SYSTEM 001 / TECHNICAL REPORT / VERSION 1.0

VELA

A consent-controlled healthcare navigation and action system

VELA turns a spoken request for care and insurance evidence into a source-backed, cost-aware path. It clarifies ambiguity, compares feasible options, separates recommendation from authorization, and records every consequential action.

FIRST PLACE WINNER	DO TRACK	NVIDIA SPARK HACK SEATTLE

Prepared by

Abishek Muralikrishna | Abhinav Ravindran | Fatima Aguilar

Decision Systems Lab

Seattle, Washington / August 2026

Prototype boundary

This report documents a hackathon research prototype. Enrollment, coverage transition, and appointment booking use controlled sandbox adapters. VELA does not diagnose, recommend treatment, guarantee prices, or act as a licensed broker.

00 / EXECUTIVE SUMMARY

The system in one page

Healthcare navigation is not a search problem alone. It is a sequence of linked decisions under incomplete information: identify the service, establish coverage, compare total cost, verify feasibility, obtain permission, and complete the next action without losing the evidence trail.

VELA tests whether that sequence can be made legible and safe in one interface. Voice and document inputs become candidate facts. A deterministic Care Journey Engine owns eligibility, network, cost, ranking, consent, and state. Local NVIDIA models interpret and explain, but they do not authorize actions or override application logic.

Core contribution

- A voice-first journey that asks when a procedure is underspecified instead of guessing a billing code.
- Source-backed comparison using hospital price transparency records and insurance context.
- Exact, action-specific consent gates between recommendation and execution.
- Idempotent sandbox adapters and append-only receipts that distinguish completion from assertion.
- Local inference on NVIDIA GB10 for speech, reasoning, and explanation inside a bounded trust environment.

Demonstrated result

In the submitted journey, VELA interpreted a request for a knee MRI without contrast, compared five Seattle hospitals, and surfaced Fred Hutchinson Cancer Center at approximately $299 as the lowest displayed option. The price was read from hospital transparency data, not generated by a model. A second deliberately vague request, 'I need an MRI,' triggered clarification rather than a guessed code.

> The project won First Place in the Do Track at NVIDIA Spark Hack Seattle in August 2026.

01 / PROBLEM AND DESIGN REQUIREMENTS

From uncertainty to controlled action

The system was designed against five requirements. Each requirement is a trust behavior that must remain observable in the interface and enforceable in code.

Requirement	Operational definition	Failure prevented
Clarify	Request missing body part, contrast, or evidence before matching.	Confidently wrong procedure or billing code.
Evidence	Attach source, time, confidence, and verification state to material facts.	Untraceable recommendations and invented certainty.
Compare	Use deterministic eligibility, network, and annual-cost logic.	Model prose silently becoming the decision rule.
Consent	Require exact approval for each consequential action and scope.	Treating analysis or conversation as permission to act.
Prove	Return timestamped, idempotent receipts and retain failures.	Claiming completion without evidence or repeating an action.

Decision principle

A useful healthcare agent should clarify ambiguity, expose evidence, compare feasible options, ask permission, and return proof of action. Model fluency is valuable only when these controls remain intact.

> The goal is not maximum autonomy. It is maximum useful progress within explicit human authority.

02 / SYSTEM ARCHITECTURE

One engine owns the truth

The Care Journey Engine is the authoritative workflow boundary. Agents are bounded capabilities around it. Models can interpret language, propose facts, detect ambiguity, and explain results. Deterministic application code decides what is feasible, ranked, permitted, and complete.

Layer	Component	Responsibility
Input	Voice, chat, camera, PDF	Normalize intent and evidence
Agent layer	Onboarding and extraction agents	Propose facts, classify documents, identify gaps
Evidence	Fact and memory ledger	Append source, time, confidence, and status
Reference	Plan and provider catalogs	Retrieve controlled healthcare records
Authority	Deterministic engine	Eligibility, network, annual cost, ranking, consent, state
Language	Nemotron explanation agent	Explain authoritative results in plain language
Execution	Sandbox adapters	Execute an approved exact scope once
Audit	Audit ledger	Record approvals, transitions, failures, and receipts

03 / DATA AND DECISION LOGIC

Evidence before recommendation

VELA distinguishes a candidate fact from a verified fact. Each material fact carries provenance, timestamp, confidence, and status. An insurance card may identify visible payer and member fields, but it cannot establish a complete benefit design. Cost-sharing logic requires a Summary of Benefits and Coverage or an equivalent source.

Price transparency

The submitted prototype read from approximately 4.2 million federally mandated hospital price transparency records. Procedure matching is constrained by the requested body part and contrast state. Provider comparison retains source context and does not treat a model-generated number as a hospital price.

Annual-cost model

Annual cost = annual premiums + expected out-of-pocket care + medication cost + remaining deductible exposure.

A low procedure price is not automatically the least expensive care path. VELA evaluates feasibility first, then compares complete paths. Ineligible or out-of-network paths cannot win merely because a component price is low.

Authoritative decision sequence

1	Normalize request	Identify service, body part, and contrast requirements.
2	Resolve ambiguity	Ask a targeted question when required information is absent.
3	Retrieve evidence	Load priced services, plans, providers, and verified document facts.
4	Test feasibility	Apply eligibility, network, referral, and evidence completeness rules.
5	Compute and rank	Calculate comparable cost components and rank feasible paths.
6	Explain	Generate a bounded explanation from the authoritative result.

04 / LOCAL AI AND NVIDIA COMPUTE

Models interpret. The engine decides.

The local inference path used in the prototype is: VELA application -> Care Journey Engine -> authenticated Hermes gateway -> NemoClaw/OpenShell policy boundary -> NVIDIA Nemotron 3.5 Lightning -> NVIDIA GB10 on an Acer Veriton GN100.

Model responsibilities

- Parakeet ASR converts voice input into text for the care journey.
- Nemotron extracts candidate facts, classifies closed-set intents and documents, normalizes names, detects ambiguity, proposes clarification, and explains authoritative comparisons.
- Magpie TTS converts the bounded response into speech.
- NemoClaw and OpenShell constrain model and network execution at the infrastructure boundary.

Closed-schema boundary

Every Nemotron response is treated as untrusted input and validated against a closed schema. Unknown actions are rejected. One bounded corrective retry is allowed. Repeated failure becomes a visible blocked or operator-review state.

Why local compute matters

- Sensitive care and insurance context can remain inside a local trust boundary.
- Model inference, retrieval context, document processing, and workflow state can operate together on one system.
- Deterministic comparison can continue when the language model is unavailable. Language-dependent steps fail closed or request review.
- Predictable latency supports a responsive voice experience.

> The model never decides eligibility, changes a ranked result, grants consent, advances workflow state, or calls an action adapter by itself.

05 / CONSENT, SAFETY, AND AUDIT

Analysis is not permission

VELA treats consent as a state transition, not a conversational implication. Each consequential action presents its exact scope. The user approves the action, target, and relevant data transfer before an adapter becomes callable.

Condition	Control	User-visible result
Missing consent	Block adapter	Explain scope and wait
Wrong workflow stage	Block adapter	Return current state and required prior step
Duplicate request	Do not repeat action	Return the existing idempotent receipt
Adapter failure	Never claim completion	Record failure and permit safe retry
Invalid model schema	Reject output	Retry once, then request operator review
Unavailable explanation service	Preserve deterministic result	Label explanation unavailable or retry locally

Receipt model

A successful sandbox action returns a timestamp, action type, exact approved scope, workflow state, idempotency key, and outcome. The audit ledger is append-only: it retains candidate facts, verification changes, approvals, rejected actions, failures, state transitions, and receipts.

> A completion message without a receipt is not completion.

06 / PROTOTYPE SCOPE AND DEMONSTRATION

What worked, what was sandboxed

The judged journey used a fictional Washington persona and controlled synthetic card, benefit, referral, plan, provider, and action data. No real patient or insurance documents were required.

Working in the prototype

- Voice and chat intake, insurance card and referral extraction, and fact-memory state.
- Procedure clarification, real hospital price lookup, three-plan deterministic comparison, and eligibility gating.
- Responsive interface across Home, Paths, Appointments, Documents, and Preferences.
- Provider verification through a controlled endpoint or call, exact consent ledger, live webhook delivery, and auditable receipts.
- Local speech recognition, reasoning, and speech generation on NVIDIA GB10.

Sandboxed

- Plan enrollment submission.
- Existing coverage transition.
- Appointment booking.

The refusal moment

When asked only, 'I need an MRI,' VELA requested the missing body part and contrast information. This behavior was intentional. During testing, a loose search for 'knee' once matched a hardware implant instead of an imaging procedure. The incident established a core invariant: missing procedure detail must trigger clarification, not confident completion.

> The strongest technical result was not the answer. It was the refusal to invent one.

07 / EVALUATION AND LIMITATIONS

A research prototype, not a finished care platform

The hackathon evaluation emphasized end-to-end behavior: intake, branching, error recovery, evidence, deterministic comparison, consent, action, and proof. The implementation included 55 passing deterministic and vertical-slice tests at submission.

Evaluation dimensions

- Correct clarification behavior for underspecified procedures.
- Deterministic cost and feasibility outputs under controlled inputs.
- Consent blocking, state-transition validation, and duplicate-action recovery.
- Visible evidence and receipts across the user journey.
- Local model execution through the authenticated NVIDIA inference route.

Current limitations

- Hospital prices require normalization and should be verified before a real care decision.
- Coverage estimates depend on complete, current plan evidence and cannot guarantee a final patient bill.
- Provider availability and appointment slots change outside the prototype.
- Real enrollment, coverage transition, and booking require licensed, authenticated, and contractually approved pathways.
- The system is not medical advice and does not diagnose or recommend treatment.

Public launch requires security review, privacy controls, regulatory analysis, licensed partner workflows, production data agreements, observability, and human escalation paths.

08 / NEXT EXPERIMENTS

From winning prototype to validated system

The next phase should test whether the same trust behaviors hold when data coverage, user needs, and operational partners become more varied.

Phase	Experiment	Evidence sought
1	Data expansion	Extend normalized procedure and hospital transparency coverage across Washington and Massachusetts.
2	Matching validation	Create expert-reviewed benchmarks for procedure, code, facility, and price matching.
3	Coverage model	Validate deductible, coinsurance, network, medication, and annual-cost logic across representative plan designs.
4	Trust study	Measure whether sources, caveats, refusal behavior, and exact consent improve comprehension and calibrated confidence.
5	Operational pilots	Integrate licensed partners for verification and booking while retaining action-specific approval and receipts.
6	Production readiness	Add identity, encryption, access controls, privacy review, monitoring, incident response, and human escalation.

Research thesis

Healthcare agents earn trust not by hiding uncertainty, but by making the path from evidence to decision to consent to action inspectable.

> VELA is an experiment in that path. Its winning result is a starting point for the harder question: can a system remain helpful, understandable, and controllable when it leaves the sandbox?

Document basis

This report reflects the verified VELA prototype and submission state from August 2026. Product claims distinguish working behavior, controlled data, and sandbox execution. Hospital transparency references describe the dataset used in the demonstration and do not constitute a price guarantee.