NeurIPS 2026 · Main Track

Synthesis Through Simulation

Generating Coherent Enterprise Data via Scalable Agent-System Interaction

Yipeng Li* · Ashutosh Hathidara* · Jane Lo · Harshavardhan Abichandani · Gunraj Singh · Atin Ghosh

SAP Labs · *Equal contribution

THE PROBLEM

The enterprise data bottleneck

Tool-calling agents need realistic enterprise databases to train and evaluate on. Today's options fall short.

Real data is locked

- Restricted by privacy and compliance
- Schemas are proprietary

Statistical generators

- CTGAN, TabDDPM, GReaT
- Model rows independently
- Can't enforce business rules
- Need seed data to learn from

Procedural synthesis

- Keeps the structure
- Unrealistic distributions
- Hand-written for every domain

KEY IDEA

Let an agent do real work through the system's own APIs

LLM agent

Sees only the tool registry, never the schema

→

Policy gate Ψ

Checks each tool's preconditions on every write

→

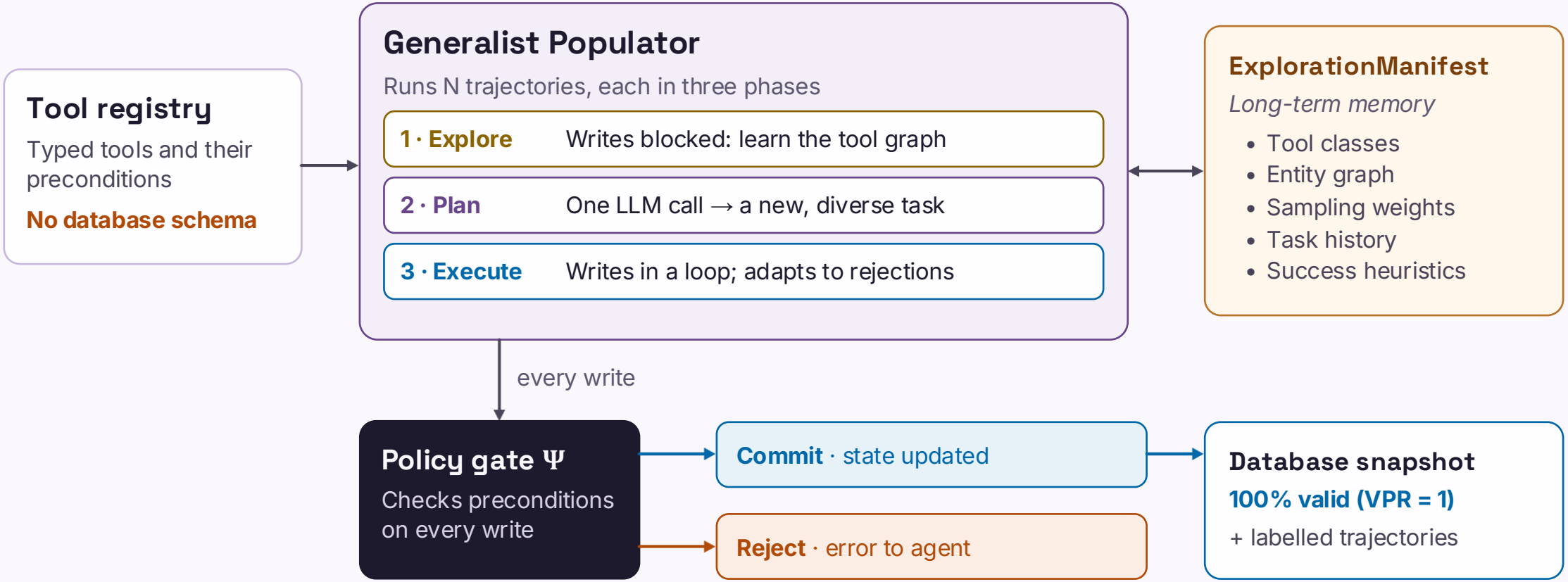
Database

The by-product: valid by construction

Proposition 1. Any trajectory ends with a validity pass rate (VPR) of exactly 1, so synthesis reduces to matching realistic distributions.

METHOD

How STS works



RESULTS

Results across ten environments

100%

validity (VPR) in all ten environments

0.88

average marginal fidelity (MDF)

No schema access · no per-domain authoring

Marginal fidelity (MDF) by environment



GP, K = 3, N = 100 trajectories each · 2–71 tables per environment · obfuscated API names (airline): 0.977

BASELINES

Comparison with baselines

7 of 10

environments have no seed data, so statistical synthesizers can't run there

≤ 0.923

validity of statistical synthesizers where they do run (STS: 1.0)

82%

of EnvScaler's airline runs write nothing, despite its schema access

Statistical synthesizers: CTGAN, TVAE, SDV-HMA, REaLTabFormer · EnvScaler: schema-privileged agent baseline

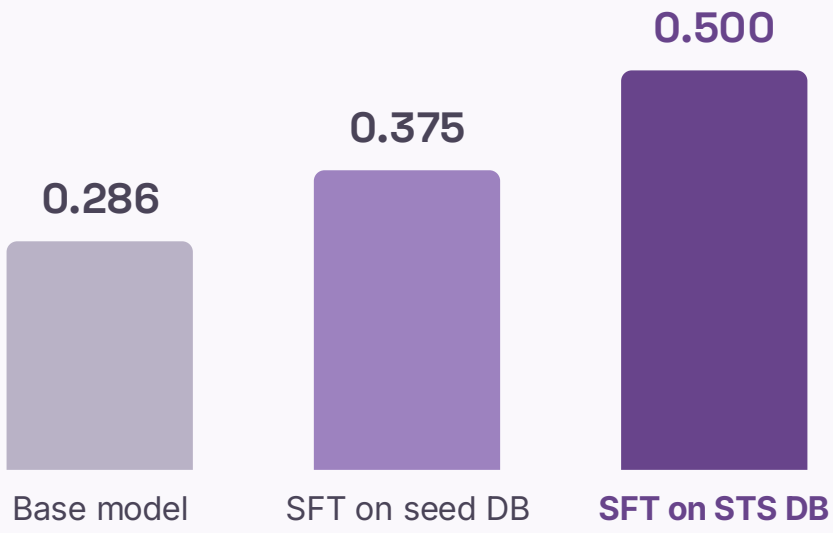
Without any schema access, STS **matches or exceeds** EnvScaler's fidelity: HR 0.939 vs. 0.939 · Cloud 0.850 vs. 0.828

DOWNSTREAM

Downstream agent training

Fine-tune Qwen2.5-7B on GPT-4.1 rollouts from each database · evaluate on τ^2-bench airline

pass@k



Metric	Base	Seed DB	STS DB
AUC	0.504	0.499	0.584
PPT	0.165	0.249	0.326
Max progress	0.632	0.611	0.695
Mean progress	0.309	0.309	0.323

STS data scores best on all five TED metrics, pass@k included

TAKEAWAY

Let the system's APIs enforce validity, and let the agent focus on realism.

Thank you

Paper arxiv.org/abs/2610.10549

Code github.com/SAP/synthesis-through-simulation

Code, environments and datasets are open source

Scan for the code