May 2025

Mauricio E. Tano, Ph.D.
Staff Scientist

Preliminary Incursions into a RELAP5-3D Agent

An AI Tool to Assist with RELAP5-3D Analyses

Battelle Energy Alliance manages INL for the
U.S. Department of Energy's Office of Nuclear Energy

Idaho National Laboratory

Vision and Motivation

CONTEXT

RELAP5-3D is the U.S. Department of Energy's industry-standard code for nuclear thermal-hydraulics safety analysis. It supports licensing submittals, design-basis accident calculations, and operational margin analysis across light-water and advanced reactor programs worldwide.

Authoring an input deck requires mastery of Volume 2, Appendix A, over 1,000 card types with strict column-positional format, arcane keywords, and component-specific parameter conventions. A single simulation may take days to weeks of expert effort.

The gap between physical understanding and RELAP5-3D execution is an analyst-expertise bottleneck that limits the pace of nuclear design and safety analysis.

VISION

"Enable any qualified engineer to design, execute, and interpret RELAP5-3D simulations through natural conversation — without writing a single Appendix A card by hand."

Accessibility

Remove the Appendix A syntax barrier. Let the analyst focus on physics, not card formatting.

Reliability

Preserve rigor through deterministic compilation. Every deck is validated before it reaches the cluster.

Productivity

Automate HPC submission, SLURM monitoring, error correction, and result plotting end-to-end.

Transparency

Session persistence, explicit validation reports, and readable Python code make every step auditable.

The agent does not replace the analyst; it removes the syntax barrier between the analyst and the physics.

Technical Challenges in Development

Volume 2 , Appendix A Card Format

- Over 1,000 card types with strict column-positional format.
- No modern tooling; arcane abbreviations.
- LLMs hallucinate card numbers, field counts, and separator characters, causing silent deck corruption.

LLM Hallucination in Numerical Domains

- Units should be exact on based on user request.
- Component numbering (100–999), IC dictionary key capitalization, and string-type selectors are all common LLM error sources with no obvious runtime warning.

Knowledge Base from Proprietary PDFs

- RELAP5-3D manual volumes contain domain vocabulary absent from LLM pre-training.
- PDFs have complex multi-column layouts, tables, and figures.
- Proprietary and export-controlled content cannot be distributed, and the RAG must encode knowledge without exposing raw text.

Python Front-End API Design

- 30+ component types, each with distinct IC dict formats and key conventions.
- Pump and pressurizer require raw card strings for complex parameterizations.
- API must be 100% card-correct.

HPC Integration & Job Lifecycle

- SSH BatchMode (no interactive passwords), SLURM partition/time-limit configuration, squeue/sacct polling, and graceful Ctrl-C detach-and-resume.
- File upload and download over SCP.
- Cluster-specific module loading.

Automatic Error Correction

- RELAP5-3D runtime error messages require parsing line numbers and card identifiers from unstructured output.
- Backed LLMs must generate targeted minimal fixes without regressing working cards, across up to 3 correction rounds per job submission.

Architecture Design Principles

1. Deterministic Before Generative

- The structured compiler path (SystemPlan → relap_python) writes no LLM tokens into Appendix A cards.
- The LLM is used only for planning intent and template-parameter inference but never for raw card generation.

2. Phase Separation & Clear Boundaries

- Each of the four phases is independently resumable via session.json. Phase transitions are explicit checkpoints.
- A failure in Phase 3 does not corrupt the Phase 2 deck.

3. Type-Driven Compilation

- User intent is first crystallized into a validated *Pydantic SystemPlan* before any code is generated.
- Validation runs at three levels: plan topology → relap_python code syntax → compiled deck cards.

4. Hybrid Retrieval

- Dense vector search (voyage-3-large, 1024-dim) combined with BM25 lexical search via Reciprocal Rank Fusion (k=60).
- Card vocabulary terms (e.g., 'TMDPVOL', 'sngljun') benefit from exact-match BM25 weighting.

5. Transparent Failure

- The structured path fails explicitly rather than silently falling back to freeform generation.
- Validation reports are shown to the user before HPC submission.
- An honest 'cannot compile' is better than a corrupted deck.

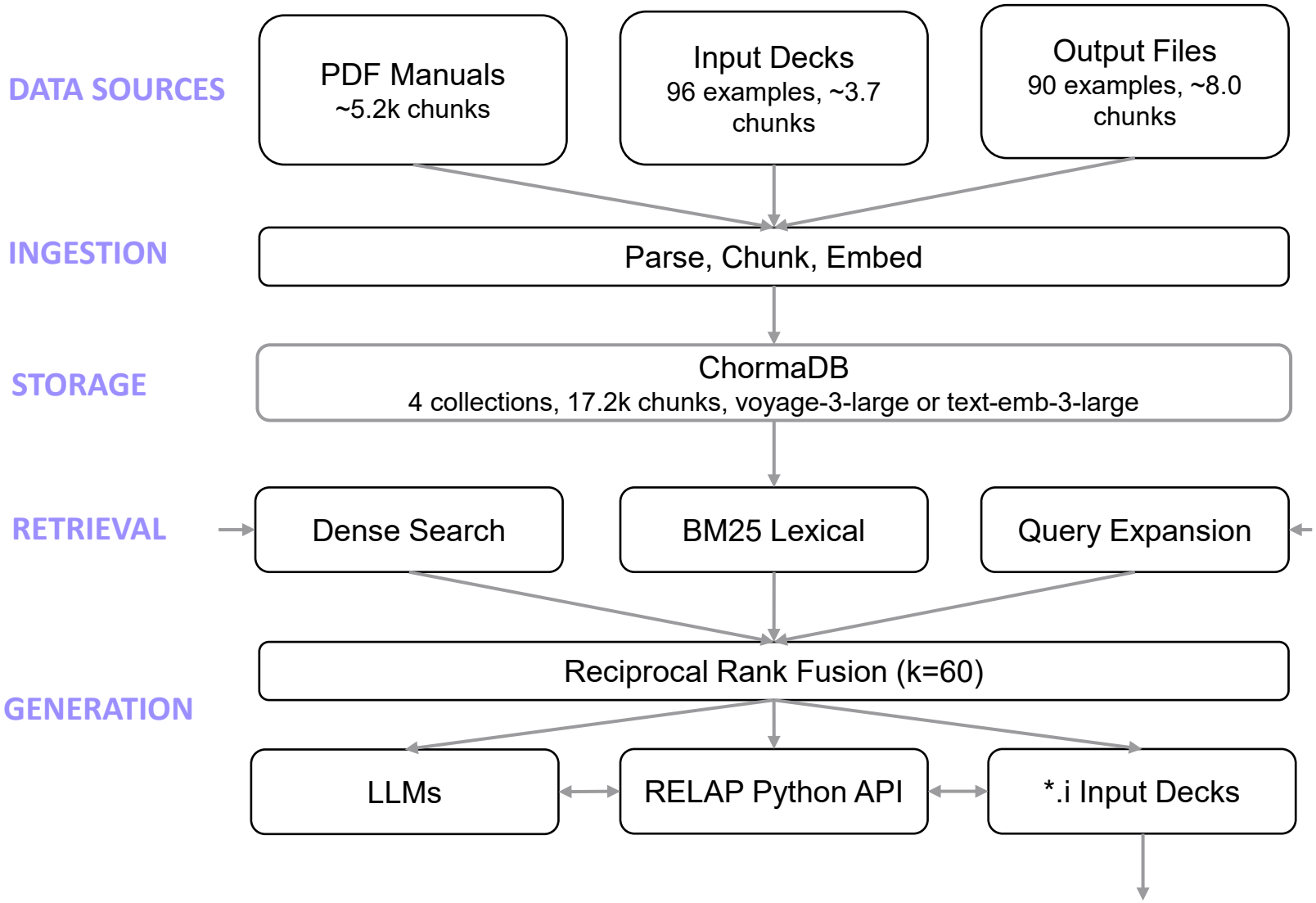
6. Session Persistence as First-Class Feature

- Long HPC workflows (queue wait + 2-hour run + correction) demand interrupt-and-resume.
- Every phase saves complete state as readable JSON.
- Manual inspection and field editing are intentionally supported.

7. Minimal External Process Dependencies

- SSH/SCP via system commands (no paramiko, etc.).
- Python-based for c130post only, auto-detected via Homebrew paths.
- No Docker, no custom daemons.
- The agent runs on any Mac/Linux with SSH keys configured.

RAG Knowledge Base: Ingestion & Retrieval Pipeline



TECHNICAL DETAILS

Chunking

Token-aware: 512 tok / 80 overlap (PDFs),
256 tok / 60 overlap (decks). tiktoken cl100k_base.

Embeddings

voyage-3-large (1024-dim, recommended)
or text-embedding-3-large (3072-dim, OpenAI).
Batch size 96.

Lexical

rank_bm25 with custom tokeniser.
Card keywords (e.g., 'TMDPVOL') preserved unsplit.
Full stop-word list.

Fusion

Reciprocal Rank Fusion k=60.
Dense weight 0.6 + BM25 weight 0.4.
Top-20 chunks passed to LLM.

Store

ChromaDB persistent client.
relay 4 named collections.
JSONL + BM25 index as backup.

Scale

~17 160 total chunks.
Ingestion: ~5 min (voyage API).
Query latency: ~120 ms per request.

The RELAP5-3D Python API Library

Hydrodynamic Components (11)

pipe · singleVolume · timeDependentVolume
singleJunction · timeDependentJunction · branch
pump · pressurizer · separator · accumulator · valve

Heat Transfer & Kinetics (4)

heat_structure · thermal_materials
reactor_kinetics · kinetics

Control & Output (7)

general_table · control_system · trips · control_variable
plot_requests · strip_requests · minor_edit_requests

Advanced Components (9)

multid · jetmixer · feedwater_heater · turbine
compressor · ecc_mixer · annulus · radionuclide
multiple_junction · pvm_coupling_junction

Core Infrastructure

relap_cards_common/{hydro,heat,control,thermal}.py — card builders
roundtrip.py — parse existing .i → Python repr → regenerate
manual_registry.json — 1.1 MB machine-readable card specs

Agent Skills and Domain Knowledge Injection

relap-geometry-closure

- Ensures hydraulic self-consistency across component network
- References: geometry-checklist.md
- Verifies: area, volume, length, and junction orientation
- Trigger: any card with area= or length= parameters

relap-model-authoring

- Guides component selection and overall deck structure
- References: manual-roadmap.md · modeling-guidelines.md · repo-roadmap.md
- Maps physical intent to RELAP5-3D component types
- Trigger: phase 1→2 transition; deck authoring session

relap-python-api

- Enforces correct relap_python API usage patterns
- References: card-number-conventions.md · common-patterns.md
- Captures capitalisation rules, IC dict formats, pump raw-card exceptions
- Trigger: any relap_python import or component construction

relap-regression-hpc

- Maintains discipline in HPC test submission and regression tracking
- References: debugging-discipline.md · error-patterns.md
- Documents the 8 most common root causes of job failure with verified fixes
- Trigger: phase 3 entry; error correction loop

relap-thermal-controls

- Validates heat structure geometry, material properties, and BC types
- References: heat-transfer-modes.md · thermal-control-checklist.md
- Covers cylindrical HS area convention (L_per_node, not 2πrL)
- Trigger: heat_structure construction or heat_structure BC assignment

Testing and Validation Framework

Unit Tests ~38 KB — Mocked — no API keys or SSH

test_agent.py

Covers all four phases with dependency injection. Validates RELAPSpec extraction, session JSON round-trip, backend selection logic, and editor command parsing.

relap_python Round-Trip 60+ tests — No data/ folder needed

relap_python/tests/test_deck_*.py

parse_deck(regenerate(fixture)) == fixture for each of 96 official sample problem decks. Catches card-format regressions immediately. ~96 fixture files committed to repo.

Deterministic Compiler Plan→deck — Requires API keys

test_deterministic_compiler.py

Full SystemPlan → relap_python code → compiled .i chain. Validates topology, IC dicts, card numbering, and deck completeness for structured generation path.

Phase 2 Backends Struct + Free — Requires API keys

test_phase2_backends.py

Structured and freeform generation for standard reactor types. Checks that generated code is syntactically valid and that the deck passes structural_validator checks.

Error Correction 3 rounds — Requires Anthropic key

test_reliability_redesign.py

Simulates RELAP5-3D error output and validates that Claude Opus generates a corrected deck in ≤ 3 rounds without regressing cards that were already correct.

HPC Regression End-to-end — Requires cluster access

test_hpc_generated_decks.py

Full pipeline: generate → SSH upload → sbatch → sacct monitor → download results. Uses marker hpc. Validated on bitterroot and teton clusters at INL.

MCP Server & Service 7 tools — No API keys

test_relap_mcp_*.py

All 7 MCP tools tested with valid and invalid inputs. Validates Pydantic schema enforcement, artifact store path isolation, and deterministic deck output.

RAG Retrieval Quality Precision@K — Requires API keys

test_hybrid_rag.py

Retrieval evaluation: annotated query set with known-relevant chunk ground truth. Measures Precision@5, Recall@10, and MRR for both dense-only and hybrid modes.

Lessons Learned from Development & HPC Testing

CARD FORMAT **"Too many numbers" is the #1 LLM error**

LLMs consistently append extra fields to component header cards (e.g., card 1200000). Solution: relap_python wrappers — never allow the LLM to write raw Appendix A cards directly.

STRUCTURED vs. FREEFORM **Structured path is 3× more reliable models**

For well-characterized reactor types (PWR, BWR, NatCirc), the structured compiler produces a valid deck on first attempt >90% of the time. Freeform LLM generation requires on average 1.8 correction rounds.

IC DICT FORMAT **Capital Pressure/Temperature keys are critical**

pipe IC dicts require 'Pressure' (capital P) and 'Temperature' (capital T). singleVolume uses the same keys but without the 'initial_conditions' selector. timeDependentVolume is time-indexed. Wrong keys → silent zero-initialisation.

RAG QUALITY **Better chunking beats more chunks**

Switching from fixed-size to token-aware chunking with metadata annotations (section, component_type, case_name) improved Precision@5 from 0.52 to 0.81. The metadata fields are critical for BM25 weighting.

TRIP NUMBERING **Trip 501 must use the logical trip band**

RELAP5-3D reserves trip numbers 500–600 for internal use. User trips must be numbered 20600500+. Logical trips must use the 20601NNN band. This was the cause of 3 separate HPC job failures.

SESSION PERSISTENCE **Interrupt-and-resume is essential in practice**

HPC queue wait + job runtime + download routinely spans 2–6 hours. Without session.json, every agent restart loses context. Users detach with Ctrl-C multiple times per workflow.

HEAT STRUCTURE AREA **Cylindrical HS: AREA = L_per_node, not 2πrL**

RELAP5-3D multiplies the supplied area by 2πr internally for cylindrical heat structures. Supplying the full surface area causes a factor-of-2πr error in heat flux. The correct value is the axial length per mesh node.

SEPARATOR TOPOLOGY **Duplicate junction in separator → fatal deck error**

A separator component cannot have a junction connecting back to itself even indirectly through a branch. The system_plan.py topology validator now catches this before deck generation.

Live Demonstration

Idaho National Laboratory

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy. INL is the nation's center for nuclear energy research and development, and also performs research in each of DOE's strategic goal areas: energy, national security, science and the environment.