

AI EVALUATION / DATA QUALITY / ANALYTICAL QA

Christopher Lyvers

AI Quality & Data Analyst

LLM evaluation, prompt and rubric design, multi-file reconciliation, calculation verification, and evidence-based analytical reasoning.

209

result QA ledger

7-10

files in core
reconciliations

2-3

responses compared

7

case studies

christopher-lyvers-portfolio.netlify.app

PORTFOLIO GUIDE

Evaluation you can audit.

I evaluate LLM outputs, design prompts and rubrics, reconcile messy multi-file records, verify calculations, and explain exactly why a conclusion is - or is not - supported.

Featured work

01 / LLM Response Evaluation

Compared candidate responses, separated grounded claims from unsupported assertions, weighted error severity, and wrote evidence-based ratings and rationales.

02 / Water-Quality Dataset QA

Reconciled 209 laboratory results across 10 files, validated timing and QC rules, corrected percentile logic, and separated screening findings from formal decision points.

03 / AI Evaluation Design

Designed failure-revealing multi-file tasks, explicit scoring criteria, reference calculations, edge cases, and grading guidance across three complex domains.

Additional work samples

04 / Medication-Order Reconciliation

Matched records across care settings, excluded a wrong-patient lab, trended renal data, checked a derived clearance value, and classified 11 active orders.

05 / Trauma Timeline & Transfusion Reconciliation

Rebuilt a crossing-midnight timeline, distinguished issued from administered products, verified ratios and thresholds, and recalculated weight- and rate-based values.

06 / Claims Evidence Audit

Audited 14 indexed records across a 19-day denied range, excluded contaminated evidence, separated defensible dates from weak ones, and removed unsupported diagnoses.

07 / Medical Record Evidence Synthesis

Aligned a referral, medication list, imaging report and laboratory data; separated confirmed findings from suspected conditions; and preserved missing-evidence boundaries.

ABOUT THESE WORK SAMPLES

Selected projects in this portfolio are adapted from completed AI evaluation and data-quality assignments. Some assignments used synthetic records, fictional organizations, and simulated real-world scenarios to test model reasoning, data reconciliation, and quality assurance. The analytical and evaluation work shown reflects work I performed. Internal platform instructions and identifiers have been removed.

01 / AI EVALUATION / MODEL QUALITY WORK SAMPLE

LLM Response Evaluation

Factuality, safety & instruction-following review

Compared candidate responses, separated grounded claims from unsupported assertions, weighted error severity, and wrote evidence-based ratings and rationales.

10 archived evaluation screenshots reviewed for this work sample	2-3 candidate responses compared in representative tasks	3 & 5 level overall-quality scales applied	Multiple criteria: grounding, completeness, safety, honesty and fit
--	--	--	---

Overview

The assignment work required more than choosing a preferred answer. Each response had to be tested against the prompt, available context, factual support, safety expectations, and the seriousness of any failure. Archived evidence shows work across factual-claim review, groundedness, completeness, harmfulness, instruction following, pairwise comparison, three-response selection, and multi-level quality scoring.

Challenge

A polished response can still fail on the criterion that matters most. The core challenge was to distinguish cosmetic weaknesses from unsupported factual claims, broken instructions, or safety failures - and to explain why the distinction changed the final rating.

Source material

- User prompts and candidate model responses
- Provided context documents or task-contained evidence
- Evaluation criteria covering groundedness, completeness, harmfulness and overall quality
- Archived rating selections and written rationales

My work

- Compared responses directly rather than grading each in isolation.
- Isolated claims that required external verification and distinguished them from statements already supplied by the context.
- Checked whether outputs followed explicit content, tone, length and format instructions.
- Identified hallucinated, unsafe, dishonest or unsupported content and weighted it by severity.
- Assigned ratings or rankings and wrote concise rationales tied to specific response evidence.

Analytical approach

- 01** Establish the task contract
Translate the prompt into observable requirements before judging style or preference.
- 02** Trace claims to evidence
Separate context-supported statements from claims that need outside verification or have no support.
- 03** Classify the failure
Distinguish factuality, grounding, completeness, instruction-following and safety errors.
- 04** Weight severity
Treat a central unsafe or false claim as more important than several minor presentation strengths.
- 05** Write an auditable rationale
Name the decisive evidence and explain why it changes the rating or ranking.

Key findings

Unsupported factual claim	A response stated that Saturn is the largest planet. I treated this as an externally verifiable factual claim and identified it as not grounded rather than accepting fluent wording at face value.
Constraint failure in creative output	For a sandwich-themed love poem, the evaluation required checking both sonnet structure and the requested balance of sincerity and whimsy - not merely whether the writing sounded pleasant.
Safety outweighed partial usefulness	In a mold-remediation comparison, one answer included useful moisture-control and protective-equipment advice but also suggested that household chemical treatment could permanently eliminate mold roots in porous drywall. I treated the central safety and judgment failure as decisive.
Audience fit mattered	For a monetary-policy explanation requested at a five-year-old level and under 400 words, I evaluated factual content together with accessibility and adherence to the requested audience.

Deliverable

Completed response selections, quality labels, criterion-level judgments, and written evidence-based rationales. Raw screenshots are not published because they contain internal interface details; the case study preserves the evaluation method and sanitized examples.

Skills demonstrated

LLM output evaluation	Factuality checking	Hallucination detection
Groundedness review	Instruction-following assessment	Safety judgment
Error severity classification	Pairwise ranking	Evaluation rationale writing

WHY IT MATTERS

Model-quality roles require defensible judgment, not taste. This work shows how I turn broad prompts and messy outputs into explicit criteria, identify the failure that actually matters, and explain the decision in language another reviewer can audit.

Source context: Completed AI evaluation assignments; internal platform identifiers and workflow details removed.

Water-Quality Dataset QA

209-result reconciliation & decision-logic audit

Reconciled 209 laboratory results across 10 files, validated timing and QC rules, corrected percentile logic, and separated screening findings from formal decision points.

10 source files across CSV, PDF and TXT	209 analyte-results in the final ledger	58 Glenmoor sample IDs reconciled 1:1	191 / 12 / 1 / 1 / 4 valid / qualified / rejected / excluded / not formal-decision
--	--	--	--

Overview

This assignment used a simulated water-system and laboratory package to test whether a model could reconcile custody records, analytical exports, quality-control evidence, program roles, and conflicting draft conclusions. I built and validated a row-level review that kept data validity separate from threshold interpretation.

Challenge

The files did not agree automatically. One reference card was obsolete, one draft used the wrong percentile method, a late analysis affected some analytes but not others, an out-of-scope row appeared in the metals export, and a QC failure required qualification without inventing a corrected result.

Source material

- Chain-of-custody and sample-receiving CSV files
- Sampling-plan/site-role records
- Ion chromatography and ICP-MS analytical exports
- Batch QC summary and instrument calibration log
- Laboratory maintenance note
- Outdated limits reference and a superseded preliminary draft

My work

- Reconciled 58 in-scope sample IDs across custody, receiving, plan and analytical records.
- Expanded 46 three-analyte ion rows into 138 results and combined them with 71 metals results to produce a 209-row ledger.
- Checked collection-to-analysis holding times by analyte instead of applying one rule to an entire sample.
- Applied QC qualifications, excluded out-of-scope data, and preserved rejected results without fabricating values.
- Recomputed lead and copper 90th percentiles using the assignment's required ordinal-rank method.

- Separated valid screening observations from the subset eligible for formal decision-making.

Analytical approach

- 01** Reconcile identity and scope
Match work order, sample IDs and program roles before interpreting any number.
- 02** Establish row disposition
Classify each result as valid, qualified, rejected, excluded or not a compliance result.
- 03** Verify calculations
Recompute elapsed times, units, rounding and ordinal percentiles from source values.
- 04** Resolve source authority
Use current governing criteria when a superseded draft or outdated card conflicts with primary records.
- 05** Separate evidence layers
Keep data validity, screening observations and formal decision criteria distinct.

Key findings

Analyte-specific holding-time decision

Sample DS-33 was analyzed 70.03 hours after collection. Nitrate and nitrite exceeded their 48-hour requirement and were qualified for exclusion from formal decisions; fluoride from the same sample remained valid because its allowed holding time was 672 hours.

Percentile calculation corrected

For 10 lead sites, the required ordinal rank was $\text{ceil}(0.90 \times 10) = 9$. The ninth ordered value was 14 microg/L (0.014 mg/L), not the draft's interpolated 0.016 mg/L. Copper's corresponding result was 0.38 mg/L; the draft had substituted the maximum, 0.52 mg/L.

Obsolete threshold rejected

A 2004 bench card listed an arsenic value of 0.05 mg/L. I did not let it override the current 0.010 mg/L criterion, which changed the interpretation of the 0.011 mg/L entry-point result and the 0.013 mg/L distribution screening result.

QC issue handled without false precision

A copper continuing-calibration verification result of 89% triggered a J qualification on 12 copper rows. I retained the reported measurements as estimated and did not invent a numerical correction.

Scope and rejection controls

OUTFALL-002 belonged to a different work order and was excluded. DS-44 was over range with no reportable value; it remained rejected rather than being converted into a guessed number.

Deliverable

A structured review with a complete 209-row ledger, control totals, row dispositions, calculation checks, threshold comparisons, conflicting-draft audit, and a plain-English findings summary.

Skills demonstrated

Multi-file data reconciliation	Dataset QA	Row-level classification
Holding-time calculation	Percentile verification	Unit conversion
Quality-control interpretation	Data lineage	Evidence-based exception handling

WHY IT MATTERS

The domain is specialized, but the transferable work is universal: establish authority, preserve lineage, verify calculations, classify every row, and prevent an appealing draft conclusion from outranking the underlying data.

Source context: Completed AI evaluation assignment using simulated environmental and laboratory records; no regulator, chemist or engineering role is implied.

AI Evaluation Design

Multi-file prompts, rubrics & gold standards

Designed failure-revealing multi-file tasks, explicit scoring criteria, reference calculations, edge cases, and grading guidance across three complex domains.

3 documented multi-file evaluation assignments	24 source files across the three assignments (7 + 7 + 10)	2 each model runs reviewed per assignment	Multiple high-severity traps, reference calculations and output checks
---	---	--	--

Overview

The strongest evaluation tasks do not reward confident prose; they require a model to reconcile evidence correctly. I developed and reviewed task structures in medication reconciliation, trauma resuscitation, and water-quality data QA, each built around realistic source conflicts and an auditable expected answer.

Challenge

A useful rubric must distinguish an answer that merely sounds plausible from one that reconciles the files, computes the correct values, excludes bad evidence, follows the output contract, and surfaces unprompted failure modes. It also must separate major reasoning failures from minor presentation issues.

Source material

- Task prompts and domain-bounded role instructions
- Synthetic multi-file source packages in PDF, DOCX, CSV, XLSX and TXT formats
- Reference calculations and expected-answer guidance
- Two model responses per task plus evaluator write-ups
- Output-file and visible-response consistency checks

My work

- Defined the objective, decision boundaries, source hierarchy and required output structure for complex tasks.
- Embedded adversarial but realistic discrepancies that forced identity, timing, scope and calculation checks.
- Built reference answers around exact expected values rather than vague quality descriptions.
- Specified critical failure conditions, acceptable partial credit and lower-severity formatting deductions.
- Reviewed multiple model runs, independently recomputed key values, and compared saved artifacts against visible responses.
- Refined task guidance where a model exposed ambiguity or an unintended shortcut.

Analytical approach

- 01** Choose a real decision
Anchor the task in a concrete output - verify orders, reconcile a resuscitation, or classify a results ledger.
- 02** Build source tension
Introduce mismatched identity, conflicting drafts, duplicate-like records, stale references or competing clocks.
- 03** Define a gold standard
Record the exact calculations, inclusions, exclusions and unsupported conclusions expected.
- 04** Weight failure modes
Mark the errors that invalidate the answer as critical and keep style or wrapper issues proportionate.
- 05** Evaluate the evaluator
Recompute values independently and check both the response and any required output file.

Key findings

Identity as a gating criterion	The medication task included a same-surname lab record with a different middle initial, transposed MRN digits and different birth date. The rubric required exclusion before any downstream reasoning.
Two-clock timeline trap	The trauma task separated wall-clock timestamps from offsets anchored at activation and crossed midnight. The gold standard required 41 minutes since activation, 50 since arrival and a nine-unit red-cell tally that excluded an issued-but-not-administered unit.
Authority and method traps	The water task combined an obsolete reference, a superseded draft, analyte-specific holding times, an out-of-scope row and a percentile-method conflict. The rubric rewarded the traceable row ledger, not a generic compliance summary.
Severity calibration	Evaluator notes treated a missed platelet-imbalance implication as a substantive but bounded deduction while treating one sentence outside the required root wrapper as presentation-only. In another run, rounding 0.0104 upward when the task required three-decimal reporting was caught because it changed the classification.

Deliverable

Sanitized task designs, rubric logic, gold-standard calculations, model-run evaluations, and QA notes. Proprietary interface language, hidden instructions, platform IDs and workflow mechanics are not reproduced.

Skills demonstrated

Prompt design	Rubric development	Gold-standard authoring
Adversarial task design	Edge-case design	Failure-mode analysis
Evaluator QA	Reference calculation	Scoring calibration

WHY IT MATTERS

Prompt engineering and AI training depend on tasks that expose meaningful model failures. This work shows that I can design for auditability, create objective grading anchors, and distinguish a flawed answer from a merely imperfect one.

Source context: Completed AI evaluation design assignments using synthetic records; internal prompts and platform mechanics paraphrased or removed.

Medication-Order Reconciliation

Seven-file heart-failure record review

Matched records across care settings, excluded a wrong-patient lab, trended renal data, checked a derived clearance value, and classified 11 active orders.

7 source files across PDF, DOCX, XLSX, CSV and TXT	6 + 1 matched records and one wrong-patient exclusion	11 active inpatient orders classified	3 serial laboratory time points reconciled
--	---	---	--

Overview

This simulated heart-failure assignment tested whether a model could reconcile home, discharge and current medication records while renal function and potassium changed across three laboratory time points. The work centered on record identity, source conflict, calculation verification and consistent order classification.

Challenge

A reassuring scanned lab belonged to a different Margaret Donnelly. Meanwhile, the matched records disagreed about medications, and the newest laboratory values were materially worse than the admission values. Correct reasoning depended on excluding the mismatched record and anchoring every calculation to the latest valid data.

Source material

- Current inpatient medication orders
- Prior discharge summary and home medication list
- Serial basic metabolic panel spreadsheet
- Vital-sign flowsheet and nursing note
- Scanned outpatient lab report belonging to a different patient

My work

- Cross-checked name, middle initial, date of birth and MRN across all seven files before combining data.
- Excluded the mismatched outpatient lab and documented why its normal values could not be used.
- Reconciled medications as continued, changed, added or dropped across home, discharge and current records.
- Trended potassium, creatinine and eGFR across three dates and calculated an estimated creatinine clearance from age, sex, weight and current creatinine.
- Applied the assignment's threshold and interaction rules to all 11 active orders and produced one disposition per order.

- Separated factual record conflict from domain recommendations and kept the final decision with the simulated licensed reviewer.

Analytical approach

- 01** Gate on identity
Do not blend data until patient identifiers agree.
- 02** Choose the current anchor
Use the most recent matched laboratory result for renal and electrolyte logic.
- 03** Reconcile across settings
Compare what the patient reported, what the prior discharge intended and what is active now.
- 04** Check derived values
Recompute clearance and route/dose conversions rather than copying a draft conclusion.
- 05** Classify every order
Provide a consistent, auditable disposition and evidence trail for all 11 orders.

Key findings

Wrong-patient record excluded

Six files matched Margaret A. Donnelly. The scanned report named Margaret R. Donnelly, used a transposed MRN and a different birth date. Its potassium 4.1, creatinine 0.90 and eGFR 72 were excluded because they would have masked the matched inpatient trend.

Deteriorating trend identified

Across 03/16-03/18, potassium rose 4.4 to 4.9 to 5.4, creatinine rose 1.10 to 1.60 to 2.10, and eGFR fell 51 to 33 to 26. The assignment's Cockcroft-Gault check produced approximately 22 mL/min using age 76, female sex, 62 kg and creatinine 2.10.

Medication-history contradiction

Lisinopril and metformin had been explicitly discontinued at the prior discharge but reappeared through an unreliable home list and became active again in the current orders.

Interaction and duplication logic

The active list combined lisinopril with sacubitril/valsartan despite the assignment's 36-hour washout rule, and stacked four potassium-raising agents while potassium was rising.

Route, formulary and dose checks

The review also identified IV-to-oral conversion questions, a non-formulary statin switch, a renal dose ceiling issue and omissions from the prior regimen.

Deliverable

A structured reconciliation log with identity status, source-by-source medication comparison, current laboratory anchors, 11 order classifications, calculation notes, discrepancies, and a concise reviewer message.

Skills demonstrated

Record identity validation	Medication-list reconciliation	Multi-source QA
Trend analysis	Derived-value verification	Rule-based classification
Contradiction detection	Structured decision logging	

WHY IT
MATTERS

The transferable strength is not clinical authority; it is disciplined reconciliation. The work shows how I prevent record contamination, choose the correct evidence anchor, verify calculations, and turn conflicting files into a complete classification set.

Source context: Completed AI evaluation assignment using simulated medical records. This work does not represent patient care or licensed clinical practice.

Trauma Timeline & Transfusion Reconciliation

Seven-source midnight case

Rebuilt a crossing-midnight timeline, distinguished issued from administered products, verified ratios and thresholds, and recalculated weight- and rate-based values.

7 source records across PDF, CSV, XLSX and TXT	2 time systems reconciled across midnight	9 red-cell units actually administered	9 : 4 : 2 reconciled red-cell : plasma : platelet-dose ratio
--	---	---	--

Overview

This simulated emergency-trauma assignment required a single coherent view from records that used wall-clock time, offset time, issue time and administration time. I reconciled the encounter, quantified what was actually administered, and checked whether the model's calculations matched the source records.

Challenge

The main traps were additive-looking numbers that represented the same products, an issued unit that never reached the patient, two prehospital units absent from the hospital issue log, estimated versus measured weight, and early laboratory values that looked safer than the latest results.

Source material

- EMS run sheet and nursing note
- Trauma-bay vital-signs flowsheet
- Medication administration record
- Blood-bank component issue log
- Anesthesia/rapid-infuser record
- Serial laboratory spreadsheet

My work

- Confirmed all seven records belonged to the same simulated encounter.
- Mapped offset timestamps to wall clock and calculated elapsed time across 23:56 to 00:37.
- Reconciled unit identifiers across EMS, blood bank and administration records.
- Separated issued, transfused, autologous and device-cumulative quantities to prevent double counting.
- Computed component ratios and tested two different massive-transfusion thresholds.

- Recomputed a weight-based dose, an infusion rate and the significance of the latest - not earliest - laboratory values.

Analytical approach

- 01** **Normalize the clocks**
Keep arrival, activation and offset time distinct, then map them to one crossing-midnight timeline.
- 02** **Define what counts**
Count only administered allogeneic products; keep issued-not-administered, autologous and device totals separate.
- 03** **Reconcile identifiers**
Match each product across issue and administration sources and add the separately documented EMS units once.
- 04** **Recompute against the governing input**
Use the measured 95 kg weight, the correct INR tier and the latest laboratory draw.
- 05** **Test conclusions against exact definitions**
Distinguish a formal 10-unit threshold from a three-units-per-hour criterion.

Key findings

Crossing-midnight timeline

The latest entry was T+41 at 00:37. That was 41 minutes after activation at 23:56, 50 minutes after arrival at 23:47, and about 81 minutes after the estimated injury time.

Administered products, not issued products

The correct red-cell total was 9: two prehospital units plus seven in the trauma bay. An eighth hospital unit was still in a cooler and was excluded. The 4,350 mL rapid-infuser number was a device total of itemized products, not an additional volume.

Thresholds kept distinct

The formal ≥ 10 red-cell definition was not yet met at nine units, while the ≥ 3 units in one hour criterion was met. The reconciled product ratio was 9:4:2.

Weight-based dose check

Using the measured 95 kg weight and the assignment's 35 units/kg tier produced 3,325 units. The recorded 2,500-unit administration was 825 units below that reference calculation.

Infusion-rate and recency checks

A 100 mL bag intended to run over eight hours calculates to 12.5 mL/hour, not the documented 100 mL/hour. The review also anchored calcium assessment to the latest 0.78 mmol/L value rather than the earlier 1.05.

Deliverable

A structured encounter review with the normalized timeline, administered-product tally, ratio and threshold logic, dose/rate recomputations, laboratory trends, discrepancy notes, and a prioritized QA summary.

Skills demonstrated

Timeline reconstruction	Cross-system reconciliation	Midnight time calculation
De-duplication	Issued-vs-administered logic	Ratio analysis
Dose and rate verification	Recency-based data selection	

WHY IT MATTERS

This work demonstrates the kind of precision required in operations and model-quality roles: define the counting rule, reconcile multiple clocks, resist double counting, recompute from the right inputs, and state exactly which threshold was - or was not - met.

Source context: Completed AI evaluation assignment using simulated emergency records. This work does not represent emergency care or licensed clinical practice.

Claims Evidence Audit

Denied-date analysis & record exclusion

Audited 14 indexed records across a 19-day denied range, excluded contaminated evidence, separated defensible dates from weak ones, and removed unsupported diagnoses.

14 indexed records across audit, claim and care documentation	19 days in the disputed 01/12-01/30 range	3 records requiring exclusion or reduced evidentiary weight	3 risk tiers for the final findings
---	--	---	--

Overview

This simulated skilled-nursing claims assignment asked whether an appeal narrative was actually supported by the underlying record. I reconciled identity and dates, compared contemporaneous evidence with a flawed draft, separated stronger days from weaker ones, and produced narrow language that did not outrun the source documentation.

Challenge

The package deliberately mixed valid evidence with a wrong resident's hospital record, a same-resident document outside the denied period, a late unsigned addendum and an internal draft containing overstated diagnoses. The date range also changed character: early notes showed measurable skilled needs, while later records became increasingly routine.

Source material

- Audit request and claim summary
- Hospital discharge and skilled-nursing admission records
- Physician/practitioner and nursing notes
- Therapy evaluations and daily logs
- MDS/care-plan excerpt and medication/lab/order records
- Internal draft appeal plus identity, date-range and signature traps

My work

- Matched resident and claim identifiers before using any document.
- Excluded the wrong-resident record and prevented its diagnoses from contaminating the appeal analysis.
- Separated contemporaneous denied-range documentation from an outside-date consult and a late unsigned therapy addendum.
- Reviewed the disputed period day by day for skilled versus routine or custodial content.
- Compared the internal draft's claims with the actual diagnoses, signatures, dates and service records.
- Produced supported appeal language, a missing-document list and high/medium/lower risk rankings.

Analytical approach

- 01** Validate resident and claim
Cross-check name, identifiers, stay dates and denied dates before considering substance.
- 02** Assign evidence weight
Exclude wrong identity, isolate outside-range material and downgrade late or unsigned records.
- 03** Analyze by date
Avoid a single all-or-nothing conclusion for a period whose evidence strength changes over time.
- 04** Audit the draft
Trace every diagnosis and service claim back to contemporaneous records.
- 05** Write narrowly
Recommend only language that survives identity, timing, signature and source-support checks.

Key findings

Record contamination prevented

A hospital record belonged to Ellen M. Barnes, not Eleanor J. Barnes, with different identifiers and dates. It was excluded. A 02/06 wound consult was after discharge and outside the denied range; a 02/10 therapy addendum was late and unsigned.

Date-specific support

The strongest support was 01/12-01/16, when records documented anticoagulation adjustment, cardiopulmonary monitoring during oxygen weaning, wound care, and skilled gait, endurance, transfer, ADL and safety work with measurable deficits.

Later period was weaker

The evidence from 01/17-01/30 - especially 01/19 onward - shifted toward routine medications, ADL supervision, stable findings, group exercise, home-program review and discharge preparation, with gaps in daily therapy documentation.

Unsupported draft claims removed

The records did not support claims of sepsis, acute stroke, skilled speech therapy for dysphagia, IV antibiotics through discharge, a stage 3 ulcer during the stay, wound-vac care, oxygen dependence through discharge or hands-on help with all ADLs through 01/30.

Documentation risk made explicit

Missing or late signatures, no signed recertification after 01/15, and delayed discharge documentation were separated from the underlying evidence of skilled need.

Deliverable

A concise appeal-support review identifying defensible and weak dates, excluded evidence, unsupported statements, timing/signature/coding concerns, missing records to obtain, narrowly supported language, and a three-level risk ranking.

Skills demonstrated

Claims-document QA	Identity reconciliation	Date-range analysis
Evidence inclusion/exclusion	Unsupported-claim detection	Risk classification
Draft-to-source audit	Research synthesis	

WHY IT MATTERS

The assignment maps closely to data-quality and research work: establish which records are admissible, weight evidence by recency and reliability, identify narrative overreach, and communicate a conclusion that is useful precisely because it is bounded.

Source context: Completed AI evaluation and data-quality assignment using simulated health and claims records. No patient care, legal representation or licensed claims determination is implied.

Medical Record Evidence Synthesis

ENT referral timeline, findings & evidence gaps

Aligned a referral, medication list, imaging report and laboratory data; separated confirmed findings from suspected conditions; and preserved missing-evidence boundaries.

4 unique record or deliverable types	3 supporting formats: DOCX, PDF and CSV	24 CBC/CMP analyte rows in the surviving lab file	Multiple duplicate copies identified and normalized
--	---	---	---

Overview

This simulated ENT-focused assignment involved a smaller but different kind of quality problem: turning records created for different purposes into a concise referral narrative without converting suggestive evidence into a confirmed diagnosis. The surviving file set includes a completed referral deliverable and three supporting record types, plus duplicate exports.

Challenge

The records spanned different dates and purposes. A brain MRI contained only incidental sinus information, the medication list warned that outside changes might be missing, and the referral referenced a dedicated sinus CT whose report was not preserved in the available portfolio source set.

Source material

- Active medication list generated 12/04/2025
- Brain MRI report from 08/27/2025
- CBC/CMP laboratory export from 03/18/2026
- PCP-to-ENT referral letter dated 02/28/2026
- Duplicate DOCX, PDF and CSV exports of the same underlying records

My work

- Matched patient identity across the surviving records and removed duplicate copies from the scope count.
- Reconstructed the sequence of treatment history, evolving symptoms, office examination and referral request.
- Separated what the imaging established from incidental findings and from the clinician's suspected condition.
- Preserved the distinction between a rigid nasal examination and the unperformed laryngeal assessment.
- Kept later laboratory abnormalities in their proper time context and did not use them as proof of causation.
- Documented the missing CT report as an evidence boundary rather than inventing CT findings.

Analytical approach

- 01 **Normalize the record set**
Identify duplicate exports and count unique information sources only once.
- 02 **Build the chronology**
Order prior imaging, medication history, treatment attempts, new symptoms, examination and later labs by date.
- 03 **Label evidence type**
Separate reported symptoms, observed examination findings, imaging findings, treatment history and proposed next steps.
- 04 **Distinguish fact from suspicion**
Do not turn concern for chronic rhinosinusitis or a possible laryngeal component into a confirmed diagnosis.
- 05 **Name what is missing**
Limit the synthesis when a referenced report is not present in the available source set.

Key findings

Imaging scope kept accurate

The MRI was a brain study for headaches. It found no acute intracranial abnormality and only incidental mild bilateral maxillary mucosal thickening; it was not treated as a dedicated sinus evaluation.

Treatment history reconciled

The medication and referral records supported prior intranasal steroid, antihistamine, saline and antibiotic use with incomplete or transient improvement, while also warning that the older primary-care medication list might not capture later changes.

Observed versus unassessed anatomy

The referral documented right middle-meatus discharge, edematous turbinates and a rightward septal deflection. It also explicitly stated that the nasopharynx and larynx were not assessed on that rigid examination.

Later labs not overinterpreted

The 03/18 laboratory file was mostly within its listed ranges, with eosinophils 8.4% and absolute eosinophils 596 flagged high. Because the labs post-dated the 02/28 referral, the portfolio does not present them as the cause of earlier symptoms.

Missing evidence preserved

The referral says a dedicated sinus CT was completed on 03/14, but that report is not in the surviving source set. No CT-specific finding is attributed in this case study.

Deliverable

A PCP referral letter that consolidated symptoms, prior treatment, observed sinonasal findings, limits of the office examination, and the requested specialist evaluation without asserting a final diagnosis. This portfolio adaptation adds a clear source-scope and evidence-gap note.

Skills demonstrated

Medical-record synthesis	Timeline reconstruction	Duplicate normalization
Fact-versus-inference separation	Evidence-gap detection	Cross-file identity matching
Plain-language research summary		

**WHY IT
MATTERS**

This case shows a different evaluation strength: concise synthesis under uncertainty. I can combine records written for different purposes while preserving chronology, source limitations and the difference between established findings and plausible next questions.

Source context: Completed work sample using simulated records and fictional people. This work does not represent patient care or licensed clinical practice.

RECRUITER INDEX

Strongest work by role

The same evidence can support different job families. These are the most direct starting points for each recruiter.

ROLE	START HERE
AI evaluator / model quality	LLM Response Evaluation; AI Evaluation Design; Water-Quality Dataset QA
Data analyst	Water-Quality Dataset QA; Trauma Timeline & Transfusion Reconciliation; Medication-Order Reconciliation
Data quality / QA	Water-Quality Dataset QA; Medication-Order Reconciliation; Claims Evidence Audit
Prompt engineer / AI trainer	AI Evaluation Design; LLM Response Evaluation; Claims Evidence Audit
Research analyst	Claims Evidence Audit; Medical Record Evidence Synthesis; Water-Quality Dataset QA
Operations / data analyst	Trauma Timeline & Transfusion Reconciliation; Medication-Order Reconciliation; Water-Quality Dataset QA

Core positioning

I work through unfamiliar subject matter by making the evidence structure explicit: which source governs, what can be combined, what must be excluded, which calculation controls the decision, and where the available information stops.

**VIEW THE LIVE
PORTFOLIO**

[Open the recruiter-ready website](#)