

FlagRelease: A Self-Verifying Agent Platform for Model Migration Across Heterogeneous Accelerators

Over 100 Models

10+ Accelerators

01 The migration challenge

Open-weight PyTorch models still require integration across the model, accelerator and software stack. A running service alone does not establish accuracy or a reproducible release.

Model × accelerator × stack

Each combination needs environment setup, component adaptation, evaluation and delivery checks. Real workloads expose gaps that isolated component tests can miss.

Migration and delivery

Prepare the environment
Enable the FlagOS stack
Launch and evaluate the model
Check accuracy against the baseline
Package and review the release

Validation under real workloads

Model-level evaluation checks whether the integrated stack meets acceptance criteria before the workflow advances.

02 Verifier-Gated Execution

Workflow engine

Ordered steps, bounded retries and checkpoints



Execution tools

Run setup, serving, evaluation and packaging



Result files

Actual outputs and evaluation evidence



Independent verifier

Read results and apply acceptance criteria



Commit verified state

Advance or retain the failure record



LLM advisor

Diagnose a failed step and propose a repair



Repair validation

Check schema, log evidence and allowed changes



Engine-controlled rerun

Retry within limits, then verify the new result

Only the verifier records PASS

Each stage is validated against predefined acceptance criteria before state transition.

Acceptance criteria remain outside the agent's edit scope.

03 Results and delivery

80%

Migration Success Rate

Across all supported accelerator platforms

Reproducible delivery

Versioned images, pinned configurations and evaluation evidence support reproduction and review.

Step records and checkpoints preserve progress for resume and rerun. Public release follows human review.

A bounded repair flow

Illustrative operator failure during evaluation

- 1 Capture logs and runtime configuration.
- 2 Diagnose the operator from log evidence.
- 3 Validate the schema and allowed changes.
- 4 Apply the patch and rerun within limits.
- 5 Verify results and retain unresolved failures.