

Cost estimate: running the workflow

Sample document. The business is fictional, and every figure here is an estimate built on stated assumptions. It shows the format of a Teknetik cost estimate.

These are estimates, not measured costs. Replace them with measured figures after the first month of real use. In live mode, the app records the actual tokens used, the time taken and an estimated cost for every AI call. The first monthly care report compares them with this page.

Summary

ITEM	LOW	HIGH
AI model use, per enquiry on average	1.3p	3.1p
AI model use, per month	£1.94	£4.75
Running cost per month without care	£1.94	£14.75
Running cost per month with Teknetik care	£151.94	£164.75

In round figures, that is about **£2 to £15 a month to run, or about £152 to £165 a month with care.** One-off pilot fees are not included.

When the workflow uses AI

- **A website form enquiry makes no AI calls.** The form collects every detail in named fields, so a standard acknowledgement is ready without AI.
- **Every other enquiry makes one call** to read the message into a job card.
- **A second call drafts a reply only when details are missing.** When nothing is missing, or the reply is the firm's fixed wording (an acknowledgement, a decline or the safety message), there is no draft call.
- **Every new message in a conversation means the whole conversation is read again.** That is one more call, and a further draft call if details are still missing.

The model is c`laude`-opus-5, asked for low effort and a structured answer (a job card or a draft in a fixed format).

Assumptions

Each assumption is labelled with where it comes from.

ASSUMPTION	VALUE	SOURCE
New enquiries a week	35	The sample review's baseline
New enquiries a month	152	$35 \times 52 \div 12 = 151.7$, rounded up
Arriving by the new website form	2 in 10, so 30 a month	Our assumption. The form is new, so there is no figure for it yet.

ASSUMPTION	VALUE	SOURCE
Email or voicemail enquiries missing a key detail	4 in 10, so 49 of the other 122	The sample review found 4 in 10 enquiries missing a key detail. We assume the same share for enquiries that don't come by form.
Customers who reply when asked for details	All of them, once, and the reply completes the card	Our assumption. It leans high, because some customers never reply.
Tokens for each AI call	1,200 to 2,400 input; 200 to 600 output	Our estimate. The app's instructions, job card format and approved facts come to roughly 900 to 1,000 tokens a call, counted at about four characters a token. The rest is the enquiry itself, which varies, and the model's answer.
Price	\$5 per million input tokens, \$25 per million output tokens	The supplier's list price for claude-opus-5 when this estimate was written. Check it before relying on it.
Exchange rate	\$1 = £0.80 (so £1 buys \$1.25)	A cautious planning rate, not today's rate. Check the current rate.

A **token** is a small piece of text, roughly three-quarters of a word. **Input** is what we send the model: the enquiry, the approved facts and the instructions. **Output** is what it writes back, including any working it does before it answers. Output costs five times as much per token as input.

The supplier bills in US dollars. These figures exclude VAT and any bank or card fees for paying in dollars.

The arithmetic

AI calls in a month (assumed mix)

ENQUIRY TYPE	A MONTH	AI CALLS EACH	CALLS A MONTH
By the new website form	30	None	0
Email or voicemail, no draft needed	73	1: read it	73
Email or voicemail, asked for details	49	3: read it, draft the reply, then read the conversation again when the customer replies	147
Total	152		220

That is about 1.45 calls for each enquiry, on average.

Per call

ITEM	LOW	HIGH
Input cost	$1,200 \times \$5 \div 1,000,000 = \0.0060	$2,400 \times \$5 \div 1,000,000 = \0.0120
Output cost	$200 \times \$25 \div 1,000,000 = \0.0050	$600 \times \$25 \div 1,000,000 = \0.0150
Total per call	\$0.0110	\$0.0270
In pounds ($\times 0.80$)	£0.0088 (0.9p)	£0.0216 (2.2p)

Per month, per enquiry and per year

PERIOD	LOW	HIGH
Per month (220 calls)	$220 \times \$0.0110 = \2.42 , which is £1.94	$220 \times \$0.0270 = \5.94 , which is £4.75
Per enquiry, on average (the month $\div 152$)	\$0.0159, which is 1.3p	\$0.0391, which is 3.1p
Per year (2,640 calls)	$2,640 \times \$0.0110 = \29.04 , which is £23.23	$2,640 \times \$0.0270 = \71.28 , which is £57.02

Monthly running costs

ITEM	LOW	HIGH	BASIS
AI model use	£1.94	£4.75	The arithmetic above
Hosting	£0	£10	Placeholder allowance. £0 if it runs on hosting the firm already pays for, or £10 for a small cloud server. Replace with the actual price.
Shared mailbox	£0	£0	Assumption: the firm's existing email subscription covers it
Job system API	£0	£0	Assumption: included in the firm's plan. [Confirm: some job systems charge extra for API access.]
Running cost without care	£1.94	£14.75	
Teknetik monthly care (optional)	£150	£150	Published price for one workflow
Total with care	£151.94	£164.75	

Monthly care covers a health check, a review of exceptions and usage, and up to 90 minutes of support or small changes. We acknowledge requests within two working days during agreed business hours. It is not a 24-hour service.

What could push costs up

CAUSE	EFFECT
More enquiries	Every extra 10 enquiries a week adds roughly 56p to £1.34 a month ($10 \times 52 \div 12 = 43.3$ more enquiries, at 1.3p to 3.1p each)
Fewer enquiries by form	If nobody used the form, the same 152 enquiries would make about 274 calls (152 readings, plus 61 drafts and 61 second readings), so £2.41 to £5.92 a month
Customers who write more than once	Each new message means the whole conversation is read again, a little longer each time, plus another draft if something is still missing
Messages that aren't new enquiries	Sales emails, spam, invoice queries and complaints are read too, at one call each. They aren't in the 35 a week.

CAUSE	EFFECT
Longer messages, or the model doing more working before it answers	More tokens per call. If measured use came out at twice our high estimate, AI model use would be about £9.50 a month (2 × £4.75).
Retries after timeouts	The app retries a failed call once. A call that times out may still be charged, and the retry is charged too.
Live test runs	A full live run of the 20-case test set makes about 24 calls, so roughly 21p to 52p a run
A different model	The TEKNETIK_MODEL setting changes the model, and with it the price. We would agree any change with the firm first.

The monthly **spend limit** on the model supplier account (see 02) limits what the workflow can spend. If it's reached, the supplier stops accepting AI calls and enquiries go to the manual queue. Nothing is lost.

Replacing these estimates

After the first month of real use:

1. Take the recorded tokens and cost from the app: the report page gives the month's totals, and each card's audit trail shows every call.
2. Check them against the model supplier's invoice.
3. Replace the assumptions above with the measured figures: the share arriving by form, the share asked for details, the calls and the tokens. Note the period and the number of enquiries.
4. Replace the hosting placeholder with the actual price.

Under care, we do this in the first monthly care report.

Not included

This page covers running costs only. It doesn't include the pilot fee (from £1,950, with the £395 review credited if the pilot is booked within 60 days), office time, or changes quoted separately.