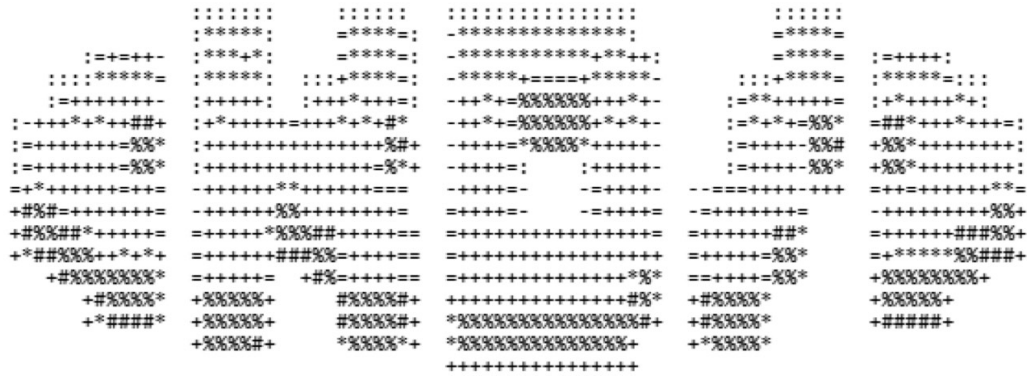




KD × Hugging Face × Gradio: A Hybrid Pipeline for Community Knowledge and Agentic Learning

Community Proposal — CrackedDevs Ecosystem
Experimental Technical Note • September 2026

Abstract

We present a practical first-step pipeline that converts the existing CrackedDevs journal (kd-penzine.pages.dev) into structured Hugging Face datasets and a Gradio space. The resulting assets enable members to load, retrieve, and integrate community knowledge into RAG systems, MCP servers, and agent frameworks. Access is gated with manual review to control API traffic. The work is released as an unofficial experimental effort under CC-BY-4.0 and is intended as infrastructure that Malaysian builders can own, extend, and use for skill development in retrieval-augmented generation, vector search, and agentic AI — without competing with frontier model providers.

1. Introduction and Motivation

CrackedDevs is a self-sustaining Malaysian builder community focused on practical AI-assisted development, bounties, and skill growth. The community already produces a monthly journal (KD Penzine) that captures builds, essays, ecosystem news, and bounty outcomes. At present this content remains largely static. Converting it into machine-readable datasets unlocks a new layer of utility: members can treat the journal as a shared knowledge base for retrieval-augmented generation (RAG), tool-using agents, and custom Model Context Protocol (MCP) servers.

The proposal deliberately stays modest. It does not claim to deliver a full autonomous-learning system. Instead it supplies the missing data layer and simple integration patterns so that members can begin experimenting immediately and later compose more sophisticated agentic workflows on top of a common foundation.

2. Current Assets

The following resources are already live on the Hugging Face Hub:

- **Datasets** — [ctaxnagomi/kd-penzine-volumes](#) (volume index) and [ctaxnagomi/kd-penzine-pages](#) (page-level markdown, 89 rows).
- **Gradio Space** — [ctaxnagomi/kd-penzine](#).

2.1 Basic Loading

```
from datasets import load_dataset
volumes = load_dataset("ctaxnagomi/kd-penzine-volumes")["train"]
pages = load_dataset("ctaxnagomi/kd-penzine-pages")["train"]
```

3. Access Control

Gated user access is **enabled** with **Manual review**. Users must share contact information and accept any stated terms before downloading. The manual gate exists to protect the datasets from uncontrolled API traffic: a misconfigured member pipeline could otherwise generate high request volume. Community contributions (Discussions and Pull Requests) remain enabled on the Space so that members can still propose improvements. The datasets contain only public journal content compiled for member use. No private or sensitive material is included. The effort is explicitly marked unofficial and experimental and is released under CC-BY-4.0.





4. Integration Patterns

4.1 MCP Server Tool

A minimal tool that can be registered inside a FastMCP server:

```
from mcp.server.fastmcp import FastMCP
from datasets import load_dataset

mcp = FastMCP("kd-penzine")
pages = load_dataset("ctaxnagomi/kd-penzine-pages")["train"]

@mcp.tool()
def search_kd_penzine(query: str, limit: int = 5) -> list[dict]:
    """Search KD Penzine pages by title or content."""
    results = []
    q = query.lower()
    for row in pages:
        if q in row["title"].lower() or q in row["markdown"].lower():
            results.append({
                "volume": row["volume"], "title": row["title"],
                "date": row["date"], "page_index": row["page_index"],
                "snippet": row["markdown"][:300] + "..."
            })
        if len(results) >= limit:
            break
    return results
```

4.2 Agent Framework Tool (LangChain-style)

The same retrieval logic can be exposed as a tool inside common agent frameworks:

```
from langchain.tools import tool
from datasets import load_dataset

pages = load_dataset("ctaxnagomi/kd-penzine-pages")["train"]

@tool
def search_kd_penzine(query: str, limit: int = 5) -> str:
    """Search the KD Penzine journal for relevant pages."""
    results = []
    q = query.lower()
    for row in pages:
        if q in row["title"].lower() or q in row["markdown"].lower():
            results.append(
                f"Vol {row['volume']} | {row['title']} ({row['date']})\n"
                f"{row['markdown'][:400]}..."
            )
        if len(results) >= limit:
            break
    return "\n\n".join(results) if results else "No matching pages found."

# agent = create_react_agent(llm, tools=[search_kd_penzine], ...)
```





5. Benefits for the Community and Malaysian AI Space

- **Local knowledge becomes usable infrastructure.** Static journal pages become structured data that any Malaysian builder can load, search, and extend. This lowers the barrier to building RAG tools, chatbots, and internal knowledge systems.
- **Practical exposure to the modern AI stack.** Members gain hands-on experience with Hugging Face Datasets and Gradio — the same tools used by the global open-source community — rather than treating AI as an abstract or purely commercial service.
- **Agentic AI as an indirect learning path.** Clean, versioned community datasets let beginners experiment with the basic agent loop (load → retrieve → generate → act) without heavy compute. The exercise demonstrates how machine learning (datasets + retrieval) and agentic reasoning coexist in practice.
- **Normalising good practices around private data.** Because the current datasets are public and gated, members can first master the full pipeline on safe content. This establishes habits of access control and careful data handling before they move to pipelines that process personal or private information.
- **Honing RAG, vector search, and agentic skills.** The datasets serve as ready training material for RAG pipelines, FAISS (or other vector stores), and small agent systems — skills that transfer directly to personal projects and commercial work.
- **Reusable pipeline for other projects.** The same data structure can be integrated into bounty submissions and other community projects, giving members a pre-configured retrieval layer instead of forcing each project to reinvent data access.
- **Building local AI culture.** Most AI activity in Malaysia still centres on consuming overseas models. Community-owned datasets and spaces create a modest but real foundation that members can point to, improve, and build upon.

6. Longer-term Vision: A KrackedDevs Harness

Over time, submissions, discussions, and experiments can be accumulated into a living community artefact — conceptually a single `krackeddevs.md` (or equivalent structured collection). This artefact would contain shared tooling, MCP servers, agent frameworks, skills, and a starter pack for a **KrackedDevs Harness**. The Harness is not intended to compete with Claude, DeepSeek, or any frontier provider. It is simply enough for the community to start: a practical, locally owned foundation that members can keep extending together. The present pipeline is a concrete first step toward that direction.

7. Conclusion

The KD × Hugging Face × Gradio pipeline converts an existing community journal into loadable infrastructure. It supplies gated datasets, a Gradio space, and ready-to-adapt snippets for MCP servers and agent frameworks. The immediate value is practical: members can begin building retrieval and agentic systems on real local content while learning responsible data practices. The longer value is cultural: it gives Malaysian builders a small piece of AI infrastructure they can own and grow. Further expansion (additional datasets, automatic approval policies, richer MCP tooling) can be driven by observed usage and community feedback.

This document is an unofficial experimental community proposal. All datasets and the associated Gradio space remain under CC-BY-4.0 and are not official KrackedDevs products.

Feedback and contributions are welcome via the Hugging Face Space discussion channels.

