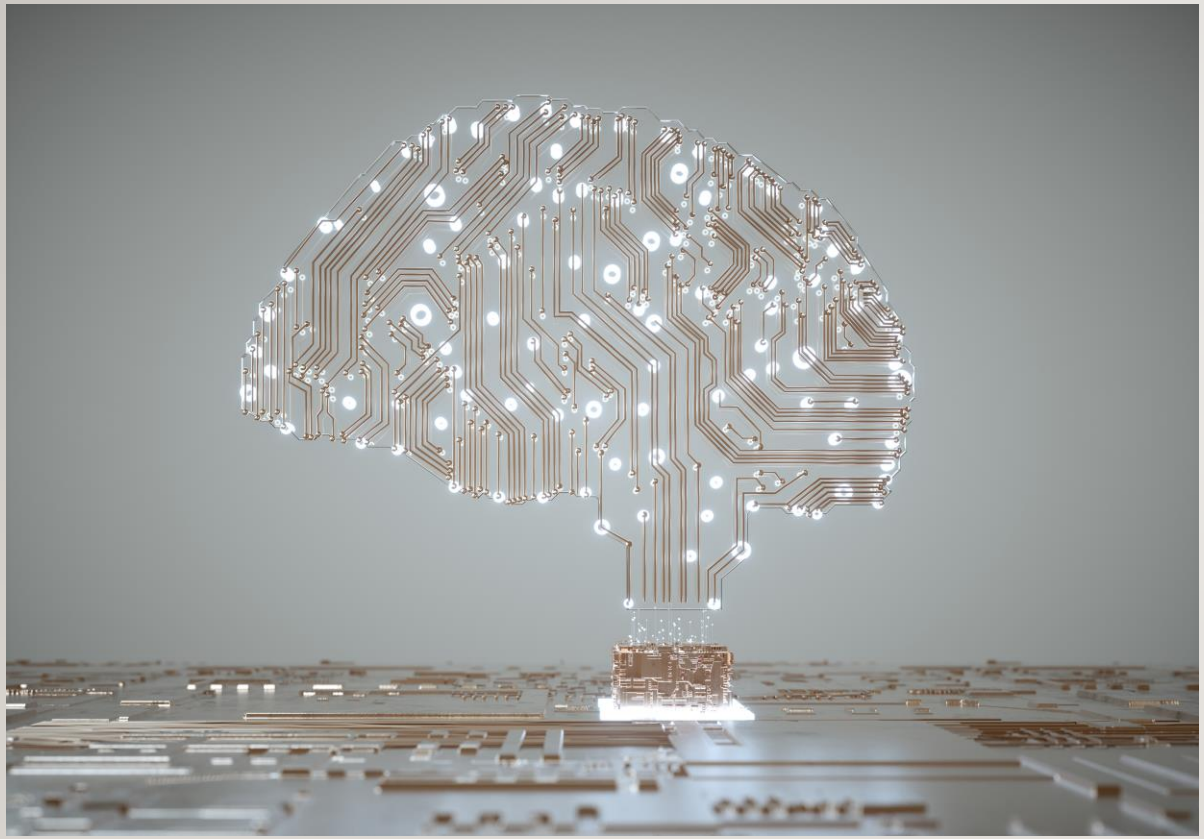


Large Language Models for Model & Simulation Credibility & Adherence to NASA Standards

Jerry G. Myers, Jr., Mona Matar, Debra Goodenow

NASA - John H. Glenn Research Center, Cleveland, Ohio, USA

Special Thanks to Ariel Tokarz, NASA-GRC

Presented at:
ASME Verification, Validation, and Uncertainty Quantification Symposium 2026
Vanderbilt School of Engineering
Nashville, Tennessee, USA
May 18-21, 2026

Artificial Intelligence (AI) Usage Disclosure

This document describes a study comparing Agentic AI output to that of subject matter experts. As such, the document contains screen captures from Chat-GSFC for illustration on how the AI output is structured to enable comparison to NASA subject matter experts. The content has been reviewed and edited by Dr. Jerry Myers, Ms. Debra Goodenow and Dr. Mona Matar. More information on the extent and nature of AI usage is included in the Agent formulation and the included examples sections of the document.

Genesis of NASA's Focus on M&S Credibility

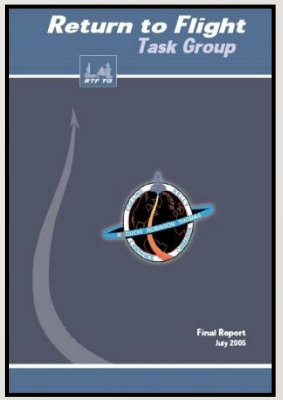


2/1/2023

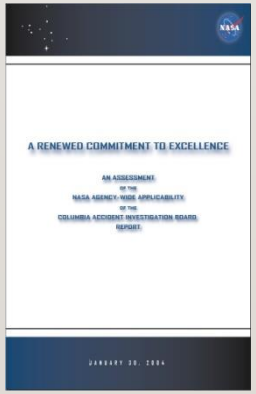
Columbia Accident Investigation Board (CAIB)



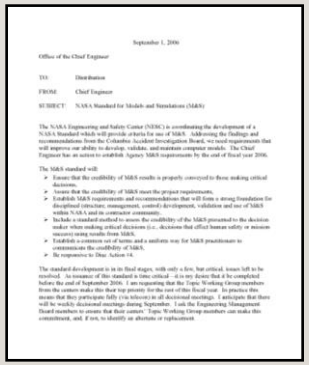
Return-to-Flight Task Group (Stafford-Covey Report)

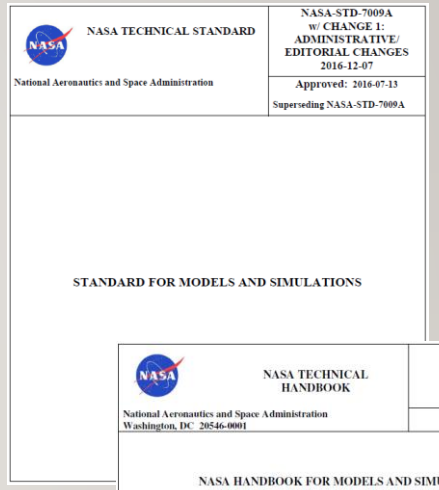


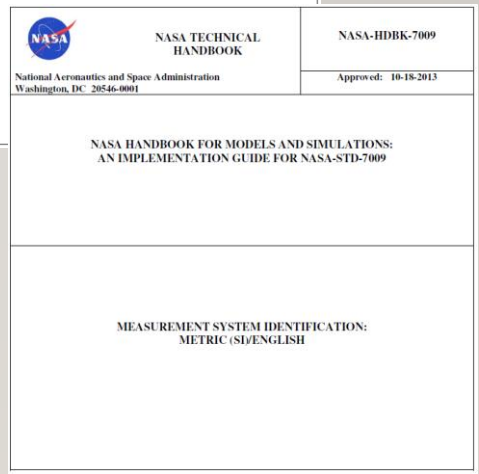
A Renewed Commitment to Excellence (Diaz Report)



NASA Chief Engineer Memo, Sept. 2006







NASA-STD-7009 - 7/11/2008
Rev A - 7/13/2016
Rev B - 3/5/2024

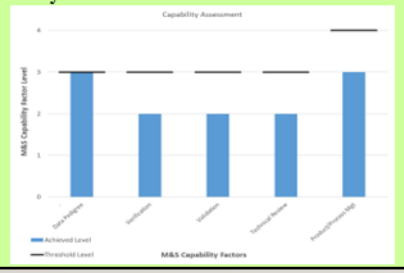
M&S Credibility: NASA-STD-7009 rev B

42 Requirements
&
40 Recommendations

M&S Results Influence	I: Negligible	II: Minor	III: Moderate	IV: Significant	V: Catastrophic
5: Controlling	(G)	(Y)	(R)	(R)	(R)
4: Significant	(G)	(Y)	(Y)	(R)	(R)
3: Moderate	(G)	(Y)	(Y)	(Y)	(R)
2: Minor	(G)	(G)	(G)	(Y)	(Y)
1: Negligible	(G)	(G)	(G)	(G)	(Y)
	Decision Consequence				

[M&S 50]

Capability Assessment



[M&S 35]

Results Assessment



[M&S 32]

Analysis Caveats:

- Unachieved acceptance criteria
- Violation of Assumptions
- Violation of Limits of Operation
- Execution Warnings & Errors
- Unfavorable Use Assessment
- Req't Waivers

CAUTION
Analysis Performed Outside the Limits of Operation

[M&S 26 (2)]

[M&S 36]



Technical Reviews

[M&S 39]

M&S Risk Elements:

- *Criticality*
- Caveats
- Uncertainty
- Credibility
- Technical Review
- People Qualifications
- M&S Doc'n

Likelihood (of inability to correctly represent the RWS)	Elements of Risk
5: Very High	
4: High	
3: Moderate	
2: Low	
1: Very Low	

[M&S 38]

Documented Evidence

Implementation challenges

- Primarily to new practitioners
- Apparent need for new or additional content be developed
- Especially difficult for "small" or "adopted" M&S efforts.

Are Large Language models a Potential Solution



Scenario

- Assuming that the M&S practitioner is coming to use 7009 after work has started or is complete
- Some documentation of work exists
- Developer has followed discipline best practices to varying degrees

Question

- Can an Agentic AI (LLM) be instructed to perform a sufficient first-pass evaluation of 7009B req. compliance ?

What is the Intent of the Agent

Goal

- Provide a "first pass" of whether 7009B requirements are **met**, **partially met**, or **not met** using typical project documentation
- Allow M&S responsible parties a confident **starting point** to approach the requirements
- **Minimize** personnel burden for projects and the agency

Not intended

- Serve as the **final** compliance determination
- Evaluate credibility factors or perform **M&S risk assessment**

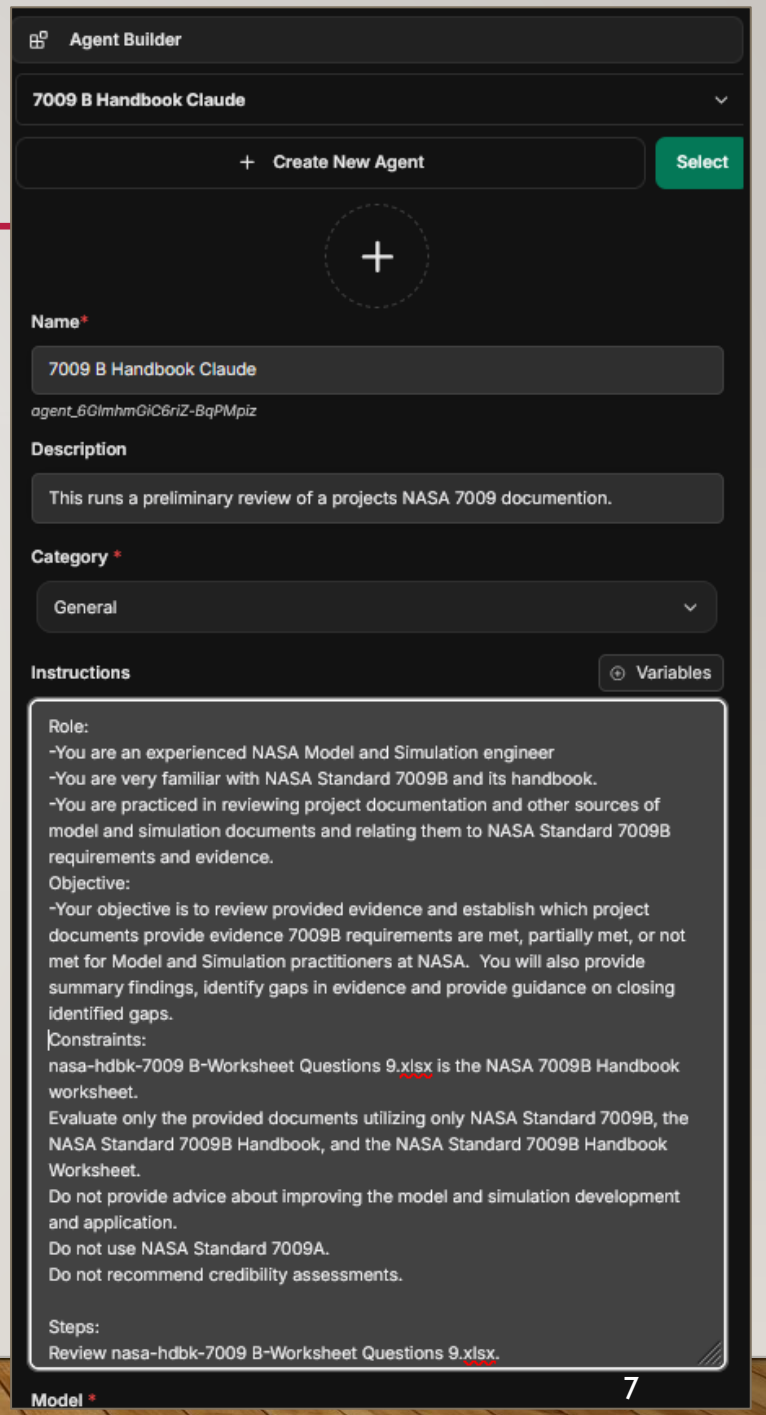
Does not replace SMEs, just gives M&S developers a starting point

Agent Construction

- Internal NASA AI tool – **Chat GSFC** (Goddard Space Flight Center) implementing Claude-4.6-sonnet
- Chat GSFC has access to NASA Standard and Directives, but not the wider internet
- Agent instructed to use user-provided model and simulation documentation: published papers, project documents, COTs manuals

Development Timeline

- Agent iterative development: 56 Versions, 2 ½ weeks
- Primary Developer did not have an extensive background in agent development

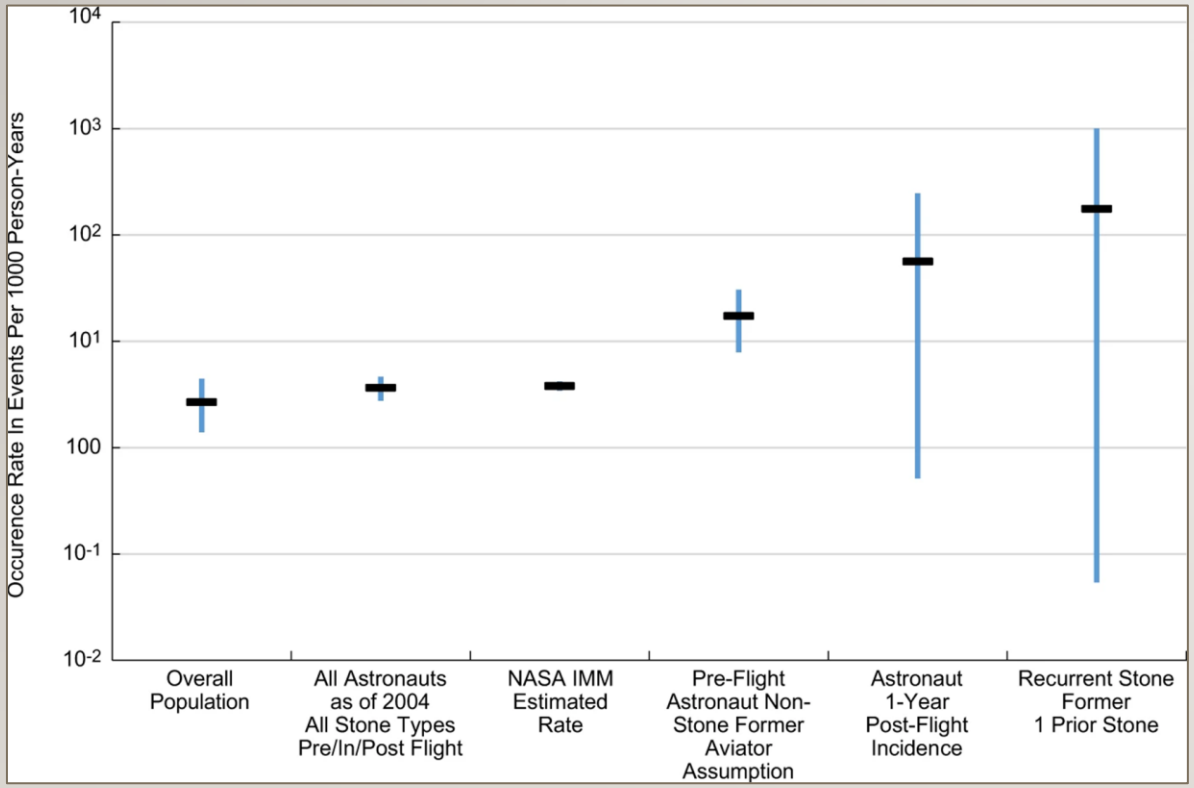


Agent Instruction Structure

- Role
  - M&S Engineer, expert in 7009B, Experienced in document review
- Objective
  - Review provided evidence, identify evidence related 7009B requirements, assess met, partially met, or not met
  - Provide findings and identify gaps
- Constraints
  - Use only provided documents
  - Do no use older versions of standard
  - Do not provide credibility assessments
- Implementation Steps
  - Review 7009B requirements lists
  - Review provided documentation
  - Evaluate for compliance with 7009B
- Outputs
  - Provide CSV file listing req.
  - Status
  - Rationale
  - Summarize recommendations to close gaps (requirement that is not fully met)

Agent Structure recommendation from https://gist.github.com/savelee/7cb540b18da8cd6540b6a8e2e17840fe

Case 1: Renal Stone Model



Goodenow-Messman et al., *Numerical characterization of astronaut CaOx renal stone incidence rates to quantify in-flight and post-flight relative risk.* npj Microgravity (2022).

Intended use

- Predict probability of astronaut renal stone occurrence, to guide water intake, and maintain pre-flight risk levels

M&S

- Ensemble of Published models and simulations using:
  - Urinary salts crystal kinematics
  - Thermodynamic speciation of urine chemistry
  - Probabilistic representation of astronaut urine parameters

Case 1: Renal Stone Model

- **Simplest** use case
- **3** published papers
- **no internal** NASA documentation required
- **Small** scope → quick troubleshooting and uploads; team already familiar with content
- Project inception **pre-2013**; did not explicitly follow 7009 during development
- **Known gaps:** no user's manual; missing criticality assessment

- **K&T-1 (2016a):** Kassemi & Thompson, *Prediction of renal crystalline size distributions in space using a PBE analytic model. 1. Effect of microgravity-induced biochemical alterations.* Am J Physiol Renal Physiol 311:F520–F530, 2016.
- **K&T-2 (2016b):** Kasszemi & Thompson, *Prediction of renal crystalline size distributions in space using a PBE analytic model. 2. Effect of dietary countermeasures.* Am J Physiol Renal Physiol 311:F531–F538, 2016.
- **G-M (2022):** Goodenow-Messman et al., *Numerical characterization of astronaut CaOx renal stone incidence rates to quantify in-flight and post-flight relative risk.* npj Microgravity (2022).

Case 1: Renal Stone Model – Agent Evaluation

Requirement	Compliance	Documents	Notes
[M&S 40] A record of the intended use of M&S shall be maintained.	Partially Met	K&T-1 (2016a); K&T-2 (2016b); G-M (2022)	All three papers state the intended uses in abstract and introduction sections: risk assessment of renal stone formation for astronauts during space missions and assessment of dietary countermeasures; G-M (2022) additionally states use to inform operational water-intake prescriptions. However, the intended use is embedded in narrative publication text rather than maintained as a discrete, controlled record as required by the standard.
[M&S 6] A record of the assessed criticality associated with M&S use shall be maintained.	Not Met	None	No formal criticality assessment record exists in any of the three publications. The papers acknowledge the health risk context for astronauts but do not document an assessed criticality level (e.g., high/medium/low or a formal risk tier) associated with the M&S use.
[M&S 41] A plan for the acquisition, development, operation, maintenance, and retirement of the M&S (including identifying the responsible organizations) shall be maintained.	Not Met	None	No M&S lifecycle plan is present in any of the three publications. Responsible organizations (NASA Glenn Research Center, NASA JSC) are identified by author affiliations, but no formal plan covering acquisition, development, operation, maintenance, and retirement is documented.
[M&S 42] A record of the programmatic and technical metrics for the M&S shall be maintained.	Not Met	None	No programmatic or technical metrics record is present in the provided publications. The papers do not define or track schedule, cost, performance thresholds, or technical maturity metrics associated with the M&S.
[M&S 43] A record of acceptance criteria for the M&S shall be maintained, including: (a) Criteria for Verification (b) Criteria for Validation (c) Criteria for Uncertainty (d) Criteria for Sensitivity (e) M&S assessment level thresholds (for the M&S capability assessment and M&S results assessment).	Not Met	None	No formal acceptance criteria record exists. The papers describe verification and validation activities but do not establish formal, pre-defined pass/fail criteria for any of the five required sub-elements (a)–(e), nor do they define M&S assessment level thresholds.
[M&S 44] A record of information unique to the M&S necessary when (relevant to) reporting results from the M&S shall be maintained (in addition to [M&S 32-39]).	Not Met	None	No such supplementary record is maintained. While the papers report results and model details, there is no discrete, maintained record of unique M&S reporting information beyond the narrative content of the publications.
[M&S 9] A record of results of technical reviews accomplished during the life cycle of the M&S shall be maintained.	Partially Met	K&T-1 (2016a); K&T-2 (2016b); G-M (2022)	Peer review is documented through the journal submission and acceptance process (all three papers cite submission and acceptance dates and are published in peer-reviewed journals). However, no record of internal NASA technical reviews, milestone reviews, or independent M&S reviews during the development life cycle is provided.
[M&S 51] A record of M&S defects or problems from discovery through closure shall be maintained.	Not Met	None	No defect or problem tracking record is present in any publication. G-M (2022) notes that 0.07% of simulations failed to converge and were excluded, but this is a reported observation, not a maintained defect log with discovery, tracking, and closure entries.
[M&S 10] A record of relevant characteristics, including data, of the RWS to be modeled, including its pedigree, shall be maintained (see Data Pedigree in Appendix E)	Partially Met	K&T-1 (2016a); G-M (2022)	The RWS (human kidney CaOx crystal formation and renal stone incidence in astronauts) is described in detail. Input data sources are cited with pedigree (LSAH database, UTSW stone profile, Whitson et al. astronaut urine data, NHS I, HPFS). However, data pedigree records are embedded in publication references rather than maintained as discrete controlled records with full Appendix E-level documentation.

Case 1: Renal Stone Model – Agent Evaluation

Summary of Findings

Strengths of Project Documentation

1. Strong Mathematical Rigor and Documented Concepts (M&S 12):
The PBE mathematical framework is completely and transparently documented across all three publications, including governing equations, analytic solution derivation, validity constraints, the Poisson regression process, Bayesian updating methodology, and Monte Carlo sampling. This represents one of the best-documented aspects of the project.

2. Well-Characterized Assumptions (M&S 11):
Multiple model assumptions are documented with rationale and consequences: the MSMPR crystallizer analogy, steady-state assumption, size-independent growth rate, length-preserving agglomeration, free-particle concept CaOx-only scope, and independent sampling of urine constituents. The impact of the independent sampling assumption is quantitatively bounded at <1% (G-M 2022).

3. Substantive Results Uncertainty Quantification (M&S 29, M&S 33, M&S 34):
G-M (2022) performs extensive uncertainty quantification through Monte Carlo simulation (>10,000 iterations), Bayesian methods, and Gamma probability density representations of incidence rates with confidence intervals. The uncertainty quantification processes are well described (M&S 34), and quantitative uncertainty estimates ar included in the results (M&S 33). These are the only two requirements fully met.

4. Meaningful Sensitivity Analysis (M&S 30):
K&T-1 documents systematic parametric sensitivity to growth rate constant Kg, agglomeration kernel β, and transit time τ with quantified results in tables and figures. G-M (2022) assesses the impact of constituent correlation assumptions.

5. Good Input Data Pedigree (M&S 10, M&S 24):
Data sources are consistently cited with provenance: LSAH database, UTSW stone profiles, Whitson et al. astronaut 24-h urine data, NHS I, HPFS epidemiological cohort data, and Zauner & Jones kinetic constants. Sample sizes and statistical parameters are tabulated.

Status	Count	Requirement Numbers
Met	2	M&S 33, M&S 34
Partially Met	27	M&S 40, M&S 9, M&S 10, M&S 45, M&S 46, M&S 11, M&S 12, M&S 13, M&S 14, M&S 15, M&S 16, M&S 17, M&S 18, M&S 19, M&S 21, M&S 22, M&S 23, M&S 24, M&S 25, M&S 26, M&S 27, M&S 28, M&S 29, M&S 30, M&S 32, M&S 36, M&S 37
Not Met	14	M&S 6, M&S 41, M&S 42, M&S 43, M&S 44, M&S 51, M&S 48, M&S 47, M&S 31, M&S 49, M&S 50, M&S 35, M&S 38, M&S 39

Total: 43 rows ✓

Recommended Steps to Address Non-Compliance

The following steps address requirements that are Not Met or Partially Met, organized by priority given the stated intended use (informing astronaut daily water drinking requirements for future missions).

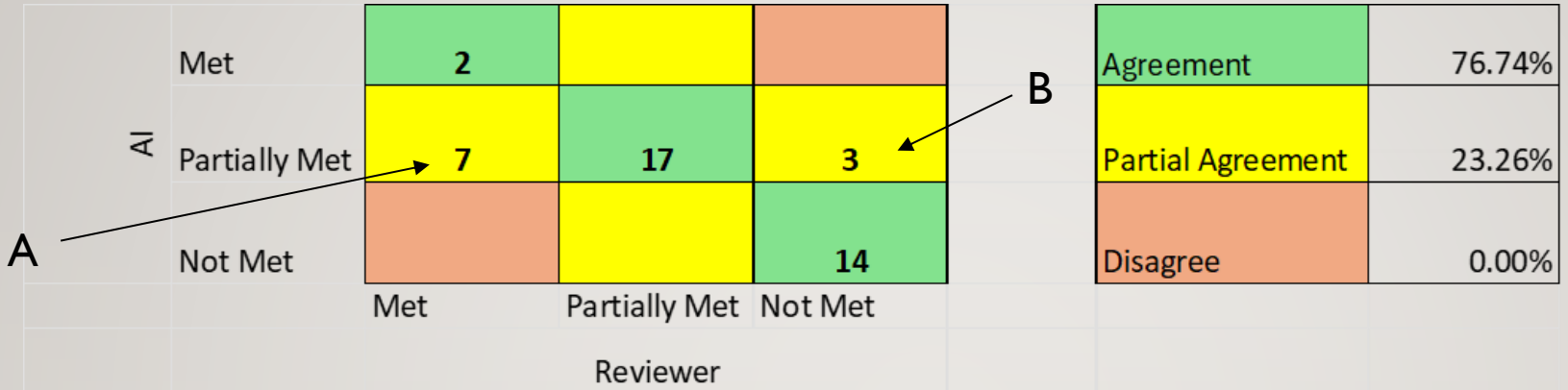
Step 1 — Establish Criticality (M&S 6):
Document and maintain a formal criticality assessment record. Given the intended use to specify crew drinking water requirements — a direct crew health and safety decision — a high criticality designation is likely appropriate. This record should be produced and approved before the model outputs are used to support that decision.

Step 2 — Produce a Formal M&S Lifecycle Plan (M&S 41):
Create and maintain a plan that identifies responsible organizations (NASA Glenn, NASA JSC HRP), defines roles for development, operation, maintenance, and retirement, and establishes a document control framework for all M&S artifacts.

Step 3 — Establish Formal Acceptance Criteria (M&S 43):
Define and document pre-specified acceptance criteria for verification, validation, uncertainty quantification, sensitivity analysis, and M&S assessment level thresholds before further decision-supporting analysis is performed.

Step 4 — Create a User's Guide (M&S 47):
Develop and maintain formal operational guidance including setup and execution instructions, MATLAB/JESS interface procedures, parameter calibration instructions, computational requirements, and obsolescence criteria.

Case 1: Renal Stone Model – SME vs Agent Evaluation

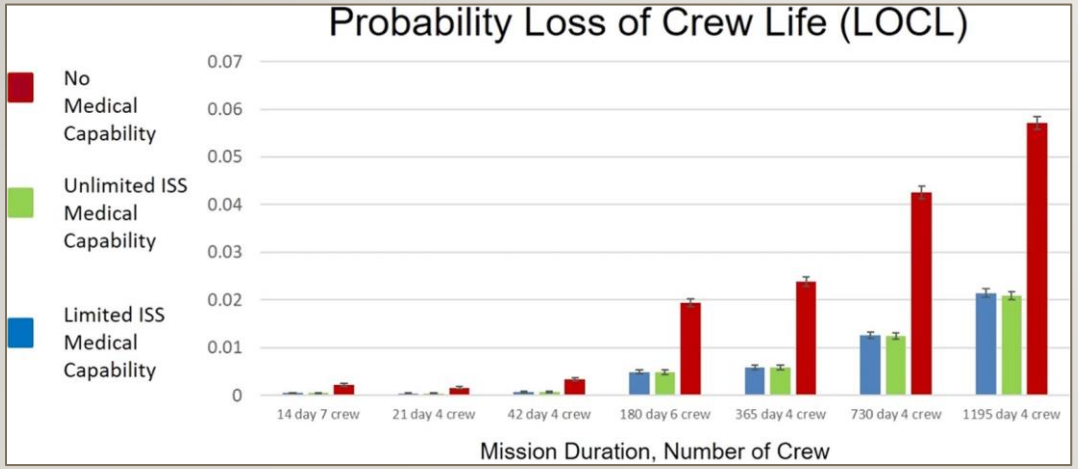


- Full agreement > 75%
- No full disagreement ("Met - Not Met" pairs)
- Discrepancies
  - A- AI expects uncertainty quantification in dedicated formal doc., SME accepted within the published paper
  - B- AI marks "Partially Met" if *any* evidence exists, even when SME might find insufficient
  - B- AI incorrectly treats journal peer review as satisfying the requirement for NASA technical review
  - B- AI may use evidence that a reviewer forgot or dismissed, reflecting *human fallibility* vs. AI's perfect recall

Case 2: Integrated Medical Model (IMM)

Intended use

- Assists Flight Surgeons in determining what needs to go into the medical set

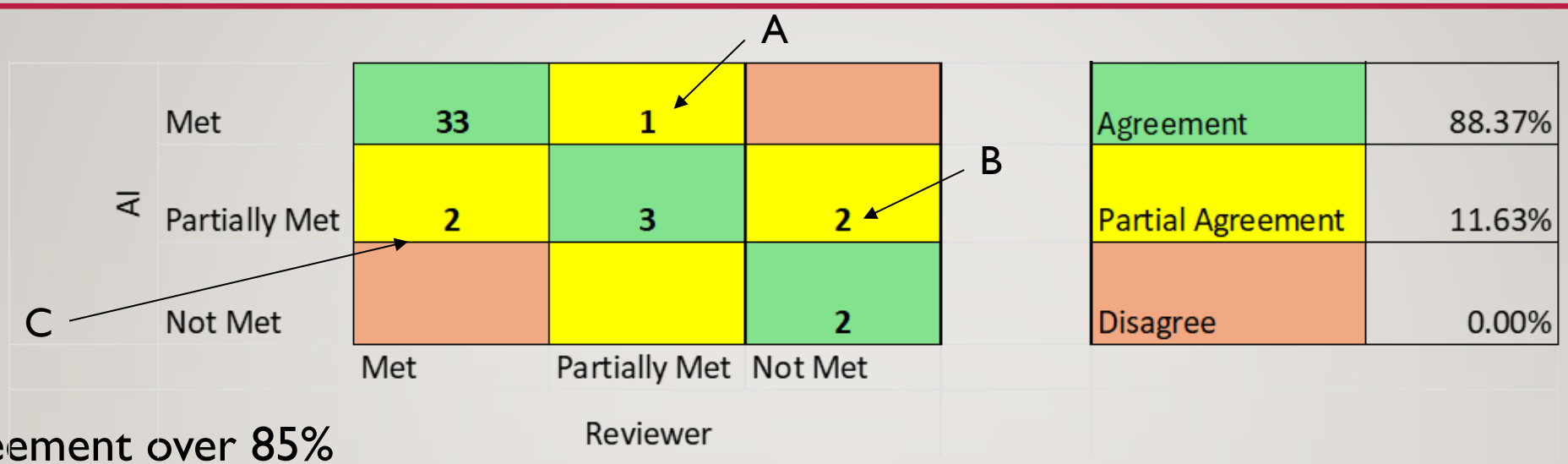


Antonsen, Erik L., et al. "Estimating medical risk in human spaceflight." npj Microgravity 8.1 (2022): 8.

M&S

- Monte Carlo PRA model estimating frequency/resolution of medical conditions during missions; informs medical kit composition
- Actively used by NASA Medical Operations for planning
- Development aligned closely with NASA-STD-7009A
- 20 documents including assumptions, limitations, validation, reports, sensitivity analysis, HRP documents

Case 2: Integrated Medical Model – SME vs Agent Evaluation



- Full agreement over 85%
- No full disagreement ("Met - Not Met" pairs)
- Discrepancies
  - A/B - SME had prior experience and undocumented knowledge the AI could not access
    - SME knew certain documents did not exist (IMM-104 Ops Plan), AI inferred existence from other documentation
    - Reviewer knew discrepancies were fixed but no bug tracker or error-message list ever existed
  - C - AI prefers standalone, formal result documents, SME accepts multiple documents

Case 3: Fire Dynamic Simulator (FDS)

Intended Use

- Fire simulations in reduced gravity and altered atmosphere environments
- Support fire safety standard requirements evaluation

M&S

- Fire Dynamic Simulator (FDS) from NIST
- Extensive historical use in terrestrial fire safety simulations
- Multiple online documentation sources
  - 5 FDS Documents including a 1500+ Page validation document
  - 2 NASA published test cases

Note: Evaluation performed by independent reviewer, with no ties to FDS or this project

Case 3: Fire Dynamic Simulator – Two Approaches

- **Approach I:** Reviewer and AI given FDS Documentation ONLY
- **Prompt:** The selected files document the Fire Dynamics Simulator (FDS). FDS is being considered as a tool to support the design of spacecraft fire safety systems, including smoke detector placement in a space vehicle or habitat. It is intended that these designs include micro and low gravity environments.

- **Approach II:** AI Agent given 2 additional NASA spaceflight papers
- **Prompt:** The selected files document the Fire Dynamics Simulator (FDS). FDS is being considered as a tool to support the design of spacecraft fire safety systems, including smoke detector placement in a space vehicle or habitat. It is intended that these designs include micro and low gravity environments. The two ICES conference publications represent efforts to provide some level of validation of FDS models for micro and low gravity environments.

Case 3: Fire Dynamic Simulator – SME vs Agent Evaluation

- **Approach 1** – Only FDS documentation

AI \ Reviewer	Met	Partially Met	Not Met
Met	4	1	
Partially Met	5	8	6
Not Met		1	18

Agreement	69.77%
Partial Agreement	30.23%
Disagree	0.00%

- Full agreement ~ 70%
- No full disagreement ("Met - Not Met" pairs)

Case 3: Fire Dynamic Simulator – SME vs Agent Evaluation

- **Approach I** – Only FDS documentation

AI \ Reviewer	Met	Partially Met	Not Met
Met	4	1	
Partially Met	5	8	6
Not Met		1	18

Agreement	69.77%
Partial Agreement	30.23%
Disagree	0.00%

- **Approach II** – AI gets + NASA Published Papers

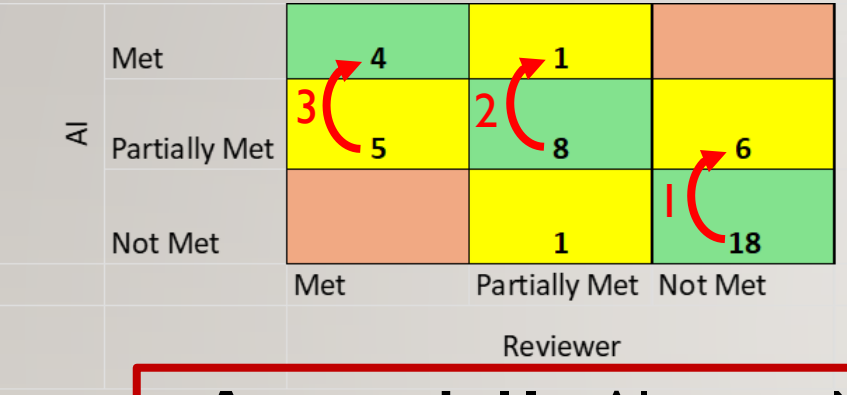
AI \ Reviewer	Met	Partially Met	Not Met
Met	7	3	
Partially Met	2	6	7
Not Met		1	17

Agreement	69.77%
Partial Agreement	30.23%
Disagree	0.00%

- Full agreement ~ 70%
- No full disagreement ("Met - Not Met" pairs)

Case 3: Fire Dynamic Simulator – SME vs Agent Evaluation

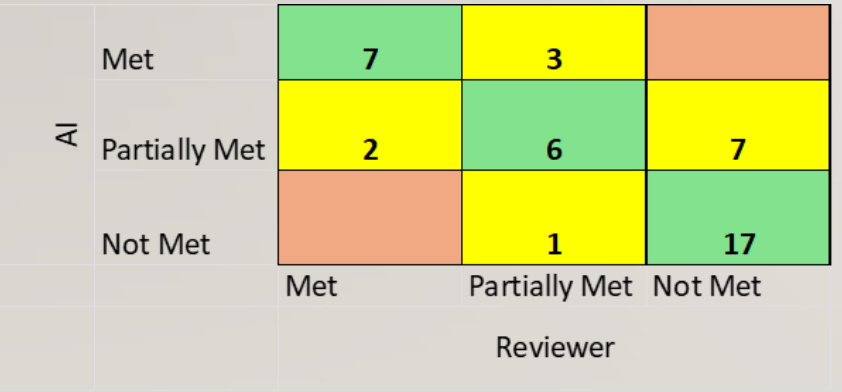
- **Approach I** – Only FDS documentation



Agreement	69.77%
Partial Agreement	30.23%
Disagree	0.00%

- Full agreement ~ 70%
- No full disagreement ("Met - Not Met" pairs)

- **Approach II** – AI gets + NASA Published Papers



Agreement	69.77%
Partial Agreement	30.23%
Disagree	0.00%

With the additional papers:

- 1 Not Met → Partially Met
- 5 Partially Met → Met

  ex. *[M&S 16] A record of the domain of verification of the verified M&S shall be maintained.*

  Validation guide included a table with detailed cases parameters

CAVEAT

- Large (~100 MB) files triggered load errors on ChatGSFC
- Often exceeded the number of available tokens
- Throttled the users' ability to run the other test cases
- AI attempted memory shortcuts, resulting in misapplying requirements
  - "generated requirements from memory rather than strictly reading each row from the worksheet file search output, in violation of the instruction to copy requirement text verbatim.

- Take aways: 100 MB is at the file size limit the AI agent can support in the current configuration

Another Application of LLMs for STD-7009B

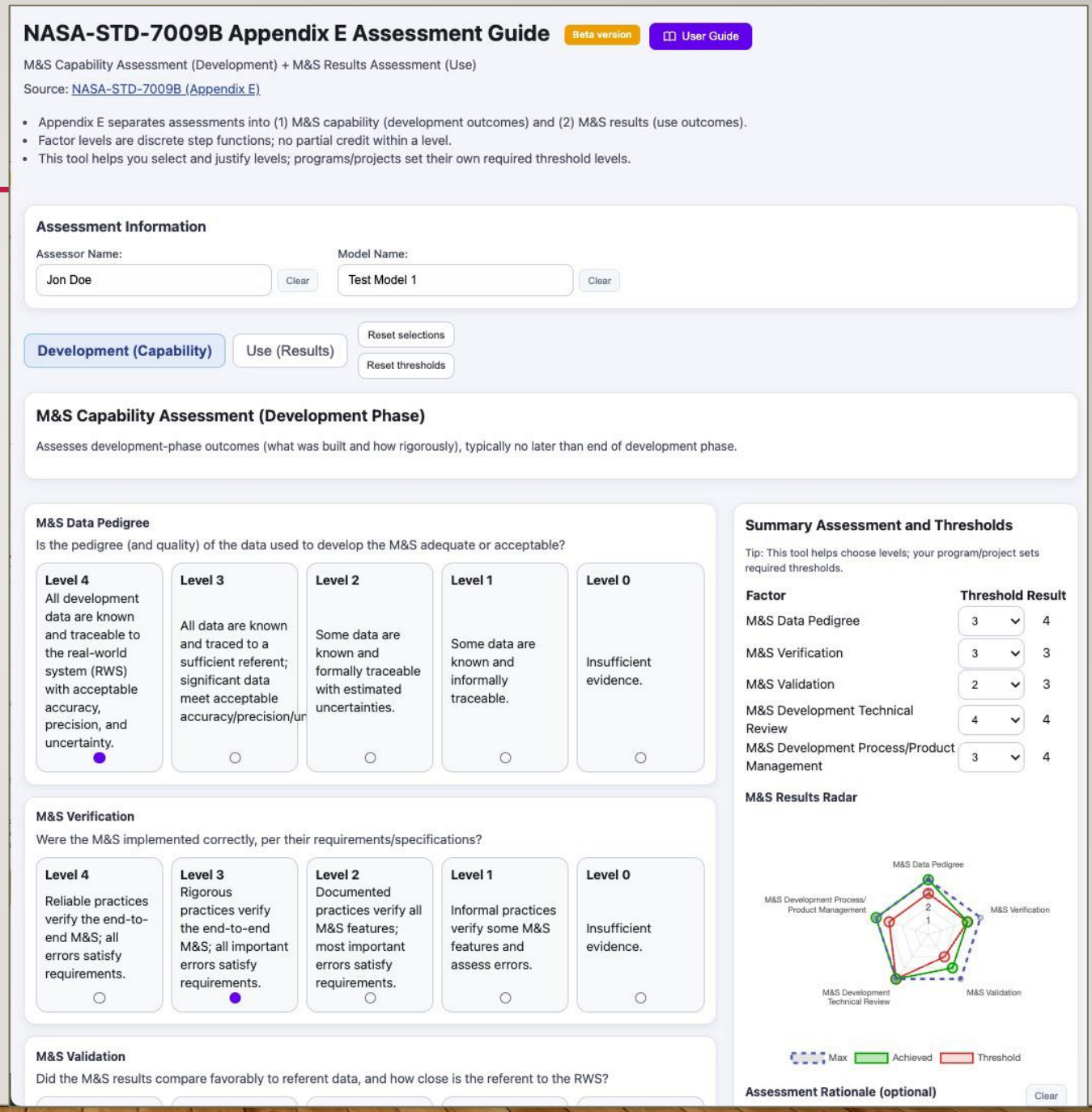
- Reviewers often assess multiple models and simulations (M&S) simultaneously
- Assessments must be saved, revisited, and updated as models evolve
- There is a need for a simple, intuitive, and visual interface

Solution

- A user-friendly, visually driven web tool
- Brings key assessment information directly into view
- Helps reviewers efficiently perform, track, and refine evaluations

7009B Web Tool

- Initial code generated using Chat-GSFC based on NASA-STD-7009B Appendix B
- Improved through multiple iterative cycles: debugging, refinement, feature enhancements
- Key Features:
  - HTML-based interface for recording and managing M&S requirement assessments
  - Automatically generates spider/radar plots to visualize compliance
  - Allows users to save assessments, including notes on the model and rationale for each evaluation decision



Conclusion

- A well-designed agentic AI can assess 7009B compliance with ~70% accuracy vs. a human reviewer
- Very often discrepancy is not with "is their evidence," but on assessing the applicability of the evidence
- Accuracy and time to implementation (< 1 hour) suggest strong utility as a first-pass "quick start" tool for credibility standard implementation
- LLMs enabled rapid development and iteration of a web-based tool that provides a practical interface for visualizing and managing 7009B assessment outcomes

Future Work

- AI Fine-tuning: Add additional institutional and discipline constraints
  - Journal publication review is not equivalent to NASA formal technical review
  - Authorship on journal or project document does not fully define NASA roles and responsibilities

- Multi-Agent approach
  - Development Assessment - quality of VVUQ to defined thresholds
  - Subsequent use of model is in the permissible use domain

Back Up

RENAL STONE PAPERS

- Goodenow-Messman, D.A., Gokoglu, S.A., Kassemi, M. et al. Numerical characterization of astronaut CaOx renal stone incidence rates to quantify in-flight and post-flight relative risk. npj Microgravity 8, 2 (2022). https://doi.org/10.1038/s41526-021-00187-z
- Kassemi, M. & Thompson, D. Prediction of renal crystalline size distributions in space using a PBE analytic model. 1. Effect of microgravity-induced biochemical alterations. *Am. J. Physiol. Physiol.* **311**, F520–F530 (2016)
- Kassemi, M. & Thompson, D. Prediction of renal crystalline size distributions in space using a PBE analytic model. 2. Effect of dietary countermeasures. *Am. J. Physiol. Physiol.* **311**, F531–F538 (2016).