ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

LLM Hallucinations Suppression in Safety-Critical Engineering Drawing Conversion

Chaoyang Zhang, Shu Wu*

University of Shanghai for Science and Technology, Shanghai, China

Abstract: *Converting engineering drawings from Chinese GB to German DIN standards demands near-perfect accuracy: even a single hallucinated value or a misplaced annotation can render a drawing unsafe for manufacturing. Large language models (LLMs) are fluent, but they lack spatial reasoning and often generate uncontrolled outputs—two fatal flaws in this domain. We therefore design a neuro-symbolic framework that couples an LLM with a knowledge-based engineering shell. The core components are a Trident Router that assigns each annotation to a risk-appropriate path, an ontology-driven retrieval-augmented generation (RAG) module with entity masking, a hard symbolic validation loop, and a spatial-aware CAD layout adapter. Before generation, structural constraints protect critical entities; after generation, symbolic checks reject any output that violates standard or format rules. Together, these mechanisms prevent technical hallucinations (e.g., altered tolerances or standard codes) and layout collisions when the translated text is reinserted into the CAD drawing. By detailing the logic and theoretical operation of each module, we show how deep neuro-symbolic integration turns an LLM from an unconstrained text generator into a reliable, auditable, and standard-compliant industrial assistant.*

Keywords: *Engineering drawing standardization; neuro-symbolic AI; large language models; hallucination control; retrieval-augmented generation; CAD spatial reasoning*

1. Introduction

Engineering drawings are the authoritative language of manufacturing. They convey geometry, design intent, tolerances, material specifications, surface finish requirements, and assembly procedures in a compact, standardized notation that must be interpreted without ambiguity anywhere in the world. As global supply chains deepen, the need to convert drawings between national standards has become routine. One of the most demanding such conversions is from the Chinese national standard family (GB) to the German standard family (DIN), a pairing that arises whenever Chinese suppliers deliver precision components to German automotive, aerospace, or machinery customers. The stakes are high: errors cause unfit parts, failed inspections, warranty disputes, and, in safety-critical applications, potential harm. Standardization itself carries geopolitical weight, as national standard systems reflect industrial policy and international influence [1].

Current practice relies on scarce bilingual engineering experts who must simultaneously understand both standard systems, the domain-specific vocabulary of each annotation, and the spatial constraints of the CAD drawing. This approach is slow, expensive, and difficult to scale. A single mistake in a tolerance band, a material grade, or a heat-treatment specification can cascade into serious downstream consequences. There is therefore strong motivation to automate the conversion process, provided that automation can match or exceed human accuracy.

Recent large language models (e.g., GPT-4 and its successors) offer a compelling starting point for automation because they are fluent in both source and target languages and have absorbed vast amounts of technical text during pre-training. However, engineering AI must be knowledge-grounded rather than purely data-driven [2]. Direct application of an LLM to annotation conversion fails for three fundamental reasons. First, near-zero hallucination tolerance: a wrong standard code or a shifted tolerance value is not merely embarrassing—it is dangerous. Unlike general-purpose translation, where a slightly imprecise word choice is acceptable, engineering annotation conversion requires exact fidelity to numerical values, standard identifiers, and material designations. Second, cross-standard semantic alignment: GB and DIN differ not only in terminology but also in underlying assumptions about measurement conventions, inspection methods, and default conditions. Bridging these differences requires a rich ontology and retrievable evidence, as knowledge-based engineering (KBE) has long emphasized [3, 4]. Third, CAD

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

spatial fidelity: German annotations are typically longer than their source-language counterparts, and naively replacing the original text with German text frequently causes collisions with adjacent annotations, broken leader lines, or text that overflows its bounding box and becomes illegible.

We therefore propose a neuro-symbolic, spatial-aware LLM framework in which the LLM serves as a controlled generator embedded within a pipeline that retrieves evidence, masks critical entities before generation, validates outputs symbolically after generation, and checks spatial feasibility before reinsertion. The main contributions of this paper are twofold: (1) reframing annotation conversion as a standard-aware, knowledge-intensive transformation that requires explicit representation of both linguistic and geometric constraints; and (2) a neuro-symbolic architecture that structurally eliminates hallucination and spatial conflict through layered, deterministic safeguards rather than relying on prompt engineering alone. A companion paper provides empirical validation on a corpus of real mechanical drawings. The remainder of this paper is organized as follows: Section 2 reviews related work in KBE, RAG, and AI for CAD; Section 3 details the system architecture; Section 4 presents theoretical application scenarios; Section 5 discusses limitations and future directions; and Section 6 concludes.

2. Theoretical Background and Related Work

2.1 Knowledge-Based Engineering

Engineering automation has long recognized that explicit, formalized knowledge is indispensable for reliable decision-making. KBE emerged as a discipline precisely to capture and reuse engineering expertise in computational form. Kügler et al. [4] trace KBE's evolution from hand-coded rule systems of the 1980s toward flexible, ontology-driven reuse frameworks suited to modern collaborative design environments. Sun et al. [5] demonstrate that structured knowledge templates can guide complex multidisciplinary design optimization, reducing both design time and error rates. Madhusudanan et al. [6] show that NLP techniques can extract formal rules from natural-language engineering documents such as aircraft assembly manuals, making it possible to populate KBE systems semi-automatically. Melluso et al. [3] find that combining knowledge graphs with NLP substantially improves interoperability across heterogeneous Industry 4.0 systems, a finding directly relevant to cross-standard conversion. Ontology-based approaches have also been applied to inspection checklist generation [7] and digital twin modeling [8], confirming the breadth of KBE's applicability.

Despite these advances, traditional KBE systems struggle with the ambiguous, free-text annotations that populate real engineering drawings. Rules written for structured data do not transfer gracefully to the open-ended linguistic variation found in manufacturing notes, process instructions, and inspection requirements. Our framework addresses this gap by wrapping LLM-based language understanding inside a KBE-inspired shell: ontology-driven retrieval provides the structured evidence, and hard symbolic validation enforces the rules.

2.2 Retrieval-Augmented Generation

RAG [9] was introduced to ground LLM outputs in external, verifiable knowledge rather than relying solely on parametric memory. By retrieving relevant documents at inference time and conditioning generation on those documents, RAG substantially reduces factual errors in knowledge-intensive tasks. RAG has since been applied across a wide range of domains, including cognitive digital twins [10], generative design review [11], product reconstruction [12], and industrial diagnostics [13]. These applications confirm that retrieval grounding is a broadly effective strategy for improving LLM reliability.

However, RAG alone does not eliminate hallucination in safety-critical settings. The LLM may still paraphrase retrieved content incorrectly, interpolate between retrieved facts in ways that introduce errors, or ignore retrieved evidence when its parametric prior is strong. Tighter controls have been proposed, including entity-linked constraints that anchor generation to specific knowledge-base entries [14] and guided decoding techniques that constrain the output token distribution to a predefined grammar [15]. Our architecture builds on these ideas by adding a hard symbolic validation loop that rejects any generated output violating entity preservation or standard compliance—a near-zero-hallucination gate that operates independently of the LLM's internal confidence.

2.3 AI for CAD and Engineering Drawing Processing

Most AI research applied to CAD and engineering drawings has been analytical rather than generative.

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

Vision-language models have been applied to extract structured information from drawing images [16]; multimodal LLMs have been used in parametric CAD modeling [17]; convolutional and graph neural networks have been developed for machining feature recognition [18]; attentive graph networks have been proposed for assembly analysis [19]; explainable AI methods have been applied to manufacturing cost estimation [20]; tensor factorization has been used for CAE model preparation [21]; automation of trimming die design inspection has been explored [22]; and geometric query methods have been developed for collision analysis [23]. These systems output labels, feature trees, or analysis results—they do not modify the drawing itself.

Our framework is fundamentally different in that it is generative: it modifies the drawing by replacing source-language annotations with target-language annotations that are both semantically correct and spatially compatible. The spatial-aware layout adapter treats annotation reinsertion as a constrained optimization problem that must preserve leader-line topology and satisfy DIN readability requirements. This generative, spatially-aware capability distinguishes our work from all prior art reviewed above.

3. System Architecture Design

3.1 Overall Architecture and Data Parsing

The proposed neuro-symbolic framework transforms a source GB drawing D_{GB} into a target DIN-compliant drawing D_{DIN} through a sequential, multi-layered pipeline. As illustrated in Figure 1, the overall architecture integrates six functional layers: CAD parsing, annotation understanding, risk-aware routing, ontology-driven RAG with entity masking, neuro-symbolic validation, and spatial-aware CAD reconstruction. Formally, the task is expressed as:

$$D_{DIN}=F(D_{GB}, K, R, S)$$

where K denotes the version-controlled standard knowledge base (containing terminology ontologies, standard clauses, and historical conversion cases), R the set of symbolic validation rules (encoding entity preservation, terminology compliance, evidence consistency, and format constraints), and S the spatial layout constraints (minimum text height, collision avoidance, and leader-line topology preservation).

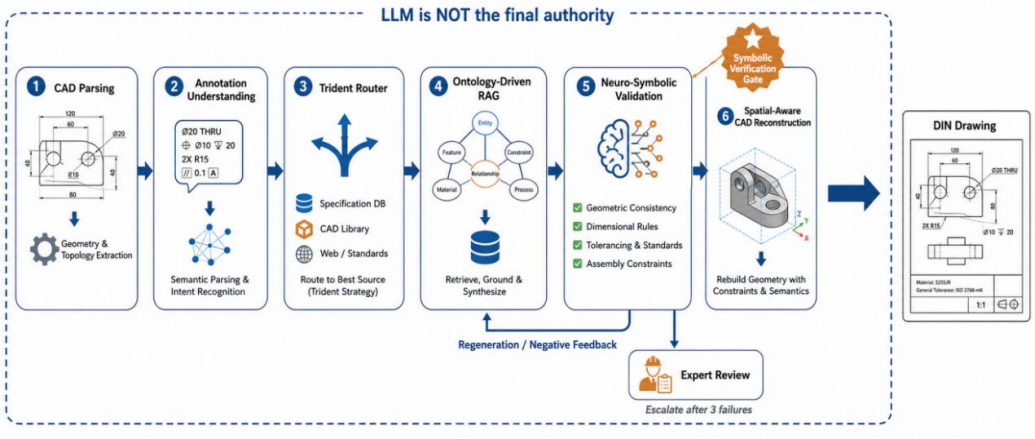

Figure 1. Overall architecture of the proposed neuro-symbolic and spatial-aware LLM framework.

The governing principle of the entire architecture is simple but non-negotiable: the LLM is never the final decision maker. Every annotation, before it is reinserted into the drawing, must pass three independent checks—it must be backed by retrieved evidence, confirmed by symbolic validators, and verified for spatial feasibility. This principle shapes every design decision in the pipeline, from the choice of a JSON output schema to the order in which spatial adjustments are attempted.

The pipeline begins with a **CAD parsing layer** that reads DXF/DWG files and extracts a structured record for each text item t_i: raw text string, layer assignment, insertion point coordinates (x_i, y_i), bounding box dimensions (w_i, h_i), rotation angle θ_i, font name and size, and a reference link ref_i to any associated leader line or geometric shape. This geometric metadata is not merely bookkeeping—it is the foundation for all subsequent spatial reasoning. Without precise knowledge of where each annotation sits in the drawing and what it is connected to, the layout adapter cannot make safe repositioning decisions.

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

Following parsing, a **technical entity recognizer** scans each text item to identify and classify key tokens: standard codes (e.g., GB/T 11354, DIN EN ISO 6507), material grades (e.g., 45 steel, C45E), tolerance specifications (e.g., H7/f6, ±0.02), surface finish symbols (e.g., Ra 1.6), thread callouts (e.g., M12×1.5-6H), and process instructions (e.g., heat treatment, coating requirements). Each recognized token receives a semantic tag and, where applicable, a normalized numerical value. The combination of geometric records and semantic annotations constitutes a rich structured representation that feeds all downstream modules and makes the pipeline's behavior fully traceable.

3.2 Risk-Aware Routing and Knowledge-Grounded Generation

Not all annotations carry the same risk, and treating them uniformly is both inefficient and unsafe. A simple general note such as "remove burrs" carries no critical numerical values and has a well-established German equivalent; translating it through a full neuro-symbolic pipeline would waste computational resources. Conversely, a complex annotation referencing multiple interdependent standards and containing critical numerical ranges whose alteration could render a component unsafe must not be routed through a simple dictionary lookup.

The **Trident Router**, whose decision flow is depicted in Figure 2, addresses this heterogeneity by computing a risk score ρ_i for each annotation based on four weighted factors: (1) the number of protected entities identified by the entity recognizer, (2) the degree of standard dependency (how many distinct standard codes are referenced), (3) wording ambiguity (measured by the entropy of the top-k retrieval results), and (4) expected German text expansion (estimated from character-count statistics of historical conversions). Two thresholds $\tau_1 < \tau_2$ partition the risk space into three routing paths.

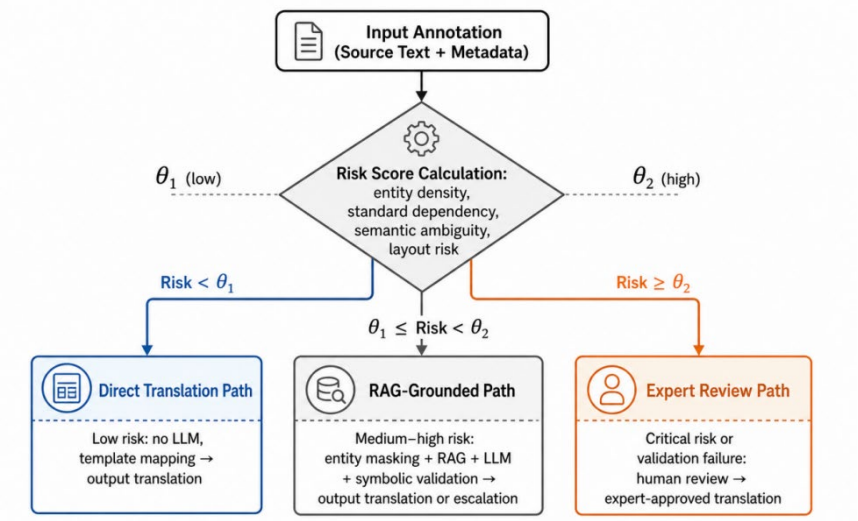

Figure 2. Trident Router decision flow.

Annotations with $\rho_i < \tau_1$ are classified as low-risk and routed to **Direct Translation**, which uses a curated bilingual dictionary without invoking the LLM. This path is fast, deterministic, and hallucination-free by construction. Annotations with $\tau_1 \leq \rho_i < \tau_2$ are classified as medium-to-high risk and routed to **RAG-Grounded Conversion**, the full neuro-symbolic pipeline described below. Annotations with $\rho_i \geq \tau_2$—those for which the knowledge base lacks a reliable mapping or the system's confidence falls below a minimum threshold—are routed to **Expert Review**. This escalation path is a deliberate architectural choice that prioritizes safety over full automation; it is not a failure mode but a designed boundary condition.

When an annotation enters the RAG path, the pipeline first applies **entity masking**: every critical entity identified by the recognizer is replaced by a typed placeholder token before the text is passed to the LLM. For example, "HRC 58–62" becomes <MASK_HARDNESS_VALUE>, "GB/T 11354" becomes <MASK_STANDARD_CODE>, and "45 steel" becomes <MASK_MATERIAL_GRADE>. This masking operation serves two purposes. First, it prevents the LLM from modifying protected values during generation, since the model never sees the actual numbers or codes and therefore cannot hallucinate alternatives. Second, it establishes a formal contract with the downstream symbolic validator: every mask token in the source must appear—either identically restored or through an approved mapping—in the generated output.

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

After masking, the system queries the version-controlled knowledge base K. Retrieval is performed using a hybrid strategy that combines dense semantic matching (via a domain-fine-tuned embedding model), lexical overlap scoring (BM25), and ontology consistency checking (verifying that retrieved clauses belong to the same standard family and version as the source annotation). The retrieval strategy explicitly favors the most recent editions of both GB and DIN standards to avoid introducing deprecated terminology. The top-k retrieved passages, together with the masked source annotation, are assembled into a structured prompt bundle that also includes: target-language grammatical rules, a list of forbidden translation patterns derived from known error cases, and a required JSON output schema specifying the fields that the LLM must populate (translated text, confidence score, referenced standard clauses, and a rationale string). The JSON schema makes the generation auditable and machine-verifiable, enabling the downstream validator to parse and check the output programmatically rather than relying on pattern matching over free text.

3.3 Symbolic Verification and Spatial Reconstruction

The generated JSON is intercepted by a deterministic symbolic validator before any CAD reinsertion can occur. As shown in Figure 3, the validator applies four categories of hard checks, each targeting a distinct class of potential error, and feeds failure reasons back to the LLM for regeneration when necessary.

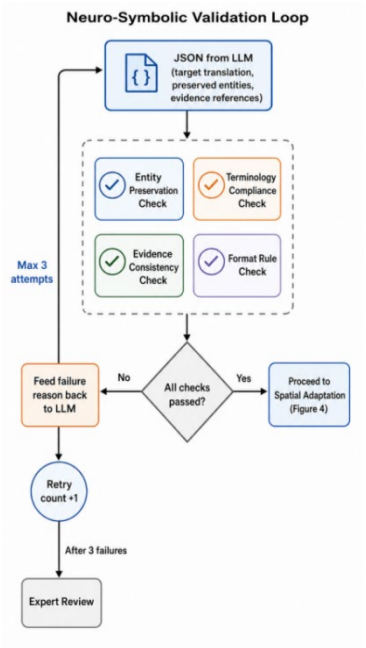

Figure 3. Neuro-symbolic validation loop.

Entity preservation verifies that every masked entity present in the source annotation appears in the target annotation either identically (for numerical values such as hardness ranges and tolerance limits, where exact numeric equality is required) or through an entry in the curated cross-standard mapping table (for material grades and standard codes, where GB designations must map to their DIN equivalents). Any missing or altered entity causes an immediate check failure.

Terminology compliance checks that every technical term in the generated output belongs to the approved DIN vocabulary for the relevant annotation category. Terms that are grammatically correct German but not part of the DIN technical lexicon are flagged as non-compliant. This check catches cases where the LLM substitutes a colloquial or obsolete term for the required standard term.

Evidence consistency requires that any standard clause referenced in the generated output has a semantically matching passage in the retrieved evidence set. This check prevents the LLM from citing standard clauses that were not retrieved—a subtle form of hallucination in which the model generates plausible-sounding but unverified references.

Format and unit rules validate that decimal separators follow DIN convention (comma, not period), that tolerance notation conforms to DIN ISO 286, that surface finish symbols use the DIN EN ISO 1302 format, and that units are expressed in SI notation consistent with DIN standards.

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

If any check fails, the failure reason is encoded as a negative constraint and appended to the prompt for a new generation attempt. The regeneration loop permits up to three attempts, with each iteration providing the LLM with increasingly specific guidance about what went wrong. After three consecutive failures, the annotation is automatically escalated to Expert Review. This dual-loop mechanism—LLM generation followed by symbolic validation, with feedback-driven regeneration—ensures that no annotation enters the final drawing without passing every symbolic gate. The combination of entity masking (pre-generation protection) and symbolic validation (post-generation verification) creates a defense-in-depth strategy against hallucination.

Passing symbolic validation is necessary but not sufficient for drawing usability, because semantic correctness does not guarantee spatial compatibility. German technical annotations are typically 20–40% longer than their source-language counterparts, and this expansion frequently causes the translated text to overflow its original bounding box, collide with adjacent annotations, or detach from its leader line. The **spatial-aware layout adapter** addresses this problem by treating annotation reinsertion as a constrained optimization, as illustrated in Figure 4.

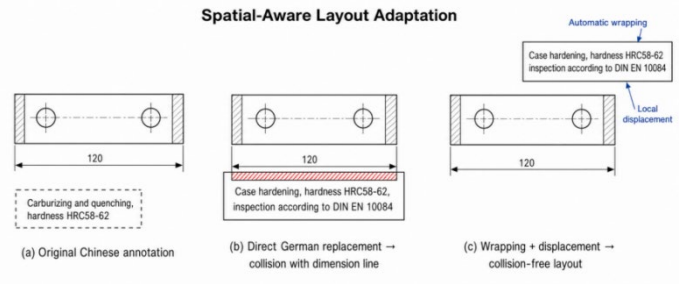

Figure 4. Spatial-aware layout adaptation example.

For each validated annotation, the adapter first estimates the new bounding box $B_i' = (w_i', h_i')$ from font metrics, per-character width statistics for the target font, and the rotation angle θ_i. It then checks whether B_i' intersects any existing drawing element—other annotations, dimension lines, geometric boundaries, or title block regions. If a collision is detected, the adapter attempts a sequence of corrective adjustments in order of increasing intrusiveness: (1) **automatic word wrapping**, which redistributes the text across multiple lines while keeping the insertion point fixed; (2) **small positional shift**, which moves the insertion point by the minimum distance needed to eliminate the collision while preserving the leader-line connection; and (3) **font size reduction**, which scales the text down by a controlled amount, subject to a minimum readable size constraint derived from DIN drawing standards. The optimization objective balances overlap penalty, displacement magnitude, leader-line distortion, and a readability score.

If none of the three adjustment strategies produces a collision-free, readable layout, the annotation is flagged for manual layout adjustment and included in the audit log with a detailed description of the conflict. Once a feasible layout is found, the system replaces the original text entity in the DXF/DWG file, updates the bounding box, rotation, and font size fields, and writes the revised drawing to disk. A complete audit log accompanies every output drawing, recording for each annotation: the original text, the routing decision and risk score, the retrieved evidence passages, the generated output and its validation results, the number of regeneration attempts (if any), and the spatial adjustments applied. This log provides full traceability for quality assurance and regulatory compliance purposes.

4. Workflow and Theoretical Application Scenarios

To make the architecture concrete, we trace three representative annotations through the pipeline, each exercising a different routing path. These scenarios are drawn from a real mechanical drawing of a carburized gear shaft and collectively demonstrate the full decision space covered by the Trident Router.

Scenario 1 — Low-risk direct translation.

The annotation reads "remove burrs." The CAD parsing layer extracts the text and records its insertion point, bounding box, and layer. The entity recognition module finds no protected entities: there are no standard codes, numerical values, material grades, or tolerance specifications. The Trident Router computes a risk score below τ_1 and routes the annotation to Direct Translation, as shown in the left branch of Figure 2. The bilingual dictionary returns the standard DIN equivalent "Entgraten." The spatial adapter computes the bounding box of the German text, confirms that it fits within the original bounding box without collision, and completes insertion without any positional adjustment. Total processing time

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

is negligible, and the result is deterministically correct.

Scenario 2 — Medium-risk RAG-grounded conversion.

The annotation specifies surface carburizing and quenching, a hardness range of HRC 58–62, a case depth range of 0.8–1.2 mm, and references GB/T 11354 as the inspection standard. The entity recognizer identifies four protected entities: a process instruction, a hardness range, a case depth range, and a standard code. The risk score falls between τ_1 and τ_2, and the annotation is routed to RAG-Grounded Conversion via the middle branch of Figure 2. Entity masking replaces the four entities with typed placeholders. Retrieval returns the DIN EN ISO 2639 clause on case-hardening depth measurement and the DIN EN ISO 6508 clause on Rockwell hardness testing, both of which are semantically consistent with the source annotation. The LLM generates a German annotation referencing DIN EN ISO 2639 and DIN EN ISO 6508 in place of GB/T 11354. The symbolic validator, operating as depicted in Figure 3, confirms that all four masked entities are correctly restored, that all terms belong to the DIN vocabulary, that the referenced clauses match the retrieved evidence, and that the format complies with DIN conventions. The spatial adapter then detects that the German text is approximately 35% longer than the original and applies automatic word wrapping across two lines, as shown in Figure 4, eliminating a collision with an adjacent dimension line. The final annotation is inserted without further adjustment.

Scenario 3 — Critical-risk escalation to expert review.

The annotation references an obsolete enterprise-specific standard for which the knowledge base contains no mapping. Retrieval returns low-confidence results from unrelated standard families. The risk score exceeds τ_2, and the Trident Router immediately routes the annotation to Expert Review via the right branch of Figure 2, without invoking the LLM. Even if the router had not escalated the annotation, the symbolic validator shown in Figure 3 would have rejected any generated output because no retrieved evidence could support an evidence-consistency check for an unknown enterprise standard. The audit log records the escalation reason, the retrieval confidence scores, and the original annotation text, providing the human reviewer with all information needed to make an informed decision. This scenario illustrates a key design philosophy: the framework is designed to recognize the boundaries of its own competence and to escalate gracefully rather than generate a plausible-sounding but unverifiable output.

5. Discussion

5.1 Why RAG Alone Is Insufficient

A natural question is whether retrieval-augmented generation, without the additional symbolic machinery, would be sufficient for this task. We argue that it would not, for two reasons. First, RAG reduces but does not eliminate hallucination: the LLM can still generate outputs that are inconsistent with the retrieved evidence, particularly for numerical values where the model's parametric prior may override the retrieved context. Entity masking removes this risk by ensuring that the LLM never has the opportunity to alter critical values. Second, RAG provides no mechanism for verifying that the generated output satisfies the formal requirements of the target standard. Symbolic validation fills this gap by applying deterministic checks that are independent of the LLM's internal state. The combination of these two mechanisms—pre-generation masking and post-generation validation—achieves a level of reliability that neither could provide alone.

5.2 Defense-in-Depth Against Hallucination

The framework implements a defense-in-depth strategy with three independent layers of hallucination control. The first layer is entity masking, which prevents hallucination of critical values before generation. The second layer is symbolic validation, which detects and rejects hallucinated outputs after generation. The third layer is expert escalation, which removes from the automated pipeline any annotation for which the first two layers cannot provide sufficient confidence. Each layer is independently effective, and their combination provides a level of reliability that no single mechanism could achieve alone. This architecture reflects a broader principle: in safety-critical systems, redundant, independent safeguards are preferable to a single highly capable but potentially fallible mechanism.

5.3 Spatial Reasoning as a First-Class Concern

Prior work on AI-assisted CAD processing has largely treated spatial layout as a post-processing

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

concern, if it is treated at all. Our framework elevates spatial reasoning to a first-class concern by integrating the layout adapter directly into the conversion pipeline, with a formal optimization objective and explicit escalation for unresolvable conflicts. The three-stage adjustment strategy—wrapping, shifting, scaling—provides a principled ordering that minimizes disruption to the drawing's geometric structure while maximizing the probability of finding a feasible layout automatically. This integration is essential for producing drawings that are not only semantically correct but also immediately usable by manufacturing engineers without manual layout correction.

5.4 Limitations

The framework has several limitations that should be acknowledged. Its performance depends critically on the quality and completeness of the manually curated knowledge base and the symbolic rule set. Gaps in the knowledge base lead to unnecessary escalations; errors in the rule set lead to incorrect validations. Maintaining these resources requires ongoing expert effort, particularly as standards are revised. The spatial adapter handles local annotation conflicts but does not perform global sheet-level layout optimization; in drawings with very dense annotation fields, local adjustments may shift conflicts rather than resolve them. Extending the framework to standard pairs other than GB–DIN, or to engineering domains beyond mechanical manufacturing (e.g., electrical schematics or civil engineering drawings), requires rebuilding both the knowledge base and the rule set from scratch, which is a significant investment. Finally, the three-attempt regeneration limit means that some annotations will always require human review; the framework does not guarantee full automation.

5.5 Future Directions

Several directions for future work are apparent. Automated knowledge base population—using information extraction pipelines to ingest new and revised standards without manual curation—would substantially reduce maintenance costs. Learning validation rules from error logs, rather than hand-coding them, would make the rule set more robust and easier to extend. Global multi-annotation layout optimization, treating the entire drawing as a single optimization problem rather than resolving conflicts annotation by annotation, would produce better layouts in dense drawings. Finally, extending the framework to handle not only text annotations but also geometric symbols (surface finish marks, welding symbols, GD&T feature control frames) would substantially broaden its applicability and bring the system closer to full drawing-level automation.

6. Conclusion

This paper proposed a neuro-symbolic and satial-aware framework for hallucination suppression in safety-critical engineering drawing conversion. Instead of treating GB-to-DIN conversion as a general translation task, the framework formulates it as a constrained engineering transformation that must preserve critical entities, comply with target standards, and maintain CAD layout feasibility. The proposed pipeline combines a Trident Router, ontology-driven RAG, entity masking, symbolic validation, and spatial-aware layout adaptation to reduce the risk of hallucinated values, unsupported standard mappings, and annotation collisions.

The framework provides three layers of control: risk-aware routing before generation, entity masking and evidence-grounded generation during conversion, and deterministic symbolic verification after generation. In addition, the CAD layout adapter checks whether the translated annotations can be safely reinserted into the drawing without violating readability or leader-line constraints. These mechanisms make the LLM a controlled component within an auditable engineering workflow rather than an unconstrained text generator.

Although the present work focuses on architectural design and theoretical workflow analysis, it provides a structured basis for reliable LLM-assisted engineering drawing conversion. Future work will focus on prototype implementation and empirical validation using real mechanical drawing datasets. Quantitative evaluation will include entity preservation rate, standard-mapping accuracy, hallucination rate, layout collision rate, and human correction effort. The framework will also be extended to support global drawing layout optimization and the conversion of engineering symbols such as surface finish marks, welding symbols, and GD&T feature control frames.

ISSN 2616-5775 Vol. 9, Issue 5: 48-56, DOI: 10.25236/AJCIS.2026.090506

References

[1] J. Freimuth, S. Kaiser, M. Schädler (Eds.), Standardization Strategies in China and India: Industrial Policy and Geopolitics and Implications for Europe, Springer Nature, 2024.
[2] R. Jiao, "From 'data-driven' to 'data-informed' design — grounding AI for design in knowledge, context and decisions," Journal of Engineering Design, vol. 37, pp. 1–22, 2026.
[3] N. Melluso, I. Grangel-González, G. Fantoni, "Enhancing Industry 4.0 standards interoperability via knowledge graphs with natural language processing," Computers in Industry, vol. 140, p. 103676, 2022.
[4] P. Kügler, F. Dworschak, B. Schleich, S. Wartzack, "The evolution of knowledge-based engineering from a design research perspective: Literature review 2012–21," Advanced Engineering Informatics, vol. 55, 2021.
[5] Z. Sun, L. Jia, J. Hao, et al., "KT-MDO: a knowledge-template-driven multidisciplinary design optimization framework," Advances in Engineering Software, vol. 214, 2026.
[6] N. Madhusudanan, B. Gurumoorthy, A. Chakrabarti, "From natural language text to rules: knowledge acquisition from formal documents for aircraft assembly," Journal of Engineering Design, vol. 30, pp. 417–444, 2019.
[7] Z. Shi, S. Ergan, "An ontology-based approach for Façades inspection checklist generation," Advanced Engineering Informatics, vol. 72, 2026.
[8] B. Patzák, S. Šulc, V. Šmilauer, "Towards digital twins: Design of an entity data model in the MuPIF simulation platform," Advanced Engineering Software, vol. 197, 2024.
[9] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP," in Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 9459–9474.
[10] D. Shi, J. Li, O. Meyer, T. Bauernhansl, "Enhancing RAG for cognitive digital twins," Computers in Industry, vol. 171, 2025.
[11] H. Xiao, J. Zhuang, B. Yang, et al., "Generative knowledge-guided review system," Advanced Engineering Informatics, vol. 68, 2025.
[12] C. Yu, Y. Chen, L. Wang, T. Yan, "Reconstructed generative design for industrial products," Journal of Engineering Design, vol. 37, pp. 1234–1254, 2026.
[13] S. Jose, K.T.P. Nguyen, K. Medjaher, et al., "Multimodal industrial diagnostics," Expert Systems with Applications, vol. 255, 2024.
[14] NLP Research Team, "Faithful to the Document or to the World?," in Proceedings of the 62nd Annual Meeting of the ACL, 2024.
[15] R. Louf, B. Willard, "Efficient guided generation for LLM," in Proceedings of the 12th ICLR, 2024.
[16] M.T. Khan, et al., "Vision-Language Model for engineering drawing extraction," in ICIAI, Springer, 2026.
[17] J. Zhou, J.D. Camba, "Multimodal LLM in parametric CAD," Expert Systems with Applications, vol. 282, 2025.
[18] R. Andukuri, "CNN–GNN machining feature recognition," Expert Systems with Applications, vol. 303, 2026.
[19] F. Wang, et al., "AAGATNet: Attentive graph network," Computer-Aided Design, vol. 193, 2026.
[20] S. Yoo, N. Kang, "Explainable AI for manufacturing cost estimation," Expert Systems with Applications, vol. 183, 2021.
[21] F. Boussuge, et al., "Tensor factorisation for CAE model preparation," Computer-Aided Design, vol. 152, 2022.
[22] J.-S. Lee, et al., "Automation of trimming die design inspection," Engineering Applications of Artificial Intelligence, vol. 127, 2024.
[23] J. Song, et al., "P2Seg: Distance query from point to segments," Computer-Aided Design, vol. 189, 2025.