn't roll over. Whatever you don't spend in one cycle doesn't carry over to the next: the allowance resets to the same size, whether you've used it up or not.

Step 4 — What happens when it runs out

When you reach 100%, the message box **is locked until the next renewal**. Your previous conversations are still there and you can read them; what you can't do is send anything new.

There's one detail that spares you some grief: **Fube estimates the cost of a message before processing it**. If it doesn't fit within what you have left, it stops it before starting and tells you, so you don't end up with a half-finished reply or burn through what little was left. In that warning, Fube also suggests the plan above yours.

There are only two ways to unlock it: **wait** for the renewal or **upgrade your plan**. You can't buy budget on its own — unlike scanning credits, which can be topped up in packs.

Step 5 — Find out where it went

If usage has shot up, the agent screen has analytics for the period you choose —by default, the last 30 days— with four sections that answer different questions:

- **The period totals**: input tokens, output tokens, cached tokens and the accumulated cost. When no cost has been recorded, an estimate is shown, and it's marked as such.
- **Daily usage**, which is where you can see the spikes. A single disproportionate day is almost always a never-ending conversation.
- **Cache savings**, with the percentage of context reused. The higher it is, the better your way of working is paying off.
- **The most used tools**, which tells you what you really ask the agent for, beyond what you think you ask it for.

And at the bottom of the page, the **interaction history**: one row per exchange, with the model used, the user who started it and the tokens on each side. It's the view you need in a team, because the budget **belongs to the company and is shared by all users**: if it's draining away, that's where you can see who's responsible.

Step 6 — Spend less without giving up the agent

Four habits, in order of impact:

1. **Open a new chat when you change topic.** It's the one that saves the most, and the one people do least. Dragging the context of the invoices conversation all the way to the question about holidays means paying for something that doesn't help answer it.

2. **Use General mode unless the task calls for something else.** The agent has three modes and they don't cost the same: General is the cheap everyday one, Reasoning the mid-range one and Pro the expensive one. For 'how much have I invoiced this month?' all three get it right.
3. **Ask specific questions.** 'Overdue invoices from Acme in 2026' uses a fraction of what 'tell me everything about Acme' uses, and gets a better answer.
4. **Close conversations that are generating at the same time.** Having several chats open costs nothing; having several **generating a response at the same time** consumes in parallel. The new chat button itself warns you about it.

What each plan includes

The four plans differ in **assistant capacity**, which appears in the comparison as four levels: **Essential**, **Advanced**, **Premium** and **Maximum**. Essential is for occasional queries; Maximum is for teams that work with the agent every day. None of them is unlimited.

You'll find the full comparison in **plans and pricing**, and the details of what the assistant can do on the **AI Agent page**.

Edge cases and common mistakes

- **The chat gets blocked halfway through the month, every month.** It's not a one-off usage problem: your plan is no longer enough for you. Before upgrading, check the daily usage and the most-used tools in case the spending comes from a single expensive habit.
- **A colleague uses up the budget.** The allowance belongs to the company, not to each user. The interaction history shows who launched each one.
- **The cost you see doesn't add up to the penny.** Figures marked as *estimated* are calculated using approximate rates per model. The budget and the amount consumed in the cycle, at the top, are exact.
- **You've renewed and the counter is still high.** Check the date: the reset follows the anniversary of your sign-up, not the 1st of the month.
- **You think receipt scanning is using up your agent allowance.** It isn't. They are two separate counters, and the scanning one is topped up separately.

Frequently asked questions

Can I buy more budget without changing plan?

No. The agent budget can only be increased by upgrading your plan. Expense scanning credits work the other way round: those can be bought in packs.

Is the budget per user or per company?

Per company. All users on the same account share the same allowance.

What exactly counts as consumption?

Every message you send and every response it generates, along with its context and reasoning. Opening the chat, reading old conversations or having chats open without typing consumes nothing.

Do the three modes consume the same?

No. The Reasoning and Pro modes answer complex questions better and empty the allowance faster. In the interaction history you can see which model was used for each one.

Do I lose my conversations when I renew?

No. The history is always kept; the only thing that resets is the cycle counter.

Check it once a month and no more nasty surprises

The agent budget isn't something you manage: you just keep an eye on it. A glance at the chat indicator every few days and a visit to the analytics when something spikes are enough to never find the text box blocked on a Tuesday morning.

If you don't use it daily yet, start with **getting started with the AI Agent**; and if you don't have an account yet, you can **create one for free** and try the agent on the free plan.

Fube

START TODAY

Put this into practice with Fube

Create your free account and issue your first Verifactu-compliant invoice in minutes. No card, nothing to install.

fube.es

Invoicing and business management for freelancers and SMEs in Spain