Integrating Neurosymbolic Systems in Advanced Product Design: A Comprehensive Review

Summary of revisions for the minor-revision resubmission

This following matrix, summarizes the substantive revisions and points to where each can be found. Locations are given by section and named paragraph rather than by page, so they remain valid if pagination shifts; figure and table numbers refer to the revised manuscript. A separate point-by-point letter reproduces each reviewer comment in full and responds to it directly; this document is the navigation summary for that letter.

Sections are now numbered in the revised manuscript, so locations below are given by section number and named paragraph; it is more convenient to follow and keeps cross-referencing proper even if pagination shifts.

A. Reviewer comments and corresponding revisions

#	Reviewer comment or concern	Raised by	What changed and how it addresses the comment	Where to find it
1	Search coverage: engineering-specific databases (Compendex, ASME, Knovel) were not consulted; corpus judged too narrow.	Jaimini	We added a supplementary search of Engineering Village/Compendex, the ASME Digital Collection and Scopus under the same eligibility criteria, retrieving 33 records of which 15 were eligible. Knovel was assessed and excluded, with the reason stated: it indexes handbooks and materials-property data rather than primary methodological literature. The eligible studies were merged into the single classification rather than appended as a separate table, and all 33 records are retained in the ORKG comparison so the eligibility boundary is inspectable.	• Sec. 2, 'Information sources and search strategy' and 'Supplementary engineering-database search' • Fig. 1 (PRISMA, two streams) • Table 3 • Sec. 10
2	Survey is descriptive; needs aggregate statistics not just method-by-method narration.	Dalal; Sarker	We have now also included a quantitative corpus profile over the 38 classified systems: distribution of integration types, symbolic substrates, neural functions and architecture families, evaluation domains, and artifact availability. The narrative claims about where loose and tight couplings concentrate are now stated against these distributions for statistical and evidence-based discussion.	• Sec. 4, 'Distribution of integration types', 'Symbolic substrates and neural functions', 'Neural configurations', 'Domain-specific patterns' • Figs. 3, 5, 6, 7 • Sec. 7.6 and Fig. 10 (artifact availability)
3	Types I–VI are not compared on scalability, interpretability, computational cost and deployment feasibility. Types IV–VI are discussed in less depth than Types I–III.	Dalal; Sarker	The comments are addressed through the evaluation and risk frameworks in sections 7,8: ➢ Sec. 7.5 sets out how computational cost should be measured across the complete hybrid pipeline, including symbolic reasoning cost as ontology size grows; Sec. 7.3 treats interpretability as explanation traceability rather than as a property inferred from architecture; ➢ Sec. 8 treats constraint verification at scale and deployment constraints as type-dependent risks. ➢ Sec. 4 now explains why the distribution looks this way: loose couplings keep the symbolic part separately testable and fit existing CAD, simulation and PLM toolchains, while tight couplings need a differentiable form of the constraint, available for geometry and physics but not for discrete manufacturing rules. Sec. 8 covers the operational side. ➢ On Types IV–VI, Sec. 5.6 identifies a specific gap: the corpus contains only couplings that check once and stop, because engineering verifiers return a pass or fail rather than a diagnosis a generator could act on. Sec. 6 pairs each transfer pattern with the condition for adoption.	• Sec. 4, 'Distribution of integration types' • Sec. 5.6, 'From checking to actionable feedback' • Sec. 6 • Sec. 7.3, 7.5 • Sec. 8.1–8.2
4	Evaluation discussion is too high-level; a concrete, standardized evaluation or	All three	We replaced the previous discussion with an operational protocol: define the task and symbolic specification, establish an unperturbed baseline, apply targeted input and symbolic interventions, compare against the baseline, and report system-level evidence. Task-specific measures and formulations are given	• Sec. 7 in full • Sec. 7.1 (protocol)

#	Reviewer comment or concern	Raised by	What changed and how it addresses the comment	Where to find it
	benchmark framework is needed.		for CAD generation, topology optimization, manufacturing and assembly planning, and prediction and quality control. A reporting rubric organises these into three levels — engineering result, symbolic contribution, and system credibility — and states, for each measure, the evidence that must accompany the number.	• Sec. 7.2 'Task-Specific Engineering Performance' • Fig. 9 • Table 4 (rubric)
5	No means of detecting reasoning shortcuts or testing whether symbolic knowledge is actually used rather than memorised.	Jaimini	We separate now constraint satisfaction from reasoning fidelity as distinct questions. Constraint compliance is quantified through the Constraint Satisfaction Rate, with the enforcement mechanism (by construction, soft penalty, or post-hoc filtering) required alongside the rate. Symbolic contribution is estimated through a symbolic-side ablation, and a stronger diagnostic revises the specification and tests whether the output changes consistently with it. Applying this to the corpus, we realized that: no reviewed system isolates the symbolic contribution through a controlled symbolic-side ablation, and none tests sensitivity to a deliberately revised specification. Existing ablation studies mostly report ablations on model or architectural choices.	• Sec. 7.3, 'Constraint compliance', 'Symbolic contribution and reasoning fidelity', 'Explanation traceability' • Eqs. (7) – (8) • Fig. 9 (diagnostic probes) • Table 4 (symbolic contribution block)
6	Reproducibility markers — code, data, hardware, model settings and symbolic artifacts — are not discussed.	Jaimini	We suggest an extended reproducibility practice beyond release of the learned model to the symbolic artifacts on which the claimed behavior depends: ontologies, knowledge graphs, rule sets, constraints, formal specifications and solver configurations, together with seeds, hardware, preprocessing and tolerances. For proprietary industrial data we state what can still be documented when the records cannot be released. We also added a corpus-level artifact-availability analysis, which shows that industrially evaluated systems in this corpus release neither code nor data. The ORKG comparison was extended with constraint/symbolic prior, dataset/benchmark, key metrics and evaluation basis, and the version accompanying this manuscript is archived under DOI: doi.org/10.48366/R1937457	• Sec. 7.6, 'Reproducibility' • Fig. 10 • Sec. 10 • Table 4 (system credibility block)
7	Industrial relevance, CAD/PLM integration and the lab-to-industry gap are underdeveloped.	Sarker; Jaimini	➢ Deployment evidence is now graded: synthetic or simulated data, retrospective evaluation on real industrial data, expert-in-the-loop evaluation, and operation in a deployed production workflow are distinguished, and the role of the engineer must be stated. ➢ Sec. 8.2 now acknowledges the "operational risks" related to CAD/PLM interoperability, latency in engineering decision loops, and human validation required for specific hybrid systems, and discusses concrete practices for dealing with each: align symbolic entities with identifiers already used by the PLM system, set the latency budget from the process rather than the model, and route cases to a reviewer by verification outcome instead of sending everything. ➢ On the point that some included methods are general AI systems: the classification now records where each system was actually evaluated, and nine are marked 'general benchmark'. Sec. 6 calls them transfer candidates and pairs each pattern with the condition that would have to hold before it could be adopted.	• Sec. 7.6, 'Deployment validity' • Sec. 5.4 • Sec. 8.2 • Fig. 11 • Table 3, 'Dom' column • Sec. 6
8	A concrete industrial case study is needed to ground the survey.	Jaimini	We now review a case study of HCP-KG, a deployed quality-control system in automotive body-shop and gear-manufacturing settings. It states the neural and symbolic roles separately, places the system in Type III, reports the task and deployment evidence, and then applies our own rubric to it. The case is further used to demonstrate the rubric of Sec.7: it scores on task performance and deployment validity, but reproducibility cannot be assessed and the symbolic contribution is never isolated, so the reported accuracy does not isolate the	• Sec. 5.4 • Fig. 8 • Sec. 7.6, 'Deployment validity' (cross-reference)

#	Reviewer comment or concern	Raised by	What changed and how it addresses the comment	Where to find it
			contribution of the symbolic component from that of the learned component.	
9	Quality assurance is discussed in the text but absent from Table 3; materials and sustainability are underrepresented.	Jaimini	Quality assurance is now a coded domain in the classification with five entries, and assembly planning has seven; both increased through the engineering-database search and are among the better-represented domains. Materials (two entries) and sustainability remain sparse, and we report that also as a finding and discuss it in detail. Some of the work in these areas is close to what we cover but does not meet the eligibility criteria. Neural models that fill gaps in life-cycle inventories, and ontologies for remanufacturing process planning, are two examples: the learning and the symbolic model sit side by side without interacting. We cite them where they give useful context, record them in the ORKG comparison as adjacent entries, and say why they are not counted.	• Sec. 4, 'Domain-specific patterns' • Fig. 7 • Sec. 5.2, 5.5 • Sec. 2, 'Supplementary engineering-database search'
10	No visual evidence of what the reviewed systems actually produce.	Jaimini	Partially addressed. Fig. 8 shows the operational interface of a deployed system, which is direct evidence of one system in use. We did not add a composite figure of generated designs, topologies or plans: reproduction of published output figures across the corpus is constrained by licensing, and rendering our own would not represent the cited work. The figures now carry statistical data or evaluation and categorization of methods. The corpus profile shows what the 38 systems actually look like, the availability figure shows who releases code and data, and the protocol figure shows the evaluation procedure step by step. Anyone wanting to see a specific system's output can follow it through the ORKG comparison to the source paper.	• Fig. 8 • Figs. 3, 5, 6, 7, 9, 10 • Sec. 10
11	Risk, ethics and regulation discussion is generic and not specific to product design or manufacturing.	Dalal	Section 8 was fully rewritten. We replaced the previous Technological/Ethical/Regulatory structure, which also applied to many generic AI systems, with a taxonomy organized by where the failure originates: Inside the hybrid architecture, at its interface with engineering infrastructure, or in the institutional context that must accept its outputs. Each of the nine items now names a failure mode specific to hybrid engineering systems and a practice that responds to it — versioning the model and knowledge base as one release, tiered constraint checking, treating ontologies as reviewed code, keying symbolic entities to PLM identifiers, decision records for accountability, and a change policy that classifies which updates require re-verification.	• Sec. 8 in full • Fig. 11 • Cross-references to Sec. 7.3, 7.5, 7.6
12	Bibliography could use more industrial AI/CAD and benchmark-oriented references; sustainability and LCA are mentioned but poorly cited.	Sarker	Added engineering work across CAD generation, assembly and process planning, quality assurance, topology optimization and industrial knowledge graphs, plus recent benchmarking practice for constraint satisfaction and for neurosymbolic description-logic reasoners. These systems sit in the same classification as the general-benchmark transfer candidates under one coding scheme, with the evaluation domain distinguishing the two so the reader can see which claims rest on engineering evidence. On sustainability, we added established life-cycle-assessment work on inventory prediction, LCA-driven multi-objective optimization in materials selection, and automated LCA for additive manufacturing. We kept Bougzime et al. and note that it appeared at NeSy 2025, the field's dedicated venue. The text now says why the LCA work is cited but not classified — the inventory model stays outside the learner — and connects it to the ontology-based sustainability reasoning it would need to support.	• Table 3 • Sec. 4 • Sec. 5.5 • Sec. 6 • Sec. 7.3, 7.5

#	Reviewer comment or concern	Raised by	What changed and how it addresses the comment	Where to find it
13	Tighten repetitive explanations and make the dense tables more accessible.	Jaimini	Shortened the classification table caption: the coding definitions it repeated now live only in the abbreviations and notation table, and the per-paper extraction detail lives in the ORKG comparison, which has no width limit. Cut repetition in the classification and applications sections, including two paragraphs that reported the same domain-coverage finding twice. Figure captions now say what each figure is for, and the lifecycle figure (also rotated for readability) is labelled a survey lens so it is not read as a proposed architecture.	• Table 3 caption • Table 1 • Sec. 4 • Caption of Fig. 4 • Sec. 10
14	Broken cross-references, inconsistent spelling of 'neurosymbolic', and presentation errors.	Sarker	➢ We have numbered the sections and subsections, which gives every cross-reference a stable target and makes the paper easier to navigate and cite. We repaired the broken references, including the one Reviewer Sarker flagged. ➢ We ran an overall manuscript consistency pass, unified the terminology on 'neurosymbolic' and the abbreviation 'NeSy', keeping variants only inside search strings and official method names, and updated the abbreviations and notation table to match. Tidied the remaining spacing and typography artifacts.	• Whole manuscript • Table 1 • Sec. 4, 5, 7, 10 • All section and subsection headings

B. Cross-cutting revisions not tied to a single comment

Revision	What changed	Where to find it
Abstract and conclusion updated to align with the revision	Both now report the corpus size, the 38-system classification, the dominant integration pattern, the evaluation gap, the proposed protocol and the industrial case.	Abstract; Sec. 9
Practical guidance on LLM components	We added a subsection covering how language-model roles, retrieval and knowledge-graph grounding, and parameter-efficient fine-tuning fit together in engineering pipelines (e.g., RAG, PEFT, LoRA, ...), and identifying the gap between execution checking and actionable symbolic feedback for iterative repair. It states explicitly that parameter-efficient fine-tuning adapts the neural component and is not itself a neurosymbolic mechanism.	Sec. 5.6
Living resource aligned with the static snapshot	The 38-system manuscript table is now aligned with an extended ORKG comparison containing the classified systems, the adjacent studies retrieved by the same searches, and the surveys cited as background. The version accompanying this manuscript is archived under DOI 10.48366/R1937457; the comparison itself remains maintained.	Sec. 2; Sec. 10; Table 3 caption