YASH SHRIVAS

AI Engineer · LLM agents, retrieval and evaluation · ML infrastructure

Jabalpur, India • yashh.real@gmail.com • github.com/Ya-shh • linkedin.com/in/yash-shrivas-b343282ba • yash-work.com

SUMMARY

AI engineer specializing in GenAI, LLM and agentic systems, with a habit of measuring what I ship. Built **Ravenhold**, a three-agent incident-response system on a FalkorDB graph with an MCP tool layer and a ground-truth evaluation harness (9/9 root causes, 100% valid citations). **Google Summer of Code 2026** contributor at Kubeflow: shipped multi-notebook composition to Kale (KEP-0812), 10 merged PRs across Kubeflow and Kornia. Built the retrieval-augmented support agent **Qdrant** hosted on its official Discord server in 2024 and awarded a bounty. ONNX weights in three **Hugging Face** embedding models with 33M+ downloads a month combined.

AI SYSTEMS

Ravenhold · Multi-agent incident response on a knowledge graph — *Oct 2026*

FalkorDB × WeMakeDevs Graph Hacks hackathon · ravenhold.onrender.com · github.com/Ya-shh/ravenhold · yash-work.com/blog/ravenhold

- Designed a Commander / Investigator / Historian agent team (Claude Sonnet 5.5 and Haiku 5.5) whose context, memory and coordination live in one graph: Task nodes for hand-offs, Episode nodes for audit, Lessons from corrections, and versioned runbooks that agents propose and humans approve.
- Exposed 23 typed graph tools through an **MCP server the agents themselves consume**; root cause found by a greedy cover over the dependency graph, with PageRank, betweenness and community detection as tools rather than model opinion.
- Built the **evaluation harness**: every answer scored against ground truth computed from the graph, never the model's claims. 9/9 root cause, 9/9 right person paged, 9/9 noise alert flagged; 100% of cited IDs real; every tool under 1 s on a 20,000-service graph; 1,600 graph operations in 3.0 s across 100 concurrent sessions; 32 deterministic tests; about $0.03 and 26 s per investigation.

Qdrant Discord support agent · Retrieval-augmented, human-in-the-loop — *2024*

Hosted by Qdrant on its official Discord server · official bounty (qdrant/qa-automation#1) · github.com/Ya-shh/Discord-AI-Agent

- **Hybrid retrieval** over the Qdrant docs: code-boundary chunking, dense snowflake-arctic-embed-l (1024-d) plus sparse SPLADE vectors via FastEmbed, indexed through Haystack and merged at query time so exact API names match alongside semantic similarity.
- Content-routed prompts: code-focused when retrieved passages contain code, explanation-focused otherwise.
- **Human in the loop** before any answer reached the community: drafts land in a private moderator forum to send, or enrich and regenerate; follow-ups from one user are queued and answered together.

OPEN SOURCE

Google Summer of Code 2026 · Kubeflow (Kale) — *May – Aug 2026*

Selected contributor, "Composable Kale Notebooks with a Visual Pipeline Editor" · project page

Mentor: **Eder Ignatowicz**, Senior Principal Software Engineer Architect, Red Hat · recommendation letter: yash-work.com/#recommendation

- Built and shipped **multi-notebook composition** (KEP-0812): a notebook: cell type that lets several notebooks compile into a single Kubeflow pipeline. 28 files, merged after review by Kale's original author and maintainers (#867).
- Recursive compilation in NotebookProcessor: each notebook becomes a separately compilable DSL module plus a nested KFP v2 GraphComponent, so every cell stays independently cacheable and schedulable.
- Automatic data-flow inference across notebook boundaries with AST analysis and PyFlakes: execution order and artifact passing derived from shared variable names, with zero dependency annotations.
- Extended the JupyterLab extension (TypeScript/React) with the Notebook cell type, path input and inline reference chips; six further merged Kale PRs across backend and frontend (#571, #646, #645, #608, #601, #593), three in review (#959, #960, #961).

Kornia · Computer vision library — *Feb – Mar 2026*

- Centralized image validation into a reusable KORNIA_CHECK_IMAGE_LAYOUT utility, cutting duplication (#3549).
- Hardened augmentation: removed redundant RNG instantiation at probability 0 or 1, strict 2D/3D crop-size validation in crop_and_resize and center_crop, fixed silent failures in parameter generators (#3605).
- Header-based PNG format detection and a shared _convert_image_type helper with RGBA/Grayscale conversion paths (#3608).

Hugging Face · ML model contributor — *Apr – Sep 2024*

- ONNX weights merged into BAAI/bge-m3 (commit), BAAI/bge-base-en (commit) and intfloat/multilingual-e5-large-instruct (commit) (33M+ downloads a month combined), enabling ONNX Runtime inference.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript
LLM systems: Claude API, agent orchestration, MCP (server and client), tool design, prompt routing, human-in-the-loop, memory
Evaluation: ground-truth eval harnesses, citation and grounding checks, concurrency and scale benchmarks, cost and latency tracking
Retrieval: Qdrant, Haystack, FastEmbed, SPLADE, snowflake-arctic-embed, hybrid dense + sparse, FalkorDB (Cypher, vector, full-text)
ML infrastructure: Kubeflow Pipelines v2, Kubernetes, Docker, ONNX Runtime, PyTorch, HF Transformers, sentence-transformers
Program analysis & frontend: Python ast, PyFlakes, Jinja2, networkx; React, TypeScript, JupyterLab extension APIs
DevOps & data: GitHub Actions, CI/CD with test and lint gates, pre-commit, pandas, NumPy

EDUCATION

Bachelor of Technology · Jabalpur Engineering College, India — *2022 – 2026*