From Unstructured Clinical Notes to Actionable Care Plans (ACPs)

Dávid Márton, PhD 11/2025

ACPs Challenge

From Doctor’s notes to actionable care plan

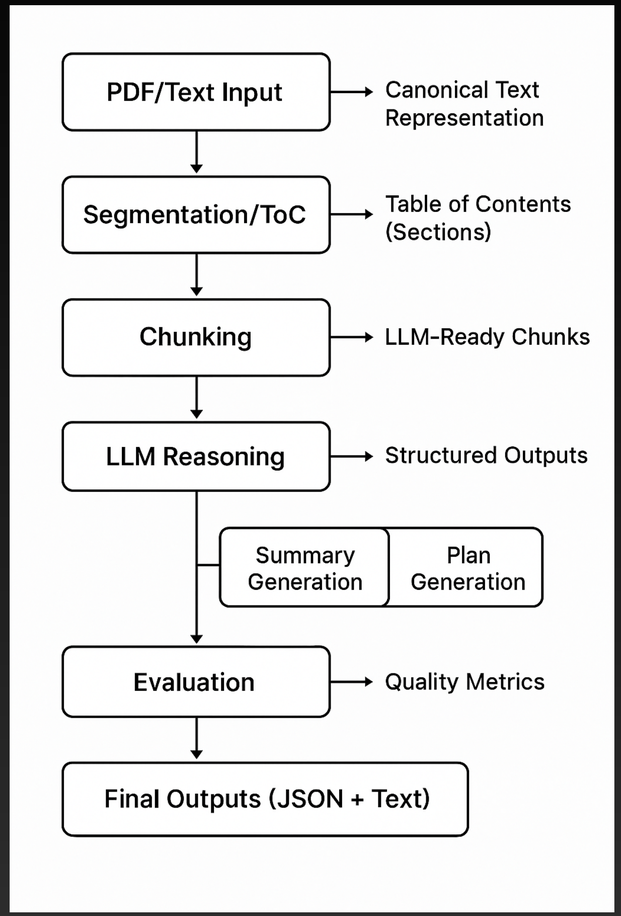
- Unstructured & variable format
  - No standardised format
  - Variable section headers/organisation
  - Information scattered across multiple sections
- Information extraction complexity
  - Distinguish observations from recommendations
  - Extract structured data from narrative text
  - Implicit vs explicit information
- Accuracy & traceability
  - Medical decisions require high accuracy
  - Claims need to be traceable
  - Cannot fabricate/infer missing info
- Privacy & compliance
  - Must protect PHI
  - Cannot hallucinate patient identifiers
  - Local processing preferred for HIPAA
- Actionability
  - Recommendations must be specific/prioritised
  - Need rationale
  - Need to provide certainty/identify gaps in information

🧠ACPs Data

- MTSamples Dataset
  - Size: ~5000 clinical note PDFs
  - Specialties: ~40 medical specialties (Allergy/Immunology, Cardiovascular, Surgery, Radiology, etc)
  - Note Length:
    - ~3400 characters on average
    - Range: 1300 - 6000 characters
    - ~9 sections per note on average
  - Documents structure
    - Overview section (with medical specialty, sample name, description)
    - Variable clinical sections (ASSESSMENT, MEDICATIONS, LABS, etc)
- De-Identification
  - Dataset is de-identified (public research dataset)
  - Additional protections
    - LLM instructions to not add/infer PHI
    - No hallucination of patient identifiers
    - Local processing for privacy
- Key Observations
  - PDFs with text only
  - Varying section names and content but standardised structure
  - Relatively short PDF length: 1-3 pages on average based on a randomised sample (n=20)
  - TOC/Summary/Plan only requires single note context

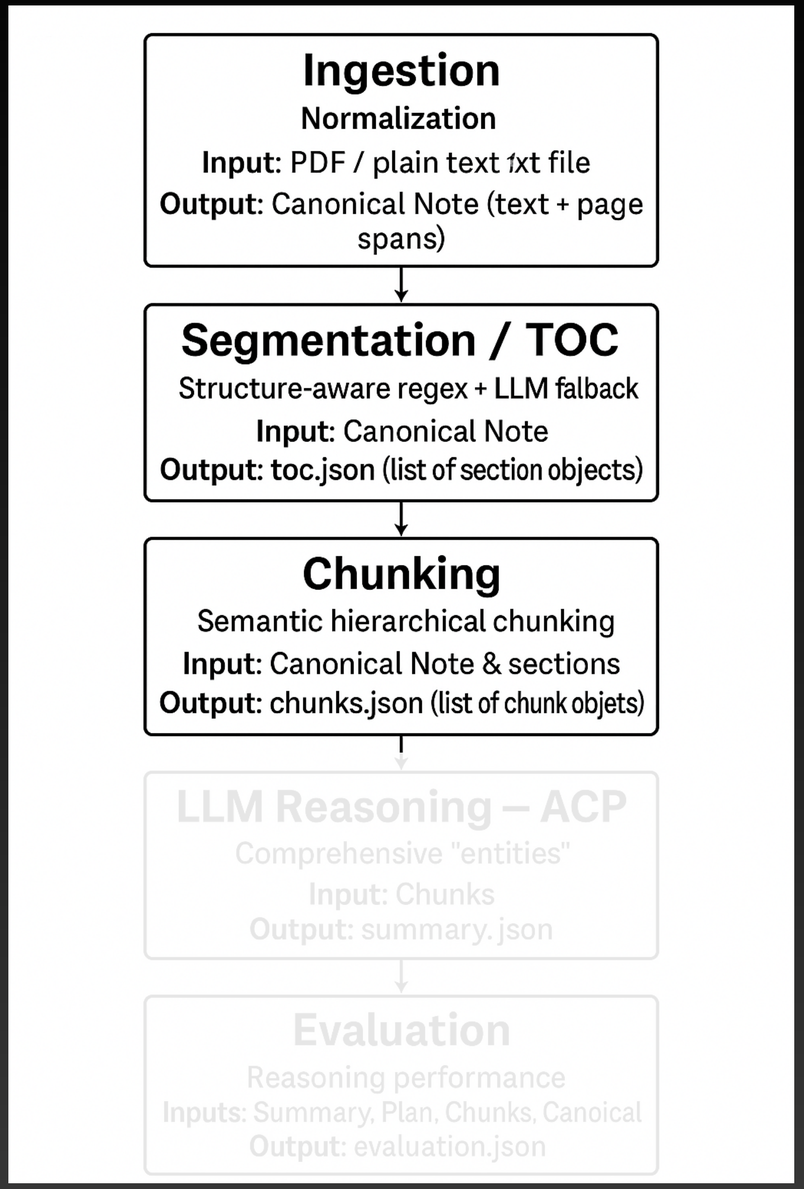
🧠ACPs Key Design Principles

1. Modular Design
   - Separation of concern
   - Components can be used independently
   - Maintable: changes to one stage don’t affect others
   - Extensible: easy to add/modify processing stages
2. Type safety
   - Pydantic models ensure type safety & validation
   - JSON schemas enable downstream processing
3. Caching & Incremental Processing
   - Plan generation from summary
   - Resume processing if interrupted
   - Selective execution: run summary or TOC only, if plan not needed
4. Robust error handling & “agentic” self-correction
   - Fast & reliable regex-based section detection
   - Rule-based first: chunking & LLM response cleaning
   - Multi-layer retry logic w ”agentic” validation retry
5. Traceability & Verification
   - Source citations with character-level precision
   - Comprehensive eval metrics for validation
   - Audit trail: complete logs of processing steps
6. Efficient local processing
   - Ollama (local) for LLM calls, privacy preserving
   - GPU acceleration (optionally)
   - Parallel processing w 2-4 workers (optionally)



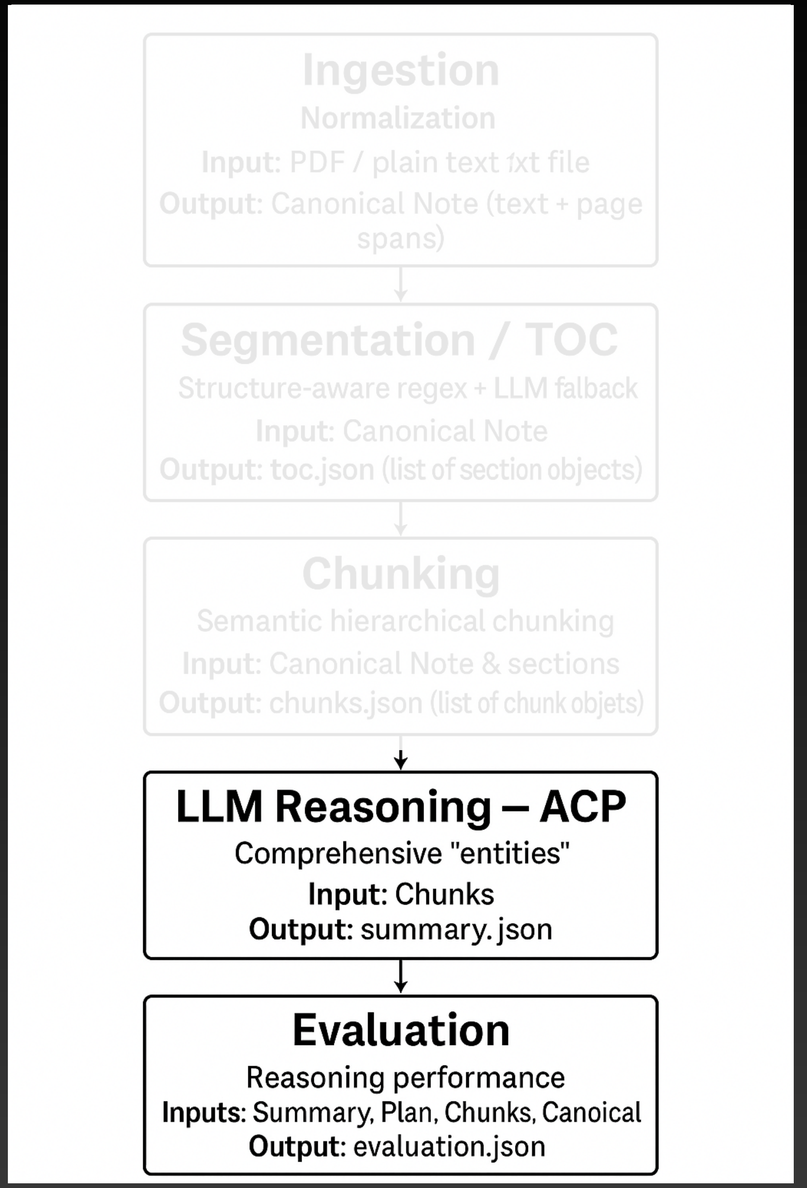
🧠ACPs Pre-Processing

- Ingestion
  - Input: PDF / plain txt file
  - Output: Canonical Note (text + page spans)
  - Process:
    - Text extraction (pypdf)
    - Normalise encoding and line endings
    - Preserve empty lines
    - Char-to-page mapping (PageSpan objects)
- Segmentation / TOC
  - Input: Canonical text
  - Output: toc.json (list of section objects)
  - Process:
    - Overview detection
    - Regex-based section header detection (Rules + Patterns) w LLM fallback
    - Section creation w char boundaries, mapped to page numbers
- Chunking
  - Inputs: Canonical Note & sections from TOC
  - Outputs: chunks.json (list of chunk objects)
  - Hierarchical chunking w semantic boundaries
  - Configuration
    - 1500 char chunk size, fits comfortably in token context windows
    - 200 char overlap for continuity at boundaries
    - 3000 char max paragraph size (2x chunk size)
  - Prompt + multiple chunks (3-10) fit into LLM token context window



🧠ACPs Reasoning

- Retrieval
  - Single LLM call instead of RAG w similarity search
  - Benefits
    - Complete context
    - No information loss
    - Faster/more efficient than multiple calls w retrieval + generation
    - Deterministic: no retrieval variability
- LLM-based reasoning
  - Prompt engineering principles
    - Anti-Hallucination instructions
    - Section definitions w Include/Exclude examples
    - Citation requirements w PHI Protection
    - Explicit output schema definition
    - Agentic validation retry logic w Pydantic error feedback
  - Summary generation
    - Input: Chunks
    - Output: summary.json
    - Structure: 7 non-overlapping sections, multiple entries w text & source fields
    - Process: Reason across chunks, parse into 7 sections, Pydantic validation
  - Plan generation
    - Input: summary.json
    - Output: plan.json
    - Structure: numbered recommendations w source, confidence, guard note fields
    - Process: using summary sections as entities, Pydantic validation



🧠ACPs Models & Tools

- LLM models
  - Qwen2.5 7B (via Ollama)
    - Good quality vs speed tradeoff
    - Sufficient context window
    - Good at following structured output formats
    - Local execution, open-source
  - Nomic-embed-text (via Ollama)
    - Lightweight, local, good for semantic accuracy eval
- Core libraries
  - Langchain: ollama support, well-tested, clean interface
  - PyPDF: lightweight, extracts font info, no OCR
  - Pydantic: schema validation
  - Typer: simple CLI
  - NumPy: efficient numerical computations
  - Matplotlib: visualization
  - Python-dotenv: env variable management
  - Pytest: testing framework

🧠ACPs Evaluation

- Eval metrics (n=500 docs)
  - Citation coverage
    - % of extracted facts that have source citations
    - Summary: 98.99% mean
    - Plan: 91.92% mean
  - Citation Validity
    - % of citations that are valid (correct ID, section name, char spans)
    - Overall: 80.2% mean, 90.5% median, 26.9% Std dev
  - Hallucination Detection
    - % of claims without source citation (orphan claims)
    - Overall: 0.04% mean. 0 - 20% range
  - Semantic Accuracy
    - Semantic similarity between plan recommendations and cited source text
    - Plan: 68% mean, 71% median, 34 - 96% range, 0.14 std dev
  - Confidence Scores
    - LLM confidence scores summary
    - Plan: 88% mean, 90% median, 0 - 100% range
  - Citation Overlap (Jaccard)
    - Similarity of extracted citations across all pairs (0 to 1)
    - Overall: 0.13 mean, 0.06 median
- Testing
  - Unit tests
    - Ingestion, section detection, chunking, schemas, config, llm client, evaluation
  - Integration tests
    - Summariser, Plan generator and pipeline tests with real LLM integration

🧠ACPs Results (Sample 3.pdf)

Original PDF

Medical Specialty:
Allergy / Immunology

Sample Name:Â Chronic Sinusitis

Description:Â Patient having severe sinusitis about two to three months ago with facial discomfort, nasal congestion, eye pain, and postnasal drip symptoms.
(Medical Transcription Sample Report)

HISTORY: I had the pleasure of meeting and evaluating the patient referred today for evaluation and treatment of chronic sinusitis. As you are well aware, she is a pleasant 50-year-old female who states she started having severe sinusitis about two to three months ago with facial discomfort, nasal congestion, eye pain, and postnasal drip symptoms. She states she really has sinus problems, but this infection has been rather severe and she notes she has not had much improvement with antibiotics. She had a CT of her paranasal sinuses identifying mild mucosal thickening of right paranasal sinuses with occlusion of the ostiomeatal complex on the right and turbinate hypertrophy was also noted when I reviewed the films and there is some minimal nasal septum deviation to the left. She currently is not taking any medication for her sinuses. She also has noted that she is having some problems with her balance and possible hearing loss or at least ear popping and fullness. Her audiogram today demonstrated mild high frequency sensorineural hearing loss, normal tympanometry, and normal speech discrimination. She has tried topical nasal corticosteroid therapy without much improvement. She tried Allegra without much improvement and she believes the Allegra may have caused problems with balance to worsen. She notes her dizziness to be much worse if she does quick positional changes such as head turning or sudden movements, no ear fullness, pressure, humming, buzzing or roaring noted in her ears. She denies any previous history of sinus surgery or nasal injury. She believes she has some degree of allergy symptoms.

PAST MEDICAL HISTORY: Seasonal allergies, possible food allergies, chronic sinusitis, hypertension and history of weight change. She is currently 180 pounds.

PAST SURGICAL HISTORY: Lower extremity vein stripping, tonsillectomy and adenoidectomy.

FAMILY HISTORY: Strong for heart disease and alcoholism.

CURRENT MEDICATIONS: DynaCirc.

ALLERGIES: Egg-based products cause hives.

SOCIAL HISTORY: The patient used to smoke cigarettes for about 20 years, one-half pack a day. She currently does not, which was encouraged to continue. She rarely drinks any alcohol-containing beverages.

PHYSICAL EXAMINATION:
VITAL SIGNS: Age 50, blood pressure is 136/74, pulse 84, temperature is 98.4, weight is 180 pounds, and height is 5 feet 3 inches.
GENERAL: The patient is healthy appearing; alert and oriented to person, place and time; responds appropriately; in no acute distress.
HEAD: Normocephalic. No masses or lesions noted.
FACE: No facial tenderness or asymmetry noted.
EYES: Pupils are equal, round and reactive to light and accommodation bilaterally. Extraocular movements are intact bilaterally. No nystagmus.
EARS: During Hallpike examination, the patient did not become dizzy until she would be placed back into sitting in the upright position. No nystagmus was appreciated; however, the patient did subjectively report dizziness, which was repeated twice. No evidence of any orthostatic hypotension was noted during the exam. Tympanic membranes were noted to be intact. No signs of middle ear effusion or ear canal inflammation.
NOSE: The patient appears congested. Turbinate hypertrophy is noted. There are no signs of any acute sinusitis. Septum is midline, slightly deviated to the left.
THROAT: There is clear postnasal drip. Oral hygiene is good. No masses or lesions noted. Both vocal cords move well to midline.
NECK: The neck is supple with no adenopathy or masses palpated. The trachea is midline. The thyroid gland is of normal size with no nodules.
LUNGS: Clear to auscultation bilaterally. No wheeze noted.
HEART: Regular rate and rhythm. No murmur noted.
NEUROLOGIC: Facial nerve is intact bilaterally. The remaining cranial nerves are intact without focal deficit.

PROCEDURE: Fiberoptic nasopharyngoscopy identifying turbinate hypertrophy and nasal septum deviation to the left, more significant posteriorly.

IMPRESSION:
1. Probable increasing problems with allergic rhinitis and chronic sinusitis, both contributing to the patient's symptoms.
2. Subjective dizziness, etiology uncertain; however, consider positional vertigo versus vestibular neuronitis as possible ear causes of dizziness, cannot rule out systemic, central or medication or causes at this time.
3. Inferior turbinate hypertrophy.
4. Nasal septum deformity.

RECOMMENDATIONS: An ENG was ordered to evaluate vestibular function. She was placed on Veramyst nasal spray two sprays each nostril daily and even twice daily if symptoms are worsening. A Medrol Dosepak was prescribed as directed. The patient was given instruction on use of nasal saline irrigation to be used twice daily and Clarinex 5 mg daily was recommended. After the patients' ENG examination, we will see the patient back for further evaluation and treatment recommendations. In light of the patient's atypical dizziness symptoms, I cannot rule out other pathology at this time, and I informed her if there are any acute changes or problems with regards to her balance or any other acute changes, which she attributes associated with her dizziness, she most likely should pursue an emergent visit to the emergency room.

Thank you for allowing me to participate with the care of your patient.

Chunks


```
{
  "chunks": [
    {
      "chunk_id": "chunk_0",
      "text": "Medical Specialty:\nAllergy / Immunology\n \nSample Name:
\nÂ Chronic Sinusitis\nDescription:\nÂ Patient having severe sinusitis
about two to three months ago with facial discomfort, nasal congestion,
eye pain, and\npostnasal drip symptoms. \n(Medical Transcription Sample
Report)",
      "start_char": 0,
      "end_char": 272,
      "section_title": "Overview"
    },
    {
      "chunk_id": "chunk_1",
      "text": "HISTORY:\n I had the pleasure of meeting and evaluating the
patient referred today for evaluation and treatment of chronic sinusitis.
As you are\nwell aware, she is a pleasant 50-year-old female who states
she started having severe sinusitis about two to three months ago with
facial\ndiscomfort, nasal congestion, eye pain, and postnasal drip
symptoms. She states she really has sinus problems, but this infection has
been rather\nsevere and she notes she has not had much improvement with
antibiotics. She had a CT of her paranasal sinuses identifying mild
mucosal\nthickening of right paranasal sinuses with occlusion of the
ostiomeatal complex on the right and turbinate hypertrophy was also noted
when I\nreviewed the films and there is some minimal nasal septum
deviation to the left. She currently is not taking any medication for her
sinuses. She also\nhas noted that she is having some problems with her
balance and possible hearing loss or at least ear popping and fullness.
Her audiogram today\ndemonstrated mild high frequency sensorineural
hearing loss, normal tympanometry, and normal speech discrimination. She
has tried topical nasal\ncorticosteroid therapy without much improvement.
She tried Allegra without much improvement and she believes the Allegra
may have caused\nproblems with balance to worsen. She notes her dizziness
to be much worse if she does quick positional changes such as head turning
or sudden\nmovements, no ear fullness, pressure, humming, buzzing or
roaring noted in her ears. She denies any previous history of sinus
surgery or nasal\ninjury. She believes she has some degree of allergy
symptoms.",
      "start_char": 272,
      "end_char": 1898,
      "section_title": "HISTORY"
    },
    {
      "chunk_id": "chunk_2",
      "text": "PAST MEDICAL HISTORY: \nSeasonal allergies, possible food
allergies, chronic sinusitis, hypertension and history of weight change.
She is\ncurrently 180 pounds.",
      "start_char": 1898,
      "end_char": 2057,
      "section_title": "PAST MEDICAL HISTORY"
    },
    {
      "chunk_id": "chunk_3",
      "text": "PAST SURGICAL HISTORY:\n Lower extremity vein stripping,
tonsillectomy and adenoidectomy.",
      "start_char": 2057,
      "end_char": 2146,
      "section_title": "PAST SURGICAL HISTORY"
    },
    {
      "chunk_id": "chunk_4",
      "text": "FAMILY HISTORY: \nStrong for heart disease and
alcoholism.",
      "start_char": 2146,
      "end_char": 2204,
      "section_title": "FAMILY HISTORY"
    },
    {
      "chunk_id": "chunk_5",
      "text": "CURRENT MEDICATIONS: \nDynaCirc.",
      "start_char": 2204,
      "end_char": 2236,
      "section_title": "CURRENT MEDICATIONS"
    },
    {
      "chunk_id": "chunk_6",
      "text": "ALLERGIES: \nEgg-based products cause hives.",
      "start_char": 2236,
      "end_char": 2280,
      "section_title": "ALLERGIES"
    },
    {
      "chunk_id": "chunk_7",
      "text": "SOCIAL HISTORY: \nThe patient used to smoke cigarettes for
about 20 years, one-half pack a day. She currently does not, which
was\nencouraged to continue. She rarely drinks any alcohol-containing
beverages.",
      "start_char": 2280,
      "end_char": 2485,
      "section_title": "SOCIAL HISTORY"
```


TOC


```
{
  "sections": [
    {
      "title": "Overview",
      "start_char": 0,
      "end_char": 272,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "HISTORY",
      "start_char": 272,
      "end_char": 1898,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "PAST MEDICAL HISTORY",
      "start_char": 1898,
      "end_char": 2057,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "PAST SURGICAL HISTORY",
      "start_char": 2057,
      "end_char": 2146,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "FAMILY HISTORY",
      "start_char": 2146,
      "end_char": 2204,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "CURRENT MEDICATIONS",
      "start_char": 2204,
      "end_char": 2236,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "ALLERGIES",
      "start_char": 2236,
      "end_char": 2280,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "SOCIAL HISTORY",
      "start_char": 2280,
      "end_char": 2485,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "PHYSICAL EXAMINATION",
      "start_char": 2485,
      "end_char": 4085,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "PROCEDURE",
      "start_char": 4085,
      "end_char": 4230,
      "start_page": 0,
      "end_page": 0
    },
    {
      "title": "IMPRESSION",
      "start_char": 4230,
      "end_char": 4648,
      "start_page": 0,
      "end_page": 1
    },
    {
      "title": "RECOMMENDATIONS",
      "start_char": 4648,
      "end_char": 5547,
      "start_page": 1,
      "end_page": 1
    }
  ]
}
```


Summary


```
{
  "patient_snapshot": [
    {
      "text": "50-year-old female",
      "source": "[OVERVIEW] section, chunk_0:12-23"
    }
  ],
  "key_problems": [
    {
      "text": "Chronic sinusitis",
      "source": "[HISTORY] section, chunk_1:49-67"
    },
    {
      "text": "Severe sinusitis with facial discomfort, nasal congestion, eye pain, and
postnasal drip symptoms",
      "source": "[HISTORY] section, chunk_1:28-53"
    },
    {
      "text": "Subjective dizziness",
      "source": "[IMPRESSION] section, chunk_10:46-52"
    }
  ],
  "pertinent_history": [
    {
      "text": "Seasonal allergies, possible food allergies, chronic sinusitis, hypertension and
history of weight change",
      "source": "[PAST MEDICAL HISTORY] section, chunk_2:19-83"
    },
    {
      "text": "Lower extremity vein stripping, tonsillectomy and adenoidectomy",
      "source": "[PAST SURGICAL HISTORY] section, chunk_3:20-54"
    },
    {
      "text": "Strong family history for heart disease and alcoholism",
      "source": "[FAMILY HISTORY] section, chunk_4:21-67"
    }
  ],
  "medicines_allergies": [
    {
      "text": "Egg-based products cause hives",
      "source": "[ALLERGIES] section, chunk_6:23-50"
    },
    {
      "text": "DynaCirc",
      "source": "[CURRENT MEDICATIONS] section, chunk_5:24-31"
    }
  ],
  "objective_findings": [
    {
      "text": "Turbinate hypertrophy and nasal septum deviation to the left, more significant
posteriorly",
      "source": "[PHYSICAL EXAMINATION] section, chunk_8:4079-4125"
    },
    {
      "text": "Normocephalic. No masses or lesions noted",
      "source": "[PHYSICAL EXAMINATION] section, chunk_8:3069-3088"
    },
    {
      "text": "No facial tenderness or asymmetry noted",
      "source": "[PHYSICAL EXAMINATION] section, chunk_8:3127-3145"
    },
    {
      "text": "Pupils are equal, round and reactive to light and accommodation bilaterally.
Extraocular movements are intact bilaterally. No nystagmus",
      "source": "[PHYSICAL EXAMINATION] section, chunk_8:3207-3245"
    },
    {
      "text": "Tympanic membranes were noted to be intact. No signs of middle ear effusion or
ear canal inflammation",
      "source": "[PHYSICAL EXAMINATION] section, chunk_8:3619-3646"
    },
    {
      "text": "The patient appears congested",
      "source": "[PHYSICAL EXAMINATION] section, chunk_8:3572-3580"
    }
  ],
  "labs_imaging": [
    {
      "text": "CT of paranasal sinuses identifying mild mucosal thickening of right paranasal
sinuses with occlusion of the ostiomeatal complex on the right and turbinate hypertrophy was
noted",
      "source": "[HISTORY] section, chunk_1:68-142"
    },
    {
      "text": "Audiogram demonstrated mild high frequency sensorineural hearing loss, normal
tympanometry, and normal speech discrimination",
      "source": "[HISTORY] section, chunk_1:153-190"
    }
  ],
```


ACP


```
{
  "recommendations": [
    {
      "number": 1,
      "recommendation": "She was placed on Veramyst nasal spray two sprays each nostril daily
and even twice daily if symptoms are worsening",
      "source": "[RECOMMENDATIONS] section, chunk_11:4683-4729",
      "confidence": 0.9,
      "hallucination_guard_note": null
    },
    {
      "number": 2,
      "recommendation": "A Medrol Dosepak was prescribed as directed",
      "source": "[RECOMMENDATIONS] section, chunk_11:4730-4750",
      "confidence": 0.9,
      "hallucination_guard_note": null
    },
    {
      "number": 3,
      "recommendation": "The patient was given instruction on use of nasal saline irrigation to
be used twice daily and Clarinex 5 mg daily was recommended",
      "source": "[RECOMMENDATIONS] section, chunk_11:4762-4803",
      "confidence": 0.9,
      "hallucination_guard_note": null
    },
    {
      "number": 4,
      "recommendation": "An ENG was ordered to evaluate vestibular function",
      "source": "[RECOMMENDATIONS] section, chunk_11:4652-4670",
      "confidence": 0.9,
      "hallucination_guard_note": null
    },
    {
      "number": 5,
      "recommendation": "After the patients' ENG examination, we will see the patient back for
further evaluation and treatment recommendations",
      "source": "[RECOMMENDATIONS] section, chunk_11:4804-4859",
      "confidence": 0.9,
      "hallucination_guard_note": null
    },
    {
      "number": 6,
      "recommendation": "In light of the patient's atypical dizziness symptoms, I cannot rule
out other pathology at this time",
      "source": "[RECOMMENDATIONS] section, chunk_11:4860-4879",
      "confidence": 0.9,
      "hallucination_guard_note": null
    },
    {
      "number": 7,
      "recommendation": "If there are any acute changes or problems with regards to her balance
or any other acute changes, which she attributes associated with her dizziness, she most likely
should pursue an emergent visit to the emergency room",
      "source": "[RECOMMENDATIONS] section, chunk_11:4880-5267",
      "confidence": 0.9,
      "hallucination_guard_note": null
    }
  ]
}
```

🧠ACPs Safety & Limits

- Pre-processing limitations
  - Need OCR support e.g. for scanned PDFs, multimodal inputs
  - Limited encoding error handling & text validation
  - Limited LLM fallback for regex failures
  - Fixed chunk size may not be optimal for all docs
- Reasoning limitations
  - Low performance for “Autopsy” type reports - need a separate pipeline
  - Good mean performance, but variability in citation validity/ semantic accuracy suggests occasional hallucinations
  - No alerts/resolution for conflicting information
  - No distinction for ‘info not present’ and ‘info not extracted’ (no info yields [] fields)
  - No temporal awareness
  - Non-standard docs may cause issues with regex (e.g. Autopsy reports): need more complex LLM fallback
  - Dealing with LLM response failures (beyond 3 retries, error logging, graceful degradation)
  - Citation validation issues: dealing w occasional name mismatches, span out of bounds, missing chunk ID
  - Exceptionally long documents may break context window limit
  - No cross-document intelligence (e.g. for multiple visits, cross-doc summary, tracking across time)
  - No medical knowledge validation
- Evaluation limitations
  - Need to validate semantic accuracy thresholds with golden answers
  - No human validation (e.g. tooling for human reviewers)
- Testing limitations
  - Unit tests to handle all edge cases above (e.g. 100+ sections, long sections, tables, etc)
  - Unit tests for further LLM failure edge cases
  - Performance testing (processing time, memory usage, latency, scalability, ...)
  - Regression testing: golden file tests, version comparisons, prompt changes
  - Complex integration testing

🧠ACPs Cost & Performance

- Cost
  - LLM inference, embeddings, pdf processing, storage all local => 0 API costs
  - No vector store solution => no exploding vector storage/subscription costs
  - Cost savings: $0.15 - $0.50 per doc with GPT-4 => $150 - $500 for 1000 docs + EC2 costs (50-100$/month)
  - Hardware Reqs: Multi-core CPU, GPU (optional), 8GB+ RAM, ~4GB storage for model
- Token Usage
  - ~3500 tokens for small, ~4700 tokens for medium, ~12000 tokens for large docs
  - Efficiency
    - Single LLM call for summary, no additional retrieval calls (as in RAG)
    - Plan generation starts with summary.json, minimising token usage
- Latency
  - Ingestion/TOC/Chunking: 0.01 - 0.05s each
  - Summary generation: 10-60s (CPU) | 2-10s (GPU)
  - Plan generation: 5-30s (CPU) | 1-5s (GPU)
  - Eval: 2-10s (CPU) | 1-3s (GPU)
  - Total: 17-101s (CPU) | 4-19s (GPU)
- Throughput
  - Sequential: 90 docs/h (CPU), 450 docs/h (GPU)
  - Parallel (4 workers): 360 docs/h (CPU), 1800 docs/hour (GPU)
- Tradeoffs
  - Full-context vs RAG: short docs/full context valuable
  - Local vs. Cloud: privacy/zero API costs with acceptable latency
  - Overlap vs. no overlap chunking: significant accuracy benefit
  - 7B vs larger models: good balance for quality, speed, hardware reqs