A Generalized AI-Powered Framework for Cross-Domain Technology Readiness Assessment

Marcelo Rea-Guamán(✉), Andrés Almeida-Jara, and Brandon Masacela-Atupaña

Departamento de Ciencias de La Computación, Universidad de las Fuerzas Armadas ESPE, Av. General Rumiñahui, Sangolquí 171103, Ecuador
amrea1@espe.edu.ec

Abstract. Technology Readiness Level (TRL) assessment remains largely domain-specific and subjective, lacking standardized cross-domain methodologies. A systematic literature review of 39 studies reveals that 84.6% of TRL adaptations target single domains, and none combines cross-domain coverage for hardware, software, and AI/ML technologies with standardized binary criteria and automated evaluation. This paper presents a generalized TRL assessment methodology grounded in ISO 16290, NASA, DoD, and EU Horizon standards, employing mandatory/supportive binary criteria and the weakest-link principle across three technology types. The methodology is operationalized through an AI system using DeepSeek-R1 with chain-of-thought reasoning deployed on HPC infrastructure. Empirical validation against independent expert assessments on 17 projects from open academic repositories demonstrates 76.5% exact concordance and 94.1% agreement within ±1 level. Results confirm that LLM-based evaluation can replicate expert judgment while providing transparent, evidence-traced justifications, enabling reproducible and scalable TRL assessment.

Keywords: Technology Readiness Level · cross-domain assessment · binary criteria · Large Language Models · systematic literature review · technology evaluation

1 Introduction

1.1 Background

Technology Readiness Levels (TRLs) were developed in the early 1970s by NASA's Stan Sadin (Sadin et al., 1989), to help quantify the level of development of technology used in space systems on a nine point ordinal scale. This concept was later formalized by Mankins (2009), who provided the nine levels and descriptors still in common use today. In subsequent years, while originally designed to be an engineering metric applicable to the space industry, it has been adopted by numerous organizations outside of the space community. For example, the U.S. Department of Defense made the use of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 476–492, 2027.
https://doi.org/10.1007/978-3-032-30488-9_30

TRL assessments mandatory through statute for all major defense acquisition programs (U.S. Government Accountability Office, 2020), the European Space Agency included the definition of TRL within their Engineering Standards (ISO, 2013), and the European Commission required TRL reporting among Horizon Research Funding Programs (European Commission, 2014). According to Héder (2017) as the concept transitioned from being an engineering tool to a cross-institutional governance tool, each new layer of interpretation progressively reduced the precision of the initial scale.

In addition to this fragmentation, it is also the creation of many domain-specific versions of TRL which are not compatible with one another. Researchers in health care have developed implementation science-based versions of TRL (Salvador-Carulla et al., 2024) while the machine learning community has created an eleven-level version of TRL to accommodate the ideas of "data readiness" and "iterative switchbacks" between maturity levels (Lavin et al., 2022) in software intensive and AI driven systems. Studies on the governance of AI have shown that there are significant differences between the hardware focused logic behind NASA's original TRL scale and how software and AI based technology behave (Martínez-Plumed et al., 2021). Each of the domains in which TRL is used is therefore developing its own evaluation process for assessing TRL independent from each other creating criteria for evaluating TRL that are neither portable nor capable of being compared with those developed by others. Even more importantly, despite these many adaptations, expert panels continue to be relied upon to evaluate TRL and these evaluations are clearly subjectively determined and difficult to replicate as well as resource intensive to accomplish (Olechowski et al., 2020; Tomaschek et al., 2016).

The computational science community feels the impact of these structural limitations the most. Evaluating modern technological tools—such as AI/ML pipelines, software platforms, and data-driven systems—remains complicated because a unified, cross-domain TRL methodology simply does not exist yet. This gap motivates the work presented in this paper.

1.2 Research Gap and Objectives

A systematic review of 39 peer-reviewed studies on TRL methodologies, conducted following PRISMA guidelines, reveals that 84.6% of existing approaches are confined to a single domain and that no prior work simultaneously addresses four characteristics that a general-purpose TRL assessment framework requires: (i) cross-domain coverage spanning hardware, software, and AI/ML technologies with criteria specific to each type; (ii) standardized binary assessment criteria grounded in internationally recognized standards (ISO, NASA, DoD, and GAO); (iii) implementation through a large language model (LLM) capable of generating evidence-traced, reproducible justifications; and (iv) application of the weakest-link principle, whereby the overall TRL of a technology is bounded by the lowest level at which all mandatory criteria are satisfied. The co-occurrence of these four properties in a single system has not been reported in the literature.

This gap gives rise to three research questions. **RQ1:** What are the prevailing patterns and open gaps in the literature on TRL assessment methodologies? **RQ2:** Can a

generalized, binary-criteria methodology cover hardware, software, and AI/ML technologies within a unified evaluative framework? **RQ3:** Can a large language model with chain-of-thought reasoning replicate expert TRL assessments with sufficient accuracy to be considered a valid evaluation instrument? The scope is limited to academic and research-stage projects; commercial deployment and defense acquisition contexts are explicitly excluded.

1.3 Contributions

This paper makes four concrete contributions to the literature on technology evaluation and AI-assisted assessment:

C1—Systematic Literature Review: A PRISMA-compliant review of 39 studies that maps the current state of TRL methodology research, quantifies the prevalence of domain-specific adaptations, and identifies the specific combination of properties absent from all existing approaches.

C2—Generalized Assessment Methodology: A unified TRL framework employing mandatory (MUST) and supportive (SHOULD) binary criteria across three technology types—hardware, software, and AI/ML—derived from a synthesis of ISO 16290:2013, NASA, DoD, and GAO standards, and directly informed by the findings of the systematic review.

C3—AI-Powered Evaluation System: An operational implementation of the methodology using DeepSeek-R1 with chain-of-thought reasoning, deployed on high-performance computing infrastructure, which produces structured TRL assignments accompanied by evidence-traced justifications grounded in the text of the evaluated documents.

C4—Empirical Validation: A comparative study of AI-generated TRL assignments against independent expert evaluations on 17 real-world projects, demonstrating 76.5% exact concordance and 94.1% agreement within ± 1 TRL level—establishing the practical viability of automated TRL assessment for research and academic contexts.

The remainder of this paper is organized as follows. Section 2 outlines related work for both TRL progression as well as the use of artificial intelligence to evaluate it. Section 3 explains the methodology used; specifically, the systematic review process; the evaluation framework; and the architectural structure of the AI system. Section 4 reports on empirical study results that integrate the results from the literature search with the validation results. Section 5 addresses implications and limitations of this study. Section 6 summarizes and draws conclusions. The overall research design is illustrated in Fig. 1.

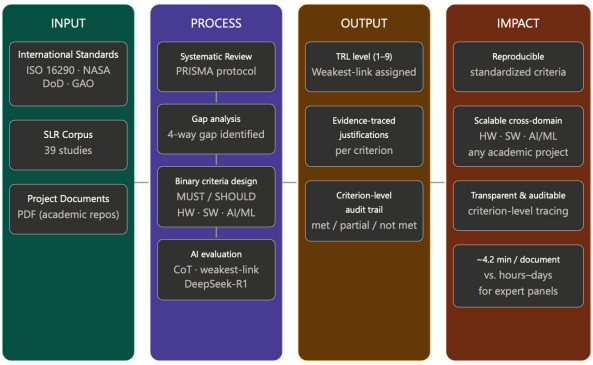

Fig. 1. Research framework overview. International standards and systematic review data are merged into a binary-criteria framework. TRLs are then assigned using an LLM-based chain-of-thought evaluation to ensure all results are traceable

2 Related Work

2.1 TRL Evolution and Standardization

The Technology Readiness Level Scale originated from NASA's early work in the late 1970s. The first official publication was published in 1989 as a seven-step readiness scale to assess space mission systems (Sadin et al., 1989). A later review of the technology scale's progression documented the institutional growth of the scale; specifically, Mankins (2009) described that the U.S. Department of Defense required the use of the TRL assessment process within all major acquisition programs as per statute. This is stated in the Technology Readiness Assessment Guidebook (U.S. Department of Defense, 2025); the European Space Agency incorporated TRL evaluation within their Engineering Standards through ISO 16290:2013 (ISO, 2013) and referenced in the Handbook ECSS-E-HB-11A (European Cooperation for Space Standardization (ECSS), 2017). The European Commission mandated reporting of TRL values under Horizon Research Programs (European Commission, 2014). Héder (2017) detailed this trend systematically demonstrating how each of the institutions' interpretations of the TRL progressively decreased the accuracy of the original NASA definitions.

This evolution highlights a major issue. The scale was designed to evaluate hardware in the aerospace industry, making it difficult to adapt its core concepts to software and modern AI technologies. The GAO-20-48G guide (U.S. Government Accountability Office, 2020), developed with over 180 experts, addresses assessment consistency across technology types but stops short of defining software- or AI-native criteria. The DoD TRA Guidebook 2025 (U.S. Department of Defense, 2025) introduced a dedicated software TRL scale, but cross-domain unification remains open. The intermediate transition zone between TRL 4 and TRL 7—the "Valley of Death"—is widely recognized as the stage where most development efforts stall due to exponential cost increases and academic-industrial misalignment (Sepúlveda et al., 2025; White et al., 2022; Clawson, 2023), and complementary readiness scales (MRL, CRL, RRL, DRL) have been proposed to address multidimensional challenges that TRL alone cannot capture (Kobos et al., 2018; H.-Y. Liu et al., 2020).

2.2 AI for Technology Evaluation

Wei et al. (2023) established that chain-of-thought prompting significantly improves LLM performance on complex multi-step reasoning tasks. Y. Liu et al. (2023), through the G-Eval framework, demonstrated that LLMs prompted with structured evaluation steps achieve higher correlation with human expert judgments than any prior automated metric—a pattern directly analogous to the criterion-by-criterion architecture of the present system. Zheng et al. (2023) established that strong LLMs used as judges achieve agreement with human preferences exceeding 80%, comparable to inter-human agreement rates.

Although significant progress has been made in applying LLMs to tasks of this nature; there are no previous studies that apply LLM-based evaluation to TRL assessments. This represents a significant void: TRL assessment represents a type of rubric-based judgment task (the same type on which architectures with LLM-as-evaluator have demonstrated their greatest success) using technical project documentation for which all of the supporting evidence must be identified, abstracted, and related to pre-defined criteria.

3 Methodology

3.1 Research Design

This research is based on a sequential mixed-methods approach with four connected stages. Phase one is a systematic literature review. It will establish the current state of research related to TRL methodologies so as to provide evidence for what combinations of attributes are missing from available methodologies. Phase two is the methodology development stage. In this phase we derive binary decision rules (i.e., pass/fail) for each type of technology (hardware, software, and artificial intelligence/machine learning technologies) using an integration of relevant international standards and consistent results from the systematic literature review; thereby making use of the systematic literature review's findings to determine methodology design, but not merely to motivate it. Phase three provides empirical realization of the methodology by means of developing an LLM-based evaluation system. Finally, Phase Four is used to validate the evaluation system using comparative assessments to independently assessed evaluations by experts.

The calibration process was done to ensure that phases three and four were connected by an iterative calibration loop before the final validation. The iterative calibration loop allowed for evaluation of the system using a defined set of cases that had already been assigned to a TRL (Technology Readiness Level) prior to the formal validation. The iterative calibration loop was used to refine the systems logic used to interpret evidence from each case until such time as consistent, criterion aligned behavior was demonstrated by the system.

3.2 Systematic Literature Review

The study followed the PRISMA 2020 Guidelines (Page et al., 2021) for conducting the systematic review. A total of five academic database searches were completed; namely,

ACM Digital Library, IEEE Xplore, Web of Science, Science Direct and Scopus. Only articles published from 2015 through 2025 were considered for the final screening. The common search string *("Technology Readiness Level" OR "TRL") AND ("adaptation" OR "customization" OR "modified") AND ("framework" OR "methodology" OR "guidelines" OR "model")*, was used on all databases using a uniform search strategy and restricting the search to the title field, abstract field and keyword field. Restricting the search to title field, abstract field and keyword field—as opposed to searching full text of articles—was a intentional methodological decision to maximize precision and eliminate retrieval of non-relevant documents that only incidently mention TRL.

Eligibility Criteria. To be included, the study must have been a complete, peer reviewed study using TRL methodology. In addition, we only considered studies published in the time frame of 2015–2025 as long as it contained TRL in the title, abstract, and/or keyword list. Articles that did not contain methodologically rigorous data, did not include a readiness framework related to TRL, were unverifiable grey literature, or duplicate data were all excluded.

Selection Process. Initially, our search produced 3,374 documents across five different databases. Upon conducting a title/abstract screen, we narrowed down to 294 potential candidates. Finally, after conducting a full-text review, we selected 39 total studies for inclusion. A detailed description of our selection process can be viewed in Fig. 2.

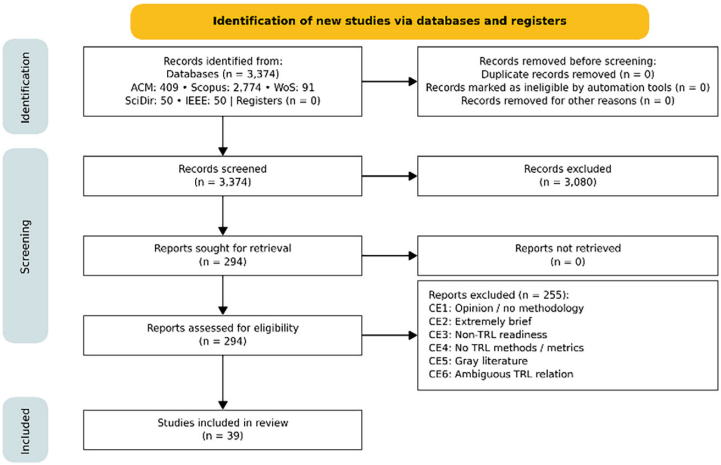

Fig. 2. PRISMA 2020 flow diagram of the systematic literature review.

3.3 TRL Assessment Framework

The assessment framework presented in this paper synthesizes criteria from four internationally recognized standards—ISO 16290:2013 (ISO, 2013), the NASA TRL Definitions (National Aeronautics and Space Administration (NASA), 2017), the DoD Technology Readiness Assessment Guidebook (U.S. Department of Defense, 2025), and

the GAO Technology Readiness Assessment Guide (U.S. Government Accountability Office, 2020)—informed by the convergent findings of the systematic review described in Sect. 3.2. The framework abandons point-based or weighted scoring in favor of binary criterion evaluation, a design decision grounded in both ISO 16290:2013—which defines conditions that must be satisfied at each level for a precise TRL assessment, not scores to accumulate—and GAO-20-48G, which was developed with over 180 experts across seven years and employs criterion-based evaluation exclusively.

Binary Criteria Structure. Each TRL level is defined by mandatory criteria (MUST), which must all be satisfied to validate the level, and supportive criteria (SHOULD), which reinforce the evaluation but do not block validation. Each criterion is assessed as satisfied, partially satisfied with justification, or not met. A level is validated when all MUST criteria are satisfied or partially satisfied.

Weakest-Link Principle. The overall TRL is the highest level for which all MUST criteria are satisfied. If any mandatory criterion is unmet at level k, the project is assigned level $k - 1$ regardless of evidence at higher levels, consistent with the NASA TRL Worksheet, AFRL TRL Calculator, and ISO 16290 (Clawson, 2023; Bruno et al., 2020).

Cross-Domain Coverage. Technology-type-specific criteria are defined for hardware, software, and AI/ML. Hardware criteria follow NASA and ISO 16290 directly. Software criteria translate each hardware concept to its computational equivalent per the DoD TRA Guidebook (U.S. Department of Defense, 2025) and NASA (National Aeronautics and Space Administration (NASA), 2017). AI/ML criteria extend the software set with mandatory additions from MLTRL (Lavin et al., 2022): data readiness, pipeline reproducibility, bias evaluation (before TRL 7), switchback provisions, model card (before TRL 6), and production monitoring (before TRL 8). Mixed projects receive dual evaluation with TRL defined as the minimum.

The present framework defines criteria natively appropriate to each type rather than adapting hardware criteria post hoc.

Illustrative Example: To illustrate the cross-domain differentiation, consider TRL 5 (system integration in a relevant environment). For hardware, MUST criteria require a complete system integrated and functional under lab conditions simulating the real environment, with full functional tests measuring key performance parameters. For software, the equivalent requires an integrated system tested with representative data and realistic load, with end-to-end performance metrics such as latency and throughput. For AI/ML, the criteria require a trained model integrated into the application pipeline with representative data, evaluated on a held-out validation set with documented metrics (accuracy, F1, AUC). SHOULD criteria at this level include technical documentation and limitation analysis, each adapted to the technology type. The complete criteria tables for TRL 1 through TRL 9 across all three technology types are available as supplementary material (Rea Guamán et al., 2026).

3.4 AI System Implementation

The assessment methodology described in Sect. 3.3 is operationalized through a two-component software system: a Python backend built on FastAPI that handles document

ingestion, criterion evaluation, and TRL assignment, and a React/TypeScript frontend that presents the evaluation workflow and results to the user. The evaluation engine interfaces with DeepSeek-R1 (32B parameters) served locally via Ollama, accessed through a REST API. Computational execution runs on high-performance computing infrastructure provided by CEDIA (Ecuador) with NVIDIA A100 GPU capacity, which supplies the GPU capacity necessary for local model inference—infrastructure that functions as an operational enabler rather than a methodological contribution.

Chain-of-thought Evaluation Pipeline. Rather than requesting a direct level assignment, the system decomposes evaluation into sequential steps: for each criterion at each candidate TRL level, the model (i) identifies specific passages constituting potential evidence, (ii) assesses whether the evidence satisfies the criterion, (iii) assigns a state (satisfied, partially satisfied, or not met), and (iv) produces a natural-language justification tracing the state to extracted textual evidence. Because of this, each assignment includes a clear and traceable explanation. This specific feature sets the approach apart from traditional scoring or regression models (Rodríguez-López et al., 2023). The weakest-link determination is performed programmatically: the system iterates from TRL 1 upward and assigns the highest level where all MUST criteria are satisfied or partially satisfied.

Technology Type Detection. The system classifies projects by analyzing domain-specific keywords in the submitted documentation, assigning one of three criterion sets: hardware, software/digital, or AI/ML. Mixed projects receive dual evaluation with the system TRL defined as the minimum of the component assessments.

Calibration Procedure. Before formal validation, the system was calibrated against a set of projects with previously established TRL assignments. Calibration focused on two aspects of the evaluation logic: the evidence extraction threshold (the minimum textual grounding required for a criterion to be considered satisfied rather than partially satisfied) and the handling of implicit evidence (cases where a document demonstrates criterion satisfaction without explicitly stating it). Calibration continued iteratively until the system produced consistent, criterion-aligned behavior across the calibration set. The calibration cases are excluded from the validation sample reported in Sect. 4.

Figure 3 details the complete system architecture. This diagram outlines the document ingestion pipeline, the structure of the evaluation prompts, and the format of the final output reports.

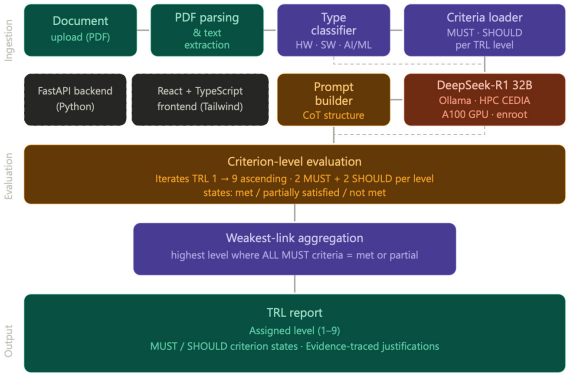

Fig. 3. TRL evaluation pipeline: document ingestion, chain-of-thought criterion evaluation, and weakest-link assignment.

4 Results

4.1 Systematic Review Findings

The 39 reviewed studies reveal a landscape dominated by domain-specific adaptations: 33 (84.6%) address TRL within a single domain, 6 (15.4%) attempt cross-domain generalization, and 5 (12.8%) incorporate an automated assessment component. Figure 4 presents the distribution by category.

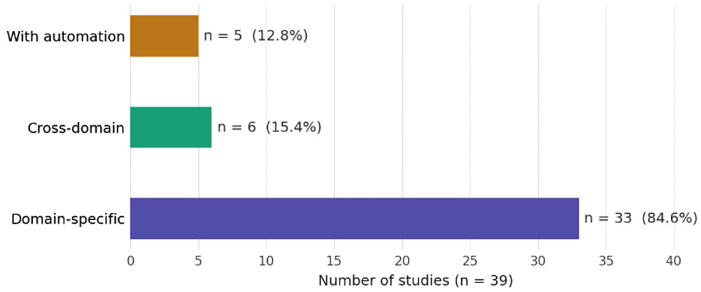

Fig. 4. Distribution of the 39 reviewed studies by domain and assessment approach category.

The 39 included studies span eight application domains: energy and ecological transition, cross-cutting policy and education, manufacturing and logistics, health sciences, environment and agriculture, aerospace, software and IoT, and chemistry and materials (Rea Guamán et al., 2026). This distribution confirms the substantial expansion of TRL beyond its aerospace origins, with energy and environmental technologies as the most active area of methodological development.

Four convergent limitations recur across the corpus and directly informed the methodology design. First, the literature identifies a persistent lack of multidimensionality: TRL evaluates technology in isolation without accounting for commercial, regulatory, or social readiness (Kobos et al., 2018; H.-Y. Liu et al., 2020; Bruno et al., 2020;

Sepúlveda et al., 2025). Second, the corpus converges on high subjectivity and limited reproducibility of expert-panel assessments; multiple authors characterize TRL evaluation without standardized checklists as potentially arbitrary (Rodríguez-López et al., 2023; White et al., 2022; Lacueva-Pérez et al., 2018; Bukar & Asif, 2024; Salvador-Carulla et al., 2024)—motivating the binary MUST/SHOULD criteria. Third, multiple studies demonstrate that technology development is iterative rather than linear, particularly for software and AI/ML (Clawson, 2023; Bruno et al., 2020; White et al., 2022). Fourth, the corpus documents difficulty assigning TRL to complex systems with subsystems at different maturity stages—addressed here through the weakest-link principle (ISO, 2013; U.S. Government Accountability Office, 2020).

Table 1 presents the five studies with automated assessment systems alongside the present work. Kobos et al. (2018) achieve cross-domain coverage through System Dynamics but produce no auditable justification. (Rodríguez-López et al., 2023) share the document-analysis intent but use regression without criterion decomposition. White et al. (2022) implement binary yes/no criteria—the closest precedent to MUST/SHOULD—but are confined to agricultural biosecurity. No prior work simultaneously achieves cross-domain coverage, binary criteria, LLM-based evaluation, and the weakest-link principle.

Table 1. Comparison of automated TRL assessment approaches identified in the systematic review.

Author(s)	Domain	Method	Cross-d	Binary	LLM
Zhong & Yang (2025)	Education	Quant. Algorithm	No	No	No
White et al. (2022)	Agri. Biosecurity	Excel calculator + binary criteria	No	Yes	No
Kobos et al. (2018)	Cross-domain	System Dynamics	Yes	No	No
Lacueva-Pérez et al. (2018)	HCI/manuf	Automated questionnaires (ESA/NASA adapted)	No	Partial	No
Rodríguez López et al. (2023)	SMEs	Text mining + regression	Yes	No	No
This work	**HW + SW + AI/ML**	**LLM + chain-of-thought**	**Yes**	**Yes**	**Yes**

4.2 Framework Design Outcomes

The methodology yields 54 mandatory and 54 supportive criteria (two MUST and two SHOULD per level per type). Technology type detection correctly identified the type in 14 of 17 validation cases (82.4%). Three documents were classified as research-type

projects lacking a deployable artifact—an emergent finding, as this category falls outside the defined scope. Notably, all three achieved exact concordance with the expert assessment, suggesting that criteria at lower TRL levels generalize naturally to investigative academic work.

4.3 System Performance

The validation of the AI evaluation system occurred with an assortment of documents (a total of 17) gathered from public university repository sites in Ecuador, consisting of three types of documents; 1.) nine software development documents, 2.) Three Hardware Development Documents, 3.) two Artificial Intelligence/Machine Learning (AI/ML) Documents, and 4.) three Research Document Types.

The processing time for reviews of a project's technical readiness level (TRL) increased as the TRL increased; specifically, the processing time was greater for evaluations that reached higher levels of TRL. For example, TRL 1–2 evaluations took 126–175 s, but those TRL 6–7 evaluations took 306–437 s. Overall, the average amount of time spent on each document during the evaluation process was about four-and-one-half minutes. This time frame is significantly less than what it takes an expert panel to complete. In fact, Olechowski et al. (2020) found that expert panels often require hours to days to evaluate research projects.

4.4 Validation Against Expert Assessment

Expert TRL assignments were obtained independently for all 17 documents by a domain expert in technology readiness assessment, applying the same binary-criteria methodology used by the system prior to comparison with system outputs. Table 2 presents the complete case-by-case comparison.

Table 2. Validation results: AI system vs. expert assessment across 17 documents.

#	Document	Type	System TRL	Expert TRL	Δ	MUST state
1	P-01	SW	2	2	0	met, met
2	P-02	SW	4	3	+1	partial, partial
3	P-03	SW	6	5	+1	partial, partial
4	P-04	SW	4	4	0	met, met
5	P-05	AI/ML	6	5	+1	partial, partial
6	P-06	SW	2	2	0	met, met
7	P-07	HW	7	7	0	met, met
8	P-08	SW	2	2	0	met, met
9	P-09	Research	2	2	0	met, met
10	P-10	AI/ML	1	1	0	met, met
11	P-11	Research	1	1	0	met, met

(*continued*)

Table 2. (*continued*)

#	Document	Type	System TRL	Expert TRL	Δ	MUST state
12	P-12	Research	4	4	0	met, met
13	P-13	SW	6	6	0	met, met
14	P-14	SW	1	1	0	met, met
15	P-15	SW	5	5	0	met, met
16	P-16	HW	2	2	0	met, met
17	P-17	HW	6	4	+2	met, met

The system achieved exact concordance in 13 of 17 cases (76.5%) and agreement within ±1 TRL level in 16 of 17 cases (94.1%). One case exhibited discordance greater than one level (5.9%), corresponding to a +2 overestimation in a hardware project. These results confirm the primary hypothesis of RQ3: a large language model with chain-of-thought reasoning can replicate expert TRL assessments with sufficient accuracy for academic and research evaluation contexts.

Analysis of the four cases which have been identified as discordant provides insight into two distinct types of failure modes. A primary failure mode found in three of the four discordant cases (#2, #3, #5) is ambiguity regarding the evidence density of the system. In these cases, the system provided a partial state of the disputed level for both of the MUST criteria. Thus, although the system recognizes some relevance of the textual evidence toward satisfying the MUST criteria, it does not recognize enough to fully satisfy them. When the expert evaluator uses his/her domain knowledge to evaluate the same criteria, he/she will judge that the evidence fails to meet the required level threshold. Consequently, this results in a consistent +1 over estimation based on the expert's evaluation when there are sparse amounts of criterion-based evidence, or the criterion-based evidence is distributed throughout the document.

A secondary failure mode has also been identified within one of the discordant cases (#17 HW, $\Delta = +2$). The failure mode identified in this case is environmental equivalence misclassification. Both of the MUST criteria at TRL 6 were evaluated as being met; however, the expert evaluator assigned a TRL value of 4. This indicates that while the system accepts testing in a controlled university environment as being equivalent to field testing in an actual operational environment (a distinction which requires domain specific knowledge beyond what can be recovered through textual analysis), it appears that the system failed to make this distinction. The failure type observed in case #17 (HW, $\Delta = +2$) is potentially more detrimental than the other failure type since it produces greater discordance and is not indicated by partial MUST states thereby providing less indication of such failures from output produced by the system (Fig. 5).

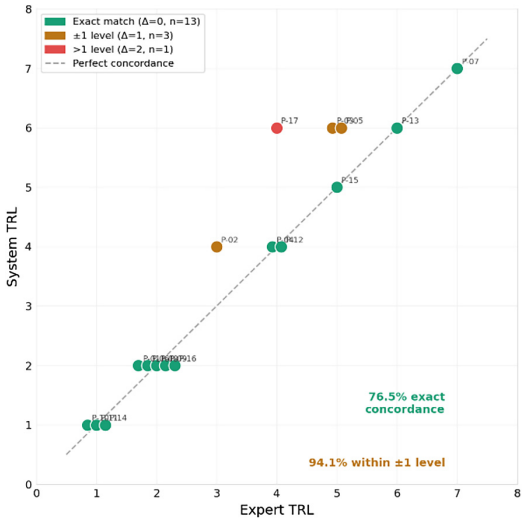

Fig. 5. Comparison between AI-generated and expert TRL assignments for the 17 validation cases. The dashed line represents exact agreement, while the colors show the degree of variation between the scores.

5 Discussion

5.1 Interpretation of Results

The empirical evidence provided in Sect. 4 provides a practical answer to each of the 3 research questions which guided this study.

RQ1 investigated "What patterns or gaps remain to be explored regarding the current state-of-the-art of TRL assessment methodologies?". A comprehensive analysis of 39 papers demonstrates that nearly all (84.6%) reviewed papers have been written about TRL for one particular area of application. No prior study has addressed cross-domains, applied binary assessment methods grounded in two international standards, used LLMs for assessing TRL, and employed the weakest link assignment method. The convergence of these limitations from the entire body of literature identified subjectivity inherent to expert panel assessments, assuming TRL is linear, difficulties with evaluating highly-complex systems and lack of multidimensional assessment as major deficiencies defining the design space for the developed methodology. More importantly, the review outcomes motivated and directly affected certain design decisions in developing the proposed methodology (as described in Sect. 3.3). The use of a binary system for categorizing MUST/SHOULD responses was designed to mitigate the problems of subjectivity documented by Rodríguez-López et al. (2023); White et al. (2022); and Bukar & Asif (2024). The use of the weakest-link principle was devised to address the challenges associated with complex systems as outlined by Clawson (2023) and Bruno et al. (2020). Finally, technology-type specific criteria were included to address issues related to the hardware centric nature of TRL assessment as described by Lavin et al. (2022); and Martínez-Plumed et al. (2021).

Validation provides evidence in support of RQ2. All three technologies were verified with either an exact or ±1 match across all three (hardware at TRL 7 demonstrate application to physical systems, software cases confirmed coverage from TRL 1 through TRL 6, and the AI/ML cases confirm that MLTRL criteria derived are easily integrated). The fourth category of "research type" is a partially covered boundary condition. Research type documents represent the potential for natural expansion of the current iteration and expand the current framework for use on all forms of scholarly output, including systematic reviews, theoretical frameworks/models, and empirical studies, excluding those artifacts related to technological development.

RQ3 receives an affirmative answer: 76.5% exact concordance and 94.1% ±1 agreement supports the viability of LLM-based TRL evaluation for academic contexts. The ambiguous evidence density failure mode is predictable and detectable through partial MUST states—a conservative post-processing rule could reduce overestimation at the cost of false negatives. The environmental equivalence misclassification mode is more consequential, requiring domain-specific context classifiers or mandatory human review of hardware cases at TRL 5+.

This system is the only automated TRL approach simultaneously achieving cross-domain coverage, binary criteria, LLM-based evaluation with evidence-traced justifications, and the weakest-link principle (Table 2). Kobos et al. (2018) achieve cross-domain coverage but require expert-parameterized simulation and produce no auditable justification. Rodríguez-López et al. (2023) share the document-analysis intent but produce statistical estimates where discordant results cannot be diagnosed at the criterion level.

Although this 76.5% concordance is as good or better than agreement between people: Olechowski et al. (2020) have documented how often experts don't agree on assigning TRLs to projects; Salvador-Carulla et al. (2024) reported an ICC of .9 for 10 evaluators and thus needed structured evaluation. This tool generates equivalent quality results through unstructured documents in about 4.2 min/document. Thus, it has improved on other methods in terms of quality and scale.

5.2 Limitations and Threats to Validity

Several limitations bound the interpretation of these results and define the agenda for future work.

Sample Size and Distribution. The validation sample consists of 17 papers that are heavily biased towards the project area for software (9/17 = 52.9%), so future validations would be improved if they employed balanced samples of at least 15 cases per technology area to enable an evaluation of the concordance within each category.

Single Expert Evaluator. One single domain expert was assigned to provide the expert judgments. To provide more confidence in the results of this study, we recommend that future research uses several independent experts, reporting on inter-rater reliabilities, e.g. Cohen's Kappa or ICC values.

Environmental Equivalence Classification. In case #17, the >1 disagreement indicates that the system cannot accurately differentiate between controlled academic test environments and real-world deployment scenarios based on the text description of the

environment. These limitations are most impactful to hardware projects with a Technology Readiness Level (TRL) of 5 or higher as it is the determination of the appropriate environment which will determine the technology readiness level. Strategies to mitigate these limitations may be achieved through the use of domain specific contextual classifiers; by using structured template forms that require an explicit description of the environment and/or have mandatory flags for human review of all hardware projects in the intermediate to high range.

LLM Sampling Variability. Although low temperature sampling was used ($T = 0.1$) there is still a document that has had two separate TRL assignment in an independent run. The fact that it could be done indicates that this program can't be determined to have the same output for documents containing limited or ambiguous foundation proof. Using an ensemble voting method using many independent runs would greatly minimize this variability at the expense of increasing the processing time.

Scope Boundary: Research-Type Documents. Our current methodology focuses specifically on three technology types, intentionally excluding theoretical research documents that lack deployable artifacts. To make this tool useful for all academic outputs, future versions of the framework will need to include specific criteria for these types of papers.

6 Conclusions

This research presents an extended AI-based technology readiness level assessment tool which addresses a previously identified gap from a systematic review of 39 studies. The gap was identified as there are no existing methods or tools to assess technology readiness levels across all three domains (hardware, software, and AI/ML), use multi-standard binary assessment criteria, evaluate technology readiness levels utilizing large language models with chain-of-thought reasoning, and apply the weakest link principle.

The PRISMA-compliant review (C1) has mapped the relevant area of research and established this four-way gap through converging limitations in all areas of the overall study which have been utilized for determining the methodology design.

The generalized methodology (C2), uses binary MUST/SHOULD criteria at nine levels and three different technology types, derived from the ISO 16290 standard, NASA, Department of Defense (DoD), and Government Accountability Office (GAO) standards to provide an explicit means to assign the weakest link based on the formal logic of assignment.

The AI-based evaluation system (C3), utilizes the methodology to evaluate by utilizing DeepSeek-R1 with chain-of-thought reasoning and produces traceable evidence-based explanations of how each criterion is justified such that every decision regarding each criterion can be audited—distinctly separating it from other methods of scoring and regression analysis.

Empirical validation (C4) was conducted on 17 projects and showed that there were 76.5% exact agreements and 94.1% ± 1 agreements. There are two failure modes; ambiguous evidence density (detectable when evaluating partial MUST states) and environmental equivalence misclassifications (domain specific classifiers are required).

Mean processing time of approximately 4.2 min per document indicates a significant decrease compared to the expert panels requiring hours or days to complete their respective processes.

Several key upgrades are planned for the next phase of this research. To improve the evaluation process, we will use balanced samples, formalize criteria for theoretical papers, and introduce multiple human reviewers to track reliability metrics. Additionally, upgrading the system with majority-voting evaluation and retrieval-augmented generation will make locating evidence much more precise.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Bruno, I., et al.: Technology readiness revisited: a proposal for extending the scope of impact assessment of European public services, 369–380 (2020). https://doi.org/10.1145/3428502.3428552

Bukar, A.M., Asif, M.: Technology readiness level assessment of carbon capture and storage technologies. Renew. Sustain. Energy Rev. **200** (2024). https://doi.org/10.1016/j.rser.2024.114578

Clawson, G.: A technology readiness level for blockchain, 249–257 (2023). https://doi.org/10.1145/3555776.3577704

European Commission. Technology readiness levels (TRL), Annex G of the General Annexes. In: Horizon 2020 Work Programme 2014–2015, Extract from Part 19—Commission Decision C(2014)4995. European Commission (2014)

European Cooperation for Space Standardization (ECSS). ECSS-E-HB-11A – Technology readiness level (TRL) guidelines. ECSS Executive Secretariat (2017)

Héder, M.: From NASA to EU: the evolution of the TRL scale in public sector innovation. Innov. J. **22**, 1 (2017)

ISO. ISO 16290:2013—Space systems—Definition of the Technology Readiness Levels (TRLs) and their criteria of assessment. International Organization for Standardization. ISO (2013). https://www.iso.org/standard/56064.html

Kobos, P.H., Malczynski, L.A., Walker, L.T.N., Borns, D.J., Klise, G.T.: Timing is everything: a technology transition framework for regulatory and market readiness levels. Technol. Forecast. Soc. Chang. **137**, 211–225 (2018). https://doi.org/10.1016/j.techfore.2018.07.052

Lacueva-Pérez, F.J., et al.: Assessing TRL of HCI technologies supporting shop floor workers, 311–318 (2018). https://doi.org/10.1145/3197768.3203175

Lavin, A., et al.: Technology readiness levels for machine learning systems. Nat. Commun. **13**(1), 6039 (2022). https://doi.org/10.1038/s41467-022-33128-9

Liu, H.-Y., Subramanian, A.M., Hang, C.-C.: Marrying the best of both worlds: an integrated framework for matching technology transfer sources and recipients. IEEE Trans. Eng. Manage. **67**(1), 70–80 (2020). https://doi.org/10.1109/TEM.2018.2858812

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: NLG evaluation using Gpt-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 2511–2522 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.153

Mankins, J.C.: Technology readiness assessments: a retrospective. Acta Astronaut. **65**(9–10), 1216–1223 (2009). https://doi.org/10.1016/j.actaastro.2009.03.058

Martínez-Plumed, F., Gómez, E., Hernández-Orallo, J.: Futures of artificial intelligence through technology readiness levels. Telematics Inform. **58**, 101525 (2021). https://doi.org/10.1016/j.tele.2020.101525

National Aeronautics and Space Administration (NASA). NASA Systems Engineering Handbook, Revision 2 (NASA/SP-2016–6105 Rev 2.). NASA Headquarters (2017)

Olechowski, A.L., Eppinger, S.D., Joglekar, N., Tomaschek, K.: Technology readiness levels: shortcomings and improvement opportunities. Syst. Eng. **23**(4), 395–408 (2020). https://doi.org/10.1002/sys.21533

Page, M.J., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, n71 (2021). https://doi.org/10.1136/bmj.n71

Rea Guamán, A.M., Almeida Jara, A.S., Masacela-Atupaña, B.: Supplementary material: complete TRL assessment criteria tables and systematic review corpus — for: a generalized AI-powered framework for cross-domain technology readiness assessment. Zenodo (2026). https://doi.org/10.5281/ZENODO.19121085

Rodríguez-López, N., Álvarez-Cabal, J.V., Cuiñas, M.C., Ortega Fernndez, F.: Applicability of technology maturity level evaluation methodologies within small- and medium-sized organizations: prospects and proposals. Systems **11**(8) (2023). https://doi.org/10.3390/systems11080387

Sadin, S.R., Povinelli, F.P., Rosen, R.: The NASA technology push towards future space mission systems. Acta Astronaut. **20**, 73–77 (1989). https://doi.org/10.1016/0094-5765(89)90054-4

Salvador-Carulla, L.B., Woods, C.E., de Miquel, C., Lukersmith, S.: Adaptation of the technology readiness levels for impact assessment in implementation sciences: the TRL-IS checklist. Heliyon **10**(9) (2024). https://doi.org/10.1016/j.heliyon.2024.e29930

Sepúlveda, J., Albis-Arrieta, A.R., Santamaría, J., Cervera, S.: University-industry-government integration in the development of sustainable and cost-efficient solutions for the shrimp sector in Colombia: Nutriaqua case. J. Sustain. Dev. Energy, Water Environ. Syst. **13**(2) (2025). https://doi.org/10.13044/j.sdewes.d13.0557

Tomaschek, K., Olechowski, A., Eppinger, S., Joglekar, N.: A survey of technology readiness level users. INCOSE Int. Symp. **26**(1), 2101–2117 (2016). https://doi.org/10.1002/j.2334-5837.2016.00283.x

U.S. Department of Defense. Technology Readiness Assessment Guidebook. Office of the Under Secretary of Defense for Research and Engineering (2025). https://www.cto.mil/wp-content/uploads/2025/03/TRA-Guide-Feb2025.v2-Cleared.pdf

U.S. Government Accountability Office. Technology Readiness Assessment Guide: Best Practices for Evaluating the Readiness of Technology for Use in Acquisition Programs and Projects. https://www.gao.gov/products/gao-20-48g. Accessed 7 Jan 2020

Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. arXiv: (arXiv:2201.11903) (2023)

White, R.M., et al.: Technology development for the early detection of plant pests: a framework for assessing technology readiness levels (TRLs) in environmental science. J. Plant Dis. Prot. **129**(5), 1249–1261 (2022). https://doi.org/10.1007/s41348-022-00599-3

Zheng, L., et al.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv: (arXiv:2306.05685) (2023)