

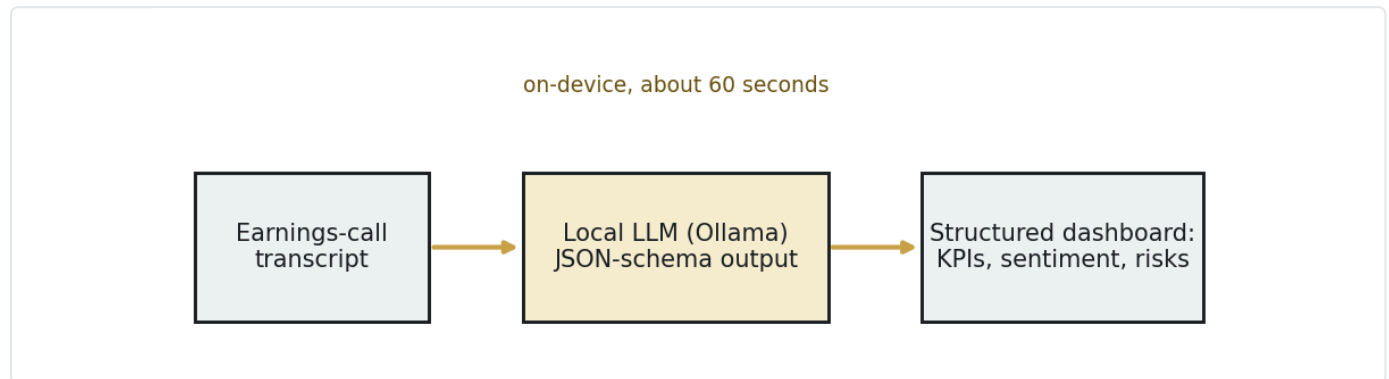
Cadence: Earnings Calls into Structured Intelligence

Personal project · Applied AI

Solo build

A local-first AI tool that turns an unstructured earnings call into a structured, queryable dashboard in about a minute.

Cadence reads an earnings-call transcript and returns a structured dashboard, running entirely on-device. It uses a local large language model with schema-constrained output, so the analysis stays private and costs nothing per run. I built it end to end in Python.



The on-device pipeline: transcript through a local LLM with schema-constrained output to a structured dashboard.

How it works

A local LLM through Ollama runs the transcript with JSON-schema-constrained decoding to pull themes, risks, and supporting evidence, which a Streamlit dashboard renders. Because the model runs locally, the transcript never leaves the machine.

Why it matters

Earnings calls are dense and slow to read at scale, and cloud models raise privacy and cost questions for market-sensitive text. A local, schema-constrained pipeline answers both, with full control of the prompt and the output shape.

SELECTED REFERENCES

- Ollama, local large language model runtime (ollama.com).
- B. Willard and R. Louf, "Efficient Guided Generation for Large Language Models" (constrained decoding), 2023.

Engineering portfolio brief. Course and team project; contribution as noted above.