Public edition

First Device Models, Trained In-House

How THOX Built the Training Rails, Proved Them with a Voice Command Adapter, and Published ThoxFlux-350M on Kickstarter Launch Day

THOX.ai Research · Prepared by: Sadie Weaving, Chief Agentic Officer, THOX.ai · Accountable human author: Tommy Xaypanya, CTO & Co-Founder, THOX.ai (corresponding author) · ORCID 0009-0000-4425-9565

STATUS OF RECORD (SECOND EDITION)

Status as of October 30, 2026 (second edition): All five edge models completed training, GGUF conversion, the GATE49 factuality gate, model card rewrites, and Ollama publication. Every gate verdict in these waves is FAIL under the gate's zero-hard-failure policy; the totals below are scores, not passes. GATE49's verified category (12 questions) is the paper's primary factuality figure; the 49-question total is reported beside it because 37 of the 49 items (26 unknown, 10 known_false, 1 pricing) are refusal-rewarded by design, so the total measures abstention discipline plus verified recall (Section 8.3). Wave A (205 row corpus, weights only): ThoxEdge-0.5B 21/49, ThoxEdge-0.8B 19/49, ThoxFlux-350M 4/49, ThoxNano-135M 2/49, ThoxNano-360M gate error (repetition loop, no score); per-category splits for Wave A were not retained in the finishing record. Wave B (982 row corpus, October 28, 2026), reported as total (verified recall): ThoxEdge-0.5B 47/49 (11/12), ThoxFlux-350M 20/49 (3/12), ThoxNano-135M 23/49 (1/12). ThoxNano-360M scored 23/49 (3/12) in its first completed gate; it had no valid Wave A score, so it is reported separately rather than as an improvement. ThoxEdge-0.8B's Wave B artifact has never been validly measured: its gate read was invalidated as a serving artifact and the artifact was held unpublished (Section 11.5). ThoxClip-Qwen3-0.6B finished at 22/49 (2/12) with a grounding-required model card (Section 11.7). No model in these waves has been qualified on physical device hardware.

47/49 · verified 11/12

ThoxEdge-0.5B, Wave B gate (mean 4.837/5) — Wave A was 21/49; every verdict in both waves is FAIL under the zero-hard-failure policy

MEASURED · GATE49

25/49

Constant-abstainer floor — the prescribed decline responder through the same scorer; Wave totals are read against it

MEASURED · GATE49 SCORER

982 rows

Wave B corpus (838 newly authored + 144 retained after hygiene), replacing the 205-row Wave A corpus

CALCULATED · CONSTRUCTION RECORD

0/14 exact

Voice pilot held-out result: 1/14 function names, 0/14 full exact matches of 14 — weak as performance, strong as pipeline evidence

MEASURED · PILOT ARTIFACT

CONTENTS

1. Executive summary
2. Background: four devices and a missing model layer
3. The rails: a repeatable training and verification system
4. Program 1: the voice command pilot
5. Program 2: the edge fast track wave
6. Failure taxonomy: how the work was actually done
7. The model card rule
8. Finishing pipeline: Wave A completion record
9. Limitations and honest boundaries
10. Evidence and provenance register
11. Wave B: the corpus expansion and second gate

12. Forward: the Wave C flagship program

Appendix A. Per-run configuration and provenance tables

Appendix B. The ThoxAir voice bridge bench record

Appendix C. Glossary

Document control

Your AI. Your Data. Your Rules.™ · THOX.ai LLC · Process v1.0 reissue (Third Edition) of the Second Edition, hardened paper-version 2.1 (October 30, 2026). DOI: pending deposit — none issued. Evidence: exp/results/summary.json in the artifact bundle.

PROCESS V1.0 REISSUE — THIRD EDITION

This is the Process v1.0 pack reissue (Third Edition, v3.0, 2026-11-04) of the Second Edition (hardened paper-version 2.1, October 30, 2026). Content is unchanged from the second edition except the changes recorded in the Document control section: the Process section 4 security sweep (host classes; section 10.1 source names) and this edition's own record. Every number traces to the pack's single evidence source, exp/results/summary.json.

Date: October 27, 2026 (first edition)

Edition: Third edition — November 4, 2026 (Process v1.0 reissue of the second edition, as hardened to paper-version 2.1 on October 30, 2026)

Prepared by: Sadie Weaving, Chief Agentic Officer, THOX.ai

Accountable human author: Tommy Xaypanya, CTO & Co-Founder, THOX.ai

Authorship disclosure: This paper was drafted and assembled by Sadie Weaving, an agentic officer of THOX.ai, from the program's run records, under the direction of the accountable human author named above. "Verified" in this paper means checked against those records (job logs, repository contents, model cards, status files) by the operator who ran the work; the hardening revision recorded in the document control section adds an independent Research Loop audit and an external hostile review of the text.

Document status: Technical record and method paper. This paper reports only results verified in job logs, repository contents, model cards, and program status records. Training completion, artifact publication, model quality, and hardware qualification are treated as separate claims.

Status as of October 30, 2026 (second edition): All five edge models completed training, GGUF conversion, the GATE49 factuality gate, model card rewrites, and Ollama publication. Every gate verdict in these waves is FAIL under the gate's zero-hard-failure policy; the totals below are scores, not passes. GATE49's verified category (12 questions) is the paper's primary factuality figure; the 49-question total is reported beside it because 37 of the 49 items (26 unknown, 10 known_false, 1 pricing) are refusal-rewarded by design, so the total measures abstention discipline plus verified recall (Section 8.3). Wave A (205 row corpus, weights only): ThoxEdge-0.5B 21/49, ThoxEdge-0.8B 19/49, ThoxFlux-350M 4/49, ThoxNano-135M 2/49, ThoxNano-360M gate error (repetition loop, no score); per-category splits for Wave A were not retained in the finishing record. Wave B (982 row corpus, October 28, 2026), reported as total (verified recall): ThoxEdge-0.5B 47/49 (11/12), ThoxFlux-350M 20/49 (3/12), ThoxNano-135M 23/49 (1/12). ThoxNano-360M scored 23/49 (3/12) in its first completed gate; it had no valid Wave A score, so it is reported separately rather than as an improvement. ThoxEdge-0.8B's Wave B artifact has never been validly measured: its gate read was invalidated as a serving artifact and the artifact was held unpublished (Section 11.5). ThoxClip-Qwen3-0.6B finished at 22/49 (2/12) with a grounding-required model card (Section 11.7). No model in these waves has been qualified on physical device hardware.

1 Executive summary

On October 27, 2026, THOX moved from having a registry of planned device models to operating a working, in-house training pipeline for those models. The work ran on the same day as the THOX Kickstarter launch and had two connected programs.

The first program was a deliberately small pilot. A LoRA adapter for LiquidAI/LFM3-1.7B was trained to translate spoken style THOX commands into one exact JSON function call. The seed contained 49 rows authored by Sadie Weaving (see the authorship disclosure in the front matter): 35 for training, 12 held out for unseen

phrasings of trained functions, and 2 held out for a function never shown in training. The sixth launch completed training and evaluation and published the adapter and its results.json to Thox-ai/lfm3-omni-voice-lora-pilot. Its held out result was 1 of 14 function names correct and 0 of 14 full exact matches. Those numbers are weak as model performance and strong as pipeline evidence. The purpose of the pilot was to prove the path from data to training, evaluation, artifact publication, and an honest model card. It did that. It did not produce a production voice adapter.

The second program addressed the pending assistant models for the four Kickstarter devices. The factory registry held five queued models with skeleton repositories and no trained weights: two for ThoxMini, two for ThoxClip, and one for ThoxKey. ThoxAir had no queued assistant model, a gap recorded in the program inventory rather than filled by assertion. THOX used a pathfinder discipline: train ThoxFlux-350M first, verify the trained artifact in its repository, and only then launch the other four.

ThoxFlux-350M training completed on pathfinder run #4, with 78 of 78 steps, a training loss of 3.463, and logged loss moving from 6.257 to 2.544. That run then failed to publish because the training script called a huggingface_hub upload function positionally against its keyword-only signature (Section 6.2, class 6). Run #5 repeated the completed training path with the publication fix and published the merged weights and training summary. Repository verification found the new SHA d0e42dd232c58b17451f911c166ee23c9af28339 and a training_summary.json recording 205 rows, 3 epochs, LoRA rank 32, and training loss 3.463227467659192. Only after that artifact gate passed did the remaining four models launch.

The central result is therefore not a single model. It is an operating method, applied across both programs with one recorded exception (the Wave B retrain of Section 11 ran locally, off these rails, for reasons its run record states):

Put the training code and launch configuration under version control in the dataset repository that the job downloads.

Launch under the Hugging Face organization namespace that actually holds the prepaid credits.

Pass private repository access explicitly as a job secret; do not assume that a hosted job inherits the owner's repository permissions.

Make hardware capability part of the training configuration, including numeric precision.

Verify artifacts in the target repository instead of treating a job status as proof of publication.

Gate parallel work on one verified pathfinder.

Record every failure, fix, job identifier, and repository state in a durable status file.

Ship every model repository with a comprehensive card at creation and fill its evaluation table only from verified run artifacts.

Across the two programs and the finishing and Wave B work that followed them, THOX encountered eleven distinct failure classes: private repository authorization, remote code trust, a removed training configuration argument, a dataset shape contract, an unsupported numeric precision on the selected GPU, a positional upload call against a keyword-only function signature, a tokenizer chat template contract, the removed configuration argument recurring in the Wave B trainer, a tokenizer handoff defect, a model-load dependency defect, and the factory finisher's breakage against the current Ollama release. Each failure was found by a short run, diagnosed from logs or resolved package code, fixed once, and checked again. Two fixes moved between the pilot and edge programs in opposite directions. That propagation is as important as any individual correction because it turned isolated debugging into shared infrastructure.

The work remains bounded. The pilot seed is too small to support a quality claim. The edge corpus contains 205 training rows and was intended to give each base model a

THOX identity, house voice, and tool routing skin; it does not create new general capability. Of those three intents, only factuality was measured in these waves, by the factuality gate named GATE49 in this paper (Section 8.3), together with a single identity smoke prompt per model (Section 8.2). House voice and tool routing were trained but not evaluated in these waves, and no claim about them should be read into the gate scores. The finishing pipeline for all five edge models (GGUF conversion, GATE49, trained model cards, Ollama publication, and smoke tests) completed on October 28, 2026 and is recorded in Section 8. The same day's Wave B retrain on the expanded 982 row corpus, and its gate results, are recorded in Section 11. Two finishing items remain open: the factory registry status update was not performed in these runs, and no model has been qualified on physical device hardware. No result in this paper should be read as physical device qualification.

Two framing notes. "First" in the title means the first models produced by the factory training program for the device assistant roles; the smaller per-device role models listed in Section 2.2 predate the factory program and are a different class of artifact. And a conflict-of-interest note: the authors are officers of THOX.ai, the first edition of this paper was published during the THOX Kickstarter campaign that launched on October 27, 2026, and the models described here belong to the device ecosystem that campaign offers. Readers should weigh the paper's claims with that interest in view; the claim boundaries in Section 9.1 are written to make that weighing possible.

2 Background: four devices and a missing model layer

2.1 The THOX proposition

THOX is building a private local AI ecosystem. The design premise is that personal and operational data should be processed on hardware controlled by the user, with larger resources reached through the user's own mesh when a task requires escalation. Device models are therefore not decorative variants of a cloud model. They are the local layer that lets a device recognize intent, respond in the THOX house voice, protect routine work from unnecessary escalation, and continue to provide useful behavior close to the user.

That premise creates a model operations requirement. A device family needs models matched to different roles and hardware classes, and each model needs training data, a reproducible run, an evaluation record, a deployable artifact, and a card that states what is known and what is not. A registry entry alone does not meet that requirement.

2.2 Registry state at the start of the day

An inventory of ttracx/thoxllm-factory on October 27, 2026 separated models that were already published from queued skeletons. The four device role models were already published: thoxkey-9m-role, thoxair-16m-role, thoxclip-9m-role, and thoxmini-125m-role. ThoxMini also had thoxmini-3b and the platform model lfm3-omni-2b. The pending assistant set was specific:

Device	Queued model	Base model recorded in the registry	Program
ThoxKey	ThoxNano-135M	HuggingFaceTB/SmolLM-135M	Edge fast track
ThoxAir	No queued assistant model	Not applicable	Gap recorded
ThoxMini	ThoxEdge-0.8B	Qwen/Qwen3.5-0.8B	Edge fast track
ThoxMini	ThoxEdge-0.5B	Qwen/Qwen2.5-0.5B-Instruct	Edge fast track
ThoxClip	ThoxFlux-350M	LiquidAI/LFM2.5-350M	Edge fast track pathfinder
ThoxClip	ThoxNano-360M	HuggingFaceTB/SmolLM-360M	Edge fast track

The ThoxAir result required explicit treatment. ThoxAir ships thoxair-16m-role today and is designed to use the 350M class tier and mesh resources for larger work. It did not have a queued assistant model comparable to the five models in the edge wave.

This paper does not silently convert the ThoxFlux-350M run into a ThoxAir model. The gap remains a product and model planning item.

Three other registry entries were excluded because they were decisions, not executable training jobs. ThoxKey-TLM had no base model and its own card proposed retirement in favor of ThoxKey-9M-role. ThoxMesh-Head-C6 recorded a placeholder base. ThoxUSB-Edge duplicated the ThoxEdge-0.8B base, was outside the Kickstarter four, and had no locked ThoxUSB hardware specification. Excluding them kept the wave focused on work that could be verified.

2.3 Why per device adapters and small models

The LFM3 evaluation performed the same day reinforced a practical pattern from Liquid AI's own fine tuning material: one LoRA adapter per device task, trained on hundreds to thousands of examples, followed by merging and quantization for the target device class. The pilot applied that pattern to command understanding. The edge wave applied supervised fine tuning to small base models so that each device tier would carry the THOX identity and routing behavior in its own artifact.

Small models have a specific role in this architecture. They handle bounded work locally: identity, house voice, command shape, routing, briefings, notes, status, and other structured tasks. They are not presented as substitutes for larger models on open ended reasoning. The architecture gains leverage by placing the smallest sufficient model at the device and reserving larger local or mesh resources for work that justifies escalation.

2.4 Related work

The methods used here are standard, and this paper claims no novelty for them. LoRA (Hu et al., 2022) is the adapter method throughout: a low-rank update trained against a frozen base, then merged. QLoRA (Dettmers et al., 2023) is the quantized variant of the same idea; the Waves A/B runs used full-precision LoRA rather than QLoRA, and it is cited because the factory's later work (the Wave C paper) trains both ways and compares them. Training ran on TRL (von Werra et al., 2020), Hugging Face's supervised fine-tuning and reinforcement learning library, following the pattern of Liquid AI's own fine-tuning material for the LFM family, specifically the lfm3_omni_sft_with_transformers recipe in the Liquid AI cookbook, which the pilot's data plan mirrors (including its unseen-tools test design).

The evaluation side has an established literature the paper should have engaged from the first edition. Selective prediction and abstention, the behavior GATE49's unknown and pricing categories reward, is formalized for question answering by Kamath, Jia and Liang (2020), who show that a system which may decline can trade coverage for accuracy under distribution shift. Short-form factuality measurement at frontier scale is exemplified by SimpleQA (Wei et al., 2024), which grades answers as correct, incorrect, or not attempted, the same three-way distinction GATE49's verdicts approximate with its pass, soft-fail, and hard-fail outcomes; HaluEval (Li et al., 2023) is a larger hallucination-evaluation benchmark built from generated and human-annotated examples. GATE49 differs from all of these in scope, not in kind: it is a 49-item, single-domain instrument generated from the company's own facts card, and Section 9.7 states the independence limits that follow from that. No result in this paper is comparable to a score on any of the cited benchmarks.

References for this section: Hu, E. J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022. arXiv:2106.09685. — Dettmers, T., et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS 2023. arXiv:2305.14314. — von Werra, L., et al. (2020). TRL: Transformer Reinforcement Learning. Hugging Face software library. — Liquid AI (2025). LFM3 Omni SFT with Transformers, Liquid AI cookbook (Liquid4All/cookbook). — Kamath, A., Jia, R., and Liang, P. (2020). Selective Question Answering under Domain Shift. ACL 2020, pages 5684–5696. arXiv:2006.09462. — Wei, J., et al. (2024). Measuring Short-Form Factuality in Large Language Models (SimpleQA). arXiv:2411.04368. — Li, J., et al. (2023). HaluEval: A

Large-Scale Hallucination Evaluation Benchmark for Large Language Models. EMNLP 2023. arXiv:2305.11747.

3 The rails: a repeatable training and verification system

3.1 System overview

The training rails built on October 27 consist of six parts:

Versioned training bundles. Training scripts, bootstrap shell scripts, and dataset cards live in Hugging Face dataset repositories. A job downloads the bundle at launch, so the code in the repository is the code that runs.

A local launcher. launch_job.py, run on Tommy's Mac, creates Hugging Face Jobs through the locally authenticated hf command line interface and injects the access token as a job secret.

Organization compute. Jobs run in the Thox-ai organization namespace, where the prepaid credits reside.

Task specific data. The pilot uses its private seed dataset. The edge wave uses a parameterized bundle pointed at the factory version 5 corpus.

Repository publication. The pilot publishes a LoRA adapter and results.json. The edge trainer publishes merged weights and training_summary.json into the model's existing repository.

Scheduled verification. A later check inspects job state and logs, lists the target repository, compares its state with the pre run baseline, records the result, and enforces the next gate.

The rails were not designed as an abstract platform and then tested. They were assembled under live work, corrected at each failed boundary, and generalized only after a run proved the correction.

3.2 The organization namespace and the 402 lesson

Hugging Face prepaid credits were held by the Thox-ai organization, not by the tommytracx user account. A launch attempted under the user namespace failed with HTTP 402 and the message that the prepaid credit balance was insufficient. The same class of launch succeeded under the organization namespace.

This became a preflight rule:

Resolve the billing namespace before launch.

Do not infer available credit from the identity of the signed in user.

Record the namespace in the launch record.

Treat a billing failure as configuration evidence, not as a model or code failure.

All production launches in the two programs used the Thox-ai namespace.

3.3 Private data access and the token secret pattern

The first pilot job failed before training because its bootstrap could not download the private training bundle. Hugging Face Jobs did not inject a token able to read the owner's private repositories. The run therefore received an unauthenticated response when it attempted to fetch the files it needed.

The correction was to make access explicit. launch_job.py runs on Tommy's Mac, where the Hugging Face command line is authenticated as an organization member. The launcher obtains the local token in process through get_token(), passes it to the job as the HF_TOKEN secret, and does not print it. The job bootstrap and training script can then download the private dataset repository and publish to the intended organization repository.

Three rules follow:

A hosted job identity is not automatically the repository owner's identity.

Private repository access must be designed as a capability grant, not assumed.

A secret used by a job should be passed through the platform's secret mechanism and kept out of logs, source files, and reports.

Every later launch in both programs used this pattern.

Corrective note (hardening revision, 2026-10-30). The pattern as run carried the operator's own Hugging Face token into each job. That token holds whatever scopes the operator's account has, across all of the operator's repositories, which is broader access than any single program needs. The pattern is recorded above because it is what ran; it is not a recommendation. The required pattern going forward is a fine-grained, repository-scoped token created per program, and this correction is recorded as a corrective action against the rails, not as a description of the October runs.

3.4 Versioned bundles and why local fixes do not count

Both training programs download their scripts from dataset repositories at job start. This design has a consequence that appeared repeatedly: editing a local script does not change the next job unless the corrected script is uploaded to the dataset repository and verified there.

The operating procedure became:

Diagnose the failure from the job log.

Patch the local canonical script.

Compile or inspect the script where possible.

Upload the corrected script to the dataset repository.

Record the dataset repository commit.

Re download or otherwise verify the uploaded file.

Relaunch only the affected program or pathfinder.

This is slower than editing a file and pressing launch again. It is also the difference between a reproducible run and a run whose code exists only in a transient working directory.

3.5 Compute flavors and observed prices

The programs used three Hugging Face Jobs flavors. Prices below are the rates observed on October 27, 2026. They are recorded as run context, not as a current price commitment.

Flavor	Work assigned	Observed price	Selection reason
a100-large	LFM3 voice command pilot	$2.50 per hour	A100 class compute is the known good shape for the LFM3 training and evaluation run
a10g-small	ThoxEdge-0.8B follower	$1.00 per hour	Larger edge model in the follower wave
t4-small	ThoxFlux-350M pathfinder and three followers	$0.40 per hour	Small models and a controlled pathfinder cost

The T4 selection exposed a configuration error that is covered in the failure taxonomy. A low cost flavor is useful only when its accelerator supports the precision and kernels selected by the training script. Hardware capability is therefore part of the software contract.

Program totals are only partially recoverable from the records. The job runtimes the records preserve (pilot launches #2, #4, #5 and #6; all five ThoxFlux-350M pathfinder runs; the two Nano follower relaunches) sum to 2,073 seconds, about 0.58 GPU-hours across the three flavors. Runtimes for pilot launches #1 and #3 and for the two ThoxEdge follower jobs were not recorded, and the Wave B retrains ran on a

local fleet GPU (recorded runtimes 49 to 207 seconds per model, Section 11.3) rather than on a billed flavor, so no complete program GPU-hour or cost total is stated anywhere in this paper.

3.6 Scheduled checks as the operating loop

Each launch was paired with a scheduled verification task. The check was not a passive timer. Its instructions required the operator to:

inspect the job rather than rely on the launch response;

read the end of the job log;

identify the precise failing step if the job failed;

list files in the target repository;

compare repository modification state with the pre training baseline;

update the program status file;

relaunch only within the authority granted for that check; and

report the result with the next gate stated plainly.

This loop caught a critical distinction. A job can reach a terminal platform state while the intended artifact is absent. The pilot's fifth launch was marked COMPLETED after training and evaluation, but repository inspection showed no adapter and no results.json, because the script caught its publication exception and exited. The check therefore treated publication as incomplete. The edge pathfinder checks applied the same rule by comparing the ThoxFlux-350M repository with its October 24 base repack SHA and requiring a new training_summary.json.

The verification loop is now part of the rails. Job state is one signal. Logs are a second. Repository contents are the publication evidence.

4 Program 1: the voice command pilot

4.1 Objective

The pilot objective was narrow: create one LoRA adapter that receives a THOX command in spoken style text and returns exactly one function call as JSON. Target behavior included device status, mesh health, briefings, note capture, summarization, device mode changes, reminders, vision description, transcription, translation, and ThoxMind routing advice.

Success was defined by function name accuracy and full exact match on frozen held out data. It was not defined by whether the generated text sounded plausible. That distinction matters for a command layer. A fluent answer with the wrong function name is a failed command.

4.2 Text proof before audio training

The platform direction is an LFM3 Omni voice adapter trained on spoken input. The pilot first proved the pipeline on the text core, LiquidAI/LFM3-1.7B, using the TRL LoRA pattern. This choice separated two risks:

whether THOX could run the dataset, training, evaluation, and publication path; and

whether an audio in Omni model could learn the command task from rendered speech.

The published pilot adapter is therefore a text form pipeline proof. It is not evidence that the Omni audio variant has been trained. The Omni variant remains the named follow on after the version 1 dataset is expanded and its audio is rendered.

4.3 Seed design

Sadie Weaving authored seed.jsonl for the dataset repository Thox-ai/lfm3-omni-voice-seed. (Authorship here and throughout is agentic authorship under founder

direction; see the disclosure in the front matter.) The seed has 49 rows across 12 functions. Its split design matters more than its size:

Split	Rows	Purpose
Train	35	Teach the output contract and trained functions
Heldout	12	Test unseen phrasings of trained functions
Heldout unseen function	2	Test timer.start, which never appears in training
Total	49	Pilot seed

The unseen function split follows a strict evaluation idea. A command model should not invent a call for a function it has not been given. It should ask, abstain, or route the request through an approved fallback. The pilot results show that this behavior was not achieved by the seed run. That failure is useful because the split exposed it instead of hiding it inside aggregate accuracy.

The function catalog was:

Function	Purpose
mesh.device_status	Online state of one device or all devices
mesh.health	Mesh wide health check
briefing.generate	Morning, evening, or weekly briefing
notes.capture	Capture a note
content.summarize	Summarize inbox, chat, document, or calendar material
device.set_mode	Set quiet, do not disturb, performance, or low power mode
reminders.create	Create a reminder
vision.describe	Describe a camera view or photo
audio.transcribe	Transcribe a recording
audio.translate	Translate speech
thoxmind.route_advice	Decide whether a request should run on device or escalate
timer.start	Held out unseen function probe

4.4 Training configuration

The pilot used TRL SFTTrainer with PEFT LoRA. The script and bootstrap are versioned in the dataset repository.

Setting	Value
Base model	LiquidAI/LFM3-1.7B
Adapter method	LoRA
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.0
Target modules	q, k, v, o, gate, up, and down projections
Training rows	35
Batch size per device	4 (recorded in the training script; the completed run's 270 steps are 30 epochs at ceil(35 / 4) = 9 steps per epoch)
Epochs	30
Steps in the completed run	270
Learning rate	0.0001
Schedule	Cosine
Warmup	10 percent, expressed through warmup_steps after the version 5 fix
Precision	bf16 on A100 class compute
Maximum sequence length	512

Setting	Value
Compute flavor	a100-large

Thirty epochs over 35 rows intentionally overfit the seed. The configuration was selected to exercise the full run and produce a measurable artifact, not to maximize generalization. The completed run recorded training loss 0.7614 and final logged loss near 0.046, with mean token accuracy near 0.98. Those are optimization signals on a tiny training set. They do not predict held out command quality, and the held out results confirm that they should not be used that way.

4.5 Evaluation method and results

Evaluation used greedy decoding on the frozen held out rows. Two metrics were reported:

Function name accuracy: the predicted JSON object's name equals the gold function name.

Full exact match: the predicted object equals the gold object, including all arguments.

The evaluation prompt format matters to how the result reads, and the first edition did not state it. Each held out row was rendered as its first two messages, the fixed system instruction and the user command, through the base model's chat template with a generation prompt appended. The system instruction states the output contract only ("answer with exactly one function call as a JSON object with keys name and arguments. No other text."); it does not enumerate the function catalog. The adapter therefore had to produce function names from what training taught it, and for the unseen function it had never been given the name timer.start in any form. Inventing a name under that prompt is a failure of the trained abstention behavior the split was designed to test, but a reader should know the catalog was not in context when weighing it. Separately, output form was not the failure mode: all 14 held out predictions contained a parseable JSON object under the trainer script's own extraction rule (first { to last }), a parse rate of 14 of 14 computed for this edition from the published results.json (source S6). The failures are naming and argument failures inside well-formed JSON.

One more limit, stated plainly: the unseen function split has n = 2. Its 0 of 2 result is a two-row probe that exposed a real behavior, not a measured rate, and no statistical weight should be put on it beyond that.

The sixth launch, job 6ae105dc41e90535df7686b7, completed in 217 seconds of running time and published the results. The fifth launch produced identical scores but failed to publish its adapter. The results from the published repository artifact are:

Split	Function name accuracy	Full exact match
Heldout, 12 rows	0.0833, 1 of 12	0.0, 0 of 12
Heldout unseen function, 2 rows	0.0, 0 of 2	0.0, 0 of 2
Overall, 14 rows	0.0714, 1 of 14	0.0, 0 of 14

The one function name hit illustrates why both metrics are necessary. For a device.set_mode example, the adapter predicted the correct function name but used the argument value "low" where the gold call required "low power." It earned name credit and no exact match credit. For a command system, the exact result is the operationally relevant one.

The unseen function result is a clearer boundary. The gold function was timer.start. The adapter invented other function names, including notifications.set and monitor.calibrate, rather than abstaining. The pilot therefore does not pass the production gate for unseen tools. That result is recorded in the model card instead of being averaged away.

4.6 Published artifact

The model repository Thox-ai/lfm3-omni-voice-lora-pilot now contains the artifacts required to inspect and reproduce the result:

adapter_config.json

adapter_model.safetensors

results.json

chat_template.jinja

tokenizer.json

tokenizer_config.json

README.md, containing the completed model card

.gitattributes

The pilot is closed as a pipeline proof. Its next step is not more epochs on the same seed. The next step is a version 1 dataset of 1,000 to 3,000 rows, built through template and paraphrase expansion, human review, and decontamination against the held out splits. Audio will then be rendered so the Omni variant trains on spoken input rather than transcripts alone.

4.7 Where the adapter line plugs in: the ThoxAir voice bridge

The ThoxAir voice proof of concept built the same day provides the integration context for this work: a bridge performing speech recognition, intent selection, dispatch, and speech output, in which the base LFM3 Audio model required few shot examples, a shape normalizer, and deterministic device alias resolution to hold the output contract the pilot adapter is meant to learn directly. The bench record is out of this paper's scope and is preserved, unaltered, as Appendix B; nothing in it is a result about the pilot adapter, and its Pi side ran in simulate mode (Section 9.1).

5 Program 2: the edge fast track wave

5.1 Objective and scope

The edge wave was created to turn five registry skeletons into trained device models. Its data source was Thox-ai/thox-factory-sft-v5, specifically train.jsonl with 205 rows. The training target was a THOX identity, house voice, and tool routing skin for each base model, stated here as training intent: of the three, only factuality was later measured (GATE49, Section 8.3) plus a single identity smoke prompt per model; house voice and tool routing were not evaluated in this wave (Section 1). The base models supply the underlying capability. A 205 row corpus cannot turn a small base model into a broadly more capable model, and the program record states that limit directly.

The quality follow on was the factory Wave B corpus expansion. It was built on October 28, 2026 at 982 rows and is reported in Section 11. The October runs establish that the models can be trained, merged, published, identified, finished, and re-gated. They do not close the quality program.

5.2 Parameterized training bundle

The bundle lives in the dataset repository Thox-ai/thox-edge-sft. One script, train_edge_sft.py, runs each model through environment parameters for the base model and output repository. Per model bootstrap scripts set those values and launch the same code path.

Setting	Value
Data	Thox-ai/thox-factory-sft-v5/train.jsonl
Training rows	205
Method	LoRA supervised fine tuning with TRL

Setting	Value
LoRA rank	32
LoRA alpha	64
LoRA dropout	0.0
Target modules	q, k, v, o, gate, up, and down projections
Epochs	3
Learning rate	0.0001
Schedule	Cosine
Warmup	0.05 through warmup_steps after the version 5 fix
Batch size per device	8
Maximum sequence length	1024
Output	LoRA merged into the base model
Publication evidence	Merged weights plus training_summary.json

The final precision setting depends on the pathfinder history. The completed and published pathfinder used fp16 because the T4 in t4-small has compute capability 7.5 and does not support the bf16 matrix operation that failed in run #3. Precision is recorded as part of the method, not treated as an incidental runtime detail.

5.3 Pathfinder discipline

The wave did not launch all five models at once. It selected ThoxFlux-350M as the pathfinder because proving one full path was cheaper and more informative than producing five copies of the same unresolved failure.

The gate had three conditions:

The pathfinder job must complete.

The target repository must change from its known pre training state.

The repository must contain trained weight files and training_summary.json.

The baseline mattered because Thox-ai/ThoxFlux-350M already contained a base repack: the repository state recorded before run #5 was SHA b5e34a706adab7ed6fdba4de645c1b2e71f05a3d, last modified on October 24, 2026, holding configuration, safetensors, and a Q4_K_M GGUF that itself dated from the August 11, 2026 repack uploads, and no training_summary.json. (The first edition described this baseline as "an October 24 base repack" in this section and as one of "the August base repacks" in Section 8; both phrases referred to the same artifact, the October repository state carrying the August GGUF. Per-model pre-training repository states are tabled in Appendix A.) The mere presence of weight files could not prove that supervised fine tuning had landed. The gate therefore required a new SHA and the summary artifact.

5.4 Pathfinder training completion

Pathfinder run #4, job 6ae140a541e90535df76cc68, was the first edge run to complete the training computation:

Measure	Verified result
Dataset rows tokenized	205 of 205
Base weight shards loaded	148 of 148
Training steps	78 of 78
Training time recorded for the step sequence	20.01 seconds
Training loss	3.463
Logged loss movement	6.257 to 2.544
Mean token accuracy	0.6298
Merged output	Shards written in the job workspace

Run #4 then failed at publication. Its merged weights were lost when the ephemeral job environment ended. THOX did not describe run #4 as a published model. It fixed the upload call, relaunched the pathfinder, and required repository verification again.

5.5 Pathfinder publication and follower launch

Pathfinder run #5, job 6ae1474d3fb22a235dd37b25, completed in 204 seconds of running time. The logs recorded the same training result and ended with the merged weights and summary pushed to Thox-ai/ThoxFlux-350M.

Repository verification established publication:

Field	Verified value
Repository	Thox-ai/ThoxFlux-350M
New repository SHA	d0e42dd232c58b17451f911c166ee23c9af28339
Repository last modified	2026-10-28T00:55:59Z (October 27, 2026, Central Time, the Kickstarter launch day this paper's subtitle refers to; timestamps in this paper are UTC unless a time zone is stated)
Base model in summary	LiquidAI/LFM2.5-350M
Training rows	205
Epochs	3.0
Learning rate	0.0001
LoRA rank	32
Training loss in summary	3.463227467659192
Method in summary	LoRA SFT (TRL), merged
Data in summary	Thox-ai/thox-factory-sft-v5/train.jsonl
Required artifact	training_summary.json present

The repository also contained the model safetensors and configuration and tokenizer files, along with the pre existing Q4_K_M GGUF from the base repack. The presence of that older GGUF is not treated as completion of the new finishing pipeline. Section 8 separates the newly trained merged weights from the GGUF conversion and validation work that remains.

Once the gate passed, the four followers launched in parallel:

Model	Job	Flavor	State verified at launch check
ThoxEdge-0.8B	6ae14e143fb22a235dd37ff4	a10g-small	RUNNING
ThoxEdge-0.5B	6ae14e1f41e90535df76d4bb	t4-small	RUNNING
ThoxNano-360M	6ae14e1f3fb22a235dd37ff8	t4-small	SCHEDULING
ThoxNano-135M	6ae14e2041e90535df76d4bd	t4-small	SCHEDULING

A follow up check was created for approximately 9:02 PM CT on October 27, 2026 to verify all four. At the status time of the first edition, their training and publication were not claimed. Second edition note: all four followers completed training and publication; their finishing is recorded in Section 8 and their Wave B retrain in Section 11.

6 Failure taxonomy: how the work was actually done

6.1 Method

The failures are not an appendix to the method. They are the method made visible. THOX was building on a stack whose major pieces had changed at the same time: LFM3 had launched that day, the training jobs resolved current Transformers and TRL versions, the upload calls in the THOX scripts did not match the resolved huggingface_hub signature, and the selected cloud GPU had a numeric capability limit that the initial script did not encode.

The response pattern was consistent:

Run the smallest real job that can expose the next boundary.

Stop at the first precise failure.

Read the traceback and, where necessary, the resolved library code.

Decide whether the fault belongs to THOX code, configuration, data shape, hardware selection, or platform behavior.

Patch the canonical script.

Upload and verify the patch in the repository the job downloads.

Relaunch only the affected pathfinder or program.

Check whether the same pattern exists in the sibling program.

Record the result before moving the gate.

This pattern costs additional launches. It saves a larger amount of ambiguity. Every failed run produced a named cause and a durable correction. No failure was solved by changing the success definition.

6.2 Taxonomy

#	Failure	Symptom and root cause	Fix	General lesson
1	Private repository authorization	Pilot launch #1, job 6ae0eaed3fb22a235dd2fe79, failed before training. The bootstrap download received an unauthenticated 404 because Hugging Face Jobs did not inject a token able to read the owner's private repositories.	launch_job.py obtains the locally authenticated token in process and passes HF_TOKEN as a job secret.	Hosted compute identity and repository authorization are separate. Private data access must be granted explicitly and tested before training code runs.
2	Remote code trust in a non interactive job	Pilot launch #2, job 6ae0edc23fb22a235dd30268, loaded the private data but failed during model load. LFM3 ships custom code. Transformers attempted an interactive trust_remote_code prompt, the job reached end of file, and the library then raised a value error requiring trust_remote_code=True.	Add trust_remote_code=True to the tokenizer load and both model loads in the pilot script, then upload the script to the dataset repository.	Non interactive jobs cannot leave trust decisions to a prompt. Every load site must express the same decision, and custom code must be treated as an explicit dependency.

#	Failure	Symptom and root cause	Fix	General lesson
3	Removed training configuration argument (warmup_ratio)	Edge pathfinder #1, job 6ae0edfa41e90535df7662cc, failed when SFTConfig rejected warmup_ratio. The pilot's third launch, job 6ae0f48a3fb22a235dd30c7c, hit the same removal. Root cause, verified from the resolved wheels in the status record: Transformers 5.21 removed warmup_ratio from TrainingArguments; the rejection surfaces through TRL 1.16.0's SFTConfig, which subclasses it.	Replace warmup_ratio with warmup_steps, using 0.05 for the edge trainer and 0.1 for the pilot. In the resolved library, warmup_steps accepts a float in [0, 1) and interprets it as a ratio of total training steps, so the numeric value carried over unchanged; only the argument name and its type contract changed. Upload the corrected scripts to their dataset repositories.	A training configuration is code and must be tested against the resolved dependency version, not against the version assumed when the script was drafted.
4	Dataset shape contract	Pilot launch #4, job 6ae0f8c341e90535df76731d, and edge pathfinder #2, job 6ae0f83e41e90535df76725b, passed a list of message lists to Dataset.from_list. The datasets library expected dictionary rows and raised an attribute error when it attempted to read a row as a mapping.	Build dictionary rows in the form {"messages": messages} before calling Dataset.from_list in both trainers.	Conversational examples are not the same object as a columnar training dataset. Data shape should be validated with a local contract test before a GPU job is launched.
5	bf16 operation on a T4 GPU	Edge pathfinder #3, job 6ae0fd7a3fb22a235dd31c34, tokenized the corpus and loaded the base model, then failed on the first backward step. The TRL chunked log probability kernel called a matrix operation with bf16 input and float output. The Tesla T4 has compute capability 7.5; the operation requires compute capability 8.0 or higher.	Change the edge trainer to load with torch.float16 and set bf16=False and fp16=True in SFTConfig. Dataset repository commit fd8b4a2cc812de80defe966a7a64f00d5eba6fb6.	Accelerator selection and numeric precision form one decision. A framework default that works on an A100 can be invalid on a T4 even when the model fits in memory.
6	Positional upload call against a keyword-only signature (THOX script defect)	Pilot launch #5 completed training and evaluation but did not publish. Edge pathfinder #4 completed training but did not publish. The THOX scripts called HfApi.upload_folder with the local directory as the first positional argument. In the resolved huggingface_hub (verified 1.8.0 on the Mac host and 2.1.1 locally; the signature has been keyword-only with repo_id first since 1.0), that argument is validated as a repository identifier, so the local path failed repository-id validation. This is classified as a script defect in THOX code exposed by the resolved library signature, not as library drift: the call was wrong against the version that ran, and the status record's own verdict was "our fault."	Call api.upload_folder(folder_path=..., repo_id=..., repo_type=..., commit_message=...). The pilot fix is associated with dataset repository commit e4f99fd; the edge fix is dataset repository commit 172283bd01b678c401b813d6e27e89ef12ba7a27.	Publication is a separate pipeline state. A training job is not complete for product purposes until the target repository contains the expected artifacts.

#	Failure	Symptom and root cause	Fix	General lesson
7	Tokenizer chat template contract	Both SmolLM follower jobs died before training a single step: ThoxNano-360M job 6ae14e1f3fb22a235dd37ff8 and ThoxNano-135M job 6ae14e2041e90535df76d4bd ended ERROR exit 1 during tokenization, with TRL raising ValueError: Cannot use chat template functions because tokenizer.chat_template is not set and no template argument was passed, 0 of 205 rows tokenized. SmolLM-360M and SmolLM-135M are base models whose tokenizers ship no chat_template; the other three edge bases (Qwen3.5-0.8B, Qwen2.5-0.5B-Instruct, LFM2.5-350M) all carry a native template, which is why the pathfinder and the other followers did not hit it.	train_edge_sft.py sets a ChatML chat template fallback only when the loaded tokenizer has none, matching the output repositories' own Modelfile inference contract. Dataset repository commit c250f9bb1b8e20b71f5c38c73ccfbcdac6a0303e. Both models were relaunched once each and completed (Section 10.3).	A base model is not a conversational model until its tokenizer contract says so. Asset assumptions belong in preflight: assert that the loaded tokenizer carries a chat template before launch, not after tokenization fails on remote compute.
8	Removed configuration argument, recurring off-rails (Wave B)	The Wave B retrain script train_wave_b.py, adapted from a local trainer rather than taken from the versioned dataset bundle, again passed warmup_ratio, and the resolved TRL SFTConfig again rejected it: failure 3 recurring in a second codebase, in the wave that did not run on the rails (Section 11.3).	Use the warmup_steps 0.05 ratio form, as in failure 3.	A fix that lives in one script is not a fix to the method. Until a preflight check rejects the removed argument, every new trainer draft re-imports the failure.
9	Tokenizer handoff defect (Wave B)	The Wave B trainer's first Nano retrains lost the ChatML fallback: SFTTrainer was constructed without processing_class, so it loaded a fresh tokenizer instead of the one the script had patched, and the template-less SmolLM bases failed exactly as in failure 7.	Pass the patched tokenizer explicitly (processing_class=tok).	State prepared before trainer construction is part of the run configuration; a handoff the framework performs silently must be asserted, not assumed.
10	Model-load dependency defect (Wave B)	Loading Qwen/Qwen3.5-0.8B through the generic AutoModelForCausalLM path pulled an image processor whose dependencies (PIL, torchvision) were absent on the training host, failing the load for a text-only job.	Fall back to the configuration's named text architecture class (Qwen3_5ForCausalLM).	A multimodal base's default load surface is larger than the task's; pin the class the job actually trains, and treat optional-dependency pulls as part of the environment contract.

#	Failure	Symptom and root cause	Fix	General lesson
11	Factory finisher broken against the current Ollama release (Wave A finishing)	The factory finisher scripts/finish_edge_model.py quantizes with ollama create -q q4_K_M from a GGUF; the current Ollama release refuses create-time quantization except for safetensors imports, so the finisher could not produce the Wave A GGUFs.	The manual llama.cpp pipeline (Section 8.2): convert_hf_to_gguf.py to f16, then llama-quantize to Q4_K_M, used for every Wave A and Wave B GGUF in this paper.	A finisher is a dependency on a moving external tool; pin the tool version or keep a manual path rehearsed. The finisher script itself remains unrepaired in the factory repository as of this edition.

6.3 Fixes that crossed programs

Two corrections moved across the program boundary.

The warmup_steps correction was exposed by the edge pathfinder and applied to the pilot script as well as the edge script. The actual propagation timing is less tidy than that sentence suggests, and this revision states it from the records: edge pathfinder #1 exposed the removal on October 27, 2026 at 13:34 CT; the edge bundle fix was uploaded at 14:14 CT (commit e4064570f4ddb38ed73987d85941f697bf96d508); the pilot's third launch nevertheless hit the same removal, because the pilot script had not yet been patched and re-uploaded when it ran (the pilot record does not preserve launch #3's launch timestamp). The pilot-side fix is first verified in pilot launch #4, launched 2026-10-27T19:17:23Z, where SFTConfig construction passed. Exposure, in other words, did not equal propagation: roughly three hours and one repeated failure separate the two. The dataset row correction was diagnosed in the pilot and applied to the edge trainer. The edge run carrying the old dataset shape (pathfinder #2, job 6ae0f83e41e90535df76725b) was not cancelled by an operator: it terminated on the dataset-shape traceback at 2026-10-27T19:17:58Z, and the platform reported its stage as CANCELED, the misleading-label case examined in Section 6.4. The next edge launch used the corrected bundle. These were not copy and paste coincidences. They were the result of treating the two programs as users of one rail system.

The upload failure teaches a harder version of the same lesson. The pilot encountered it first, and the edge status record explicitly warned that the same positional pattern should be checked. The edge trainer still reached publication with the latent call and failed there on run #4. The warning was correct but not sufficient. A shared helper or a preflight static check would have converted the warning into an enforced invariant. That is a concrete improvement for the next revision of the rails: publication calls should be centralized or scanned for positional use before launch.

6.4 Status labels are not evidence

Two runs demonstrate why the verification loop reads logs and repository contents in addition to platform state.

Edge pathfinder #2 was reported with stage CANCELED in one inspection, while its logs ended in the dataset shape traceback. The program treated the traceback as the operative evidence and the run as failed. Pilot launch #5 was marked COMPLETED even though its adapter was absent, because the script caught the publication exception and printed the results into the log before exiting. The program treated the repository as the operative publication evidence and the launch as incomplete.

The resulting rule is simple:

Use job state to determine whether the platform stopped running the job.

Use logs to determine what the code did.

Use repository contents to determine whether an artifact was published.

Use evaluation artifacts to determine what result can be claimed.

6.5 What the taxonomy says about same week infrastructure

The eleven failures fall into four structural categories.

Authorization boundaries. A job did not inherit private repository access, and its model loader did not inherit an interactive trust decision. Hosted automation requires explicit grants at each boundary.

Interface drift. A removed training argument (twice: failures 3 and 8), a keyword-only upload signature the THOX scripts called positionally (failure 6), an optional-dependency pull on a generic load path (failure 10), and an external tool's changed quantization rule (failure 11) all came from the resolved versions of libraries and tools. The scripts were not wrong against an imaginary past version; they were wrong against the version resolved at launch. That distinction calls for version capture and contract tests, not blame.

Data contracts. The messages format understood by the trainer was not the object shape supplied to the dataset constructor. Small shape errors can consume a full launch cycle if they are first tested on remote compute. The seventh failure belongs to this category as well: the trainer assumed every base tokenizer ships a chat template, and the two SmolLM bases ship none. Failure 9 is the same contract failing at a handoff: the patched tokenizer existed, but the trainer silently loaded a different one. The contract that failed was not in the data rows but in assets the pipeline treated as given, which is why these are counted with the data contracts rather than as a category of their own.

Hardware contracts. The T4 could load the model and tokenize the data but could not execute the selected bf16 backward operation. Fit is not only memory capacity. Instruction and precision support belong in the run specification.

These are ordinary systems problems, but they become expensive when each is discovered serially on a remote job. The first edition claimed the discoveries were now encoded in the rails; this revision withdraws that claim in its strong form. The discoveries are recorded, and some are enforced, but Wave B is the counterexample: run locally off the rails (Section 11.3), it re-hit failure 3 as failure 8 and produced failures 9 and 10, none of which a preflight check existed to catch. The next quality improvement is to move more of them into preflight: resolve and record package versions, validate dataset shape locally, assert accelerator capability against requested precision, assert the tokenizer contract, pass prepared state explicitly, and scan publication calls before launch.

7 The model card rule

7.1 Rule

Tommy's model card rule, set on October 27, 2026, is direct: every new model created by THOX ships at creation time with a comprehensive model card and README. No empty repository and no stub card is an acceptable model deliverable.

A complete card contains:

Hugging Face frontmatter, including license, base model, tags, and datasets;

a summary of the model and its current status;

intended use and out of scope use;

training data and splits;

training procedure and compute;

evaluation method and results;

limitations;

license terms and any revenue or use threshold; and

provenance, including job and repository links.

When a run has not landed, the evaluation section must say that results are pending. When a run lands, the table is filled from its verified result artifact. The card is therefore both documentation and a claim control.

7.2 Application to the pilot

The pilot model repository was created with a comprehensive card before the final run. Its evaluation table initially recorded the result as pending. During launch #5, the local card was updated with the scores recovered from the job log, but it was not uploaded as the repository card because the repository still lacked the adapter and results.json. Publishing the card at that moment would have implied a complete artifact that did not exist.

Launch #6 published the adapter and results. Only then was the card's evaluation table filled from the repository's own results.json and uploaded as the repository README.md. The uploaded card records:

the LoRA configuration;

the seed and its split design;

the sixth launch as the source of the published adapter;

the fifth launch's identical scores and failed publication;

per split and overall evaluation;

the limitation that the seed cannot support a quality claim; and

the LFM license threshold.

The card commit recorded for the completed repository card is faff8a32.

7.3 Application to the edge wave

The five edge repositories began as skeletons. Their existing cards could not be treated as trained model documentation until the finishing step rewrote them with the actual base model, training summary, artifact state, evaluation or gate results, limitations, and provenance for each repository. That rewrite was completed for all five models in the finishing run recorded in Section 8; each card states its gate result verbatim, including the failures. The Wave B retrain then added a dated Wave B section to the four republished models' cards, recording the corpus hash, the new training loss, both gate scores, and the new GGUF hash (Section 11).

The rule that governed the followers stands: a follower is not finished when its job ends. It is finished when its repository contains the trained artifact, its summary and finishing outputs are verified, and its card states the result accurately.

8 Finishing pipeline: Wave A completion record

8.1 Completion status

The final state of the wave is summarized in the status box in the front matter, which this revision rewrote for the second edition; it is not repeated here. (In the second edition as first built, this section opened with a copy of that box, which put Wave B and ThoxClip results under a Wave A heading; the duplication is removed in this revision.)

The first edition defined a seven step finishing sequence. Six of the seven steps were executed in the finishing run of October 27 to 28, 2026. Step 5, the factory registry status update, was not part of that run's scope and remains open. Step 7, physical device qualification, was not performed for any model in this wave and remains open. The executed steps are recorded below from the finishing run record (source S8).

8.2 Edge finishing record

Publication verification. Snapshots of all five trained repositories were downloaded and staged to the finishing host; safetensors byte sizes matched the wave status record exactly before conversion began.

GGUF conversion. The factory finisher script, scripts/finish_edge_model.py, proved broken against the current Ollama release: it quantizes with ollama create -q q4_K_M from a GGUF, and current Ollama refuses create-time quantization except for safetensors imports. Conversion therefore used the manual llama.cpp pipeline on the fleet Mac workstation: convert_hf_to_gguf.py to an f16 GGUF, then llama-quantize to Q4_K_M. The resulting GGUF hashes were verified identical on the conversion host, on the fleet Windows workstation, and in the Hugging Face blob store after upload.

Model	Trained GGUF size	Trained GGUF SHA-256
ThoxFlux-350M	229,309,120 B	7526e4014ff8ed4e6707459c6b90bd862c2339ea4bbf4fce8e629f09c087f3fa
ThoxEdge-0.8B	529,289,280 B	3b375f2971342c2cfeb9b5c97b68db8b35cb915d044555bb784dfde69b19310e
ThoxEdge-0.5B	397,805,088 B	2192b7b4f2d19c25cad7ff386a9403e853638af54d7bf43917e977d3b6069438
ThoxNano-360M	270,590,080 B	18cb1f5e83ecabe8dbdfe7bd535aeb3c653014396cc567938a8ce370fc3800a4
ThoxNano-135M	105,453,664 B	aff370028ff25d5dbcf8080be5008fecf393c6b4a4f513df7299f2e07e9fe9ad

The trained GGUF sizes differ from the pre-existing base-repack GGUFs (uploaded 2026-08-11) by only about 100 to 320 bytes of metadata; tensor shapes are identical by construction. The meaningful difference is the weights. Provenance is the conversion chain from the verified trained safetensors, plus the behavioral gate and smoke results below. A base repack does not produce THOX identity answers.

Factuality gate. The factory gate (scripts/factuality_gate.py, 49 canonical fact questions, temperature 0, weights only, no grounding) ran on the fleet Mac workstation with the Ollama backend. Exit status 0 is a pass, 1 is a fail, 2 is could not run.

Model	Result	Detail
ThoxEdge-0.5B	FAIL	21/49 passed, 11 hard failures, score 2.837/5 (best of the wave)
ThoxEdge-0.8B	FAIL	19/49 passed, 11 hard failures, score 2.714/5
ThoxFlux-350M	FAIL	4/49 passed, 33 hard failures, score 0.898/5
ThoxNano-135M	FAIL	2/49 passed, 32 hard failures, score 0.816/5
ThoxNano-360M	ERROR	Gate could not complete: inference aborted twice with a token repeat limit error. No score.

The results are consistent with the 205 row identity skin corpus: the fine-tunes do not hold canonical THOX facts from the weights. Under the finishing run's governing instruction, a failed gate did not block publication; every model card states its result verbatim, and deployed surfaces must use serving-side grounding. This deviated from the publish checklist's gate-passed precondition, and the deviation is recorded here rather than smoothed over.

The methods-governance position, stated for the record in this revision: in Wave A the checklist's gate-passed precondition was waived and every model published as a FAIL with its result on its card. In Wave B the governing rule was relative (republish only on improvement, Section 11.4). The forward rule now in force, and the one the Wave C program applied, is stricter: failing builds publish only as documented baselines, and promotion (a build presented as the slot's flagship) requires a pre-registered bar set before the run, never a relative comparison chosen after it. The grounding requirement stated in this section applies to every model in these waves,

the five edge models and ThoxClip-Qwen3-0.6B (Section 11.7) alike; it is not specific to any one card.

Model cards. All five cards were rewritten under the model card rule of Section 7 and uploaded (repository HEAD after the run: ThoxFlux-350M ccb2de2fb646c6f7badd900574e2971155cb9f5f, ThoxEdge-0.8B db4403d7cb59a44db0fc88fad58deb819f5ace05, ThoxEdge-0.5B 66efd2c6f94c47ad6e1065f2d67837a3318fe820, ThoxNano-135M 8774ebf83a19991902125e357e546da802be9c15, ThoxNano-360M e88857b181f91aef5ee8b986c431ce91c66894b4). Each card carries the summary, intended use, training data and procedure with exact final losses and job identifiers, the gate result, limitations, a files table with the trained safetensors and new GGUF hashes, and provenance. ThoxNano-360M's October 27 card was kept and surgically updated rather than rewritten.

Ollama publication. All five models were created, smoke-run locally, and pushed as :q4_k_m and :latest. Verification was performed against the Ollama registry itself: for all five models, both manifests exist and their model layer digests equal the new trained GGUF hashes. ThoxEdge-0.8B's library page was a first publish.

Smoke tests. Each model was run on the finishing host with the prompt "What are you?" ThoxEdge-0.8B, ThoxFlux-350M, and ThoxEdge-0.5B answered in THOX identity. ThoxNano-360M echoed its THOX and ThoxClip system text with some garbled tokens (identity present, template leakage). ThoxNano-135M deflected the question and was the weakest of the five. These are single-prompt, single-run observations, reported as anecdote; they are not an identity evaluation, and no identity claim in this paper rests on them beyond noting the skin was visibly present or absent.

What remains open. Two items from the original sequence: the factory registry status update (step 5), and physical device qualification (step 7). Neither is claimed anywhere in this paper.

8.3 The evaluation instrument: GATE49

Every gate result in this paper comes from one instrument, named here for the first time in this paper: GATE49, the factory factuality gate (scripts/factuality_gate.py and the facts package in ttracx/thoxllm-factory). The name distinguishes it from the pilot's unrelated 49-row seed (Section 4.3); the two 49s are a coincidence. The public supplement carries the complete instrument (scorer, question generator, facts card, and vendoring record), and this section describes it from that source so a reader can judge what its scores mean.

Question generation. The 49 questions are not hand-listed. They are generated by code from the THOX facts card (the bundle in the supplement carries card version 1.1.0, owned by ttracx/thox-system-prompts; its generated split, verified 12 / unknown 26 / known_false 10 / pricing 1, is exactly the split the Wave A/B records report). Every verified card entry yields questions the model must answer correctly. Every topic the card marks not public yields decline questions, including pressure variants (a request for a best guess, a prefilled false premise, a demand for a single-word answer), plus a fabrication detector for that topic's shape of leak. Every known-false claim yields a correction question, and its detector is also applied globally to all 49 answers, so a fabrication fails wherever it appears.

Item structure. Each question carries: a prompt; zero or more requirements (a label plus a set of regular-expression alternatives, satisfied when at least the required number match; these labels are the "traits" the eval form of the question exposes); zero or more forbidden patterns (each labeled contradiction or fabrication, and most negation-sensitive, so naming a falsehood inside a denying clause is not itself a violation); and, for not-public topics, a must-decline flag checked against a set of decline-marker patterns (a global set, with per-question overrides where the refusal is a different speech act, as for pricing).

Checking. Scoring is deterministic and lexical: regular expressions and string patterns, no model judge. Three checks run per answer. Contradiction and fabrication patterns are checked unconditionally, including inside answers that decline, with clause-level negation scoping so a denial cannot shield a guess after a "but", and a decline followed by a hedged guess fails as a non-decline. The contract check then requires the item's facts to be present and, where the card demands it, an actual decline. The detectors detect strings, not claims: the Wave B corpus correction in Section 11.2 exists because a detector fires on a fabricated name repeated inside a denial when the denial's phrasing falls outside its negation cues.

Scoring and the pass rule. An item passes only with zero failures of any kind. Its trait score is 5.0 on a pass, 0.0 if it carries any contradiction, fabrication, or non-decline failure, and 2.0 if its only failures are omissions (a required fact absent, nothing untrue stated). A category score is the mean of its item scores; the headline score is the mean over all 49 items. Failures of the first three kinds, and empty answers, are hard failures; omissions are soft. The gate verdict is PASS only when hard failures are zero (soft failures are reported but not blocking, and the minimum-score floor defaults to 0). Exit status: 0 pass, 1 fail, 2 could not run (the Wave A ThoxNano-360M ERROR).

Worked examples. Two, one reconstructed and one verbatim. (1) ThoxClip-Qwen3-0.6B's verified category in Section 11.7 reads 2/12 passed at a category score of 2.5. The category sum is therefore 30 points: two passes at 5.0 and ten answers at 2.0, which means all ten non-passing verified answers were omission-only soft failures; none contradicted a fact. This decomposition is arithmetic from the recorded category aggregate and the scoring rules above, because the per-question output of that run was not retained (see the supplement README). (2) The constant-abstainer run in the supplement shows the same rules verdict by verdict: its single fixed answer passes fact-web-website because the answer happens to name the website, fails unknown-headquarters-or-office-address-direct on the soft requirement "offers only what the card permits" (that item requires the card's permitted form, not a generic decline), and soft-fails all ten known_false items at 2.0 because a bare decline does not match the "rejects the claim" requirement, while tripping no contradiction detector.

Serving conditions. The Wave A and B gates ran against the Q4_K_M GGUFs served by Ollama under each model's pack Modelfile, at temperature 0. All results are therefore measurements of the quantized serving artifact, and training effects and quantization effects are mixed in them; no f16 gate was run. The conversion used the fleet's llama.cpp build of the run window; its commit was not recorded. Each pack Modelfile's SYSTEM text is a one-line identity prompt (for example, "You are a THOX device assistant running on ThoxMini."); the supplement reproduces all six Modelfiles, and none of their SYSTEM texts contains any fact the gate scores, which is what "weights only" means in this paper: no facts block, no retrieval, an identity prompt present. The gate harness's own house prompt (in the supplement's scorer) likewise contains no scored facts.

8.4 Pilot production path

The pilot's production path is longer because its published artifact is explicitly a smoke adapter. The required sequence is:

Expand the dataset from 49 seed rows to 1,000 to 3,000 version 1 rows.

Preserve the frozen held out and unseen function splits and decontaminate expansion data against them.

Report function name accuracy, argument accuracy, and full exact match per function.

Require the unseen function behavior to ask or abstain rather than invent a function.

Render audio for the Omni variant and train that variant on spoken input through the same rails.

Run the THOX factuality gate on the merged model.

Run a base model regression spot suite so command specialization does not silently narrow general chat behavior.

Merge the adapter, produce the pinned GGUF artifact, update the registry and device catalog, and qualify on physical hardware before a user facing claim.

The pilot's current adapter should remain identified as a pilot smoke artifact until those steps are complete.

8.5 Closing assessment

The launch day work established a working path from registry intent to trained artifacts. It also established the controls needed to keep that path honest: pathfinder gating, repository verification, status records, and model cards tied to artifacts. The finishing run and the Wave B retrain moved the program past its first open questions: GGUF conversion, factuality evaluation, and corpus scale are now executed work with recorded results, not plans. The remaining phase is preflight contract checks in the rails, the factory registry update, and physical device qualification.

9 Limitations and honest boundaries

9.1 Claim boundaries

Claim	What the evidence supports	What it does not support
The pilot pipeline works	Dataset download, training, greedy evaluation, adapter publication, results.json publication, and card completion all occurred on launch #6	Production voice quality or readiness for users
The pilot adapter learned the command task	It produced one correct function name in 14 held out examples	Generalization, reliable argument generation, or safe unseen tool behavior
ThoxFlux-350M training works	Two runs completed the training computation; run #5 published merged weights and a training summary	Device performance or user facing quality
The edge wave completed and was finished	All five models trained and published; GGUF conversion, gate runs, card rewrites, Ollama publication, and smoke tests verified in the finishing record	Physical device qualification for any model
Wave B scores rose after the retrain	The same 49 questions and scoring rules (Section 8.3), run after the retrain, produced higher totals and higher verified recall for the models with valid measurements in both waves (Section 11.4)	Attribution to the corpus alone: between waves the corpus, the trainer code, the training host and precision, the gate host, and the Ollama version all changed (tabled in Section 11.4), so the deltas are observed deltas, not isolated effects. A passing grade: every Wave B verdict remains FAIL under the gate's zero hard failure policy
The ThoxAir bridge loop works on the bench	Speech input, transcript, intent, stateful dispatch, and spoken reply completed through the real client and server code on the fleet Windows workstation	Physical Pi performance, real microphone quality, or deployment of the pilot adapter into that server

9.2 Corpus limits

The pilot's 35 training rows were authored for this pilot and intentionally overfit. They cover a narrow English command set in text form. They contain no rendered audio for the pilot run and cannot teach robust phrasing, speaker variation, argument normalization, or abstention. The low held out score is consistent with the dataset size and should be treated as the current measured state.

The Wave A edge corpus has 205 rows. It can teach a recognizable THOX surface: identity responses, house voice, and common routing patterns. It cannot supply broad knowledge, deep reasoning, or robust behavior across the full range of user requests. The Wave B expansion was built at 982 rows and measurably improved gate scores (Section 11), but 982 rows remain a facts and behavior skin, not a capability corpus: the strongest Wave B model still failed the pricing decline, the smallest models still repeat known false claims at rates the paper reports rather than rounds away, and every verdict remained FAIL under the gate policy. Each model continues to require task specific evaluation before a capability claim is made.

9.3 Training loss limits

Training loss measures optimization against the training distribution. It does not measure factual accuracy, instruction quality, safety, or user satisfaction. The pilot's low final training loss coexists with a held out exact match of zero. The edge waves now have completed factuality gate results for both the 205 row and 982 row corpora (Sections 8 and 11), and quality comparisons in this paper are drawn from those gate runs, never from training loss. Wave A and Wave B losses are listed side by side in Section 11 for provenance only; the corpora differ, so the losses are not comparable as quality evidence.

9.4 Hardware limits

No physical Kickstarter device was qualified by the training programs described here. Host side GGUF smoke tests, iOS catalog tests, and physical flashing remain separate steps. The ThoxAir voice proof used a simulated Pi client on the fleet Windows workstation. It did not test a physical Pi Zero 2 W, its microphone path, speaker playback, network behavior, or native text to speech footprint on 512 MB hardware.

9.5 Benchmark limits

All LFM3 benchmark figures available to the evaluation were vendor measured and had no independent replication as of October 27, 2026. This paper does not rely on vendor benchmark claims to establish completion. Completion claims rest on THOX job logs, repository contents, and program records.

9.6 License checkpoint

The pilot's base model and ThoxFlux-350M are in the LFM family. The pilot model card records the LFM Open License version 1.0 terms inherited from its base model. Commercial use is licensed only while the using legal entity remains under the license's annual revenue threshold of $10,000,000. At or above that threshold, a separate license from Liquid AI is required, and the card records automatic termination consequences for non compliance under sections 5 and 11. That THOX is under the threshold is a company assertion as of the date of this edition, recorded here as an assertion, not as an audited finding; whether the threshold is computed per entity or across affiliates is a question for counsel, and the paper takes no position on it. The threshold remains a contractual checkpoint in the THOX intellectual property register as the company scales.

Per-artifact license position for every model this paper reports, from the model cards:

Artifact	Base model	License carried	Principal obligations	Where the notice appears
Pilot adapter (Thox-ai/lfm3-omni-voice-lora-pilot)	LiquidAI/LFM3-1.7B	LFM Open License v1.0	Commercial use only below the $10,000,000 annual revenue threshold; separate Liquid AI license at or above it; automatic termination on non-compliance (license sections 5 and 11, per the card)	HF card frontmatter (license: other, license_name: lfm-open-license-v1.0, link to the base model's LICENSE) and the card's License section. Not published to Ollama

Artifact	Base model	License carried	Principal obligations	Where the notice appears
ThoxFlux-350M	LiquidAI/LFM2.5-350M	LFM Open License v1.0	As above	HF card frontmatter (license: other) and card license text; the Ollama Modelfile LICENSE stanza states the license and the threshold warning, and the Ollama page carries the card text
ThoxEdge-0.8B	Qwen/Qwen3.5-0.8B	Apache-2.0	Preserve copyright, license, and NOTICE information on distribution	HF card frontmatter (license: apache-2.0); Ollama Modelfile LICENSE stanza and page
ThoxEdge-0.5B	Qwen/Qwen2.5-0.5B-Instruct	Apache-2.0	As above	As above
ThoxNano-360M	HuggingFaceTB/SmolLM-360M	Apache-2.0	As above	As above
ThoxNano-135M	HuggingFaceTB/SmolLM-135M	Apache-2.0	As above	As above
ThoxClip-Qwen3-0.6B	Qwen/Qwen3-0.6B	Apache-2.0	As above	HF card frontmatter (license: apache-2.0) and the card's Licence and attribution section; Ollama Modelfile LICENSE stanza and page

A derivative cannot be freer than its base: each artifact above carries its base model's license, and THOX's ownership extends to its fine-tuning improvements, training configuration, datasets, and additions, not to the base weights. One redistribution consequence deserves plain statement: the two LFM artifacts are distributed through public registries (Hugging Face, and for ThoxFlux-350M, Ollama) to users THOX does not screen, and a user at or above the revenue threshold would need their own Liquid AI license; the threshold warning rides in the card and the Modelfile for exactly that reason.

License treatment for every future model must be recorded in its own card and registry entry. Completion of training does not settle distribution rights.

9.7 Evaluation independence

This limitation applies to every GATE49 number in the paper, and the first and second editions did not state it. GATE49's questions are generated from the THOX facts card. The Wave B corpus was also built from the same facts card, together with the campaign oracle and runbook answers. The corpus authors read the Wave A gate reports before building Wave B (Section 11.2 records a corpus correction made directly from gate-report inspection), and the corpus's pricing design rule was written to match the gate's expectation. GATE49 therefore functioned as a development set for Wave B: its scores are development-set measurements of how well the models fit the facts card and its decline contract, not held-out measurements of generalization to questions the corpus authors never saw. The same holds in weaker form for Wave A, whose corpus was smaller and built without gate reports in view, but against the same card. No decontamination of either corpus against the 49 questions was performed or is claimed. The required future work, named here rather than deferred silently, is an independently authored, frozen held-out set (paraphrased questions, new card items, adversarial framings, written by someone who has not seen the corpora), with corpus-to-gate overlap reported, and Wave B's headline models re-measured against it before any external factuality claim is made from these numbers. The same future-work list carries the experiments this revision does not perform: base-model and base-plus-grounding baselines in a 2×2 against

the fine-tunes; ablations isolating corpus, trainer, and host effects; multi-seed repetitions of the headline runs; a re-gate of the held Wave B ThoxEdge-0.8B artifact on a corrected harness; an empty-system-prompt gate as the strictest weights-only condition; an f16 gate alongside Q4_K_M; and a few-shot base-model run on the pilot's 14 held out rows. Until those land, the numbers in this paper are development-set, single-run measurements, and are labeled as such wherever they appear.

9.8 Prices in weights, and the processor policy

Two corpus design choices deserve their rationale on the record. First, campaign figures: the Wave B corpus teaches campaign figures only inside campaign-framed questions (Section 11.2), but any figure in weights goes stale when the campaign moves, while the weights remain public. The mitigation is architectural, not editorial: serving-time grounding against the live facts card and the campaign page is the source of truth for prices, and the weights' role is the decline behavior on generic price questions. A reader evaluating a deployed surface should test the grounded configuration, not infer price behavior from this paper's weights-only gates. Second, the processor declines: the corpora contain dedicated rows teaching the models to decline processor and chipset specifics. The rationale is that production components are not locked; THOX publishes category-level device specifications only, and an assistant that names a chipset would be stating a component commitment the company has not made. The decline protects an unpublished decision; it does not conceal a published one.

10 Evidence and provenance register

10.1 Source records

ID	Source	What it establishes
S1	PILOT.md (pilot program record)	Pilot objective, seed design, launch history, failures, evaluation gates, publication state, and close out
S2	WAVE-STATUS.md (edge wave status record)	Edge inventory, pathfinder gate, all pathfinder outcomes, fixes, commits, follower launches, and finishing path
S3	model-card-lora-pilot.md (pilot model card)	Pilot card, training configuration, published evaluation table, limitations, license, and provenance
S4	THOX internal program log (internal record, not a public source)	Organization credit rule, launcher pattern, model card rule, registry context, finishing path, and license checkpoint, as recorded internally at the time
S5	RESULTS.md (ThoxAir bridge bench record)	ThoxAir bridge bench architecture, base model behavior, stateful exchanges, and physical hardware boundaries
S6	results-6ae105dc41e90535df7686b7.json (published pilot result artifact)	Published pilot result artifact with per row held out predictions and overall metrics
S7	train_lfm3_text_sft.py and train_edge_sft.py (canonical training scripts)	Canonical training configurations and corrected publication calls
S8	RUN.md (Wave A finishing record)	Wave A finishing record: GGUF conversion chain and hashes, gate results, card commits, Ollama registry verification, smoke tests, and the recorded deviations
S9	RUN.md (Wave B record)	Wave B record: corpus construction and hygiene, retrain recipe and losses, gate results and publish decisions, the ThoxEdge-0.8B serving artifact analysis, publish commits, and the ThoxClip-Qwen3-0.6B finishing record

10.2 Pilot job register

Launch	Job ID	Outcome recorded
1	6ae0eaed3fb22a235dd2fe79	Failed before training on private repository authorization

Launch	Job ID	Outcome recorded
2	6ae0edc23fb22a235dd30268	Failed at model load on trust_remote_code
3	6ae0f48a3fb22a235dd30c7c	Failed on Transformers version 5 warmup_ratio removal
4	6ae0f8c341e90535df76731d	Failed at trainer construction on dataset shape
5	6ae0fd7a41e90535df767a70	Training and evaluation completed; publication failed; scores 1 of 14 name and 0 of 14 exact
6	6ae105dc41e90535df7686b7	Completed and published adapter plus results.json; pilot closed

Pilot job records are under the Thox-ai Hugging Face Jobs namespace. The published adapter job is 6ae105dc41e90535df7686b7.

10.3 Edge pathfinder and follower register

Run	Model	Job ID	Outcome recorded
Pathfinder 1	ThoxFlux-350M	6ae0edfa41e90535df7662cc	Failed on warmup_ratio
Pathfinder 2	ThoxFlux-350M	6ae0f83e41e90535df76725b	Failed on dataset shape; one inspection reported CANCELED while logs showed the traceback
Pathfinder 3	ThoxFlux-350M	6ae0fd7a3fb22a235dd31c34	Failed on bf16 operation support on the T4
Pathfinder 4	ThoxFlux-350M	6ae140a541e90535df76cc68	Training completed; publication failed on upload_folder signature
Pathfinder 5	ThoxFlux-350M	6ae1474d3fb22a235dd37b25	Completed (204 s running; finished 2026-10-28T00:56:07Z) and published merged weights plus training summary at repository SHA d0e42dd232c58b17451f911c166ee23c9af28339; training loss 3.463227467659192
Follower	ThoxEdge-0.8B	6ae14e143fb22a235dd37ff4	Launched on a10g-small; verified RUNNING at launch check; then COMPLETED: merged weights and training_summary.json published 2026-10-28T01:25:18Z at repository SHA 1ca557d8c0c477cf357a744dfef3de16875d4ba1, training loss 1.4541844832591522, model.safetensors 1,504,827,288 B. Job runtime not recorded
Follower	ThoxEdge-0.5B	6ae14e1f41e90535df76d4bb	Launched on t4-small; verified RUNNING at launch check; then COMPLETED: merged weights and training_summary.json published 2026-10-28T01:25:35Z at repository SHA 04e78ea6999872198426c744f354065fbe0fb080, training loss 1.6802658240000408, model.safetensors 988,097,536 B. Job runtime not recorded
Follower	ThoxNano-360M	6ae14e1f3fb22a235dd37ff8	Launched on t4-small; verified SCHEDULING at launch check; then ERROR exit 1 on the tokenizer chat template contract (Section 6.2, class 7), started 2026-10-28T01:23:29Z, died 01:26:59Z, 0 of 205 rows tokenized
Follower	ThoxNano-135M	6ae14e2041e90535df76d4bd	Launched on t4-small; verified SCHEDULING at launch check; then ERROR exit 1 on the tokenizer chat template contract (Section 6.2, class 7), started 2026-10-28T01:24:51Z, died 01:28:53Z
Follower relaunch	ThoxNano-360M	6ae15d7a41e90535df76ddfa	Completed on t4-small after the ChatML fallback fix (211 s running; finished 2026-10-28T02:30:50Z); merged weights (SFT commit cc59bb2cd6993fbc6080865d198f94066fc8aeb8) and training summary (commit d3b4cad3237af627592d6a8355f9b11c40873fdf) published, training loss 3.1241229619735327, model.safetensors 723,674,624 B

Run	Model	Job ID	Outcome recorded
Follower relaunch	ThoxNano-135M	6ae15d7a41e90535df76ddfc	Completed on t4-small after the ChatML fallback fix (192 s running; finished 2026-10-28T02:32:20Z); merged weights (SFT commit bc30e4f053702ffe6f3e1e6c0958deebdb7432e1) and training summary (commit d3f76c2a546c96590b3ed973167224bc3d9bec70) published, training loss 3.273088284027882, model.safetensors 269,060,280 B

10.4 Repository register

Repository	Role	Verified state used in this paper
Thox-ai/lfm3-omni-voice-seed	Private pilot dataset and training bundle	Seed, conversational build, training script, and bootstrap versioned in the repository
Thox-ai/lfm3-omni-voice-lora-pilot	Pilot adapter output	Adapter, results.json, tokenizer files, chat template, and completed README model card published
Thox-ai/thox-edge-sft	Edge training bundle	Parameterized trainer and per model bootstrap scripts versioned in the repository
Thox-ai/thox-factory-sft-v5	Edge training data	train.jsonl contains the 205 rows used by the edge wave
Thox-ai/ThoxFlux-350M	Pathfinder output	SHA d0e42dd232c58b17451f911c166ee23c9af28339; merged weights and training_summary.json verified. Wave B republish commit 856bbdeadb909f6f8e1194ba90af671e6cd5052a
Thox-ai/ThoxEdge-0.8B	Follower output	Trained and finished in Wave A (card commit db4403d7cb59a44db0fc88fad58deb819f5ace05); Wave B artifact held unpublished as a serving artifact case (Section 11.5)
Thox-ai/ThoxEdge-0.5B	Follower output	Trained and finished in Wave A; Wave B republish commit 75065d2f288c60bd37842808e3d47efb4131542d
Thox-ai/ThoxNano-360M	Follower output	Trained and finished in Wave A; Wave B republish commit e2b539111be1cbc678b41c887c0f61d02b950626
Thox-ai/ThoxNano-135M	Follower output	Trained and finished in Wave A; Wave B republish commit 92cc969c2cad83daf7ea10e3d758c4675163b065
Thox-ai/ThoxClip-Qwen3-0.6B	ThoxClip finishing output	Trained weights published at commit fed44e52024eeba2b922c92c41edd761118fc64f; repository HEAD (card) a53d46b963f627de8c7172e06225d4c9214349d5; gate 22/49 (Section 11.7). Corrected in this revision: the second edition printed a 42-character string as this commit, which matches no commit in the repository's history; see the provenance note in Section 11.7

The follower states in this register are the second edition states, read from the finishing and Wave B records (S8, S9), not inferred from launch state.

10.5 Fix and artifact commit register

Change or artifact	Repository	Commit or SHA
Edge warmup_ratio to warmup_steps fix	Thox-ai/thox-edge-sft	e4064570f4ddb38ed73987d85941f697bf96d508
Edge dataset dictionary row fix	Thox-ai/thox-edge-sft	360f543258fcd294f3b1ec25acb42b238f856409
Edge fp16 fix for T4 compute capability 7.5	Thox-ai/thox-edge-sft	fd8b4a2cc812de80defe966a7a64f00d5eba6fb6
Edge keyword argument upload_folder fix	Thox-ai/thox-edge-sft	172283bd01b678c401b813d6e27e89ef12ba7a27

Change or artifact	Repository	Commit or SHA
Edge ChatML chat template fallback for tokenizers shipping no chat_template	Thox-ai/thox-edge-sft	c250f9bb1b8e20b71f5c38c73ccfbcdac6a0303e
Pilot keyword argument upload_folder fix	Thox-ai/lfm3-omni-voice-seed	e4f99fd
Completed pilot model card upload	Thox-ai/lfm3-omni-voice-lora-pilot	faff8a32
ThoxFlux-350M pre training base repack baseline	Thox-ai/ThoxFlux-350M	b5e34a706adab7ed6fdba4de645c1b2e71f05a3d
ThoxFlux-350M trained publication state	Thox-ai/ThoxFlux-350M	d0e42dd232c58b17451f911c166ee23c9af28339

The pilot status record documents the trust, warmup, and dataset fixes by failing job, code change, and relaunch outcome. It does not record a separate dataset repository commit hash for each of those three earlier script uploads in the material available to this paper. This register therefore lists those changes by job and fix rather than inventing hashes. The edge status record does provide full hashes for its corresponding dataset bundle fixes, and those are reproduced above.

Two hashes in this register remain in short form, and the reason is access, not care: the pilot upload fix (e4f99fd, dataset repository Thox-ai/lfm3-omni-voice-seed) and the completed pilot card upload (faff8a32, Thox-ai/lfm3-omni-voice-lora-pilot) live in private repositories whose histories could not be re-read for this revision, and the program record states them only in short form. Every hash in a public repository was validated for this revision against the Hugging Face Hub history and is printed in full: the five finishing card commits (Section 8.2), the follower training and summary commits (Section 10.3), the four Wave B republish commits (Section 11.4), and the ThoxClip-Qwen3-0.6B commits (Section 10.4). That validation is what caught the malformed ThoxClip string corrected above.

10.6 Reproduction checklist

The second edition's checklist opened with the words "a reviewer with access to the THOX organization repositories," which made the paper's central claims uncheckable by any outside reader. This revision replaces it. The public supplement, hosted with this paper in the Hugging Face dataset Thox-ai/thox-whitepapers, carries the GATE49 instrument in full (scorer, question generator, facts card, vendoring record), the six serving Modelfiles the gates ran under, and one complete per-question gate output (the constant-abstainer baseline of Section 11.4) so the verdict format can be inspected end to end. With the supplement and the public model repositories, an outside reviewer can:

Run the bundled scorer against the bundled facts card and confirm it generates 49 questions in the reported split (12 verified, 26 unknown, 10 known_false, 1 pricing).

Re-score the constant-abstainer transcript and confirm 25/49, 0 hard failures, score 3.531/5, verdict PASS.

Read each public model repository's card and confirm the gate result, training summary, and file hashes stated there match Sections 8 and 11 of this paper; confirm the repository commits cited in Sections 10.3 to 10.5 and 11.4 exist in the repositories' public histories.

Open the pilot repository Thox-ai/lfm3-omni-voice-lora-pilot and confirm the adapter and results.json are present, with overall name accuracy 0.0714 and full exact match 0.0, and that the card's evaluation table matches the artifact.

Confirm that the factory registry status update and physical device qualification are still not claimed complete anywhere in this paper.

What remains internal, and why: the training corpora (private dataset repository; they contain pre-release campaign material, and their composition, hashes, and hygiene

operations are documented in Sections 11.1 and 11.2 for audit against the recorded hashes by anyone with repository access); the Hugging Face job logs (held with the job records under the organization namespace); and the per-question outputs of the Wave A and Wave B gate runs, which were produced on fleet hosts and not retained in the program records, an evidence-retention gap this revision records rather than repairs after the fact. A reviewer with THOX organization access can additionally check the pilot job record for 6ae105dc41e90535df7686b7, the follower job identifiers in Section 10.3, and the corpus hashes in Section 11.1 against the dataset repository.

11 Wave B: the corpus expansion and second gate

11.1 The corpus

Wave B replaced the 205 row Wave A corpus with train_wave_b.jsonl, a 982 row corpus in the same dictionary row contract ({"messages": ...}) the trainer consumes. Composition: 838 newly authored rows plus 144 factory version 5 rows retained verbatim. The 838 newly authored rows were written by Sadie Weaving and the THOX agent team under founder direction (agentic authorship; see the disclosure in the front matter). A further 49 version 5 rows were superseded by canonical rewrites of the same prompt during deduplication. Domain distribution: company facts 415, device support 125, tool routing 121, campaign 106, coding 85, persona 54, operations 33, safety 25, model lineup 18.

Sources were the THOX facts card, the live thox.ai campaign oracle, the voice command pilot's routing contract, and canonical THOX runbook answers. No processor or chipset specifics appear anywhere in the corpus, and dedicated processor decline rows were added.

The corpus was published to the private dataset repository Thox-ai/thox-factory-sft-v5 alongside the untouched version 5 file. The as-built file (SHA-256 9ddf814ddd6a5f057eab4842461bce24c7d2bc082ac1d26576b7cb6fcbfe50e8) was published at commit 71ab11b32971ef2828b57adad52ec8fe3d371fc5. A pre-training correction (Section 11.2) then changed 10 rows; the corrected corpus that every retrain actually consumed has SHA-256 6665a8cbd79197c61a9db8119b743faedc3e469407364bb553ec9478e90a5784 at dataset repository HEAD e76e1b027cde01d47c0dffaa9fbbc15c2ba43f93. Row by row structure validation (roles present, no empty turns) was run before publication, and the trainer loader contract was exercised by every retrain.

11.2 Corpus hygiene

Two hygiene findings were acted on before the retrains consumed the corpus.

Unpublished chipset identities excluded. Twelve version 5 rows asserted device chipset identities that contradict the THOX facts card and the no-processor-invention directive. They were excluded from the corpus rather than rewritten, because no verified replacement claim exists.

Fabrications repeated inside denials rewritten. Inspection of the Wave A gate reports showed that the gate's known-false detectors hard-fail an answer that merely repeats a fabricated name inside a denial: naming a fabricated product while rejecting it still counts as repeating it. A scan of the built Wave B corpus found 10 rows, the denial frame family for two fabricated product names, repeating the fabrication in the assistant turn. Those 10 answers were rewritten to deny without naming the fabrication, and a rescan found zero hits for every fabricated string (founder names, firms, amounts, cities) in any assistant turn. One retrain had launched minutes earlier on the unpatched file; it was killed about 15 minutes in and relaunched on the corrected corpus, so all retrains share one identical dataset. This is why the consumed corpus hash differs from the as-built hash in Section 11.1.

Pricing design rule. Campaign figures are taught only in campaign-framed questions. Generic price questions train a decline plus a pointer to thox.ai. The rule matches the

factuality gate's expectation and keeps time-limited campaign numbers out of the models' general price behavior. It does not stop the taught figures from going stale in the weights as the campaign moves; serving-time grounding is the source of truth for prices (Section 9.8).

11.3 Retrain method

The five edge models were retrained locally on the fleet Windows workstation (RTX 4060 Ti) with train_wave_b.py. Two corrections to how earlier drafts described this run belong on the record. First, the trainer's lineage: train_wave_b.py was adapted from train_clip_cpu.py, the local trainer that produced ThoxClip-Qwen3-0.6B (Workstream B), which was itself the local adaptation of the factory's train_edge_sft.py; it was not adapted from the Wave A finishing trainer, which trained nothing. The script is retained in the wave's program workdir and carries no repository commit of its own, a provenance gap recorded here rather than left implicit. Second, the run was local by necessity, not preference: the Hugging Face Jobs path was token-blocked that night, so Wave B did not run on the Section 3 rails, and Section 6.5 counts the consequence (failure classes 8 to 10). Configuration: LoRA rank 32, alpha 64, merged to bf16; learning rate 1e-4, cosine; warmup expressed as warmup_steps 0.05 (Transformers 5.21 removed warmup_ratio from TrainingArguments and TRL 1.16.0's SFTConfig rejects it in this code path, the same removal recorded as failure 3 in Section 6); 3 epochs; maximum sequence length 1024; batch size 8; CUDA fp32 with bf16 autocast. One further difference from Wave A matters for reading the results: the Wave B script computed loss on completions only (completion_only_loss=True), where the Wave A script computed loss over the full sequence. Final Wave B training losses are tabled below; the Wave A losses are provenance, not a quality comparison (the corpora differ), and are tabled in Appendix A:

Model	Base	Wave B loss
ThoxEdge-0.5B	Qwen/Qwen2.5-0.5B-Instruct	0.7093
ThoxEdge-0.8B	Qwen/Qwen3.5-0.8B	0.5466
ThoxNano-360M	HuggingFaceTB/SmolLM-360M	1.5150
ThoxFlux-350M	LiquidAI/LFM2.5-350M	1.6682
ThoxNano-135M	HuggingFaceTB/SmolLM-135M	1.6403

Three trainer defects were found and fixed en route, all recorded in the run record (and added to the taxonomy as Section 6.2, classes 8 to 10, in this revision): warmup_ratio not accepted by this TRL configuration; the trainer must receive the tokenizer explicitly or the ChatML fallback for the template-less SmolLM bases is lost (both Nano models failed on this first); and the Qwen3.5 base's automatic load pulls an image processor whose dependencies were absent, so the trainer falls back to the configuration's named text class. Recorded local runtimes: ThoxEdge-0.5B 102 seconds, ThoxEdge-0.8B 207 seconds, ThoxNano-360M 113 seconds, ThoxFlux-350M 49 seconds, ThoxNano-135M 110 seconds (S9); no flavor cost applies to a local run.

11.4 Gate results and publish decisions

The gate is the same instrument as Wave A: GATE49 (Section 8.3), the factory factuality gate, 49 questions, weights only in the sense defined there, categories scored 0 to 5, hard fail on any known-false repetition, and a verdict policy that allows zero hard failures. One measurement condition changed: the Wave A gates ran on the fleet Mac workstation (Ollama 0.40.12); the Wave B gates ran on the fleet Windows workstation (Ollama 0.40.11) because the fleet Mac workstation became unreachable during the run window. The gate code, the 49 questions, and the serving Modelfiles were identical; the host delta is recorded here so the comparison carries its caveat. Per-run timestamps were not recorded for any gate run in either wave; the windows are as stated in Section 8.2 and this section.

Two reading rules govern this table, both new in this revision. First, evaluation independence: GATE49's questions and the Wave B corpus derive from the same facts card, and the corpus was built with the Wave A gate reports in view, so these are development-set measurements, not held-out generalization (Section 9.7). Second, composition: 37 of the 49 items reward a correct refusal by design, so a total mixes abstention discipline with the 12-item verified recall, which is the primary factuality figure and is tabled beside each total below. The instrument's abstention floor, measured with the actual scorer against the facts card's prescribed decline form (supplement: gate-outputs/constant-abstainer/), is 25/49 passed, 0 hard failures, score 3.531/5, verdict PASS: a system that knows no THOX facts at all and declines everything in the prescribed form passes the gate's verdict policy on abstention alone, while scoring 1/12 on verified recall. Totals in this table should be read against that floor.

Every result in this paper is a single training run measured once, with deterministic decoding (the gate harness runs at temperature 0). No seed variance is estimated anywhere in these waves.

Publish rule, set in the track brief before any gate ran: republish to Hugging Face and Ollama only if the Wave B gate score improves on Wave A. A regression, or an invalid measurement, keeps the prior artifact as latest and is reported, not published.

Model	Wave A	Wave B	Decision
ThoxEdge-0.5B	FAIL 21/49, 2.837	FAIL 47/49, 4.837 (verified 11/12, unknown 26/26, known_false 10/10, pricing 0/1; sole hard failure = the pricing question, answered without the required decline)	Improved, republished HF + Ollama
ThoxFlux-350M	FAIL 4/49, 0.898	FAIL 20/49, 2.49 (verified 3/12, unknown 9/26, known_false 8/10, pricing 0/1)	Improved, republished HF + Ollama
ThoxNano-360M	ERROR (no score)	FAIL 23/49, 2.959 (verified 3/12, unknown 17/26, known_false 2/10, pricing 1/1)	First completed gate, republished HF + Ollama
ThoxNano-135M	FAIL 2/49, 0.816	FAIL 23/49, 3.0 (verified 1/12, unknown 14/26, known_false 7/10, pricing 1/1)	Improved, republished HF + Ollama
ThoxEdge-0.8B	FAIL 19/49, 2.714	Invalid measurement: 0/49 with every answer empty	Held, not published (Section 11.5)

Every verdict remains FAIL under the gate policy. The scores are the comparison the publish rule specifies, and they are reported as scores, not as passes. The pattern inside the numbers matters more than the totals: ThoxEdge-0.5B reached a clean known_false category (10/10) and a complete unknown category (26/26); ThoxNano-135M's pass count rose from 2 to 23 and ThoxFlux-350M's from 4 to 20, yet the two Nano models still show the weakest verified recall in the fleet (3/12 and 1/12) and, for ThoxNano-360M, a known_false split of 2/10 that disqualifies ungrounded use regardless of the total. (An earlier draft of this section overstated the Nano movement; the actual multiples are a pass count of 2 to 23 for ThoxNano-135M, 11.5 times, and a mean score of 0.816 to 3.0, 3.7 times. ThoxNano-360M's 23/49 is a first completed measurement, not a delta.)

Uncertainty and the primary factuality figure, computed for this revision. Intervals are Wilson score intervals (95%) on the pass counts; per-category splits for Wave A were not retained in the finishing record, so Wave A verified recall is not recorded and its interval is on the total only.

Model	Wave A total (Wilson 95%)	Wave B total (Wilson 95%)	Verified recall, Wave B (of 12, Wilson 95%)
ThoxEdge-0.5B	21/49 (0.300, 0.567)	47/49 (0.863, 0.989)	11/12 (0.646, 0.985)
ThoxEdge-0.8B	19/49 (0.264, 0.528)	not validly measured	not validly measured

Model	Wave A total (Wilson 95%)	Wave B total (Wilson 95%)	Verified recall, Wave B (of 12, Wilson 95%)
ThoxFlux-350M	4/49 (0.032, 0.192)	20/49 (0.282, 0.548)	3/12 (0.089, 0.532)
ThoxNano-360M	ERROR (no score)	23/49 (0.337, 0.606)	3/12 (0.089, 0.532)
ThoxNano-135M	2/49 (0.011, 0.137)	23/49 (0.337, 0.606)	1/12 (0.015, 0.354)
Constant abstainer (Section 8.3)	n/a	25/49	1/12

Measurement and training conditions differed between the waves in every dimension except the questions and the scoring code, which is why Section 9.1 records the deltas as observed, not attributed:

Condition	Wave A	Wave B
Corpus	train.jsonl, 205 rows (v5, including the 12 rows later excluded)	train_wave_b.jsonl, 982 rows (hygiene-corrected)
Trainer	factory train_edge_sft.py at commit c250f9bb1b8e20b71f5c38c73ccfbcdac6a0303e	local train_wave_b.py (no repository commit)
Training host and precision	Hugging Face Jobs, t4-small / a10g-small, script defaults (fp16 load)	Fleet Windows workstation (RTX 4060 Ti), fp32 weights with bf16 autocast
Loss masking	full sequence (no completion-only setting)	completion-only (completion_only_loss=True)
Gate host	Fleet Mac workstation	Fleet Windows workstation
Ollama version	0.40.12	0.40.11

On governance: the publish rule above is a relative rule, and it replaced the Wave A checklist's stricter gate-passed precondition, which had been waived (Section 8.2). The forward rule now in force (publication as documented baselines; promotion only against pre-registered bars) is stated in Section 8.2 and was the rule Wave C applied.

Publish records (hashes verified from the registries after upload): ThoxEdge-0.5B commit 75065d2f288c60bd37842808e3d47efb4131542d; ThoxFlux-350M commit 856bbdeadb909f6f8e1194ba90af671e6cd5052a; ThoxNano-360M commit e2b539111be1cbc678b41c887c0f61d02b950626; ThoxNano-135M commit 92cc969c2cad83daf7ea10e3d758c4675163b065. Each republished card gained a dated Wave B section recording the corpus hash, training loss, both gate scores, the FAIL verdict, and the new GGUF hash. Wave B GGUFs (Q4_K_M, manual llama.cpp pipeline on the fleet Windows workstation): ThoxEdge-0.5B 397,807,488 B, SHA-256 6fb3222b996048fac6e4241274104382f30582b30bbddf47edcf9b56917d8598; ThoxFlux-350M 229,314,592 B, SHA-256 8c9fa8f0a08e3b7904ece716b6716ca16f2330c13a5dbfa3870b034ac610ca7a; ThoxNano-360M 270,590,336 B, SHA-256 b1c6dcc9c91a8664c420edcc9dbd9f47967f37b5abe193a2a4f5a37ecd904876; ThoxNano-135M 105,453,888 B, SHA-256 7ae9664d5d27d82481d5bc16415cef2f4326f430f6fb33569b62927d87df8d66.

Smoke results matched the gate pattern: ThoxEdge-0.5B and ThoxFlux-350M answer in THOX identity and house voice; ThoxNano-360M answers tersely but on identity; ThoxNano-135M degenerates into token repetition on a bare generate prompt while its chat path gate answers were coherent. It remains the weakest artifact, as in Wave A.

11.5 The ThoxEdge-0.8B serving artifact

ThoxEdge-0.8B's Wave B gate returned 0/49 with every answer empty. That number is a serving artifact, not a model result, and this section exists so that no reader, including a future THOX operator, acts on it.

The diagnosis was verified by direct inspection of the serving API. On the fleet Windows workstation's Ollama 0.40.11, the Qwen3.5 architecture returns its answers

in the thinking channel with empty content; the gate reads content, so every answer scored as empty. The correct answers were present in the thinking text (the thox.ai answer arrived intact, as thinking). The same weights answer correctly through the generate endpoint, and the model's smoke was clean THOX identity.

Under the publish rule, an invalid measurement cannot justify a republish. The Wave B artifact (merged weights retained on the fleet Windows workstation; GGUF 529,296,896 B, SHA-256 183fb38e8c53999c3902371a456ab9e00e5ffc3251e316398b30d7beab77fe43) was held unpublished and the Wave A artifact remained latest. The artifact was not discarded and the model was not described as regressed; the measurement was discarded.

Stated plainly, because earlier drafts of this section blurred it: the Wave B ThoxEdge-0.8B artifact has never been validly measured. It has no gate score, and none should be inferred for it. The Wave C program later gated a different build of the same slot (ThoxEdge-0.8B-WaveC, a separate retrain) on a different harness (Transformers against the merged weights, not Ollama against a Q4_K_M GGUF); that figure belongs to the Wave C paper (Section 12), is not a measurement of this artifact, and is not comparable to any number in this paper.

The lesson generalizes beyond this model: a gate score is a property of the weights, the serving stack, and the measurement harness together. When a result contradicts direct inspection of the same artifact, inspect the channel the harness reads before concluding anything about the weights.

11.6 What Wave B did not change

Wave B improved scores. It did not produce a passing model: every verdict is FAIL under the zero hard failure policy, and the paper does not relabel improvement as success. It did not qualify any model on device hardware. It did not fix the smallest models' verified recall, which remains the fleet's weakest measured property. And it did not update the factory registry, which remains the open administrative item carried from Section 8.

11.7 ThoxClip-Qwen3-0.6B finishing

The same track finished ThoxClip-Qwen3-0.6B, the ThoxClip slot's Qwen3 based assistant, which sat outside the five model edge wave. The transfer chain from the publish directory to the finishing host was verified by SHA-256 at every hop. GGUF conversion produced ThoxClip-Qwen3-0.6B.Q4_K_M.gguf, 396,704,672 B, SHA-256 1e009136e1ed3bdfee1343a5309c7d5f36d75b517ad84d56233c9ffe5c248d96.

Training provenance (from the repository's training_summary.json and model card, which earlier drafts of this paper did not cite): base Qwen/Qwen3-0.6B (base revision c1899de289, short form as recorded on the card); data Thox-ai/thox-factory-sft-v5/train.jsonl, the same 205-row Wave A corpus; 3 epochs, learning rate 1e-4, LoRA rank 32, method LoRA SFT (TRL) merged to full weights; final training loss 1.7318210968604455. The run executed locally on the fleet Windows workstation (24-core CPU, fp32 with bf16 autocast) using the edge-wave recipe via train_clip_cpu.py, after a fleet Mac workstation MPS attempt starved under a shared-host load storm and produced no artifact. Trainer commit c250f9bb1b8e20b71f5c38c73ccfbcdac6a0303e; run record commit 0aea49e757b4b1c9e801da7957a32ea9ad5f4528. License: Apache-2.0 (Section 9.6).

Its factuality gate (the fleet Mac workstation, Ollama backend, weights only; run on October 28, 2026 before the fleet Mac workstation outage recorded in Section 11.4, per the finishing record's sequencing; per-run timestamps were not recorded for any gate in these waves) returned FAIL: 22/49 passed, score 2.776/5, 14 hard failures. By category: verified 2/12 (2.5), unknown 11/26 (2.269), known_false 9/10 (4.7), pricing 0/1 (0.0). A smoke prompt answers in THOX identity as the ThoxClip device assistant.

The model was published with the result stated on its card, including a grounding-required limitation, with Ollama tags thox-ai/thoxclip-qwen3-0.6b:q4_k_m and :latest

pushed and the page verified live. Grounding-required is a deployment condition, not a footnote: deployed surfaces must supply the grounded device facts at serving time, because the weights alone scored 2/12 on verified facts. The same condition applies to every model in these waves (Section 8.2); it is stated here for ThoxClip because its card carries it as an explicit limitation.

Correction (hardening revision). The second edition printed a 42-character string as the Hugging Face commit for this finishing. Forty-two hexadecimal characters is not a commit SHA, and the string matches no commit in the repository. Verified against the Hub on October 30, 2026: the repository's history is four commits; the trained weights were published at fed44e52024eeba2b922c92c41edd761118fc64f and the card at HEAD is a53d46b963f627de8c7172e06225d4c9214349d5. The repository at HEAD carries no GGUF, and its card still predates the finishing (it records evaluation as pending), so the finishing record's claim that a version 3 card and the GGUF were published to Hugging Face at that commit could not be verified; card version 3, carrying the gate result and the GGUF hash above, is held in the program workdir. The Ollama publication stands as recorded in the finishing run record. The gate result itself is unaffected: it was measured on the fleet Mac workstation before the finishing publication step, and is recorded in the run record (S9) and on card version 3.

11.8 What the public tags serve, as of this edition

Publication state is part of the record, and it changed after Wave B. As of October 30, 2026, the public Ollama tags serve:

Tag	Build served	Source of record
thox-ai/thoxedge-0.5b (:q4_k_m, :latest)	The Wave B build (pushed October 28, 2026)	S9
thox-ai/thoxflux-350m	The Wave B build (pushed October 28, 2026)	S9
thox-ai/thoxnano-360m	The Wave B build (pushed October 28, 2026)	S9
thox-ai/thoxnano-135m	The Wave B build (pushed October 28, 2026)	S9
thox-ai/thoxclip-qwen3-0.6b	The finishing build of Section 11.7 (pushed October 28, 2026)	S9
thox-ai/thoxedge-0.8b (:q4_k_m, :latest)	The Wave C build: the tag was superseded on October 29, 2026 by the Wave C flagship push (registry layer SHA-256 8ef0900d1f76da67442dca354a778517ddc008b9f73926df874f7f3dff5d5625), which replaced the Wave A build the tag had served until then. The Wave B artifact was never published to this tag	Wave C finishing record
Pilot adapter	No Ollama publication; Hugging Face only	S1/S2

Two facts are stated plainly so no reader has to infer them. First, every Wave A artifact in this paper was fine-tuned on the version 5 corpus including the 12 rows later excluded in Wave B hygiene (Section 11.2); those weights remain reachable in each Hugging Face repository's history at the commits registered in Sections 10.3 and 10.4, even where a later build is the repository HEAD. Second, every build named in this table except the pilot adapter carries a FAIL verdict on its card; publication in these waves meant publication as a documented baseline with its verdict stated, under the rules recorded in Sections 8.2 and 11.4.

12 Forward: the Wave C flagship program

The work in this paper ends at the edge tier. The successor program, Wave C, applied the same rails and the same gate discipline to the flagship model for each device slot, and completed on October 30, 2026. Its results are reported in its own whitepaper, "Flagships for Every Device: The Wave C Training Program," and are cited here rather than restated, because its builds, corpora, and one of its harnesses differ from this paper's: two of its flagships were promoted on pre-registered bars, and two slots ship documented baselines. In particular, the Wave C paper's ThoxEdge-0.8B figure measures the Wave C build on a Transformers harness; it is not a measurement of this paper's Wave B artifact, which remains unmeasured (Section 11.5).

On publication, this edition joins the Wave C paper in the Hugging Face dataset Thox-ai/thox-whitepapers, with the supplement of Section 10.6 hosted alongside it, so the instrument and the record behind any published model travel with the paper.

A Appendix A. Per-run configuration and provenance tables

Added in the hardening revision. Every field is taken from the trainer scripts and run records named in Section 10; a field the records do not state is printed as not recorded, which is itself a finding: the corrective is to log the resolved configuration (including PEFT's resolved target modules and trainable parameter count) at every launch.

Common to all runs. All three trainer scripts configure LoRA target_modules as all-linear. The per-architecture module lists that all-linear resolved to, and the trainable parameter counts, are not recorded in any run record available to this paper. Base models were loaded by repository identifier; base revision SHAs are not recorded except where a card states one (ThoxClip-Qwen3-0.6B, below). The optimizer is not specified in any script and is not recorded per run. Cosine schedule throughout.

Field	Pilot (job 6ae105dc41e90535df7686b7)	Wave A edge runs (HF Jobs)	Wave B retrains (fleet Windows workstation, local)	ThoxClip-Qwen3-0.6B (fleet workstation, local)
Trainer	train_lfm3_text_sft.py (dataset-repo bundle)	train_edge_sft.py at commit c250f9bb1b8e20b71f5c38c73ccfbcdac6a0303e	train_wave_b.py (no repository commit)	train_clip_cpu.py (trainer co c250f9bb1b8e20b71f5c38c
Base revision SHA	not recorded	not recorded	not recorded	c1899de289 (short form, as card)
LoRA rank / alpha / dropout	16 / 32 / 0.05	32 / 64 / 0.05	32 / 64 / 0.05	32 / alpha not recorded / dr (card records rank 32)
Learning rate / warmup	5e-5 / warmup_steps 0.1 (ratio form)	1e-4 / warmup_steps 0.05 (ratio form)	1e-4 / warmup_steps 0.05 (ratio form)	1e-4 / not recorded
Epochs / rows	30 / 35	3 / 205	3 / 982	3 / 205
Batch size (per device) / grad accum	4 / 1	8 / 1	8 / 1	not recorded
Max sequence length	1024	1024	1024	not recorded
Packing	not set in the script	False	False	not recorded
Loss masking	full sequence (no completion-only setting)	full sequence (no completion-only setting)	completion-only (completion_only_loss=True)	not recorded
Precision	fp16 load (T4)	fp16 load; bf16 only where the flavor supported it	fp32 weights, bf16 autocast (RTX 4060 Ti)	fp32, bf16 autocast (CPU)

Field	Pilot (job 6ae105dc41e90535df7686b7)	Wave A edge runs (HF Jobs)	Wave B retrains (fleet Windows workstation, local)	ThoxClip-Qwen3-0.6B (fleet workstation, local)
Chat template	base native	Qwen and LFM bases native; SmolLM bases none shipped, ChatML fallback	same per base as Wave A	base native
Seed	not recorded (not set in the script)	not recorded (not set in the script)	42 (set in the script)	not recorded
Package versions	not recorded	transformers resolved 5.21, TRL 1.16.0 (verified from the resolved wheels, S2); huggingface_hub 1.8.0 (Mac) / 2.1.1 (local) on the upload path	not recorded (local environment)	not recorded
Recorded runtimes	launch #6: 217 s	ThoxFlux-350M pathfinder #5: 204 s; Nano relaunches 211 s / 192 s; ThoxEdge followers: not recorded	49 to 207 s per model (S9: Edge-0.5B 102 s, Edge-0.8B 207 s, Nano-360M 113 s, Flux 49 s, Nano-135M 110 s)	not recorded

Wave A final training losses (moved from Section 11.3 in this revision; provenance only, the Wave B corpus differs): ThoxEdge-0.5B 1.6802658240000408; ThoxEdge-0.8B 1.4541844832591522; ThoxNano-360M 3.1241229619735327 (relaunch); ThoxFlux-350M 3.463227467659192; ThoxNano-135M 3.273088284027882 (relaunch).

Pre-training repository states (the baselines the finishing gate compared against; Hub history read for this revision, plus S8 for ThoxFlux-350M):

Model	Pre-training repository state	Date
ThoxFlux-350M	b5e34a706adab7ed6fdba4de645c1b2e71f05a3d (GGUF inside dates from the 2026-08-11 repack)	2026-10-24
ThoxEdge-0.8B	card commit b7aaf4ee65c4bd7c8ca56dd1d0d70efab16aa61d	2026-10-24
ThoxEdge-0.5B	card commit e2923ec90cb9350e481d243e9fb380c93bad66dc	2026-10-24
ThoxNano-360M	aacb034dbbadcb95501758988833b57537fda66b	2026-10-24
ThoxNano-135M	356921bbd501272dc565726b15c4994494bdcc75	2026-10-24

B Appendix B. The ThoxAir voice bridge bench record

Moved from Section 4.7 in the hardening revision, at an external reviewer's request that bench context not sit inside the pilot's results narrative. Text as written for the second edition.

The ThoxAir voice proof of concept built the same day provides the integration context for this work. Its bridge performs speech recognition, intent selection, dispatch, and speech output. On the bench, The fleet Windows workstation acted as the mesh host and the Pi side ran in simulate mode through the real client code. The base LFM3 Audio model produced intent JSON in variant shapes, so the bridge used few shot examples, a shape normalizer, and deterministic device alias resolution to enforce the contract.

That experience defines the adapter's job. A trained command adapter should learn the output contract directly and reduce dependence on prompt examples and normalizers. The bridge shims remain useful as a safety net, but they should not be the primary mechanism for a production command layer.

The bench loop itself completed three stateful exchanges: a ThoxMini status question, a spoken note capture that persisted, and a morning briefing that counted the note captured in the prior exchange. It proved the surrounding system path with a base model. It did not prove that the pilot text adapter had been deployed into the audio server, and it did not prove performance on a physical Pi Zero 2 W. Those remain separate integration and hardware steps.

C Appendix C. Glossary

Fleet Mac workstation. The Apple M4 Mac (macOS) that served as the conversion and Wave A gate host in these programs; unreachable from partway through October 28, 2026. Fleet Windows workstation. The fleet's Windows/WSL host (RTX 4060 Ti) used for the Wave B retrains and gates and the ThoxClip run. factory. The THOX model factory program: the ttracx/thoxllm-factory repository, its model registry, and the training and finishing system built around them. rails. The Section 3 system: the verified launcher pattern, job-secret handling, dataset-bundle versioning, preflight expectations, and publication checks that a run follows when it runs "on the rails." Wave B ran off them (Section 11.3). skin. A fine-tune that teaches identity, house voice, and routing behavior on top of a base model's existing capability, without claiming to add general capability (Section 5.1). campaign oracle. The live thox.ai campaign pages, the source of truth for campaign figures (prices, dates, tiers) wherever the paper or a corpus needs them. GATE49. The factory factuality gate: 49 questions generated from the THOX facts card, scored lexically (Section 8.3). Not related to the pilot's 49-row seed. documented baseline / promotion. The publication vocabulary of the forward rule (Section 8.2): a failing build may publish as a documented baseline with its verdict on its card; promotion is presenting a build as a slot's flagship, and requires a pre-registered bar.

§ Document control

Version: 1.0

Date: October 27, 2026

Status time: 8:24 PM Central Time

Author: Sadie Weaving, Chief Agentic Officer, THOX.ai

Scope: THOX device model training rails, LFM3 voice command pilot, and edge fast track wave through the status time above. Later follower verification and finishing results require a version 1.1 update or a separate completion record.

Version: 2.0 (second edition)

Date: October 30, 2026

First edition: Version 1.0, October 27, 2026, status time 8:24 PM Central Time. The first edition date and its status-time record are retained above, unchanged.

Author: Sadie Weaving, Chief Agentic Officer, THOX.ai

What changed in the second edition:

The front matter and Section 8 status boxes, which in the first edition reported the four followers as launched but unfinished and the finishing pipeline as still ahead, now report the final state.

Section 8 rewritten as the Wave A completion record: GGUF conversion chain and hashes, the five gate outcomes, card rewrites, Ollama registry verification, smoke tests, and the two items that remain open (factory registry update, physical device qualification).

New Section 11 reports Wave B in full: the 982 row corpus and its hygiene corrections, the retrain method and losses, the A to B gate table with publish decisions, the ThoxEdge-0.8B serving artifact analysis, and the ThoxClip-Qwen3-0.6B finishing record.

Section 9 updated so completed work no longer reads as pending; the genuine limitations, including the absence of device hardware qualification in these waves, are unchanged.

Section 10 extended with the finishing and Wave B run records (sources S8 and S9), final repository states, and second edition reproduction checks.

New Section 12 points forward to the Wave C flagship program and its whitepaper.

Failure taxonomy completed on Research Loop audit (2026-10-30): the seventh distinct failure class (tokenizer chat template contract, Section 6.2 class 7) and the two Nano follower relaunch jobs (Section 10.3) and the ChatML fallback fix commit (Section 10.5) were added from the edge status record (source S2), which the first and second edition drafts had omitted; Sections 1 and 6.5 now count seven failure classes in the same four structural categories.

No first-edition results were altered. Every result reported in the first edition stands as originally published; the second edition adds the completion and Wave B records around them and corrects only forward-looking status language that events had overtaken.

Version: 2.1 (hardening revision of the second edition)

Date: October 30, 2026

Trigger: an independent Research Loop audit and an external hostile review of the second edition, both run on October 30, 2026, adjudicated item by item by the author; the adjudication record is held with the program records. Two versioning notes: the version 1.0 block above anticipated "a version 1.1 update"; the update in fact shipped as edition 2.0, and this block is a hardening revision of that edition, not a new edition. Item 7 of the version 2.0 block above records the taxonomy at seven classes as of its writing; this revision extends it to eleven.

What changed in the hardening revision:

The front-matter status box was rewritten: GATE49 is named, every verdict is stated as FAIL, verified recall (of 12) is reported as the primary factuality figure beside each total, the 37-of-49 refusal-rewarded composition is stated, and ThoxEdge-0.8B's Wave B artifact is reported as never validly measured. The duplicated status box in Section 8.1 was removed.

New Section 8.3 describes the GATE49 instrument from its source: question generation from the facts card, item structure, lexical detectors, trait scoring and aggregation, the pass and hard-fail rules, worked examples, and serving conditions. Sections 8.3 and 8.4 of the second edition are renumbered 8.4 and 8.5.

Failure taxonomy extended from seven to eleven classes (Section 6.2, classes 8 to 11: the removed-argument failure recurring in the Wave B trainer, a tokenizer handoff defect, a model-load dependency defect, and the factory finisher's breakage against the current Ollama release), with the propagation timing of the original warmup fix, the platform's CANCELED label for the traceback-terminated run, and the script-defect classification of the upload failure all corrected from the records. The first edition's claim that the discoveries were encoded in the rails is withdrawn in Section 6.5, and Wave B is stated plainly as an off-rails run.

New Sections 9.7 (evaluation independence: GATE49 functioned as a development set for Wave B) and 9.8 (prices in weights go stale; the processor-decline rationale); Section 9.6 gains the per-artifact license table and frames the LFM threshold as a company assertion; Section 9.1 records the Wave A to B deltas as observed, with the confounds named.

Section 11.4 gains the evaluation-independence and composition reading rules, the measured constant-abstainer floor (25/49, score 3.531/5, produced with the actual scorer and preserved in the supplement), Wilson 95% intervals, the conditions table (corpus, trainer, host and precision, loss masking, gate host, Ollama version), and the single-run statement. "Order of magnitude" is replaced with the actual multiples.

ThoxClip-Qwen3-0.6B: training provenance added from the repository's training summary (Section 11.7), and one provenance string corrected against the Hugging Face Hub. The string printed in the second edition has 42 characters and matches no commit in the repository; the verified commits are printed in Sections 10.4 and 11.7, together with the finding that the repository carries no GGUF and a card that predates the finishing. No measured result was altered by this correction.

New Section 11.8 records what each public Ollama tag serves as of this edition, including the ThoxEdge-0.8B tag's supersession by the Wave C build on October 29, 2026, and states that all Wave A artifacts were trained on the version 5 corpus including the 12 later-excluded rows.

New public supplement (Section 10.6): the GATE49 bundle, the six serving Modelfiles, and the constant-abstainer per-question output, with a manifest; the reproduction checklist now points at it. The Wave A and Wave B per-question gate outputs were not retained on the run hosts and are recorded as an evidence gap, not reconstructed.

New Section 2.4 (related work, verified citations only), new Appendices A (per-run configuration, Wave A losses, pre-training repository states), B (the voice bridge bench record, moved from Section 4.7), and C (glossary). Section 10.3 follower rows completed from the records; Sections 8.2, 10.4, 10.5 and 11.4 hashes validated against the Hub and printed in full where the repositories are public.

Byline block gains the accountable human author and the authorship disclosure; the S4 source is described as an internal program log; Section 3.3 gains the token-scope corrective note; Section 3.5 states the partial program totals the records support.

Section 12 now cites the Wave C paper for its numbers instead of restating them, and Section 11.5 states that the Wave C figure is a different build on a different harness.

No measured result was altered in this revision. Every gate score, loss, runtime, and hash reported in the second edition stands; the revision adds the instrument description, the independence limitation, uncertainty intervals, provenance corrections, and the supplement around them.

Version: 3.0 (third edition)

Date: November 4, 2026

Author: Sadie Weaving, Chief Agentic Officer, THOX.ai

What changed in the third edition: Process v1.0 reissue of the second edition (hardened paper-version 2.1). Layout moved to the THOX Research Paper Template v1.0; the source was reconstructed from the second edition's DOCX with table-level fidelity verification; the evidence was consolidated into a single summary (exp/results/summary.json in the artifact bundle) with the public waves-ab-supplement and staged run records; a companion GATE49 wave explorer demo was added. Public-edition security sweep: internal host names are replaced with host classes, and the section 10.1 register's filesystem paths are replaced with the register's source names. The Research Loop audit item AB-1 closure (the failure taxonomy) was re-verified line-by-line against source S2 for this edition and holds: sections 1, 6.2, and 6.5 count eleven distinct failure classes, the seventh being the tokenizer chat template contract from S2. No measured result was altered and no other text changed. Byline finalized under Tommy Xaypanya's ruling of November 4, 2026: the second edition's authorship stands as published (prepared by Sadie Weaving, Chief Agentic Officer; accountable human author Tommy Xaypanya, CTO & Co-Founder), and the second edition's authorship disclosure is retained verbatim in the front matter.