wxrks

RESEARCH PAPER

Buying decisions NOT WORDS

The economics of AI post-editing, run two ways for two years. One localization program, simulated from published rate research, and a real file run through four languages. Working paper WP-2026-02.

2×	½	¼
per-word rate paid to the expert	the program cost in year two	the MQM penalty points that ship

Gabriel Fairman · Founder and CEO, wxrks

Abstract. "AI post-editing" now names two different ways of running a localization program. In one, a post-editor reads every new word of an AI draft. In the other, automated checks read every segment and send the ones they doubt to a senior language expert, with a written finding, and the expert's confirmations are written back into the decisions the next draft is made from. This paper compares the two routes on one simulated program: 150,000 English words a month into four languages for 24 months, with rates taken from published pricing research. It tests the parts the comparison rests on with a real help article run through the wxrks platform in German, French, Brazilian Portuguese and Japanese, and with 105 seeded errors. Every seeded error of major or critical severity reached the expert with a finding that named it. In the base case the expert is paid about twice the post-editing rate per word, and by the second year the Augmented Route costs 49% of post-editing and ships about a quarter of the MQM penalty points. The result depends mainly on the share of text a person reads: the routes cost the same when about half of every month's new words reach a person, against 19% measured. Only one scenario, built with every assumption against the Augmented Route, reverses the cost result. The paper also reports what the comparison cannot settle: how many real errors the checks find, how fast the routed share falls, and how many errors people introduce while they edit.

A note on sources. The amounts come from published rate research and a simulation of one program, and are for illustration; a real program's will differ. The proportions rest on how much of the text a person reads, not on the rates. Every figure is marked as sourced, industry, measured, assumed or simulated, and the method, sources and scenarios are in the appendices. wxrks sells the platform tested here and built the model. Before publication the model and the draft were given to an AI model instructed to find the weakest assumptions and the strongest case for post-editing, and the scenarios include the cases it raised.

01

One sentence, three setups

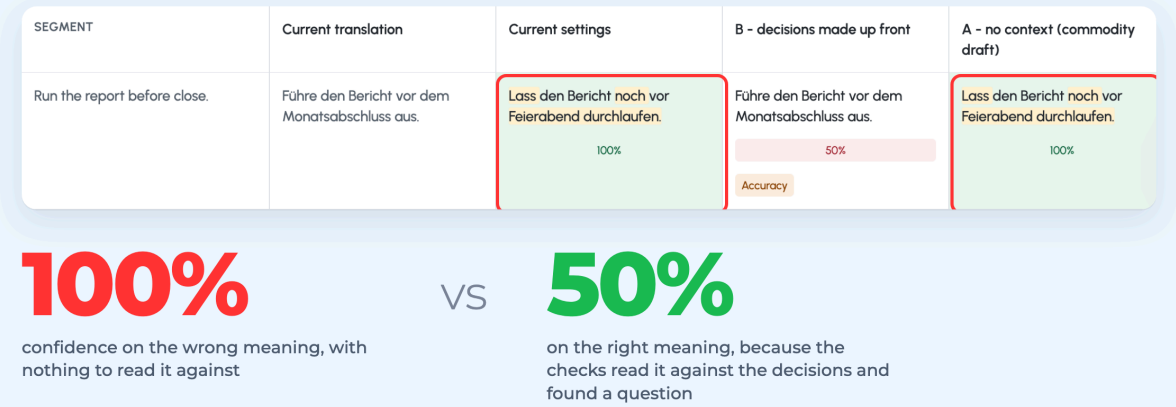
A help article explains the month-end financial close. One sentence reads "Run the report before close." The same article was run through three configurations of one translation setup in the wxrks Translation Lab, and the platform's quality checks scored each draft.

Two configurations told the German reader to run the report *vor Feierabend*, before the workday ends. Both scored 100%. The third wrote *vor dem Monatsabschluss*, before the month-end close, which is what the article is about. It scored 50%, because the check doubted it.

THE SAME SENTENCE, THREE SETUPS

Without decisions, “close” became the end of the working day, at 100% confidence

“Run the report before close”, from the Translation Lab run on the test file. The account's current settings and the no-context draft both write *vor Feierabend*, before the workday ends. One paragraph of decisions writes *vor dem Monatsabschluss*, before the month-end close, and the checks still send it to a person.



100% confidence on the wrong meaning, with nothing to read it against

VS

50% on the right meaning, because the checks read it against the decisions and found a question

Real output, wxrks Translation Lab, demo account, run 5 Nov 2026, captured 6 Nov 2026. Highlights mark differences from the current translation.

The wrong meaning scored 100% because nothing told the check what "close" means here. The right one scored 50% because the check read it against the decisions and found a question for a person.

Nothing was broken. Each draft was scored against its own brief, and the two wrong ones did what their briefs asked. The result locates where the consequential decisions in an AI workflow are made. In a workflow built around reading every sentence, they are made at the end, segment by segment, by whoever reads. In an AI workflow they are made at the start, in the glossary, the instructions and the rule for what a person reads, and the quality of the output follows from them.

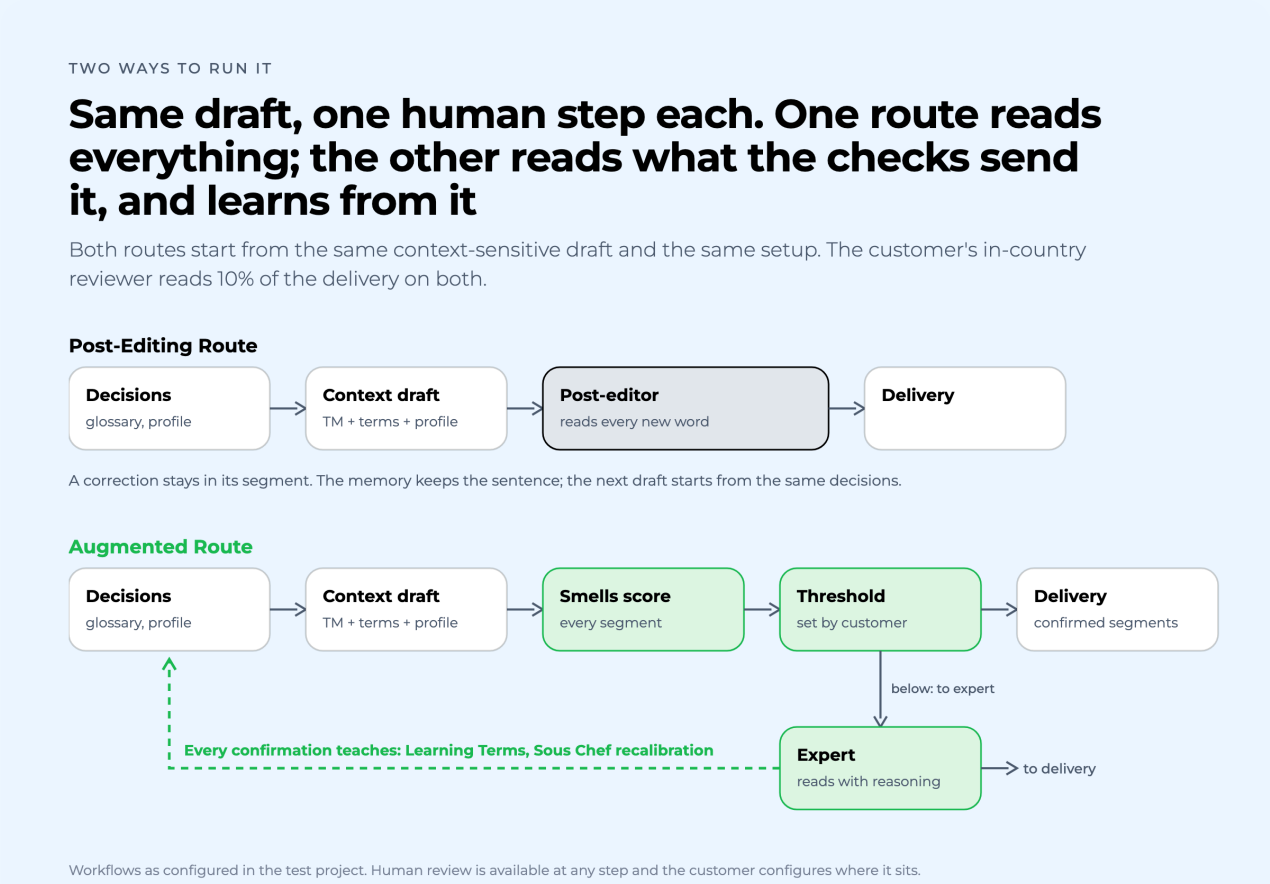
02

Two programs with the same name

The simulated program is a software company localizing about 150,000 English words a month into German, French, Brazilian Portuguese and Japanese. Both routes start from the same AI draft, made with the same glossary and the same instructions, and both have one human step. Both would be called AI post-editing in a request for proposals.

On the **Post-Editing Route**, a post-editor reads every new word and fixes what needs fixing, and rule-based QA (spelling, terminology, tags) runs on what they deliver.

On the **Augmented Route**, a check reads every segment first and sends the ones it doubts to a senior language expert, with a written reason and the model's account of what it decided. Everything the expert confirms goes back into the system, so the next draft starts from it. The route is called augmented because the person in it is supported at each step: told where to look and why, shown the model's reasoning, checked while they work, and remembered once they decide.



Same draft, one human step each. The routes differ in what decides where a person looks, and in whether the system learns from what they do.

The routes differ in what the person does all day, what they need to know, what the program costs in its second year, what reaches the reader, and what a fifth language costs. A score on a segment and a score that decides who reads what, sits next to the model's reasoning and feeds the next draft are different ways of working.

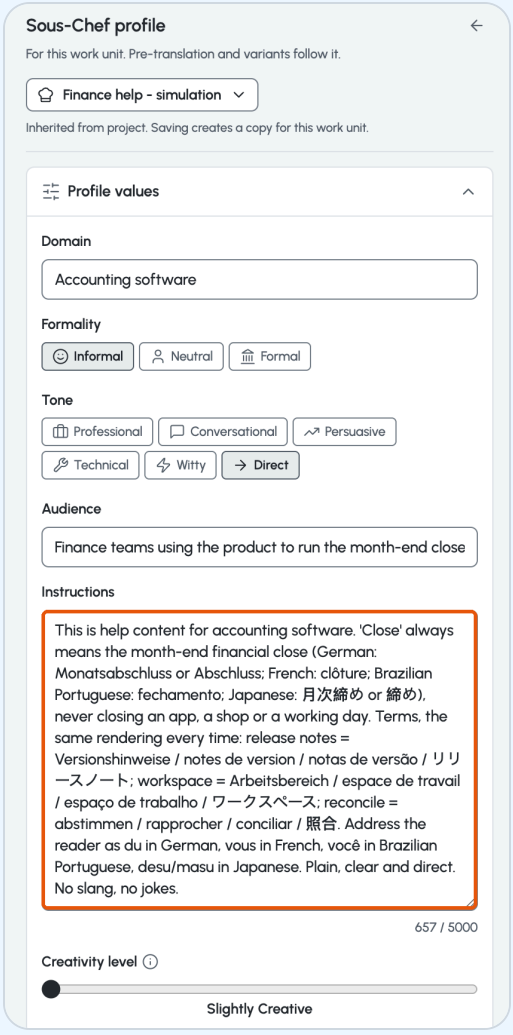
03

Decisions made at the source

The test article uses "close" eight times, always for the month-end close. Machine translation rendered it five ways in German, including closing time and the end of the working day, and five in Japanese, including a deadline and a shop closing. A generic AI pass corrected the grammar and did not fix the meaning, because nothing in its input said what the word meant here.

One paragraph of instructions was then added to the project profile: what "close" means in this product and how to render it in each language, fixed renderings for three other terms, and how to address the reader. With that paragraph, "close" came out one way in all eight places in all four languages.

Without that paragraph, the problem is left to the reviewer at the end of the chain, who has to catch it as it passes. A correction made there stays in its segment. It does not reach the next segment, and translation memory stores the corrected sentence, not the decision behind it, so the same correction recurs in the next file.

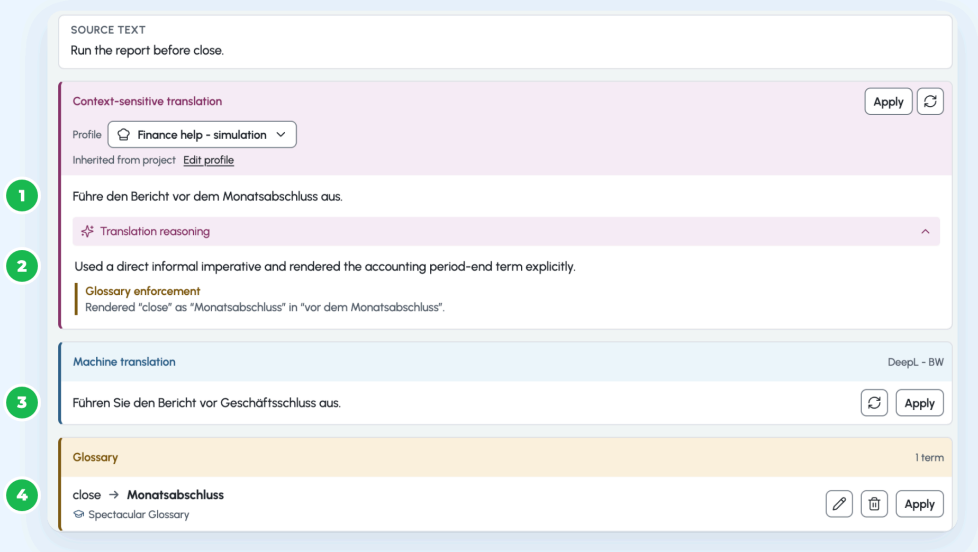


The whole setup for the test file: one profile, and one paragraph of decisions. The reader is addressed as du in German, a choice another company would make differently.

WHERE THE DRAFT COMES FROM

The machine translation says end of business. The context-sensitive draft says month-end close, and says why

The suggestions panel for the same segment in the editor. The draft is written against the project's profile and glossary, and its reasoning names the decision it applied. The machine translation is shown beside it for comparison.



1. The draft, from the **Finance help** profile the project inherits.
2. Its reasoning: the form of address it used, and the glossary term it enforced.
3. Machine translation alone: *vor Geschäftsschluss*, before end of business.
4. The customer's glossary decision: close → Monatsabschluss.

Real output, wxrks editor, demo account, 6 Nov 2026. Test project "Month-end close: getting started", German.

The draft names the decision it applied. A reviewer who disagrees changes the decision once, instead of the sentence every time.

The model gives the Post-Editing Route the same paragraph, so the comparison isolates the human step. The paragraph plays a different role on each route. On the Augmented Route, decisions of this kind are the core of the work. On the Post-Editing Route they are optional, because a reader downstream catches what goes wrong, and the model does not price any difference in how much attention they receive.

ONE WORD, EIGHT TIMES PER LANGUAGE

A generic AI pass cleaned up the draft. One paragraph of decisions made “close” mean one thing

Distinct renderings of the source word “close” across its eight occurrences in the test file.

6

renderings of “close” in Japanese after a generic AI pass

VS

1

with one paragraph of decisions, in every language

	Machine draft	After a generic AI pass	One paragraph of decisions
German	●●●●● 5	●●● 3	● 1
French	●● 2	● 1	● 1
Brazilian Portuguese	●● 2	● 1	● 1
Japanese	●●●●● 5	●●●●●● 6	● 1

Real output from wxrks, test project “Month-end close: getting started”, 21 segments, run 5 Nov 2026.

One paragraph, written once, did what a reviewer would otherwise do in every segment of every file.

04

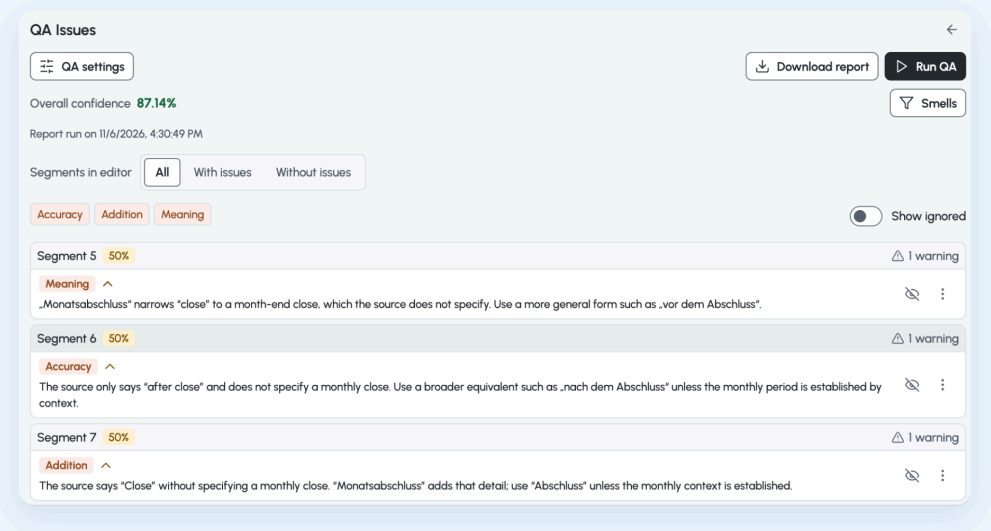
The setup work

The Augmented Route moves work to the setup. Someone audits the glossary, decides the register, writes and tests the Sous-Chef instructions that tell the model who the reader is and what the words mean, chooses which model handles which content, and finds words like "close" before they become questions. Someone then monitors the program: the share of segments reaching a person each month, which kinds of finding are falling, which language drifts, and whether a check keeps raising a doubt the customer has already answered. Quality is read as trends rather than sampled at the end.

THE QUESTION THAT REACHES A PERSON

The checks question the glossary. Whether “close” means month-end close is a judgment about this customer

The QA report on the test file. Six segments are flagged because the English never says the close is monthly, though the glossary says it is. No score settles that. Someone who knows the customer does, and their answer goes back into the decisions.



Real output, wxrks QA report, demo account, run 6 Nov 2026, German.

The checks question the glossary in six places. Whether this "close" is monthly is a fact about the customer, and only someone who knows the customer can settle it.

The model prices that work. The Augmented Route gets twice the setup hours and a quality and data lead at about half a full-time role, which post-editing does not need. It is the largest line the Augmented Route adds, about $91,000 a year in this program *(assumed)*.

The role carries risk, because a configuration error reaches every segment. The first run of the test file picked up a profile the demo account already had, written for teenagers ("super witty, slangy and easy to understand"). German came back with *Feierabend* for "close" and told the reader to "check die Versionshinweise oder frag deinen Admin", and every segment scored 100%. On the Post-Editing Route a bad sentence is one bad sentence. On the Augmented Route a bad brief produces a bad file that scores well.

The Translation Lab runs the same content through several configurations side by side, with the checks' scores beside each, before any of it reaches production. It shows the consequences of a brief; choosing the brief stays with the customer. The skill the route requires is reading text and data together, and recognizing a high score that is high for the wrong reason.

05

Findings and scores

Which segments a person reads can be decided in more than one way, and the choice matters more than any other variable in the comparison.

A common approach is statistical. A quality estimation model gives each segment a number, roughly the likelihood that it needs editing, and segments below a threshold go to a person. The number has three limits. It does not say what might be wrong, so the person reads the segment from scratch to find out. It does not separate a segment that will be edited for preference from one that reverses the meaning of the source, so any threshold either sends too much or lets serious errors through. And unless it is retrained, it does not reflect what the customer decided last week.

The checks tested here work differently. Smells reads each segment against the source, the glossary and the customer's instructions, and reports findings: a category, a severity and a sentence stating what is wrong. The rule for what a person reads is set by severity, so a customer can let a spelling issue pass and send every reversed meaning to the expert, which a single number cannot express. The expert opens a segment that arrives with its diagnosis.

The seeded test. Correct translations of the help article in all four languages were each given one planted error: "reconcile" rendered as closing the account, "the totals can be edited" where the source says they cannot, a wrong number, a dropped clause, untranslated English, an off-glossary term, the wrong form of address, a spelling mistake, the wrong quotation marks. The editor's own check then ran on each.

All 65 errors that change the meaning reached the expert, each with a finding that named it, for example "This says the totals can be edited, reversing the source meaning." Of 84 correct segments, 3 reached the expert anyway, two of them because a demo-account glossary entry told the check that "you" in Portuguese should be the Spanish "Tú": the check enforced a bad decision faithfully. Seeded errors are cleaner than natural ones, and the first run of the file had a natural error the check missed, so the model assumes the checks send 90% of serious errors to a person, not all of them *(measured, then discounted)*.

TESTING THE CHECKS

Every seeded error of Major severity or worse was sent to the expert, in four languages

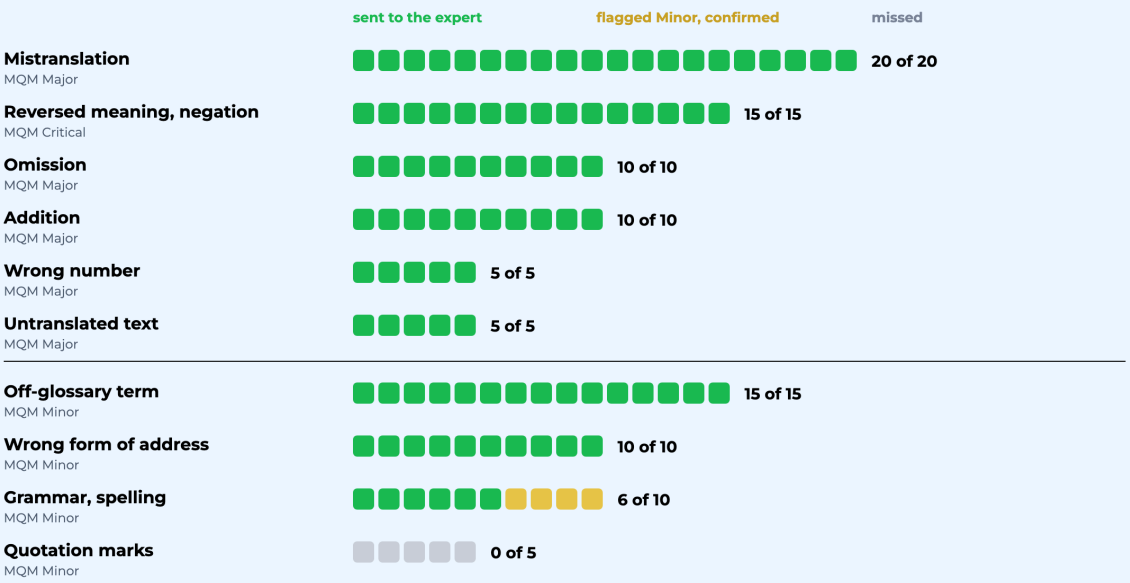
105 targets, each with one MQM error, checked by Smells under the project's finance profile. Each square is one seeded segment. Errors above the line are accuracy errors, Major or Critical; below are Minor.

65 of 65

Major and Critical errors sent to the expert

0

meaning errors that slipped through



Real output, demo account, 6 Nov 2026. German and French run twice, Portuguese and Japanese once. Seeded errors are clearer than natural ones. Set and results: smells-seeded-test.json.

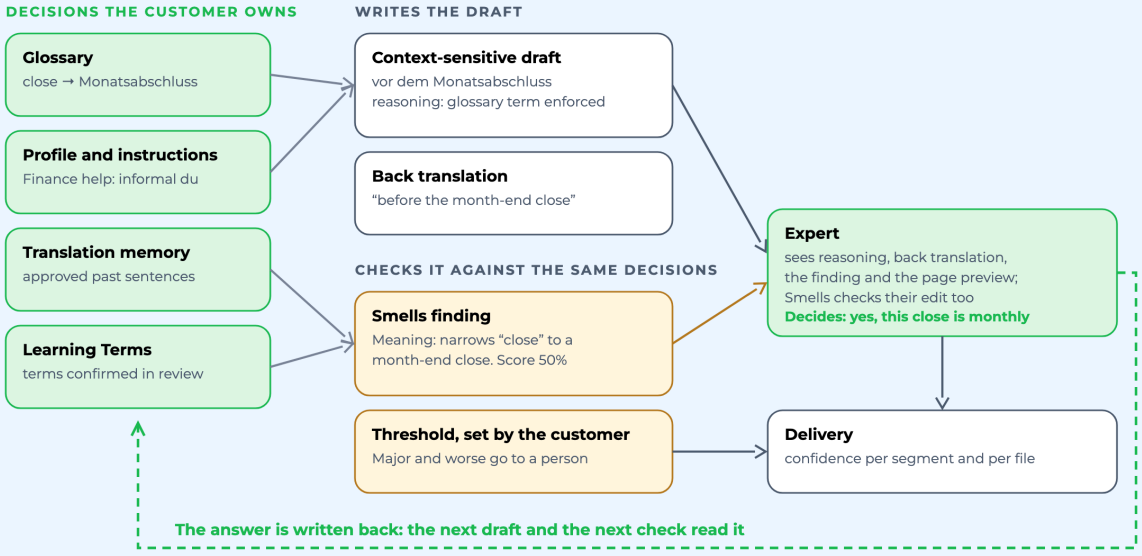
Every seeded error that changes the meaning reached the expert, with a finding that named it. Spelling and punctuation reached delivery unread.

The check also runs on the expert's own work. It runs again each time a segment is saved, so an error the expert introduces is reported too, and minor findings that do not send a segment to a person are still listed for the expert. On the Post-Editing Route, rule-based QA runs once on the delivery.

HOW THE PIECES MEET THE REVIEWER

The draft and the checks read the same decisions, so what reaches a person is a question about the customer

One segment from the test file, traced through the Augmented Route. Context sensitivity means the draft is written against the customer's decisions; augmentation means the person reviewing it gets the reasoning, the findings and the page with it.



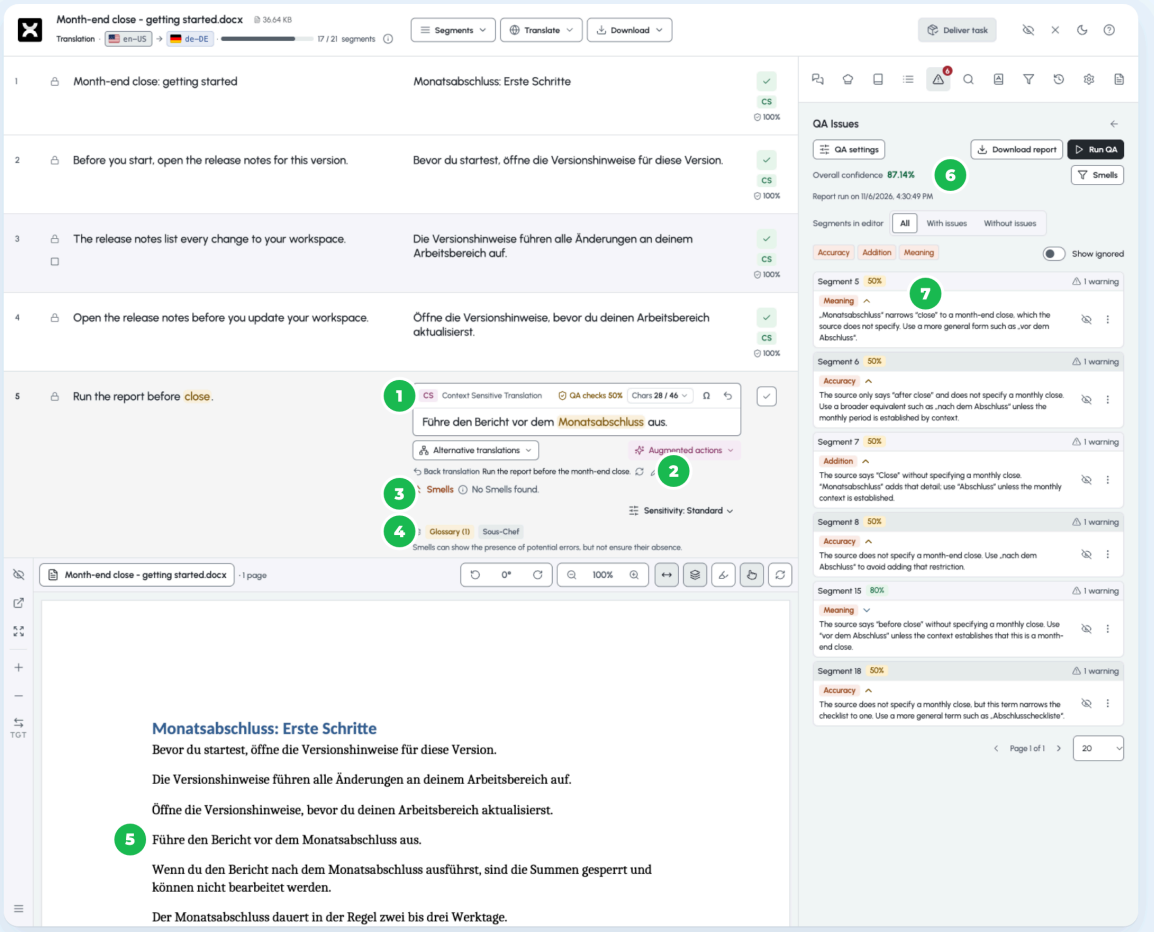
Example values are real output from the German test file, demo account, 5 and 6 Nov 2026. Which checks run, their sensitivity and the threshold are customer settings.

Context sensitivity is the draft reading the customer's decisions. Augmentation is the reviewer receiving what the system knows about the segment. The loop is what lowers the next month's cost.

WHAT THE EXPERT WORKS WITH

A segment sent to the expert arrives with its reasoning, its back translation, its findings and its place on the page

The editor on segment 5, with the in-context preview open below and the QA report on the right. Every aid is a setting the customer controls: which checks run, at what sensitivity, and what reaches a person.



1. The context-sensitive draft and its QA confidence for this segment, 50%.
2. Back translation, to read the German as English.
3. Smells on the expert's own text as they edit, at the sensitivity set for the project.
4. The glossary entry and Sous-Chef profile in play for this segment.
5. The sentence in its page, in the in-context preview.
6. Overall confidence for the file before delivery, 87.14%.
7. Each finding with its category, score and written reason.

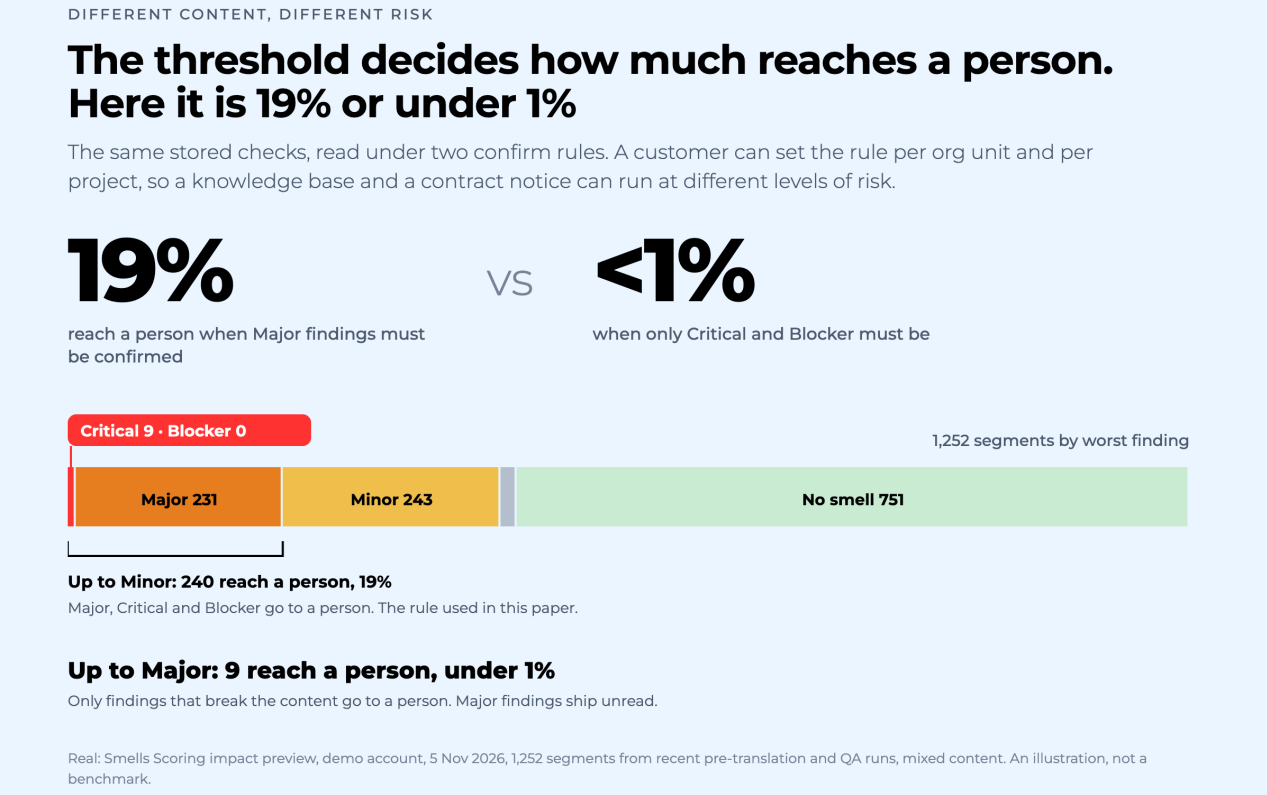
Real screen, wxrks editor, demo account, 6 Nov 2026. Pins added.

What the expert works with. Each aid is a setting the customer controls.

06

Scrutiny by content type

Once a finding decides who reads what, the rule can differ by content. The confirm rule, the sensitivity of the checks and the workflow can be set per organizational unit and per project, so a product release, a contract notice and a knowledge base article can each run under the scrutiny the customer chooses for them.



The same stored checks under two rules, in the account's own data. The customer can preview the effect before changing anything.

The rule used in this paper sent 19% of segments to a person. A rule that sends only critical findings sent under 1%. Where the rule sits for each content type is the customer's decision, and the impact preview shows its effect on stored results before it is applied. When the only lever is how many words a person reads, this choice is not available.

07

Cost

In the second year the expert is paid nearly twice the post-editor's rate per word, the program costs half as much, and it ships about a quarter of the MQM penalty points *(simulated)*.



The second year costs $375,569 against $767,611 for post-editing, and the first year $511,462 against $844,631. Spend on linguists falls from about $550,000 a year to about $130,000. These amounts illustrate one program; the proportions are the finding. With cheaper post-editing, enterprise rate cards or more expensive experts, the second year costs between 52% and 67% of post-editing *(simulated)*. Only the first month costs more, $119,000 against $111,000, because the Augmented Route pays for more setup and for the quality lead. From the second month it costs less, and the gap widens as the draft learns.

The driver is the share of text a person reads. The routes cost the same when a person reads about half of every month's new words. The measured share is 19% in the first month, assumed to fall toward 8% as the draft improves. If the share stays at 19%, the Augmented Route costs about two-thirds of post-editing. At 40%, it still costs less. If the expert also reads 30% of the surrounding text for context, it costs 75% in the first year. The one case where it costs more is the scenario built with every assumption against it at once.

WHAT THE PROGRAM COSTS

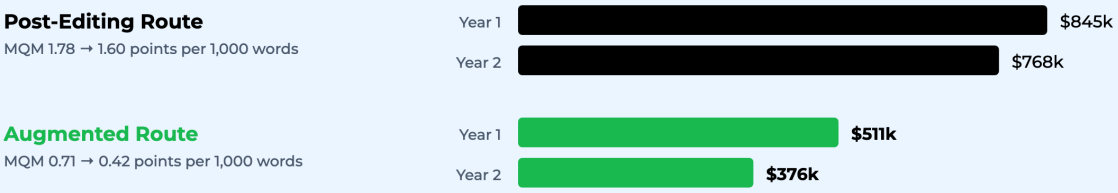
The Augmented Route costs 61% of post-editing in year one and 49% in year two, and scores better in MQM

Cost of the whole program, four languages, and MQM penalty points per 1,000 delivered words (lower is better) in year one and year two.

$333k
less in year one

$392k
less in year two, as the draft learns

3.8x
fewer MQM penalty points in year two



Simulated, base case. Per-word rates from Nimdzi's 2020 UK pricing study (post-editing £0.079, full translation £0.121 to £0.134, at ~$1.28), wxrks list prices, assumed hours and error rates. Model: simulation-commodity-vs-agentic/.

The whole program, four languages, including platform, project management, the quality lead and the customer's own review time.

08

Quality

Quality is scored in MQM: every issue in the delivered text counts, by category and severity, at 1 point for a minor issue, 5 for a major one and 25 for a critical one, per 1,000 words [7]. Inconsistent renderings count as issues, because three post-editors each choosing a defensible word for the same term produce three readings of one decision.

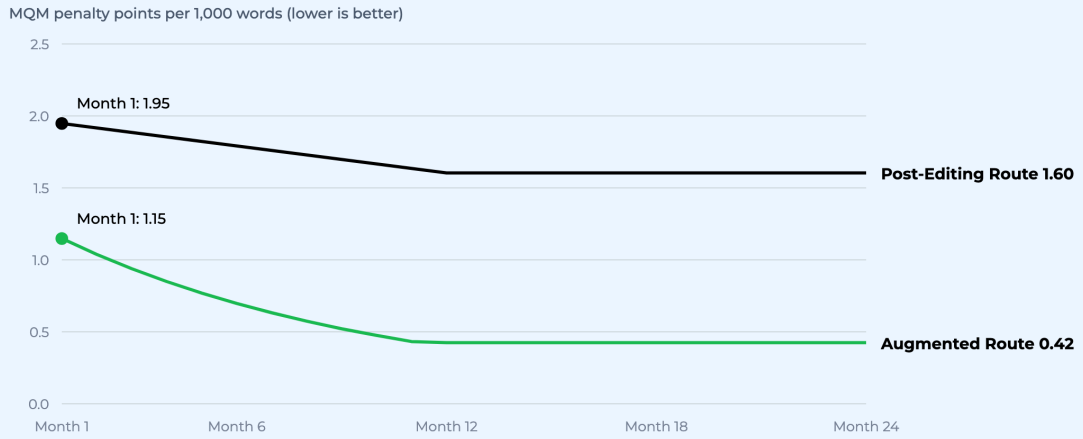
In the second year the Augmented Route ships 0.42 penalty points per 1,000 words against 1.60, about a quarter, and a third of the accuracy errors *(simulated)*. In the first month the gap depends on how many errors people introduce while editing, which the published literature does not measure well. At the assumed rate the first month scores 1.15 against 1.95. With no introduced errors and no checks on the expert's edits, the routes are level. From there the augmented draft improves and the post-editing draft improves only through the memory.

WHAT SHIPS, SCORED

The routes start between level and well apart, depending on how many errors people introduce. After that the Augmented Route pulls away, because its draft learns

MQM penalty points per 1,000 delivered words, all four languages: accuracy, terminology, style, fluency, locale and inconsistent renderings, weighted by severity. The expert is checked by Smells as they work; the post-editor is not.

1.60 post-editing, month 24 VS **0.42** augmented, month 24 **3.8x** fewer penalty points per 1,000 words



Simulated. MQM penalty points per 1,000 delivered words, Minor 1, Major 5, Critical 25 (themqm.org scoring model). Check recall from the seeded test (65 of 65 Major and Critical errors sent to the expert), discounted to 90%; error rates assumed.

Lower is better. How far apart the routes start depends on an assumption not yet measured; after that, the Augmented Route pulls away as its draft learns.

09

Scale

Per-word quotes do not show the cost of adding a language. On the Post-Editing Route each new language adds a team reading every word. On the Augmented Route it adds an expert who reads what the checks send, and decisions made once at the source (the do-not-translate list, answered ambiguities, account instructions) apply to the new language

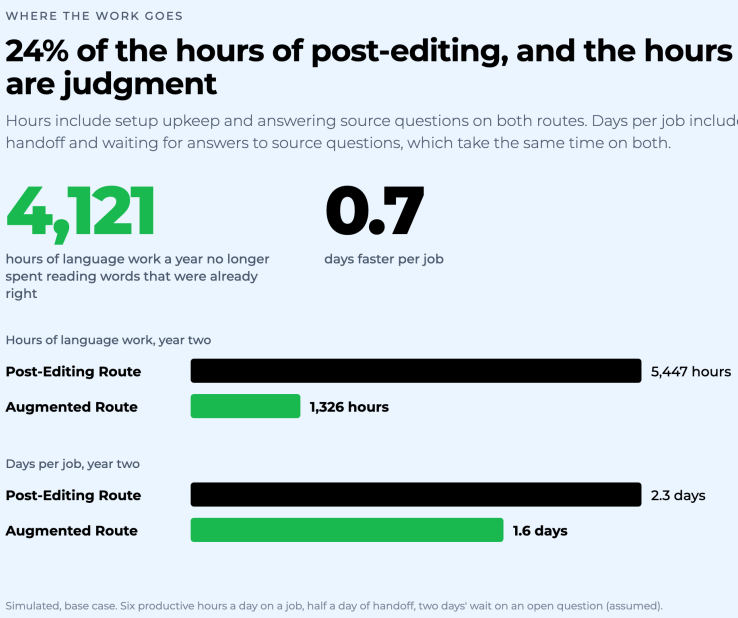
from its first job. The quality lead is a fixed cost, so it falls per language as the program grows: in the second year the Augmented Route costs about half of post-editing per language at four languages, and under 40% at twelve *(simulated)*.

The model holds volume fixed. Whether a lower cost per language leads programs to add languages, content types or markets is not modeled.

10

Labor

In the second year the Augmented Route needs about 1,300 expert hours against 5,400 hours of post-editing, and spend on linguists falls from about $550,000 to about $130,000 *(simulated)*. Each word the expert reads is paid nearly twice the post-editing rate, each hour up to 60% more, and there are about a quarter as many hours.



A quarter of the hours, each paid more.

The work itself changes. A post-editor works through a stream of sentences. The expert receives segments the checks doubted, opens each without having read the ones before it, and so needs product knowledge and often the surrounding text or the whole article. Many of their decisions are written back and apply across later segments, files and languages. The role asks for more qualification than post-editing and the model pays it more per hour.

Two consequences sit outside the model. The route depends on one expert per language, and the model does not price that person leaving. And most of the industry pays by the word, some providers by edit distance. When an expert opens a doubted segment, checks it against the product and confirms it unchanged, the edit distance is zero. Paying for that decision is the subject of a separate paper, *Zero edits, one decision*.

11

Tokens

The model is the cheapest line in the program. Sending all 7.2 million translated words a year directly to a frontier model, with a full glossary, style guide and reference translations attached to every request, costs between about $1,000 and $12,000 a year in tokens at list prices [8, 9]. Against $511,000 for the Augmented Route's first year, that is under 3% of spend.

The rest of the money runs the program: deciding what "close" means and keeping that decision in force across four languages and every later job, knowing which file shipped with which term, and paying the people whose judgment the model does not have. What it costs to run that without a system built for it is outside the scope of this paper.

12

What this paper does not settle

Three quantities decide the result, and each can be measured on a few weeks of a program's own content:

1. How many natural serious errors the checks send to a person under the program's settings.
2. How fast that share falls as decisions accumulate.
3. How many errors people introduce while they edit.

A sample of production segments, scored in MQM by a reviewer blind to which ones reached the expert, would settle the first and the third.

The learning curve is borrowed. It comes from accounts that stayed with wxrks to month ten, which are the accounts where it worked [1]. The learning test in Appendix A confirms the mechanism for one term in one language, not the rate.

13

Appendix A. Method, sources and scenarios

The program. 150,000 English source words a month into German, French, Brazilian Portuguese and Japanese, about twenty jobs a month per language, for 24 months. Translation memory leverage grows from 15% to 30% over the first year, the same on both routes. An in-country reviewer on the customer side reads 10% of every delivery on both routes, and the customer's time is counted as a cost. Both routes get the same shared setup (100 expert hours plus 20 customer hours) and the same monthly upkeep, use the same draft and run on the same platform tier. The Augmented Route adds 60 setup hours per language and a quality and data lead at 80 hours a month *(assumed)*. Post-editing gets rule-based QA that catches half of the post-editor's own minor slips and remaining term misses *(assumed)*.

Kinds of figure. *Sourced:* the wxrks Edit Distance Report [1], wxrks list pricing, the MQM scoring models [7], token list prices [8, 9]. *Industry:* research-firm rate averages [2–6]. *Measured:* tests run for this paper, with their sample sizes. *Assumed:* modeling choices, all editable. *Simulated:* what the model produces.

Rates. Nimdzi's UK Language Services Pricing Study (2020): £0.079 a word for post-editing "irrespective of the domain or language pair", £0.121 for full translation into Western European languages and £0.134 into Asian ones, at about $1.28 to the pound *(industry [2])*. Post-editing a context-sensitive draft is paid at 85% of the post-editing average, $0.086 a word, at 1,000 words an hour *(assumed)*. Every word sent to the expert is paid at the full-translation average, $0.155 to $0.172, and the expert's hours at $95 to $140 *(assumed)*. Other references:

SOURCE	WHAT IT MEASURED	FIGURE
Nimdzi, 2020 [3]	Post-editing pricing analysis	Discounts of up to 35% to 40% make sense for light post-editing, since it still takes 60% to 85% of the effort
Slator, 2018 [4]	584 price points from 59 providers' US federal rate cards	Full translation averaged $0.215 a word; Japanese $0.251
Slator, 2024 [5]	Survey of language service providers	91% still charge per word; 12% charge by edit distance
CSA Research, 2020 [6]	5,286 linguists	Translation productivity averages 574 words an hour

Cost by line *(simulated, base case)*:

	POST-EDITING, YEAR 1	AUGMENTED, YEAR 1	POST-EDITING, YEAR 2	AUGMENTED, YEAR 2
Linguists or experts, including upkeep	$580,633	$178,808	$550,859	$127,204
Project management	$35,737	$35,737	$35,623	$35,623
Quality and data lead		$91,200		$91,200
Platform subscription (pulsAR, list price)	$39,600	$39,600	$39,600	$39,600
Rework on what the customer finds	$12,354	$4,167	$11,158	$2,707
Customer time (review, questions)	$41,231	$41,231	$40,725	$40,725
Setup	$35,450	$64,550		
Overhead (15% of labor)	$99,626	$56,169	$89,646	$38,510
Total	**$844,631**	**$511,462**	**$767,611**	**$375,569**
MQM penalty points per 1,000 words	1.78	0.71	1.60	0.42
Days per job	2.5	1.8	2.3	1.6

MONTH BY MONTH

Post-editing gets cheaper only as the memory grows. The Augmented Route keeps falling until the learning curve flattens

Monthly cost of the whole program, four languages. Both routes pay the same setup in month one.

$64k post-editing, month 24 VS **$31k** augmented, month 24



Simulated, base case. The fall follows the Edit Distance Report learning curve (41.7% to near 16% edit distance by month ten); the curve between its endpoints is assumed.

What a person reads. The share sent to the expert is set by Smells under the confirm rule "up to Minor": 16 of 84 segments on the test file and 240 of 1,252 in the account's impact preview *(measured)*. It falls toward an 8% floor as the draft improves *(assumed)*. The expert

also reads 5% of the remaining text for context *(assumed; 30% runs as a scenario)*. The Augmented Route costs the same as post-editing in year two when a person reads 51% of new words every month.

THE NUMBER THAT DECIDES THE COST

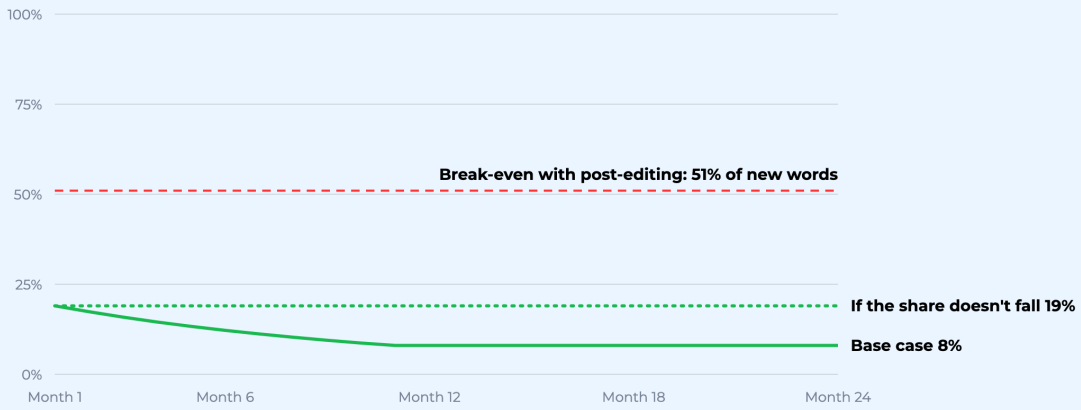
Below 51% of new words reaching a person, the Augmented Route costs less than post-editing

Share of new words Smells sends to the expert each month under the confirm rule "up to Minor". It starts at the 19% measured on the test file and in the account's impact preview, and falls toward 8% as the draft improves.

19% measured share sent to the expert today

VS

51% where post-editing becomes cheaper



Start: measured, 16 of 84 test segments and 240 of 1,252 preview segments sent. Fall and floor: assumed. Break-even: simulated, year two.

How fast the draft improves. Across wxrks accounts, average edit distance (how much of each segment a person changed) went from 41.7% in an account's first month to near 16% by month ten *(sourced [1], per-account means)*. The model uses that curve only for how fast the augmented draft needs less changing. The post-editing draft improves only through the memory.

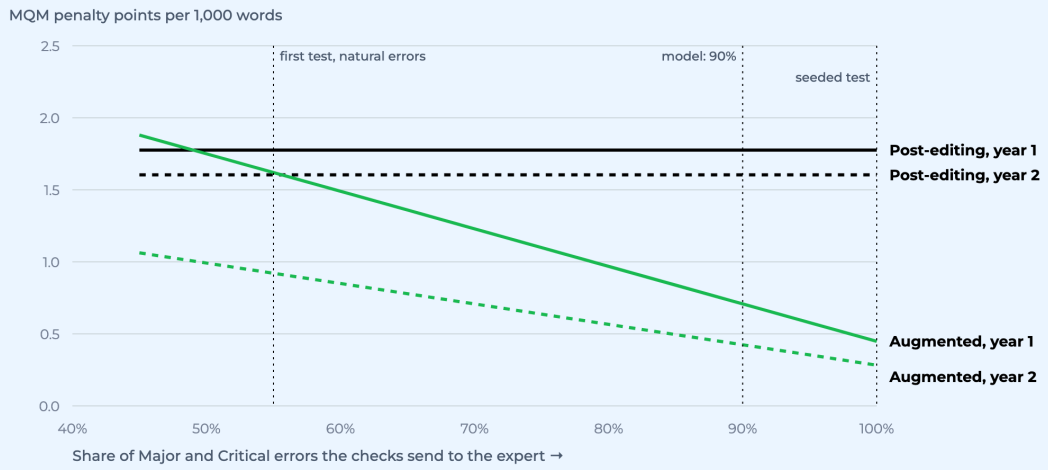
Quality. Errors in the context-sensitive draft are assumed: 0.78 accuracy errors per 1,000 new words, a tenth of them critical, the ratio to plain machine translation taken from the Edit Distance Report (8.1% against 15.5% TER) [1], plus 2 minor issues per 1,000 words and term misses of 12% on undecided terms and 3% on decided ones. Post-editors fix 85% of accuracy errors, 88% of term misses and 70% of minor issues; people introduce 0.3 minor issues and 0.03 major errors per 1,000 words they edit *(all assumed)*. On the Augmented Route the checks send 90% of major and critical errors and 60% of minor ones to the

expert, flag 87.5% of minor ones, and re-check every save *(measured on seeded errors, then discounted)*. In month one the routes score 1.95 and 1.15; with no introduced errors and no checks on the expert's edits, 1.64 and 1.68.

THE NUMBER THAT DECIDES QUALITY

The Augmented Route ships fewer penalty points than post-editing once the checks catch about half the serious errors in year one, and 45% in year two

MQM penalty points per 1,000 words for each route, as the share of Major and Critical errors the checks send to the expert changes. Post-editing does not depend on the checks, so its lines are flat.



Simulated. Measured points: 65 of 65 seeded Major and Critical errors sent to the expert (6 Nov 2026, four languages); 6 and 4 of 9 natural errors sent in the first test (5 Nov 2026, mostly register errors).

The seeded test. 21 correct targets and 21 seeded targets per language, one MQM error each, 13 major or critical and 8 minor; German and French run twice. Results: 65 of 65 major and critical errors sent to the expert, 31 of 40 minor sent, 35 of 40 flagged, 3 of 84 correct segments sent.

The learning test. German segment 12 confirmed a term for "chart of accounts"; the confirmation created a learned glossary term; the second context-sensitive pre-translation rerun used it in segment 13. One term, one language.

Scenarios. Each row changes the assumptions named; everything else stays at base.

TRYING TO BREAK IT

In year two the Augmented Route costs less than post-editing in every scenario but one, and scores better in all of them

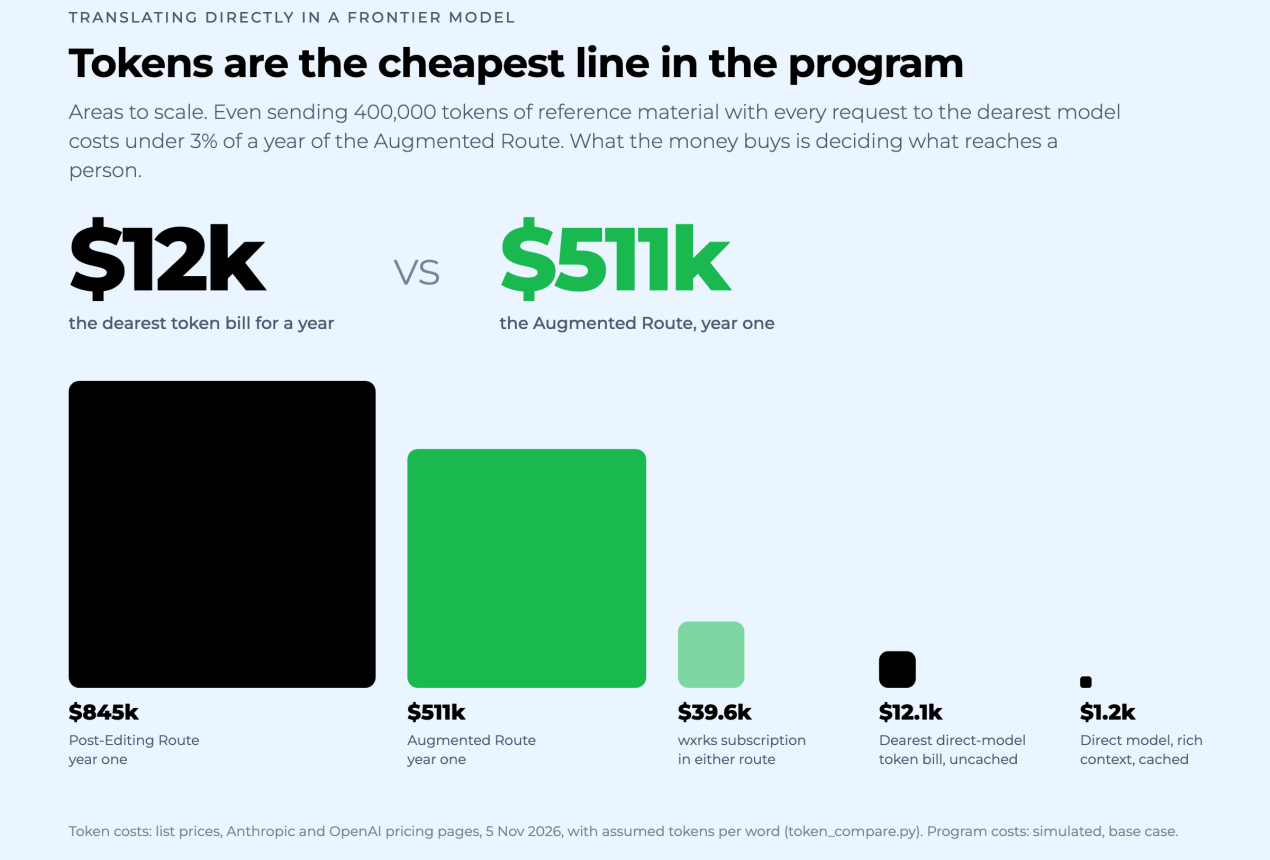
Augmented Route cost as a share of post-editing, year one and year two, and MQM penalty points per 1,000 words in year two. Bold marks the cases where it costs more or scores worse, by any margin.



Simulated. Each row changes the assumptions named; everything else stays at base.

Tokens. The same program sent directly to a frontier model in 1,000-word requests, 1.7 tokens per English word, 2.5 per translated word, 1,000 reasoning tokens per request *(assumed)*. A year of tokens, without caching / with caching, list prices of 5 November 2026 [8, 9]:

MODEL	LEAN CONTEXT (10K TOKENS)	RICH (100K)	EVERYTHING (400K)
Claude Opus 5.5	$841 / $615	$3,433 / $1,177	$12,073 / $3,049
Claude Sonnet 5.5	$420 / $308	$1,716 / $588	$6,036 / $1,524
Claude Haiku 5.5	$21 / $16	$429 / $163	$1,509 / $443
GPT-6.1 Sol	$420 / $303	$1,716 / $538	over the short-context limit
GPT-6 Astra	$2,102 / $1,544	$8,582 / $3,002	over the short-context limit



On the demo account in the test week the platform metered about 87 AI tokens per AI-processed word *(measured: 1,909,158 tokens over 22,003 words, 1 to 5 November 2026, mostly test runs)*. At 87 both routes need a custom tier and the Augmented Route costs 62% of post-editing in year one and 51% in year two *(custom tier price assumed)*.

14

Appendix B. Sources

1. wxrks Research (2026). The Edit Distance Report. Working paper WP-2026-01. https://resources.wxrks.com/papers/context-sensitivity-research/
2. Nimdzi Insights (2020). UK Language Services Pricing Study. https://atc.org.uk/wp-content/uploads/UK-Language-Services-Pricing-Study-2020-by-Nimdzi-Insights.pdf
3. Nimdzi (2020). Machine translation post-editing: how much is the fish? https://www.nimdzi.com/machine-translation-post-editing-how-much-is-the-fish-nimdzi-finger-food/
4. Slator (2018). America's translation rate holds firm at USD 0.22. https://slator.com/americas-translation-rate-holds-firm-at-usd-0-22/
5. Slator (2024). How much does translation cost? https://slator.com/resources/how-much-does-translation-cost/
6. CSA Research (2020). The State of the Linguist Supply Chain. https://cdn2.hubspot.net/hubfs/4041721/Newsletter/The%20State%20of%20the%20Linguist%20Supply%20Chain%202020.pdf
7. MQM Council. The MQM scoring models. https://themqm.org/error-types-2/the-mqm-scoring-models/
8. Anthropic. Pricing. https://platform.claude.com/docs/en/about-claude/pricing
9. OpenAI. API pricing. https://developers.openai.com/api/docs/pricing