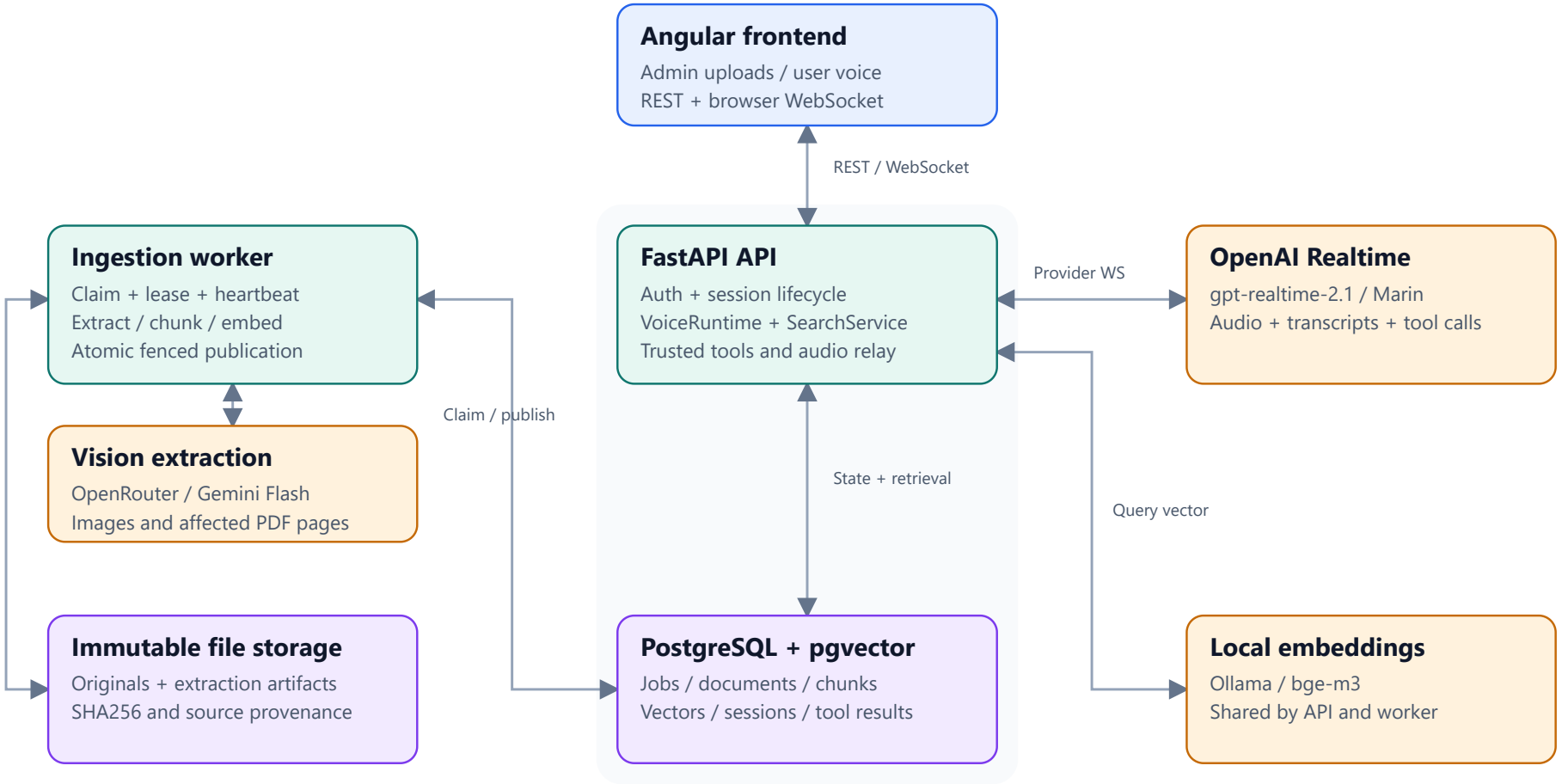
System architecture

One API, one ingestion worker, shared knowledge storage

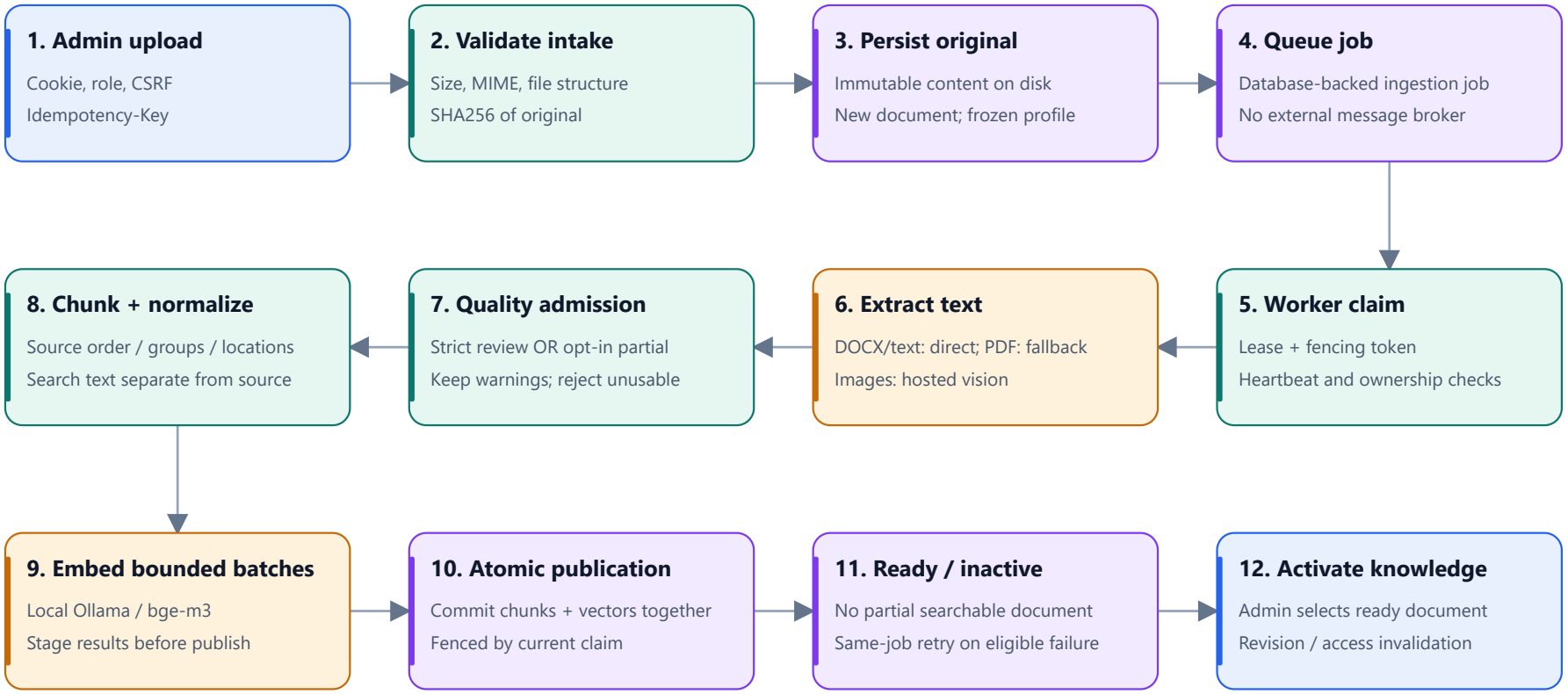


Voice: browser → API → OpenAI. Ingestion: worker → extraction → embeddings → atomic publication.

Principal connections only. Detailed job, evidence, activation and interruption flows are shown in the other three diagrams.

Document ingestion pipeline

Independent uploads, bounded extraction and atomic publication

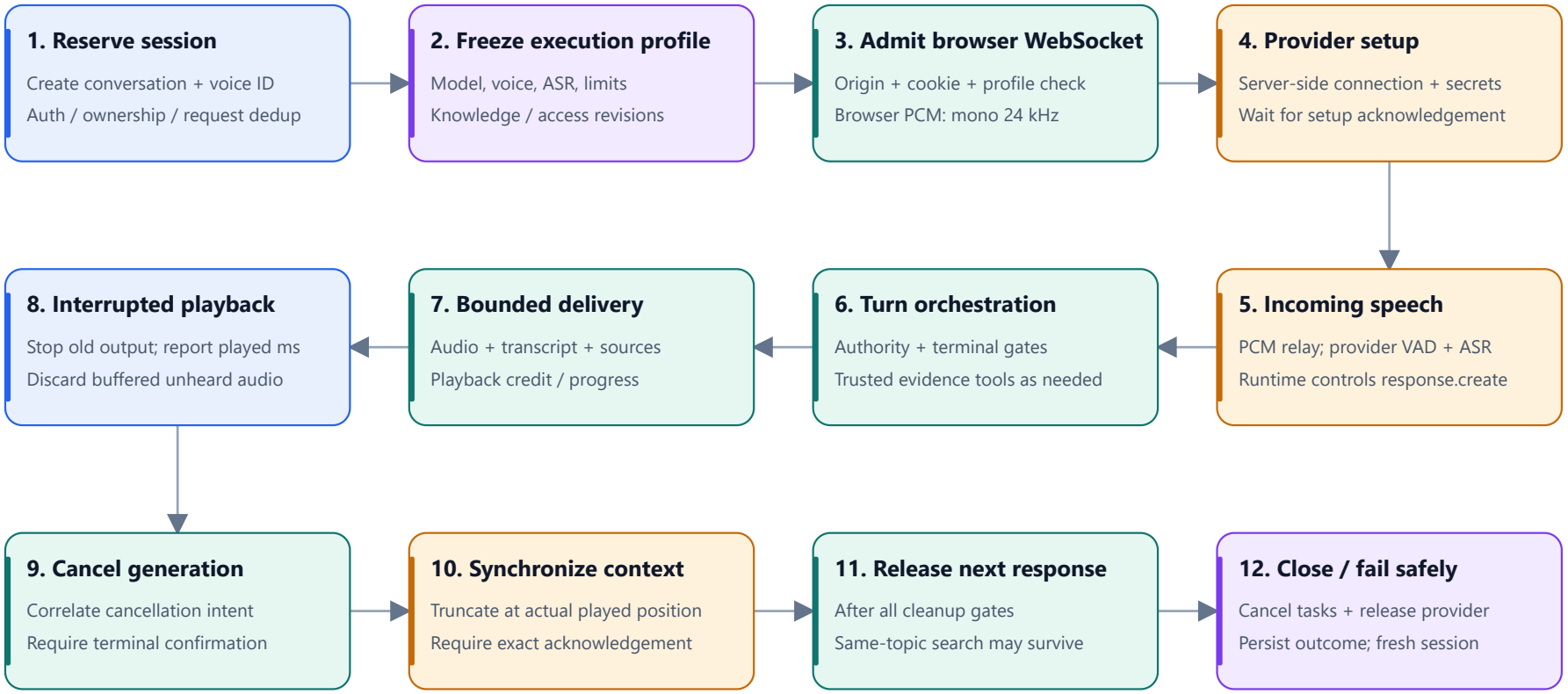


Review is conditional. Local partial-admission mode bypasses manual approval; malformed data, identity and bounds still fail.

Frontend / admin · Backend logic · Persistent state · AI / embedding provider

Live voice session and streaming

Control lifecycle plus continuous bidirectional media

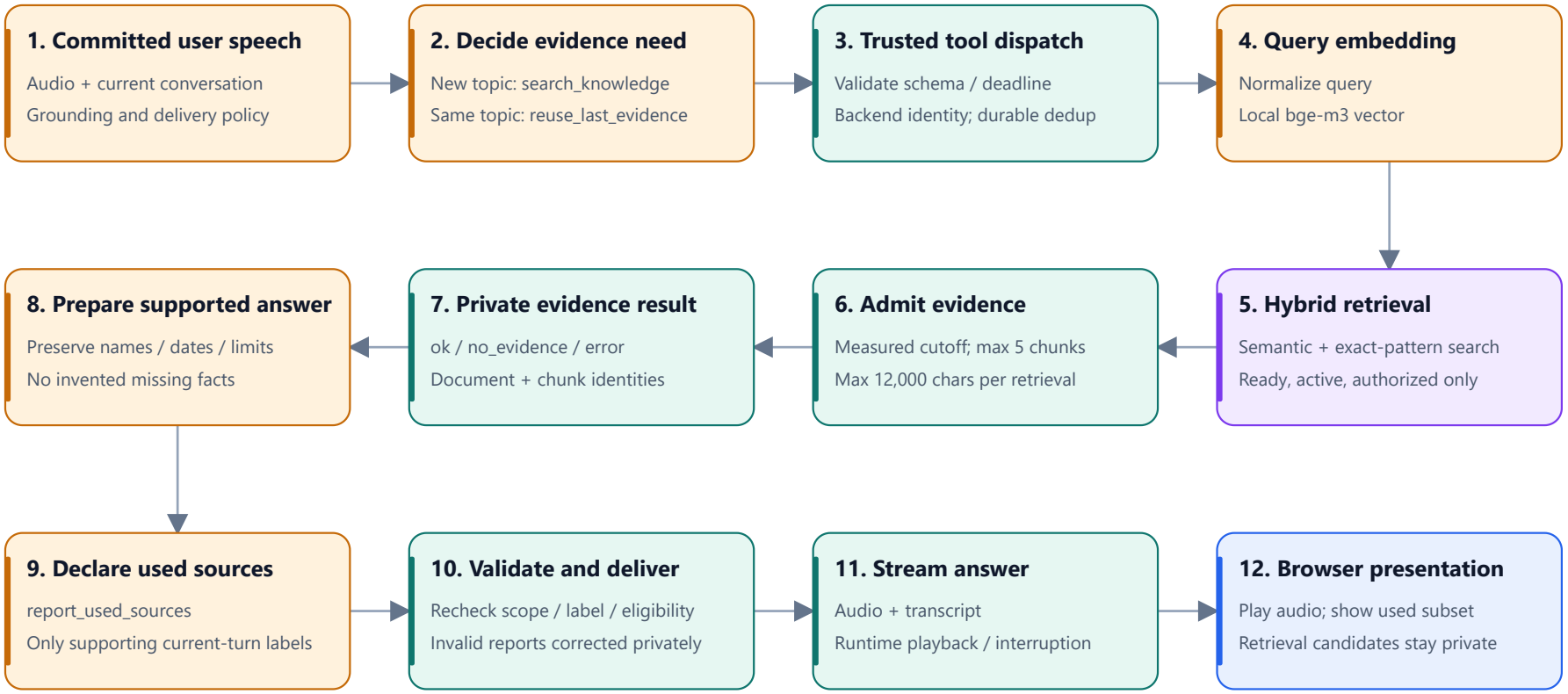


This is a lifecycle summary, not serial audio processing. See the sequence diagram for asynchronous events and acknowledgements.

Frontend / admin · Backend logic · Persistent state · AI / embedding provider

AI and grounded retrieval pipeline

The model chooses tools; the backend owns identity and evidence



Retrieval branches: 50 semantic + up to 50 exact-pattern candidates, deduplicated. Greetings need no retrieval; failures never authorize fallback facts.

- Frontend / admin
- Backend logic
- Persistent state
- AI / embedding provider