

AITOOLWORX

THE TOPS TRAP: EXECUTIVE SUMMARY

How Edge AI Optimized for Token Efficiency—and Forgot to Engineer for Truth

“Our assertion is not that AI sometimes gets facts wrong. It is that an AI system can produce a well-formulated, seemingly well-reasoned, and convincingly supported answer while contradicting or misrepresenting the very evidence it was explicitly instructed to follow.”

James Childers | Founder, AIToolworx

Author: James Childers | Founder, AIToolworx

Document identifier: AITW-AITR-2026-001-ES

Version 1.01 — First Public Edition, revised | October 2026

SECURE AI. LOCAL CONTROL. TRUSTED DATA.

DOCUMENT METADATA

Field	Entry
Title	The TOPS Trap: Executive Summary
Subtitle	How Edge AI Optimized for Token Efficiency—and Forgot to Engineer for Truth
Author	James Childers, AIToolworx
Document type	Executive summary (companion to the full technical research white paper)
Companion document	The TOPS Trap (AITW-AITR-2026-001), https://tops-trap.aitoolworx.com
Edition / version	Version 1.01 — First Public Edition, revised, companion to Version 1.0 of the full paper
Publication date	October 2026
Document identifier	AITW-AITR-2026-001-ES
Document SHA-256	Listed in the release manifest at https://tops-trap.aitoolworx.com/manifest
Reproducibility package	https://tops-trap.aitoolworx.com/evidence (archive hash in the release manifest)
Suggested citation	Childers, James. The TOPS Trap: Executive Summary. AIToolworx, 2026. Companion to https://doi.org/10.5281/zenodo.23423297
Source verification	Outside facts cited here were checked against primary sources on October 2 and 3, 2026. Exceptions are listed in the full paper’s Evidence and Source Register.
Source register	Appendix A of the full paper
Commercial interest disclosure	AIToolworx has a commercial interest in AI architecture, security consulting, and application of the trust-assessment methodology proposed in the full report. The proposed benchmark is an AIToolworx-developed framework and has not been adopted as an independent industry standard or accredited certification program.

TABLE OF CONTENTS

YOUR AI IS CONFIDENTLY WRONG, AND IT CITES SOURCES TO PROVE IT..... 4

 SIX FINDINGS 5

WHAT WAS TESTED AND HOW 6

 THE SETUP 6

 HOW THE TESTS WORKED 6

 WHY DEEPSEEK WAS EXCLUDED 7

WHAT WAS FOUND 8

 1. EVIDENCE DID NOT GUARANTEE GROUNDING..... 8

 2. HIDDEN INSTRUCTIONS SHAPED THE EFFECTIVE PROMPT 8

 3. FLUENCY AND CITATIONS WERE NOT PROOF 8

 4. BEHAVIOR CHANGED WITH THE TOPIC 8

 5. SPEED AND HARDWARE DID NOT PREDICT TRUST 9

 6. THE BOX SAYS TAIWAN. THE MODEL SAYS ALIBABA. 9

WHAT THE RESULTS DO AND DO NOT ESTABLISH 10

WHAT TO DO ABOUT IT 11

 ENGINEER TRUST AROUND THE MODEL..... 11

 A TRUST BENCHMARK 11

 RECOMMENDATIONS 11

EVIDENCE, LIMITS AND HOW TO CHECK OUR WORK 12

YOUR AI IS CONFIDENTLY WRONG, AND IT CITES SOURCES TO PROVE IT

We handed local AI models a folder of evidence and one rule: answer only from this. Some of those models were also carrying instructions we never wrote, never approved, and could not see: text added automatically by the model's own template or runtime, invisible to the person typing.

Then came the answers. Some broke our rule. One added casualty figures that were nowhere in the folder. One said the folder held nothing relevant when it did. One cited four sources in a single answer and still added claims those sources did not support.

The wrong answers sounded as fluent and sure as the right ones. Confidence did not track the evidence.

The sharpest case: in one model's answers, 45 of 93 claims (48.4%) were unsupported or contradicted by the evidence it was given. Of the 37 claims that carried a citation, 18 (48.6%) cited a source that did not support them. In one answer the model credited a rationale to Andrew Jackson and cited source S1 for it. S1 does not say that.

These were local models running on a Raspberry Pi 5. They are not every AI. But the test is simple, and every run is preserved so that others can check it.

None of this shows up on the spec sheet. Edge AI is usually judged on speed: how efficient, how fast, and how many trillion operations per second (TOPS) the chip delivers. A board that Raspberry Pi lists at \$200 and 40 TOPS (INT4) can run a language model on your desk. Those numbers tell you whether a model can run. They say nothing about whether you can believe it.¹

We call that gap the TOPS Trap.

Performance answers "Can it run?" Trust answers "Can it be trusted?"

AIToolworx tested seven local language-model configurations, with and without an AI accelerator, using fixed evidence packets, preserved prompts and logged runs. Six configurations were tested under controlled conditions. The seventh, DeepSeek R1 Distill Qwen 1.5B, failed a basic instruction-following check during early exploration and was excluded before controlled testing.

Observation first. Attribution only where the evidence supports it.

¹ Raspberry Pi Ltd, "Raspberry Pi AI HAT+ 2," product page, <https://www.raspberrypi.com/products/ai-hat-plus-2/>, accessed October 2, 2026. The page lists the Hailo-10H accelerator at "40 TOPS (INT4)" with 8GB of on-board RAM.

SIX FINDINGS

1. **Evidence does not guarantee grounding.** Giving a model sources and telling it to answer only from them did not reliably keep it inside those sources.
2. **The prompt you type is not the prompt the model receives.** Chat templates and runtimes add instructions the user never sees.
3. **Fluency and citations are not proof.** Well-written answers and citation markers still carried unsupported claims.
4. **Behavior changed with the topic.** Matched China and United States tests produced different behavior in some configurations. The study measures this and does not assume a cause.
5. **Speed did not predict trust.** The configuration with the strongest source discipline ran at about the same speed as accelerator-hosted models that failed more often.

One result points the other way. The tested Gemma configuration repeatedly said when the evidence did not establish a requested fact. That shows source discipline is achievable for a compact local model. It does not show that Gemma is trustworthy in general.

The engineering lesson: the model may reason about the evidence, but the evidence remains authoritative.

WHAT WAS TESTED AND HOW

THE SETUP

The hardware was a Raspberry Pi 5 with the Raspberry Pi AI HAT+ 2, whose Hailo-10H accelerator is listed at 40 TOPS (INT4). Some models ran on the accelerator and some on the CPU alone. Findings apply first to the exact configuration tested, because model revision, quantization, runtime, chat template, system instruction and packaging can each change behavior.

Configuration	Where it ran	What was observed
Qwen 2.5 Instruct 1.5B	Hailo accelerator	Repeated source blending and missing citations; response mode differed between the China and United States conditions.
Qwen 2.5 Coder 1.5B	Hailo accelerator	Said the evidence was absent when it was present (China grounded); unsupported expansion in the United States condition.
Qwen 2 1.5B	Hailo accelerator	Source blending and unsupported additions; repeated factual hallucination and contradiction.
Llama 3.2 3B	Hailo accelerator	Partly grounded when it could be evaluated; runtime failures impaired key comparisons.
Gemma (gemma-4-e2b)	CPU	Strongest source discipline observed; repeatedly identified the limits of the evidence.
Ministral 3 3B Instruct 2512 (Q4_K_M)	CPU (llama.cpp)	Fluent answers with significant source blending; citation behavior differed by topic.
DeepSeek R1 Distill Qwen 1.5B	Hailo accelerator	Failed a basic three-sentence instruction check and was excluded before controlled testing. Not scored.

Table 1. Tested configurations, summarized from Section 12 of the full paper.

HOW THE TESTS WORKED

- Fixed evidence packets, identified by SHA-256 hash, held the only sources the model was allowed to use.
- Prompts told the model to answer only from the packet and to cite packet entries as [S1], [S2], and so on.
- Two matched topic packets, one China-related and one United States-related, used the same test structure.

- Each model was tested in four conditions: with and without the study’s system message, each closed-book and evidence-grounded.
- Requests, responses, runtime logs and model hashes were preserved for every run. The Ministral record was run at temperature 0 with three trials per condition.
- Exploratory observations are reported separately from the controlled tests.

WHY DEEPSEEK WAS EXCLUDED

DeepSeek R1 Distill Qwen 1.5B ran on the Hailo accelerator and was a candidate for the study. It never reached the controlled tests. Before any evidence-grounding began, it was given a deliberately simple task: explain in three concise sentences why a local AI assistant is useful without Internet access. That was a minimal instruction-following check, not a test of politics or sources.

It failed on several counts at once. It did not stay within three sentences. It exposed a long stream of unrequested step-by-step reasoning text, including a visible `</think>` marker. It then gave a second answer, and it used far more output than the task needed.

A configuration that cannot meet that minimum cannot support a meaningful controlled comparison, so it was excluded before confirmatory testing. The exclusion is kept in the record instead of the model being quietly dropped. A separate exploratory prompt on a historical topic also produced a reasoning trace with factual confusion; that is not a controlled result.

What this does not show is why it happened. The cause could lie in the underlying model, the distilled architecture, the accelerator packaging, the chat template or the runtime. DeepSeek receives no grounding, topic or reliability score, and nothing here ranks it against the six controlled configurations.

WHAT WAS FOUND

1. EVIDENCE DID NOT GUARANTEE GROUNDING

Models added facts that were not in the packet, ignored the citation requirement, or claimed the packet lacked information it contained. Qwen 2.5 Coder showed the last error: in the China grounded condition it said the evidence was absent when it was present.

2. HIDDEN INSTRUCTIONS SHAPED THE EFFECTIVE PROMPT

In all 32 runs of Qwen 2.5 Instruct and Qwen 2.5 Coder with no system message, the runtime log shows the sentence “You are Qwen, created by Alibaba Cloud. You are a helpful assistant.” added to the prompt. It was absent when the study’s system message was supplied, and absent from all 64 Qwen 2 and Llama 3.2 logs.²

The same sentence is in the official upstream Qwen2.5-1.5B-Instruct chat template. That shows the sentence exists upstream. It does not show which template the accelerator-hosted build used, or who put it there. In a separate test, a request for an exact response was overridden by a phrase from the system layer.³

3. FLUENCY AND CITATIONS WERE NOT PROOF

Minstral’s answers were fluent and organized, which made the provenance failures harder to see. Its locked record was adjudicated claim by claim: of 93 claims, 15 (16.1%) were supported by the packet and 45 (48.4%) were unsupported or contradicted. Of the 37 claims that carried a citation, 5 (13.5%) cited correctly and 18 (48.6%) cited a source that did not support the claim. In the three China grounded trials shown, it gave no citations at all.⁴

A citation shows only that the model produced a citation marker. Source adherence requires a verifiable link between the claim and the evidence.

4. BEHAVIOR CHANGED WITH THE TOPIC

In matched tests, some configurations answered one topic substantively and gave a non-answer on the other. Qwen 2.5 Instruct reversed its pattern when the system message changed. That does

² Counts come from 128 runtime logs of the four accelerator-hosted models. Each log is listed with its SHA-256 in the full paper, Appendix C (claim C3).

³ Qwen/Qwen2.5-1.5B-Instruct, tokenizer_config.json (chat_template), Hugging Face repository revision 989aa7980e4cf806f80c7fef2b1adb7bc71aa306, https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct/raw/main/tokenizer_config.json, accessed October 2, 2026.

⁴ Claim-level adjudication of the locked Minstral record, summary file in the reproducibility package; see the full paper, Section 7 and Appendix C (claims C1 and C2).

not fit a simple model in which one topic is permanently suppressed. The results show that topic-conditioned behavior has to be measured. They do not show why it occurs.

5. SPEED AND HARDWARE DID NOT PREDICT TRUST

Gemma ran on the CPU alone at roughly 6 to 7 tokens per second (mean 6.7) and showed the strongest source discipline. Several accelerator-hosted 1.5B models ran at similar speeds with repeated grounding failures. Ministral ran at roughly 2.3 to 3.6 tokens per second and produced fluent answers with weak source separation. Computational efficiency and reliability behaved as separate dimensions.

6. THE BOX SAYS TAIWAN. THE MODEL SAYS ALIBABA.

Check the box. Raspberry Pi. Empowered by Hailo. Made in Taiwan.⁵ Not one mention of Alibaba or China: not on the box, Raspberry Pi’s product page or Hailo’s launch announcement, and Raspberry Pi’s launch article lists Qwen as a model name without saying whose it is.⁶ Yet the Qwen models running on that board were built by Alibaba, in China,⁷ and the instruction the study found in the prompt names Alibaba.⁸ The label says where the hardware was made. It does not say who trained the model, who wrote its templates, or who can change its instructions.

Where something was made, or where its maker is headquartered, is not its supply chain. The paper tracks three.⁹ Physical: the hardware. Software and model: the weights, runtime and chat template. Control: the instructions and policies that steer behavior, and who can change them. A trusted device is not a trusted model, and local does not mean every rule was written locally.

This is a supply-chain question, not a nationality test. The study does not show who inserted the sentence, or why the models behaved as they did.

⁵ Author’s photographs of the retail box, taken October 2, 2026: “Made in Taiwan,” “Raspberry Pi Ltd, CB4 0AB (GB), D01 H104 (IE),” “Empowered by HAILO,” “For Raspberry Pi 5.” Evidence archive, 05_Hardware-Evidence; full paper, Appendix A, source S9.

⁶ Read October 16, 2026. Raspberry Pi Ltd, AI HAT+ 2 product page, raspberrypi.com/products/ai-hat-plus-2/ (names no model), and launch article (full paper, Appendix A, source S22; lists Qwen2.5-Coder, Qwen2.5-Instruct and Qwen2; no mention of Alibaba or China). Hailo’s July 22, 2025 press release, as syndicated by EE Journal, eejournal.com/industry_news/hailo-announces-general-availability-of-hailo-10h-edge-ai-accelerator-with-generative-ai-capabilities/ (no mention of Qwen, Alibaba or China; Hailo’s own copy blocks automated access).

⁷ QwenLM/Qwen3 repository description: “Qwen3 is the large language model series developed by Qwen team, Alibaba Cloud,” github.com/QwenLM/Qwen3. Alibaba Group, “Corporate Information FAQ”: principal executive offices in Hangzhou, China, alibabagroup.com/en-US/faqs-corporate-information. Read October 16, 2026.

⁸ In 32 of 32 runs of Qwen 2.5 Instruct and Qwen 2.5 Coder with no system message, the prompt contains “You are Qwen, created by Alibaba Cloud. You are a helpful assistant.” (full paper, Appendix C, claim C3). The same sentence is in the upstream template, Qwen/Qwen2.5-1.5B-Instruct tokenizer_config.json, revision 989aa7980e4cf806f80c7fef2b1adb7bc71aa306, huggingface.co/Qwen/Qwen2.5-1.5B-Instruct.

⁹ Full paper, Section 9, pages 187 to 193: “The Country Printed on the Box Does Not Identify Who Shaped the AI,” “Three Supply Chains Must Be Tracked,” “This Is a Supply-Chain Risk, Not a Nationality Test.”

WHAT THE RESULTS DO AND DO NOT ESTABLISH

The study separates what it observed from why it happened. A model behavior shows that the behavior occurred. It does not, by itself, show the cause.

Supported by the tests	Not established by the tests
Source grounding was not reliable across the tested configurations.	CCP direction of the observed outputs.
Retrieval is not provenance.	Malicious intent by any vendor.
The user-visible prompt was not always the effective prompt.	That Alibaba authored every observed template element.
Model provenance alone was insufficient; execution and control provenance also mattered.	That Hailo caused the model failures.
Some configurations showed topic-conditioned behavior.	That Hailo has no engineering value.
Confidence did not reliably track evidence or correctness.	That local models are inherently unreliable.
Gemma showed the strongest source discipline of the configurations tested (a limited claim).	That CPU inference is inherently more trustworthy.

Table 2. Summarized from Section 14 of the full paper.

A lack of response from a vendor is recorded as a fact. It is not treated as evidence of intent or responsibility.

WHAT TO DO ABOUT IT

ENGINEER TRUST AROUND THE MODEL

The investigation began as a search for a capable local model and ended as a requirement: assume every model can fail, and build the system so that failure does not become authoritative.

1. The model can propose; policy and evidence constrain it.
2. Verification happens outside the model.
3. Authorization is separate from generation.
4. Logging does not depend on the model.
5. Track all three supply chains, and know who can change the instructions, not just who built the hardware.

Retrieval is not provenance, citations are not proof, local control is not complete control, and TOPS are not trust.

A TRUST BENCHMARK

The full paper proposes an Edge-AI Trust Benchmark: 16 trust domains in five families (evidence integrity, epistemic discipline, control and provenance, operational reliability, and system governance). It reports a trust profile beside a separate performance profile, and critical failures cannot be averaged away. The benchmark is an AIToolworx framework and has not been adopted as an independent standard.

RECOMMENDATIONS

- Stop selling compute as trust. Publish trust characteristics alongside performance characteristics.
- Identify the exact artifact: model, revision, quantization, conversion path, tokenizer and chat template.
- Treat chat templates as versioned software, and make default system instructions visible to operators.
- Document the full model-conversion chain and keep a model bill of materials.
- Do not make the model responsible for enforcing its own evidence boundary.
- Verify citations independently; a citation is not proof of grounding.
- Test for false absence as well as hallucination.
- Separate observation from attribution when reporting results.

EVIDENCE, LIMITS AND HOW TO CHECK OUR WORK

- **Scope.** The findings apply to the exact configurations tested, on one hardware platform and two topic packets. They are not claims about all local models or all edge AI.
- **Reproducibility.** Prompts, evidence packets, raw outputs, runtime logs, scoring records and model hashes accompany the full paper: <https://tops-trap.aitoolworx.com/evidence>. Hashes for every file are in the release manifest at <https://tops-trap.aitoolworx.com/manifest>. Model binaries are included only where licenses permit; otherwise the package gives hashes and acquisition information.
- **Sources.** Every outside fact is listed with its primary source, the date it was checked, and its status in the full paper's Evidence and Source Register.
- **Vendor requests.** Requests for technical comment were sent on September 12, 2026 to Raspberry Pi, Hailo, Qwen/Alibaba, DeepSeek and Meta. For the Ministral finding, the model itself reviewed the claim-level adjudication in-app, and Mistral AI was notified by email on October 6, 2026. No response had been received from any of these organizations as of October 16, 2026.
- **Adjudication.** The Ministral adjudication was done manually, claim by claim. The claim-by-claim file is part of the package so that others can re-score it.
- **DeepSeek.** DeepSeek R1 Distill Qwen 1.5B failed a basic preliminary check and was excluded before controlled testing. Its preserved test artifact remains in the record.
- **Read more.** The full paper gives the method, the complete results, and the context for every finding above: The TOPS Trap, AITW-AITR-2026-001, <https://tops-trap.aitoolworx.com>.
- **Version.** Version 1.01 — First Public Edition, revised, October 2026; companion to Version 1.0 of the full paper.