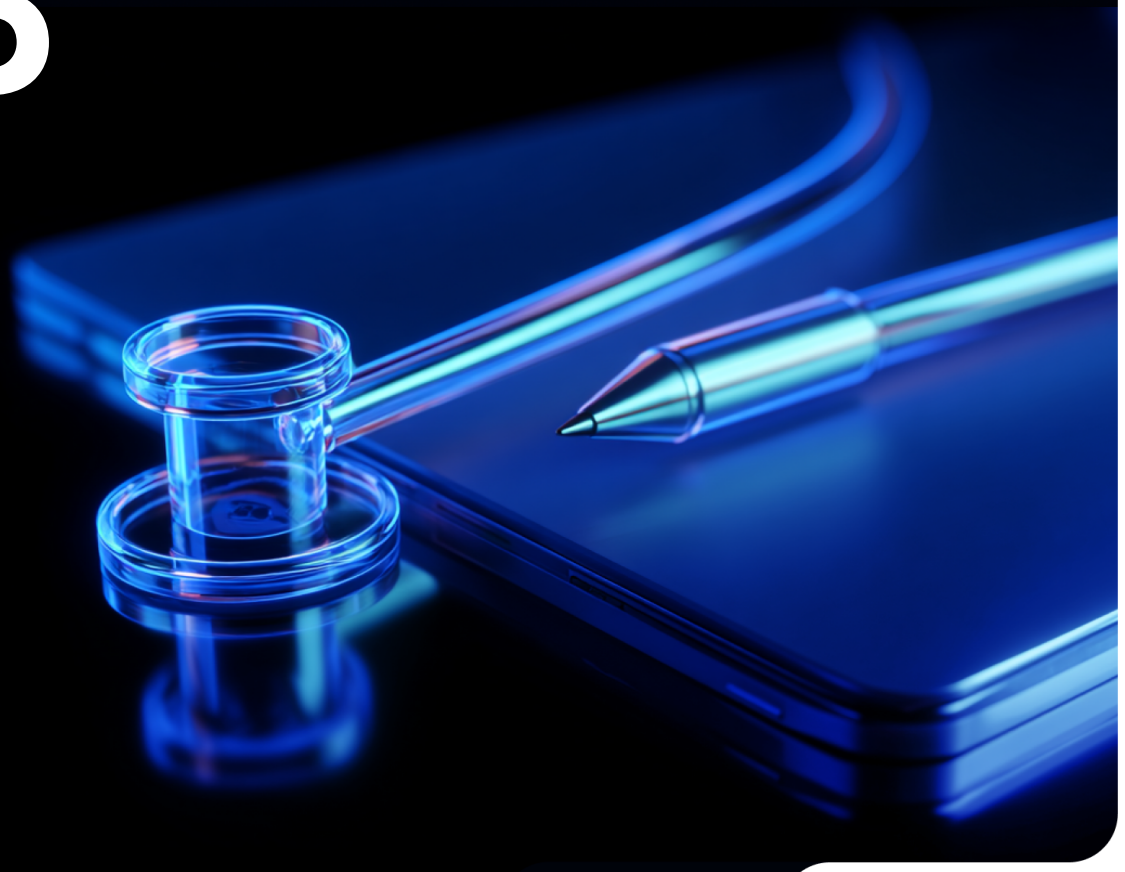
Case Study

REDUCE DOCUMENTATION BURDEN WITH AI-POWERED WORKFLOW

A healthcare client optimized clinical documentation for scale and accuracy

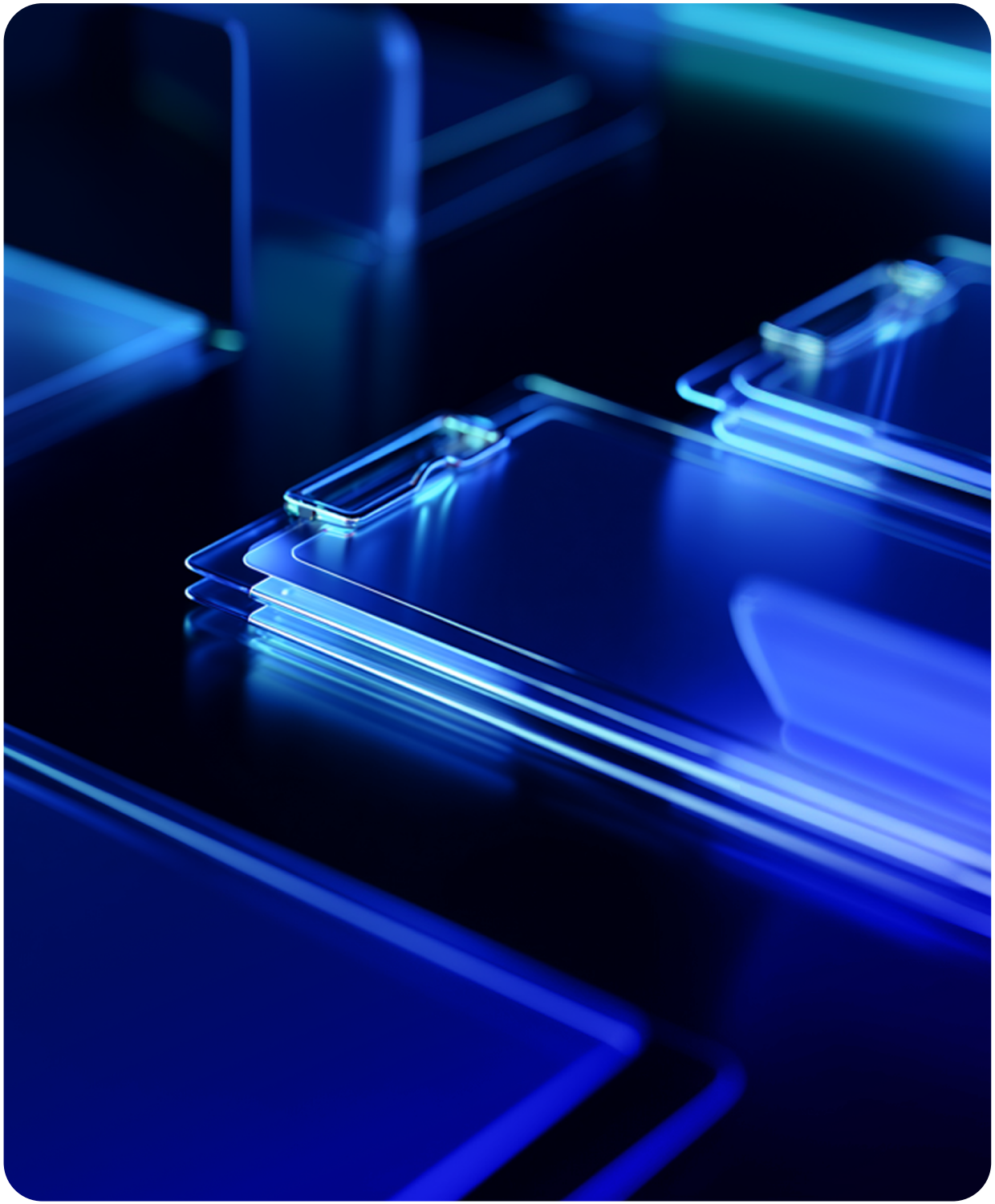


softserve

Managing kidney care at scale demands precision and time. For a nationwide healthcare provider serving over 200,000 patients, clinical documentation had become a major obstacle to efficiency. The most time-intensive artifact was the **comprehensive health evaluation (CHE)**, a multi-section summary consolidating lab results, medications, diagnoses, comorbidities, and care plans into a single longitudinal view.

Preparing one CHE required nearly four hours of detective work across multiple PDFs, EHR screens, and scanned faxes. With **8,000 CHEs due monthly**, this translates into **32,000 clinician hours**. That time is equivalent to **40 full-time nurse practitioners focused solely on paperwork**. Beyond the **$800,000 monthly labor cost**, the hidden toll was diminishing patient interaction and declining job satisfaction.

The goal was to **reclaim time for caregivers without compromising clinical rigor**. The approach centered on integrating a large language model (LLM) into the workflow to handle repetitive extraction and drafting tasks while keeping final judgment in human hands. This strategy cut CHE preparation time by one hour per case, a **25% improvement** that leads to significant savings and improves patient engagement.



Challenges

Preparing thousands of comprehensive health evaluations each month placed a heavy strain on clinical teams. The process demanded accuracy, speed, and cost control. Yet, existing workflows were slow, labor-intensive, and vulnerable to errors. Addressing these issues required a solution that balanced scale with clinical integrity.

Clinical-grade accuracy

Generative models can produce fluent text yet omit or fabricate critical details. In healthcare, even one missing chronic condition or incorrect lab value can distort care decisions. To mitigate this risk, the solution adopted a human-in-the-loop design: The AI drafts each section, cites sources for every fact, and routes the document back to the nurse practitioner for review and sign-off. This pairing of speed with oversight ensures safety and trust.

High-volume throughput

The operational target was 400 CHEs per business day, each requiring data from 10 PDFs. Early prototypes using local OCR and synchronous LLM calls stalled under load. The architecture evolved into stateless microservices for OCR, chunking, indexing, prompt orchestration, and verification. These services communicate via message queues and scale in the cloud, allowing the system to absorb spikes in patient intake without delays.

Cost justification for enterprise rollout

A productivity tool must deliver measurable savings. By applying retrieval-augmented generation (RAG), batching prompts, and lightweight self-verification, the cost per CHE dropped to about $3 — far below the $100 manual preparation cost. This margin supports broad adoption without constant ROI recalculations.

Solution

The project ran as a two-quarter initiative blending rapid experimentation with structured scale-up. SoftServe's AI experts began by **benchmarking models against real patient documents to balance accuracy, latency, and cost**. Gemini Pro on Google Vertex AI emerged as the preferred model for its ability to process PDF fragments without prior OCR. An MVP was deployed to a pilot group of nurse practitioners, validating clinical fidelity and usability.

Following successful trials, the system was **re-platformed into a microservices architecture**. Each service (OCR, chunking, vector indexing, prompt orchestration, and assembly) operates independently, coordinated through queues for resilience and elasticity. Parallel processing ensures uninterrupted throughput even during maintenance windows.

Specialists in MLOps, data engineering, and UI/UX contributed to production readiness. The clinician-facing console highlights every AI-extracted fact with its source, reinforcing transparency and supporting the review process. The technology stack includes **Google Vertex AI** for managed pipelines, **Gemini Pro** for multimodal reasoning, and a **NoSQL document store** for fast, flexible data handling.

The result is a cloud-native workflow assistant that reduces time spent on each CHE while scaling seamlessly with patient volume.

Value Delivered

Early production metrics confirm substantial impact:

- **One hour saved per CHE**, reducing preparation time from four hours to three.
- **8,000 clinician hours reclaimed monthly**, equivalent to 10 full-time nurse practitioners redirected to patient care.
- **$800,000 monthly labor savings**, projected to exceed $10 million annually.
- Positive staff feedback: Nurses report spending reclaimed time on patient conversations rather than document assembly.

Future plans include extending the platform to other document types, such as dialysis change-over reports and discharge summaries.

Contact SoftServe

Contact SoftServe to learn how and where to implement AI to streamline your clinical workflows with the best ROI.

About SoftServe

SoftServe is a premier IT consulting and digital services provider. We expand the horizon of modern technologies to solve today's complex business challenges and achieve meaningful outcomes for our clients.

Social Links

info@softserveinc.com
www.softserveinc.com

Contact

NORTH AMERICAN HQ

201 W 5th Street, Suite 1550
Austin, TX 78701
+1 866 687 3588 (USA)
+1 647 948 7638 (Canada)

EUROPEAN HQ

30 Cannon Street
London EC4 6XH
United Kingdom
+44 333 006 4341

softserve