

Human-in-the-Loop AI for Requirements-to-Risk-Control Traceability in Medical Device Systems Engineering

Esteban M. Solórzano Zeledón



COLORADO STATE
UNIVERSITY



September 22-23, 2026 | Washington, DC
Theme: Speed and Innovation through AI4SE & SE4AI

CASE STUDY | COGNITIVE ASSISTANT

**Boston
Scientific**
Advancing science for life™

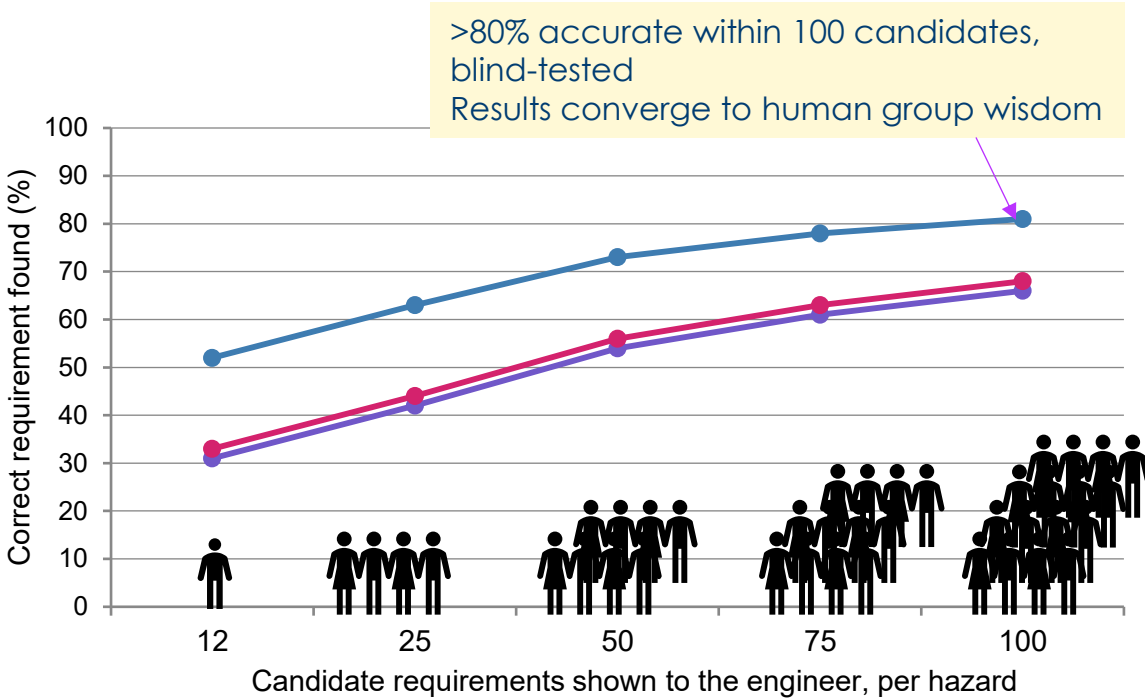
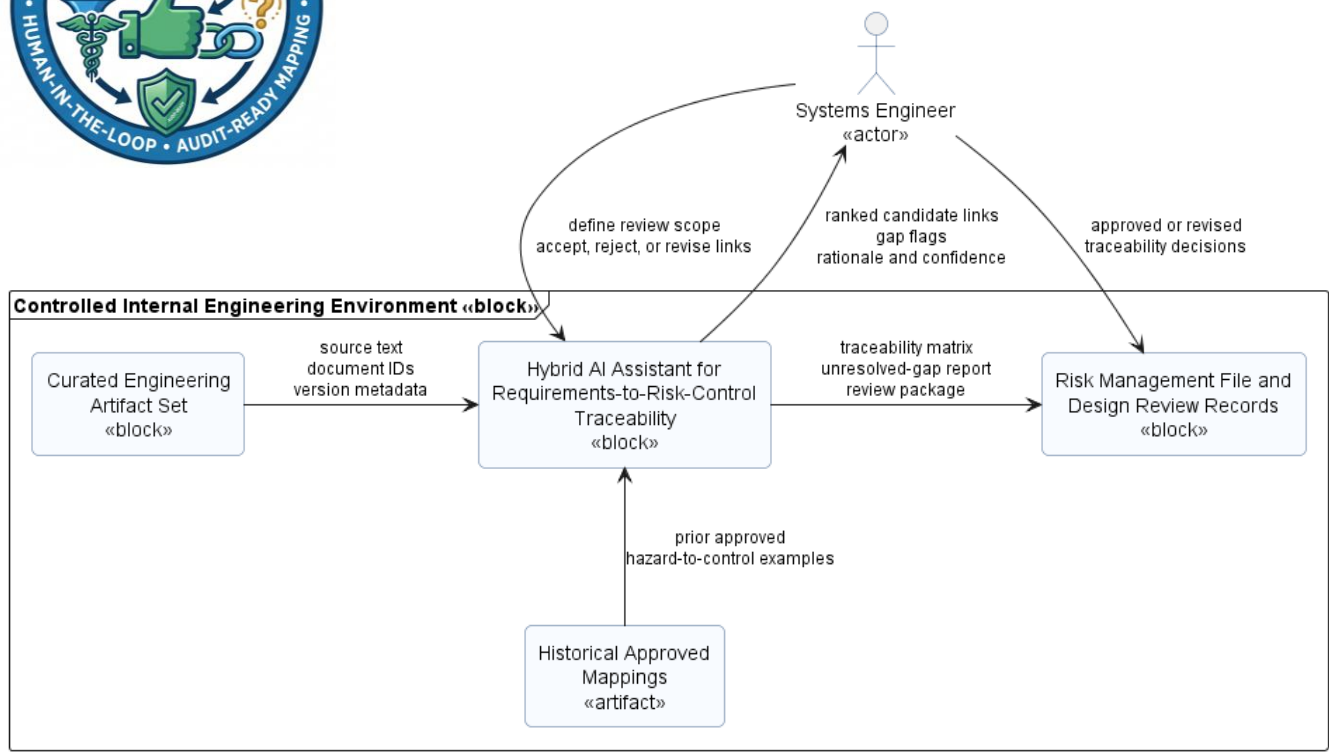




Hybrid AI Assistant for Risk-Control Traceability



Speed and Security do not need to be in conflict



AI architectures tested

— Risk Trace (TF-IDF + LLM) — AI embeddings (MiniLM)
— AI embeddings (MPNet)



Presentation Outline

1



Mission need:
Why traceability breaks?

2



Bounded assistance:
How Risk Trace works?

3



Evidence and transfer:
What it proves?



Mission need → Traceability problem → Human-AI workflow → Evidence

What is a medical device?



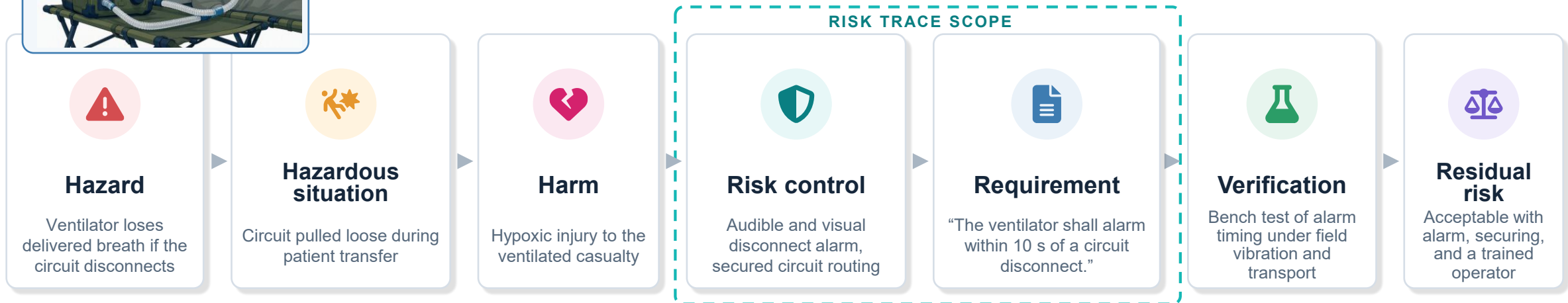
An instrument, apparatus, machine, implant, in vitro reagent, software function, component, or accessory intended to diagnose, cure, mitigate, treat, prevent, or affect body structure/function - without achieving its primary purpose through chemical action or metabolism.

Statutory Source: Section 201(h) of the FD&C Act (21 U.S.C. § 321(h)).



Intended use defines the medical device and everything that must be proven

Risk control thread



 **Gap type 1:**
a hazard has no mapped risk-control requirement.

 **Gap type 2:**
a requirement is unused or lacks risk rationale.

 **Approved trace:**
link, rationale, provenance, and reviewer status



A link is only useful if it carries rationale, provenance, and a reviewer decision.



In plain terms



Every hazard needs a control

- Hazard: anything about a device that could hurt someone: an alarm that fails, a battery that dies mid-transport, a confusing setup screen.
- For each hazard: design a control and write it as a requirement.



Every control needs a paper trail

- Regulators & reviewers ask: show me the hazard, the requirement that controls it, and the test that proves it works.
- That chain of links is traceability, and it is the evidence a device is safe to accomplish mission.



The links are the hard part

- Risk control pieces live in separate documents written by separate teams in different words
- Finding the right match is slow manual searching, and a broken link can hide a hazard nobody controlled.



Finding candidate links is a search problem. Deciding they are adequate is an engineering judgment.



Research question



Can AI help a systems engineer find and organize plausible hazard-to-requirement links while preserving human authority over the final traceability decision?



In scope

- Requirements definition into hazard analysis
- Text and semi-structured engineering artifacts
- Candidate trace links, confidence, rationale, and gaps
- Human approval and exportable review artifact



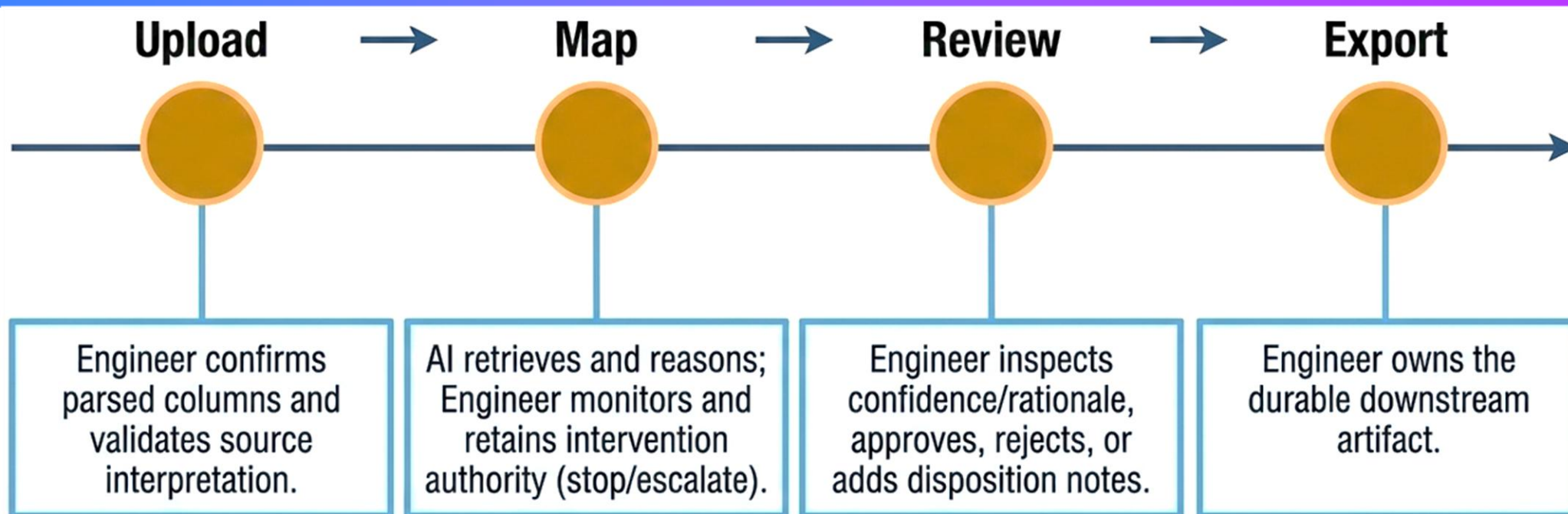
Not in scope

- Generate or approve final requirements
- Declare a risk control adequate
- Accept residual risk
- Close a quality-system action
- Act as a field-deployed medical device function



Question bounded on purpose: help the search, not the approval.

Strategy: Recall over precision & Human in the “control” loop



False negative — invisible:

Plausible control relationship is never inspected
A weakly controlled hazard or missing verification obligation can stay hidden.

False positive — visible:

Engineer spends time rejecting a weak candidate and the queue gets noisier,
The error is caught whenever human approval is required.



A missed link is invisible, a weak candidate is visible. Tune for recall, review hard
Systems engineer remains the approval authority.



AI method selection and the Risk Trace architecture

Supervised ML / deep learning **DEFERRED**

Needs a large, clean history of approved links that does not exist yet.

Rule-based matching **NOT ALONE**

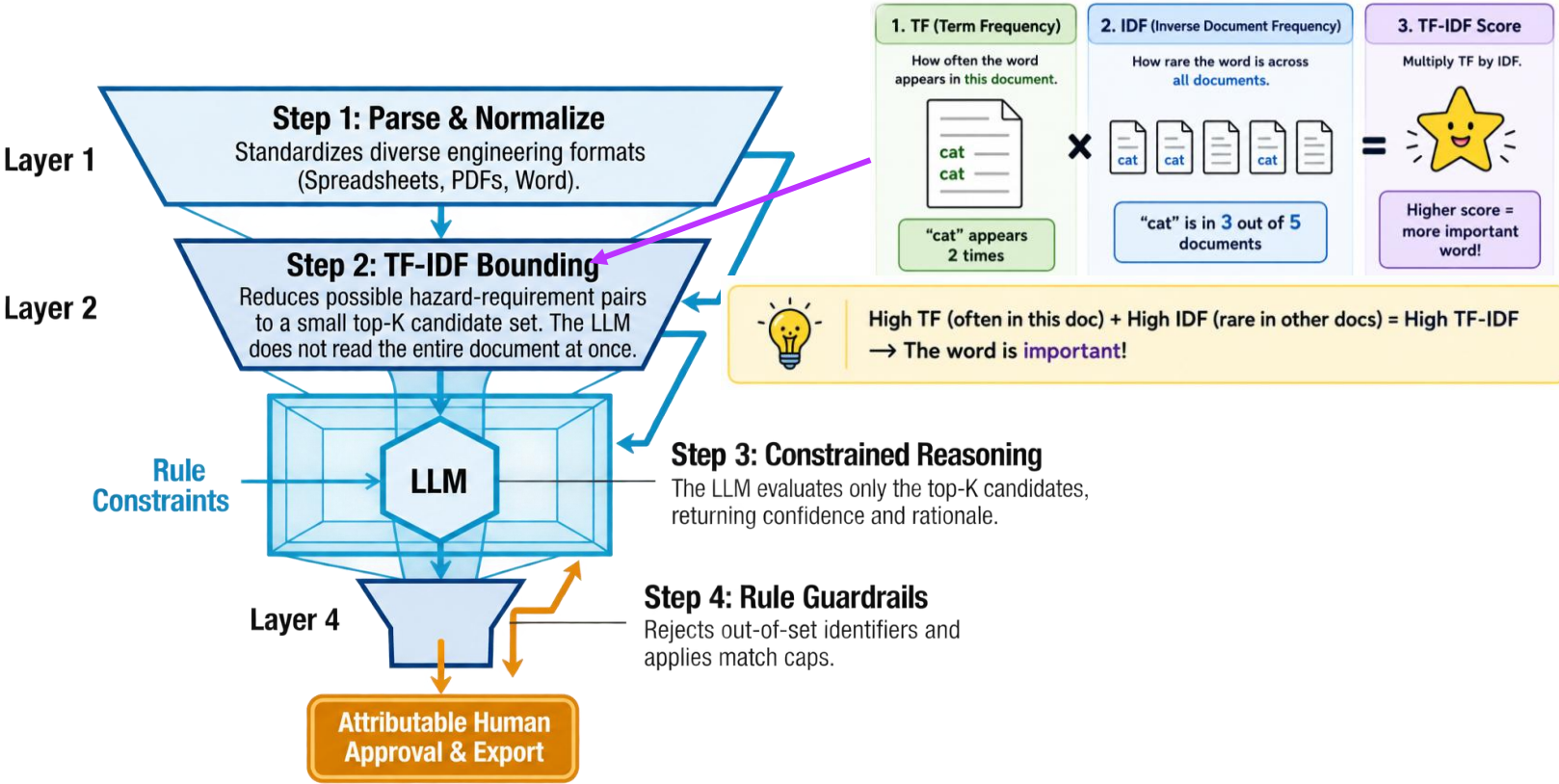
Strong on identifiers, brittle when two teams describe the same hazard differently.

Agentic / LLM-only **HIGH RISK**

Fluent, persuasive errors with no repeatable evidence; adds autonomy without validation.

Hybrid (Risk Trace) **SELECTED**

Retrieval for recall, bounded LLM reasoning, rule guardrails, and human approval.



No single method fits: layering retrieval, bounded reasoning, and rules covers each gap and every stage is bounded so the engineer, not the model, holds the decision.

Walkthrough: Hazards & Requirements Mapping Preview

Requirements Documents

Success

337 requirements added from Firmware_Requirements.docx.

Click to upload or drag and drop requirements documents

Supports: .xlsx, .xls, .csv, .docx, .pdf - select multiple files, then confirm each parsed preview

84 requirements parsed

Column Mapping - verify these are correct

Requirement Id

→

ID#

Requirement Text

→

Design Requirement

Requirement Category

→

Requirement Category

Source Rationale

→

Source / Rationale

ID#	DESIGN REQUIREMENT	REQUIREMENT CATEGORY	SOURCE / RATIONALE
RF-HWR2144: Enclosure Flammability Rating 60601-1 [] [Legacy ID: N/A]	The generator console enclosure shall have a flammability rating of FV-2 or better. Traces From: RF-SYR1439, RF-SYR2833 Traces To: RF-DO8492 Related Hazardous Situations: RF-HAZSIT0422 Related Hazards: RF-HAZ0610, RF-HAZ0356 Related Failure Modes: RF-FM1898	Environmental Life-Related Maintainability and Support Reliability	STDP_JEC60601-1_2012_(Edition_3.1)_Clause_11.3 OR STDP_JEC60601-1_2020_(Edition_3.2)_Clause_11.3: Constructional requirements for fire ENCLOSURES of ME EQUIPMENT Enclosures with flammability rating reduces the risk of a hazardous situation.
RF-HWR2145: Internal Insulated Wiring Flammability Rating 60601-1 [] [Legacy ID: N/A]	The insulated wires within the generator console enclosure shall have a flammability rating of FV-1 or better. Traces From: RF-SYR1439 Traces To: None	Environmental Life-Related Maintainability and Support Reliability	STDP_JEC60601-1_2012_(Edition_3.1)_Clause_11.3 OR STDP_JEC60601-1_2020_(Edition_3.2)_Clause_11.3: Constructional requirements for fire ENCLOSURES of

Confirm Requirements

Firmware_Requirements.docx

- 1 Parser returns headers, preview rows, and a proposed mapping.
- 2 Reviewer verifies requirement ID, text, category, and source/rationale fields.
- 3 Analysis remains disabled until both input groups are confirmed.

The tool proposes the field mapping. The reviewer approves it.

Walkthrough: review evidence, approve or reject, then export gaps

131
HAZARDS

114
COVERED

17
GAPS

91/1459
REQS USED

Sort: Gaps first

17 / 131 reviewed

Gaps only

Low confidence

Unreviewed

Approved

Rejected

Hazard Table Evidence

Clear

Next gap

Next unreviewed

Approve All

Export

MAPPED HAZARD

RF-HAZ0240: RF Generator - Loss of Output [\$RFGeneratorCore OR \$Unibipolar]

RF-HAZSIT0362: Less energy is applied than required to form an effective, therapeutic lesion (Fault) [\$RFGeneratorCore]

HARM

HARM0343: Prolonged Procedure - Limited

HAZARD TABLE EVIDENCE

No explicit source fields were mapped for this hazard.

Proposed requirement traces

RF-FWR2600: Open Circuit Impedance []

Firmware_Requirements.docx

If the channel's measures an impedance greater than the Open Circuit value the firmware shall trigger a RF-FWR3609: Single Channel Cut-Off - GX1-SYS-JP condition.

Related Hazards: RF-HAZ0254, RF-HAZ0240

Req Crossref Text Match

WHY THIS WAS PROPOSED

- Requirement metadata explicitly references this hazard or hazardous situation.
- Hazard and requirement text have strong keyword overlap.

LLM RATIONALE

Detecting an open circuit impedance and cutting off the RF output directly prevents the delivery of insufficient energy, mitigating the hazard of an ineffective lesion.

Approved

MAPPED HAZARD

RF-HAZ0246: Use Error - Attentional Failure [\$RFGeneratorCore OR \$Unibipolar]

RF-HAZSIT0346: User accidentally selects unintended UI element [\$RFGeneratorCore]

HARM

HARM0011: Additional Intervention Required - Moderate; HARM0342: Prolonged Procedure - Negligible; HARM0343: Prolonged Procedure - Limited

HAZARD TABLE EVIDENCE

No explicit source fields were mapped for this hazard.

Proposed requirement traces

RF-FWR21593: Intraprobe SUD misuse in GX1 Monopolar and Bipolar Therapy [\$Unibipolar]

Firmware_Requirements.docx

When a RF therapy is started with an Intraprobe probe connected to the output port, firmware shall stop the therapy and trigger an error as specified by RF-SYD7760: Get Detailed Error - GX1_system_design_specification.
Note: If CQM is not plugged in, the CQM warning LED blinks and CQM error is reported. Or else, the Intraprobe SUD misuse error will be reported.
Note: If the impedance value is in the range of 90 Ohm to 110 Ohm in Monopolar and 180 Ohm to 220 Ohm in Bipolar they are considered as test plugs.

Related Hazards: RF-HAZ5256

Text Match

WHY THIS WAS PROPOSED

- Hazard and requirement text have strong keyword overlap.

LLM RATIONALE

This requirement directly prevents unintended therapy by halting it when an incompatible Intraprobe probe is detected, mitigating the use error and associated harms.

☐ Not applicable

Rejected



Every proposed link arrives with its evidence attached and a decision required
An approved mapping records who decided and on what basis.
A rejected mapping is evidence too, and it is kept.



Test and Evaluation Metrics



Recall@K: did we even show it?

- The tool hands the reviewer a short list of K candidate requirements per hazard. Recall@K asks how often the right one made that list.
- Recall@75 of 78% means that in 78 of 100 cases, the correct requirement was somewhere in the top 75 shown.



Precision and recall: a trade

- Precision = how much of what we accepted was right.
- Recall = how much of what was right we found..



F1: the two in one number

Combines precision and recall into a single score between 0 and 100%. Compares methods on the same test set, but says nothing about whether the workflow is safe for its intended use.



These metrics measure the candidate list, not the engineering decision made from it.

Test and Evaluation Outcomes

5 of 5 core workflow criteria met

- ✓ Multi-format ingestion
- ✓ Bounded candidate generation
- ✓ Reviewable AI output
- ✓ Gap surfacing
- ✓ Human approval retained

Blind evaluation output

131

hazards analyzed

Blind evaluation

128

hazards with matches

97.7% coverage

3

hazards with no match

2.3% gap rate

2,059

accepted matches

Pipeline output count

16.1

matches / covered hazard

Average

0.737

average confidence

Model-reported score

54.1%

current best reported F1, blind

Good enough for review preparation, not for autonomous approval.

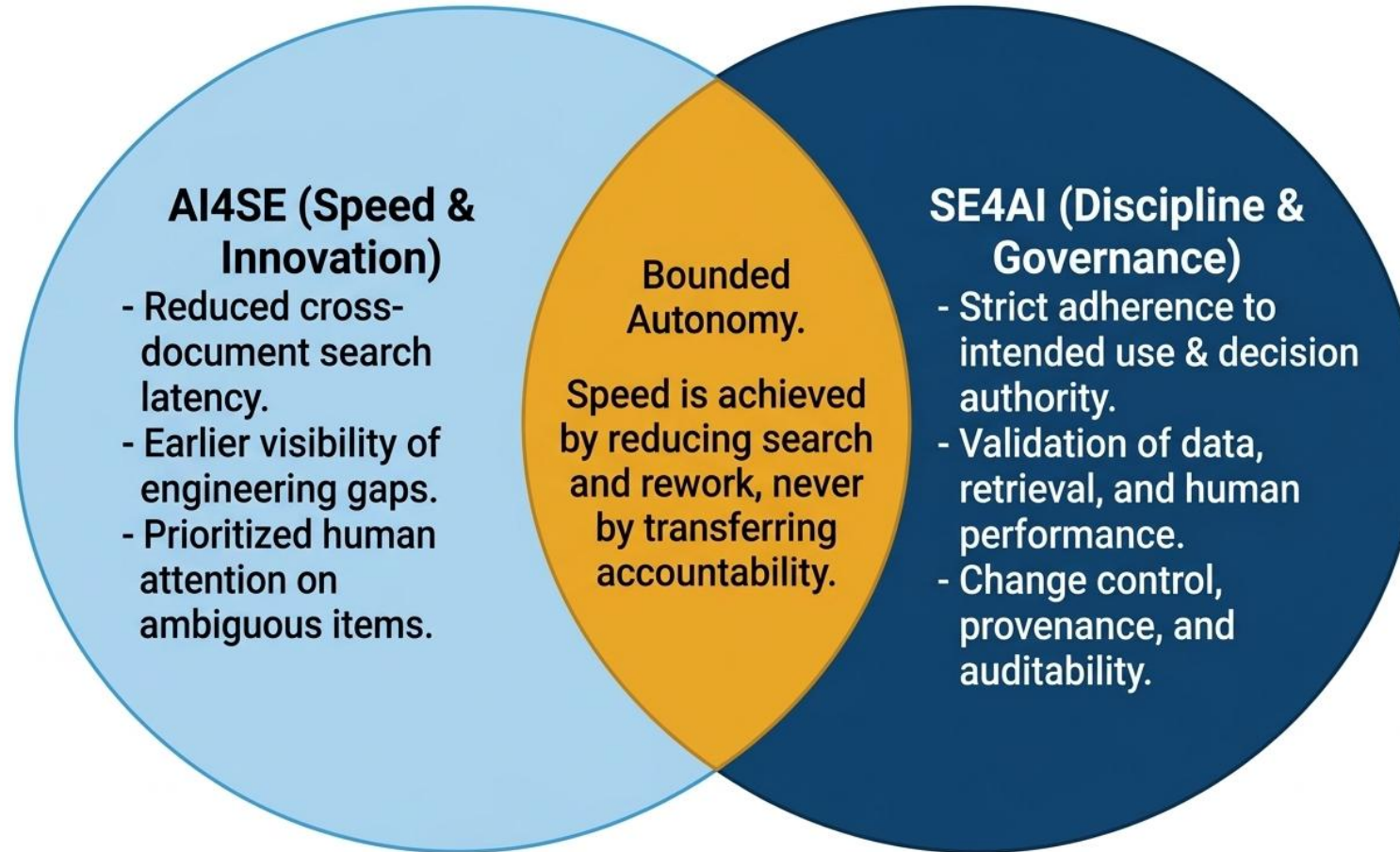
- Functional prototype** **Demonstrated**
- Software verification** **Tested**
- Accuracy benchmark** **Baseline**
- Human-system validation** **Not completed**
- Intended-use validation** **Not completed**



Coverage and confidence describe pipeline output, not correctness. The assistant compresses the search. The systems engineer retains approval authority.

Closing Message

AI4SE speed requires SE4AI discipline



Speed comes from less searching, not from less accountability.
Where in your lifecycle is search burden high, but approval must remain human?



Acknowledgments



Dr. Jeffrey Meganck
Boston Scientific,
Industry advisor



Dr. Steve Simske
Colorado State University,
Hewlett Packard,
Academic advisor



Dr. Steven Conrad
Colorado State University,
*AI-Augmented Systems
Engineering SME*



Lanceng Waise
Boston Scientific,
AI Engineer



You Can Contribute to This Research

Boston
Scientific

Your expertise can help ensure that future medical cyber-physical systems are not just innovative, but safe, interoperable, and regulatory-ready.



Let's Connect:

Esteban.SolorzanoZeledon@bsci.com

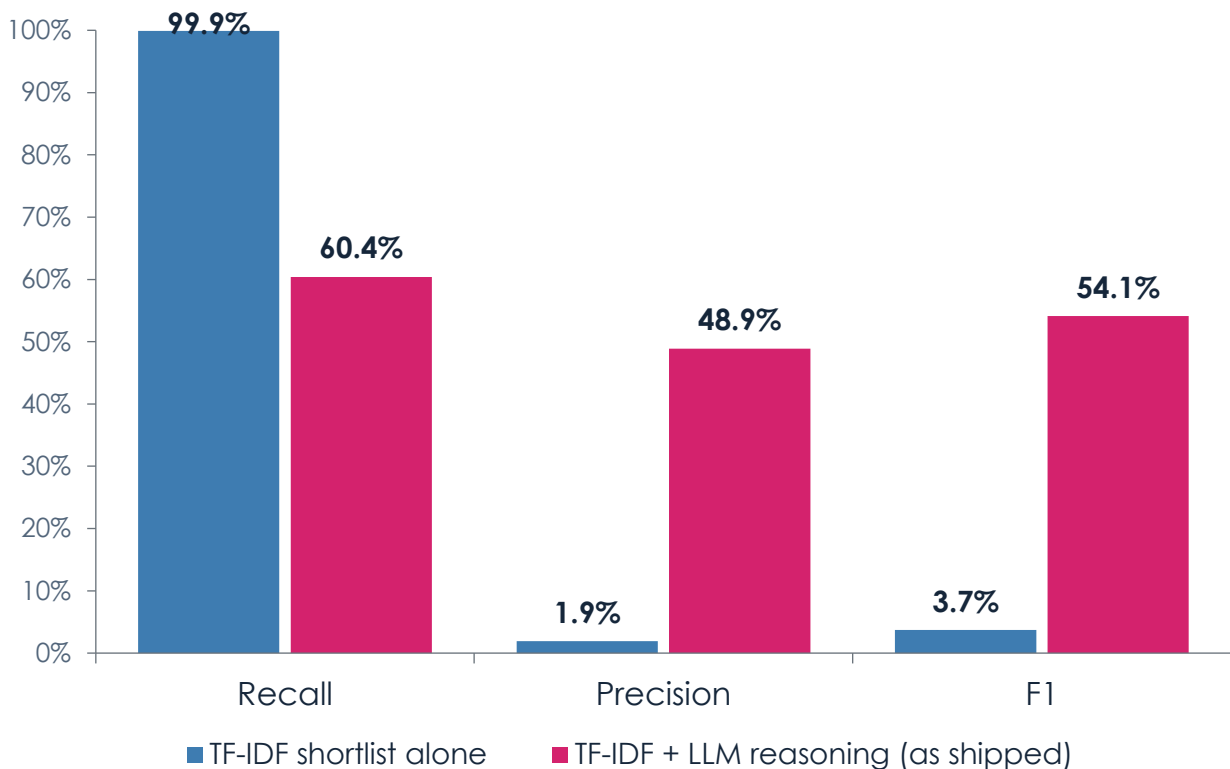
esteban.solorzanozeledon@colostate.edu



Backup Slides



Solution: Term Freq. Inverse Document Freq. shortlist + LLM reasoning



Recall is how many real links surface, precision is how much of what is shown is real, F1 balances the two.

Neither stage works alone

43x

less reading: 620 candidates per hazard become 14, each with a written rationale.

26x

more precise: 1.9% of the shortlist is a real link, versus 48.9% of what the LLM keeps.

179k

pairs exist; the LLM never sees them all. The bounded shortlist is the guardrail.



Blind evaluation, 131 hazards x 1,368 requirements: retrieval alone F1 3.7%, with LLM reasoning F1 54.1%.