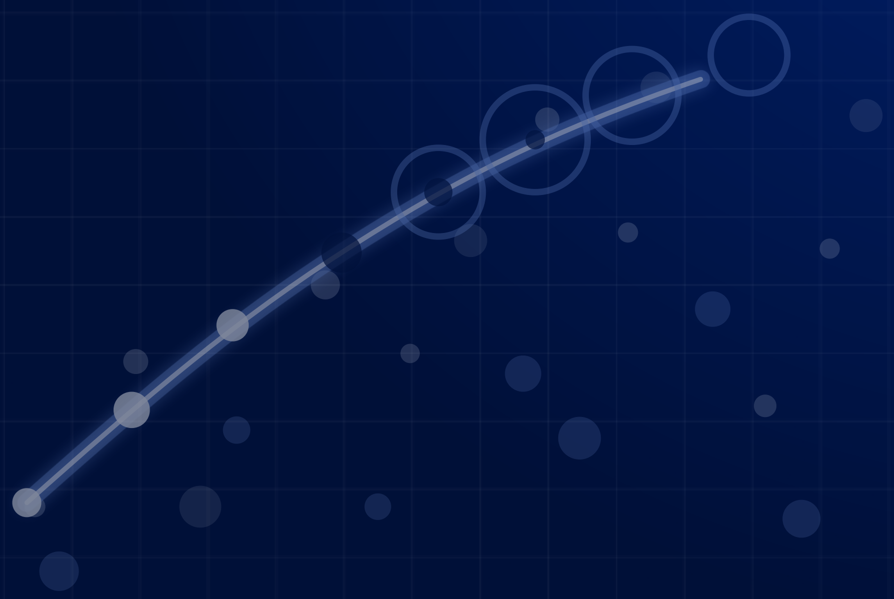
PHYLOOP

PHYLOOP RESEARCH

WHITEPAPER · NO. 002
SOLVER-VALIDATED EVIDENCE



PHYSICS-AI QUALIFICATION WHITEPAPER

When Can a Physics-AI Surrogate Be Trusted?

Solver-Validated Qualification with Field Uncertainty and Solver Fallback

DOCUMENT TYPE	PUBLISHED	VERSION	REFERENCE
Technical whitepaper solver validation	2026-06-14	v1.0	PL-WP-002

WHITEPAPER NO. 002

When Can a Physics-AI Surrogate Be Trusted?

SOLVER-VALIDATED QUALIFICATION WITH FIELD UNCERTAINTY AND SOLVER FALLBACK

Prepared by Aditya Tiwari

Phyloop Research · legal entity: Culturiq Research Private Limited

Scope: 2D fluid-side forced convection · solver: OpenFOAM simpleFoam + thermalEnergyFoam · run date: 14 June 2026

ABSTRACT

This whitepaper reports a solver-validated qualification workflow for AI field surrogates. The workflow creates parameterized OpenFOAM thermal-fluid cases, validates solver outputs, extracts fixed-grid fields, trains a compact neural field operator, calibrates uncertainty, and evaluates when a workflow should serve the surrogate or route a case back to the solver.

PL-WP-002 reports the full corrected validation campaign: a steady thermal solver contract, retained OpenFOAM archives, physical plausibility gates, field-level prediction, uncertainty calibration, and routing diagnostics.

The reported campaign requires base-flow convergence, dedicated thermal solve convergence, terminal-field stability, raw outlet enthalpy sanity checks, and raw local-temperature extrema checks before a case enters the learning dataset. The retained dataset contains 848 solver-qualified cases, all extracted at 128 x 128 resolution and all passing fixed-grid physical plausibility validation.

The trained CompactFNO ensemble predicts temperature and velocity-magnitude fields from geometry, mask, signed-distance, coordinates, and operating parameters. The evidence supports a solver-validated surrogate qualification workflow. It is not a production digital twin, not a replacement for OpenFOAM, and not a general hidden-regime detector.

KEYWORDS

01	02	03	04
multiphysics AI	OpenFOAM	neural operator	uncertainty calibration
05	06	07	
solver fallback	OOD stress test	solver-validated evidence	

CLAIM BOUNDARY

The result qualifies a solver-validated surrogate workflow for known envelope shifts. It does not claim production coolant fidelity, OpenFOAM replacement, general hidden-regime detection, or early detection of hidden physics-regime changes inside the declared envelope.

-- Contents

01 Executive Summary

The result is a solver-validated qualification workflow for physics-AI surrogates. It shows how field models can be evaluated against retained solver evidence, calibrated for uncertainty, stress-tested outside the training envelope, and paired with solver fallback rather than deployed as unchecked replacements.

RETAINED SOLVER CASES	FIXED-GRID FIELDS	PHYSICS VALIDATION
848 solver-qualified and archived	848 128 x 128, fluid-mask metrics	848/848 0 enthalpy/extrema failures
ID 90% COVERAGE	**OOD 90% COVERAGE**	**OOD BAND INFLATION**
0.908 target band [0.87, 0.93]	0.948 reported, not hidden	3.664x uncertainty expands under shift

Thesis. This is credible evidence for a solver-validated surrogate qualification workflow: solver provenance, physical rejection gates, calibrated in-distribution uncertainty, regime stress testing, and conservative routing.

Boundary. The out-of-distribution pools are constructed by known parameter-envelope shifts. Explicit envelope gating is therefore the primary baseline; learned scores are supporting telemetry and research signals.

PRIMARY RESULT

The surrogate is calibrated on in-distribution (ID) cases at the target 90% level (0.908), while out-of-distribution (OOD) error rises materially: ID-test mean relative L2 is 0.207, OOD-test is 0.458, and fresh-OOD is 0.453. The 90% uncertainty band inflates by 3.664x on OOD, but band width ranks OOD error weakly (0.123 Spearman). This is evidence for controlled regime-shift behavior, not a general OOD detector.

CAPABILITIES DEMONSTRATED

- → Generate and retain OpenFOAM cases under a documented steady thermal solver contract.
- → Extract solver fields to a fixed-grid operator dataset with mask/SDF/parameter channels.
- → Train and evaluate a compact FNO ensemble on temperature and velocity-magnitude fields.
- → Report calibrated uncertainty, OOD error degradation, band inflation, and routing diagnostics.

02 Workflow Overview

PL-WP-002 is a solver-to-surrogate qualification workflow. The sequence starts with a defined engineering envelope, keeps only validated solver outputs, trains a field operator on retained archives, calibrates uncertainty, and routes cases back to the solver when the surrogate is outside its validated operating envelope.

PHASE 1	PHASE 2	PHASE 3	PHASE 4
Define and solve	**Validate and retain**	**Train and calibrate**	**Stress and route**
Benchmark geometry, operating envelope, OpenFOAM base flow, and thermal solve.	Convergence, stability, enthalpy, extrema, and fixed-grid plausibility gates.	CompactFNO ensemble, field metrics, split-conformal uncertainty calibration.	ID/OOD evaluation, uncertainty expansion, explicit envelope gate, solver fallback.

End-to-end PL-WP-002 workflow sequence.

STEP	STAGE	ACTION	OUTPUT
01	Envelope	Define geometry, operating parameters, and known OOD stress cases.	case design ledger
02	Base flow	Solve steady velocity field with OpenFOAM simpleFoam.	converged flow archive
03	Thermal solve	Solve steady fluid-side thermal transport over frozen velocity.	temperature field archive
04	Raw checks	Apply residual, stability, enthalpy, and extrema guardrails.	validated or rejected case
05	Field extraction	Resample retained archives to fixed grid with mask, SDF, coordinates, and parameters.	operator-ready NPZ
06	Plausibility	Check fixed-grid physical consistency before learning.	retained field dataset
07	Model fit	Train CompactFNO ensemble on temperature and velocity magnitude.	field surrogate ensemble
08	Calibration	Apply split-conformal uncertainty calibration on held-out ID cases.	90% prediction bands
09	Evaluation	Measure field error and coverage on ID, constructed OOD, and fresh-OOD pools.	performance tables
10	Stress test	Measure whether uncertainty expands under regime shift.	OOD telemetry
11	Routing	Apply explicit envelope gate and compare learned telemetry signals.	serve-or-solver policy
12	Evidence	Publish traceable metrics, figures, and evidence-pack paths.	whitepaper evidence map

03 Experiment Contract

The experiment is a controlled 2D finned cooling-channel benchmark. The study prioritizes solver-validated field data, uncertainty behavior, and evidence reconciliation over production-grade thermal fidelity.

PHYSICS AND SOLVER

Each case solves a steady base flow with OpenFOAM simpleFoam, then solves a dedicated fluid-side thermal energy scalar over the frozen converged velocity field with the project-local thermalEnergyFoam solver. The scalar semantics are temperature_K_steady_transport_stability_window.

The model output fields are temperature T and velocity magnitude |U|. Metrics and loss are computed over the fluid mask; solid target pixels are zero-filled and are not allowed to inflate coverage.

RETENTION GATES

- OpenFOAM base-flow convergence and residual target.
- Thermal residual target and terminal field stability.
- Raw outlet enthalpy sanity check.
- Raw local-temperature extrema check.
- Fixed-grid plausibility validation after extraction.

Varied inputs for the PL-WP-002 validation benchmark.

PARAMETER	ID RANGE	OOD CONSTRUCTION
h_ch	4-10 mm	pushed 25-40% beyond ID range
n_fin	4,5,6,7,8	3,10,12 outside trained discrete set
h_fin	2-6 mm	retained within physically filtered candidate geometry
d_in	2-6 mm	pushed 25-40% beyond ID range
u_in	0.2-1.5 m/s	sampled operating condition
q_w	5-50 kW/m^2	positive wall heat flux heats fluid

SCOPE BOUNDARY

This is still fluid-side forced convection over a frozen steady velocity field. It is not a two-region CHT model, turbulence validation, production coolant qualification, or mesh-sensitivity study.

04 Dataset and Validation

The retained dataset contains every solver case that passed the corrected validation contract. The fresh-OOD pool contains 52 validated cases, so it is reported as directional stress-test evidence rather than a fully powered external benchmark.

CANDIDATE DESIGNS	VALIDATED CASES	REJECTED BY GATES
1,157 acquisition ledger	848 retained solver archives	227 convergence or physical validation

NOT EXECUTED	FRESH-OOD CASES	GRID RESOLUTION
82 not required after retained targets	52 directional stress-test pool	128 x 128 nearest-cell resampling

Retained cases and fixed-grid physical plausibility by pool.

POOL	RETAINED	PHYSICS PASSED	PHYSICS FAILED
id_train	446	446	0
id_calibration	100	100	0
id_test	100	100	0
ood_test	150	150	0
ood_fresh	52	52	0

PHYSICAL PLAUSIBILITY RESULT

Strict validation passed all retained fixed-grid cases: 848 checked, 0 failed. The enthalpy guardrail checked 848 cases and failed 0. Allowed temperature range was 291 K to 493 K.

WHY THIS MATTERS

The model result inherits credibility from the solver archive. Only validated solver outputs enter the learning dataset, and rejected cases remain traceable through convergence and final-status ledgers.

05 Field Operator: Temperature

The field model is a compact Fourier neural operator ensemble. It predicts full fixed-grid fields from geometry/mask/SDF/coordinate channels and broadcast operating parameters, with all metrics taken over the fluid region.

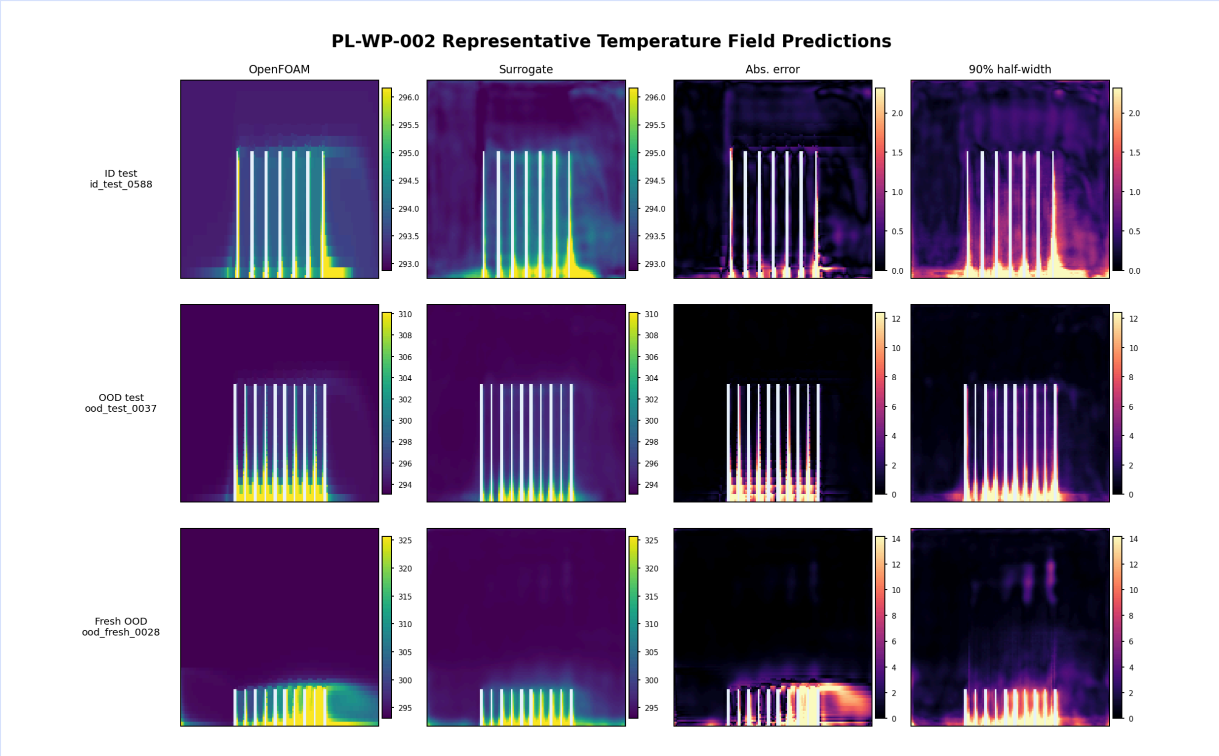


FIG. 1 Rows are representative ID-test, OOD-test, and fresh-OOD cases selected near median relative L2 error within each pool. Columns show OpenFOAM target, surrogate prediction, absolute error, and calibrated 90% half-width. Bright error and half-width bands identify where thermal-field mismatch and uncertainty concentrate, especially near fins and shifted regimes.

White fin pixels are solid geometry and are excluded from fluid-mask metrics. Figure source: evidence_pack_tier2_steady/05_figures/tier2_physical_operator/field_prediction_examples_temperature.png.

05 Field Operator: Velocity Magnitude

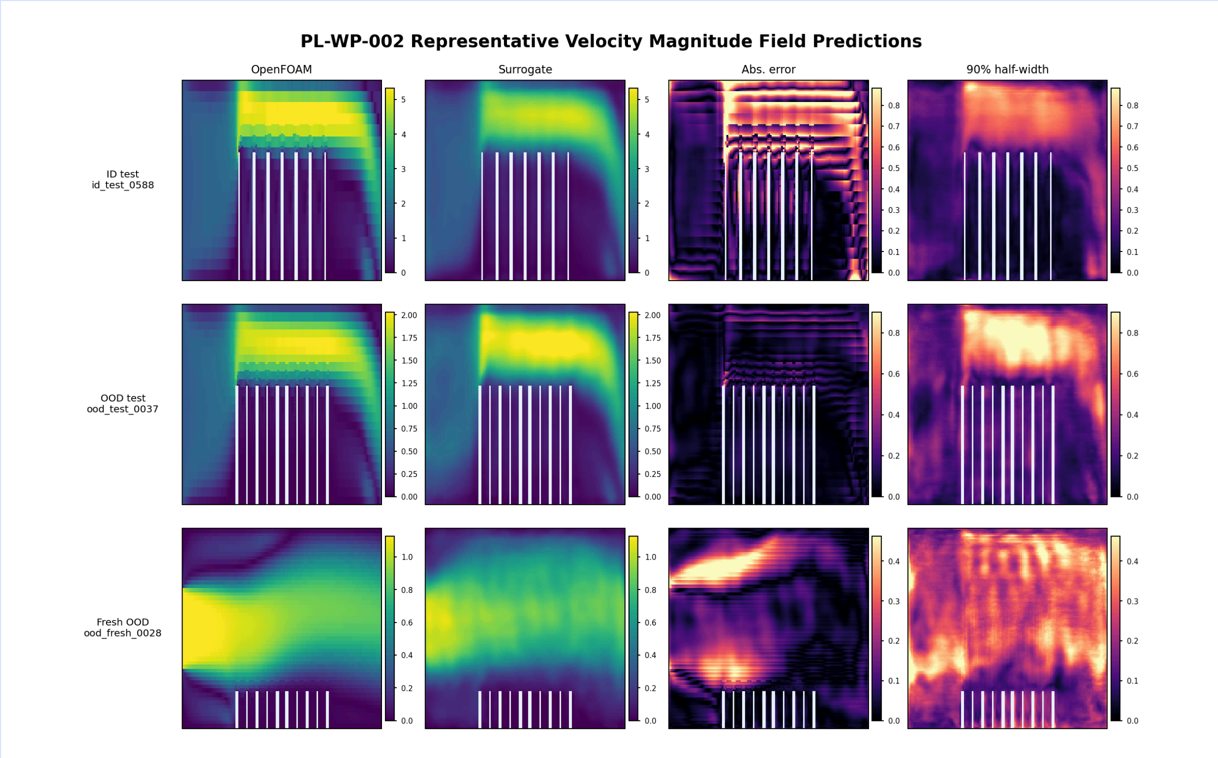


FIG. 2 Rows use the same representative cases as FIG. 1. Columns compare OpenFOAM velocity magnitude, surrogate prediction, absolute error, and calibrated 90% half-width. The panels show whether the model reproduces geometry-driven flow structure around fins; larger OOD residuals mark where the field operator smooths or misses local flow features.

05 Field Operator: Evaluation Summary

The held-out evaluation compares in-distribution cases against constructed and fresh regime shifts. The field surrogate remains usable inside the validated envelope, while error rises under OOD stress and uncertainty bands expand.

Evaluation summary from tier2_physical_operator/prediction_index.csv. Coverage in this table is the mean of per-case field coverage; aggregate ID coverage is reported separately in the calibration table.

POOL	CASES	MEAN REL L2	MEAN PER-CASE COVERAGE 90	MEAN BAND WIDTH
id_test	100	0.207	0.907	0.380
ood_test	150	0.458	0.949	2.458
ood_fresh	52	0.453	0.928	1.543

ARCHITECTURE

Small Fourier neural operator with local convolution skip paths; predicts full fixed-grid T and |U| fields from geometry/mask/SDF/parameter channels. Ensemble size 5; 688,346 parameters per member; 30 training epochs.

06 Uncertainty Calibration

The calibrated 90% band lands inside the ID acceptance interval and expands on OOD. The important limitation is that the band width does not rank OOD error strongly enough to claim a standalone learned OOD detector.

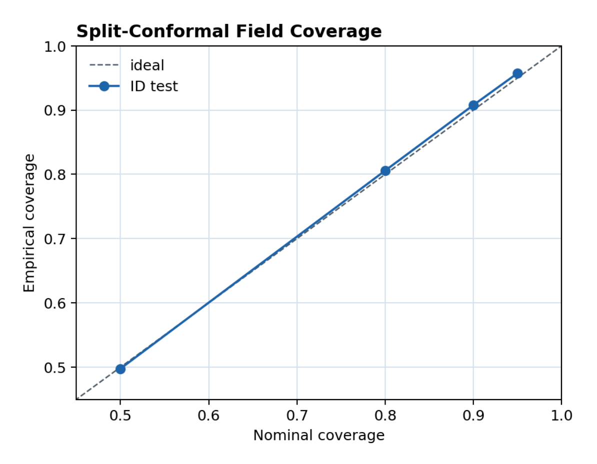


FIG. 3A Empirical field coverage versus requested nominal coverage on ID-test cases. The dashed diagonal is ideal calibration; blue points close to the diagonal mean the calibrated bands achieve their intended coverage inside the validated envelope.

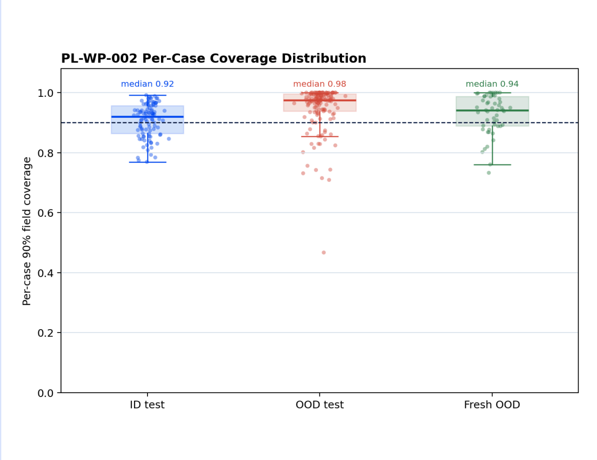


FIG. 3B Each dot is one evaluated case; boxes summarize per-case 90% field coverage by pool. The dashed line is the target 90% level. Medians remain above target, while lower-tail cases show where uncertainty coverage can still fail.

Primary 90% uncertainty result.

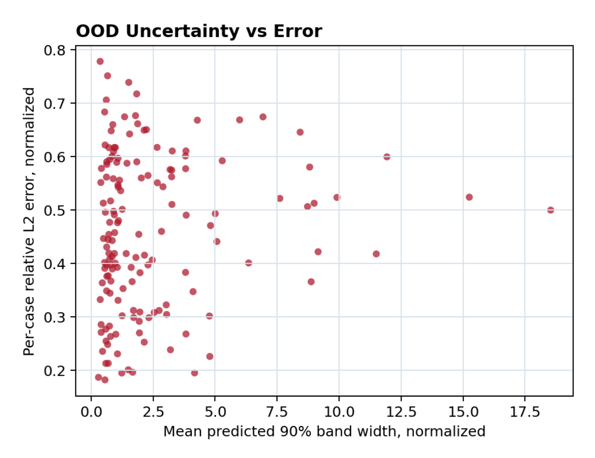
NOMINAL	ID COVERAGE	OOD COVERAGE	BAND INFLATION	OOD SPEARMAN
0.90	0.908	0.948	3.664x	0.123

INTERPRETATION

Coverage and band inflation are encouraging. Error ranking is weak: Spearman 0.123 is below the level needed for band width to serve as a standalone error-ranking detector, so learned uncertainty should be treated as telemetry rather than the primary deployment gate.

06 Regime Shift and Solver Fallback

The regime-shift analysis checks whether uncertainty expands when the field operator is moved outside the training envelope. The result is useful for triage, but the hard deployment control remains explicit solver fallback.



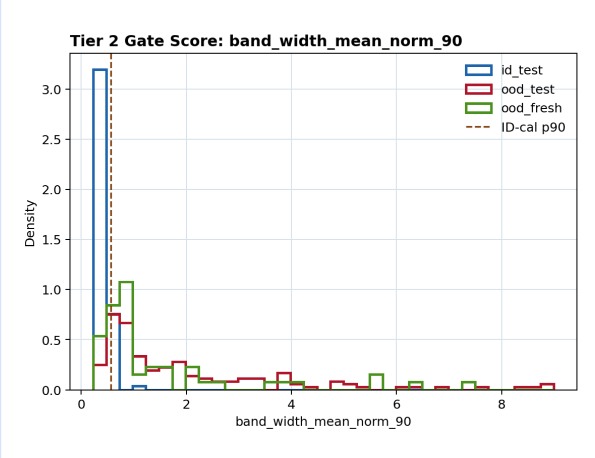


FIG. 4A Each point is an OOD case. The x-axis is mean normalized 90% band width and the y-axis is true per-case relative L2 error. The broad scatter means uncertainty expands under shift, but band width alone does not reliably rank the worst errors.

FIG. 4B Histogram of the learned band-width score for ID-test, OOD-test, and fresh-OOD cases. The dashed line is the ID-calibration 90th percentile threshold. OOD scores shift right, but overlap remains, so this is telemetry rather than a standalone gate.

06 Solver Fallback Policy

The qualification workflow is intentionally conservative. Known envelope violations route back to the solver; learned scores are reported as supporting telemetry rather than as the sole authority for serving a surrogate result.

Gate diagnostics across the retained evaluation pools.

POLICY	ID SERVED	OOD ROUTED	SERVED COVERAGE	ROLE
explicit envelope	1.000	1.000	0.908	primary baseline for known parameter shifts
band width score	0.900	0.866	0.903	learned uncertainty telemetry; useful but incomplete
risk score	0.930	0.822	0.892	experimental composite risk signal
nearest ID distance	0.850	1.000	0.910	parameter-distance signal, not latent-regime detection

Deployment baseline. Explicit envelope gating serves all ID-test cases and routes all constructed OOD cases, with served 90% coverage 0.908.

Research signal. Learned band-width routing has AUROC 0.953 for combined OOD, but routes only 86.6% of OOD at the ID-calibration p90 threshold.

06 Gate Diagnostics

The learned scores separate in-distribution cases from constructed regime shifts reasonably well, but that distinction is not the same as latent-regime detection. The nearest-ID distance score is especially strong here because the OOD definition is a parameter-envelope shift.

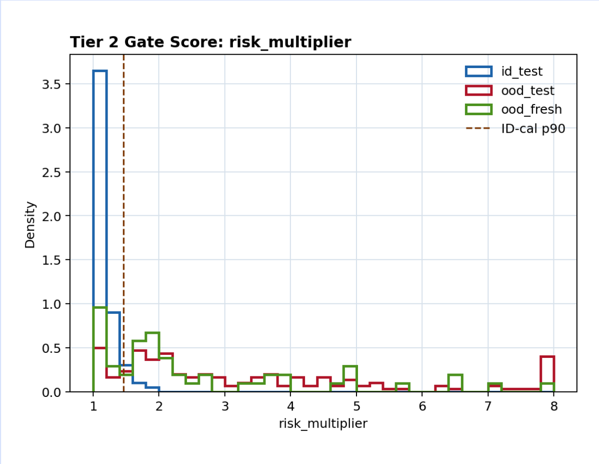


FIG. 5A Histogram of an experimental risk multiplier across ID-test and OOD pools. The dashed line is the ID-calibration 90th percentile threshold. Right-shifted OOD tails are useful warning signals, but distribution overlap limits direct deployment use.

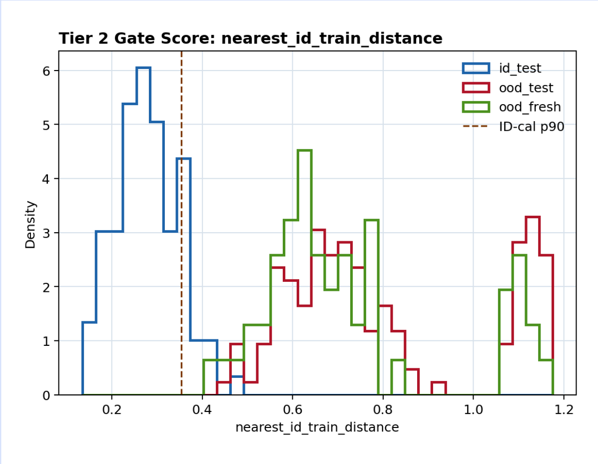


FIG. 5B Histogram of distance to the nearest ID training case in parameter space. OOD pools separate strongly because they were constructed by parameter displacement; this validates explicit envelope routing, not hidden-regime discovery.

Score AUROC where higher score means more OOD.

SCORE	COMBINED OOD AUROC	CORRECT READING
band width	0.953	useful learned uncertainty telemetry
risk score	0.916	experimental composite risk signal
nearest ID distance	1.000	parameter-envelope detector, not latent-regime detector

07 Engineering Adoption

Physics-AI adoption depends on proving when a surrogate is credible for a specific engineering decision, not only on producing fast predictions. PL-WP-002 addresses that adoption requirement by putting solver provenance, context-of-use boundaries, uncertainty evidence, and conservative solver fallback ahead of raw inference speed.

The value of this benchmark is not that it replaces a validated solver. The value is a reproducible qualification pattern that can sit between solver archives, field operators, uncertainty calibration, and engineering review rules.

Alignment between qualification needs and the PL-WP-002 evidence package.

ENGINEERING QUESTION	PL-WP-002 EVIDENCE	QUALIFICATION VALUE
Can the data be trusted?	OpenFOAM archives are retained only after convergence, thermal stability, enthalpy, extrema, and plausibility checks.	reviewable evidence before model training
Can use be bounded?	ID ranges, constructed OOD pools, fresh-OOD cases, and explicit envelope routing are reported separately.	controlled surrogate use inside a declared operating envelope
Can uncertainty be interpreted?	Split-conformal 90% coverage, per-case coverage, and OOD band inflation are measured at field level.	uncertainty becomes a reviewable deployment signal
Can the claim be reviewed?	Metrics, figures, prediction indices, run summaries, and gate diagnostics are linked to an evidence pack.	traceable technical due diligence

Client outcome. The workflow gives engineering teams a solver-validated way to decide when a physics-AI field operator can support bounded design decisions and when a case should return to the full solver.

Boundary. This is not a certified surrogate, not a homologation artifact, and not a production coolant model. It is a solver-validated evidence pattern that can be adapted to higher-consequence client programs.

08 Use Context and Boundaries

This workflow is most relevant to engineering teams already using solver data to train physics-AI surrogates and concerned about silent extrapolation. The demonstrated value is evidence-backed workflow control, not substitution for a validated solver.

Context-of-use interpretation for client programs.

CONTEXT OF USE	EVIDENCE FIT	DECISION RULE
Concept screening	Supported as a qualification pattern within a declared envelope.	serve surrogate only when solver gates and envelope checks pass
Design optimization	Appropriate as bounded design-loop acceleration evidence.	use surrogate for iteration, route boundary cases to solver
Release support	Not supported by this paper alone.	requires client-specific solver fidelity, mesh studies, thresholds, and sign-off
Certification support	Outside current scope.	requires independent review, sector evidence, and formal approval pathway

APPROPRIATE USE

- → Qualify a surrogate against solver provenance, physical plausibility, calibration, and fallback gates.
- → Build an evidence-pack workflow where every reported metric traces to solver archives and machine-readable tables.
- → Evaluate envelope routing plus learned uncertainty telemetry for engineering design loops.

OUT OF SCOPE

- → Production-grade thermal-fluid surrogate.
- → OpenFOAM replacement.
- → General latent-regime detector.
- → Full conjugate-heat-transfer or coolant-fidelity claim.

Before this pattern is used in higher-consequence workflows, it should be repeated across additional geometry families, solver fidelities, mesh studies, and residual/regime probes, with earlier failure-detection signals beyond explicit envelope checks alone.

09 Evidence Map

The whitepaper is backed by a machine-readable evidence pack. The primary validation artifacts are JSON, CSV, NPZ, model checkpoint, and PNG files under evidence_pack_tier2_steady.

Primary evidence sources used by this document.

ARTIFACT	PATH	ROLE
Config	configs/study_tier2_steady.toml	physics, dataset, solver, and UQ contract
Retained manifest	03_data/physical_acquisition_retained_summary.json	case counts, rejected cases, retained pool sizes
Fixed grid	04_predictions/physical_acquisition_fixed_grid_summary.json	128 x 128 extraction summary
Plausibility	02_metrics/physical_acquisition_physical_plausibility_summary.json	strict physical checks
Operator	02_metrics/tier2_physical_operator/run_summary.json	model and UQ metadata
Predictions	04_predictions/tier2_physical_operator/prediction_index.csv	per-case field metrics
Gate diagnostics	02_metrics/tier2_gate_diagnostics/tier2_gate_diagnostics.json	routing policies and AUROC

The paper reports the corrected PL-WP-002 validation campaign. The evidence files remain available for technical due diligence and reproducibility checks.

PHYLOOP

PHYLOOP RESEARCH

WHITEPAPER · NO. 002

DOCUMENT CONTROL

ARTIFACT

When Can a Physics-AI Surrogate Be Trusted?

EVIDENCE PACK

evidence_pack_tier2_steady

CLASSIFICATION

Technical whitepaper

LEGAL ENTITY

Culturiq Research Private Limited

This whitepaper reports a solver-validated surrogate qualification workflow and its supporting evidence. It does not claim production optimization, certified component performance, or deployable hidden-regime detection.