A $34 fine-tune brings a 27B model to frontier-level performance on Houdini repair benchmarks

Dmytro Rohovyk, Founder Pre-seed · August 2026

The problem

VFX studios under client security terms often cannot send scene files to a third-party API. This measures what a studio gives up by running a specialised model on its own hardware instead. Six models, identical conditions, three worked examples each, and code compiled in a live Houdini session.

Fine-tuning setup

Qwen/Qwen3.8-27B

Base model, recent release from Alibaba Cloud.

QLoRA r=32, 1 epoch, 17.2M tokens

Training setup to fit into limited production budget

$34

7.1h on one NVIDIA A100-large via Hugging Face

~$80

Total including evaluation of tuned and base models

HEADLINE

Across four measured capabilities, the tuned 27B leads on isolated VEX compilation, shows no statistically detectable difference from the strongest tested frontier model on graph repair and diagnosis at the current sample size, and leads every tested untuned open-weight model across all four metrics.

HEADLINE

Model	VEX Compilation	Graph repair	Minimal Patch	Diagnosis (silent)
Qwen 3.8 27B TUNED	**89.6%**	**72.0%**	**72.0%**	**54.5%**
Claude Opus 5	**76.4%**	**67.4%**	**64.1%**	**52.2%**
Claude Sonnet 5	63.8%	50.0%	58.0%	40.0%
Gemma 4 31B	62.5%	40.0%	41.0%	14.3%
Llama 3.3 70B	53.2%	40.0%	38.0%	13.3%
Qwen 3.8 Base (untuned)	47.9%	46.0%	38.0%	36.4%

What can and cannot be claimed

Can

leads on isolated VEX compilation; shows no statistically detectable difference from Opus 5 on graph repair and diagnosis at this sample size; leads every tested untuned open-weight model on every reported cross-model metric.

Cannot

Across the ten paired comparisons, eight reach raw $p < 0.05$ and seven remain significant after Holm correction. The only two comparisons that are not significant are both against Opus 5. On the paired graph-repair subset, Opus scores marginally higher. The defensible conclusion is that no statistically detectable difference from Opus 5 was found at this sample size — not that equivalence or superiority has been established.

A studio working on unreleased content is frequently prohibited by its client's security terms from sending scene files to a third-party API. For that segment the frontier models are not a more expensive option — they are not an eligible one.

The question a technical director actually faces is therefore not "which model is best" but "how much capability do I forfeit by keeping the work inside the building?" On these four measures, the answer is: on isolated VEX compilation, none — the on-premise model leads. On graph repair and diagnosis, no statistically detectable capability gap was observed at this sample size.

Secondary consequences, not measured here but structural: inference runs on hardware the studio already owns, so marginal cost per task is zero and no per-token bill scales with usage; and a 27B at 4-bit fits a single workstation GPU, so deployment does not require new infrastructure.

Why parity is the commercially interesting result

Every number here **comes from a public forum corpus**. The commercial claim is that the same pipeline, pointed at a studio's own HDAs, naming conventions and shot history, produces the same lift. That is untested. What is demonstrated is that the pipeline transfers across base models; what is assumed is that it transfers across corpora. Those are different claims, and only the first has evidence behind it.

The open question this report does not answer

Code that runs
Comparison

Every generated VEX snippet was compiled and cooked inside a live Houdini 22 session. Either it executes or it does not.

Model	Snippets	Compiled	Rate
Qwen 3.8 27B TUNED (our model)	48	43	89.6%
reference answers (ceiling)	96	80	83.3%
Claude Opus 5	89	68	76.4%
Claude Sonnet 5	47	30	58.0%
Gemma 4 31B	48	30	62.5%
Llama 3.3 70B	47	25	53.2%
Qwen 3.8 Base (untuned)	48	23	47.9%

Code that runs
Conclusion

The tuned model exceeds the reference ceiling. The **ground-truth answers** themselves **compile only 83.3%** of the time in this harness.

Reference snippets come from real scenes and often depend on attributes an isolated wrangle cannot supply, whereas the tuned model writes self-contained code. It should be read as "at or above the practical maximum of this isolated compilation harness," not as a claim of superhuman VEX code output.

Snippets were compiled **in a SOP attribwrangle regardless of original context**, so POP- and volume-specific builtins fail here. Both absolute rates and the ceiling are therefore conservative.

Evaluation by how much the prompt gives away

Comparison

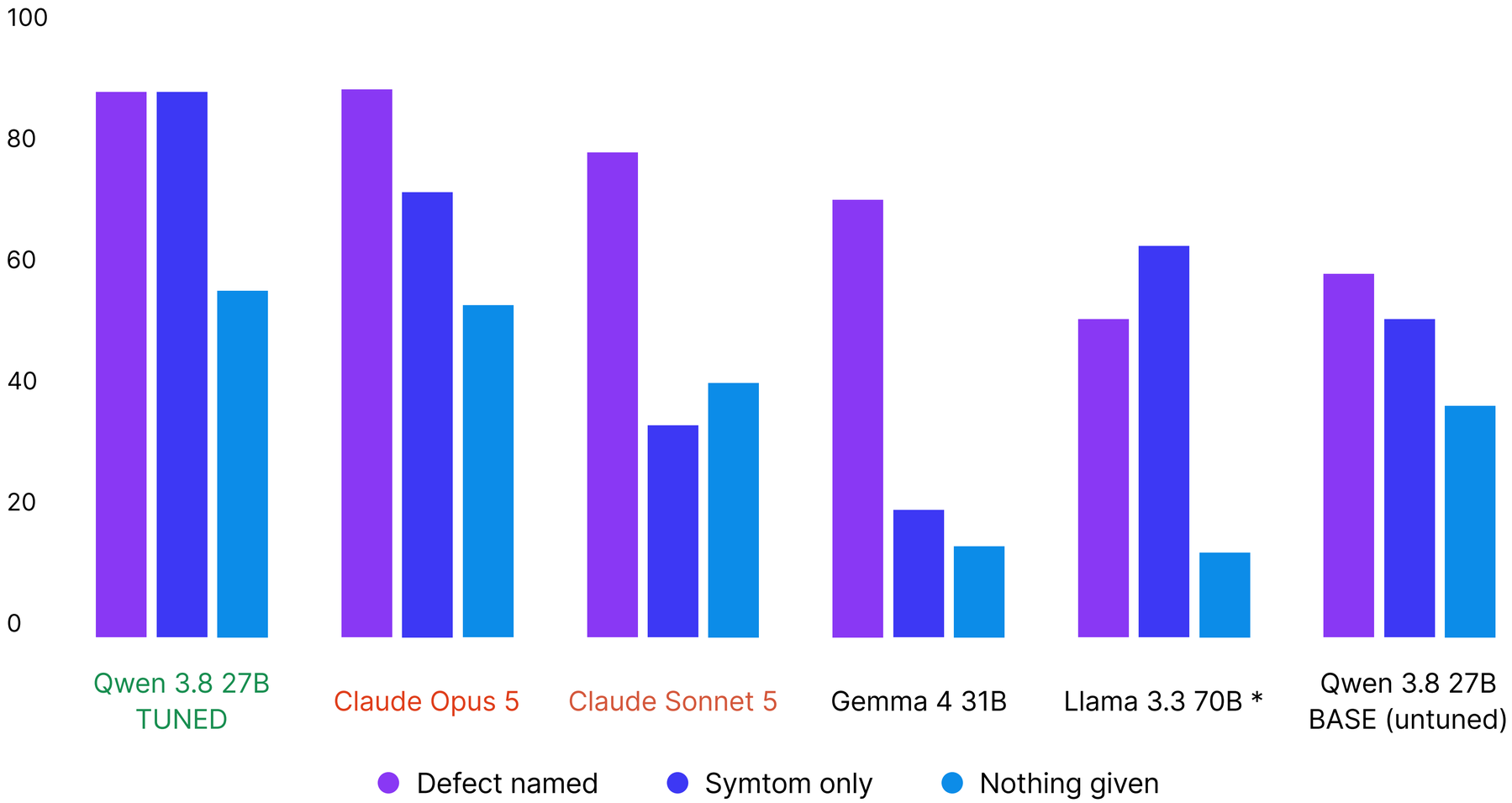
Every repair prompt presents a broken node graph. The framing is different - some name the defect outright, some describe only a visible symptom, and some say nothing beyond "this isn't working." Only the last two test diagnosis.

Model	Defect named	Symptom only	Nothing given
Qwen 3.8 27B TUNED (our model)	85.7% (18/21)	**85.7% (12/14)**	**54.5% (6/11)**
Claude Opus 5	**86.1% (31/36)**	**70.0% (21/30)**	**52.2% (12/23)**
Claude Sonnet 5	76.2% (16/21)	33.3% (5/15)	40.0% (4/10)
Gemma 4 31B	68.8% (11/16)	20.0% (2/10)	14.3% (1/7)
Llama 3.3 70B	50.0% (10/20)	61.5% (8/13)	13.3% (2/15)
Qwen 3.8 Base (untuned)	57.1% (12/21)	50.0% (7/14)	36.4% (4/11)

Evaluation by how much the prompt gives away

Comparison

Visualization for better comparison



*likely sample-size noise, see Limits

Evaluation by how much the prompt gives away
Conclusions

The open-weight models collapse when information is withheld.

Gemma from 64.7% to 14.3%, Llama from 47.6% to 13.3%. The tuned model and Opus 5 both hold up, and are level with each other.

This improvement **is consistent with the intended effect of a pipeline change**. The corruption generator was rebalanced from 91% explicitly stated defects to 35% stated / 35% symptom-only / 30% silent. The resulting model performs substantially better on diagnosis-style prompts, although this research does not isolate the rebalancing as the sole causal factor.

Important note*

An initial zero-shot comparison showed the tuned model far ahead of every other model. That result was largely an **artefact**: comparators had simply **never seen the output format**. while the tuned model **learned it from thousands of training examples**.

Giving the model 3 successfull answers before running the test helped it adapt to the domain specific approaches the test determines.

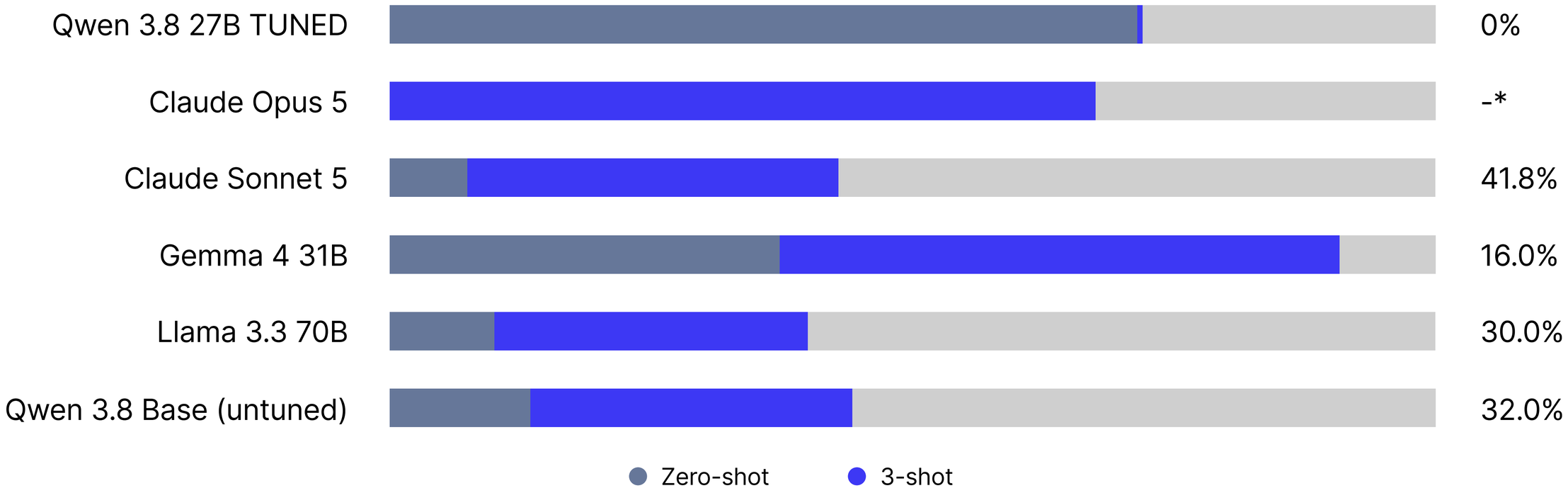
Introduction to

The measurement that changed the result

The measurement that changed the result

Comparison

Re-running every model with three worked examples closed most of the gap.



*- Claude Opus 5 didn’t have zero-shot test due to feature being implemented during evaluation process of Opus

The measurement that changed the result

Opus 5 has no zero-shot figure because it was never run that way. It entered the comparison after few-shot had been adopted as the fair condition, so it was only ever measured with examples. Reporting a zero-shot Opus number would require a run that was not performed.

The tuned model is the control: its **score is identical at 72.0%** in both conditions, because it learned the format during training and gains nothing from being shown it again. Every other model **moves by 16 to 41 points**. Format effect is what made the isolated difference.

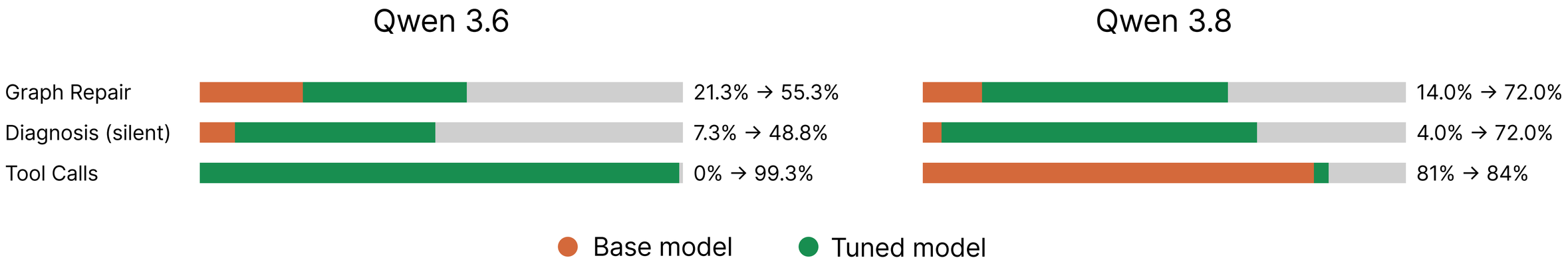
Roughly two-thirds of the apparent advantage in the original zero-shot comparison **was format familiarity rather than Houdini context knowledge**. All headline figures in this report use the 3-shot condition, which is the fair one and deliberately favours the comparators. The remaining difference is what survives that correction.

Portability across two Qwen generations

The same pipeline applied unchanged to Qwen 3.6 and Qwen 3.8

The pipeline was applied unchanged to Qwen 3.6 and Qwen 3.8, released months apart. 3 days after completing Qwen 3.6 training, Alibaba released their latest Qwen 3.8 as open weights model, available on HuggingFace. Better agentic workflows tool calling and stronger core model out of pocket.

Evaluated on different corpus versions, so deltas rather than absolutes are the comparable quantity.



Portability: the same pipeline on two base models

Conclusions

Two conclusions. The pipeline transfers without modification between two generations of the Qwen family, **producing large gains on both**.

And tool-calling is no longer something this pipeline needs to teach — Qwen 3.6 base could not emit a valid tool call at all, while Qwen 3.8 base does so 81% of the time unaided. The pipeline's value has now moved from teaching tool syntax and how to use it directly to teaching diagnosis and fixing things.

Portability: the same pipeline on two base models

Conclusions

So, does that mean base models will absorb the rest?

Partly, and the direction should be stated plainly rather than argued away. Generic capabilities — instruction following, tool syntax, general code. All are being absorbed by each base generation, and anything built on those has a short half-life.

What is not converging is the part that depends on data the model cannot have seen: a specific studio's node conventions, its HDAs, the failure modes of its own pipelines. A frontier model trained on the public internet has no access to that, and no amount of scale supplies it. The defensible asset is therefore not the fine-tune itself but the process that converts a studio's private corpus into one — and, per the caveat above, that is the part still to be demonstrated.

A practical read of the same evidence: because each new base arrives more capable, the cost of re-running this pipeline falls over time while the private-data advantage does not. Qwen 3.8 was released, tuned, and **independently measured against five models for under $100.**

Limits

were faced during the process

- **Small samples.** 50 records per form per model (35 for Gemma, which was cut short). 95% confidence intervals run ±12–16 points, so only large differences are resolved. Most pairwise gaps against Opus 5 are not statistically significant.
- **Record sets diverged.** Each model's tokenizer filtered the eval set differently before sampling, so models did not all see identical records. Comparisons against the tuned model are computed pairwise on shared records (n = 24–47) to correct for this. Headline aggregate rates should therefore be read descriptively; paired shared-record comparisons are the stronger basis for cross-model claims and stats.
- **Claude arms use default sampling.** temperature is deprecated on Claude 5, while all self-hosted arms ran greedy decoding. Claude's figures carry sampling variance the others do not. Inference conditions therefore could not be made perfectly equivalent across providers.
- **Half the repair data is synthetic corruption.** Half the repair data is synthetic corruption. The model may therefore have partly learned to invert patterns produced by a known corruption generator rather than to diagnose arbitrary production faults. Generalisation to genuinely broken production scenes remains untested and is the most important open question.
- **Prose was not evaluated.** Open-ended Houdini Q&A requires a human or model judge and was excluded. Notably the tuned model's prose answers are far terser than the base model's (30 vs 627 median words) — an unmeasured behavioural change.
- **Tool calling is not cross-model comparable.** The scorer keys on Qwen tool-call syntax; other models would be unfairly zeroed for using their own correct format.

Every metric collected, both prompting conditions, with sample sizes. The terminology below is used throughout this report and in every chart.

The five task types ("forms")

The dataset is built from public Houdini forum threads and the scene files attached to them. Each thread yields different kinds of training example, labelled A through T. Held-out evaluation records: 2,221 total, drawn from threads that appear nowhere in the 35,010-record training set.

Introduction to

Complete measurements

Complete measurements

Labeling

Form	The model is asked to…	Answer looks like	train	eval	Scored?
A — graph repair	find the fault in a broken node network and return the whole corrected graph	diagnosis sentence + full 40–400 line scene listing	3,909	300	automatic
D — minimal patch	same fault, but return only the lines that change	-exportcontext="cop2_BAD"	5,152	287	automatic
C — code generation	write the VEX or Python for a described task	vector up = set(0,1,0); …	5,424	350	compiled in Houdini
T — tool trajectory	operate a live Houdini session through MCP tool calls	<function=get_network_overview>	4,148	431	Qwen-family only
E / F — forum prose	answer an open Houdini question in plain language				

Complete measurements

Labeling

- **A and D are the same faults in two output formats.** A tests diagnosis plus faithful transcription; D tests diagnosis alone. D exists because transcription, not fault-finding, was the measured bottleneck — and because a real integration applies a patch rather than rewriting the artist's scene.
- **Forms T and E/F are excluded from cross-model comparison**. The tool-call scorer keys on Qwen's syntax and would unfairly zero other models for using their own correct format, and prose needs a human or model judge that was never run.

Complete measurements

How much the prompt gives away — "disclosure"

Every repair prompt (forms A and D) presents a broken graph, but varies in how much it reveals. This split is what separates diagnosis from targeted editing.

- **stated** — the defect is named outright: "fuse13 is bypassed so its effect is missing." Tests targeted repair only.
- **vague** — only an observable consequence: "one step in the chain seems to have no effect any more." Tests diagnosis.
- **silent** — nothing but "This Houdini setup isn't working." Tests diagnosis with no assistance at all.

Complete measurements

Graph repair (form A)

Model	Condition	N	Fixed	95% CI	Exact	No block	Graph not reproduced
Qwen 3.8 TUNED	3-shot	50	72.0%	58–83	72.0%	0.0%	14.0%
Claude Opus 5	3-shot	95	67.4%	57–76	62.1%	0.0%	18.9%
Claude Sonnet 5	3-shot	50	50.0%	37–63	42.0%	0.0%	20.0%
Qwen 3.8 base	3-shot	50	46.0%	33–60	40.0%	2.0%	18.0%
Gemma 4 31B	3-shot	35	40.0%	26–56	37.1%	0.0%	5.7%
Llama 3.3 70B	3-shot	50	40.0%	28–54	30.0%	0.0%	20.0%
Qwen 3.8 TUNED	0-shot	100	72.0%	63–80	71.0%	0.0%	18.0%
Gemma 4 31B	0-shot	100	24.0%	17–33	12.0%	5.0%	68.0%
Qwen 3.8 base	0-shot	100	14.0%	9–22	6.0%	27.0%	74.0%
Llama 3.3 70B	0-shot	100	10.0%	6–17	4.0%	16.0%	79.0%
Claude Sonnet 5	0-shot	91	8.8%	5–16	2.2%	0.0%	85.7%

Complete measurements

Code generation (form C) — format and execution

Model	Cond.	n	fenced block	tagged vex	code extractable	compiles	95% CI
reference (ceiling)	—	96	—	—	96/96	83.3%	74–89
Qwen 3.8 TUNED	3-shot	50	100.0%	100.0%	48/48	89.6%	78–95
Claude Opus 5	3-shot	93	98.9%	96.8%	89/90	76.4%	66–84
Claude Sonnet 5	3-shot	50	100.0%	96.0%	47/47	63.8%	49–76
Gemma 4 31B	3-shot	50	100.0%	88.0%	48/48	62.5%	48–75
Llama 3.3 70B	3-shot	50	98.0%	96.0%	47/48	53.2%	39–67
Qwen 3.8 base	3-shot	50	100.0%	100.0%	48/48	47.9%	34–62
Qwen 3.8 TUNED	0-shot	100	100.0%	99.0%	96/96	79.2%	70–86
Gemma 4 31B	0-shot	100	98.0%	0.0%	94/96	56.2%	46–66
Claude Sonnet 5	0-shot	97	95.9%	0.0%	90/94	47.8%	38–58
Llama 3.3 70B	0-shot	100	95.0%	80.0%	91/96	27.1%	19–37
Qwen 3.8 base	0-shot	100	72.0%	65.0%	68/96	18.8%	12–28

Complete measurements

Minimal-edit patch (form D)

Model	Cond.	n	correct	95% CI	exact	dumped graph	no block
Qwen 3.8 TUNED	3-shot	50	72.0%	58–83	68.0%	0.0%	0.0%
Claude Opus 5	3-shot	92	64.1%	54–73	59.8%	1.1%	1.1%
Claude Sonnet 5	3-shot	50	58.0%	44–71	50.0%	0.0%	0.0%
Gemma 4 31B	3-shot	39	41.0%	27–57	35.9%	0.0%	0.0%
Qwen 3.8 base	3-shot	50	38.0%	26–52	34.0%	2.0%	2.0%
Llama 3.3 70B	3-shot	50	38.0%	26–52	30.0%	0.0%	0.0%
Qwen 3.8 TUNED	0-shot	100	82.0%	73–88	79.0%	0.0%	0.0%
Claude Sonnet 5	0-shot	96	61.5%	51–71	49.0%	0.0%	0.0%
Gemma 4 31B	0-shot	100	47.0%	38–57	36.0%	0.0%	0.0%
Qwen 3.8 base	0-shot	100	46.0%	37–56	42.0%	0.0%	39.0%
Llama 3.3 70B	0-shot	100	25.0%	18–34	19.0%	2.0%	19.0%

The two zero-shot 0.0% entries under "tagged vex" are a formatting artefact and not a capability gap at all, where Gemma and Sonnet emitted fenced blocks but did not label them VEX. Their code still compiled 56.2% and 47.8% of the time. Compilation is treated here as an execution metric, not a complete measure of semantic task correctness.

Under few-shot the column ceases to discriminate at all (88–100% for every model), which is why no claim in this report rests on it.

code extractable counts responses from which any code block could be recovered at all. Zero-shot Qwen 3.8 base failed on 28 of 96; every 3-shot arm is near-perfect.

Conclusion

Complete measurements

Scored only on records both models actually saw, correcting for the record-set divergence. Both models answer the same records, because this is paired data.

The approach chosen to utilize McNemar's exact test — a non-parametric statistical method used for paired nominal (categorical) data with binary outcomes.

It examines only the records where the two models disagree. b = tuned correct where the competitor failed; c = the reverse.

Because ten pairwise hypotheses are tested, Holm correction is also reported to control family-wise error.

Introduction

Paired comparisons tuned model vs each competitor

Paired comparisons

Visualization

Comparator	Task	n	TUNED	competitor	b	c	p	raw p	Holm-significant	Verdict
Qwen 3.8 base	minimal patch	50	72.0%	38.0%	18	1	0.0001	0.0001	Yes	tuned wins
Qwen 3.8 base	graph repair	50	72.0%	46.0%	13	0	0.0002	0.0002	Yes	tuned wins
Gemma 4 31B	minimal patch	35	74.3%	40.0%	12	0	0.0005	0.0005	Yes	tuned wins
Llama 3.3 70B	graph repair	26	80.8%	42.3%	10	0	0.0020	0.0020	Yes	tuned wins
Gemma 4 31B	graph repair	24	83.3%	45.8%	9	0	0.0039	0.0039	Yes	tuned wins
Claude Sonnet 5	minimal patch	28	82.1%	53.6%	8	0	0.0078	0.0078	Yes	tuned wins
Llama 3.3 70B	minimal patch	26	73.1%	38.5%	10	1	0.0117	0.0117	Yes	tuned wins
Claude Sonnet 5	graph repair	25	72.0%	48.0%	6	0	0.0312	0.0312	No	not significant after correction
Claude Opus 5	minimal patch	45	71.1%	57.8%	12	6	0.238	0.238	No	not significant
Claude Opus 5	graph repair	47	70.2%	72.3%	2	3	1.000	1.000	No	not significant

Eight of ten paired comparisons reach raw $p < 0.05$, and seven remain significant after Holm correction for multiple comparisons. The two raw comparisons that are not significant are both against Opus 5; neither Opus comparison approaches significance after correction. The tuned model is measurably better than the tested base and smaller comparator models on most paired tasks, while its difference from Opus 5 remains statistically unresolved at this sample size.

The b and c columns show why. Against Gemma on graph repair, there are 9 records the tuned model fixed and Gemma did not, and zero the other way that is a one-sided result that a 24-record sample resolves easily. Against Opus the same columns read 2 and 3: the two models succeed and fail on almost exactly the same records.

Conclusion

Paired comparisons tuned model vs each competitor

Qwen-family only

tool calls and prose

Model	Tool call emitted	Tool name correct	Prose E	Prose F
Qwen 3.8 tuned	98.0%	84.0%	30 words	44 words
Qwen 3.8 base	100.0%	81.0%	627 words	592 words
Qwen 3.6 tuned	—	99.3%	—	—
Qwen 3.6 base	—	0.0%	—	—

Tuning reduced tool-name accuracy relative to the 3.6 generation (84% vs 99.3%) because the 3.8 base already arrives competent and the training signal adds little. The tuned model's prose is roughly 20× shorter than the base model's. It is a substantial behavioural shift whose effect on correctness, completeness and user preference was not evaluated here.

Cost

Item	Detail	Cost
Training	1 epoch, 9,500 records, 17.2M tokens, 7.1h on one A100 Large	$34
Evaluation	six model arms, two conditions, live Houdini validation	~$46
Total cloud spend	self-funded, single developer	~$80

Dataset preparation, pipeline repair and all scoring ran on local hardware at no cost.

One full iteration that includes data build, training, and a six-model evaluation costs under $100. Roughly $15 of the evaluation spend was lost to configuration errors found and fixed during the run.

What would resolve the open questions

- Corpus portability. Run the pipeline end-to-end on a partner studio's own scene archive and measure the same four metrics. This is the load-bearing experiment; nothing else in the roadmap matters as much. Needs a design partner, not more compute.
- Statistical resolution. Sample sizes of 24–50 leave every frontier comparison unresolved. Running the full 2,221-record held-out set across all six models would settle them and costs roughly $400 of GPU time — trivial relative to the decision it informs.
- Generalisation beyond synthetic faults. Half the repair training data is script-generated corruption. Testing against genuinely broken production scenes would show whether the model diagnoses faults or inverts a known generator.
- Open-ended Houdini questions. Prose was never scored. The tuned model's answers are 20× shorter than the base model's, which may be better or worse for an artist; it is currently unknown.

The evidence here establishes that the method works on one corpus, transfers across base models, and **produces results statistically unresolved from the strongest tested frontier model** on selected measurable Houdini tasks, while outperforming the tested untuned open-weight models on most paired comparisons. It does not yet establish that it works on the data a customer would actually pay to have trained on.

Post postscript

Base Qwen/Qwen3.8-27B · adapter houdini-qwen38-27b-lora · dataset 9d77f4e3c905d4b5 · 35,010 train / 2,221 held-out records, budget capped training on ~10k records + eval savings, zero thread-level overlap between splits. All evaluation inputs, selected records, raw generations and scoring outputs are retained. Deterministic parts of the pipeline are reproducible from a fixed seed. Few-shot exemplars drawn from the training split only. Zero of them appear in the evaluation set.