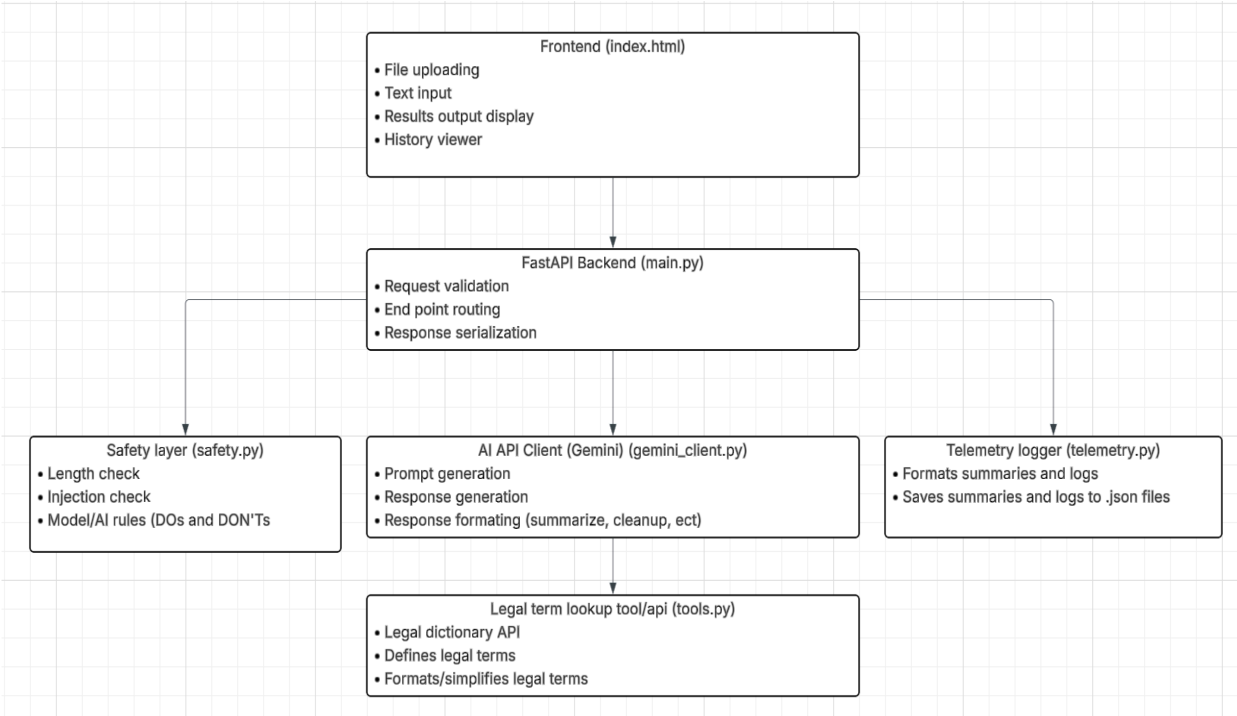
Architecture Diagram:



Guardrails

Input Validation:
- Length: 50-50,000 chars (prevents trivial inputs & DoS)
- Prompt injection: Common attack patterns blocked ("ignore previous instructions", "you are now", etc.)

API Safety:
- Content filters: BLOCK_NONE (legal terms like "breach" trigger false positives)
- Fallback: Generic summary if API blocks (finish_reason: 12)

System Prompt:
- DO: Summarize, identify terms, highlight risks, etc
- DON'T: Provide legal advice, make decisions for user, etc

Evaluation

Test Suite: 16 cases (13 valid contracts, 2 injection attacks, 1 edge case)

Metrics:
- Pattern matching: ≥50% expected terms found
- Tool usage: Correct legal term lookup triggering
- Tokens, latency, success rate

Typical Pass Rate: 93.8% (15/16) >80% requirement

Known Limitations

1. API False Positives: some requests blocked by Gemini safety filter (e.g., "copyright" triggers block).

2. Term Detection: AI-driven (flexible) but relies on Gemini's knowledge. Covers most of the common legal terms.

3. **Rate Limiting:** 10 req/min (free version). The test suite uses 7s delays.

4. **File Support:** TXT (full), PDF (text only). No DOCX, or images.

5. **Language:** English only (Dictionary API constraint).

6. **Max Input:** 500k chars. Longer documents must be split.