SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

ResEd: A Platform for Radiology Resident Education

Steven Rothenberg, MD, Thomas Jefferson University; Paras Lakhani, MD; Shahir Monsuruddin, MD; Baskaran Sundaram, MD; Trevor Vent, PhD; Adam Flanders, MD, FSIIM

Briefly describe what attendees will see during your demo.
Live demo of the ResEd platform and an overview of the development and deployment process

Problem & Motivation: What clinical or operational problem does your MVP address?
Residents write preliminary reports and attendings finalize them. The edits between the two (missed findings, changed impressions, critical misses) are where residents learn the most. Traditionally that happened at the workstation, but rising RVU demands and the shift toward remote reading increasingly separate residents from attendings, so the direct feedback loop often doesn't happen. Reviewing prelim-vs-final across thousands of exams by hand isn't practical, so the feedback rarely reaches the resident, and program directors lack per-resident data on how trainees can improve.

What You Built: Describe your MVP, tool, or workflow in plain language
ResEd ingests paired preliminary and final reports and runs an LLM comparison that scores each exam and flags the ones where the attending's final meaningfully changed the prelim. The main output is a resident dashboard that surfaces those flagged exams first, so a trainee sees what changed and why it mattered without scrolling through every read. Each resident sees only their own exams, with counts, a clinical-agreement score and trend, and a per-finding sensitivity/specificity tracker over time. A weekly email digest pushes each resident their recent flagged exams and score movement, so they don't have to remember to check in. Users sign in through a five-role login (Flask-Login with scrypt, optional Microsoft SSO, and passwordless magic links). For admins, a security-health dashboard runs automated checks (TLS, secrets, PHI-in-logs scans) at startup and daily. Inputs are RIS report feeds (JSON-URL, a read-only source-DB view, or the PowerScribe 360 API). Raw dictation is de-identified at import and never stored.

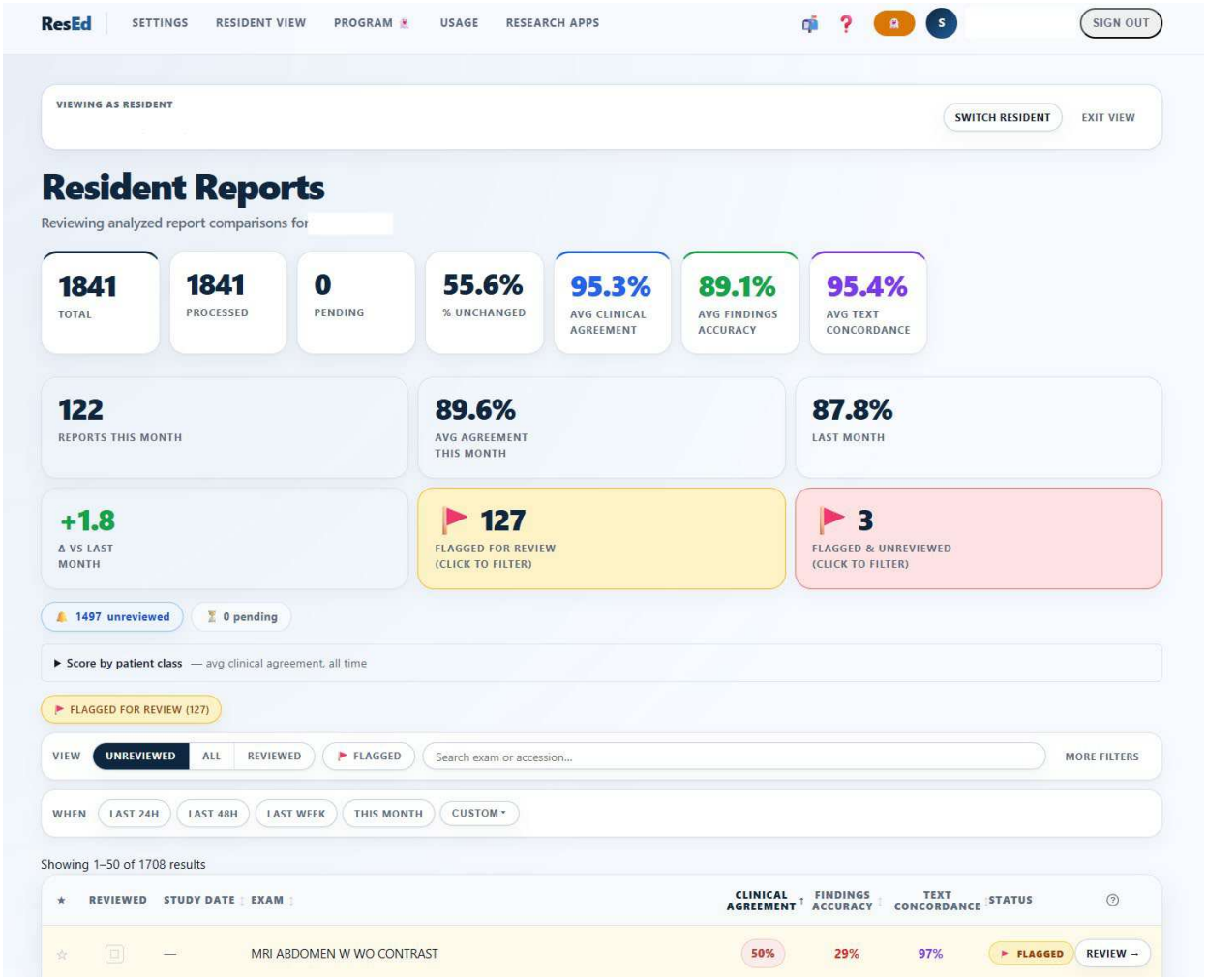
Tech Stack & Development Approach: What tools, frameworks, or methods did you use to build it?
Python 3.11 / Flask, PostgreSQL (SQLAlchemy Core), Gunicorn, Apache/TLS. LLM analysis runs against Azure OpenAI or local Ollama, switchable at runtime. Auth is Flask-Login with scrypt, optional Entra ID SSO, and magic-link login, over a five-role RBAC model. The application was mostly written using Claude Code, across 1,395 commits, with a DESIGN_HISTORY.md capturing 63 dated architecture and privacy decisions and 40 specs. Development happens in a separate environment; code is pushed to GitHub and Azure DevOps. CI runs bandit, semgrep, detect-secrets, a coverage floor, and PHI-egress tests. Beyond that, the hospital security team audits the codebase with Mend (SCA/SAST), and the live configuration is penetration-tested with Burp.

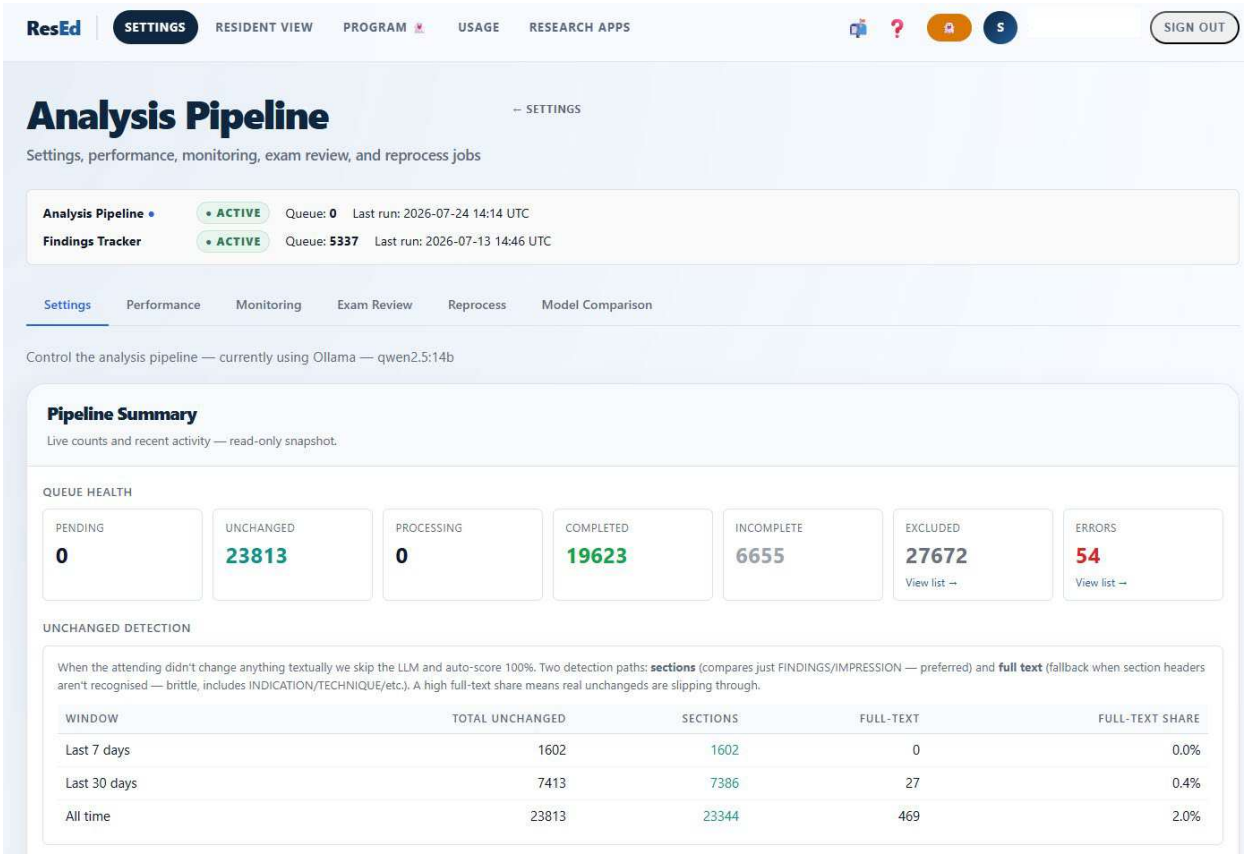
Validation & Evidence: Have you done any informal testing or validation? If so, what did you find?
The system has been running in production against live data since April 17, 2026. Daily RIS imports have received 75,170 reports to date, of which 43,409 exams have been processed across 40 residents. IT has completed two formal security audits of the application. Testing also includes a Postgres-backed pytest suite with security and PHI-exposure gates. We have not yet run a formal accuracy study of the LLM's discrepancy detection against radiologist adjudication.

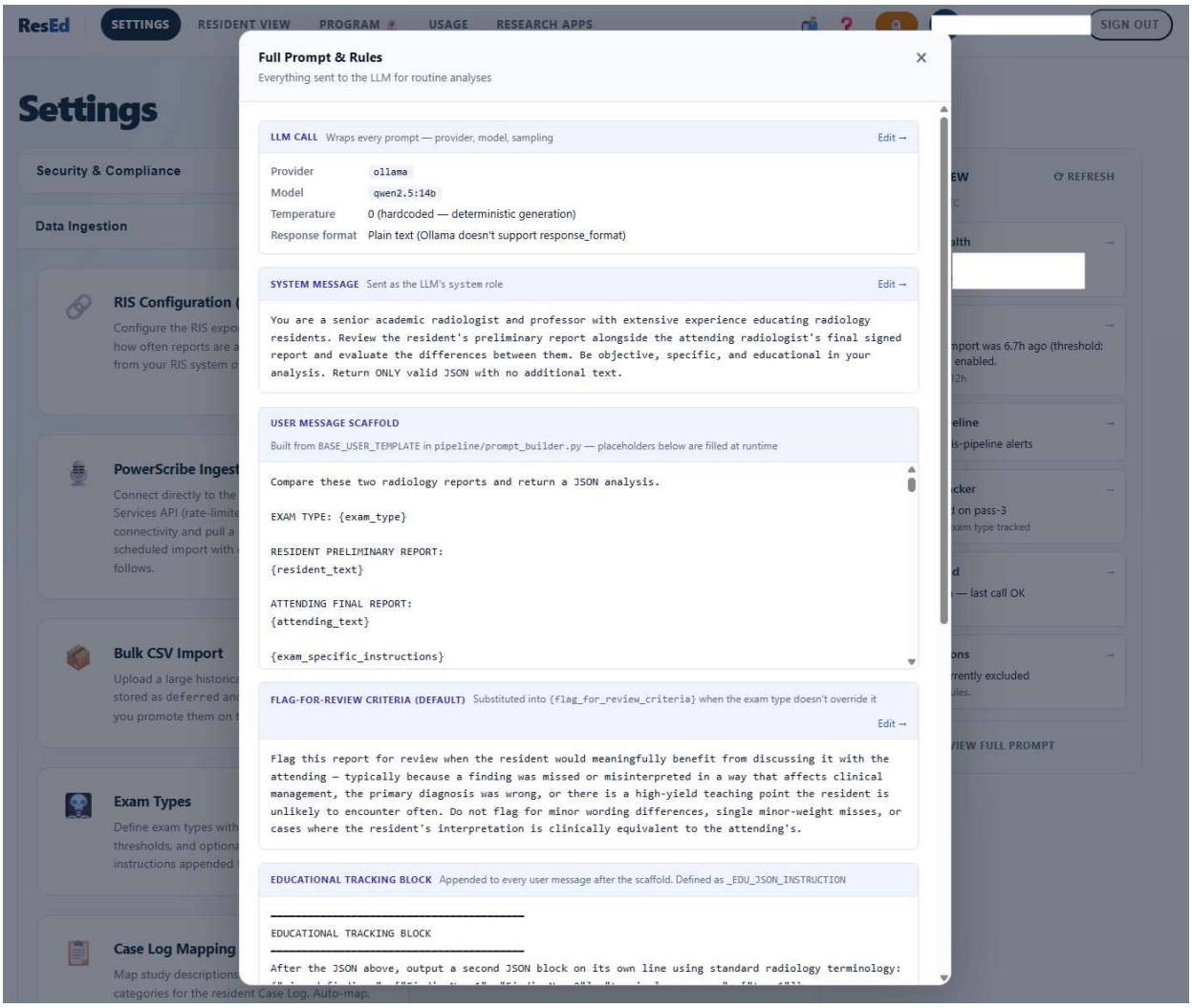
Feedback You're Seeking: What specific feedback, help, or collaboration are you hoping to get from the CAIMI26 community?
Mainly interest and help. If anyone wants to try it, we're happy to share the code, and we'd welcome contributors on new features.

Known Limitations & Honest Failures: What isn't working yet, or what have you already tried that failed?
Ingestion is the main weakness. We currently pull from an administrative database that collects preliminary and final reports, but it misses draft preliminary reports, and those exams score as a perfect match when in reality no prelim was ever captured. We can pull data directly from PowerScribe 360, but haven't yet configured the audit-trail ingestion that would recover those missed drafts. We're working on a more direct design to fix this.

Potential for Impact: If this MVP were fully developed and deployed, who would benefit and how?
The immediate benefit is to trainees, who get objective feedback on their reads over time. The platform is also built to expand beyond education: its RBAC foundation supports other areas such as AI peer review and quality, which we're already working on.


ResEd
SETTINGS
RESIDENT VIEW
PROGRAM
USAGE
RESEARCH APPS
S
SIGN OUT
VIEWING AS RESIDENT
SWITCH RESIDENT
EXIT VIEW
Resident Reports
Reviewing analyzed report comparisons for
1841
TOTAL
1841
PROCESSED
0
PENDING
55.6%
% UNCHANGED
95.3%
AVG CLINICAL AGREEMENT
89.1%
AVG FINDINGS ACCURACY
95.4%
AVG TEXT CONCORDANCE
122
REPORTS THIS MONTH
89.6%
AVG AGREEMENT THIS MONTH
87.8%
LAST MONTH
+1.8
Δ VS LAST MONTH
127
FLAGGED FOR REVIEW (CLICK TO FILTER)
3
FLAGGED & UNREVIEWED (CLICK TO FILTER)
1497 unreviewed
0 pending
Score by patient class — avg clinical agreement, all time
FLAGGED FOR REVIEW (127)
VIEW
UNREVIEWED
ALL
REVIEWED
FLAGGED
Search exam or accession...
MORE FILTERS
WHEN
LAST 24H
LAST 48H
LAST WEEK
THIS MONTH
CUSTOM
Showing 1–50 of 1708 results
REVIEWED
STUDY DATE
EXAM
CLINICAL AGREEMENT
FINDINGS ACCURACY
TEXT CONCORDANCE
STATUS
MRI ABDOMEN W WO CONTRAST
50%
29%
97%
FLAGGED
REVIEW →


ResEd
SETTINGS
RESIDENT VIEW
PROGRAM
USAGE
RESEARCH APPS
S
SIGN OUT
Analysis Pipeline
← SETTINGS
Settings, performance, monitoring, exam review, and reprocess jobs
Analysis Pipeline • ACTIVE Queue: 0 Last run: 2026-07-24 14:14 UTC
Findings Tracker • ACTIVE Queue: 5337 Last run: 2026-07-13 14:46 UTC
Settings
Performance
Monitoring
Exam Review
Reprocess
Model Comparison
Control the analysis pipeline — currently using Ollama — qwen2.5:14b
Pipeline Summary
Live counts and recent activity — read-only snapshot.
QUEUE HEALTH
PENDING 0
UNCHANGED 23813
PROCESSING 0
COMPLETED 19623
INCOMPLETE 6655
EXCLUDED 27672
View list →
ERRORS 54
View list →
UNCHANGED DETECTION
When the attending didn't change anything textually we skip the LLM and auto-score 100%. Two detection paths: sections (compares just FINDINGS/IMPRESSION — preferred) and full text (fallback when section headers aren't recognised — brittle, includes INDICATION/TECHNIQUE/etc.). A high full-text share means real unchangeds are slipping through.
WINDOW
TOTAL UNCHANGED
SECTIONS
FULL-TEXT
FULL-TEXT SHARE
Last 7 days 1602 1602 0 0.0%
Last 30 days 7413 7386 27 0.4%
All time 23813 23344 469 2.0%


ResEd
SETTINGS
RESIDENT VIEW
PROGRAM
USAGE
RESEARCH APPS
SIGN OUT
Settings
Security & Compliance
Data Ingestion
RIS Configuration
PowerScribe Ingest
Bulk CSV Import
Exam Types
Case Log Mapping
REFRESH
Full Prompt & Rules
Everything sent to the LLM for routine analyses
LLM CALL Wraps every prompt — provider, model, sampling
Edit →
Provider ollama
Model qwen2.5:14b
Temperature 0 (hardcoded — deterministic generation)
Response format Plain text (Ollama doesn't support response_format)
SYSTEM MESSAGE Sent as the LLM's system role
Edit →
You are a senior academic radiologist and professor with extensive experience educating radiology residents. Review the resident's preliminary report alongside the attending radiologist's final signed report and evaluate the differences between them. Be objective, specific, and educational in your analysis. Return ONLY valid JSON with no additional text.
USER MESSAGE SCAFFOLD
Built from BASE_USER_TEMPLATE in pipeline/prompt_builder.py — placeholders below are filled at runtime
Compare these two radiology reports and return a JSON analysis.
EXAM TYPE: {exam_type}
RESIDENT PRELIMINARY REPORT:
{resident_text}
ATTENDING FINAL REPORT:
{attending_text}
{exam_specific_instructions}
FLAG-FOR-REVIEW CRITERIA (DEFAULT) Substituted into {flag_for_review_criteria} when the exam type doesn't override it
Edit →
Flag this report for review when the resident would meaningfully benefit from discussing it with the attending — typically because a finding was missed or misinterpreted in a way that affects clinical management, the primary diagnosis was wrong, or there is a high-yield teaching point the resident is unlikely to encounter often. Do not flag for minor wording differences, single minor-weight misses, or cases where the resident's interpretation is clinically equivalent to the attending's.
EDUCATIONAL TRACKING BLOCK Appended to every user message after the scaffold. Defined as _EDU_JSON_INSTRUCTION
EDUCATIONAL TRACKING BLOCK
After the JSON above, output a second JSON block on its own line using standard radiology terminology:
VIEW FULL PROMPT