

Autonomous hardware security evaluation from reconnaissance to physical validation

Raphael Viera

`raphael.viera@emse.fr`

Mines Saint-Etienne, CEA, Centre CMP

Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11061085/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Autonomous hardware security evaluation from reconnaissance to physical validation

Raphael Viera

Mines Saint-Etienne, CEA, Leti, Centre CMP, F-13541 Gardanne, France
raphael.viera@emse.fr

Abstract Physical hardware-security evaluation requires both target-specific reasoning and control of laboratory instruments. This study presents a large language model (LLM) agent that links reconnaissance, vulnerability hypotheses, laser fault-injection experiments and reporting through human approval gates. The agent reverse-engineers firmware and instruction behaviour, combines this evidence with documentation and chip imagery, and generates bounded experiments with predefined rules for classifying outcomes. On an ATmega328P microcontroller, agent-derived experiments produced a skipped processor load instruction, acceptance of an incorrect personal identification number (PIN), and the same two-byte Advanced Encryption Standard (AES) encrypted-output difference in 14 trials, with 29 paired fault-free controls. Photon-emission imaging narrowed the physical search region. A second campaign on an Artix-7 field-programmable gate array (FPGA) tested transfer to programmable logic using firmware-binary analysis, optical-to-layout registration and cryptographic-output classification. Recorded agent and operator decisions connect observations to experimental revisions and documented conclusions, demonstrating an AI-driven workflow across two hardware architectures.

Physical hardware-security evaluation begins before an instrument command: identify the security property and enforcing operation, select a plausible fault mechanism, derive where and when to intervene, and define an observation that separates that mechanism from protocol or instrumentation error. Laser benches can move a stage, synchronize a pulse and read a device, but useful evaluation remains target-specific [1], [2]. Electromagnetic fault injection poses the same evaluation problem through a different physical interaction and resulting computational errors [3], [4].

Increasingly, laboratory agents connect large language models (LLMs) to documentation, code and physical tools [5], [6]. An agentic X-ray scientist planned scans, interpreted observations and adapted to an unexpected motor offset [7]. In hardware security, this loop must test adversarial but authorized hypotheses, distinguish security effects from instrumentation error, and respect external bounds the agent cannot change. Laser-fault detection and LLM-assisted system-on-chip (SoC) security research address complementary aspects of this problem [8], [9]. The open problem is connecting target-specific security reasoning to physical validation and measurement-driven follow-up.

To the best of current knowledge, this is the first reported LLM-agent methodology to span reconnaissance, reverse engineering, security-hypothesis generation, closed-loop physical laser experiments, measurement-driven adaptation and reporting under human approval gates. The agent owns routine experimental decisions; humans set scope and safety limits, authorize campaigns and adjudicate findings. Within those limits, the agent selects actions from measured outcomes rather than relying on a person to prescribe each shot.

To organize this reasoning, the agent builds a target evaluation surface: an evidence-to-experiment map connecting target components and operations to candidate errors, security consequences, observable outcomes and bounded tests. For each entry, it records spatial and timing hypotheses, a decision rule (classifier), negative controls, uncertainty and a stopping rule. The map combines die imagery, literature, firmware and documentation so that evidence can become a testable hypothesis and executable experiment.

Physical validation comprises three ATmega328P evaluations of register-load behaviour, personal identification number (PIN) verification and software Advanced Encryption Standard (AES), followed by an Artix-7 campaign that tests transfer to programmable logic through binary reverse engineering and optical-to-layout localization. The ATmega328P exercises the complete agent workflow; the Artix-7 case extends it to a second architecture. Table 2 distinguishes validated findings from follow-on evaluations.

Results

Autonomous evaluation spans the hardware-security workflow

The workflow comprises reconnaissance, enumeration, hypothesis generation, closed-loop experimentation and reporting (Figure 1). C0 fixes the human-approved scope, safety limits, shot budget and stop conditions. During **reconnaissance**, the agent combines literature, documentation, firmware and imagery to recover operations, state, timing references and candidate block assignments, retaining their uncertainty. During **enumeration**, it

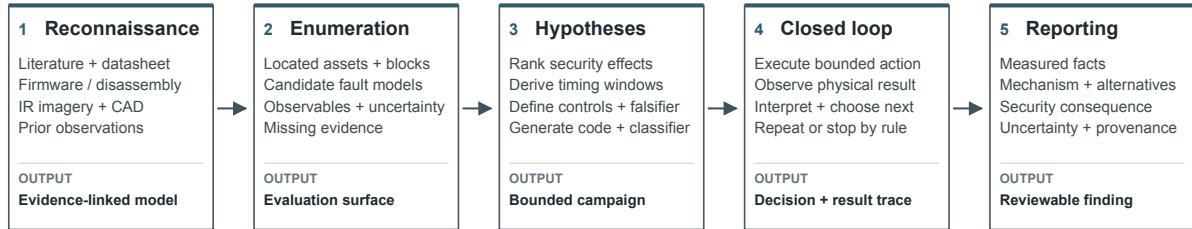
links these assets to candidate fault models, security effects and distinguishing observables. The proposed evaluation determines which regions and operations to prioritize; at C1, the human checks the target mapping and admits the evaluation.

During **hypothesis generation**, the agent ranks candidate explanations, derives spatial and timing windows, and implements campaign code and classifiers. Each campaign specifies a protected property, a success criterion, negative controls, observations that would reject the explanation and a stopping rule. C2 authorizes a bounded campaign block after code and classifier gates pass.

The **closed loop** connects execution to the next scientific decision. The controller checks approved bounds and workload preconditions, and records actions and readbacks. The agent interprets these observations before selecting the next experimental condition, control or stopping decision. In the ATmega record, the agent revised bounded proposals between campaign blocks; within each block, scripts executed the authorized sequence and stop rules. C3 covers anomalies, safety stops and human redirection of a region or parameter limit; routine choices within the authorized block remain with the agent.

During **reporting**, the agent relates outcomes and denominators to the hypothesis, distinguishes measured effects from inferred mechanisms and untested proposals, and identifies the next discriminating experiment. C4 reserves acceptance, qualification or rejection of the conclusion for the human. These checkpoints define the arbitration contract; the campaign series and readout endpoints are specified separately below. Extended Data Fig. 1 shows the physical bench and campaign software layers.

LLM AGENT: EVIDENCE TO PHYSICAL FINDING



SPARSE HUMAN ARBITRATION

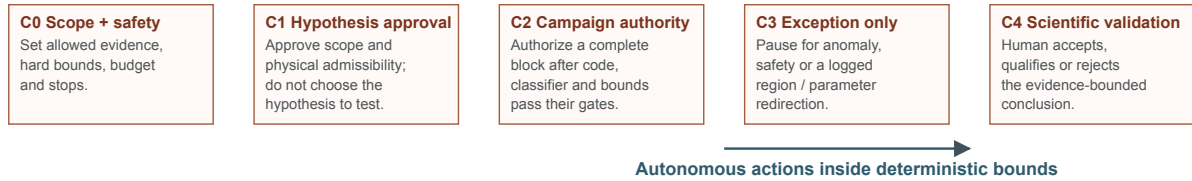


Figure 1: **Agent workflow with sparse human arbitration.** Evidence becomes a bounded campaign and reviewable finding. The agent selects actions from physical outcomes within deterministic limits; C0-C4 reserve authorization, exceptions and conclusions for humans.

The agent turns target evidence into an experimental map

For the ATmega328P, the agent combined the datasheet, AVR instruction set, firmware and an infrared (IR) image of the exposed die. The target is an 8-bit Harvard architecture with 32 registers, 32 KiB Flash program memory, 2 KiB static random-access memory (SRAM), 1 KiB electrically erasable programmable read-only memory (EEPROM) and a two-stage fetch/execute pipeline [10]. AVR instruction semantics and cycle counts follow the instruction-set manual [11]. At 16 MHz, one clock period is 62.5 ns. The ATmega firmware implements three evaluated workloads: a sequence of ten two-cycle `ld rK, Z+` instructions, PIN verification and triggered software AES. A missing load leaves the destination unchanged and suppresses the pointer post-increment, shifting the subsequent register tail. This structural sequence distinguishes a skip from an arbitrary wrong byte; the complete-response endpoint and characterization classifiers are specified separately in Methods.

The infrared image supports schematic macro-block assignments (Figure 2 a). The agent linked Flash/fetch, SRAM, EEPROM, register and mixed-signal/control regions to candidate transient or retained effects (Table 1). ATmega photon-emission microscopy (PEM) recorded 20 idle and 20 active frames. The differential image localized workload-related activity to candidate digital logic (Figure 2 b). Combined with the macro-block exclusions, this activity map seeded the physical search, and the subsequent responses established fault sensitivity. Optical registration provides a functional localization search prior rather than a calibrated pixel-to-stage transform.

The second target is a 28-nm Artix-7 XC7A200T on a Nexys Video board. The agent performed firmware-binary reverse engineering with Ghidra [12] and supporting analysis tools, identifying and interpreting functions alongside the device documentation. A backside IR mosaic covers approximately 11 mm by 12 mm. The agent aligned die edges, backside markings and clock regions to a mirrored Vivado Device view, then used implemented placement and routing to locate the AES datapath and the two responsive delay sensors shown in Figure 2 c [13],

[14], [15]. Vivado placement supplies slice-level coordinates; optical registration is at clock-region level. Table 1 records the resulting FPGA evaluation surface, including configuration readback as the discriminator required before attributing a repeated wrong output to configuration SRAM.

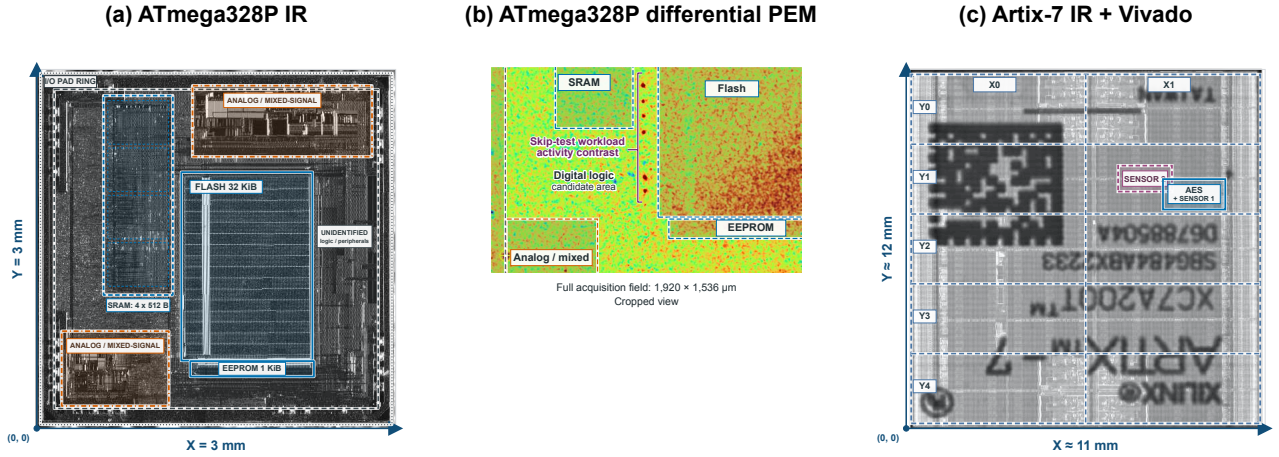


Figure 2: **Target localization.** **a**, Schematic ATmega328P blocks over infrared imagery. **b**, Active-minus-idle PEM during the instruction-skip workload (20 two-second frames per state); red denotes greater active-state contrast. Shading marks excluded blocks, and the clear digital region seeds laser testing. The crop comes from a 1,920 μm by 1,536 μm x5 field. **c**, Artix-7 clock-region registration with Vivado-located AES logic and responsive sensors.

Table 1: **Cross-target evaluation surface.** Candidate mechanisms, security effects and required discriminators derived from target artifacts. Optical assignments are schematic; FPGA slice coordinates come from Vivado placement.

Block or asset	Candidate fault model	Security consequence	Distinguishing evidence
ATmega fetch/execute	Transient instruction omission or modification	Altered accesses, comparisons or branches	Destination, pointer and control-flow signatures; distinguish execution from stored-code changes.
ATmega Flash	Retained program-memory corruption	Changed code or constants	Address-specific readback after reset and power cycle; exclude intended writes.
ATmega SRAM/stack	Volatile state corruption or faulty access	Altered PIN, counter, AES state or return data	State readback with verified address/access; distinguish storage upset from skipped access.
ATmega registers/status register (SREG)	Architectural-state corruption	Altered predicates, loops or intermediates	Held-state observation paired with dependent computation or branch.
ATmega EEPROM	Access error or retained nonvolatile corruption	Configuration change in an EEPROM-using workload	Sacrificial record and post-recovery readback; requires an additional workload here.
ATmega clock/reset/I/O	Timing, reset or communication disturbance	Control disruption or misleading readout	Independent timing/reset observations and complete valid responses.
Artix AES logic/state	Transient logic/timing fault or captured state error	Incorrect encryption	Input-specific ciphertext decision rule, cycle-resolved observation and recovery behavior.
Artix configuration/routing	Retained volatile configuration upset	Modified logic or connectivity	Masked configuration-frame difference before recovery, correlated with behavior.
Artix S1-S3 sensors	Delay-path, sensor-state or sampling disturbance	Missed indication or false alarm	Complete sensor sequence, per-sensor decisions, placement and laser-off baselines.
Artix clock/UART/buffers	Sequencing, framing or buffer corruption	Wrong or stale reported result	Input-output correspondence, frame integrity and independent operation markers.

Turning a candidate into a testable experiment

An evaluation-surface entry becomes testable when its classifier distinguishes a security outcome from alternatives. The ATmega analysis produced two strict decision rules: a complete register dump matching a shifted-load template, and a wrong-PIN acceptance response followed by a separate authenticated-state read. Counter changes, malformed replies and resets were separate classes. AES evaluation compared complete ciphertexts against a public known-answer vector and retained fault-free, altered, empty, reset and malformed responses as separate classes.

The additional entries in Table 2 and Extended Data Fig. 2 define follow-on questions and their required discriminators, separately from the measured findings. Extended Data Table 2 records the additional evidence needed for each.

A bounded search identifies omission of one load instruction

The agent combined firmware and instruction-set analysis with operator-supplied die-region exclusions and a previously established faulting-power range to bound the physical search. Two tested positions returned the fault-free register sequence. The third returned a complete shifted-load signature and activated the automatic stop. The unchanged destination and one-register shift in the remaining load sequence establish an instruction-level skip of the targeted `ld rK, Z+` operation.

A subsequent fixed-site, one-factor characterization completed 35 shots: 18 matched exact templates and 17 returned the recorded `gold` class (Extended Data Fig. 4). The delay, width and drive-current series contained 18, eight and nine shots, respectively. Three repetitions of the nominal reference condition produced template load counts of two, two and one. These single-pass transitions characterize the tested conditions; Extended Data Fig. 4 shows the ordered observations.

A waveform audit identified duplicate acquisitions across shots with different register outcomes. The agent rejected shot-specific timing interpretation, corrected acquisition handling and required fresh laser-off traces before continuing. The register responses remain valid. Figure 3a therefore places the timing hypothesis beside a separate normalized oscilloscope reference; Methods and Extended Data Table 1 specify its provenance and permitted interpretation.

A separate state read corroborates a wrong-PIN bypass

Initial tests exposed denial-path and retry-counter behavior but did not satisfy the bypass objective. The agent consequently required a wrong-PIN acceptance response and a separate authenticated-state read as the success criterion. The final search used the non-secret attempt 0000 and reset the retry counter before each shot. Its three resumed campaign blocks contain 84 physical pulses. Four laser-off recovery controls restored and verified the denied state after corrupted responses.

The final shot returned the acceptance token, and a separate status transaction recorded the authenticated state, with the retry counter still equal to three. The automatic stop then ended the search (Figure 3 b). The acceptance response and separate authenticated-state query confirm the wrong-PIN bypass.

AES output corruption is reproduced with paired controls

The operator connected the AES trigger after a routing gate failed. The agent then identified missing output initialization in the firmware and corrected instrumentation without changing AES, authentication or key-expansion logic. Eight fresh laser-off validation traces verified known-answer behavior and the measurement path before physical testing resumed.

The initial AES pilot produced fault-free, invalid and reset responses alongside a valid altered ciphertext affecting 15 bytes and 51 bits. The mixed outcomes led the agent to retain complete response classes and require a verified laser-off fault-free result before refinement.

In lower-stimulus refinement, the identical two-byte ciphertext difference recurred in 14/14 shots at the selected condition across five campaign blocks. One adjacent sampled condition produced the same output. All 29/29 paired laser-off controls in R2b-R2e matched the fault-free ciphertext. The changed zero-based byte indices were 10 and 13, totaling ten changed bits (Figure 3 c). Four distinct, correctly centred R2d acquisitions support controller-pulse provenance (Figure 3 d).

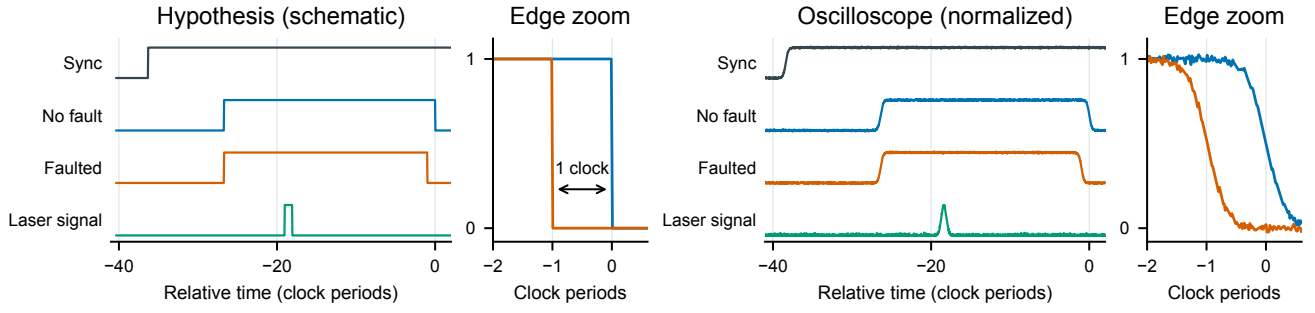
Across the initial pilot and refinement, 35 physical AES pulses are retained. Extended Data Fig. 5 accounts for every outcome: ten fault-free responses, 15 occurrences of the repeated pattern, two other altered ciphertexts, four empty responses, three reset messages and one malformed response. The 15 pattern occurrences comprise the 14 selected-condition observations and one adjacent-condition observation; these denominators are not pooled into an estimated success probability. The operator stopped when the physical proof-of-concept objective was met. The two-byte pattern establishes repeatable physical output corruption, but does not match the four-byte output expected from a classical single-byte fault before round-9 MixColumns [16]. Exact round and internal mechanism remain unresolved; no key recovery was attempted.

The workflow transfers to programmable hardware

The retained FPGA campaign used visual inspection and Vivado placement, not photon emission, to select the AES and sensor region. After checking fault-free UART, ciphertext and monitor behavior, the campaign localized responses with a coarse scan and executed a denser AES-region campaign (Figure 4 a). The independently parsed log contains 182,985 physical records across 5,545 coordinates, eleven nominal currents and three pulse widths, with 11,896 ciphertext mismatches (6.50%). For a compact cross-architecture result, Figure 4 reports the 150-ns series: 60,995 records with 7,319 mismatches. This series shows a localized response and rising mismatch rates at high nominal current (Figure 4 a-b); full campaign data will accompany the public research package.

The controller immediately repeated a wrong ciphertext at low current. The same wrong output recurred for 8,874 mismatches (74.60%), while 3,022 did not; 8,929 of all mismatches coincided with at least one monitor flag (Figure 4 c). Of all mismatches, 2,967 (24.94%; 1.62% of all shots) had no coincident final-sample monitor flag. Because the controller retained the final sensor sample rather than the complete indication window, this class measures final-sample concurrence. The **repeated-output** and **non-repeated-output** classes describe

a Timing hypothesis and separate oscilloscope reference



b Wrong-PIN search and stop c AES ciphertext differences

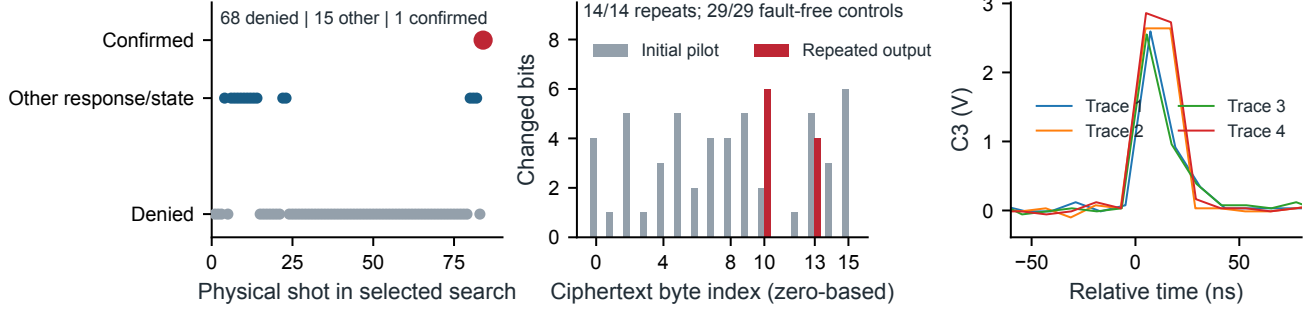


Figure 3: **ATmega328P timing and physical results.** **a**, Schematic one-clock hypothesis and a separate normalized oscilloscope reference. **b**, Outcomes from 84 physical pulses; red marks the wrong-PIN bypass confirmed by acceptance and authenticated-state readback. **c**, AES differences: 14/14 selected-condition outputs were identical and 29/29 paired controls matched the fault-free ciphertext. **d**, Four distinct, correctly centred R2d controller traces.

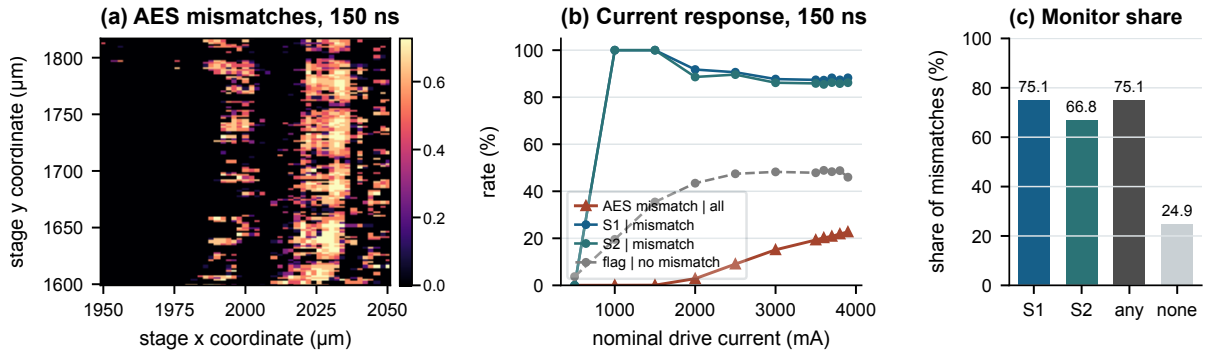


Figure 4: **Artix-7 AES campaign.** **a**, Mean mismatch rate at 150 ns across eleven nominal currents (5,545 coordinates, 60,995 records; 900-ns delay). **b**, Current response at 150 ns. **c**, Monitor concurrence among 11,896 mismatches; Sensor 3 is omitted.

immediate retests; configuration readback is the required discriminator for attributing repeated outputs to configuration SRAM.

The Artix case extends the workflow from instruction effects in a microcontroller to optical-to-layout localization, cryptographic fault classification and detector assessment in programmable logic. Immediate repeat readback provides a concrete follow-up discriminator; configuration readback is the next proposed test for identifying the underlying mechanism.

Table 2: **Agent-generated follow-on evaluations.** Measured or architectural evidence is linked to each question and decisive control; Extended Data Tables 2-3 provide the evidence requirements and decision trace.

ID / surface	Existing basis	Follow-on question	Decisive readout and control
P1-P2 / ATmega state	Security globals, registers and stack	Can held-state corruption be distinguished from a skipped access?	State mismatch with verified access and independent readback; laser-off and shifted-timing controls.
P6 / AES state	Repeatable two-byte output corruption	What evidence distinguishes transient computation from retained state?	Fresh public vectors and independent recovery checks; internal mechanism and key-state effects remain untested.
P8 / Authorization	PIN/key-change commands require authentication	Can faulted authorization permit a protected state change?	Verify key change by independent AES vectors or PIN change by later verification; laser-off rejection baseline.
F2 / Configuration	Repeated wrong FPGA ciphertexts	Does retained configuration contribute to repeated failure?	Masked readback before recovery; fresh inputs and separate logic-reset/reconfiguration controls.
F3-F4 / Sensors	Placement, LUT/carry and clock variants	How do placement and implementation affect coverage?	Matched designs and separately calibrated thresholds; missed indications and laser-off false alarms.
F8 / Readout	Only final sensor samples retained	Can final-sample classification miss an earlier indication?	Compare final-sample and full-window decisions on valid frames with matched false-alarm evaluation.
T1 / Triggerless	Present experiments use firmware timing observables	How does reduced observability affect evaluation coverage?	Measure synchronization uncertainty and validated output coverage under matched budgets against the triggered baseline.

Discussion

Across the two targets, the methodology connects a security hypothesis to an observation that can test it. The ATmega record links firmware analysis to register and authentication decision rules, followed by repeatable AES output corruption. The Artix case adds optical-to-layout localization and distinguishes faulty computation from the final recorded alarm state. The evaluation surface is shared; the evidence needed to resolve each mechanism remains target-specific.

The ATmega sequence also shows why the loop must interpret evidence rather than merely execute a sweep. Conflicting register waveforms caused the agent to reject a timing claim and request fresh controls; failed AES measurement gates prompted an instrumentation correction; and the first altered ciphertext motivated a lower-stimulus refinement with paired controls. The operator supplied the physical trigger, authorized each new block and stopped acquisition at the proof-of-concept objective. Extended Data Table 3 preserves this observation-to-revision chain and the division of responsibility.

The wrong-PIN acceptance and separate authenticated-state read establish an authentication bypass. Repetition of an identical ciphertext difference with fault-free controls establishes reproducible physical AES output corruption. The observed two-byte pattern does not by itself identify the internal electrical mechanism, establish compatibility with classical differential fault analysis (DFA) or enable key recovery. PEM supplies spatial localization; physical injection and functional readout establish the security effect.

Reverse engineering and recovery of the spatial-temporal operating envelope can dominate preparation. The injection search spans position, power, pulse width and delay. IR structure and PEM can replace a full-die sweep with local refinement; in evaluator practice, favorable targets can move localization from hours or days to minutes, although the gain depends on circuit structure, workload observability and registration quality [17]. Trigger-synchronized or time-resolved PEM can also identify an active execution phase. Delay is usually the hardest parameter for dynamic faults because scheduling, variable execution paths and jitter make a host transaction an uncertain reference. Without a dedicated trigger, repeated I/O, power or electromagnetic patterns can provide a synthetic timing reference, whose phase error must be measured. Static faults in SRAM or retained registers relax this constraint because the state can be illuminated and read back over a broader interval. Held-state and execution-dependent models nevertheless require different observables; persistence in software cannot be assumed for combinational crypto hardware [18].

Together, the ATmega328P and Artix-7 studies support a campaign-decision closed loop within the C0-C4 framework, with human authority over physical safety and conclusions.

A retrospective practitioner comparison estimated a roughly fivefold reduction in preparation time: about one working day per agent-assisted target versus one working week or longer for comparable manual evaluations. The estimate uses the elapsed-time, expertise, target-knowledge, access and equipment dimensions of JIL attack-potential assessment [19]. The agent also generated the broader follow-on evaluation surface in Table 2, including hypotheses beyond the researcher's usual specialization. Recorded AGENT and OPERATOR events

make the decisions inspectable. These results establish the physical feasibility of the evidence-to-experiment methodology across two hardware architectures.

Evaluation reporting can connect this record to Common Criteria and its Common Evaluation Methodology (CEM) [20], [21], while retaining the distinction between demonstrated effects, unresolved mechanisms and proposed countermeasure tests. No attack-potential rating or certification-level coverage is assigned.

Methods

Agent, operator and evidence provenance

The physical ATmega study used an LLM coding agent, Python instrument-control and analysis software, and Arduino/AVR C firmware. PyLabSAS instrument drivers provided the laboratory interfaces [22]. A trained operator retained control of the class-4 laser, interlocks and emergency stop. The operator authorized physical phases, set the proof-of-concept objective and ended AES acquisition. The agent analyzed firmware, proposed bounded experiments, checked instrument and target preconditions, classified responses and recorded stop decisions.

The authoritative ATmega source is a frozen evidence package containing 453 SHA-256-indexed files, including raw camera frames, run tables and logs, retained waveforms, firmware inputs, approvals and decision streams. The observable records preserve AGENT and OPERATOR actor labels. Extended Data Table 3 reports paraphrased observable decisions and approvals while preserving source AGENT and OPERATOR labels. Event identifiers within each stage are distinct from the methodology's C0-C4 checkpoint labels. The five retained stage streams contain 40 AGENT and 17 OPERATOR events. These counts describe provenance records, not intervention duration; human actions were limited to approval gates, physical connections, bounded redirection and acceptance or rejection of conclusions.

The demonstrated implementation used Qwen3.5-35B-A3B-FP8 served through vLLM for multimodal reasoning over documentation, firmware and disassembly, code and images. Inference ran on one NVIDIA L40S GPU with 48 GB VRAM. The study uses this model to test end-to-end methodology feasibility; the evidence, decision and control interfaces are specified independently of model internals. An AI assistant supported evidence organization and language revision of this manuscript. The author verified the source evidence, calculations, citations and final text and retains responsibility for all scientific judgements and conclusions. The agent received the security objective, target artifacts, approved bounds and localization evidence where available, then generated shot sequences and vulnerability hypotheses within each approved block. Low-level actions and stop rules followed agent-generated campaign scripts; the agent interpreted the resulting observations and selected or revised subsequent blocks. Campaign-decision autonomy operated under C0-C4 authority; low-level actuation remained deterministic, and physical operation remained subject to bench controls and operator authorization.

The retained software and observation interface is file-based: campaign-wide decisions, per-stage JSON Lines (JSONL) events, approval records, run logs and comma-separated value (CSV) tables, and classification summaries link inputs and outputs through evidence paths and run identifiers. Approval records define the permitted block; campaign scripts perform its deterministic actions and return observations for interpretation. Subsequent decision records state the revised proposal or stopping decision. The package manifest identifies the retained campaign and analysis source by hash. Observable event records preserve evidence, proposals, approvals, actions, revisions and stop decisions.

All package files and publication-manifest hashes were checked before analysis. The new-result figures are regenerated from the frozen observations; their sidecar records input hashes, output hashes and display transformations. Earlier ATmega teaching datasets, reconstructed waveforms and simulation outputs are segregated from the new results and denominators. Extended Data Table 1 enumerates the included evidence and its permitted interpretation. Figure 3a is a separate reference recovered from a retained normalized oscilloscope SVG, not from the authoritative campaign package. Its four vector traces are hash-checked and replotted without synthesizing samples. A common no-fault edge origin within each panel and the 16-MHz period define the displayed time units; the laser signal is not clipped or treated as calibrated optical power. The source does not supply the original raw acquisition export or timing uncertainty, and the figure makes no numerical measured-separation claim.

Firmware and optical evidence

For ATmega, the agent reverse-engineered the retained HEX firmware image using generated disassembly, documentation and AVR instruction semantics [11] to recover security-relevant functions and instruction-level endpoints. The later AES instrumentation correction added output initialization to the existing trigger; the retained source comparison leaves AES, authentication and key expansion unchanged. Fresh known-answer controls

were required after the upload. For Artix, function interpretation used Ghidra [12] and supporting binary analysis, while implemented Vivado placement and routing supplied slice positions.

The physical bench used a 1,064-nm laser with approximately 2-W maximum source power. The available objectives ranged from 5x (approximately 15- μ m spot) to 50x (approximately 1.5- μ m spot). ATmega access was frontside after acid decapsulation; Artix access was backside after mechanical opening. Artix registration used visual inspection and mirrored Vivado placement, without photon-emission acquisition. Related optical work is described by Lima et al. [23]. Nominal source commands and direct drive-current values are retained in their native units; drive current is not converted to incident optical power without calibration.

ATmega PEM used a Raptor Ninox 640 at high gain with bench illumination off. Twenty idle and twenty assembly-active frames, each exposed for 2 s, were organized in ten paired blocks with alternating acquisition order. The active-minus-idle image identifies workload-related activity that, together with macro-block exclusions, prioritizes candidate digital logic for physical testing. Camera contrast is a relative activity measurement rather than a photon count; fault injection and functional readout confirm security relevance.

Figure 2b uses the same acquisition's deposited smoothed difference (Gaussian sigma 1.5 pixels). Colour limits use the symmetric 99.8th percentile of absolute contrast. For display, the image is rotated 180 degrees and shifted vertically by 54 camera pixels following the operator's crop guide. The visible spot line marks the activity contrast used to focus the candidate region. The full x5 acquisition field (1,920 μ m by 1,536 μ m) is stated separately from the cropped view, without assigning die-stage coordinates. The block outlines and exclusion shading in Figure 2b are schematic visible-field annotations based on Figure 2a and function as a search prior rather than a calibrated pixel-to-stage transform. They leave the source camera measurements unchanged.

Register and authentication endpoints

The register endpoint requires a complete ten-byte response matching an exact shifted-load template. A marker byte alone is insufficient; malformed and unclassified responses remain separate. The selected evidence-informed localization contains three pulses, stopped at the first instruction-level skip. Fixed-site characterization then varies one factor per series: delay (18 shots), width (eight) or drive current (nine). Three repeated reference observations yielded template load counts two, two and one. Counts describe tested settings, not a fitted transition or a population fault probability.

The characterization's aborted attempt emitted one pulse before the completed 35-shot run. Its record lacks a persisted response row and waveform; the run log and decision event retain the additional pulse. Total exposure is therefore 36 against the recorded 35-pulse cap. Fourteen duplicate waveform-hash groups in the completed run invalidated shot-to-waveform pairing. The subsequent four-trace laser-off freshness gate passed after the acquisition-handling correction; it does not retrospectively validate rejected traces.

The selected wrong-PIN search comprises three resumed campaign blocks (7, 2 and 75 pulses), all using the non-secret attempt 0000 without supplying the correct PIN to that search. The earlier correct-PIN lockout calibration remains a different endpoint. Each selected attempt starts with the retry counter reset to three and authentication denied. Confirmation requires both an immediate AA response and a separate status read with `authenticated=0xAA`. Four laser-off recovery controls checked normal denial before continuation after anomalous responses. The search ended at its first confirmed event. Results at untested settings are not inferred.

Figure 3b classifies the actual response and separately read state, rather than relying on a saved candidate flag. The functional response and state query, not the corresponding acquisition traces, establish the endpoint; no clock-shortening estimate is derived from those traces.

AES controls and acquisition quality

A public FIPS-197 known-answer vector supplied the fault-free AES reference. The initial routing attempt passed functional encryption but failed scope-density and routing gates, with no laser pulse. After the operator's wiring change and the agent's instrumentation correction, four fresh full-window traces and four short-window traces passed the laser-off validation gates. These eight traces are baseline measurements rather than injected trials. The initial pilot includes four physical pulses despite only three persisted response rows: the AGENT decision and pilot summary document an empty-response pulse before a logging failure. Its exploratory waveform is listed as saved in the original record but is excluded from the share package. The event is retained in exposure accounting without a new waveform-based inference. A separate laser-off recovery restored the known-answer result before resumption. The first valid altered ciphertext differs from the fault-free result in 15 bytes and 51 bits. Its C4 signal remains below 0.33 V, so that channel is not treated as independently thresholded optical-delivery evidence.

The subsequent R2a-R2e tables contain 31 physical pulses. Same-condition observations contribute 1, 4, 4, 4 and 1 identical two-byte results, respectively. One adjacent R2e condition adds a separate matching observation. All 29 paired laser-off controls in R2b-R2e match the fault-free ciphertext, including recovery controls after invalid

replies. The total AES exposure is 35 physical pulses, including the initial pilot's missing-row event. Invalid output and reset responses remain in the run counts.

R2b overwrote repeated-condition waveform filenames; R2c retained distinct files but used the wrong scope-centering sign. Both retain functional responses. Only R2d supplies the four distinct, correctly centred 16,668-point controller traces used in Figure 3d. The plot changes the time origin by a constant and does not rescale voltage or infer an optical pulse width. Legacy CSV labels containing DFA were re-evaluated from complete ciphertext differences, yielding the response classes reported in Results.

Artix campaigns and analysis

The AES-region campaign crossed 5,545 coordinates with eleven nominal currents and 50-, 100- and 150-ns pulse widths at a 900-ns delay. It logged 182,985 primary records, each compared with the fault-free ciphertext. Figure 4 presents the 60,995-record 150-ns series as a compact cross-target demonstration.

Monitor concurrence means that at least one of three sensor readings fell outside the inclusive interval 104-108 (instrument units), consistent with the recorded flag. The controller retained the final sensor sample of each sequence; F8 proposes full-window coverage as a follow-on evaluation.

Wrong ciphertexts were immediately retested at 100-mA drive current, classified by whether the same wrong output occurred, and followed by a target power cycle. These retests are separate from the primary records. Aggregate output and monitor classes retain all 182,985 primary records.

Evaluation reporting and attack potential

Extended Data Fig. 3 summarizes the physical evaluation workflow and its feedback loop.

Joint Interpretation Library (JIL) attack-potential assessment distinguishes identification from exploitation on another target specimen, accounting for elapsed time, expertise, target knowledge, access, equipment, and open samples or samples with known secrets [19]. EUCC supplies scheme-specific guidance [24]. A rating requires documented factors and evaluator judgement; none is assigned here.

Statistical units

The experimental unit is a target specimen; observations within a campaign are repeated measures on that specimen. PEM supplies an activity-localization prior tested by subsequent fault injection. Register and AES counts are descriptive observations from adaptive, single-specimen campaigns rather than population estimates, so no binomial intervals are assigned. Complete response classification retains empty, reset and malformed outcomes, with laser-off controls separated from injected-pulse totals. Artix spatial and parameter-response rates use the selected 60,995-record 150-ns series; pooled output classes retain the 182,985-record audit denominator, and immediate retests remain separate from primary records.

Data availability

The public research package contains the figures, supporting result tables and CSV data, and observable AI-agent reasoning and decision records underlying the reported findings. It is available at partage.imt.fr/index.php/s/CNqoaybmQ7zSfRp.

Code availability

The analysis and figure-generation code is shared through the same research package. The agent, controller and PyLabSAS campaign-control code that operates the laser bench is available to editors and referees during review and to qualified researchers on request; it is not publicly released because it directly controls a class-4 laser bench.

References

- [1] T. Sugawara and K. Sakiyama, "An Introduction to Laser Fault Injection Attack against Secure Embedded Devices," *The Review of Laser Engineering*, vol. 47, no. 6, p. 305, 2019, doi: 10.2184/lsej.47.6.305.
- [2] C. O'Flynn and Z. Chen, "ChipWhisperer: An Open-Source Platform for Hardware Embedded Security Research," in *Constructive Side-Channel Analysis and Secure Design (COSADE)*, in Lecture Notes in Computer Science, vol. 8622. Springer, 2014, pp. 243–260. doi: 10.1007/978-3-319-10175-0_17.
- [3] P. Maurine, "Techniques for EM Fault Injection: Equipments and Experimental Results," in *2012 Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC)*, IEEE, 2012, pp. 3–4. doi: 10.1109/FDTC.2012.21.
- [4] S. Ordas, L. Guillaume-Sage, and P. Maurine, "Electromagnetic fault injection: the curse of flip-flops," *Journal of Cryptographic Engineering*, vol. 7, no. 3, pp. 183–197, 2017, doi: 10.1007/s13389-016-0128-3.
- [5] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous Chemical Research with Large Language Models," *Nature*, vol. 624, pp. 570–578, 2023, doi: 10.1038/s41586-023-06792-0.
- [6] Y. Xie, K. He, and A. Castellanos-Gomez, "Toward Full Autonomous Laboratory Instrumentation Control with Large Language Models," *Small Structures*, vol. 6, no. 8, p. 2500173, 2025, doi: 10.1002/ssstr.202500173.
- [7] Z. Chen *et al.*, "An Agentic Artificially Intelligent X-ray Scientist," *Nature Machine Intelligence*, vol. 8, pp. 1075–1086, 2026, doi: 10.1038/s42256-026-01261-5.
- [8] M. Ebrahimabadi *et al.*, "DELFINES: Detecting Laser Fault Injection Attacks via Digital Sensors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 3, pp. 774–787, 2024, doi: 10.1109/TCAD.2023.3322623.
- [9] D. Saha *et al.*, "LLM for SoC Security: A Paradigm Shift," technical report 2023/1561, 2023. [Online]. Available: <https://eprint.iacr.org/2023/1561>
- [10] Atmel Corporation, "ATmega328P: 8-bit AVR Microcontroller with 32K Bytes In-System Programmable Flash." 2015. [Online]. Available: https://ww1.microchip.com/downloads/aemDocuments/documents/MCU08/ProductDocuments/DataSheets/Atmel-7810-Automotive-Microcontrollers-ATmega328P_Datasheet.pdf
- [11] Microchip Technology Inc., "AVR Instruction Set Manual." 2021. [Online]. Available: <https://ww1.microchip.com/downloads/en/DeviceDoc/AVR-InstructionSet-Manual-DS40002198.pdf>
- [12] National Security Agency Research Directorate, "Ghidra Software Reverse Engineering Framework." 2019. [Online]. Available: <https://github.com/NationalSecurityAgency/ghidra>
- [13] Xilinx, Inc., "7 Series FPGAs Data Sheet: Overview (DS180)." 2020. [Online]. Available: <https://docs.amd.com/api/khubs/documents/2LByHKO~nSZXcei2D55fTg/content>
- [14] Xilinx, Inc., "7 Series FPGAs Configurable Logic Block User Guide (UG474)." 2016. [Online]. Available: https://docs.amd.com/v/v1.8-English/ug474_7Series_CLB
- [15] Advanced Micro Devices, Inc., "Vivado Design Suite User Guide: Design Analysis and Closure Techniques (UG906)." 2024. [Online]. Available: <https://docs.amd.com/r/2024.1-English/ug906-vivado-design-analysis/Navigating-in-the-Device-Window>
- [16] G. Piret and J.-J. Quisquater, "A Differential Fault Attack Technique against SPN Structures, with Application to the AES and KHAZAD," in *Cryptographic Hardware and Embedded Systems (CHES)*, in Lecture Notes in Computer Science, vol. 2779. Springer, 2003, pp. 77–88. doi: 10.1007/978-3-540-45238-6_7.
- [17] H. Perrin, J.-M. Dutertre, and J.-B. Rigaud, "Betrayed by Light: How Photon Emission Microscopy Empowers Register Bit-Level Laser Attacks on Microcontrollers," in *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2025, pp. 35–45. doi: 10.1109/HOST64725.2025.11050055.
- [18] F. Zhang *et al.*, "Persistent Fault Attack in Practice," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2020, no. 2, pp. 172–195, 2020, doi: 10.46586/tches.v2020.i2.172-195.
- [19] Joint Interpretation Library, "Application of Attack Potential to Smartcards and Similar Devices (Version 3.2.1)." Feb. 2024. [Online]. Available: <https://www.commoncriteriaportal.org/files/ccfiles/JIL-Application-of-Attack-Potential-to-Smartcards-v3.2.1.pdf>
- [20] Common Criteria Maintenance Board, "Common Criteria for Information Technology Security Evaluation, Part 1: Introduction and General Model (CC:2022, Revision 1)." Nov. 2022. [Online]. Available: <https://www.commoncriteriaportal.org/files/ccfiles/CC2022PART1R1.pdf>
- [21] Common Criteria Maintenance Board, "Common Methodology for Information Technology Security Evaluation: Evaluation Methodology (CEM:2022, Revision 1)." Nov. 2022. [Online]. Available: <https://www.commoncriteriaportal.org/files/ccfiles/CEM2022R1.pdf>
- [22] R. Viera, "PyLabSAS: laboratory-automation framework for bench control." 2026. [Online]. Available: <https://pylabsas.myresearch.ac/>
- [23] R. S. Lima, R. Viera, J.-M. Dutertre, W. Magrini, M. Pommies, and A. Bertrand, "Reverse-Engineering and Data Extraction from SRAM using Photon Emission Analysis," in *2024 IEEE Physical Assurance and Inspection of Electronics (PAINE)*, Huntsville, AL, USA, 2024, pp. 1–7. doi: 10.1109/PAINE62042.2024.10792702.
- [24] European Union Agency for Cybersecurity, "EUCC Scheme State-of-the-Art Document: Application of Attack Potential to Smartcards and Similar Devices (Version 2)." Dec. 2025. [Online]. Available: https://certification.enisa.europa.eu/publications/application-attack-potential-smartcards_en
- [25] Advanced Micro Devices, Inc., "7 Series FPGAs Configuration User Guide (UG470)." 2023. [Online]. Available: https://docs.amd.com/v/v1.17-English/ug470_7Series_Config

Funding

The author declares no relevant funding.

Author contributions

R.V. conceived the study, performed the experiments, adapted and used the agent, and wrote the manuscript.

Competing interests

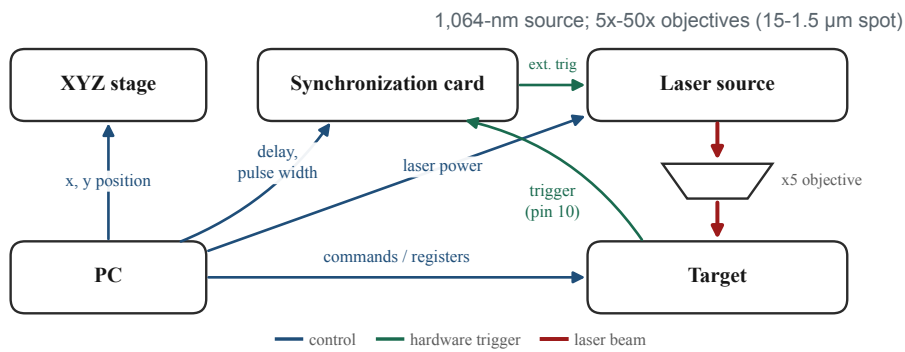
The author declares no competing interests.

Additional information

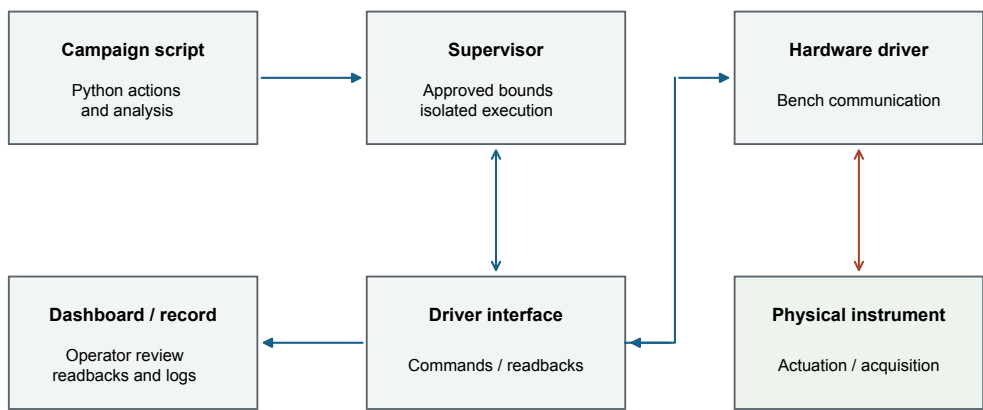
Correspondence and requests for materials should be addressed to R.V.

Extended Data

(a) Physical laser fault-injection bench



(b) Hardware campaign software



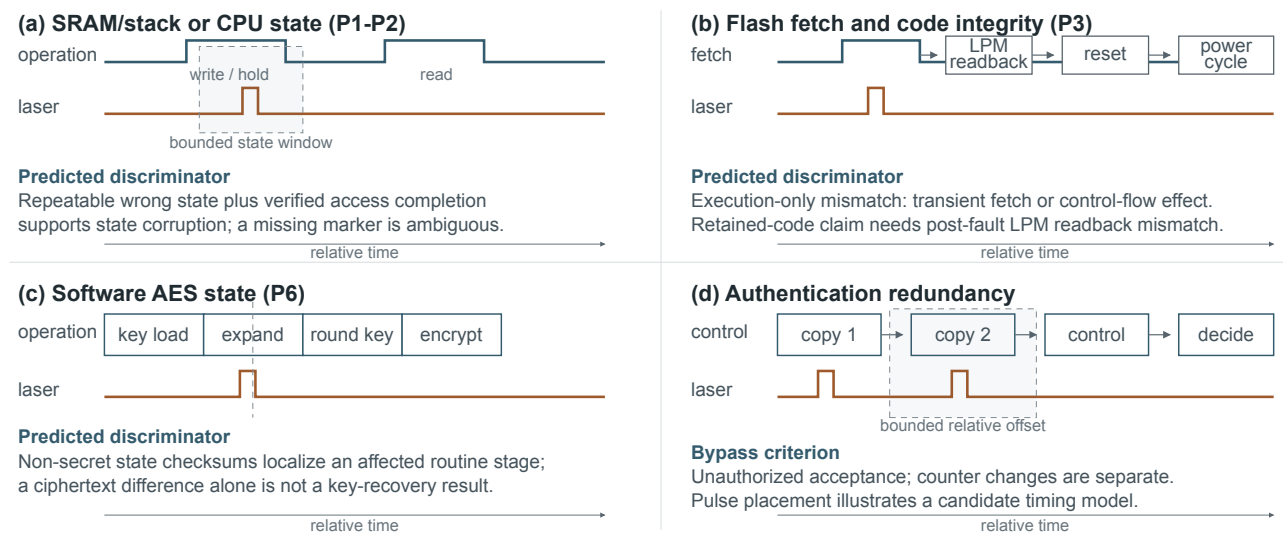
Extended Data Fig. 1 | Physical bench and campaign control. **a**, ATmega bench and trigger path. **b**, Campaign supervision, driver and readback layers.

Extended Data Table 1 | Physical evidence and trial counts. A1-A6 use the authoritative ATmega package; F1 uses the separate Artix acquisition. Units are endpoint-specific, and every documented exposure is retained in accounting.

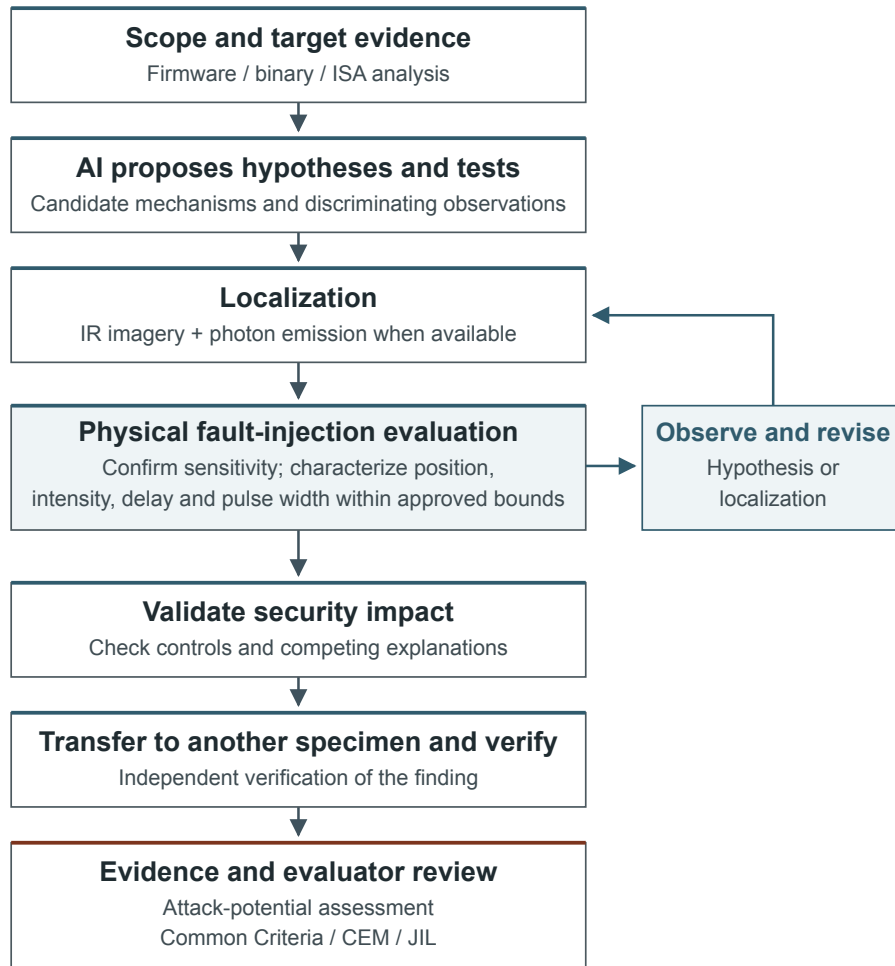
ID	Evidence series	Retained observations	Permitted interpretation
A1	ATmega PEM	20 idle and 20 active camera frames; ten paired blocks.	Workload-related activity and macro-block exclusions narrow the physical search to candidate digital logic; Fig. 2b.
A2	Register localization	Three physical pulses: two fault-free responses and one complete shifted-load template.	The complete shifted-load response establishes an instruction-level register-load skip.
A3	Register characterization	35 completed shots: 18 exact templates, 17 fault-free; one additional pulse in an aborted run.	Single-pass observations. Total 36 exceeds the recorded 35-pulse cap. Duplicate acquisitions invalidate shot-specific scope timing; Extended Data Fig. 4.
A4	Wrong-PIN search	Selected search: 84 pulses in three resumed campaign blocks; four recovery controls; an acceptance response with a separate authenticated-state read. Earlier calibration runs remain separate.	Confirmed wrong-PIN authentication bypass; Fig. 3b. The correct-PIN lockout calibration is excluded from this endpoint.
A5	AES routing gates and pilot	Eight initial laser-off validation traces. Four physical pilot pulses; only three persisted response rows. The missing-row event remains in decisions and chronology.	The first valid altered ciphertext changes 15 bytes / 51 bits. Its C4 trace does not establish independently thresholded optical delivery. The omitted exploratory waveform is not reanalyzed.
A6	AES refinement	31 pulses over R2a-R2e; 14 identical same-condition two-byte outputs plus one at an adjacent condition; 29/29 paired controls match the fault-free ciphertext.	Repeatable physical output corruption. Four unique, centred R2d traces support controller timing; R2b overwrite and R2c centering defects remain recorded. No classical differential fault analysis match or key recovery; Fig. 3c-d.
R1	Oscilloscope reference	Four normalized vector traces retained from a separate acquisition; shown beside a schematic hypothesis in Fig. 3a.	Explanatory reference only, outside the authoritative campaign counts. No new-campaign timing validation or numerical measured-separation claim.
F1	Artix AES and sensors	182,985 primary records over 5,545 coordinates; 11,896 mismatches; immediate retests separate.	Spatial and current-response panels present the 60,995-record 150-ns series. Repeated-output and monitor classes use the full primary series; Fig. 4.

Extended Data Table 2 | Follow-on questions and discriminators. Each row links existing evidence to the observations needed for attribution; C0-C4 checkpoints apply throughout.

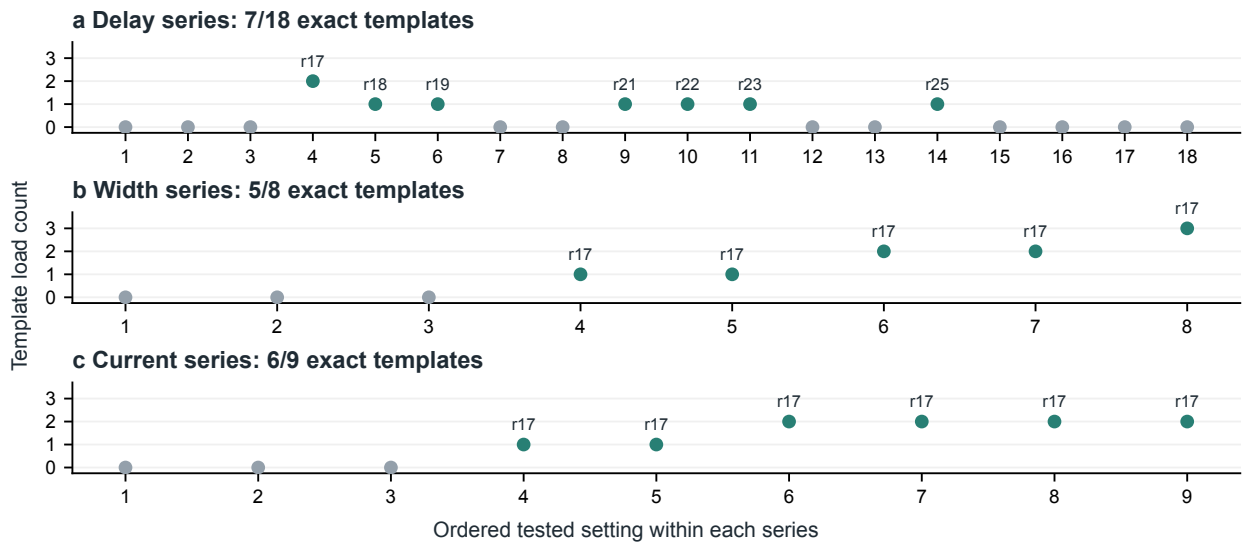
ID	Question	Existing evidence / status	Additional evidence required
P1	SRAM and stack integrity	Architectural candidate requiring a dedicated workload.	Independent state and access observations, negative controls and recovery behavior that distinguish a stored-state change from execution or readout error.
P2	Register file and SREG	Shifted-load templates establish an execution effect without localizing register-file or flag storage.	Agreement between state readback and dependent behavior, with baseline and sham comparisons.
P3	Flash fetch and code integrity	Architectural candidate requiring retained-code integrity measurements.	Independent memory integrity and execution evidence before and after recovery, excluding intended writes.
P4	EEPROM configuration	Evaluation requires an EEPROM-using workload.	A diagnostic workload, independent retained-data checks and baseline variability measurements.
P6	AES mechanism and state	Repeatable two-byte ciphertext corruption motivates mechanism and state attribution.	Fresh public vectors and recovery evidence that distinguish a computation error from retained-state effects; separate any claim of key recovery.
P7	Timing randomization	Countermeasure evaluation requires an implemented timing-randomization target.	Matched baseline and implemented-countermeasure measurements, a declared timing distribution, functional overhead and endpoint-specific coverage.
P8	Protected state changes	Authenticated status is measured; protected updates require separate confirmation.	Independent confirmation of the protected state and subsequent legitimate behavior, with rejection controls.
F2	FPGA retained failure	Repeated wrong output motivates configuration readback.	Matched configuration integrity and recovery observations [25]; exclude ordinary user-memory changes.
F3-F4	Detector coverage	Placement and implementation variants are proposed.	Matched designs, frozen thresholds, separately measured laser-off false alarms and missed indications.
F8	Sensor sequence coverage	The recorded controller returns final sensor values.	Complete valid sequences and a comparison of final-sample and full-window decisions at matched false-alarm rates.
T1	Reduced timing observability	Present physical studies use timing observables; triggerless assessment requires dedicated measurements.	Measured synchronization uncertainty, matched budgets, independent output validation and an explicit statement of available interfaces.



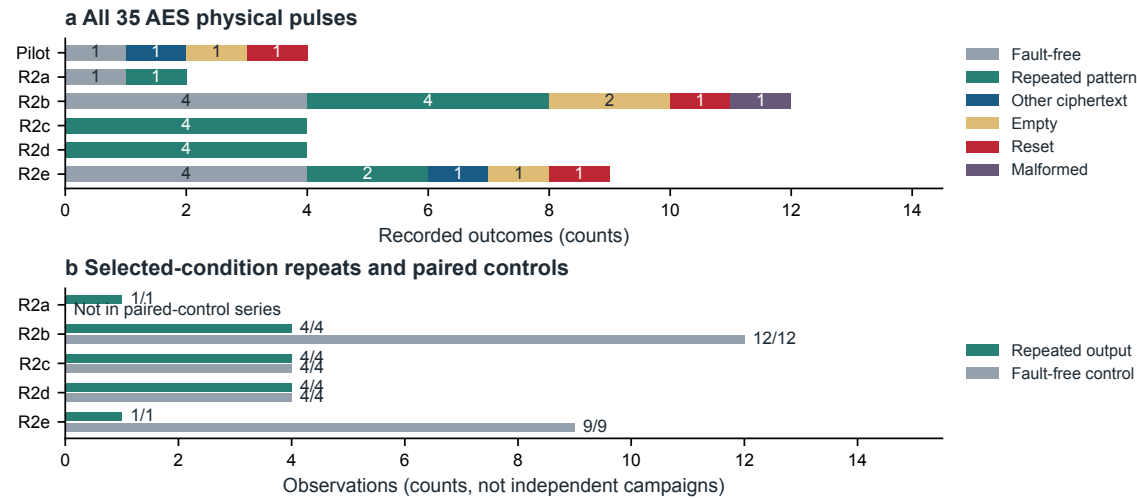
Extended Data Fig. 2 | Schematic evaluation blueprints. a, SRAM/register state; b, Flash fetch; c, software-AES state; d, redundant authentication. Extended Data Table 2 gives the discriminators.



Extended Data Fig. 3 | Physical evaluation workflow. Evidence guides hypotheses, localization, bounded tests and revision. C0-C4 follow Fig. 1; CC/CEM/JIL provides the reporting framework [19], [20], [21].



Extended Data Fig. 4 | ATmega register characterization. Exact-template counts across ordered **a**, delay, **b**, width and **c**, drive-current settings. Labels mark the first retained register; all 35 completed responses are shown without interpolation.



Extended Data Fig. 5 | AES outcome accounting. **a**, All 35 pulses by campaign block, including the documented missing-row exposure. **b**, Selected-condition repeats and paired laser-off controls. The repeated class comprises 14 selected-condition observations and one adjacent match.

Extended Data Table 3 | Agent decisions and human arbitration. Paraphrased decisions and approvals preserve source AGENT/OPERATOR labels and links to physical records. Source event identifiers are distinct from C0-C4.

Phase	AGENT record	Observation / conclusion	OPERATOR intervention	Evidence event / checkpoint
Reconnaissance	Reverse-engineer the retained HEX firmware image; analyse the reverse-engineered firmware using disassembly, documentation and instruction semantics; require strict status and ciphertext parsing.	Binary-derived security functions and instruction-level endpoints.	Supply scope, target artifacts and initial admissible laser-power range; confirm exclusions.	Firmware F0-F4; decisions D0-D3; C0-C1.
Enumeration and hypotheses	Connect the proposed evaluation to candidate regions and a complete-response classifier.	PEM activity and macro-block exclusions prioritize candidate digital logic for physical confirmation.	Provide the block interpretation and approve masked physical testing.	Localization decision stream and approval; C1-C2.
Localization	Execute the authorized bounded sequence and stop at the first complete shifted-load template.	Two fault-free responses and one instruction-level skip in three pulses; evidence-guided.	Authorize the block and bounded search region.	Stage-04 run and approval; C2-C4.
Characterization audit	Record 18 templates / 35 shots; reject duplicated shot-specific traces; validate corrected acquisition without laser.	One extra pulse across interrupted/resumed runs; 14 duplicate-hash groups; four fresh post-fix controls.	Approve the 35-shot characterization and revise the minimum trace-density criterion.	Characterization C0-C8; C2-C3.
Authentication objective	Separate counter/state anomalies from authenticated acceptance; retain earlier failed or differently conditioned runs.	A correct-PIN lockout result does not satisfy the wrong-PIN objective.	Reject that result as the main endpoint and require a secret-independent attempt.	PIN P18; earlier analysis; C3.
Authentication confirmation	Run the amended bounded search, retain negative and anomalous responses, query authenticated state separately and stop.	84 selected-search pulses; acceptance plus authenticated-state confirmation; four recovery controls.	Request broader timing coverage and authorize continuation only after a laser-off recovery control passes.	PIN P21-P23 and approval; C2-C4.
AES measurement gates	Identify trigger-initialization defect; correct instrumentation; require fresh known-answer acquisitions.	Eight initial laser-off validation traces pass. Later empty-response logging defect is retained in pulse accounting.	Connect the trigger and authorize the AES phase.	AES A0-A7; C2-C3.
AES confirmation and reporting	Request controls after a sparse output; audit filename/centering defects; obtain four distinct centred traces; classify output differences.	14 identical same-condition outputs; 29 fault-free controls; no classical differential fault analysis match.	Request a byte-diversity check, then decline further spatial work and stop at the proof-of-concept objective.	AES A8-A12; C2-C4.