SYR-RAG

A grounded retrieval pipeline for Syracuse research

From prototype to grounded RAG pipeline

LAST YEAR

Vision and prototype

- Distill SU papers into a chatbot
- Fine-tuned summarisation and chat models
- Metadata only — no vector store

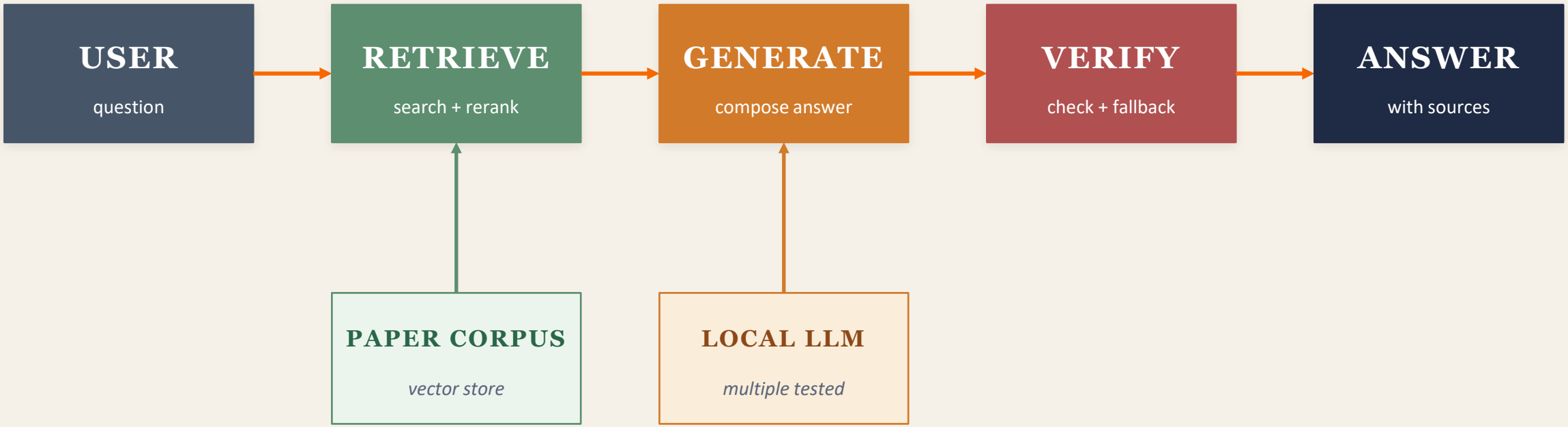
THIS YEAR

Retrieval-first and grounded

- Seven-stage pipeline per turn
- Vector retrieval with reranking
- Follow-ups stay on topic
- Hallucination checks before answers ship

ARCHITECTURE

How a question becomes a grounded answer



PIPELINE

Per-turn execution: seven stages

Meta-commands and weak-evidence cases short-circuit the pipeline.

Loading state, reading intent, finding evidence

1 Entry

Load and gate

- Restore conversation state
- Handle resets and topic switches
- Answer simple meta-queries directly

2 Intent

Read the question

- Classify the question type
- Detect follow-ups and pronouns
- Add prior context only when stable

3 Retrieval

Find evidence

- Vector search across the corpus
- Deduplicate and filter
- Rerank by relevance and name match

STAGE 4

Confidence drives the prompt budget

4 Five confidence levels

How much we trust the evidence

- high → keep context
- medium → keep context
- low → narrow the prompt
- weak → stricter guard
- inconsistent → prefer fallback

Prompt budget

More evidence when confident, less when not

WHEN CONFIDENT

16 docs × 600 chars / doc

WHEN UNCERTAIN

10 docs × 450 chars / doc

GROUNDING

Verifying every answer before it ships

CHECK 1

Citation scrub

every cited title

must match

Lines that reference papers not in the retrieved set are dropped from the answer.

CHECK 2

Name check

every named person

verified both ways

People in the answer are matched against retrieved metadata in both directions.

BACKSTOP

Extractive answer

if a check fails

rebuilt from docs

When verification fails, the answer is rebuilt directly from the retrieved set.

DATA

Three corpora, three tradeoffs

MODE A

full

Most detail

- Full PDFs ingested
- Best evidence per query
- Slowest to build

MODE B

openalex

Broadest coverage

- Open metadata source
- Largest researcher set
- Lighter on full text

MODE C

abstracts

Lightest footprint

- Abstracts only
- Fast to build and query
- Surface-level answers

Tested with multiple local LLMs. Build challenges: publisher access, format drift, coverage vs depth tradeoffs.

Thank you

Questions welcome.