Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Ruiqi Guo

ruiqiguo@um.edu.mo

University of Macau

Xinrong Ouyang

University of Macau

Yuan Wang

University of Macau https://orcid.org/0000-0003-3814-4520

Zishen Dai

University of Macau

Pan Zhang

Peking University https://orcid.org/0000-0001-9130-5982

Pui-In Mak

University of Macau

Rui P. Martins

State-Key Laboratory of Analog and Mixed-Signal VLSI, University of Macau

Jiawei Xu

University of Macau

Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11644084/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no conflict of interest

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.

Official email addresses of all authors

Xinrong Ouyang: yc67301@connect.um.edu.mo

Ruiqi Guo: ruiqiguo@um.edu.mo

Zishen Dai: mc66309@connect.um.edu.mo

Jiawei Xu: jiaweixu@um.edu.mo

Pan Zhang: panzhang1989@pku.edu.cn

Yuan Wang: yuanwang@um.edu.mo

Rui Paulo Martins: rmartins@um.edu.mo

Pui-In Mak: pimak@um.edu.mo

*** Corresponding authors**

Ruiqi Guo

Email: ruiqiguo@um.edu.mo

Postal address: State Key Laboratory of Analog and Mixed-Signal VLSI, AMSV, Research

Building N21, 3/F, University of Macau, Avenida da Universidade, Taipa, Macau, China
Telephone: +853 8822 8287

Pan Zhang
Email: panzhang1989@pku.edu.cn
Postal address: School of Integrated Circuits, Peking University, Beijing, China
Telephone: +8610 6275 5145

Yuan Wang
Email: yuanwang@um.edu.mo
Postal address: State Key Laboratory of Analog and Mixed-Signal VLSI, AMSV, Research
Building N21, 3/F, University of Macau, Avenida da Universidade, Taipa, Macau, China
Telephone: +853 8822 9039

Abstract

Microelectromechanical systems (MEMS) research requires translating device concepts into geometries with physically meaningful behavior under specified structural constraints. Although research agents increasingly automate scientific workflows, successful numerical execution alone does not establish operating-mode identity and dynamic behavior, identify loss mechanisms, or verify geometric compliance. Here we introduce the MEMS Research Agent Framework, which brings agent-based research and experience-informed procedure refinement into an expert-guided MEMS workflow. Agents execute investigations using reusable instructions and scripts, while experts define objectives, assess physical interpretations, and review methodological changes. In a researcher-guided study of a silicon double-ended tuning-fork resonator, the proposed framework carries a literature-derived thermoelastic-dissipation hypothesis through geometry construction, coupled-field simulation, and operating-mode identification. The selected segmented-slot design achieves a modeled thermoelastic quality factor approximately 5.2 times the baseline at a resonant frequency 20% lower. The slotting mechanism is consistent with a published experimental study reporting a 5.9-fold increase in measured quality factor in a different geometry. Development histories trace a procedural route toward self-improvement: recorded failures inform expert-reviewed changes to research instructions and diagnostic checks reused in subsequent runs. A registered comparison across 25 runs evaluates research outcomes and evidence completeness under five frozen configurations. The final author assessment gives 1/5 successful runs without the proposed framework and 5/5 with it; the initial model-judge assessment gives 1/5 and 2/5, respectively, reflecting differences in evidence acceptance. These studies demonstrate the feasibility of delegating parts of MEMS inverse-design iterations to agents, laying a foundation for self-improving research automation in device design and discovery.

Introduction

Autonomous research agents are now capable of generating hypotheses, executing experiments, and composing manuscripts within closed computational loops. The AI Scientist exemplifies such coordination in a bounded machine-learning setting[1], while other frameworks extend this capacity to dataset-driven inquiry, with increasing degrees of human copiloting as complexity rises[2]. These circumscribed cycles exemplify a regime of full delegation, in which agents execute complete research cycles without sustained human intervention; on levels adapted from the laboratory-autonomy scale proposed for materials research[3], we assign this to level 4 (L4) (Supplementary Table S1). Closed research loops have likewise been reported in mathematics[4], chemistry[5], and biomedicine[6], each operating within a supplied problem, evaluator, or laboratory environment. In the MEMS domain, however, comparable agent-based research remains comparatively scarce, insofar as the computational output must be validated against the intended device behavior, which is a judgment that cannot be delegated to the agent alone.

Within MEMS, both the physical model and the evaluative criteria are contingent upon a device's operating principle and intended function. For resonators, this validation gap is particularly acute: the quality factor is constrained by competing dissipation mechanisms, including thermoelastic dissipation (TED), anchor loss, surface effects, and fluidic damping, whose relative contributions depend on geometry, material, and environment[7]. TED originates from thermal gradients induced by inhomogeneous strain during vibration[8–10] and is highly sensitive to structural dimensions and modal shape; consequently, a geometry that reduces TED may inadvertently shift modal order, amplify anchor loss, or violate fabrication constraints. A converged simulation therefore does not, in itself, constitute validated behavior: the governing equations may omit dominant physics, boundary conditions may idealize as-fabricated reality, and the mesh may fail to resolve the thermal gradients that drive

TED. Studies of mechanical resonators have demonstrated learned structural prediction and search[11,12], data-efficient optimization[13], and workflows that compare predicted performance with measurements obtained from fabricated devices[14]. A research investigation must therefore employ such capabilities not merely to establish numerical superiority but to test a physical hypothesis and evaluate whether the computed output corresponds to the intended physical behavior under the relevant operating conditions. These difficulties concern not only which device to design, but also how a research agent decides what to simulate, what evidence to inspect, and whether a result supports a physical interpretation. Experience can therefore inform revisions to the research procedures themselves, including the diagnostic checks used to assess subsequent results. This raises the possibility of MEMS research agents that accumulate explicit methodological experience through repeated investigations.

We use a silicon double-ended tuning-fork (DETF) resonator to demonstrate the MEMS Research Agent Framework (MRAF) within a concrete MEMS design problem. As reported in the literature[15], slotting the mechanical structure of the resonator can effectively regulate the TED-associated Q-factor while preserving the device size factor and baseline features. Under prescribed footprint and structural constraints, the agent must translate this mechanism into candidate geometries, perform coupled-field simulations, and determine whether the calculated improvements belong to the intended operating mode. This case therefore allows us to examine how a proposed research agent framework carries established physical knowledge through a design investigation and how difficulties encountered during research inform revisions to its procedures. In summary, three distinctive attributes of the proposed MRAF are emphasized: (i) a closed-loop autonomous research methodology that remains grounded in practice; (ii) an experience-informed procedure-refinement process in which agents draft reusable instructions and scripts for expert review, providing an initial

implementation of a virtual research apprentice; and (iii) relatively restrained proposal generation that supports real-world implementation.

Research experience may alter agent procedures, as several existing frameworks have shown. ChemAgent maintains and updates experiential knowledge for chemical reasoning[16], whereas SkillPCF couples reusable design skills with simulator-guided revision of its skill bank[17]. AIDE² renders the code of research agents itself an object of revision, such that evaluated rewrites determine the implementation edited in subsequent rounds[18]. These frameworks motivate a distinction between improving candidate designs and refining the research procedures used to formulate hypotheses, execute simulations, inspect evidence, and qualify conclusions. We investigate this second process in MEMS as experience-informed procedure refinement: agents draft methodological revisions, experts review their adoption, and subsequent tasks or repeated runs reuse the retained procedures.

The MRAF starts with a researcher-defined objective and physical constraints (Fig. 1a). Here, a mission denotes the overarching research objective, instantiated in this study by improving the operating-mode thermoelastic quality factor of a silicon DETF resonator under specified structural constraints. A task is a bounded assignment within that mission, with defined objectives and completion conditions; a run is one agent execution of a task. Multiple tasks and repeated runs may therefore belong to the same mission. A central agent coordinates six stages—hypothesis formulation, literature search, device geometry, MEMS simulation, result analysis, and agent update—and invokes external tools. Simulation and result analysis iterate within a task until completion criteria are met or the researcher stops the task. The procedures require candidates and their expected and refuting results to be recorded before solving, operating modes to be identified from saved displacement fields, and decisions to be accompanied by reasons. Researchers supply physical guidance, interpret results, and determine which modifications to research procedures should be retained. We

designate this division of responsibility as L3.5, an author-defined operational category in which agents execute the research cycle and researchers approve procedure revisions between runs (Supplementary Table S1). Fig. 1b places L3.5 among published research systems; these levels describe task scope.

Between tasks, difficulties and reviewed outcomes inform revisions that require human approval before use in later steps (Fig. 1c). “Agent Update” refers to expert-reviewed changes to reusable instructions, scripts, and diagnostic requirements; language-model weights remain fixed throughout. These internal methodological revisions are distinct from the external outcome criteria used to assess the frozen configurations. Methods are intended to remain fixed within a task; historical amendments are documented in Methods and Supplementary Note S1. Figs. 2 and 3 document the earlier development of the MRAF, using earlier procedures, and only the registered experiment (Fig. 4) tests its effect. Supplementary Fig. S1 details the research roles, tools, and outputs.

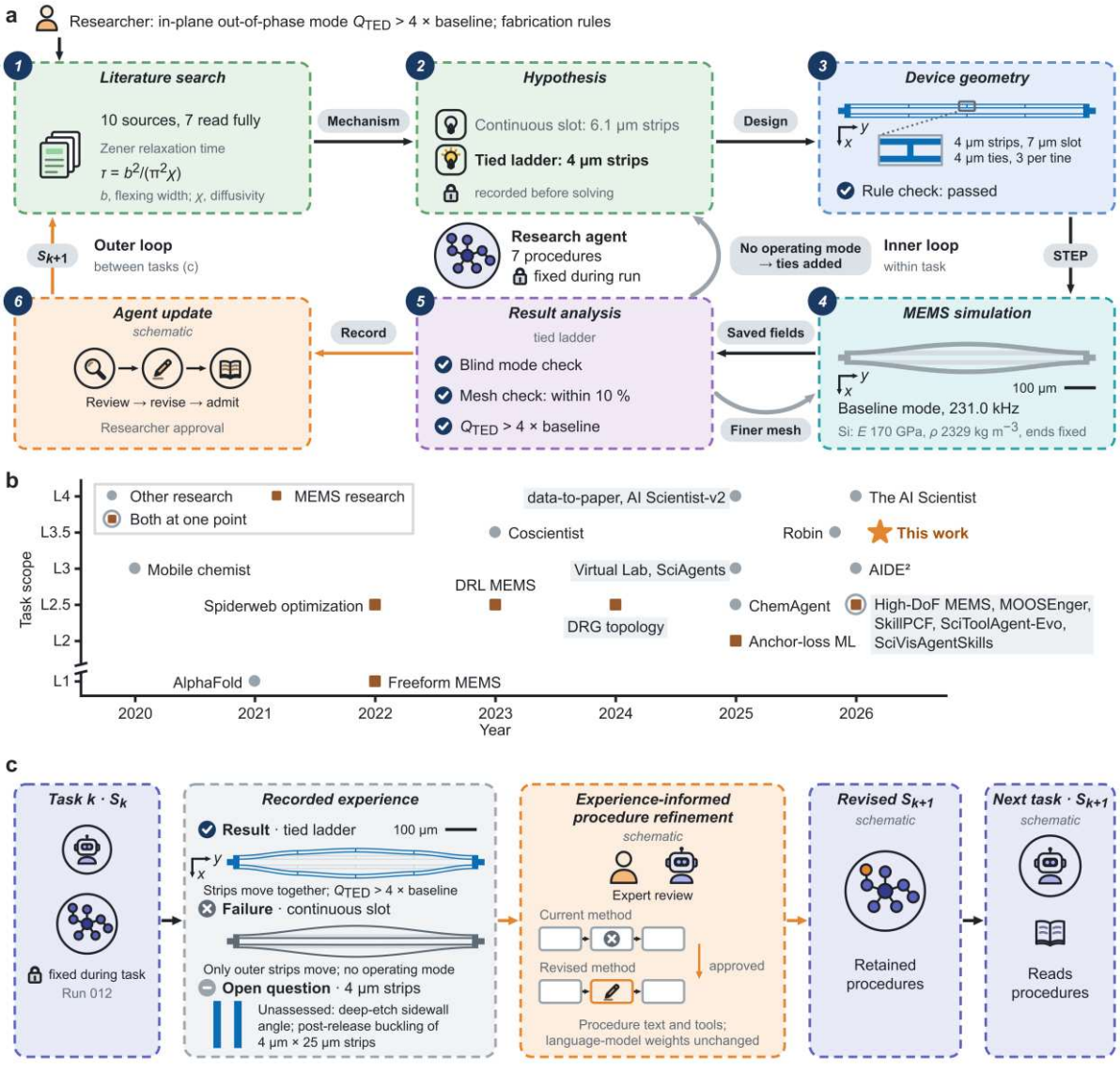


Fig. 1. Research organization and method revision under expert review. **(a)** Stages 1–5 show the recorded sequence of registered Run 20261023-012 (n = 1), in which the inner loop assesses candidates within the task, and stage 6 illustrates expert-reviewed revision of reusable procedures between tasks. **(b)** Published research systems[1,2,5,11–14,16–18,23,31,32,37–44] by year and author-assigned task scope (Supplementary Table S1); L3.5 is an author-defined division of responsibility, and the levels provide qualitative context rather than a calibrated performance scale. **(c)** Procedures S_k remain fixed within a task, and reviewed experience informs S_{k+1} for the next task; the first two boxes show records from Run 20261023-012. Mode shapes are normalized and exaggerated for visibility. Q_{TED}, thermoelastic quality factor; E, Young's modulus; ρ, density.

This study presents the MRAF and addresses two questions: whether a research agent operating at L3.5 autonomy can carry a physical hypothesis through candidate construction and qualified computational comparison, and how recorded difficulties are translated into expert-reviewed procedural revisions and documented in subsequent use within this MEMS research setting. The device case is provided by an early, researcher-guided design-discovery study of a silicon MEMS double-ended tuning-fork resonator, selected because a fabricated and measured device of the same architecture was available[15], which links thermoelastic-dissipation hypotheses to geometry, modal interpretation, and quality-factor–frequency trade-offs. A separate development sequence—consisting of dozens of iterations conducted over more than three months—examines reviewed procedural change, its subsequent use, recurring failures, and documented resource demands. A registered experiment of 25 runs, in five configurations that differ in supplied procedures, scripts, and orchestration instructions, tests whether these configurations change the research process and the completeness of its evidence.

Results

Case study using a MEMS DETF resonator

Quality factor affects the noise performance and stability of MEMS resonators. We examined whether material removal from a DETF resonator could increase the thermoelastic quality factor, Q_{TED}, of its fundamental in-plane flexural mode, in which the two tines vibrate out of phase. The baseline corresponds to the unslotted design fabricated and measured by Wang et al.[15]. Slots modify heat paths and strain fields[19], as investigated previously in slotted tuning forks and microbeams[20,21]. This early study, Run 20260814-012, used the August 2026 procedures. Researchers who knew the published layout supplied the objective and fixed geometry constraints, and asked the agent to consider one relief segment per band of high

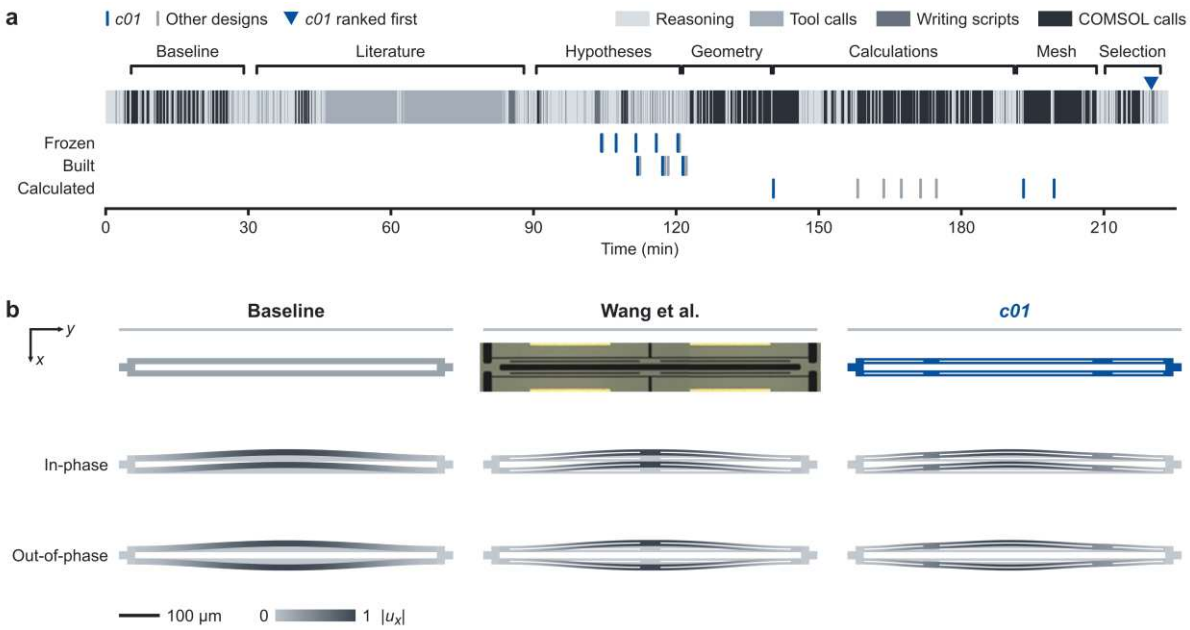
thermoelastic loss (Fig. 5b; Supplementary Fig. S2). It is separate from the registered experiment, whose task prescribed no design direction.

The event log spans 217 min; the session timeline spans 224 min (Fig. 2a). No human message arrived during the run (Methods). The agent first verified the supplied baseline model by re-solving it (resonant frequency of the operating mode = 231.046 kHz, Q_{TED} = 8045); thereafter, it conducted a literature search and proposed six candidates across five structural families, fixing this candidate set before any candidate was computed. The set was revised after geometry checks. It examined saved fields, refined *c01*, and recommended it for human review. After calculating *c01*, it read earlier-run records and copied checking scripts; its task allowed this, whereas the registered experiment would void such a run. The authors later corrected one *c01* field export (Supplementary Note S1).

The selected design, designated *c01*, incorporates three slot segments per tine (Fig. 2b; Supplementary Table S2). Following the guidance, the agent placed one 3 μm slot in each high-loss band, leaving a 6 μm strip on either side. Its out-of-phase mode is located at 184.743 kHz with Q_{TED} = 41499, corresponding to approximately 5.2 times the baseline value at a resonant frequency 20% lower. The in-phase mode shifts from 217.394 kHz (Q_{TED} = 9453) to 177.992 kHz (Q_{TED} = 43176). The run identified the operating mode from saved fields; the authors assigned the in-phase mode and drew the mode shapes without new solves. The Q_{TED} ratio compares the candidate's operating mode with the baseline at the same temperature. Four other candidates gave ratios of 2.0 or less; one gave a larger ratio but was not ranked because its operating mode could not be identified under every required condition (Supplementary Fig. S3). Author recalculations on finer meshes changed the ratio from 5.16 to 5.05 (Supplementary Tables S3 and S4). The later driven responses in Fig. 2c were calculated by another run for the baseline and by the authors for *c01*; half-power quality factors agreed with Q_{TED} within 0.6%.

A second development investigation produced *c02*, comprising two rows of twelve short slots per tine (Supplementary Table S2; Supplementary Fig. S4). This design yields a resonant frequency of 207.780 kHz and Q_{TED} = 31449, corresponding to 3.9 times the baseline value at a resonant frequency 10% lower; its calculated driven response exhibits the matching quality factor within 1%. Upon recalculation at two finer mesh settings, the ratio was 3.93 and 3.97, respectively. The two designs thus exhibit different trade-offs between quality factor and resonant frequency. This *c02* is distinct from candidate 2 of the first study (Supplementary Table S2).

By contrast, fabricated devices furnish an experimental reference. The slotted device reported in reference[15], which incorporates two 320 μm slots per tine, exhibits a measured Q-factor of 43000 at 100.7 kHz, corresponding to 5.9 times that of the unslotted device at a resonant frequency 54% lower. Application of the same model to this layout yields 110.1 kHz and Q_{TED} = 55202, equivalent to 6.9 times the baseline at a resonant frequency 52% lower. The authors made this calculation before the recorded run, as a calibration check (Run 20260814-001). Calculation and measurement agree in direction, but the calculated Q_{TED} includes thermoelastic loss alone: its ratio exceeds the measured 5.9 and the published simulated 5.1, and its resonant frequency is approximately 9% above the measurement. On the other hand, for the in-phase mode, the calculated Q_{TED} = 59265, which is approximately five times the measured Q-factor of 12000. This discrepancy is consistent with additional losses, including anchor loss, because the support reactions of the in-phase mode do not cancel.



	f (kHz)		Q		f (kHz)		Q	
	Calculated	Measured	Calculated	Measured	Calculated	Measured	Calculated	Measured
Baseline	217.4	196.9	9453	5500	231.0	220.2	8045	7300
Wang et al.	108.7	99.6	59265	12000	110.1	100.7	55202	43000
c01	178.0		43176		184.7		41499	

Fig. 2. An early, researcher-guided device study and the selected design *c01*. **(a)** Candidate freezing, geometry construction, and coupled calculations over 224 min in Run 20260814-012, which used the August 2026 procedures before the registered experiment (Fig. 4); the triangle marks the message ranking *c01* first. **(b)** Baseline, the published slotted device (micrograph from reference[15]), and *c01*, with calculated in-phase and out-of-phase (operating) mode shapes from saved displacement fields, each normalized and exaggerated for visibility and shaded by transverse displacement $|u_x|$. **(c)** Calculated driven-response power from checks made after the study, normalized to its peak, with the dotted line at half power; table entries distinguish calculated eigenfrequency and Q_{TED} from published measured frequency and total Q-factor. Calculations are at 293.15 K, with one solve per value (n = 1); Q_{TED} includes thermoelastic loss alone (Methods; Supplementary Note S1).

Expert-reviewed method revision and its later use

The development records trace the procedural component of self-improvement, linking recorded difficulties to methodological revisions and subsequent reuse. The development record divides into two periods according to who wrote the methods (Supplementary Table S5). Researchers wrote them during the ten August runs of Fig. 3b, ending with the device study of Fig. 2. Thereafter, agents drafted changes for researcher approval, while researchers continued infrastructure repairs. Task conditions, models, and sub-agent use changed; the historical counts describe development, not the system effect tested in Fig. 4.

One earlier method history connects the two device studies (Supplementary Fig. S5; Note S1). During a researcher-directed check, a frequency-response tool saved zero responses because its load did not enter the equations (Fig. 3a). The revision required checking the effective load before solving and after reopening, and diagnosing a zero response before declaring a mode absent. The plan had prior approval; no separate sign-off followed the check, although a further export fix was approved. The later *c02* study read the revision and completed four driven sweeps with half-power quality factors within 1% of Q_{TED} (Supplementary Fig. S4).

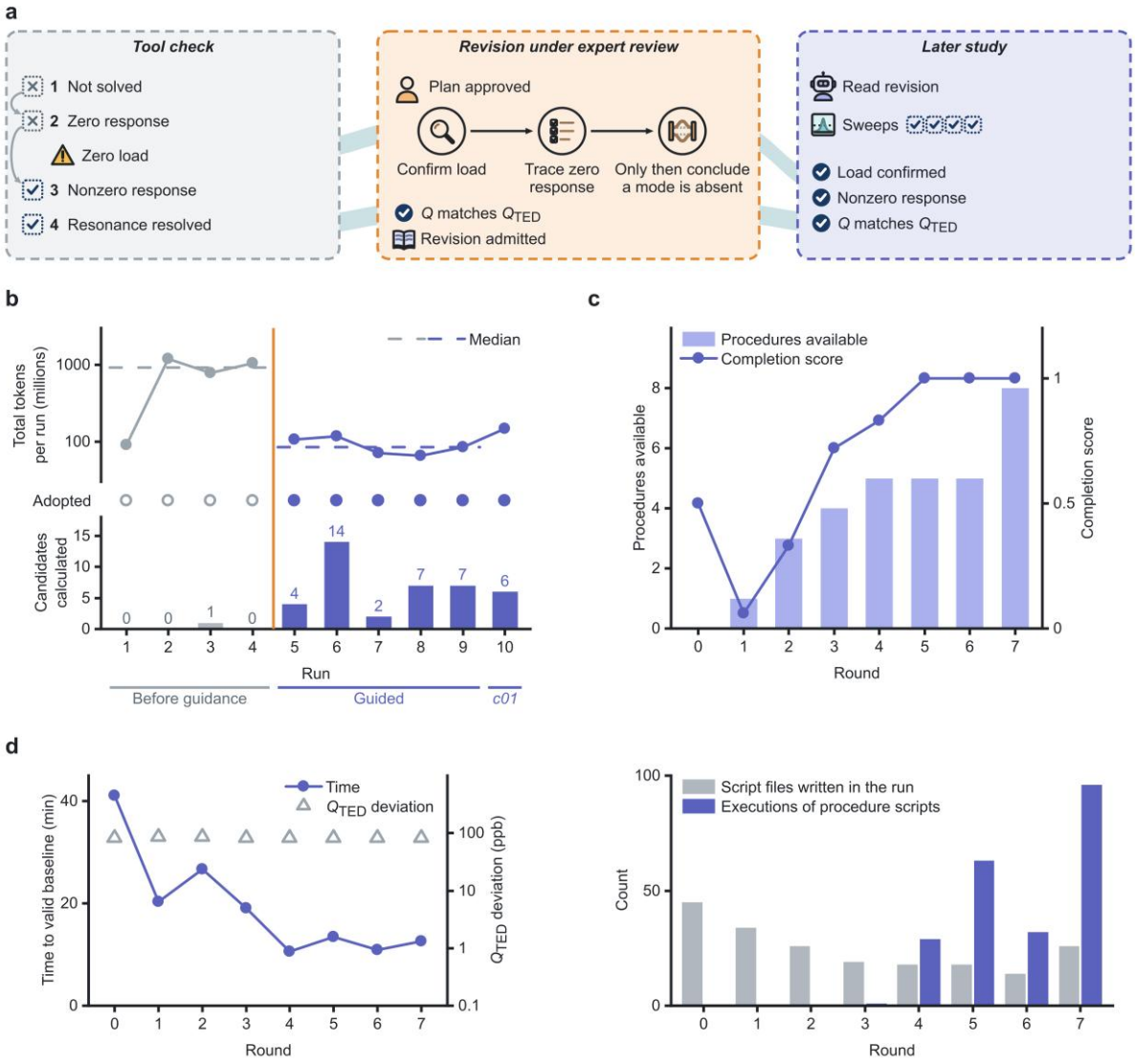


Fig. 3. Method development before the registered experiment. **(a)** A researcher-directed tool check revealed an ineffective load, so load verification and zero-response diagnosis were added to the procedure, which a later run used for four driven sweeps; Q is obtained from the half-power bandwidth and Q_{TED} from the eigenfrequency study. **(b)** Total tokens per run (input, including cached input, and output across main and sub-agent sessions; logarithmic axis), adoption of candidates in the run registry (filled markers), and candidates calculated in ten development runs in launch order (n = 10); dashed lines show medians of Runs 1–4 and 5–9, and Run 10 is the *c01* study. **(c)** Eight rounds of one task, one run per round, with the procedures available and the completion score, the mean of six 0–1 criteria (Supplementary Table S6). **(d)** Time to first baseline reproduction, maximum Q_{TED} deviation (parts per billion) from prompt values supplied by earlier scripted solves of the same model, and counts of script files written and of procedure-script executions in the same rounds.

Eight developmental rounds, addressing the common objective of increasing baseline Q_{TED} through material removal, were conducted between 29 September and 1 October 2026 (Round 0–Round 7; Fig. 3c, d; Supplementary Fig. S6; Note S2). Round 0 was conducted without any procedures. Separate agent sessions reviewed each round, and a researcher approved changes implemented by a coding assistant. Round 1, Round 2, and Round 4 also introduced procedures already present in the system (Supplementary Note S2). All eight recorded decisions were approvals; each round represents one reviewed run.

The first three modifications concern the reports. A finalization procedure followed Round 2, after which reports and final checks were completed. A source-extraction check followed Round 4, but scanned a different folder from the original defect; non-recurrence rests on later reviews. A numeric-range check passed in Round 6 and Round 7 only because their reports contained no readable range, while reviews still found untraceable values. The associated single-value check approved after Round 6 was not implemented in Round 7 (Supplementary Note S2).

The fourth modification pertains to the operating mode (Supplementary Figs. S6 and S7). Round 6 ranked a full-length slot at 10.0 times baseline Q_{TED} although its strips did not move together; its report recommended a 2.1-fold alternative without correcting the ranking. The revision required saved-field identification before ranking, making the agent's assessment procedure itself an object of refinement. Round 7 ranked two segmented designs, the second at 6.1 times baseline Q_{TED}, but followed only one line per tine without comparing strips; its mesh check, using tolerances chosen after solving, failed for the first design, whose 13.1-fold value is therefore not converged. It also changed the task, procedures, reasoning setting, and scripts, and read reference[15] before freezing candidates. Its result cannot be attributed to the mode check or called blind discovery.

Fig. 3 and Supplementary Table S6 report completion scores, baseline re-solve times, and script counts. The checks examine the text of the package, while reviews establish whether problems recur. Passing a check does not establish physical validity. Round 1–Round 4 could read prior folders, and Round 7 changed several settings, so the rounds are not independent repeats or a controlled learning curve. A further check approved after Round 7 has no subsequent test in this sequence.

Research outputs and resource demands across configurations

To test whether the procedures change the way a run reaches the research goal and the completeness of the evidence it leaves, we registered a controlled experiment before any run started (Methods). The 25 runs shared the task, baseline, language model, and host setup. Five runs received each cumulative configuration: Without MRAF; Stage 1: Solve; Stage 2: Design; Stage 3: Verify; and With MRAF (Fig. 4a). These are subsets of one frozen version, not revision rounds; agents could implement methods not supplied. Each run used a fresh session and host without human input. All stopped within 4.2 h of the first launch, before the individual 8 h limit (Supplementary Fig. S8).

Each run was assessed on one of its reported designs against three requirements: operating-mode Q_{TED} above four times baseline, mesh and driven-response evidence, and geometry satisfying fabrication constraints (Methods). No run was void. The final author assessment counts the mesh requirement as met and accepts a complex driven response through the operating-mode resonance without a separate harmonic-temperature record. These rules were set after neutral-code summaries were seen and frozen before configuration unblinding. Across the five configurations, 1, 1, 4, 5, and 5 of 5 runs succeeded, respectively: 16 of 25 overall, including all five with the MRAF (Fig. 4b; Supplementary Fig. S9). The main comparison is 1 of 5 without the system against 5 of 5 with it (two-sided Fisher's exact

P = 0.048). The test was selected after unblinding and applies to the final author assessment, which differs from the original registered interpretation (Supplementary Note S3).

All 25 runs were judged to meet the Q_{TED} condition. For Run 20261023-018, two judges accepted the run's own mesh after the registered-mesh re-solve failed. Both assessments' counts include a correction to Run 20261023-022, a Stage 3: Verify run, after unblinding, when complete judging files became available. Five runs differ between initial and final assessments because of mesh checks or separate harmonic-temperature records. For the main comparison, the initial model-judge assessment gives 1 of 5 runs without the system and 2 of 5 with it (two-sided Fisher's exact P = 1.0; Supplementary Note S3).

The unsuccessful runs fell short at different steps (Supplementary Table S7). Seven of the eight failures without the system or in Stage 1 lacked a driven sweep retaining thermoelastic loss; the eighth cut an outer side face. The Stage 2 failure cut about 10 µm into both anchor regions. The runs' own claims of completion, 2, 1, 5, 5, and 5 of 5, did not enter the outcomes. During re-judging, DeepSeek-V4.1-Flash, the model that ran the experiment, replaced a judge whose usage limit had been reached in four runs. Its answers contributed to three initial model-judge successes, two disputed by the third judge (Methods). Run 20261023-012, the earliest-launched successful run with the system, is the registered illustrative case in Figs. 1 and 5a; its 2:1 judgment is included in the five runs, with missing driven-peak fields noted by the dissenting judge.

Means (SD) without and with the MRAF were 149 (72.6) and 66.9 (24.8) min from the first model reply to the recorded stop, 0.334 (0.0654) and 0.584 (0.156) million output tokens, 319 (99.9) and 496 (107) tool calls, and 0.867 (0.506) and 0.467 (0.298) problems in review per judge, over five runs each (Fig. 4c). The tool calls, the problems per judge, and the further measures of all five configurations in Supplementary Fig. S10 were not part of the registered analysis; they were counted after unblinding and are descriptive. The resource

counts must be interpreted alongside the outcomes reported in Fig. 4b; they do not compare equivalent amounts of completed work. Full counts and definitions are provided in Supplementary Table S8 and Note S3.

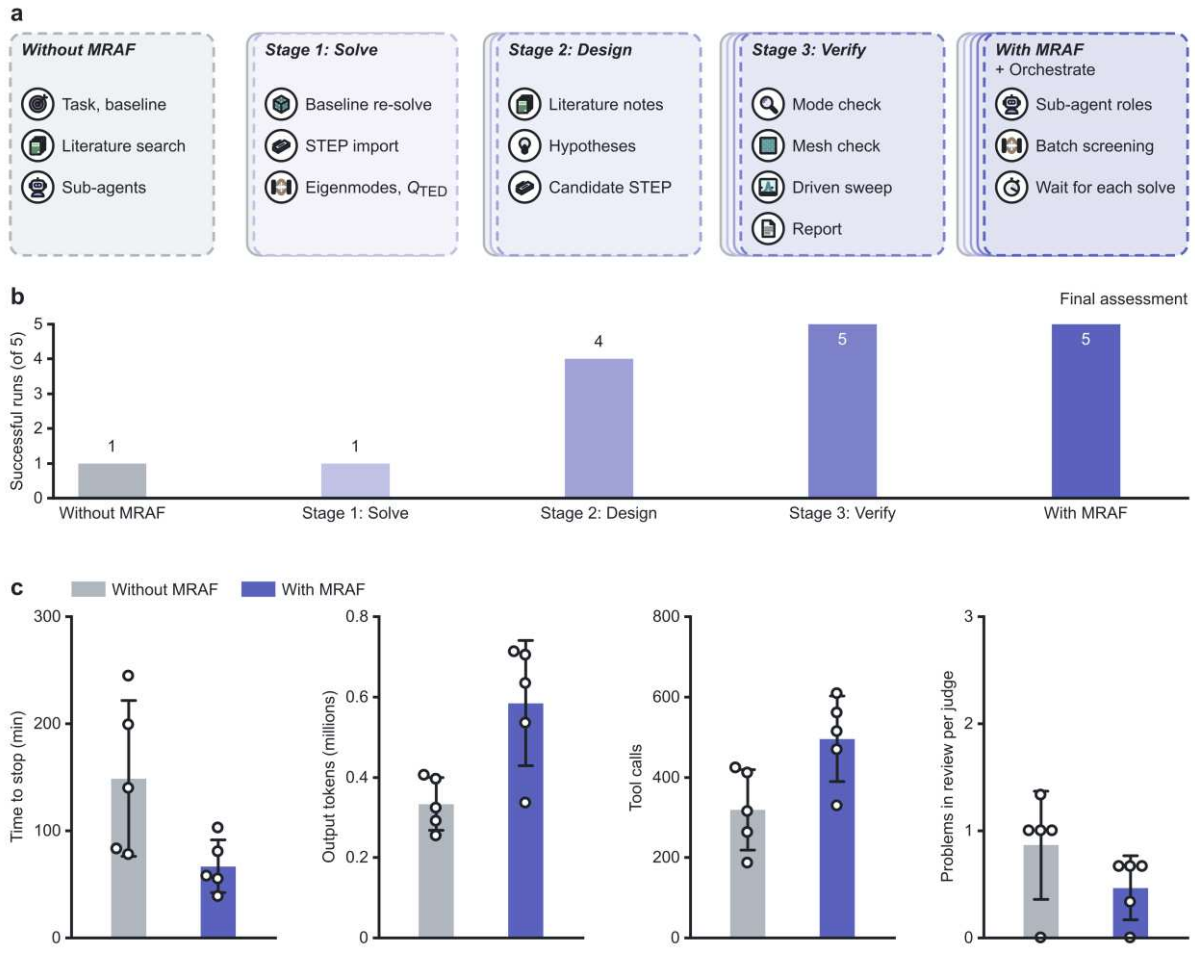


Fig. 4. Outcomes and resource use in 25 registered runs (five per configuration) that shared the task, baseline, and model and ran in fresh sessions without post-launch human input. **(a)** The five cumulative configurations progressively add procedures and, in the full system, extended instructions and orchestration (Methods; Supplementary Table S9). **(b)** Successful runs under the final author assessment, in which success requires operating-mode Q_{TED} above four times baseline, mesh and driven-response evidence, and geometry meeting the specified fabrication rules (Methods; Supplementary Note S3 and Table S7). **(c)** Time from first reply to stop, output tokens, tool calls, and review problems per judge in the registered comparison without and with the MRAF; columns show means, error bars sample SD (n − 1), and open circles individual runs (n = 5 per group).

Discussion

Agents carry out and record each study, while experts decide which procedural revisions subsequent tasks will use. The case study using a MEMS DETF resonator demonstrates how an expert-defined objective may be carried through structural design, coupled-field calculation, and physical interpretation: the agent compared six geometries under shared constraints, selected *c01* on the basis of the Q_{TED} of its operating mode, and the slotting trend it identified corresponds to that measured in fabricated devices[15]. Both designs enhance the modeled thermoelastic quality factor at the expense of a reduced resonant frequency; the calculated Q_{TED} accounts solely for thermoelastic dissipation, whereas the measured total Q-factor incorporates additional losses[7]. Across the eight rounds of one task, reviewed problems led to approved procedures used in later rounds. Related systems accumulate skills during inference[22] or compress research time[6]; here, changes drafted by agents are retained for subsequent use after expert review. These records illustrate an interpretable route toward self-improving MEMS research: experience is retained as explicit, revisable procedures whose subsequent use can be inspected. Sustained gains and transfer to other missions remain questions for future evaluation.

At this prototype stage, three features of the MRAF constitute its contribution. First, revised methods enter practice through expert review, although the inspected records lack the author's approval message for the pre-experiment freeze (Fig. 5a). Second, under the final author assessment, successful outcomes require a recomputed operating-mode Q_{TED} above four times the baseline, accepted mesh and driven-response evidence retaining thermoelastic loss, and geometry meeting the specified fabrication rules. In the 25-run experiment, the final author assessment found that all 5 of 5 runs with the MRAF met these conditions, compared with 1 of 5 without it (Fig. 4b); the initial model-judge assessment is retained in

Supplementary Table S7. Third, the records document the cost: runs with the system stopped sooner but used more output tokens and tool calls (Fig. 4c; Supplementary Table S8).

The study has several limitations. It concerns one device and one task, with five runs per configuration. The cumulative configurations cannot isolate individual procedures or test their revision; only large differences can be detected, and a nonsignificant result does not establish equivalence. All 25 runs were judged to meet the Q_{TED} condition, so it did not separate the configurations; resonant frequency was recorded but not constrained, and reported designs were 8%–70% lower in frequency than the baseline (Supplementary Table S7). The rules of the final author assessment were fixed before unblinding but after the author had twice inspected neutral-code summaries. It counts the mesh requirement as met and accepts a complex driven response through the operating-mode resonance without separately recorded harmonic temperature perturbation. The main-comparison Fisher P = 0.048 depends on this interpretation and a test chosen after unblinding. Two successes, including Run 20261023-012, met fabrication rule (b) only under the scripted check's treatment of narrow ties joining two strips (Methods; Supplementary Note S3.5). The only success without the system was assessed on a design other than the one its run recommended (Methods). Unblinding proceeded without the registered signature, and one outcome was corrected afterward. The executing model replaced an unavailable judge in four runs, contributing to three initial-assessment successes, including one with the system. One run was accepted on its own mesh after the registered-mesh re-solve failed. Judges did not receive configuration labels, but masked procedure names could reveal whether procedures were present. The guided development case replaced the registered main case (Methods). The frozen frequency-response tool has a documented convergence limitation above $Q_{\mathrm{TED}} \approx 2 \times 10^5$; no reported design from a run supplied with it reached this range. All candidates used a previously calibrated baseline, and sensitivity to thermal conductivity, thermal expansion, thermal

boundaries, etched linewidth, and sidewall angle was not tested. Development-round checks do not establish physical validity; all eight review decisions were approvals, leaving review selectivity untested. Calculated gains concern thermoelastic loss alone; under matched conditions with additive losses, the total quality factor cannot exceed Q_{TED}. Held-out-device retrieval was guarded only in registered runs, and the guard does not cover what the language models may have learned about reference[15] during training (Methods; Supplementary Notes S1 and S3). Complete records accompany every registered run; we do not claim that repeated runs would follow the same course.

Extending the system to fabrication and measurement would permit observed device behavior to test both the proposed structures and the procedures employed to assess them. The arrangement could also be tested in other device families requiring physical judgment and in MEMS interface circuitry. An AI-assisted platform that iterates over the device and its ASIC together, toward the system-level design of intelligent microsystems on a chip, is the direction of our future work.

Materials and methods

Research tasks and the MRAF

Comparative outcome estimates are based on the 25 registered runs. The early device study, research and revision histories, and eight developmental rounds explain method construction (Supplementary Notes S1–S2 and S4). The guided device study, Run 20260814-012, was executed in Codex CLI on DeepSeek-V4-Flash at maximum reasoning after the requested GPT-5.6-Sol was rejected. Permission requests were decided by the automatic reviewer. Its task carried guidance written by researchers who knew reference[15], without a held-out rule. After the initial instruction to continue, no human message arrived before its report. Later calculations supporting this study are detailed in Supplementary Note S1.

Seven roles cover coordination, literature, hypotheses, geometry, simulation, mode interpretation, and reporting. Procedures are stored as agent skills; frozen version 2026-10-23.1 was written to contain no expected design or result, and its COMSOL scripts work on model copies. The agent that loads the entry procedure coordinates the study. In Stage 3: Verify and With MRAF, it passes the blind mode check to a separate sub-agent that does not see the main session; this sub-agent works from saved displacement fields and per-segment amplitude and phase records. In With MRAF, it also passes mode screening and literature search to sub-agents, and in some runs command execution and the cross-check of deliverables. The 20 runs supplied procedures had a hook instructing the agent to load the entry skill by name, without retrieval, at session start and after restart or context compaction; the five runs without the MRAF had none (Supplementary Table S9). L3.5 is our author-defined division of responsibility, adapted from reference[3].

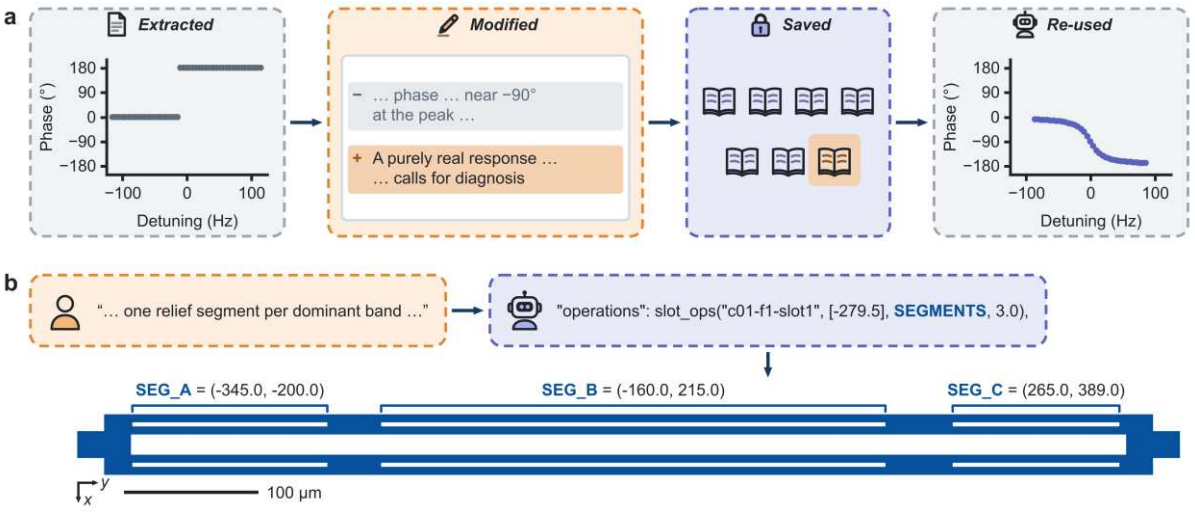


Fig. 5. Method revision and device design. **(a)** A trial run's purely real driven response prompted a diagnostic requirement, and a later registered run read the revised procedure and produced a complex response with phase crossing −90°; each plot shows one baseline sweep from a different run (n = 1 per plot), with detuning relative to 231.046 kHz. **(b)** For the early *c01* study (Fig. 2), the researcher suggested one relief segment per loss band, and the agent defined three full-depth slots per tine, each 3 µm wide, with lengths of 145, 375, and 124 µm, before building the STEP geometry.

Development archive

Inspected messages from Round 0–Round 7 name GLM-5.3. Procedures and rules changed; Round 7 also changed the task and reasoning setting. Reviews, analogous to the reviewer and critic agents of earlier systems[2,23], recorded severity and subsequent recurrence. These records and earlier paired runs are development history, excluded from effectiveness estimates (Supplementary Note S2; Tables S6 and S10; Fig. S11).

The revision in Fig. 5a followed five pilots whose driven calculations or reporting were incomplete. Run 20261022-004 claimed complete deliverables despite 1112 purely real driven-response rows. Commit b11bf570 revised three procedures, changed one script, and added another; a purely real response required diagnosis. Kimi reviewed the revision before freeze 3277b693; the author's approval message is not retained. Run 20261023-012 read the revised COMSOL procedure 11 s after its task. The common registered prompt also rejected purely real responses, so this chronology does not establish a revision effect. The earlier tool revision of Fig. 3a is documented in Supplementary Note S1.3.

Registered experiment and blind assessment

On 23 October 2026, before any run started, the author signed the registration, an internal document whose SHA-256 hash is stored in the launch records. Five cumulative configurations—Without MRAF; Stage 1: Solve; Stage 2: Design; Stage 3: Verify; and With MRAF—each comprised five new sessions with the same task, baseline, DeepSeek-V4.1-Flash at maximum reasoning, and Claude Code 2.1.310. The task did not supply the held-out slotted-device design from reference[15]. A fixed seed assigned configurations and launch order across new Alibaba Cloud hosts, each with 4 vCPUs, 16 GB memory, and one COMSOL 6.3 image. All launched within 6.2 min from 00:44 UTC on 24 October. Unsupplied methods could be implemented by the agent. No run received post-launch human input or was resumed. Unfinished runs would count as unsuccessful. All ended within 4.2 h of the first

launch, before the individual 8 h limit; all 25 were retained without host replacement (Supplementary Note S3).

The original registered outcome was to assess the final recommended design, or the highest reported Q_{TED} ratio design if none was recommended. In practice, when a run listed the comparisons that met the target, the frozen evidence-collection script took the design in the first such row of the run's comparison table. In four runs this was not the recommended design: in Runs 20261023-005 and -009, no design had a driven sweep retaining thermoelastic loss, and Runs 20261023-013 and -016 were counted as successes in the final assessment on designs they had not recommended (Supplementary Note S3.1). The original registered success criteria required: (1) recalculated operating-mode Q_{TED} > 4× matched baseline at $\Delta T = 0$; (2) the run's two-level mesh check within its tolerance declared before solving (default 10%), and complex driven responses of baseline and design through resonance, with non-zero harmonic temperature and operating-mode displacement fields at the peaks; and (3) full-depth material removal preserving task invariants and meeting the two fabrication rules that the task specified for the silicon-on-insulator multi-user MEMS process (SOIMUMPs): (a) minimum silicon feature/gap 2 μm; (b) features narrower than 6 μm spanning ≤ 500 μm with each end merging into silicon ≥ 6 μm wide, or ≤ 100 μm with one such end (the scripted check counted a narrow tie joining two strips as such an end; Supplementary Note S3.5). Baseline alteration, reading another run, or using the held-out device to construct/rank candidates would void a run while retaining it in the denominator; none was void. Retrieval alone was recorded, and frequency changes and self-reported completeness did not decide success.

The conditions distinguish improvement[24] from valid completion[25,26], with separate evidence interpretations[27], violation rules[28,29], mesh verification[30], mode identification[12], and

driven-response checking[11,31]. Separate recalculation[32] and model judges[33,34] implement a three-condition rubric[35].

Scripts frozen before the first run ended recalculated baseline and design at Finer (h_{auto} = 3), adding Extra fine (h_{auto} = 2) for walls <3 μm or ratios 3.6–4.4; Extra fine decided when available. Ratios straddling four were undetermined. Twenty modes were requested near 200 kHz; fields determined operating-mode identity. Kimi K3.1 Pro and GLM-5.3 judged neutral-coded files independently; GPT-6-ASTRA resolved disagreements by majority. No unresolved disagreement reached the author. Masked procedure names could reveal whether procedures were present, limiting blinding.

Three amendments preceded opening outcomes. The first two repaired evidence collection and judge calls. After seeing two neutral-code summaries, the author allowed import-setting changes to be recorded, failed preparations to reach judges with available evidence, and a separate relaxed model-judge reading. After the operator (AI coding assistants that prepared the judging files and ran the judges under the author's direction) had read judgments, clarified instructions treated tool flags as failures only when results showed errors, condition 1 as undetermined only when no result existed, and mode criteria beyond the task definition as reference aids. Six runs were rejudged; earlier judgments remain. For Runs 20261023-013 and -018, the operator selected STEP files by content after filename mismatches. Run -018's prescribed refinement failed; two judges accepted its Normal-mesh result over GPT-6-ASTRA's objection, using a Q_{TED} ratio of 6.57 against the matched Normal-mesh baseline.

When Kimi reached its usage limit, the author authorized DeepSeek-V4.1-Flash, the experiment's own model, to supply missing judgments for Runs -012, -018, -023, and corrected -022. Its answers supported three initial-assessment successes. For Run -012, GPT-

6-ASTRA dissented because the judging files lacked driven-peak displacement fields. This substitution limits judge independence (Supplementary Table S7).

The initial model-judge and author's outcome tables were frozen at 10:43 UTC on 24 October, before configuration unblinding at 10:44 without the prescribed signature. The author's interpretation, set after neutral-code summaries, counts the mesh requirement as met and accepts a complex driven resonance without a separate harmonic-temperature record. It differs from both the initial and relaxed model-judge readings. After unblinding, Run 20261023-022 was found to have been judged from 161 of 326 files; GLM-5.3 and DeepSeek-V4.1-Flash reassessed the complete files and accepted all conditions. Frozen tables remain unchanged. The author subsequently adopted this corrected final assessment for reporting: 1, 1, 4, 5, and 5 successes. It is not the original preregistered interpretation or unanimous independent adjudication. Amendment timings, preliminary readings, and masking/held-out-search limitations remain in Supplementary Note S3. The preliminary reading does not enter outcomes; the held-out scan did not cover all tool returns or reasoning.

The frozen driven-response tool had a documented convergence limitation above about $Q_{\mathrm{TED}} = 2 \times 10^5$; it was retained unchanged. Its pilot checks and scope are in Supplementary Note S3.10.

Statistics and descriptive measures

We report successes out of five with exact Clopper–Pearson 95% intervals (Supplementary Note S3.9). The registered main comparison was without versus with the MRAF; a two-sided Fisher test was selected after unblinding. The final assessment gives P = 0.048; the initial assessment remains in Supplementary Note S3 and Table S7. Intermediate configurations are descriptive and do not isolate procedure effects.

Run 20261023-012 is the earliest-launched successful run with the MRAF under both readings, as required for the registered illustrative case, and appears in Figs. 1 and 5a. The author selected the early *c01* study for Fig. 2 after unblinding; this does not change the analysis population.

Time and tokens were registered secondary quantities; their pre-outcome definitions lack recorded formal approval. Time spans first reply to stop. Tokens and tool calls include main and sub-agent sessions; Fig. 4c uses output tokens. Problems per judge average unmet/undetermined conditions across the final responding judges. Tool calls, review counts, and mean/SD displays were selected after unblinding without significance tests. Token counts, reported as cost or efficiency measures in earlier agent studies[16,24], here describe unequal amounts of completed work and do not establish efficiency (Supplementary Table S8).

Selected baseline and material model

The early study copied a 668089-byte baseline; the registered runs used a 510254-byte file with identical settings, including supports, thermal boundaries, coupling, and temperature sweep. Archived selection records and Supplementary Note S3 identify hashes. Tine width, thickness, and gap were 15, 25, and 15 μm. Material parameters were $E = 170$ GPa, $\nu = 0.28$, $\rho = 2329$ kg m^{-3}, $C_p = 700$ J kg^{-1} K^{-1}, effective $k = 70$ W m^{-1} K^{-1} (130 in the material record), and $\alpha = 3.0 \times 10^{-6}$ K^{-1}. Before the early study, α was adjusted from 2.6×10^{-6} K^{-1} to approach the published simulated baseline Q_{TED} of 8100 (reference[15]). All candidates used this calibrated model. Temperatures were 293.15 K + {0,20,40,60} K; registered comparisons and unspecified figure values use $\Delta T = 0$.

Geometry construction and coupled-field calculations

Baseline re-solves had to reproduce supplied frequency and Q at every temperature within 1% in the early study/Round 0–Round 6 and 5% in Round 7/registered runs. Candidate definitions preceded geometry imports and solves; the early study froze its candidate set five times, whereas the registered task required immutable records. Outer shape, anchors, boundary faces, connectivity, member count, and mirror symmetry were invariant. The six full-depth *c01* slots were 3 μm wide and removed 15.58% of baseline volume (Fig. 5b; Supplementary Table S2).

For a lightly damped complex eigenfrequency f, $Q_{\mathrm{TED}} = |\mathrm{Re}(f)|/[2|\mathrm{Im}(f)|]$ (reference[8]). The early study ranked the minimum candidate/baseline Q_{TED} ratio over four temperatures, requiring the fundamental operating mode at each: transverse in-plane motion, opposing tines, and single-lobe bending. Post-run tracking used the modal assurance criterion[36]. Authors made in-phase assignments and mode drawings after the runs; refinement calculations are retained in Supplementary Tables S3–S4, and author Run 20261002-002 recalculated *c02* at the same two finer mesh settings.

Driven $Q = f_{\mathrm{res}}/(f_+ - f_-)$, where $|H|^2$ falls to half its peak at f±. The authors' post-run *c01* check applied 1 nN to each facing inner wall, oppositely for the operating mode or together for the in-phase mode, and read the difference/sum of transverse midspan displacement on the inner strips with thermoelastic coupling retained. Responses are normalized to their own peaks.

Published-device comparison and figure data

Measured Q-factor values and the micrograph come from reference[15]; unslotted-device frequencies were read approximately from its plot axes. The authors modeled the published geometry from the same reference before the early study, as a calibration check. Baseline

driven curves came from later Run 20260906-001 and *c01* curves from author Run 20261002-001. For additive losses, $Q_{\text{total}}^{-1} = Q_{\text{TED}}^{-1} + Q_{\text{anchor}}^{-1} + Q_{\text{gas}}^{-1} + Q_{\text{surface}}^{-1} + \cdots$ (reference[7]). Fig. 2a spans 224 min from the first reply after the researcher's “go on” to the report; stage boundaries follow file creation times. Fig. 1a,c use Run 20261023-012, but their remaining elements, including boxes 3–5 of panel c, are schematic. Apart from the judges' recalculations and author *c01* driven responses, figures required no new COMSOL solve. Supplementary Note S4 distinguishes archived fields, subsequent calculations, and figure processing; these contributions are not credited to autonomous execution of the early study.

Data availability

The data and code that support the findings of this study are available from the corresponding authors upon reasonable request.

Acknowledgments

This work was supported by the Science and Technology Development Fund, Macau SAR (FDCT), 0078/2026/ITP1, the University of Macau, MYRG/GRG2024/00025/IME, the Funding Scheme for Scientific Research Platforms, National Platforms, 0001/2025/NRP.

Author contributions

X.O. and R.G. contributed equally to this work. R.G. conceived the MRAF concept, Y.W. selected the device case and designed the comparison with and without MRAF, and X.O. designed the registered experiment. X.O. wrote the manuscript, and R.G., P.Z., J.X. and Y.W. revised it. X.O. developed the MEMS Research Agent Framework, performed the experiments, and analyzed the data. Z.D. conducted the literature survey, revised Figs. 1-3, and assisted with the experiments. Y.W. provided the fabricated and measured resonator data. R.M. and P.-I.M. supervised the project. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that they have no competing interests.

Supplementary information accompanies the manuscript on the Microsystems & Nanoengineering website (http://www.nature.com/micronano).

References

1. Lu, C. et al. Towards end-to-end automation of AI research. Nature 651, 914–919 (2026).
2. Ifargan, T., Hafner, L., Kern, M., Alcalay, O. & Kishony, R. Autonomous LLM-driven research — from data to human-verifiable research papers. NEJM AI 2, AIoa2400555 (2025).
3. Hung, L. et al. Autonomous laboratories for accelerated materials discovery: a community survey and practical insights. Digit. Discov. 3, 1273–1279 (2024).
4. Romera-Paredes, B. et al. Mathematical discoveries from program search with large language models. Nature 625, 468–475 (2024).
5. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. Nature 624, 570–578 (2023).
6. Mitchener, L. et al. Kosmos: an AI scientist for autonomous discovery. Preprint at https://arxiv.org/abs/2511.02824 (2025).
7. Miller, J. M. L. et al. Effective quality factor tuning mechanisms in micromechanical resonators. Appl. Phys. Rev. 5, 041307 (2018).
8. Lifshitz, R. & Roukes, M. L. Thermoelastic damping in micro- and nanomechanical systems. Phys. Rev. B 61, 5600–5609 (2000).
9. Zener, C. Internal friction in solids. I. Theory of internal friction in reeds. Phys. Rev. 52, 230–235 (1937).
10. Zener, C. Internal friction in solids II. General theory of thermoelastic internal friction. Phys. Rev. 53, 90–99 (1938).

11. Guo, R. et al. Deep learning for non-parameterized MEMS structural design. Microsyst. Nanoeng. 8, 91 (2022).

12. Chen, C. et al. Machine learning-driven discovery of high-performance MEMS disk resonator gyroscope structural topologies. Microsyst. Nanoeng. 10, 161 (2024).

13. Shin, D. et al. Spiderweb nanomechanical resonators via Bayesian optimization: inspired by nature and guided by machine learning. Adv. Mater. 34, 2106248 (2022).

14. Jia, Z. et al. Deep learning-driven performance prediction and design of high-DoF MEMS resonators. Microsyst. Nanoeng. 12, 194 (2026).

15. Wang, Y. et al. A resonant Lorentz-force magnetometer featuring slotted double-ended tuning fork capable of operating in a bias magnetic field. J. Microelectromech. Syst. 30, 958–967 (2021).

16. Tang, X. et al. ChemAgent: self-updating memories in large language models improves chemical reasoning. In Proc. 13th International Conference on Learning Representations (eds Yue, Y. et al.) 100922–100958 (ICLR, 2025).

17. Chen, S., Shu, T. & Ren, S. Learning design skills as memory policies for agentic photonic inverse design. Knowl.-Based Syst. 351, 116726 (2026).

18. Srikanth, D., Zhao, B., Xu, D., Wu, Y. & Jiang, Z. Recursive self-improvement of AI research agents. Preprint at https://arxiv.org/abs/2609.26457 (2026).

19. Candler, R. N. et al. Impact of geometry on thermoelastic dissipation in micromechanical resonant beams. J. Microelectromech. Syst. 15, 927–934 (2006).

20. Zega, V., Frangi, A., Guercilena, A. & Gattere, G. Analysis of frequency stability and thermoelastic effects for slotted tuning fork MEMS resonators. Sensors 18, 2157 (2018).

21. Cheng, R., Song, P., Chen, J., Huang, Q. & Zhang, L. Thermoelastic damping suppression method of micro-beam resonators with basically constant resonant frequency. J. Therm. Stresses 45, 960–973 (2022).

22. Huang, X. et al. CASCADE: cumulative agentic skill creation through autonomous development and evolution. Preprint at https://arxiv.org/abs/2512.23880 (2025).

23. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. Nature 646, 716–723 (2025).

24. Huang, Q., Vora, J., Liang, P. & Leskovec, J. MLAgentBench: evaluating language agents on machine learning experimentation. In Proc. 41st International Conference on Machine Learning 20271–20309 (PMLR, 2024).

25. Chen, Z. et al. ScienceAgentBench: toward rigorous assessment of language agents for data-driven scientific discovery. In Proc. 13th International Conference on Learning Representations (eds Yue, Y. et al.) 96934–96990 (ICLR, 2025).

26. Soleymanibrojeni, M. et al. GENIUS: an agentic AI framework for autonomous design and execution of simulation protocols. Commun. Mater. 7, 115 (2026).

27. Meissner, J. A., Kuboth, P. & Meisner, J. Evolving transition-state search with agentic large language models. npj Comput. Mater. 12, 298 (2026).

28. Chan, J. S. et al. MLE-bench: evaluating machine learning agents on machine learning engineering. In Proc. 13th International Conference on Learning Representations (eds Yue, Y. et al.) 50466–50494 (ICLR, 2025).

29. Mandal, I. et al. Evaluating large language model agents for automation of atomic force microscopy. Nat. Commun. 16, 9104 (2025).

30. Oberkampf, W. L. & Trucano, T. G. Verification and validation in computational fluid dynamics. Prog. Aerosp. Sci. 38, 209–272 (2002).

31. Sui, F., Yue, W., Zhang, Z., Guo, R. & Lin, L. Trial-and-error learning for MEMS structural design enabled by deep reinforcement learning. In Proc. 2023 IEEE 36th International Conference on Micro Electro Mechanical Systems (MEMS) 503–506 (IEEE, 2023).

32. Ghareeb, A. E. et al. A multi-agent system for automating scientific discovery. Nature 655, 497–505 (2026).

33. Xu, F. F. et al. TheAgentCompany: benchmarking LLM agents on consequential real world tasks. In Advances in Neural Information Processing Systems Vol. 38 10291–10321 (NeurIPS, 2025).

685 34. Zhuge, M. et al. Agent-as-a-Judge: evaluate agents with agents. In Proc. 42nd International
686 Conference on Machine Learning 80569–80611 (PMLR, 2025).
687 35. Starace, G. et al. PaperBench: evaluating AI's ability to replicate AI research. In Proc. 42nd
688 International Conference on Machine Learning 56843–56873 (PMLR, 2025).
689 36. Allemang, R. J. The modal assurance criterion – twenty years of use and abuse. Sound Vib. 37,
690 14–21 (2003).
691 37. Abulawi, Z., Li, M., Giudicelli, G., Liu, Y. & Permann, C. Deploying frontier agentic technology
692 in MOOSEnger, a multiphysics-capable AI assistant. Preprint at
693 https://arxiv.org/abs/2608.15881 (2026).
694 38. Tang, Y. et al. SciToolAgent-Evo: an ontology-aware self-evolving agent for open-world
695 scientific tool acquisition. Preprint at https://arxiv.org/abs/2607.28692 (2026).
696 39. Ai, K., Miao, H., Tang, K., Liu, S. & Wang, C. SciVisAgentSkills: design and evaluation of agent
697 skills for scientific data analysis and visualization. Preprint at https://arxiv.org/abs/2606.05525
698 (2026).
699 40. Burger, B. et al. A mobile robotic chemist. Nature 583, 237–241 (2020).
700 41. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–
701 589 (2021).
702 42. Ghafarollahi, A. & Buehler, M. J. SciAgents: automating scientific discovery through bioinspired
703 multi-agent intelligent graph reasoning. Adv. Mater. 37, 2413523 (2025).
704 43. Yamada, Y. et al. The AI Scientist-v2: workshop-level automated scientific discovery via agentic
705 tree search. Preprint at https://arxiv.org/abs/2504.08066 (2025).
706 44. Chen, W., Xiao, Y., Zhu, K., Li, L. & Wu, G. Machine-learning-assisted anchor loss reduction of
707 MEMS resonator with one-dimensional phononic crystal tether. IEEE Trans. Electron Devices
708 72, 3127–3132 (2025).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf