# LLMs for unstructured claims data

Rob Lieberthal, PhD

Lieberthal & Associates, LLC

July 28, 2026

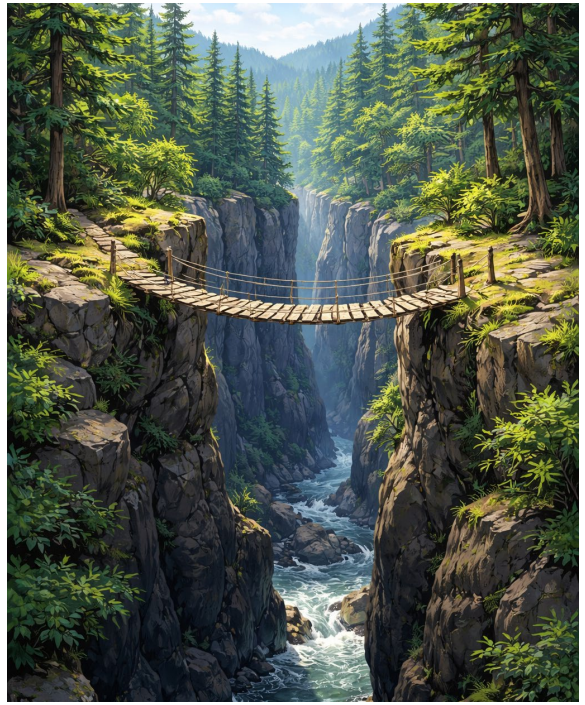
# Disclosure and about me


- This project was funded by the Casualty Actuarial Society

- All views are my own and do not represent the SOA or the other organizations I work for

- Thanks to the project team and CAS staff and AI Working Group
    - Team members: Richard Tran (MDSight), Bao Phan, Jawand Singh, and Lizzie Sottung

# The unstructured data gap



**Actuaries rely on structured data:** diagnosis codes, claim amounts, dates

Claims generate rich narrative throughout the life cycle: medical records, adjuster notes, phone transcripts, claimant statements, settlement documents

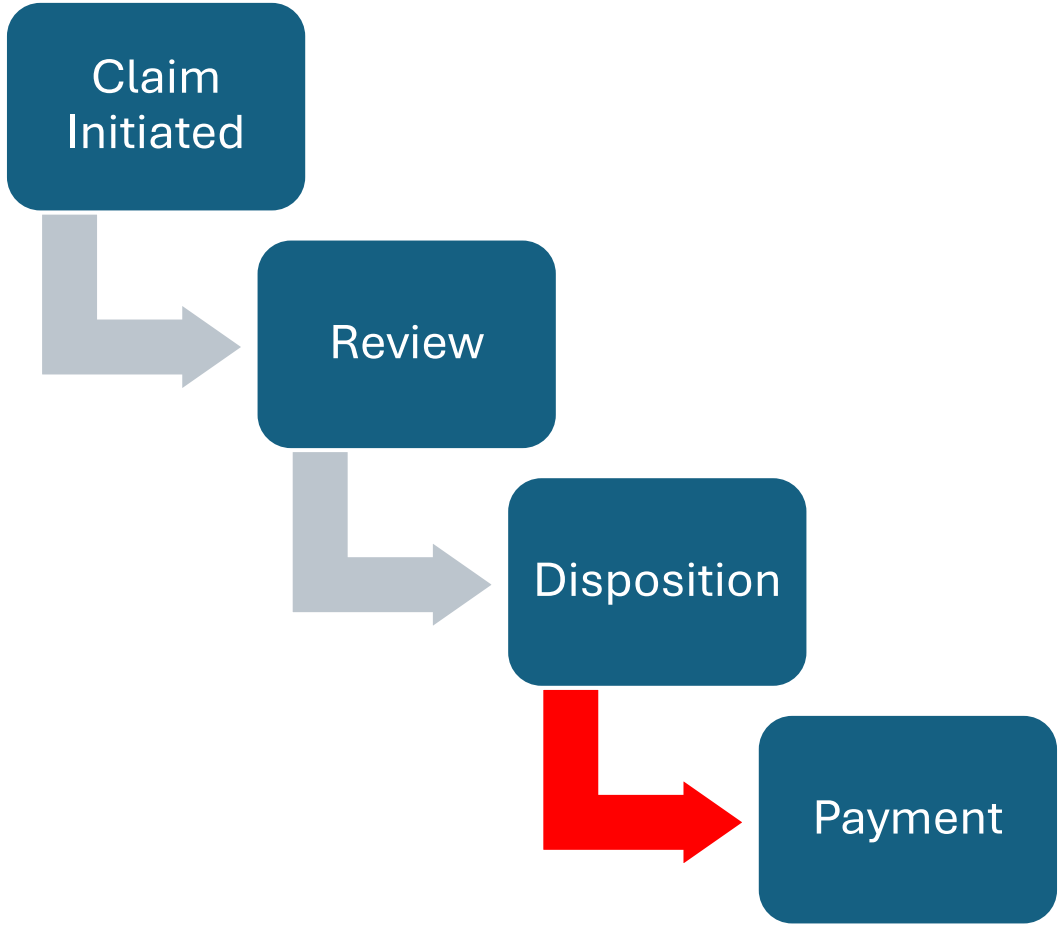
Manual extraction is slow, inconsistent across reviewers, and not scalable

Over 80% of potentially useful business information exists in unstructured form

*Image via ChatGPT

# LLMs offer a repeatable assessment

- **Insurance companies have LLMs in their workflow**
  - Summarization, search, document triage
  - This extends them to structured extraction with validation
- **The key value is reliability and consistency**
  - Same  process for every claim
  - Structured output for actuarial systems
  - Systems that get better as you work



# Research objectives

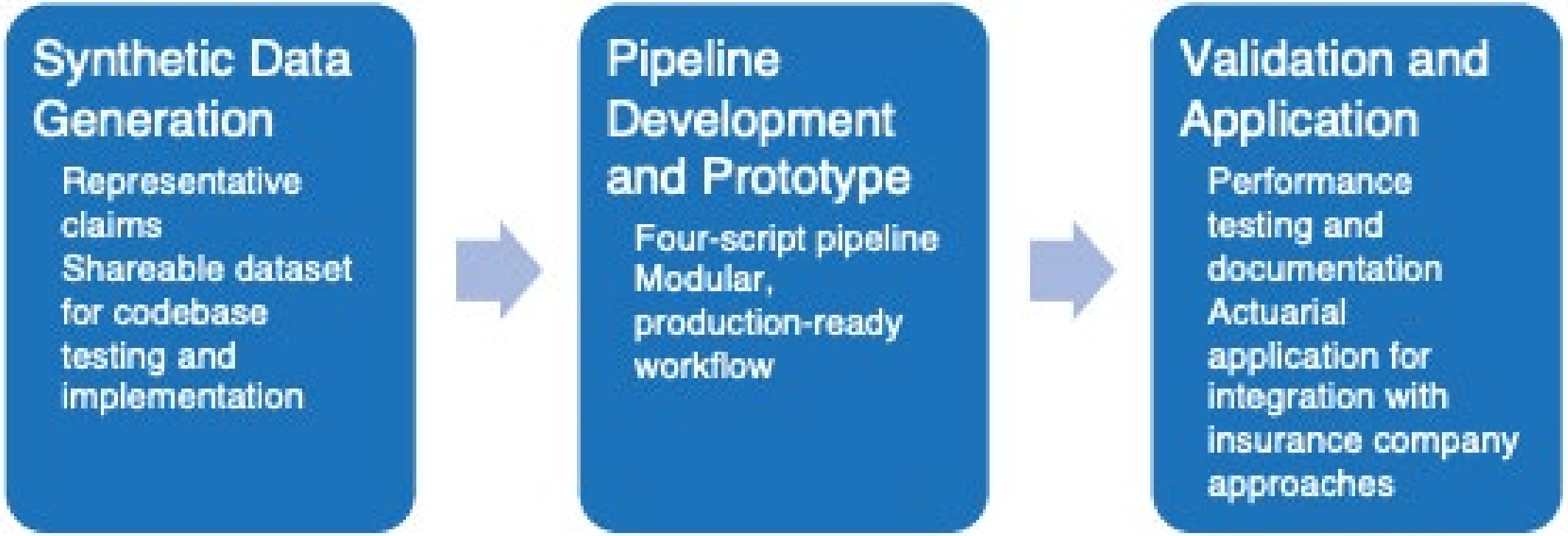
**Reproducible methodology** — establish a systematic, replicable approach for applying LLMs to extract actuarial variables

**Practitioner-ready tools** — translate LLM capabilities into practical tools with clear documentation

**Validated integration** — benchmark LLM-extracted variables and demonstrate that outputs work with existing actuarial systems

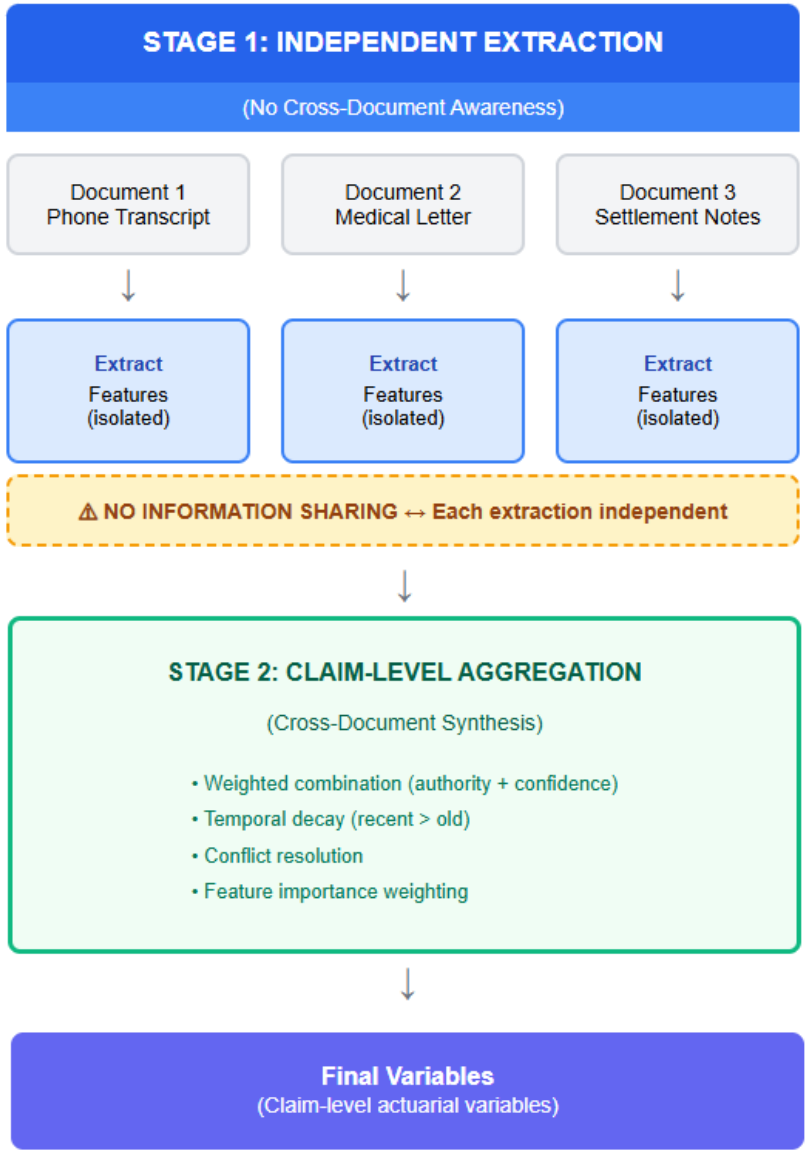
**Practical usability** — deliver a working demonstration, a simplified medical-records pipeline

# Project workflow and data processing pipeline

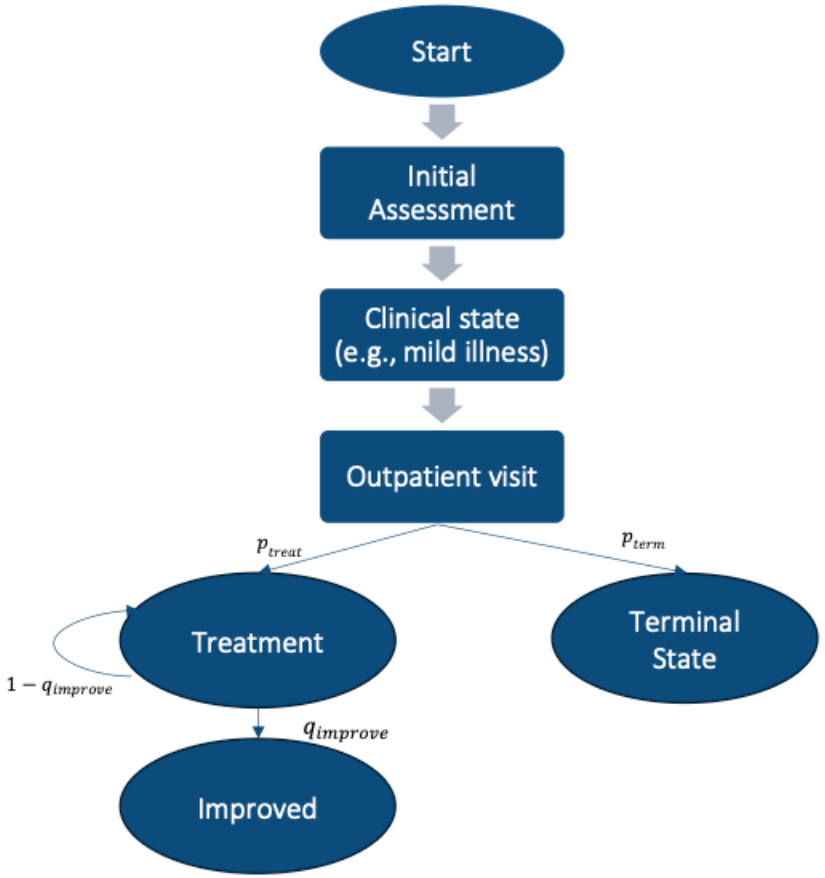


# Prototype workflow

- Stage 1 processes documents sequentially but independently
- Stage 2 provides the case context where the temporal nature of claims becomes relevant
- Four factors combine multiplicatively to produce final feature weights for aggregation
- Full technical details on the compound scoring algorithm are on the CAS website



# Synthetic data approach




- Realistic healthcare data
- Generated based on Markov state transition
- Represents aspects of patient care and coverage
- Non-sensitive; Can be shared
- Used to generate “ground truth” for bias assessment methods

# Variable taxonomy: 36 variables, 14 in this prototype

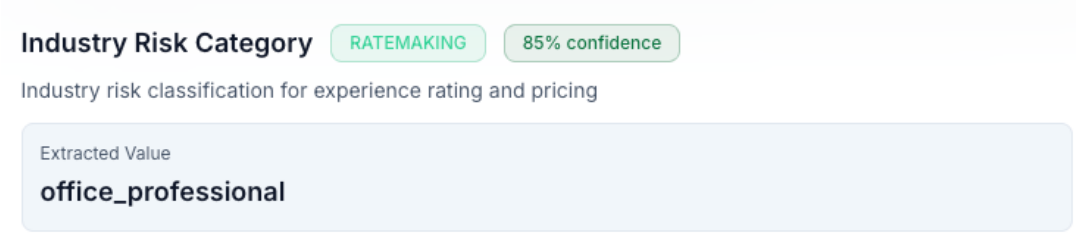
| | |
|---|---|
| **Reserving** | Injury severity, medical complexity, recovery trajectory, ultimate cost category |
| **Ratemaking** | Causation type, safety compliance, pre-existing conditions, risk modifiers |
| **Claims Management** | Litigation risk, settlement likelihood, return-to-work prognosis, intervention priority |

*Each variable includes a categorical value, confidence score (0–1), and mandatory rationale field citing specific document evidence.*

# LLM results – how does the machine score

## Aggregation Logic

- 1. Extract occupation from `claimant_age_occupation`
- 2. Map to North American Industrial Classification Scheme (NAICS) industry code
- 3. Map NAICS to standard `industry_classification` categories




Allowed values = `construction_heavy, construction_light, manufacturing_heavy, manufacturing_light, healthcare_hospital, healthcare_outpatient, transportation, retail, office_professional, agriculture, utilities, public_safety, education, other`

# LLM rationale – what did the LLM “think”

**Stated Rationale**

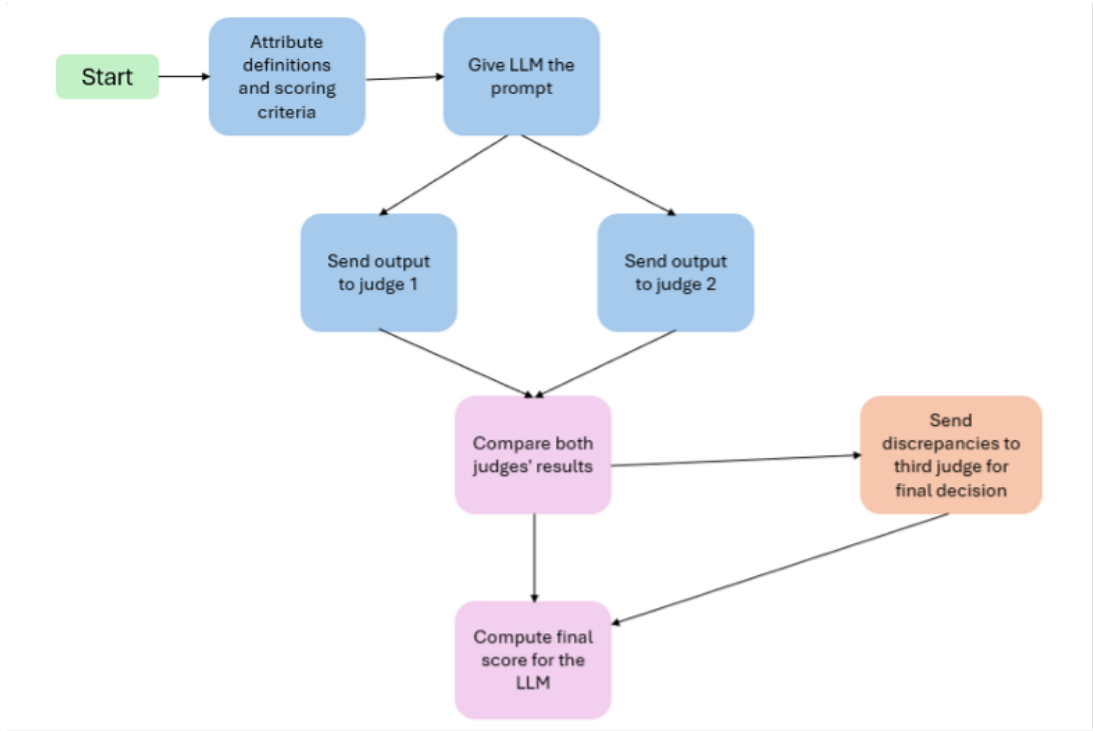
- “The industry risk category is classified as office professional based on the claimant's occupation as an Administrative Assistant, which typically involves office work with moderate risk factors. This classification aligns with the nature of the incident and the work environment.”

**Stage 1 Source Contributions (Partial)**

| Source | Mapping | Confidence |
|---|---|---|
| Adjuster Notes | Industry Type | 0.9 |
| Adjuster Notes | Employer Classification | 0.85 |
| Phone Transcript | Employer Name | 0.7 |
| Phone Transcript | Job Title | 0.68 |
| Phone Transcript | Work Description | 0.65 |
| Claimant Statement | Employer Information | 0.6 |

# Validation design

- **Synthetic data:** Synthea FHIR records augmented with LLM-generated documents (adjuster notes, transcripts, claimant statements). Ground truth is known – no PHI required.
- Two independent clinician reviewers score LLM extractions on a Likert 1–5 scale using a structured rubric
- Disagreements adjudicated by an independent third reviewer
- Rubric and scoring tool available at **validate.jawand.dev** (the same tool you will use today)



# Validation results

**Average extraction score: >4 out of 5**

**Weighted kappa between reviewers: 0.53** (moderate inter-rater agreement)

**Divergence rate: 19%**

Extractions with confidence >0.85: **>95% agreement** with expert judgment

Occupational risk category and medical complexity: strongest performance.

*Even trained clinicians disagreed 19% of the time. You will experience this shortly.*

# Integration with chain ladder reserving

SEVERITY BREAKDOWN

## Segmented chainladder results

| SEGMENT | CLAIMS | LATEST PAID | ULTIMATE | IBNR | SHARE OF SEGMENTED IBNR |
|---|---|---|---|---|---|
| Major | 92 | $16,058,199 | **$16,746,356** | $688,157 | 82.2% |
| Moderate | 158 | $8,666,700 | **$8,801,153** | $134,453 | 16.1% |
| Minor | 170 | $2,598,770 | **$2,613,225** | $14,454 | 1.7% |
| **Total** | **420** | **$27,323,670** | $28,160,733 | $837,063 | **100%** |

**Minor**
Fast emergence, low noise, minimal tail.

**Moderate**
Steady emergence, balanced noise profile.

**Major**
Longer tail, materially heteroskedastic increments.

# Limitations of our research approach

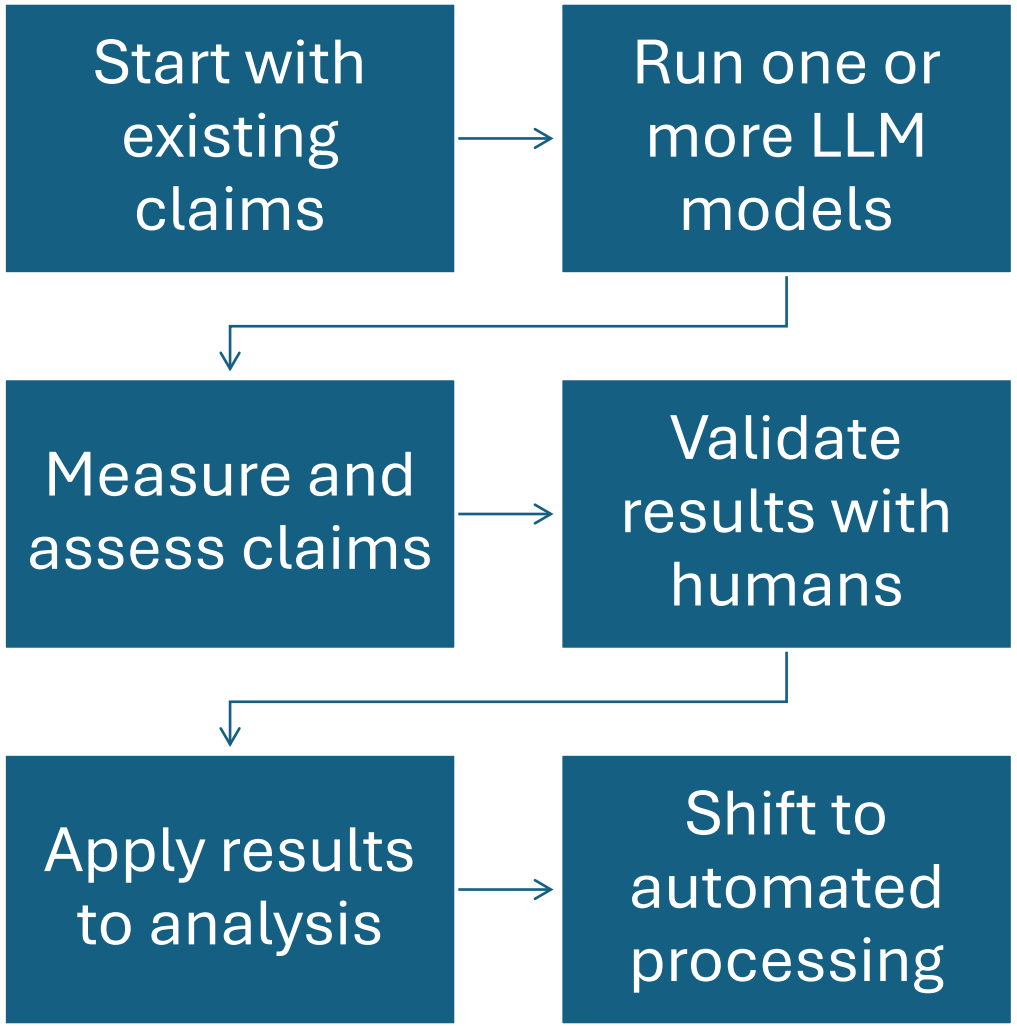
- Validated on synthetic data; real-world performance may differ
- Litigation risk requires additional domain-specific study beyond medical records
- Timestamp and maturity milestone tracking not yet implemented for production use
- Production deployment requires HIPAA deidentification and controls
- The modular architecture supports phased adoption: validate against real claims in parallel before full rollout

# What does this look like at scale?

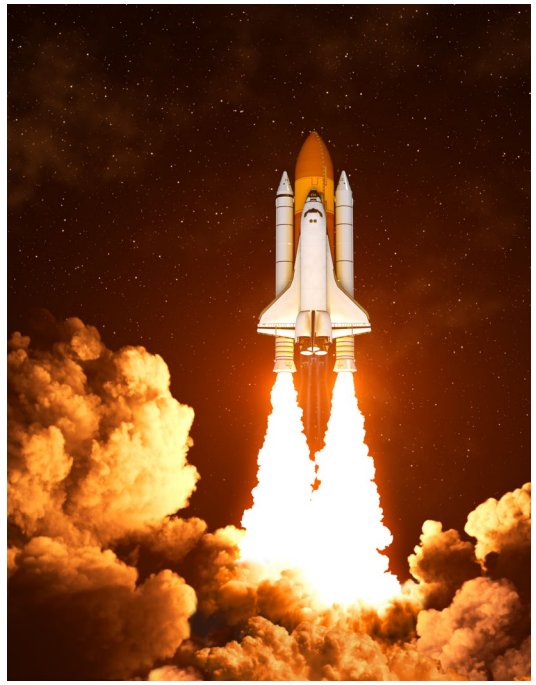
**A commercial version at scale**

- Same examination on every claim
- Privacy and security enabled
- Plugs into existing claims systems
- Continuous validation loop built in

**Catch errors and exceptions, refine as needed**



# Key takeaways



**Unstructured data is the untapped frontier** — Claims data often lives in free-text notes

**LLMs enable repeatable, scalable extraction** — Every claim gets the same consistent examination

**Human-in-the-loop validation is essential** — Reinforcement and continuous feedback

**Ready to scale** — Prototype plugs into existing claims systems with a built-in validation loop

# Thank you!

**Rob Lieberthal, PhD**
Lieberthal & Associates, LLC

rob@lieberthalandassociates.com
LinkedIn (Personal) | LinkedIn (Company)
917-756-8508

*Questions & Discussion Welcome*

# Backup

# Scoring rubric

| Principal | Dimension | Definition |
|---|---|---|
| Quality and Accuracy | Relevance | Correctness of information |
| Quality and Accuracy | Accuracy | Accuracy of information |
| Understanding and Reasoning | Contextual Understanding | Ability to interpret claims |
| Understanding and Reasoning | Consistency with Code Sets | Use of specialized terms and codes |
| Safety and Ethics | Bias and Fairness | Output is free from unfair assumptions |
| Safety and Ethics | Hallucination | No fabricated information |

All should be scored on a “Likert scale” (1-5)

- 1 = Strongly disagree
- 5 = Strongly agree

# Validation metrics

- Overall quadratic weighted kappa was 0.53 (linear: 0.40), consistent with moderate agreement
- Exact agreement occurred on 51.5% of items and within-one-point agreement on 81.0%
- 19.0% of items (38 of 200, across 17 of 20 claims) diverged by two or more points.
- Mean absolute difference was 0.74 points.
- Reviewer 1 had a mean score of 4.01 (SD 1.14) and Reviewer 2 a mean of 4.23 (SD 1.25).

| Principle | Exact | %≥2 | $\kappa_w$ |
|---|---|---|---|
| Quality and Accuracy | 46% | 15% | 0.54 |
| Safety and Ethics | 62% | 18% | 0.60 |
| Understanding and Reasoning | 40% | 30% | 0.34 |
| Overall | 52% | 19% | 0.53 |