Supplementary Information

for

Toward Self-Improving MEMS Research Agents through Experience-Informed Procedure Refinement

Xinrong Ouyang[1,†], Ruiqi Guo[1,†,*], Zishen Dai[1], Jiawei Xu[1], Pan Zhang[2,*], Yuan Wang[1,*], Rui Paulo Martins[1], Pui-In Mak[1]

[1] Institute of Microelectronics, State Key Laboratory of Analog and Mixed-Signal VLSI, University of Macau, Macau, China.

[2] School of Integrated Circuits, Peking University, Beijing, China.

† These authors contributed equally to this work: Xinrong Ouyang and Ruiqi Guo.

*Correspondence: Ruiqi Guo, ruiqiguo@um.edu.mo; Pan Zhang, panzhang1989@pku.edu.cn; Yuan Wang, yuanwang@um.edu.mo

The 25 registered runs in Fig. 4 supply the comparative outcome evidence. The final author assessment gives success counts across the five configurations of 1, 1, 4, 5, and 5 out of five per configuration. Results of the initial model-judge assessment are retained for provenance and sensitivity in Table S7 and Note S3. The author interpretation was set after neutral-code summaries and frozen before unblinding; final counts include the disclosed later correction to Run 20261023-022. Early studies document how methods were developed and used; they are not pooled with the registered runs or used to estimate a procedure effect.

The configurations, in order, are Without MEMS Research Agent Framework (MRAF); Stage 1: Solve; Stage 2: Design; Stage 3: Verify; and With MRAF. References use the main bibliography. Unless specified, values are at 293.15 K and Q_{TED} includes thermoelastic loss only.

Contents

1. Supplementary Table S1. Source publications and author-assigned scope for Fig. 1b.
2. Supplementary Note S1. The early guided device study and its supporting calculations
3. Supplementary Fig. S1. Organization and execution of a MEMS research task.
4. Supplementary Fig. S2. Baseline geometry, operating mode, and thermoelastic-loss distribution in the early study.
5. Supplementary Table S2. Geometry of the baseline, *c01*, candidate 2 of the device study, *c02* (the two-row slot design), and the two-slot layout of reference[15], measured from the STEP files.
6. Supplementary Fig. S3. Candidates of the early device study.

7. Supplementary Table S3. Discretization and stored solutions in the six calculations made after the run.
8. Supplementary Table S4. Operating mode of the baseline and *c01* on three meshes, identified from displacement fields.
9. Supplementary Table S5. Development periods and the ten early research runs.
10. Supplementary Fig. S4. The frequency-response revision and its later use.
11. Supplementary Fig. S5. Earlier procedure revisions and subsequent use.
12. Supplementary Fig. S6. Eight developmental rounds of one task.
13. Supplementary Note S2. Method revision across eight developmental rounds
14. Supplementary Fig. S7. Saved modes in the Round 6 and Round 7 development records.
15. Supplementary Table S6. Process records of the eight rounds of one task: problems recorded in the reviews, final check, completion score, time to the first valid baseline, and scripts written and reused.
16. Supplementary Fig. S8. Timing and reported designs of all 25 registered runs.
17. Supplementary Table S7. The 25 registered runs: configuration, host, language model, procedures, outcome under each registered condition and under each reading, judges, and recomputed values of the reported design.
18. Supplementary Fig. S9. Cumulative outcome conditions in the 25 registered runs.
19. Supplementary Note S3. Registered protocol, assessment history, and results
20. Supplementary Fig. S10. Descriptive measures of the 25 registered runs.
21. Supplementary Table S8. Principal descriptive measures for all 25 registered runs.
22. Supplementary Table S9. Frozen procedures and the checks they supply.
23. Supplementary Fig. S11. Earlier paired runs retained as development records.
24. Supplementary Table S10. Earlier paired runs and archival outcome counts.
25. Supplementary Note S4. Figure provenance and scope

Supplementary Table S1 | Source publications and author-assigned scope for Fig. 1b.

Levels describe task scope, not scientific quality, autonomy in all settings, or comparative performance. They are author assignments adapted from reference[3]; the cited works are contextual examples rather than a systematic survey.

(a) Scope definitions

Level	Scope of the reported task	Human contribution
L1	One prediction or screening task	Question, inputs, and subsequent study
L2	A defined multistep workflow	Objective, workflow, and design space
L3	Adaptive coordination of research steps within a supplied problem	Objectives and specified review or experimental steps
L4	A complete bounded task from hypothesis to analysis and manuscript	Topic, data, and initial scaffold

L2.5 denotes adaptive search or tool use within a supplied objective. L3.5 is an author-defined operational category: agents coordinate research steps, with defined human contributions during or between cycles. In this work, agents execute the research cycle within each task, while humans set the objectives and approve procedure revisions between runs. The levels are qualitative categories, not a calibrated scale. Initial objectives and constraints bound the setting; they do not alone lower a level. When a publication offers both autonomous and assisted modes, the autonomous mode supplies the plotted scope.

(b) Publications

Study (label in Fig. 1b)	Ref.	Year	Field	Type	Level	Reason for the level
Mobile chemist	40	2020	Chemistry	Other	L3	Adaptive experiments in a predefined chemistry space.
AlphaFold	41	2021	Structural biology	Other	L1	Protein-structure prediction from a supplied sequence.
Freeform MEMS	11	2022	MEMS	MEMS	L1	Modal-frequency and anchor-loss prediction for supplied geometries.
Spiderweb optimization	13	2022*	Nanomechanical resonators	MEMS	L2.5	Bayesian design search; fabrication and measurement remain human.
DRL MEMS	31	2023	MEMS	MEMS	L2.5†	Adaptive structural search within specified MEMS objectives.
Coscientist	5	2023	Chemistry	Other	L3.5	Literature, code, and robotic experiments within supplied objectives.
DRG topology	12	2024	MEMS	MEMS	L2.5	Surrogate-guided topology search within a supplied design space.
Anchor-loss ML	44	2025	MEMS	MEMS	L2†	Forward/inverse design with finite-element verification.
ChemAgent	16	2025	Chemistry	Other	L2.5	Experience retrieval and revision for bounded chemical reasoning.
SciAgents	42	2025*	Materials science	Other	L3	Materials hypotheses and plans, without new measurements.
Virtual Lab	23	2025	Biomedicine	Other	L3	Agent-designed nanobody workflow with human guidance and laboratory testing.
AI Scientist-v2	43	2025§	Machine learning	Other	L4	Hypothesis-to-manuscript automation in scoped computational research.
data-to-paper	2	2025*	Data-driven research	Other	L4	Automatic data-to-manuscript mode; complex datasets use human copiloting.
High-DoF MEMS	14	2026	MEMS	MEMS	L2.5	Learned prediction and particle-swarm search with experimental comparison.
MOOSEnger	37	2026§	Multiphysics simulation	Other	L2.5	Adaptive preparation and repair of multiphysics input files.
SciToolAgent-Evo	38	2026§	Scientific tool use	Other	L2.5	Acquisition of tool capability within supplied evaluation tasks.
SciVisAgentSkills	39	2026§	Scientific data analysis and visualization	Other	L2.5	Reusable analysis and visualization procedures, not whole-study autonomy.

Study (label in Fig. 1b)	Ref.	Year	Field	Type	Level	Reason for the level
SkillPCF	17	2026	Photonics	Other	L2.5	Simulator-guided revision of skills within photonic inverse design.
AIDE[2]	18	2026§	Machine learning	Other	L3	Evaluated revisions of research-agent code.
Robin	32	2026	Biomedicine	Other	L3.5	Literature, experiments, and analysis; humans perform and transfer experiments.
This work	—	2026	MEMS	Star	L3.5‡	Agent execution of the research cycle within a task; human objectives and approval of procedure revisions between runs.
The AI Scientist	1	2026	Machine learning	Other	L4	Idea-to-manuscript automation within supplied computational scaffolds.

Ref. numbers follow the main bibliography. Years follow the cited edition or preprint submission. Square, MEMS; circle, other research; star, this work. Same-year, same-level studies share a marker; an enclosing ring indicates mixed fields. The offset between Robin and this work only separates labels. Abbreviations in the study labels: DRL, deep reinforcement learning; DRG, disk resonator gyroscope; DoF, degrees of freedom; ML, machine learning.

* Year of the journal issue; the article appeared online in the preceding year. † Provisional because accessible workflow details are limited. ‡ Author-defined operational category; see the definition after part a. § Preprint submission year.

Supplementary Note S1. The early guided device study and its supporting calculations

S1.1 Task and run boundaries

Run 20260814-012 used the August procedures and is the guided development study of Fig. 2. The researcher supplied the baseline, fixed dimensions and anchors, and guidance to place one relief segment in each dominant loss band. The researchers knew the published slotted layout; the task had no held-out restriction. After an endpoint rejected the requested model, the researcher switched to DeepSeek-V4-Flash and said "go on." No further human message arrived before the report. The study therefore documents autonomous execution after guidance, rather than prior-blind discovery under the later registration.

The event record begins with baseline preflight at 05:54:49 UTC on 15 August. The Fig. 2a axis starts 5.3 min earlier, at the first reply after "go on," and ends 224 min later. The run verified the baseline, searched the literature, defined six candidates, calculated them, inspected saved fields, and ranked five. Candidate 2 was withheld because its operating mode was not identified at the required 60 K temperature rise. The run froze the candidate set five times after geometry checks; one revision increased its removed-volume cap from 0.10 to 0.35. Thus the sequence is not an immutable preregistered candidate comparison.

After freezing and solving *c01*, the agent read records of Runs 20260814-010 and -011 and copied seven checking/export scripts. Printed reads contained no earlier frequency or Q values, but folder names exposed earlier candidate families. This access was permitted by its task, although the contemporary entry procedure discouraged exposing prior outputs; it would void a registered run. After unblinding, the author initially chose *c01* instead of the registered main case for Fig. 2. Here it remains an early study; Run 20261023-012 remains the case selected by the registered rule.

S1.2 Fields, meshes, and later calculations

The run identified *c01*'s operating mode from saved displacements. Its field export at 60 K used a neighboring in-phase solution, although the ranking used the correct operating-mode eigenvalue. On 30 September the authors corrected that export from the saved model without solving; the corrected mode's displacement correlation with the 293.15 K operating mode was 0.9986, versus 0.0007 for the erroneous export. The original export is retained.

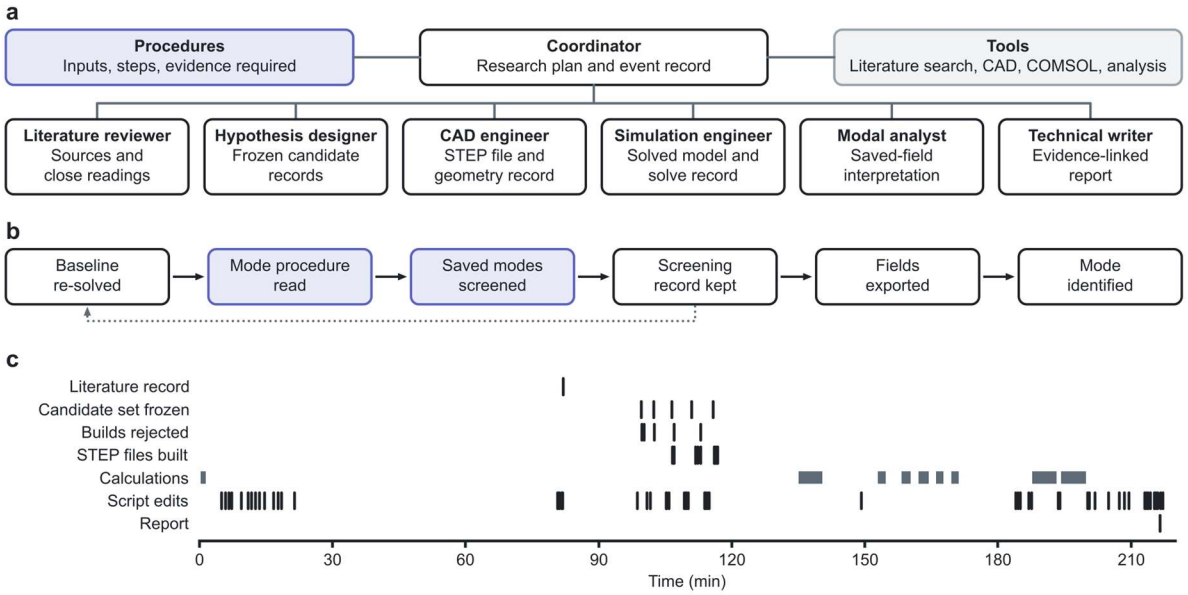
The run refined *c01* alone from 3844 to 5438 elements; Q_{TED} changed by about 0.2%. Author calculations after the run refined both structures at Normal, Fine, and Finer (Tables S3–S4). The room-temperature ratio decreased from 5.159 to 5.046; this sensitivity study is not a rigorous convergence error bound. Fig. 2c combines baseline driven responses calculated in later Run 20260906-001 and *c01* responses calculated by the authors in Run 20261002-001. These supporting calculations were not part of the original autonomous run.

S1.3 Procedure revision and subsequent use

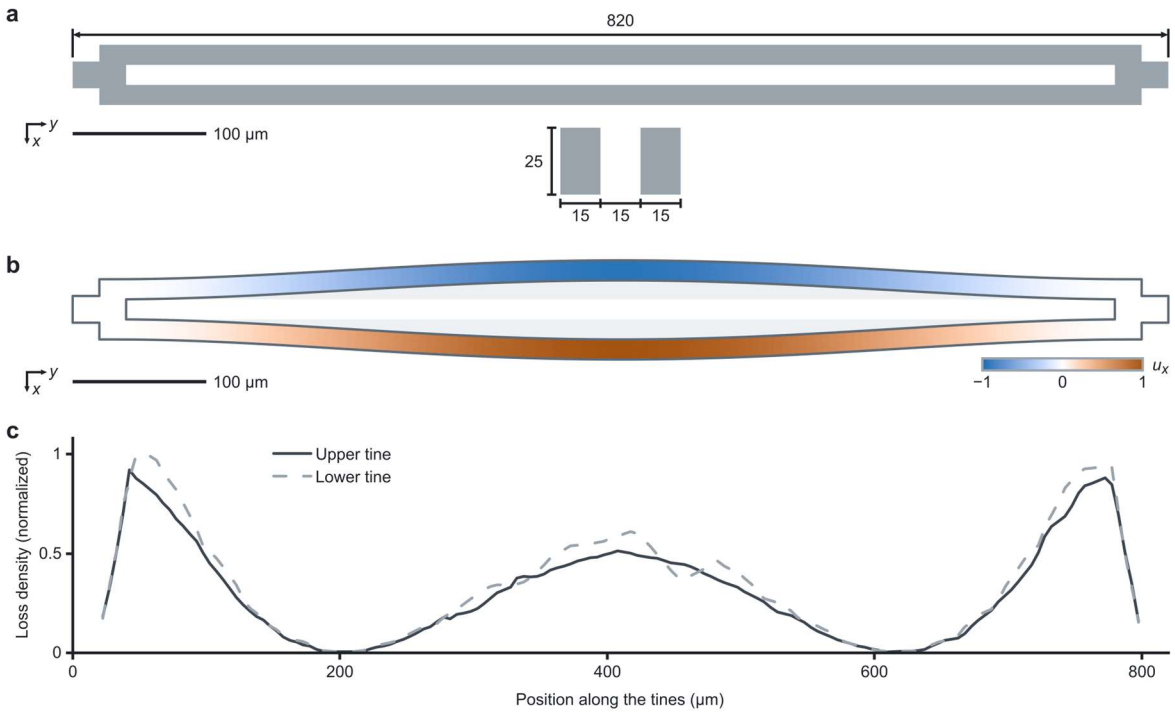
In a 5 September tool check (Run 20260905-003), a load written to an inactive property produced zero response. With the load corrected, the third attempt returned a nonzero response, and the fourth resolved the baseline resonance at 35 frequencies, giving half-power $Q = 8023$ versus $Q_{\mathrm{TED}} = 8045$. The revised procedure required checking the effective load before solving and after reopening, and diagnosis of zero response before declaring a mode absent. A coding assistant saved the revision under an earlier approved plan; no separate sign-off was recorded. The researcher later approved a solved-frequency export fix. Run 20260906-001 read the revision and used the corrected tool for four sweeps totaling 148 frequencies, including the two-row design called *c02* here. These records establish revision and reuse but do not establish the revision's effect on research success.

S1.4 Published reference and calibration

The published fabricated devices had operating-mode total Q-factor values of 7300 and 43000 near 220.2 and 100.7 kHz. Before the early study, the authors adjusted the baseline thermal-expansion coefficient from 2.6 to $3.0 \times 10^{-6}\ \mathrm{K^{-1}}$ to approach the published simulated baseline Q_{TED} of 8100. The resulting baseline Q_{TED} was 8045; effective thermal conductivity was 70 $\mathrm{W\ m^{-1}\ K^{-1}}$ despite 130 in the material record. The modeled published layout gave 110.064 kHz and $Q_{\mathrm{TED}} = 55202$. Its ratio, 6.862, was unchanged by this α calibration. The total measured Q-factor includes other losses and does not validate the simulated candidates.



Supplementary Fig. S1 | Organization and execution of a MEMS research task. Research roles and tools are shown with records of the early device study. **(a)** The coordinator connects six scientific roles with tools and outputs; roles need not be separate agents. **(b)** In the early study, the agent re-solved the baseline, loaded the mode procedure, screened saved fields, and recorded its export selection; screening guides inspection but does not establish mode identity. **(c)** The same run's research timeline. Arrows show information or execution links; indigo denotes procedures.



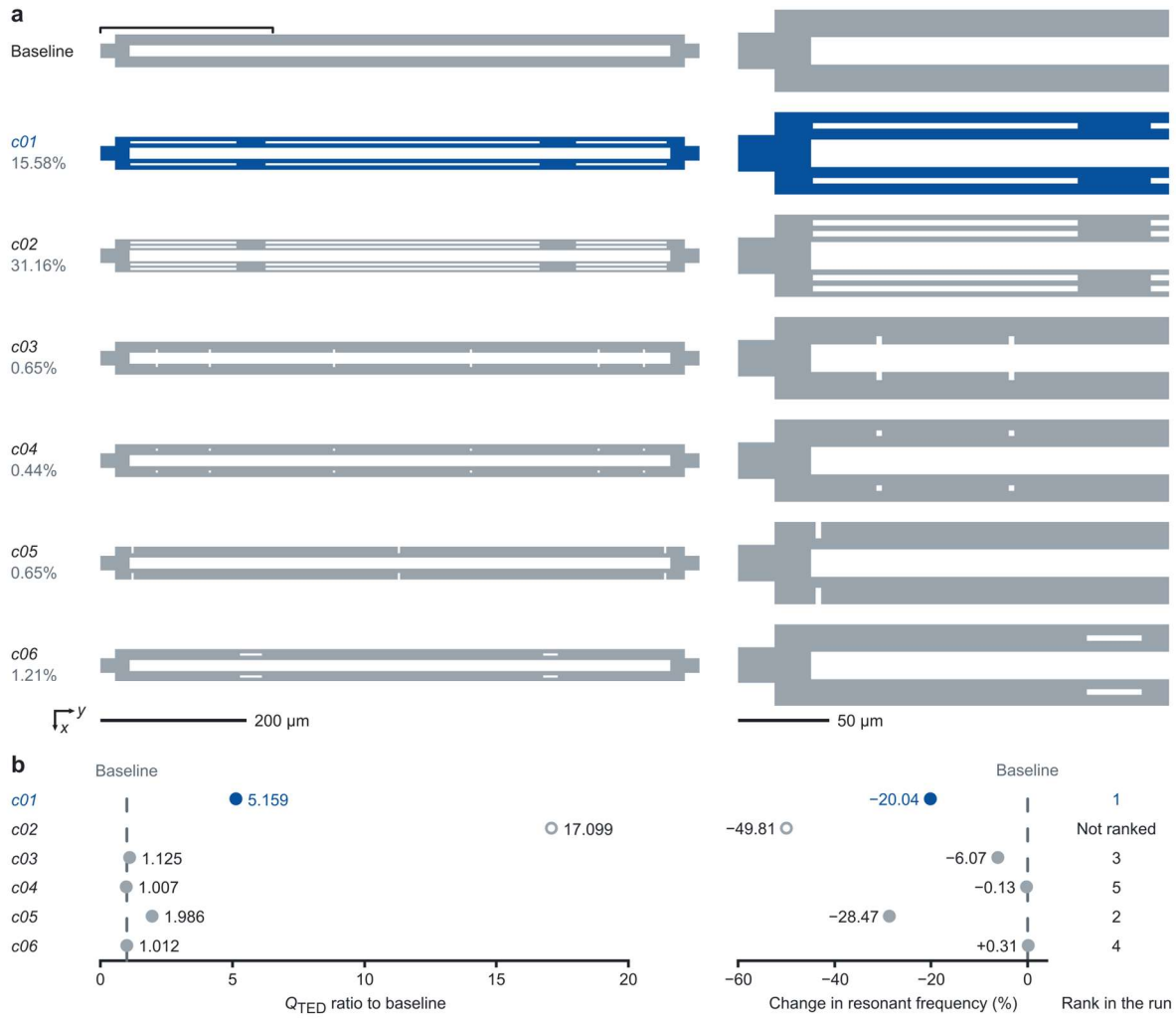
Supplementary Fig. S2 | Baseline geometry, operating mode, and thermoelastic-loss distribution in the early study. Following the task's guidance, the agent defined three relief segments from three dominant loss bands of the baseline, producing *c01* (Figs. 2 and 5b). **(a)** STEP geometry of the supplied DETF baseline (dimensions in µm). **(b)** Its fundamental in-plane flexural mode, in which the two tines vibrate out of phase. **(c)** Conduction-loss indicator $0.5k|\nabla T|^2/T$ along the tines at 293.15 K. In b, color shows normalized transverse displacement; the deformation is normalized and exaggerated for visibility. The indicator in c locates high-loss regions but does not partition integrated modal dissipation.

Supplementary Table S2 | Geometry of the baseline, *c01*, candidate 2 of the device study, *c02* (the two-row slot design), and the two-slot layout of reference[15], measured from the STEP files.

Quantity	Baseline†	*c01*	Candidate 2 of the device study	*c02* (two-row slot design)‖	Two-slot layout of reference[15]*
Source of the geometry	Supplied with the tasks	Run 20260814-012 (device study)	Run 20260814-012 (device study)	Run 20260906-001	Run 20260814-001, by the authors
Overall length × width × thickness‡	820 µm × 45 µm × 25 µm	Same	Same	Same	Same
Tine width; gap between the tines, width × length‡	15 µm; 15 µm × 740 µm	Same	Same	Same	Same
Outer edge of each tine, between the anchor stems‡	780 µm	Same	Same	Same	Same
End block joining the two tines, length at each end	20 µm	Same	Same	Same	Same
Anchor stem at each end, width × length	20 µm × 20 µm	Same	Same	Same	Same
Slots per tine (both tines)	None	3 (6)	6, in two rows of 3 (12)	24, in two rows of 12 (48)	2 (4)
Slot width; depth	—	3 µm; full thickness	3 µm; full thickness	3 µm; full thickness	3 µm; full thickness
Slot lengths along the gap§	—	145, 375, and 124 µm	145, 375, and 124 µm in each row	55 µm each	320 and 320 µm
Uncut length between successive slots§	—	40 and 50 µm	40 and 50 µm in each row	5 µm each, 11 per row	50 µm
Uncut length at the two ends of the gap§	—	1 and 5 µm	1 and 5 µm	12.5 and 12.5 µm	25 and 25 µm
Solid strips across a slotted section of a tine, outer edge to gap side	—	6 and 6 µm	3, 3, and 3 µm	3, 3, and 3 µm	6 and 6 µm
Solid volume	620000 µm³	523400 µm³	426800 µm³	422000 µm³	524000 µm³
Volume removed from the baseline	—	15.58%	31.16%	31.94%	15.48%

Geometry was measured from the imported STEP files with CadQuery. All cuts are rectangular full-depth removals, mirrored between tines; the outer outline, anchors, spacing, and thickness are retained. The 820, 780, and 740 µm lengths denote respectively anchor-end separation, outer tine edge, and inter-tine gap; none is an effective beam length. Along the 740 µm gap, cuts and uncut segments sum to that span. The closest agent-generated cut is 41 µm from an anchor end face.

* Author-built model of the published layout, not an agent-generated candidate. † Model of the published unslotted device. ‡ Outline spans defined above. § Measured along the gap. ‖ The later two-row design is from Run 20260906-001; candidate 2 of the August study is a different design despite its original filename *c02*.



Supplementary Fig. S3 | Candidates of the early device study. The run recommended *c01*. **(a)** Baseline and six candidate STEP geometries at true scale, with enlarged anchor ends; labels *c02–c06* mean candidates 2–6, not the later two-row *c02* design. **(b)** Operating-mode Q_{TED} ratios and frequency changes at 293.15 K, and ranking; candidate 2 was excluded because mode identification was incomplete across the required temperatures. In a, gray denotes silicon, blue *c01*, and white full-depth removal, and percentages give the fraction of baseline volume removed. Each candidate has one original calculation.

Supplementary Table S3 | Discretization and stored solutions in the six calculations made after the run.

Run	Structure	COMSOL element size	Mesh elements	Degrees of freedom*	Stored eigensolutions at each temperature step†
20260930-031	Baseline	Normal‡	495	4140	8, 8, 9, 7
20260930-032	*c01*	Normal‡	3844	27997	10, 11, 10, 11
20260930-033	Baseline	Fine	1043	7566	12, 6, 10, 12
20260930-034	*c01*	Fine	6587	46689	12, 11, 8, 9
20260930-035	Baseline	Finer	1205	8884	6, 7, 9, 9
20260930-036	*c01*	Finer	10608	71320	11, 10, 10, 10

These author calculations were not part of the original autonomous run. Only predefined element size changed: Normal, Fine, and Finer correspond to h_{auto} = 5, 4, and 3. Fields were subsequently exported without solving to identify modes for Table S4. The agent's own *c01*-only mesh refinement is separate (Note S1.2).

* Solved degrees of freedom, excluding separately logged internal degrees of freedom. † Counts at ΔT = 0, 20, 40, and 60 K; 226 eigensolutions comprise 144 vibration modes and 82 non-oscillatory decays. ‡ Original predefined size; these runs re-solved the saved models unchanged.

Supplementary Table S4 | Operating mode of the baseline and *c01* on three meshes, identified from displacement fields.

Temperature rise above 293.15 K (K)*	COMSOL element size	Baseline: solution†	Baseline: resonant frequency (kHz)	Baseline: Q_{TED}	*c01*: solution†	*c01*: resonant frequency (kHz)	*c01*: Q_{TED}	Q_{TED} ratio, *c01* to baseline
0	Normal	6	231.046	8045	4	184.743	41499	5.159
0	Fine	10	229.999	8155	6	184.253	41367	5.073
0	Finer	4	230.151	8162	5	183.872	41179	5.046
20	Normal	6	226.034	7272	5	178.087	37035	5.093
20	Fine	4	224.985	7369	5	177.591	36926	5.011
20	Finer	5	225.166	7377	4	177.227	36765	4.984
40	Normal	7	220.898	6592	4	171.060	32997	5.006
40	Fine	8	219.846	6677	2	170.557	32907	4.928
40	Finer	7	220.060	6686	4	170.213	32771	4.902
60	Normal	5	215.631	5988	5	163.588	29288	4.891
60	Fine	10	214.576	6063	3	163.078	29214	4.818
60	Finer	7	214.823	6073	4	162.755	29101	4.792

Operating modes were identified from saved fields before attaching numerical values: predominantly transverse in-plane displacement, opposite tine motion, and one lobe per tine. The author review was not blind to earlier results, and each mesh was inspected separately rather than matched by modal-shape correlation. Twenty-two broader or mixed competitors were retained; all were non-oscillatory decays. $Q_{TED} = |\mathrm{Re}(f)|/[2|\mathrm{Im}(f)|]$, with thermoelastic loss alone; ratios use unrounded values.

At 293.15 K the ratio falls from 5.159 to 5.046 across these meshes. Over the four temperatures its minimum falls from 4.891 to 4.792. Nonmonotonic frequency changes and unequal refinement increments make these sensitivity checks, not converged values or rigorous error bounds.

* Temperatures retained because the early task ranked the lowest ratio over all four. † Index in the saved solution set at that temperature; modes were not selected by index, frequency, or *Q*.

Supplementary Table S5 | Development periods and the ten early research runs.

(a) Essential development history

Dates in 2026	Activity and retained record	Interpretation
26 July–8 August	Exploration followed by scripted baseline/launch preparation	Early setup; no common outcome assessment
9–15 August	Ten research runs while researchers built procedures, ending with *c01*	Guidance and model usage changed; not controlled repeats
21 August–30 September	Procedure use, infrastructure repairs, and geometry/response revisions	Records trace revision and reuse
29 September–1 October	Eight reviewed rounds, each followed by a recorded approval	Method-development sequence, not independent evaluation

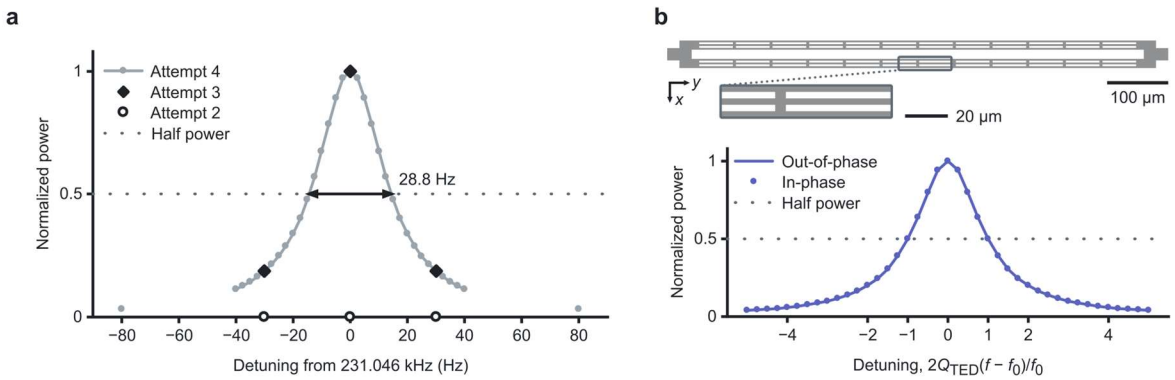
The last two phases overlap. Researchers continued infrastructure repairs; records do not attribute every change to an agent. Earlier paired comparisons are summarized in Table S10. These histories explain the origin of the frozen procedures but do not enter the 25-run comparison.

(b) Runs underlying Fig. 3b

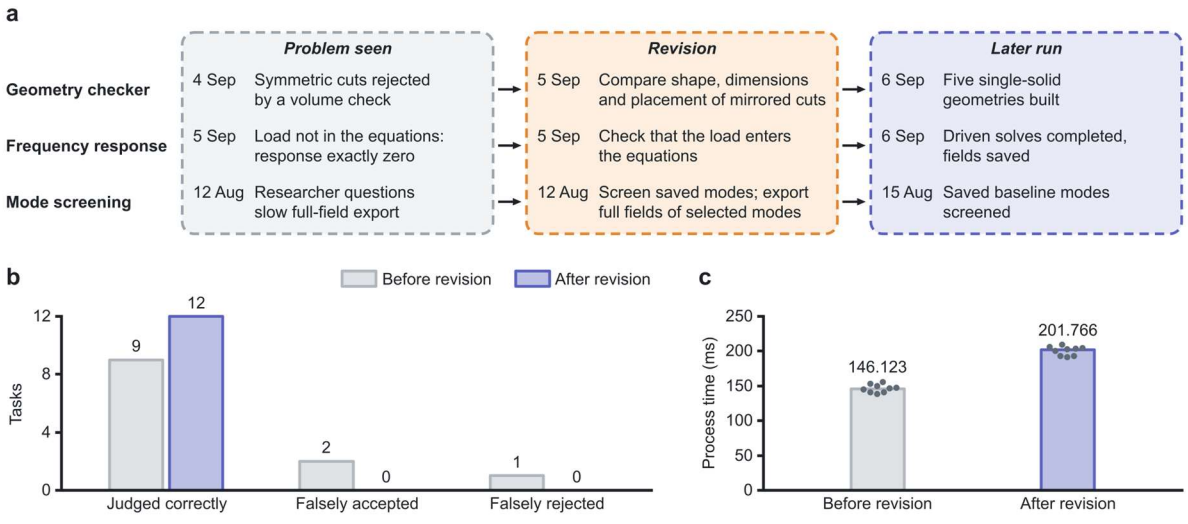
Run in Fig. 3b	Run	Group in Fig. 3b	Total tokens (M)§	Candidates calculated¶	Result adopted‖
1	20260809-002	Before guidance	90	0	No
2	20260810-001	Before guidance	1192	0	No
3	20260810-003	Before guidance	782	1	No
4	20260811-001	Before guidance	1045	0	No
5	20260812-001	Guided	106	4	Yes
6	20260813-002	Guided	116	14	Yes
7	20260814-009	Guided	71	2	Yes
8	20260814-010	Guided	65	7	Yes
9	20260814-011	Guided	85	7	Yes
10	20260814-012	Device study of Fig. 2 (*c01*)	147	6	Yes

Tokens include cached input and all main/sub-agent sessions, unlike the output-only measure in Fig. 4c. The 914 M versus 85 M medians reflect, among other changes, extensive early sub-agent use and a model change; main-session medians were 93 M and 85 M. They cannot estimate efficiency gained from guidance. Dates in run IDs are preparation dates; Runs 8–10 launched on 15 August UTC. Runs 8–9 supplied the prior records later read by the *c01* study. The larger phase registry includes scripted/support runs and unlaunched preparations, not additional research repeats.

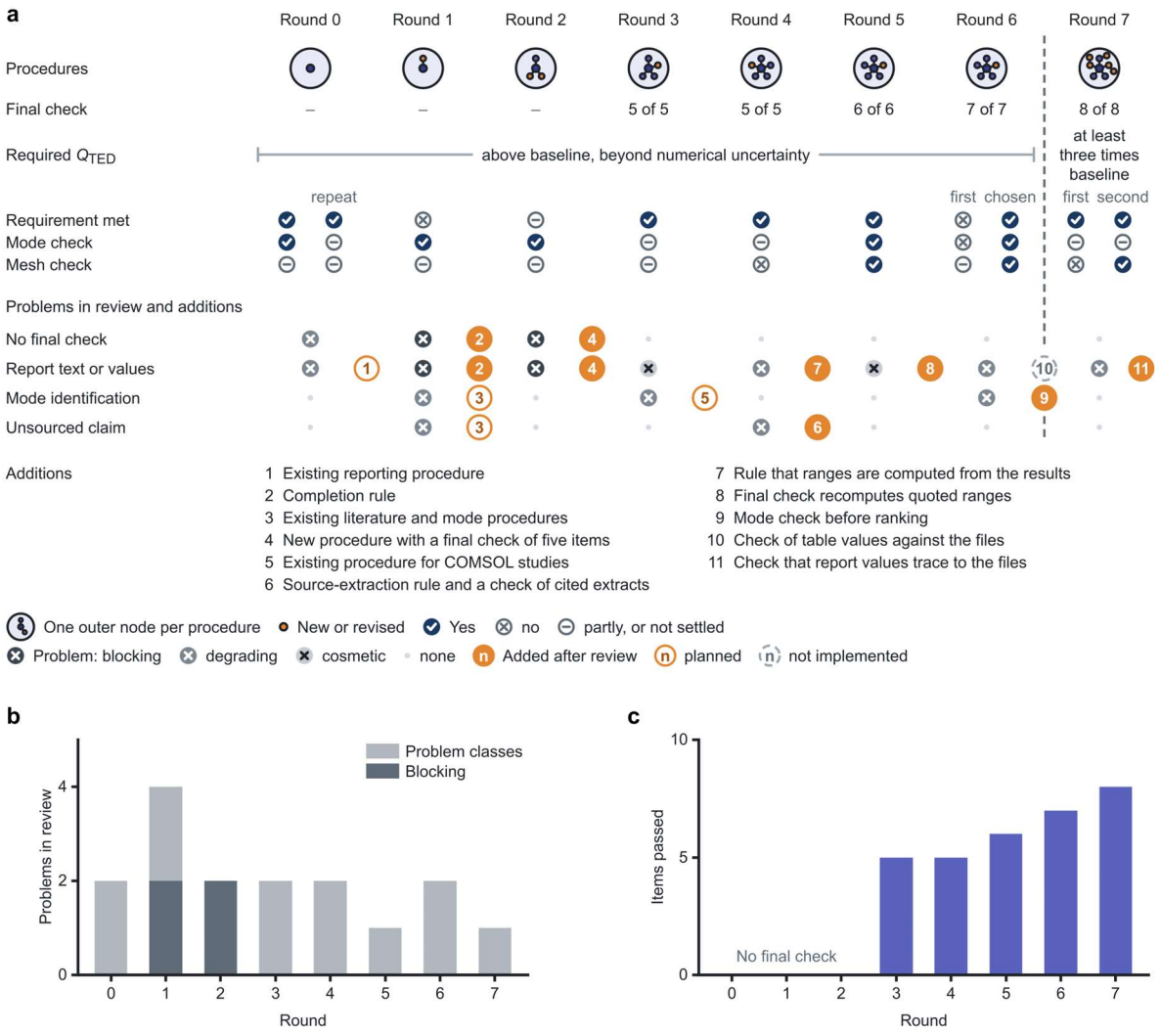
§ Rounded millions. ¶ Distinct candidates with a completed solve. ‖ Registry adoption, not independent physical acceptance.



Supplementary Fig. S4 | The frequency-response revision and its later use. Calculated driven responses come from the tool check that prompted the revision and from a later run that used it (Note S1.3). **(a)** Four baseline tool-check attempts; attempt 1 stopped before solving, and the second returned zero because its load property was inactive; power is normalized to attempt 4's peak. **(b)** Out-of-phase and in-phase modes of the two-row design *c02* from Run 20260906-001, each normalized to its own peak; f_0 is the mode's eigenfrequency. The run in b made four sweeps of the baseline and design, with half-power and eigenfrequency Q agreeing within 1%; the baseline sweeps appear in Fig. 2c. Each curve is one sweep with thermoelastic coupling retained.



Supplementary Fig. S5 | Earlier procedure revisions and subsequent use. Revision records are shown with author tests of the revised geometry checker. **(a)** Three revision histories connect a recorded problem to a revised procedure and later use. **(b)** Classifications by the geometry checker before and after revision, in author tests on six symmetric and six asymmetric tasks; three fresh-process repeats per task and version agreed. **(c)** Process time of the two checker versions for nine tasks correct in both versions; dots show task medians and columns their median. These software tests do not measure effects on research success or time savings.



Supplementary Fig. S6 | Eight developmental rounds of one task. The rounds document development and are not independent repeats or a controlled learning curve (Table S6; Note S2). **(a)** One reviewed run per round; the separate unreviewed Round 0 repeat is marked; columns summarize declared attainment, mode/mesh checks, file checks, review findings, and approved changes. **(b)** Review problem classes, with blocking findings distinguished. **(c)** File-check items passed. Round 7, separated by a dashed line, changed the task, procedures, reasoning, and scripts, retrieved reference[15] before freezing candidates, and set mesh tolerances after solving.

Supplementary Note S2. Method revision across eight developmental rounds

S2.1 Task and records

Round 0–Round 7 ran between 29 September and 1 October 2026 using GLM-5.3 in Claude Code. The task sought higher operating-mode Q_{TED} by material removal from the same baseline. Round 0–Round 6 requested at least five candidates, frozen before geometry construction and compared over four temperatures; Round 7 changed the objective to at least three times the baseline at its stated conditions. Each round is one reviewed run. Separate repeats of Round 0 and Round 7 lack completed reviews and are excluded from the sequence. Two earlier attempts without procedures were excluded after reading externally stored procedures; one was deleted at the researcher's request and survives as an incident record.

S2.2 Review and approval

A separate session reviewed each round for seven classes: unsupported claims, missing candidate freeze, geometry mismatch, baseline re-solve failure, incorrect mode correspondence, untraceable values, and incomplete finalization. Findings required file evidence and severity labels. The reviewer used the same language model and was not blinded. The researcher decided in chat; an AI coding assistant recorded the decision and prepared the next round. All eight decisions were recorded as approved, seven before another round; the records identify “user” as approver but do not contain the original chat. Consequently they do not test selective expert rejection. Weights were never updated.

S2.3 Revisions that entered later rounds

Finalization was added after Round 2, followed by completed reports and passing file checks from Round 3. After Round 4, a source-extraction requirement was added; later reviews did not repeat that finding, but its implemented check searched a different folder from the original defect. The absence of recurrence therefore rests on the reviews alone. A numeric-range check approved after Round 5 passed in later rounds because neither report contained a readable range; reviews still found untraceable values. A single-value check approved after Round 6 was not implemented in Round 7. Thus passing a checker did not necessarily remove the underlying problem.

S2.4 Operating-mode revision

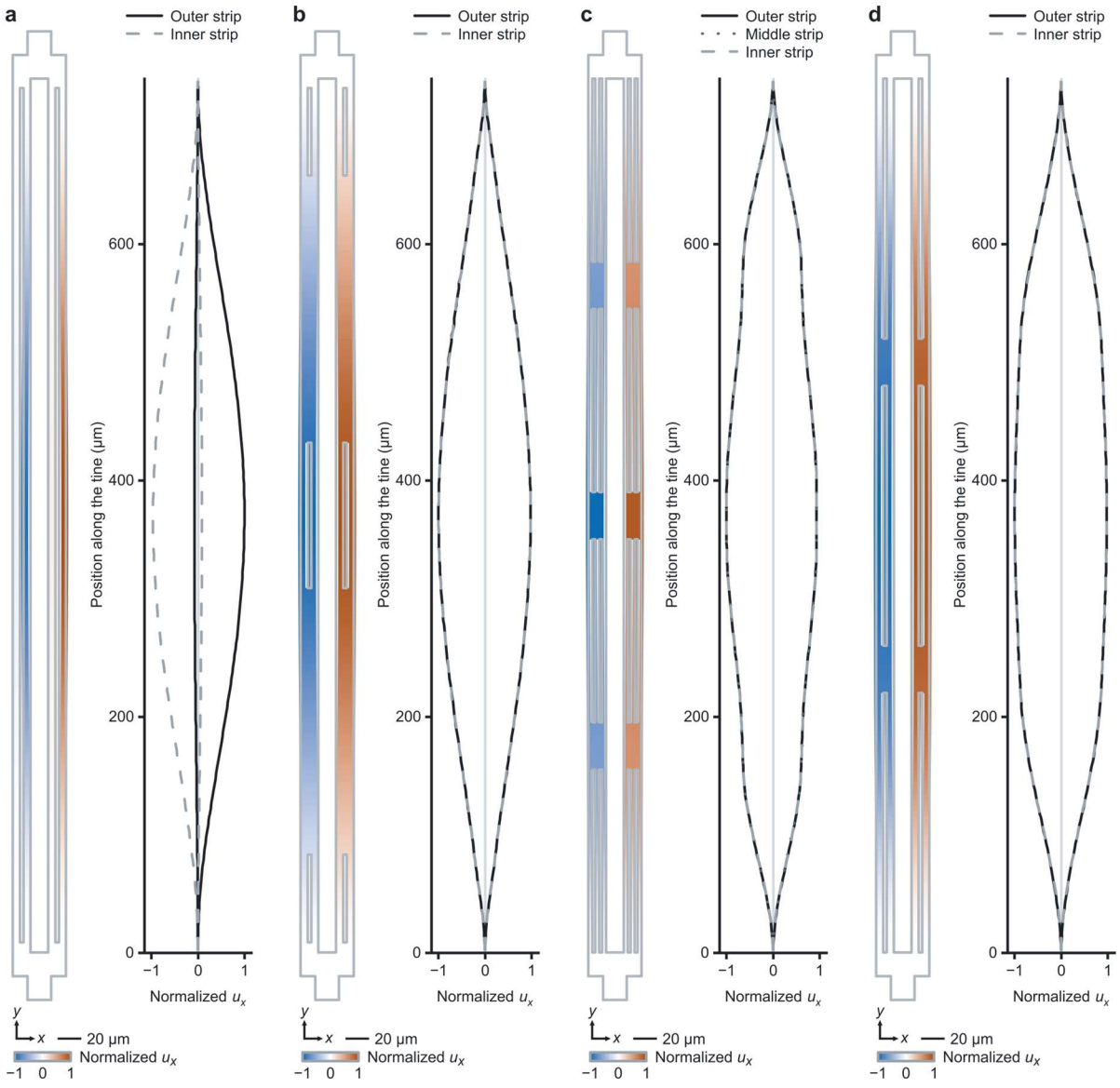
Round 6 ranked a full-length slot at Q_{TED} 10.0× baseline, although strip amplitudes showed that it was not the intended operating mode. Its report instead recommended a 2.1× design, but the ranking was not corrected. A revision required saved-field identification before ranking. Round 7 ranked two segmented designs, the second at 6.1×, yet its checks followed one line per tine without comparing strips. The first design failed its own mesh tolerance, chosen after solves, so its 13.1× value is not converged. The task, procedures, reasoning setting, and scripts also changed, and related literature including reference[15] was read before freezing candidates. The result cannot be attributed to the mode rule or called blind discovery.

S2.5 What the measures mean

Table S6 retains completion scores, baseline re-solve times, and script counts. A score of 1.00 means six reporting/file criteria were met and does not establish physical validity. Baseline timing used 1% tolerance in Round 0–Round 6 and 5% in Round 7; early times are upper bounds. Reproducing the same scripted baseline to parts per billion largely reflects rounded reference values, not agreement with an experiment. Stored-script calls and new script files measure different activities; inline code is not counted. The observed code mix does not establish time or token efficiency.

S2.6 Limits

Round 1–Round 4 accessed earlier folders; multiple settings changed together, and only one reviewed run represents each round. Several procedures were introduced on a schedule rather than derived from the preceding failure. The revision approved after Round 7 has no subsequent test in this sequence. Earlier method histories and historical configuration comparisons supply complementary evidence. This evidence is limited to recorded method development and subsequent use; the 25 registered runs alone support comparisons of research outcomes across configurations.



Supplementary Fig. S7 | Saved modes in the Round 6 and Round 7 development records. Saved displacement fields and author-extracted strip profiles for four designs. **(a)** Round 6's first-ranked full-length slot: its 10.0-fold Q_{TED} belongs to a non-operating mode. **(b)** Its recommended design: 2.1-fold Q_{TED} and 2% lower frequency. **(c)** Round 7's first design: 13.1-fold Q_{TED} (not converged) and 50% lower frequency; its mesh check failed the declared frequency tolerance. **(d)** Its second design: 6.1-fold Q_{TED} and 39% lower frequency, within mesh tolerance. Values are relative to the baseline. Color shows transverse displacement, gray the undeformed geometry, and deflection is normalized and exaggerated for visibility. Round 7 did not compare strips itself.

Supplementary Table S6 | Process records of the eight rounds of one task: problems recorded in the reviews, final check, completion score, time to the first valid baseline, and scripts written and reused.

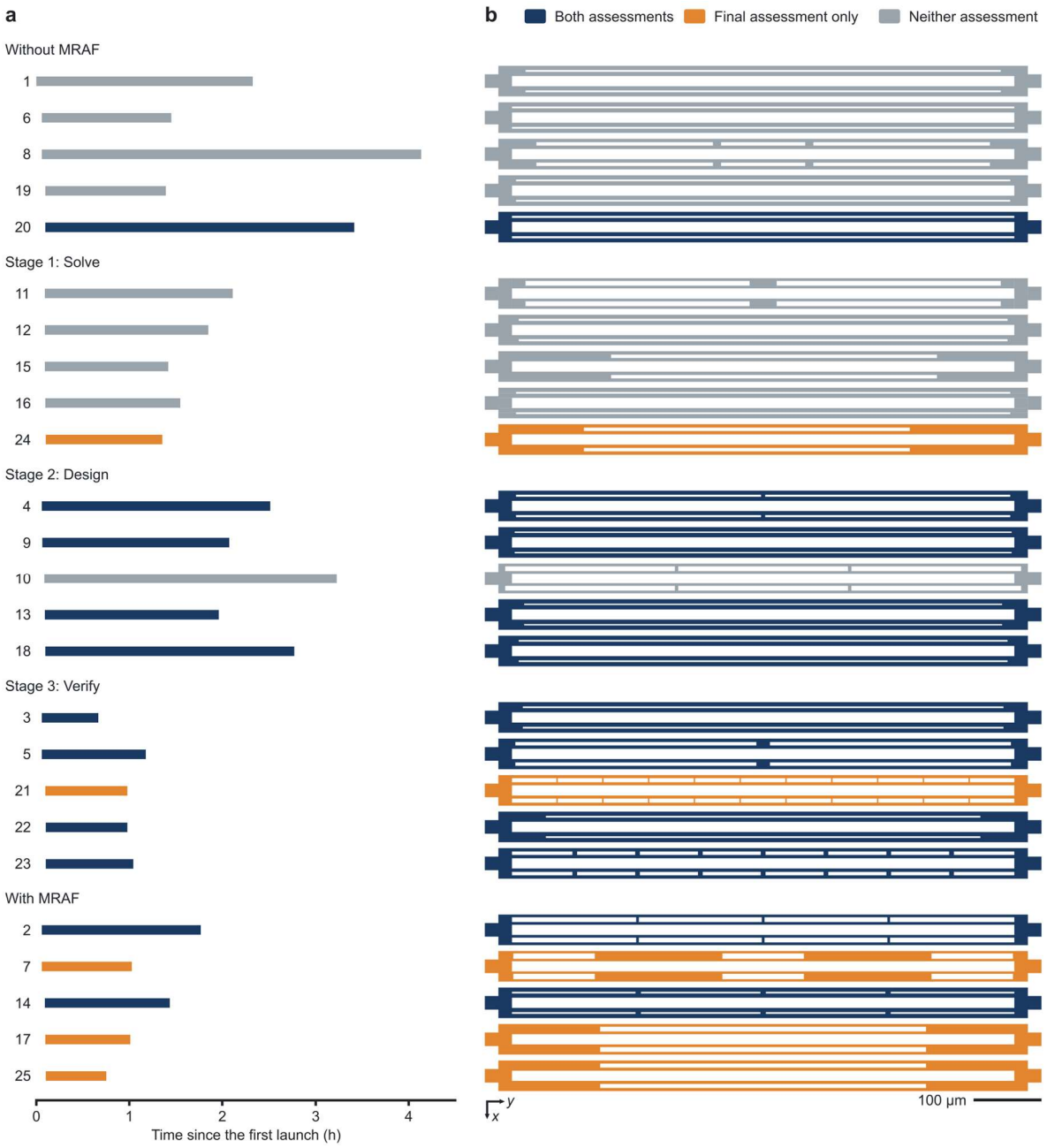
Round	Procedures available	Problem classes recorded in the review (severity)	Problems recorded (blocking)	Final check, items passed	Completion score‡	Time to first valid baseline (min)‖	Script files written by the agent§	Runs of supplied scripts¶
Round 0	0	Report traceability (degrading); final check (degrading)	2 (0)	None	0.50	41.1	45	0
Round 1	1	Literature source (degrading); mode correspondence (degrading); report traceability (blocking); final check (blocking)	4 (2)	None	0.06	20.3	34	0
Round 2	3	Report traceability (blocking); final check (blocking)	2 (2)	None	0.33	26.7	26	0
Round 3	4	Mode correspondence (degrading); report traceability (cosmetic)	2 (0)	5/5	0.72	19.1	19	1
Round 4	5	Literature source (degrading); report traceability (degrading)	2 (0)	5/5	0.83	10.6	18	29
Round 5	5	Report traceability (cosmetic)	1 (0)	6/6*	1.00	13.4	18	63
Round 6	5	Mode correspondence (degrading); report traceability (degrading)	2 (0)	7/7	1.00	10.9	14	32
Round 7	8	Report traceability (degrading)	1 (0)	8/8†	1.00	12.6	26	96

One reviewed run represents each round. Procedure counts include existing procedures introduced on schedule. Review findings describe recorded problem classes; they are not complete error counts. "Final check" counts passed file/report checks. Problem counts give blocking findings in parentheses. The completion score is the mean of six criteria: report present, no placeholders, manifest present, final-check pass rate, frozen-candidate coverage, and no workspace escape. Neither metric validates physical results.

Baseline time is from the first message to the first baseline re-solve within task tolerance: 1% in Round 0–Round 6 and 5% in Round 7. Round 0–Round 2 values are upper bounds. Earlier supplied baseline values came from scripted solves; their parts-per-billion reproduction is not a measurement comparison. New script files exclude inline code; supplied-script executions count calls and are a different quantity. The total new code did not fall consistently.

Round 1 ended without a report. Round 1–Round 4 could access earlier folders, and Round 7 changed task, reasoning setting, scripts, and procedures and retrieved reference[15] before freezing candidates. These development records are not independent repeats or a controlled improvement curve. Note S2 gives the material limitations; Fig. S6 shows the underlying check sequence.

* Round 5's session-log result is retained because a later retest overwrote its stored check. † Round 7 passed 8/8 during the run; an item added after review later gave 9/9 and is not credited to that run. ‡ Scores were calculated after the sequence; an operator resource cap restricted Round 3 to two of six candidates. § Script counts include two repair scripts in each of Round 4–Round 6. ¶ Supplied-script counts differ by up to 10% from an independent counting rule. ‖ Early timing values are upper bounds as defined above.



Supplementary Fig. S8 | Timing and reported designs of all 25 registered runs. Five runs per configuration used fresh sessions and separate hosts from one image; all runs stopped within 4.2 h of the first launch, before their individual 8 h limits. **(a)** Timing of each run; rows follow launch time, and numbers give the seeded launch order (Table S7a), not run numbers. **(b)** Assessed STEP geometries at a common scale. Colors distinguish success under both initial model-judge and final author assessments, the final assessment only, or neither. Reported geometries need not meet the evidence requirements or the specified fabrication rules.

Supplementary Table S7 | The 25 registered runs: configuration, host, language model, procedures, outcome under each registered condition and under each reading, judges, and recomputed values of the reported design.

The final author assessment is the main reported outcome: 1, 1, 4, 5, and 5 successes across the five configurations. Results of the initial model-judge assessment remain for provenance and sensitivity. The source labels "the author's reading" and "registered judging" correspond to the final author assessment and initial model-judge assessment, respectively. Both neutral-code tables were frozen before unblinding; Run 20261023-022 (J19) was corrected afterwards from complete files (Note S3.2). Rows are ordered by configuration and launch time.

(a) Run configuration and launch record

Configuration*	Run†	Launch order and host‡	Start (UTC)§	Language model‖	Procedures¶
Without MRAF	20261023-004	1	00:44:31	DeepSeek-V4.1-Flash	None supplied
	20261023-021	6	00:48:08	DeepSeek-V4.1-Flash	None supplied
	20261023-006	8	00:48:13	DeepSeek-V4.1-Flash	None supplied
	20261023-010	19	00:50:20	DeepSeek-V4.1-Flash	None supplied
	20261023-016	20	00:50:23	DeepSeek-V4.1-Flash	None supplied
Stage 1: Solve	20261023-023	11	00:49:50	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-005	12	00:49:55	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-009	15	00:50:06	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-007	16	00:50:09	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-013	24	00:50:38	DeepSeek-V4.1-Flash	Version 2026-10-23.1
Stage 2: Design	20261023-017	4	00:48:04	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-020	9	00:48:14	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-015	10	00:49:47	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-018	13	00:50:00	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-011	18	00:50:16	DeepSeek-V4.1-Flash	Version 2026-10-23.1
Stage 3: Verify	20261023-014	3	00:48:01	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-022	5	00:48:06	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-025	21	00:50:27	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-008	22	00:50:30	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-019	23	00:50:36	DeepSeek-V4.1-Flash	Version 2026-10-23.1
With MRAF	20261023-012	2	00:48:00	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-026	7	00:48:11	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-024	14	00:50:03	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-028	17	00:50:13	DeepSeek-V4.1-Flash	Version 2026-10-23.1
	20261023-027	25	00:50:42	DeepSeek-V4.1-Flash	Version 2026-10-23.1

(b) Conditions and outcomes

Configuration	Run	Q_{TED} > 4× baseline**	Evidence complete**	Geometry meeting the specified fabrication rules**	Void**	Outcome, initial model-judge assessment**	First unmet condition**	Outcome, final author assessment††	Outcome, the judges' relaxed reading‡‡	Run's own completion claim§§
Without MRAF	20261023-004	Met	Not met	Met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; blocked; partial
	20261023-021	Met (2 of 3)	Not met	Met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; blocked; partial

Configuration	Run	Q_{TED} > 4× baseline**	Evidence complete**	Geometry meeting the specified fabrication rules**	Void**	Outcome, initial model-judge assessment**	First unmet condition**	Outcome, final author assessment††	Outcome, the judges' relaxed reading‡‡	Run's own completion claim§§
	20261023-006	Met	Met	Not met	No	Not successful	Geometry meeting the specified fabrication rules: outer side face of the baseline cut	Not successful	Not successful	Target met; goal completed; complete
	20261023-010	Met	Not met	Met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; blocked; partial
	20261023-016	Met	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete
Stage 1: Solve	20261023-023	Met	Not met	Met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; blocked; partial
	20261023-005	Met	Not met	Met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; blocked; partial
	20261023-009	Met	Not met	Met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; goal completed; partial
	20261023-007	Met (2 of 3)	Not met	Not met	No	Not successful	Evidence complete: driven sweep without thermoelastic loss	Not successful	Not successful	Target met; goal completed; partial
	20261023-013	Met	Not met	Met	No	Not successful	Evidence complete: no two-level mesh check of the assessed design	Success	Success	Target met; goal completed; complete
Stage 2: Design	20261023-017	Met	Met (2 of 3)	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-020	Met	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-015	Met	Met	Not met	No	Not successful	Geometry meeting the specified fabrication rules: slots reaching about 10 μm into both anchor regions	Not successful	Not successful	Target met; goal completed; complete
	20261023-018	Met (2 of 3)¶¶	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-011	Met	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete
Stage 3: Verify	20261023-014	Met	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete

Configuration	Run	Q_{TED} > 4× baseline**	Evidence complete**	Geometry meeting the specified fabrication rules**	Void**	Outcome, initial model-judge assessment**	First unmet condition**	Outcome, final author assessment††	Outcome, the judges' relaxed reading‡‡	Run's own completion claim§§
With MRAF	20261023-022	Met	Met	Met	No	Success‖	—	Success‖	Success‖	Target met; goal completed; complete
	20261023-025	Met	Not met	Met	No	Not successful	Evidence complete: mesh check outside the declared tolerance; no record of the harmonic temperature perturbation of the design	Success	Not successful	Target met; goal completed; complete
	20261023-008	Met	Met (2 of 3)	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-019	Met	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-012	Met	Met (2 of 3)	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-026	Met	Not met (2 of 3)	Met	No	Not successful	Evidence complete: no record of the harmonic temperature perturbation of the design	Success	Not successful	Target met; goal completed; complete
	20261023-024	Met	Met	Met	No	Success	—	Success	Success	Target met; goal completed; complete
	20261023-028	Met	Not met (2 of 3)	Met	No	Not successful	Evidence complete: no record of the harmonic temperature perturbation of the design	Success	Not successful	Target met; goal completed; complete
	20261023-027	Met	Not met (2 of 3)	Met	No	Not successful	Evidence complete: no record of the harmonic temperature perturbation of the design	Success	Not successful	Target met; goal completed; complete

(c) Judges and recalculated values

Configuration	Run	Judges who answered	Mesh	Q_{TED} of the design	Q_{TED} ratio	Change in resonant frequency (%)	Removed volume fraction (%)
Without MRAF	20261023-004	K3.1 Pro, GLM-5.3	Finer	59280	7.26	−53.3	14.7
	20261023-021	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Finer	73496	9.01	−57.5	15.5
	20261023-006	K3.1 Pro, GLM-5.3	Finer	66559	8.16	−50.7	26.0
	20261023-010	K3.1 Pro, GLM-5.3	Finer	66775	8.18	−55.9	14.7
	20261023-016	GLM-5.3, GPT-6-ASTRA***	Extra fine	80667 / 80731	9.90 / 9.91†††	−58.7 / −58.3	17.9
Stage 1: Solve	20261023-023	GLM-5.3, GPT-6-ASTRA, DeepSeek-V4.1-Flash***	Finer	205266	25.15	−69.7	37.3
	20261023-005	K3.1 Pro, GLM-5.3	Finer	72053	8.83	−56.8	17.4
	20261023-009	K3.1 Pro, GLM-5.3	Extra fine	33798	4.15	−29.2	19.4
	20261023-007	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Finer	66262	8.12	−56.0	Not computed‡‡‡
	20261023-013	GLM-5.3, GPT-6-ASTRA***	Extra fine	33380	4.10	−29.4	19.4
Stage 2: Design	20261023-017	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Extra fine	74421	9.13	−57.4	17.5
	20261023-020	K3.1 Pro, GLM-5.3	Finer	60728	7.44	−54.6	11.8

Configuration	Run	Judges who answered	Mesh	Q_{TED} of the design	Q_{TED} ratio	Change in resonant frequency (%)	Removed volume fraction (%)
	20261023-015	K3.1 Pro, GLM-5.3	Finer	164701	20.18	−61.0	42.3
	20261023-018	GLM-5.3, GPT-6-ASTRA, DeepSeek-V4.1-Flash***	Normal¶¶	52892	6.57	−51.7	11.4
	20261023-011	K3.1 Pro, GLM-5.3	Finer	64096	7.85	−55.1	14.5
Stage 3: Verify	20261023-014	K3.1 Pro, GLM-5.3	Finer	60084	7.36	−53.8	14.3
	20261023-022	GLM-5.3, DeepSeek-V4.1-Flash***	Extra fine	126263	15.50	−64.4	28.6
	20261023-025	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Extra fine	50949	6.25	−10.2	34.8
	20261023-008	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Finer	42803	5.24	−45.8	12.9
	20261023-019	K3.1 Pro, GLM-5.3	Extra fine	38014	4.67	−8.3	27.9
With MRAF	20261023-012	GLM-5.3, GPT-6-ASTRA, DeepSeek-V4.1-Flash***	Finer	111737	13.69	−49.3	41.1
	20261023-026	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Extra fine	58497	7.18	−41.1	23.2
	20261023-024	K3.1 Pro, GLM-5.3	Finer	42396	5.19	−29.5	14.4
	20261023-028	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Finer	58885	7.21	−39.4	27.1
	20261023-027	K3.1 Pro, GLM-5.3, GPT-6-ASTRA	Finer	43542	5.34	−34.0	23.2

* Configurations are cumulative subsets of frozen version 2026-10-23.1, not developmental rounds. † IDs use preparation date 23 October; launches were 24 October UTC. ‡ Launch position and host assignment follow the fixed seed. § Every run returned a model reply and ended before its individual 8 h limit; no host replacement or run exclusion occurred. ‖ All used DeepSeek-V4.1-Flash in Claude Code 2.1.310 at the highest main-session reasoning setting. ¶ The no-procedure arm had no skill files or loading hook.

** The initial model-judge assessment uses majority decisions on three registered conditions, with two or three judges as needed. Condition 1 requires operating-mode Q_{TED} >4× matched baseline; condition 2 requires the run's mesh and driven-response evidence; condition 3 requires full-depth removal respecting protected geometry and process rules. Fractions identify disagreement. A run would be void for altered baseline, access to another run, or use of the held-out design; none was void. “First unmet” follows condition order. Own completion claims do not decide outcomes.

†† The accepted author interpretation counts the mesh requirement as met and accepts a complex driven response through the operating-mode resonance without a separate harmonic-temperature record. It was set after viewing neutral-code summaries and frozen before unblinding, then adopted for final reporting. It does not imply unanimous independent judgments. In Runs 20261023-026 and -027, two of three judges also found no driven-peak displacement fields in the judging files; the run folders contained them. ‡‡ The judges' separate relaxed reading, introduced by amendment 3, differs from their registered outcome only for Run 20261023-013. §§ Self-reported target, stop, and completeness fields are descriptive only.

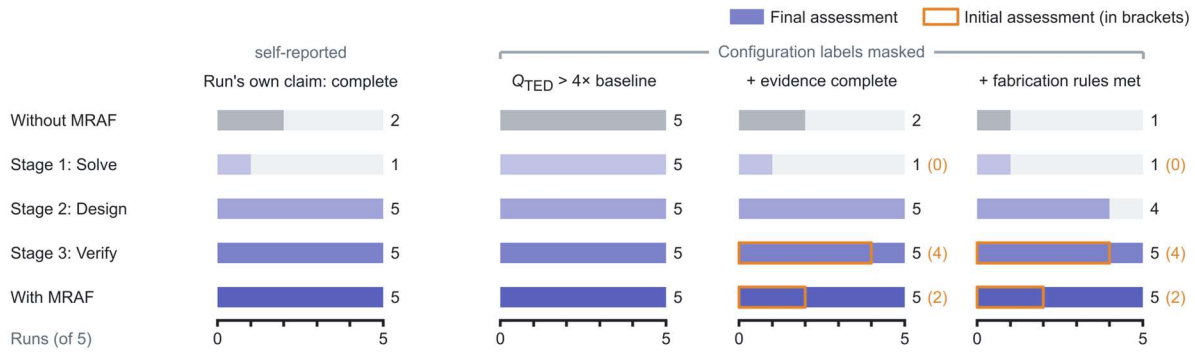
‖‖ Run 20261023-022 (J19) was frozen as unsuccessful after its second judgment used 161 of 326 files. With the complete judging files, GLM and DeepSeek found all conditions met after unblinding. Both assessment totals include this correction; the frozen table is unchanged.

¶¶ Run 20261023-018 failed the registered refined-mesh solve. GLM and DeepSeek accepted condition 1 from the Normal-mesh result; GPT-6-ASTRA dissented. Part c reports Q_{TED} = 52892 and ratio 6.57 against the same Normal-mesh baseline (Q_{TED} = 8045).

The deciding mesh is Extra fine where calculated, otherwise Finer, with the exception above. Matched baseline Q_{TED} values are 8148 on Extra fine and 8162 on Finer; ratios use unrounded values. Values and frequency changes use the operating mode identified from saved fields, never frequency or Q alone. For Runs 20261023-021 and -007, part c follows the mode accepted by two judges over the third's objection. Frequency change is $100 \times (f_{design}/f_{baseline} - 1)$; it is not a success criterion. Removed volume is descriptive.

*** Rejudged after clarification; DeepSeek replaced Kimi where its usage limit prevented an answer. Its votes support three model-judge successes, including the corrected Run -022. ††† In Run -016 the judges selected two different modes; both values are retained. ‡‡‡ Run -007's volume removal was not computed because the geometry check stopped at 12 disconnected solids.

Run 20261023-012 is the earliest-launched success with the MRAF under both assessments and the registered illustrative case. Fig. 2's post-unblinding *c01* selection contributes only development history.



Supplementary Fig. S9 | Cumulative outcome conditions in the 25 registered runs. Each configuration contains five runs. Self-reported completeness did not decide success; subsequent columns cumulatively count designs meeting Q_{TED} > 4× baseline, evidence requirements, and fabrication rules. Filled bars show the main final author assessment; bracketed values show the initial model-judge assessment where different. Final assessment counts the mesh requirement as met and accepts a complex driven resonance without a separate harmonic-temperature record. Note S3 gives confidence intervals and assessment history; Table S7 gives individual results.

Supplementary Note S3. Registered protocol, assessment history, and results

This note retains the decisions that affect interpretation of the 25-run experiment. Times are UTC. Full prompts, amendments, script freezes, first and later judgments, and copied-back run records remain archived. No development or pilot run is pooled with these 25.

S3.1 Registration and outcome

The author signed the registration at 16:54 on 23 October, before preparation. It fixed five configurations with five fresh runs each, seeded assignment, common inputs, one assessed design per run, and an 8 h limit. The design to be assessed was the final recommendation, or the highest reported Q_{TED}-ratio design when none was recommended. In practice, the frozen evidence collector took the first design in the run's comparison table among the comparisons the run listed as meeting the target; it used the design the run named as best only when no such list existed, and it did not read the report. In four runs the assessed design was therefore not the recommended one. In Runs 20261023-005 and -009 (Stage 1: Solve), no design had a driven sweep retaining thermoelastic loss, so the outcome did not depend on the choice. Run 20261023-013 (Stage 1: Solve) was assessed on a design without a mesh check, whereas its recommended design had a two-level check. The recommended design of Run 20261023-016, the only Without MRAF success in both assessments, was not assessed; assessing it would leave that count at one or reduce it to zero. A successful run had to meet three conditions: operating-mode Q_{TED} above four times baseline after recalculation; complete run-generated mesh and driven-response evidence; and geometry meeting the task's invariants and fabrication rules. Frequency was recorded, not scored. A changed baseline, access to another run, or use of the held-out device for candidate construction/ranking would void a run, but all runs remained in the denominator. None was void. Host failure before the first reply could be replaced; none occurred.

S3.2 Amendments and final acceptance

The pre-launch amendment at 00:23 on 24 October fixed the cloud host and Claude Code version, retained the known response-tool limit and held-out arrangement, and required judging scripts to be frozen before the first run ended. Their first freeze was at 00:55. The threshold of four was unchanged throughout.

At 01:28, after a run ended but before any registered result was opened, amendment 1 expanded cited folders/patterns during evidence collection because a pilot rehearsal had missed temperature and peak-field files. It selected Kimi K3.1 Pro and temporarily substituted Claude Sonnet 5.5 for the unavailable third judge. Amendment 2 at 06:48 corrected inaccessible cloud paths and judge calls, restored GPT-6-ASTRA as third judge, and masked a shared configuration-bearing path. The operator (AI coding assistants that prepared the judging files and ran the judges under the author's direction) had seen technical preparation flags and two unmatched design names. At the author's request, two

summaries by neutral code then showed each design type, self-reported Q_{TED} ratio, and completion/stop status, but no configuration assignment.

After these summaries, amendment 3 at 07:10 allowed import-setting changes to be recorded instead of rejecting the recalculation, submitted failed preparations with their available evidence, and added a separate relaxed model-judge reading. Three runs prepared under earlier instructions were judged again; earlier judgments remain. The author had not opened outcomes when this amendment was made. Three addenda corrected read-only access and scheduling.

At 09:45, after all runs had been judged and the operator had read their judgments, instructions were clarified: tool flags count only when they indicate an erroneous result; participation criteria beyond the task's mode definition are aids; condition 1 is undetermined only if recalculation yields no result. For Runs -013 and -018, the operator selected the STEP of the assessed design by content rather than unmatched filename and recalculated it. Ten runs were named for reconsideration; six were actually rejudged (-012, -013, -016, -018, -022, -023), while four (-004, -005, -007, -021) retained their first judgments. All records remain.

Kimi's usage limit prevented answers during this second assessment. At 10:06 the author authorized DeepSeek-V4.1-Flash, the experiment's own model, to supply missing answers while retaining valid ones. It answered for four runs and supports three accepted model-judge successes: -012, -018, and corrected -022. GPT-6-ASTRA dissented on -012 because the judging files lacked driven-peak displacement fields and on -018 because the registered refined-mesh solution was unavailable. These are material limits on judge independence and agreement.

Two neutral-code tables were frozen at 10:43:52 and unblinded at 10:44 at the author's request, without the signature required by the registration. Assessment A preserves the model-judge result. Assessment B applies the author's evidence-acceptance interpretation: count the mesh requirement as met and accept a complex driven response through resonance without a separate harmonic-temperature record. This interpretation was set after the neutral-code summaries and frozen before unblinding. The frozen counts were A = 1, 0, 4, 3, 2 and B = 1, 1, 4, 4, 5.

After unblinding, Run 20261023-022 (J19) was found to have been rejudged from only 161 of 326 files. Its correction was completed at 11:14: GLM and DeepSeek reassessed its complete judging files and accepted all conditions. The correction adds one success to Stage 3: Verify in both assessments; frozen files were not rewritten. The author subsequently adopted B for final reporting, giving 1, 1, 4, 5, 5 and 5/5 with the MRAF. A is retained as provenance/sensitivity and is not an equally weighted main conclusion. Final acceptance does not imply unanimous independent judges or make B the original preregistered primary endpoint.

S3.3 Frozen inputs and configurations

The registration froze the stage manifest, task, deliverables template, and 510254-byte baseline. Their SHA-256 values and the shared runtime inputs are retained here to identify the exact experiment.

Input	Bytes	SHA-256
Stage manifest: the five configurations as subsets of procedure version 2026-10-23.1	1187	24ea59eef6c77ebd317fd04a34e271044b40faf4459aec398fe5496fe9939b20
Task prompt, also the first message of every run	24891	e7aaf91bf5f34a16ffc826f8d9a8e20e68887c2d47d28473c2a9c3d00b663159
Deliverables template	9639	3c774b92df6c98b0a6d406d21d17afa2cced79ba26b794cbc98540c5e93620e9
Baseline model	510254	a9bbae7529b67e89d67ffb07cfcec9451af01f291a31e84cc61de6e3d8f1ba0a
Baseline geometry (STEP)*	54682	b66824b9540f749134b6a74679405d7790392ee5ae48414cb887a2a1f419ea81
Session hook, in the 20 runs with procedures*	567	166c0c9496d8d9866492972aae415a6fa76daa2f15322255f1a0998ea4004563
Continuation message, in all 25 runs*	529	d371f77e324d0f2720e4170ff3118b48136554bfb00d97790adccc8eae096238
Configuration of the literature tools, in all 25 runs*	1007	f16862330cb42865f9188c8654d88d171ea0cd5b4d50f566b199738ce182130b

* Not individually listed in the registration, but identical across the prepared folders concerned. The five cumulative procedure trees are identified by prefixes 1ca095a6e5a9, d4fa29282f7c, 4309b8af803c, f8fe689713ab, and 8069db30385b. Stage 3: Verify and With MRAF contained the same 35 files and scripts, differing only in six instruction texts concerning orchestration. All runs received the same prompt and literature tools; unsupplied methods could be implemented by the agent. Twenty runs had the skill-loading hook. Every main session used DeepSeek-V4.1-Flash-1017 at max reasoning in Claude Code 2.1.310; COMSOL was 6.3. Although a fixed neutral continuation message was prepared, no run received it or any other post-launch researcher message, and no session was resumed.

S3.4 Assignment and hosts

Seed 2660037412793056957 is the integer formed from the stage-manifest hash's first 16 hexadecimal digits. Python random.Random shuffled five copies of each configuration across Runs -004 to -028; seed + 1 shuffled launch order. Each run had a new non-preemptible Alibaba Cloud instance, 4 vCPUs and 16 GB, from one image. The first started at 00:44:31; the others at 00:48:00–00:50:42. All ended by 04:57, and copied-back files were hash-checked by 05:21 before hosts stopped. Judges' recalculations used a local COMSOL installation. Table S7 retains assignment and launch identities.

S3.5 Geometry and fabrication

The common task specified the silicon-on-insulator multi-user MEMS process (SOIMUMPs) and allowed only full-depth removal from its 25 μm silicon layer. Under rule (a), every silicon feature and etch gap had to be at least 2 μm. Under rule (b), a feature narrower than 6 μm required both ends merged into silicon at least 6 μm wide and span ≤500 μm, or one such supported end and span ≤100 μm. Anchors, supports, protected faces, connectivity, member count, outer widths, spacing, thickness, envelope, and mirrored tines were invariant. The scripted check tested solid count, removal-only/full-depth geometry, retained side faces and anchor regions, and mirror symmetry. Widths and spans were scanned every 0.25 μm along both axes at mid-thickness; sub-2 μm features required persistence over two scan lines. For this check, a tie joining two strips ended each strip: the scan line through the tie crossed at least 6 μm of continuous silicon (strip, tie, and strip), although the tie itself could be

narrower than 6 µm (4 µm in Run 20261023-012 and 2 µm in Run 20261023-025). Spans were measured between ties or between a tie and a slot end. Whether such ties limit curling as well as wider silicon does was not assessed. Runs -012 and -025 met rule (b) only under this reading. A post hoc screening of the scan records shows that, if their ties are counted as part of the narrow strips, the strips are about 740 µm long between slot ends, over the 500 µm limit. The final assessment would then give 1, 1, 4, 4, and 4 successes (1 of 5 against 4 of 5, P = 0.21), and the initial assessment 1, 0, 4, 4, and 1 (P = 1.0). The other successful designs are unaffected. Oblique/curved features were listed for judge inspection. Geometry checks were unchanged after the first freeze. Removed volume was reported, not scored.

S3.6 Recalculation and judging

Each delivered STEP was imported into a copy of the unmodified baseline. The model was solved at 293.15 K on its native mesh and at Finer, with Extra fine added when the minimum wall was below 3 µm or a Q_{TED} ratio was 3.6–4.4. Twenty modes were requested near 200 kHz; saved fields and participation records, not the requested frequency, determined operating-mode identity. Extra fine decided when available, otherwise Finer. Ratios straddling four were undetermined. Run -018 is the exception: prescribed refinement failed, and two judges accepted its Normal-mesh result over the third's objection. Table S7 reports that exception explicitly.

The evidence collector read baseline hashes, held-out/prior-run contacts, the selected design's geometry, mesh checks, complex responses, harmonic-temperature evidence, and cited deliverables without changing the run. Judging files used fixed-seed neutral codes and masked procedure/tool names. Tokens for masked procedures still revealed whether procedures were present, so masking was incomplete. Kimi and GLM answered independently; GPT-6-ASTRA resolved disagreements by majority. Eleven runs had three answers; no unresolved case reached the author. Files were fingerprinted before and after judging, and no judge modified them. The original condition 2 required a declared mesh tolerance, default 10%, and complex responses retaining thermal perturbation with displacement fields at the operating-mode peaks; half-power Q was not required. The final author assessment uses the interpretation in S3.2.

S3.7 Preliminary reading and prior information

Outside the recorded outcomes, GLM and DeepSeek made a quick reading of coded packages at 07:17–07:52, after amendment 3, with mesh differences set aside unless evidently wrong. This preliminary reading does not enter outcomes. The operator knew technical preparation flags; the author had seen two neutral-code summaries before setting the acceptance interpretation. Configuration unblinding occurred only after the outcome-table freeze, but this prior exposure means the interpretation was not specified before any outcome-related information was seen.

S3.8 Held-out publication

Open literature search was permitted. Retrieval of reference[15] was logged; using its geometry, dimensions, or values for candidate construction/ranking would void the run. The frozen scan searched tool-call inputs, literature maps, and reports for its DOI, title phrases, and author/year patterns. Because it did not cover tool-return text or all reasoning, it does not establish complete absence of exposure. Only Run -024 matched, in an open literature question; both judges found no use of the held-out device and did not void it. No other run matched. The two October 23 pilots underwent a broader read of their complete sessions, with no contact found.

S3.9 Statistics and accepted results

All 25 runs are retained. Each was judged to meet condition 1, including the disclosed -018 exception; complete evidence and the specified fabrication rules distinguish success. The final author assessment is B, while A preserves the initial model-judge assessment. The table reports both assessments, with the source-label mapping above and the post-unblinding correction.

Configuration	Successes, initial model-judge assessment	95% confidence interval	Successes, final author assessment	95% confidence interval	Run's own claim of complete deliverables
Without MRAF	1/5	0.005–0.716	1/5	0.005–0.716	2/5
Stage 1: Solve	0/5	0–0.522	1/5	0.005–0.716	1/5
Stage 2: Design	4/5	0.284–0.995	4/5	0.284–0.995	5/5
Stage 3: Verify	4/5	0.284–0.995	5/5	0.478–1	5/5
With MRAF	2/5	0.053–0.853	5/5	0.478–1	5/5
All configurations	11/25	—	16/25	—	18/25

Intervals are exact Clopper–Pearson 95% intervals, based on five runs per configuration. The registered main comparison was Without MRAF versus With MRAF using an unspecified exact test. A two-sided Fisher test was selected after unblinding: B gives 1/5 versus 5/5, P = 0.048; A gives 1/5 versus 2/5, P = 1.0. B's acceptance interpretation is distinct from the original criteria; small samples and cumulative configurations prevent attribution to one procedure. Failure to reject under A is not evidence of equivalence.

Time and tokens were secondary registered quantities, with operational definitions in a pre-outcome note that has no recorded author approval or formal amendment. Other action measures and the mean/SD displays were chosen after unblinding and are descriptive; no significance tests were performed. Several anticipated quantities, including baseline-pass time, the number of designs meeting the specified fabrication rules, and ratio geometric means/medians, were not reported. Resource counts describe different amounts of completed work and cannot establish efficiency by themselves.

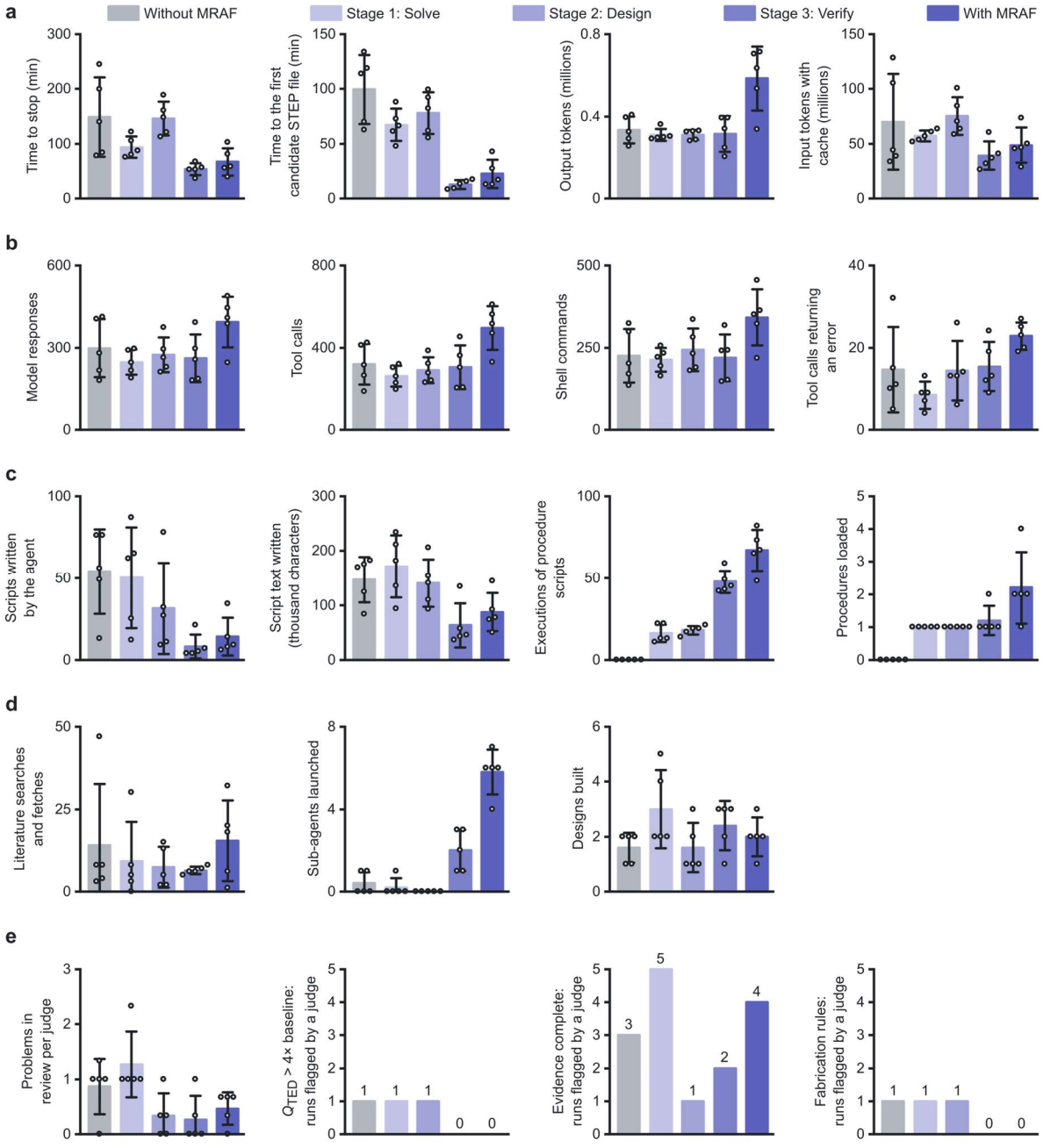
S3.10 Response-tool limit

Before launch, a pilot design near $Q_{\mathrm{TED}} = 2.6 \times 10^5$ failed the frozen driven solver's default 10^{-3} tolerance, while a design near 10^5 and the baseline converged. The author retained the tool unchanged, recording a limitation above about 2×10^5. Only Run -023's reported design reached this range (2.05

$\times 10^5$); it was not supplied that tool and its own responses were purely real. No reported design from a tool-supplied configuration exceeded 1.26×10^5. This documented implementation limit is not a physical bound on Q.

S3.11 Pilots excluded from the experiment

Five October 22 pilots used earlier prompts/procedures, heterogeneous local hardware, and no advance outcome rule; their review preceded revision and registration. After signing, October 23 Runs -029/-030 tested frozen inputs on local virtual machines for launch, host sizing, held-out handling, solver limits, and assessment. Judging-script rehearsals used pilots only. Pilots contribute no registered observations or evidence.



Supplementary Fig. S10 | Descriptive measures of the 25 registered runs. Configurations and colors follow Fig. 4 (n = 5 each). **(a)** Time and tokens. **(b)** Tool activity. **(c)** Scripts and procedure use; "Procedures loaded" counts only calls to the skill-loading tool, in main and sub-agent sessions, and most runs given procedures also opened every procedure's main file directly. **(d)** Literature access, sub-agents, and candidate construction. **(e)** Model reviews; the last three plots are review records, not final-author failure counts. In the 16 plots with circles, columns show means, error bars sample SD, and circles individual runs, shifted horizontally for visibility. Lower SD bars are clipped at zero. No significance tests were performed (Table S8).

Supplementary Table S8 | Principal descriptive measures for all 25 registered runs.

The table retains every run, in the order of Table S7. Time is from the first model reply to the written stop reason. Tokens and tool calls include the main and all sub-agent sessions, with each response counted once. Cache writes were zero throughout. Problems per judge is the mean number of the three conditions marked unmet or undetermined by answering model judges; it describes their review, not the final author assessment.

Configuration	Run	Time to stop (min)	Output tokens (10^6)	Uncached input (10^6)	Cache read (10^6)	Tool calls	Problems per judge
Without MRAF	20261023-004	139.5	0.291	0.229	43.8	314	1.00
	20261023-021	83.3	0.254	0.129	33.6	186	1.33
	20261023-006	244.4	0.405	0.199	128.1	423	1.00
	20261023-010	77.7	0.323	0.287	38.2	263	1.00
	20261023-016	199.1	0.395	0.175	104.9	410	0.00
Stage 1: Solve	20261023-023	121.4	0.303	0.155	63.5	303	1.00
	20261023-005	105.5	0.298	0.166	53.3	228	1.00
	20261023-009	79.4	0.294	0.171	51.3	196	1.00
	20261023-007	87.2	0.360	0.292	60.2	320	2.33
	20261023-013	75.0	0.289	0.162	56.6	258	1.00
Stage 2: Design	20261023-017	147.3	0.285	0.235	63.0	229	0.33
	20261023-020	120.6	0.328	0.206	100.3	384	0.00
	20261023-015	188.5	0.325	0.208	69.9	239	1.00
	20261023-018	112.1	0.281	0.159	58.1	276	0.33
	20261023-011	160.4	0.326	0.208	84.2	322	0.00
Stage 3: Verify	20261023-014	36.6	0.202	0.224	26.3	203	0.00
	20261023-022	67.1	0.338	0.282	36.5	301	0.00
	20261023-025	52.7	0.392	0.416	59.7	452	1.00
	20261023-008	52.8	0.392	0.357	40.3	361	0.33
	20261023-019	56.3	0.246	0.232	31.3	205	0.00
With MRAF	20261023-012	102.4	0.712	0.876	45.2	560	0.33
	20261023-026	57.8	0.535	0.648	49.1	468	0.67
	20261023-024	80.5	0.634	0.677	44.8	514	0.00
	20261023-028	54.9	0.704	0.821	72.5	608	0.67
	20261023-027	38.8	0.336	0.445	28.0	329	0.67

Configuration summaries used in Fig. 4c

Measure	Without MRAF	Stage 1: Solve	Stage 2: Design	Stage 3: Verify	With MRAF
Time to stop (min)	149 (72.6)	93.7 (19.4)	146 (30.9)	53.1 (10.9)	66.9 (24.8)
Output tokens (10^6)	0.334 (0.0654)	0.309 (0.0291)	0.309 (0.0238)	0.314 (0.0865)	0.584 (0.156)
Tool calls	319 (99.9)	261 (51.4)	290 (64.0)	304 (106)	496 (107)
Problems in review per judge	0.867 (0.506)	1.27 (0.596)	0.333 (0.408)	0.267 (0.435)	0.467 (0.298)

Values are mean (sample SD), n = 5 per configuration. Summary statistics use the archived per-run values. Output tokens were rounded to 0.001 million before aggregation; displayed run values are rounded. Time and tokens were registered secondary quantities, with the operational-definition limitation in Note S3.9. Tool calls and review counts were chosen after unblinding. No significance test was made, and runs performed different amounts of work supported by the required evidence. Lower time cannot alone establish higher efficiency.

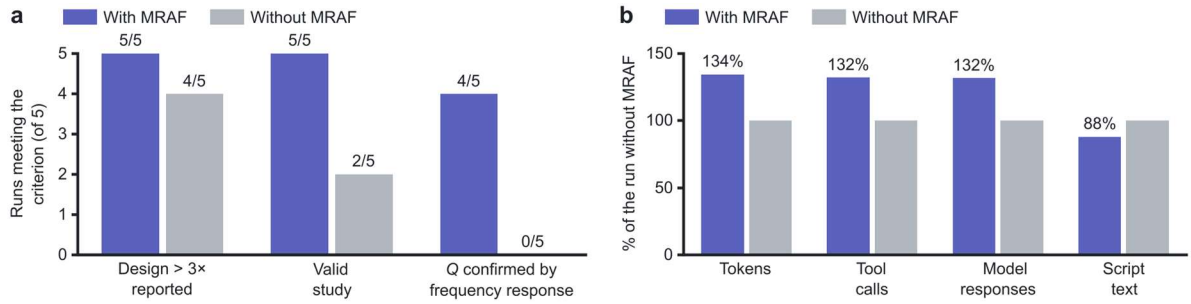
Exploratory action measures appear in Fig. S10. First STEP time depends on export practice, and supplied-script counts omit indirect calls; neither measure determines the outcome.

Supplementary Table S9 | Frozen procedures and the checks they supply.

Procedure	Scientific requirement	Supplied method	Configurations
Coordination	Separate target attainment, evidence completeness, and stop reason	Load the relevant procedures; collect outstanding solves and checks; full form assigns sub-agents and batch mode screening	Short entry: Stage 1: Solve–Stage 3: Verify; full orchestration: With MRAF
Literature	Trace constants, mechanisms, and process limits to sources and conditions	Source table with reported/derived/assumed labels; no copying a published device into candidates	Stage 2: Design–With MRAF
Hypotheses and fabrication	State mechanism, predicted *f*/*Q*/mode changes, and falsification before geometry	Immutable candidate record and process-rule check	Stage 2: Design–With MRAF
Geometry	Retain invariants, connectivity, symmetry, and minimum features	STEP construction, removal checks, and re-import	Stage 2: Design–With MRAF
COMSOL studies	Fresh solves, effective loss/load, matched modes and mesh evidence	Baseline re-solve, STEP import, eigenvalues; mesh/readout/driven tools added in Stage 3: Verify	Stage 1: Solve–With MRAF
Mode identification	Identify the intended mode from fields rather than scalar *Q* or frequency	Field export/images from Stage 1: Solve; participation and a fresh blind assessor from Stage 3: Verify	Stage 1: Solve–With MRAF
Reporting	Every result traces to a file and its qualification	Design status, tabular checks, and independent cross-check in full orchestration	Stage 3: Verify; With MRAF

Stage 1: Solve has 3 procedures and 7 command-line tools, Stage 2: Design has 6 and 9, and Stage 3: Verify has 7 and 16. With MRAF has the same scripts as Stage 3: Verify; six instruction texts add orchestration, not a new solver. The version is 2026-10-23.1, source commit f79078b4, full tree SHA-256 8069db30385bc92b1093109d455406ce628d8606b83fbfa16bd5e328345f463c. Twenty runs had the hook loading the entry skill; runs without the MRAF had no procedures or hook, but their agents could implement methods themselves.

The mode procedure's defaults require dominant transverse motion, antiphase tines, one lobe, and coherent cross-section motion: transverse amplitude at least twice other components; phase within 30°; local moving segments at least half the largest amplitude. These are supplied research checks, distinct from the registered operating-mode definition; the assessment clarification treated additional participation criteria as aids (Note S3.2). The driven tool checks effective load and thermoelastic coupling before solving; its known high-*Q* limitation remains in Note S3.10. Procedures state no expected candidate design or result.



582

583 **Supplementary Fig. S11 | Earlier paired runs retained as development records.** Ten archived runs form five
584 pairs with and without the MRAF (one run per condition; Table S10). **(a)** Cumulative counts meeting a reported
585 Q_{TED} ratio >3, study-validity checks, and half-power confirmation. **(b)** Median paired resource ratios; script-text
586 length is missing for Pair 5. Criteria were retrospective, judges knew the arm, and the second reviewer saw the
587 first judgment. These development runs completed different amounts of qualified work, are excluded from the
588 25-run comparison, and cannot separate procedure effects from prompt effects.

Supplementary Table S10 | Earlier paired runs and archival outcome counts.

These 35 developmental runs comprise two early batches, one preliminary pair, and the five pairs below. The historical threefold threshold, tasks, procedure versions, prompts, and assessment differ from the registered experiment. They are retained for transparency and method history; they are not pooled or treated as independent replication of the 25 runs.

(a) Five pairs

Pair	Model label	Reasoning setting	Command-line program	Run with the MRAF	Run without the MRAF
Pair 1	GLM-5.3-Flash	Default	Claude Code CLI	20260906-001	20260906-002
Pair 2	GLM-5.3-Flash	Default	Claude Code CLI	20260906-003	20260906-004
Pair 3	K3	High*	Claude Code CLI	20260906-005	20260906-006
Pair 4	GLM-5.3	Max*	Claude Code CLI	20260906-007	20260906-008
Pair 5	GPT-6-ASTRA	Ultra	Codex CLI	20260906-009	20260906-010

* High for K3 and Max for GLM both set a 32000-token thinking budget. Labels reproduce archived launch requests. Prompts for runs with the MRAF additionally supplied research guidance; paired contrasts therefore do not isolate procedure effects. Each configuration was observed once.

(b) Historical criteria and counts

Group	With MRAF	Runs	Design with Q_{TED} > 3× baseline reported	Valid study	Q confirmed by frequency response
Early runs, first batch	With	7	2/7	1/7	0/7
	Without	6	1/6	0/6	0/6
Early runs, second batch	With	5	0/5	0/5	0/5
	Without	5	2/5	1/5	0/5
Early runs, both batches	With	12	2/12	1/12	0/12
	Without	11	3/11	1/11	0/11
Preliminary pair	With	1	0/1	0/1	0/1
	Without	1	1/1	0/1	0/1
Five pairs (part a; Supplementary Fig. S11a)	With	5	5/5	5/5	4/5
	Without	5	4/5	2/5	0/5
All runs	With	18	7/18	6/18	4/18
	Without	17	8/17	3/17	0/17

Cumulative criteria were a recommended/reported ratio above three; a valid study with mode, baseline, and mesh checks; and half-power confirmation. Early batches lacked the threefold objective and used the minimum over four temperatures; later tasks used $\Delta T = 0$. Criteria were written on 5 October after outcomes were known, with the judge aware of the arm and the second reviewer aware of the first judgment. These are retrospective counts.

(c) Pair 5 designs cited in the development history

Run	Design	Q_{TED} ratio	Change in resonant frequency (%)	Status in the run's record
20260906-009	R1	26.064	−69.25	Reported, not recommended: no independent mode audit, and both of its driven calculations failed
20260906-009	R5	6.801	−54.13	Conditionally recommended
20260906-010	R2	2.452†	−26.03	No recommendation; the largest estimate among its designs matched to the baseline operating mode

In Pair 5 (GPT-6-ASTRA), Run 009 conditionally recommended its design R5 (the run's own numbering of its designs), with an eigenfrequency-derived Q_{TED} ratio of 6.801 and a 54.13% resonant frequency reduction (Supplementary Table S10c). Run 010 reported a largest matched estimate of 2.452 for its design R2, with a 26.03% resonant frequency reduction, and ended without a qualified recommendation.

† Not converged: Run 20260906-010's baseline mesh check exceeded its 1% frequency tolerance.

Supplementary Note S4. Figure provenance and scope

The archive retains prompts, frozen inputs, candidates, geometry and simulation models, numerical results, mode fields, reports, session records, registration/amendments, judging scripts, launch records, frozen outcome tables, and the subsequent correction. Figure data and scripts are kept with these records.

S4.1 Fig. 1

Registered Run 20261023-012 supplies the workflow examples. Its task was fixed before launch, no human message followed, and its 34 procedure files remained unchanged. Only outer strips moved in the first continuous-slot candidate, so it lacked an operating mode, prompting the tied-ladder design and a finer-mesh check; the reported tied-ladder candidate had a field-identified operating mode. Its mesh changes, −2.15% in frequency and +2.36% in Q_{TED}, met the run's declared 10% tolerance. These are run-level checks. Review, revised procedures, and the next task in panel c are schematic; the actual pre-experiment revision appears in Fig. 5a.

The blind mode check in panel a used saved fields without design names, targets, or Q values; each field file still recorded the imaginary part of the eigenfrequency, from which Q follows. In panel b, studies sharing a year and level are grouped, except Robin and this work, which are separated for readability. No method revision followed Run 20261023-012, and model weights were unchanged. Device views use the files of Run 20261023-012. Mode shapes are drawn from saved displacement fields normalized to their largest value; the drawn deflection is exaggerated for visibility and is not a vibration amplitude.

S4.2 Fig. 2

This is the guided August *c01* study, not Run 20261023-012. Geometries come from the retained STEP files and modes from saved displacement fields. Baseline/*c01* eigenvalues are from the original run; published-layout eigenvalues come from an earlier author calibration. Driven baseline curves came from later Run 20260906-001, and *c01* curves from author Run 20261002-001. These later contributions must not be credited to the original autonomous execution. The micrograph is from reference[15], cropped and rotated by 90°; brightness, contrast, and color were unchanged. Scaling used the published 740 μm tine gap because its printed scale bar implied dimensions about 21% larger. Modeled Q_{TED} includes only thermoelastic loss; the measured Q-factor belongs to the fabricated reference devices.

Geometries in panel b share one scale. Mode shapes in panel b are drawn from saved displacement fields normalized separately for each mode; the drawn deflection is exaggerated for visibility and is not a vibration amplitude. In panel c, points are calculated frequencies. In the table in panel c, *c01* has calculations only.

S4.3 Fig. 3

Panel a follows a coding-assistant check, procedure revision, and later use. Panels b–d contain distinct development series with changing conditions (Tables S5–S6) and do not form a controlled learning curve. Panel b includes cached input and sub-agent tokens; Fig. 4c uses output tokens only.

Baseline reproduction in panel d used a tolerance of 1% in Round 0–Round 6 and 5% in Round 7; times for Round 0–Round 2 are upper bounds. Script files written and executions of procedure scripts in panel d are separate counts. The series in panel b and in panels c and d share no runs. Some rounds could read earlier runs, and Round 7 changed the task and reasoning setting, so these development records are not independent repeats (Note S2).

S4.4 Fig. 4

All quantities concern the same 25 registered runs. The final author assessment supplies the success counts. Resource measures come from session records. Display choices and exploratory measures were selected after unblinding. Development runs contribute no observations or missing evidence.

Under the final author assessment, panel b shows 1, 1, 4, 5, and 5 successful runs. Assessment rules, initial judgments, masking limitations, and one post-unblinding correction are documented in Methods and Note S3; individual outcomes are in Table S7. Output tokens and tool calls in panel c include all sessions. Review problems in panel c count unmet or undetermined conditions (of three), averaged over model judges; they are separate from final author outcomes in panel b. Resource use reflects unequal amounts of completed work and does not establish efficiency. All configurations are shown in Fig. S10.

S4.5 Fig. 5

A pilot claimed completeness despite 1112 purely real driven-response rows. Commit b11bf570 revised three procedures; Kimi reviewed the changes before freeze 3277b693. The author's approval message is not retained. Run 20261023-012 read the revised COMSOL procedure 11 s after its task. Because the common prompt also rejected purely real responses, this chronology cannot isolate the revision's effect. Panel b separately traces the early *c01* task guidance to code and STEP geometry; the panels do not depict one registered run learning during execution. Dimensions in panel b are in μm.

S4.6 Supplementary figures

Fig. S1. Panels b and c show development records. The early device study used one session. In Stage 3: Verify and With MRAF, registered runs passed the blind mode check to a separate sub-agent, and With MRAF runs also passed mode screening to sub-agents; agents in any configuration could also start sub-agents themselves (Fig. S10d). Procedure versions are retained per run; Table S9 defines the registered configurations, compared only through the 25 runs in Fig. 4.

Fig. S2. The mode in panel b is from the saved displacement field, and the indicator in panel c from saved thermal-gradient fields. This single guided development study used its original mesh and calibrated model; it is neither a registered outcome nor independent validation of the mechanism.

Fig. S4. In panel a, attempt 4's half-power Q and eigenfrequency Q_{TED} agree within 0.3% (Note S1.3); this checks the calculation and does not show a device Q increase.

Fig. S5. In the author tests, four boundary cases were unscored. These software tests involved neither a research agent nor COMSOL.

Fig. S7. In panel a, normalized strip amplitudes are 0.08 and 0.97 in one tine. Q_{TED} and frequency values are at 293.15 K; the baseline has a resonant frequency of 231.0 kHz and $Q_{\mathrm{TED}} = 8045$.

Fig. S8. The first run launched alone; the rest followed its first reply. In panel a, bars start at launch but use durations from first reply to stop; reply lag was 20–55 s. Tables S7–S8 give outcomes, mesh exceptions, and resource definitions.

Fig. S10. The last three plots in panel e count runs flagged as unmet or undetermined by at least one model judge.

Fig. S11. Baselines and remaining task text were matched within pairs; the system arm added seven procedures and prompt guidance. Budgets and interventions varied, and Pairs 1–2 used different prompt/procedure versions despite the same model label. Tokens were 6–83% higher with the system; tool calls and model responses fell only in Pair 2.