From Manuals to Worlds: Synthesizing Executable Enterprise Environments for LLM Agents

Harsh Vishwakarma*, K.N. Ajay Shastry*, Pranoy Panda*, Vartika Sengar*, Ankush Agarwal*, Suraj Nagaje*, Ganesh Senrayan*, Shrey Satapara*, Sanika Wagh, Pranav Bhagat, Mohit Meena, Ishan Jindal, Mahesh Chandran, Satoshi Nakashima, Hiromichi Kobashi · **Fujitsu Research** · *Primary contributors

Introduction

Enterprise agents must complete multi-step workflows, such as submitting and approving a leave request, that depend on app state, business rules, access policies and dependencies between operations.

This knowledge already lives in user manuals. The **Enterprise Emulation System (EES)** turns a manual into two outputs:

> **System Functional Graph (SFG)**: documented workflows linked to the APIs and data entities that realize them
>
> **Executable environment**: a runnable replica with a seeded database and APIs exposed as MCP tools

Motivation

Closed source

Business logic, roles and workflow constraints are hidden from agents.

Restricted access

Trial-and-error on live deployments is limited by access, safety and cost.

Generic worlds don’t transfer

Training Qwen3-8B on general-purpose synthetic environments: only 0.31 → 0.34.

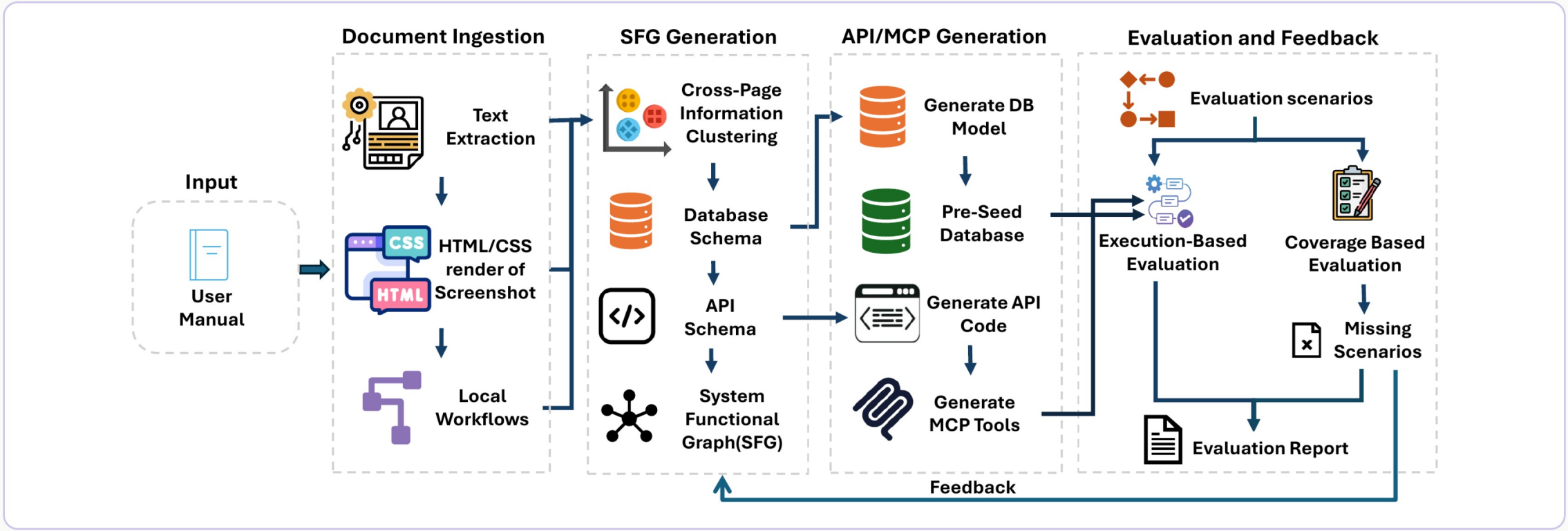
Raw manuals don’t help

Knowledge is spread across pages and screenshots; manual RAG adds at most +0.02.

Related work, gaps, and how EES addresses them

Related work	Gap	How EES addresses it
Tool learning: ToolBench, API-Bank, Toolformer	Assume the target tools and environment are already available	Recovers the app's own tool surface from its manual, exposed as MCP tools
Enterprise environments: CRMArena, EnterpriseBench, EnterpriseArena, EnterpriseOps-Gym	Built by hand around fixed apps; cannot reproduce a new, closed-source app	Automatically reconstructs the target app as an executable replica
Environment synthesis: EnvScaler, Agent World Model	Generic worlds miss app-specific business rules: Qwen3-8B gains only 0.31 → 0.34	Synthesizes the environment from the target app's own documentation
Retrieval over documents: RAG, agent memory	Knowledge is scattered across pages and screenshots: manual RAG adds ≤ +0.02	Structures workflows, APIs and entities into the SFG, which helps every model

Overview: from a user manual to an executable world



Document Ingestion

Text + screenshots rebuilt as HTML → page-grounded local workflows

SFG Construction

Merge workflows across pages; build DB and API schemas; map steps to APIs

Executable Env.

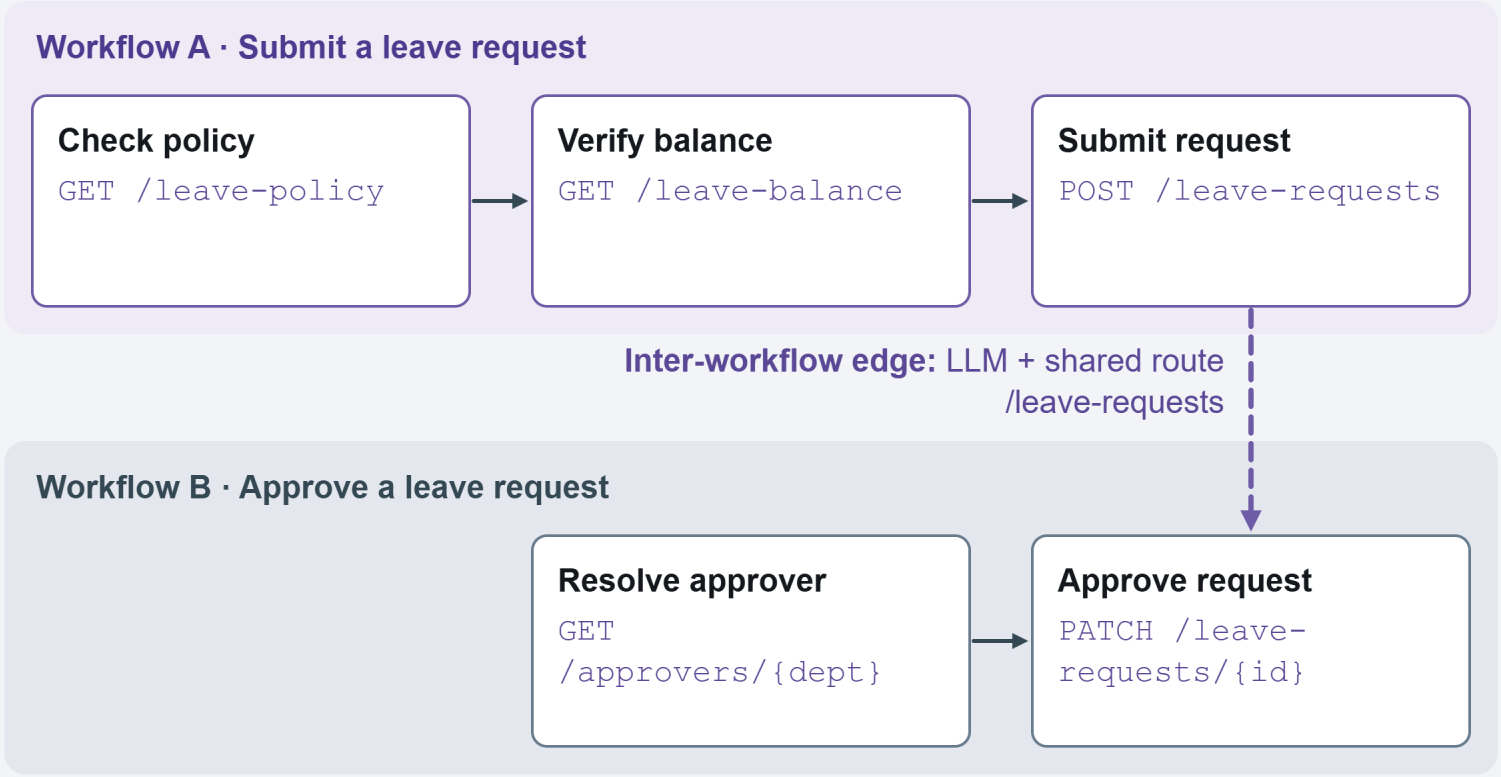
Seeded database + LLM-written API logic, tested and exposed as MCP tools

Evaluation & Feedback

Missing endpoints fed back to refine the SFG; integration tests run whole workflows

Inside the System Functional Graph (SFG)

Illustrative example: leave-request workflows in an HR app (step and route names are examples)



Node

A workflow step plus the API call that realizes it.

Intra-workflow edge

The documented order of steps within one workflow.

Inter-workflow edge

Links workflows on shared data: an LLM matches them, guided by API route structure.

What it captures

Rules spread across pages: policy and balance checks precede a request, which must exist before approval.

Results and analysis

Execution accuracy on EnterpriseArena (MCPEval Phase 2, LLM judge vs. ground truth)

Context	GPT-5.1	Qwen3-8B	Arctic-AWM	EB+GRPO	EES+GRPO
No context	0.61	0.31	0.34	0.37	0.39
+Manual RAG	0.61	0.30	0.36	0.36	0.38
+SFG RAG	0.67	0.34	0.40	0.42	0.47

Target-grounded training transfers better

No context: EES 0.39 vs. EnterpriseBench 0.37 vs. generic AWM 0.34 vs. base 0.31.

The manual helps only once structured

Same retriever: +SFG improves every model, while +Manual adds at most +0.02.

The two outputs are complementary

EES training + SFG context lift Qwen3-8B from 0.31 to 0.47 (+52% relative).

Fidelity of the emulation

88.9%

of 27 manually verified Frappe-HR API operations recovered (24)

97.8%

Frappe-HR functional coverage after two feedback rounds (from 88.5%)

66.7%

of the gain from training on the real apps (0.39 vs. 0.43)

Cost: ≈ $114 and 5.5–10.5 hours per 50-page manual

Conclusion

EES turns user manuals into a System Functional Graph and an executable emulation. Agents learn from both: Qwen3-8B rises from **0.31 to 0.47** on EnterpriseArena.

Takeaways

- Structure beats raw text
- Training on the target app's emulation transfers
- SFG context and EES training are complementary
- 88.9% of documented Frappe-HR APIs recovered

Limitations

- Relies on a strong LLM backbone
- Captures only what is documented
- API names and arguments may differ from the real system
- No human-in-the-loop refinement yet

Future work

Self-evolution: agents continuously learn from interaction and feedback inside the emulated environment, to improve on the actual system.