

Better Be Sure: A Verification and Validation Taxonomy for Industrial KG Construction

Wilma J. Schmidt^{1,2,*}, Evgeny Kharlamov^{1,3} and Adrian Paschke^{2,4}

¹Robert Bosch GmbH, Corporate Research, Renningen, Germany

²Freie Universität Berlin, AG Semantic Web, Berlin, Germany

³SIRIUS, Centre for Scalable Data Access, University of Oslo, Norway

⁴Data Analytic Center (DANA), Fraunhofer Institute FOKUS, Berlin, Germany

Abstract

LLMs are increasingly applied to Knowledge Graph (KG) construction in manufacturing as an alternative to manual, symbolic or neuro-symbolic approaches. However, neural methods including LLMs are naturally opaque and, hence, pose challenges to safety-critical domains such as manufacturing. While verification and validation (V&V) frameworks are common in software engineering, there is no structured V&V approach for KG construction in safety-critical industrial domains. To close this gap, we propose a *five-dimensional* V&V taxonomy for KG construction in the manufacturing domain. We evaluate our taxonomy on 50 reports from this field and support future work in this field by identifying gaps in reported V&V evaluation steps. This taxonomy offers researchers and practitioners a taxonomy structuring this research field to compare V&V evaluations of neuro-symbolic approaches for manufacturing KG construction.

Keywords

Knowledge Graph, KG Construction, Verification and Validation, Taxonomy, Neuro-symbolic AI

1. Introduction

Safety-critical domains, such as manufacturing, are increasingly confronted with neural methods, among others large language models (LLMs), which, by their nature, come with opaqueness and cannot fully ensure factual accuracy. To support adoption of these methods, verification and validation (V&V) of these approaches, such as evaluation frameworks for LLM-based RAG systems [1], are needed. Some of the current trends in the manufacturing domain are small language models (SLMs), agentic AI and physical AI. This is an indication that opaque elements, with the challenges they pose, will remain in the manufacturing domain. Knowledge Graphs (KGs) as a semantic technology are recognized elements in the manufacturing research field [2, 3] as well as in the manufacturing system landscape [4] as they bridge distributed data sources and offer semantic context to retrieved entities. This is especially valuable in complex domains where an answer of $p = 0.7$ means nothing without context: label of p , unit of the value 0.7, defined thresholds for values of p , etc. However, KG construction still is a bottleneck in the manufacturing domain.

To scale KG construction, (semi-)automation of the construction steps is needed. Research on neuro-symbolic, with specifically, neural methods such as *Bilateral Encoder Representations from Transformers* (BERT) or LLM-supported KG construction in manufacturing is, unsurprisingly, increasing [5]. The development in manufacturing is similar to trends of KG construction in other domains, e.g., scientific literature. Lu *et al.* [6] see an evolution from neural approaches with learned representations from 2010 to 2022 towards LLM-supported KG construction from 2022 onward. The manufacturing domain expects the data quality of the KG to be high as safety-critical downstream applications expect reliability and factuality of the KG. To evaluate data quality and reliability of a KG is non-trivial, see e.g. [7].

SGKi 2026: Scaling Knowledge Graphs for Industry Workshop, co-located with SEMANTiCS'26: 22nd International Conference on Semantic Systems, September 15–17, 2026, Ghent, Belgium

*Corresponding author.

✉ wilma.schmidt@bosch.com (W. J. Schmidt); evgeny.kharlamov@de.bosch.com (E. Kharlamov); adrian.paschke@fokus.fraunhofer.de (A. Paschke)

ORCID 0009-0001-0872-4578 (W. J. Schmidt); 0000-0003-3247-4166 (E. Kharlamov); 0000-0003-3156-9040 (A. Paschke)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

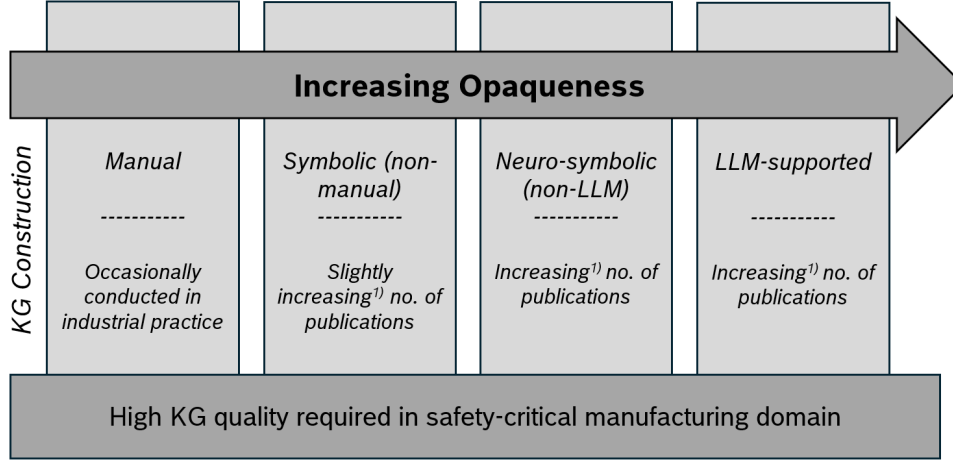


Figure 1: Illustration of Increasing Opacity across Different KG Construction Approaches which motivates the need for structured V&V in this field. 1) according to [5].

Further, the evaluation of “automatically constructed knowledge graphs is challenging” [8]. As opaque methods in KG construction methods in manufacturing increase, see Figure 1, there is an urgent need for structured V&V of these approaches to ensure trust in their results: industrial KGs which are reliable semantic sources for downstream applications such as information retrieval and search engines [9], recommendation systems [9], supply chain management [9], or digital twins [10].

Multiple neuro-symbolic frameworks for KG construction in manufacturing are proposed already with different quality evaluation steps. While there are overlaps in support for a high quality KG output, there is no adequate structural approach for this evaluation. It is not enough to evaluate correct triples with precision, recall and $f1$ -score, as we need a deeper understanding and adjusting of intermediate steps than plain output evaluation of KG construction to actually improve it. Neuro-symbolic KG construction might further contain intermediate steps with ontological constraints, ergo the verification is not the same as of a blackbox LLM model. Approaches vary across the AI research landscape including standard metrics such as *precision*, *recall* and $f1$ -score, as well as hybrid strategies with a mix of rule-based, semantic and statistical verification steps (see, e.g., [11]), and *LLM-as-a-Judge* approaches with *Evaluator Agents* (see, e.g., [6]) and more. In summary, there is no sufficient clarity on how to verify and validate a neuro-symbolic KG construction process. Following [12], there is a need for V&V in safety-critical domains. Driven by the need of high quality to serve safety-critical applications, a structured V&V is needed for this task.

The contributions of this work are the following:

- **Taxonomy for V&V of KG Construction.** We propose a taxonomy to structure V&V of neuro-symbolic AI approaches for KG construction. This structure is especially needed in safety-critical domains such as manufacturing.
- **Application of Taxonomy.** Randomly selecting 50 reports from a recently conducted systematic literature review, we assess V&V steps in neuro-symbolic AI approaches for KG construction in the manufacturing domain with our taxonomy.
- **Identification of Research Gaps.** We identify gaps in reporting V&V evaluation steps of neuro-symbolic AI approaches for KG construction in manufacturing. We excavate these gaps from the evaluation of 50 reports on neuro-symbolic AI for manufacturing KG construction.

The remainder of the paper is structured as follows: We discuss related works in Section 2 and identify a research gap that motivates our work. As this research is at an intersection of multiple research fields (KG construction, V&V, neuro-symbolic approaches, manufacturing AI), we briefly introduce the background and definitions relevant for our research in Section 3. We propose a taxonomy for V&V for neuro-symbolic AI approaches in industrial KG construction in Section 4. We put this taxonomy to practice in a case study format and apply it onto a search corpus from a systematic literature review

on neuro-symbolic AI for KG construction in the manufacturing domain. We present the results in Section 5. Reflecting on our work, we identify limitations of this study and point out research gaps in manufacturing KG construction in Section 6. In a final chapter, we conclude this work and briefly discuss future work.

2. Related Work

In this section, we review related works on taxonomies in ontology engineering as well as knowledge engineering and selected works with taxonomies on neuro-symbolic AI, including LLM and Agentic AI. We further include selected relevant surveys to show the importance and dynamic development of this research field.

2.1. Taxonomies for Verification and Validation in Ontology Engineering

Much research has been conducted already years ago on ontology V&V. Ontology quality initiatives such as *Linked Open Data*¹ were proposed (and still exist) as well as many terms and definitions for domain ontology evaluation, see, e.g., an overview in [13]. While *McDaniel and Storey* observe in their work that “[m]ost ontology evaluation research focuses on a single method to determine quality” [13], they propose a classification into *five* ontology evaluation research classes: *Domain/Task Fit*, *Error Checking*, *Libraries*, *Metrics*, and *Modularization*. [14]) proposes factors on *inconsistency*, *incompleteness*, and *redundancy* for ontology verification. *Gangemi et al.* [15] motivated their earlier work with *unsystematic* basic approaches of ontology evaluation in research. The authors subsequently proposed three measure types for verification, as there are the *structural* dimension, *functional dimension*, and *usability-profiling* dimension of ontologies [15]. Ontology validation is formulated as a diagnosis task by the authors. Additionally, [16] assesses and evaluates ontology evaluation works. However, we will not go into this level of detail with our taxonomy and case study, but remain on a level that identifies whether an ontological V&V step has been conducted at all or not. Our goal is to provide a holistic V&V taxonomy for neuro-symbolic AI on KG construction, and not dive too deep into details of the respective steps in this work.

2.2. Taxonomies for Verification and Validation in Knowledge Engineering

Assessing knowledge engineering taxonomies examines also the data layer. *Huaman et al.* [7] assess KG validation. The authors do not propose a taxonomy, but briefly name two categorizations for validation studies of KGs: the validation target and the source against which is validated (KG or external). For neuro-symbolic approaches for KG construction, we see many more aspects, such as the temporal aspect of when the validation takes place (e.g., during construction). These aspects are critical to consider in order to compare and assess V&V frameworks of KG construction because neuro-symbolic approaches might evaluate and correct neural methods during runtime with symbolic verifiers.

On knowledge engineering and assessing bias in systems of LLM-supported KGs, *Zabihi et al.* [17] offer a taxonomy which considers bias as a pipeline emergence. While the authors identify also structural bias, originating i.a. in KG construction, the focus of their work differs from ours and cannot close the research gap for V&V of KG construction with neuro-symbolic AI.

With focus on the task of knowledge-augmented question answering (KAGA), *Samir et al.* [18] recently introduce a taxonomy with two foundational axes: knowledge sources and knowledge selection/grounding. Building on these axes, they further refine categories that characterizes KAQA

¹<https://lov.linkeddata.es/>

Table 1

Comparison of selected similar works.

Work	Goal of KG Construction	Neuro-/Symbolic Evaluation Objects	Proposes a Taxonomy of ...
Min et al. [1]	-	LLM-RAG	none (instead: V&V framework)
Sapkota et al. [20]	-	Generative AI, AI agents, agentic AI	Generative AI, AI agents, and agentic AI
Samir et al. [18]	-	KG, RAG, LLM	Knowledge-augmented question answering
Renkhoff et al. [12]	-	Neuro-symbolic AI	none (instead: survey on V&V)
Ibrahim et al. [9]	(√)	KGs augmented with LLMs	none
Tiwari et al. [19]	-	LLM	LLM evaluation across capability, methodology, metric layers
Zabihi et al. [17]	√	LLM	LLM bias
Our work	√	Neuro-symbolic AI	V & V for Neuro-symbolic AI

methods. Our work focuses on KG construction, hence, knowledge selection and grounding, as needed when retrieving the correct answer to a question, are not the most relevant dimensions during KG construction to assess.

2.3. Taxonomies for Verification and Validation of Neuro-Symbolic Approaches

Next, we review selected taxonomies and strongly relevant evaluation works with assess neuro-symbolic or LLM-based approaches. Along with a survey, Tiwari *et al.* [19] propose a taxonomy on LLM evaluation including a capability layer which distinguishes between knowledge reasoning, code generation and more. Very insightful in their work is further the observed *evaluation trilemma* which illustrates the challenges of benchmarks for LLM evaluation as authenticity, contamination resistance, and scalability are difficult to achieve all at once. In contrast to Tiwari *et al.* [19], we focus on one specific task (KG construction) and instead of one element type (LLM), our taxonomy addresses a multi-element neuro-symbolic structure. This is why, e.g., the dimension “When” is needed for the evaluation. In a recent work, Sapkota *et al.* [20] propose a taxonomy of AI agent paradigms which structures generative AI, AI agents, agentic AI, and inferred generative agents into multiple dimension such as initiation type and temporal continuity. Renkhoff *et al.* [12] survey neuro-symbolic verification and evaluation and discusses two taxonomies of neuro-symbolic AI, one by Kautz [21] and one by Yu *et al.* [22]. The authors distinguish in their assessment the V&V of neuro-symbolic AI between external frameworks and the respective, own neuro-symbolic architecture. This comprehensive survey, posing multiple open research challenges, underlines the importance of this research field. Ibrahim *et al.* [9] introduce in a comprehensive survey on augmenting KGs with LLMs commonly used evaluation metrics such as accuracy, ROUGE [23], BLEU [24], *f1*-score, and human evaluation. It is a great challenge to structure such a field and not focus of their work. Our work aims to structure for the specific task of KG construction the evaluation steps of V&V by providing a systematic taxonomy.

2.4. Summary of Related Work

All in all, we observe that the increase of neural methods, especially LLMs, raise the pressure on structured V&V of KG construction with neural or neuro-symbolic methods. Research that introduces new frameworks evaluates them thoroughly [6, 1], yet, not standardized, and hence, challenging to compare. To the best of our knowledge, no work has proposed a taxonomy which systematically organizes V&V of KG construction. Our research fills this gap, see also Table 1.

3. Background and Foundations

In this section, we introduce the relevant foundations for our work.

3.1. Verification and Validation

Verification and validation is a term which is often used in software engineering. Sargent [25] defines *computerized model verification* “as ensuring that the computer programming and implementation of the conceptual model is correct”. Sargent proposes further validity definitions from three different perspectives: conceptual model validity, operational validity and data validity. In summary, *validity* describes that the evaluation object is correct (e.g., for data) and its output behaviour shows “sufficient accuracy for its intended purpose” (e.g., for a model).

To be more precise for a task of knowledge engineering, we define what *validation* means in our context of KG construction. Following [7], “knowledge validation [...] measures whether statements from KGs are semantically correct and correspond to the so-called “real” world.”

There are different types of KGs and to address as many of them as possible with our taxonomy, we follow an open definition of KGs in this paper: A KG follows this set of characteristics according to [26], but we omit the fourth factor: A KG (1) mainly describes real world entities and their interrelations, organized in a graph, (2) defines possible classes and relations of entities in a schema, (3) allows for potentially interrelating arbitrary entities with each other, and (4) covers various topical domains. We expect KG focus on the manufacturing domain.

3.2. KG Construction

The KG construction process is understood as in Zhong *et al.* [27] where KG construction is a “procedure” which maps data with “background knowledge of the target”, e.g., “domain knowledge” into a KG. Structured data is mapped into a KG, i.e., rules are defined how the structured data is assigned onto KG elements and then these rules can be executed to create the KG. For unstructured data, there are two main different approaches for KG construction: sequential construction and triple extraction. Sequential construction is a pipeline approach in which entities and relations are sequentially extracted. On the other hand, triple extraction extract the KG triples in one step. Both approaches are almost equally represented in the manufacturing domain when construction contains neuro-symbolic approaches [5].

4. Verification & Validation Taxonomy for KG Construction

We propose a taxonomy for V&V of neuro-symbolic AI approaches for KG construction in manufacturing, see Figure 2. The proposed taxonomy spans across *five* dimensions: *Evaluation Artefact*, *Reference Basis*, *Evaluation Perspective*, *Evaluation Executive*, and *Temporal Execution*. This taxonomy enables a structured perspective onto a dynamic environment as models and approaches evolve. This contribution may further serve as an anchor for comparable V&V of neuro-symbolic KG construction in a complex domain with expert knowledge and high quality requirements.

We briefly state relevant decisions and constraints on our taxonomy. Please note that this taxonomy is blending aspects of KG construction V&V with KG V&V as one can best, but not exclusively, judge a process (construction) by its outcome (KG). Each dimension can unfold into several aspects. An evaluation framework might implement multiple of those verification aspects within one dimension. We introduce all dimensions and their aspects in detail in the following. Further, there is a difference between proposing an additional method for improvement of the construction versus verifying and validating a method. Proposing an additional method can be an additional step such as a common

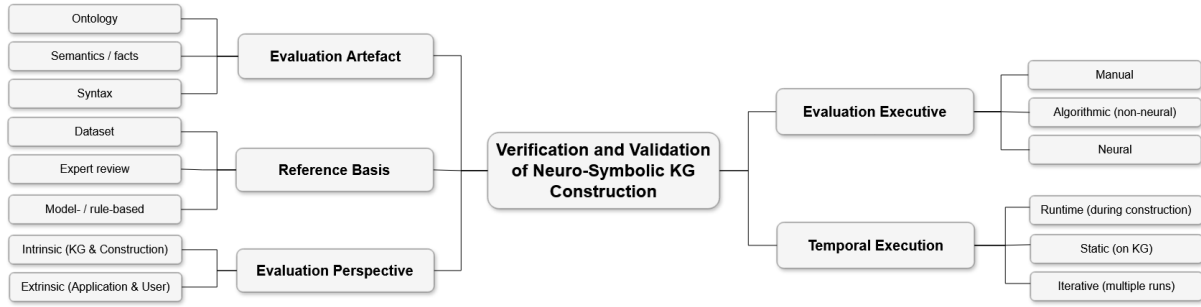


Figure 2: Taxonomy of Verification and Validation of Neuro-Symbolic KG Construction.

Conditional Random Field (CRF) after a *Bidirectional Long Short-Term Memory (BiLSTM)*. Verifying or validating a proposed method is often rare or costly such as labelled data or expert review. Please note that it is not always possible to identify exactly how a research paper verified and validated its steps, especially if explanations on this are omitted. However, we will give some examples of how we expect and interpret scenarios in a research paper when mapping them onto our taxonomy.

4.1. Dimension: Evaluation Artefact

The first dimension, the *Evaluation Artefact*, is probably the most intuitive to consider: *what* aspect to verify and validate. In KG construction, the evaluation aspects can be of *syntactical*, *ontological*, or *semantical* nature. The semantic fact checks are verifying and validating the data layer of the KG. This could be by *f1*-score of extracted entities, relations, or triples. Or, e.g., in an application-oriented evaluation in which users give feedback on the answers they receive to queries. Ontological checks comprise both the verification and validity of the ontology itself, e.g., with a validating step in the seven-step method or a pre-assessment before ontology refinement, as well as top-down KG construction approaches or similar which ensure that extracted data conforms to the ontological layer of the KG. For ontological checks it can be interesting to further differentiate between whether the final KG or another, external source is employed [7]. Syntactical aspects evaluate the syntactical correctness of the chosen knowledge representation format and serialization, e.g., the *Resource Description Framework (RDF)* with *turtle (ttl)* serialization for KGs.

4.2. Dimension: Reference Basis

The *Reference Basis* dimension distinguishes between a dataset, e.g., a gold expert-labelled ground truth or a silver KG generated with a powerful LLM, an expert review in which experts manually review the KG elements, and a model- or rule-based reference, e.g., a random wandering method or a set of rules. Case studies in a research paper often use a static dataset for validation and are included in the aspect *dataset* if they contain an evaluation step on this data.

4.3. Dimension: Evaluation Perspective

The *Evaluation Perspective* dimension structures the nature from which the V&V is motivated: intrinsically or extrinsically. The aspect *Intrinsic (KG & Construction)* comprises ontological quality evaluations, e.g., soundness, as well as evaluations on the KG, its elements, and of the KG construction steps. Next to standard metrics of, e.g., accuracy measurement of KG elements, also KG construction evaluation such as ablation studies are included, justifying that each element during construction is proposing a value add. The other aspect of this dimension is to extrinsically evaluate in an application- or user-oriented fashion, i.e., define tasks which the KG should serve and evaluate the performance on these tasks.

4.4. Dimension: Evaluation Executive

The *Evaluation Executive* is identifying who is conducting the V&V. We differentiate between manual evaluation, algorithmic (non-neural) evaluation and neural methods. Manual approaches are very often still used in manufacturing, and likely also other safety-critical domains, to address the challenge that someone has to be responsible for any potential failure that might happen. Algorithmic approaches cover deterministic, e.g., rule-based, probabilistic and statistical approaches, in which the evaluation steps are traceable. These approaches can for example be implementations of standard metrics *precision*, *recall*, and *f1-score*, and are attractive due to their scaling potential. Neural methods, such as LLM-as-a-judge, are included as an aspect to pave the way for assessing whether LLMs can in the future bring the best of both worlds: human judgement with scalability of fully-automated systems. However, caution is needed when assessing opaque methods with opaque methods. First works are already implementing neural methods for evaluation, see, e.g., [28, 6].

4.5. Dimension: Temporal Execution

The *Temporal Execution* is opened up by the fact that KG construction is a process. We distinguish between V&V steps implemented in runtime during construction, statically on the final KG (post-hoc) or iteratively in a multi-run fashion. Design choices to improve accuracy are included in *Runtime*, if they contain any active part such as an evaluation or if they are rule-based to ensure via constraints that only correct data is allowed. Further, common techniques such as *Knowledge Fusion* with *Entity Resolution* are considered as choices during runtime to conduct verification and improvement of data quality. To choose *BERT* because of its known good performance in *Named Entity Recognition (NER)* is not a verification step, but evaluating several different parameter settings or models on a test data set and then choosing the best performing one, e.g., *BERT*, is. The repeated execution of KG construction to observe reproducibility aspects is especially interesting with the increasing adoption of neural methods to assess their robustness for this task.

4.6. Verification and Validation across the Taxonomy

Each of the *five* taxonomy dimensions covers both verification and validation aspects. Most aspects apply to both V&V evaluation sides, yet *four* of them do not or, at least, strongly focus on one side. In dimension *Evaluation Artefact*, the *syntax* aspect clearly addresses verification steps. In dimension *Evaluation Perspective*, the *Intrinsic* aspect targets verification steps, the *Extrinsic* aspect targets validation steps. Further, in dimension *Temporal Execution*, the *Iterative* aspect focusses mainly on verification of the construction. Defining a threshold within which variances of results of re-runs are acceptable, is an aspect of validation. On an additional note, the aspects *ontology* and *semantic / facts* in dimension *Evaluation Artefact* may appear at first glance as assignable: *ontology* as a conceptual model verification and *semantic* checks as a validation. However, an ontology can be assessed from the intended purpose of modeling the, e.g., specified domain, which is a validation; and the semantic checks encompass also triple evaluations in a multi-run assessment which is purely on verification of the construction approach. Hence, both *ontology* and *semantic / facts* can address verification as well as validation steps.

5. Case Study: Manufacturing

We put this taxonomy to practice by assessing works from a recent systematic literature review on neuro-symbolic AI for KG construction in the manufacturing domain [5]. Next, we explain the configurations of the case study and, subsequently, present observations and results.

5.1. Configurations

From a review search corpus of neuro-symbolic AI for manufacturing KG construction [5] with 105 reports, we randomly select 50 for our case study by using a random number generator tool². Each report is partially reviewed, i.e., only sections such as *Experiments*, *Results*, *Evaluation* and *Case Study* are assessed as well as partially the description of the proposed *Framework*, *Methodology* or *Solution*, if the paper contains such a section. We do *not* assess the *Abstract*, *Introduction*, *Related Work* or *Conclusion*. In our assessment, we look for statements from the authors on V&V steps before, during or after KG construction (*Reference Basis*) and on details on these steps. We do not assess primarily via a keyword search as there are too many synonyms in practice on V&V and sometimes these steps are not declared explicitly as V&V at all. However, we perform a supporting verification step by additionally searching for “seven-step” and “top-down” to identify V&V steps on ontological level.

If authors conduct several V&V steps, it is possible that multiple aspects of a dimension are fulfilled. For example, a study can evaluate three methods during *Runtime* to identify the best performing one and, after construction, validate the KG in a *Static* fashion on a case study. Please note that a V&V step does not have to apply to the whole construction process and final KG to be marked as fulfilled. For example, a KG construction process may adopt both a top-down and a bottom-up approach. During bottom-up construction, it is possible that no ontological evaluations are conducted. Additionally, despite impressive steps on improving *NER*, *Relation Extraction (RE)*, or *joint extraction of triples*, the V&V taxonomy organizes only V&V steps, not model improvements.

5.2. Observations and Results

Next, we discuss the taxonomy evaluation observations on studies with neuro-symbolic approaches for manufacturing KG construction, see Figures 3 and 4.

Almost all constructed KGs are evaluated on a semantic fact level (92%). One of the few exceptions constructs an ontology without explicit population steps, hence, only reports the ontological schema evaluation. More than half of the studies show ontological evaluations, e.g., the seven-step method for ontology development, top-down KG construction approaches, knowledge reasoning to identify errors, or mapping evaluation against a gold-mapping while checking also the ontological mapping elements. We discuss the low number of studies with syntax checks (8%) in Section 6.2. The vast majority of studies (84%) reports V&V on a dataset. We observe several case studies, which is nearby due to the domain-proximity. Model- or rule-based reference bases include an annotation verifier, rule-based knowledge reasoning, a graph random wandering method, or syntactic and semantic constraints to filter candidate-entity pairs. The reported V&V steps target more often intrinsically the ontology, the KG or its construction (72%) than extrinsically a (potential) downstream application or user (54%). Most studies report a *static* V&V step after KG construction (86%) and half of the studies report a V&V step during *runtime* of KG construction (52%). Only 8% report reproducibility evaluations.

Algorithmic V&V steps are very commonly reported (90%) as the standard metrics *precision*, *recall* and *f1-score* are easy to implement and execute if there is a reference dataset. A neural method, specifically *LLM-as-a-judge*, is reported in one study for assistance in data augmentation as a V&V step [29] with the results stressing the potential of LLMs for industrial maintenance in this task. Manual steps with a human-in-the-loop, often as expert review, are also reported (36%).

6. Limitations and Research Gaps

We carefully build a frame around our work in the following section. Assessing the limitations of our own work is necessary to properly integrate our contributions, especially the proposed taxonomy, into the research landscape. Further, observing gaps in reporting V&V evaluation steps can spark new or support existing research topics within neuro-symbolic KG construction in safety-critical or industrial domains. We undertake both tasks in the following.

²<https://www.randomizer.org/>

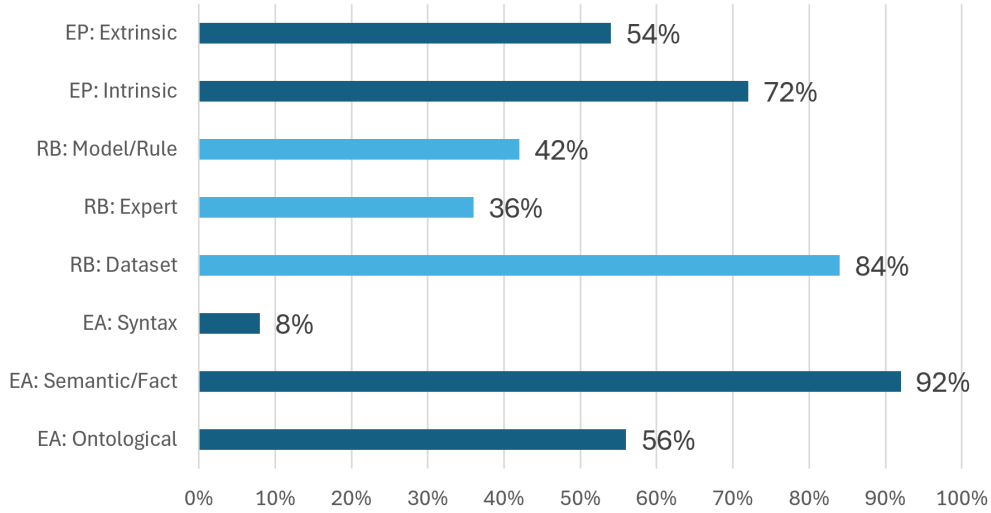


Figure 3: Case Study Results on Evaluation Artefact (EA), Reference Basis (RB), and Evaluation Perspective (EP). Percentage (%) of studies reporting the respective V&V aspect.

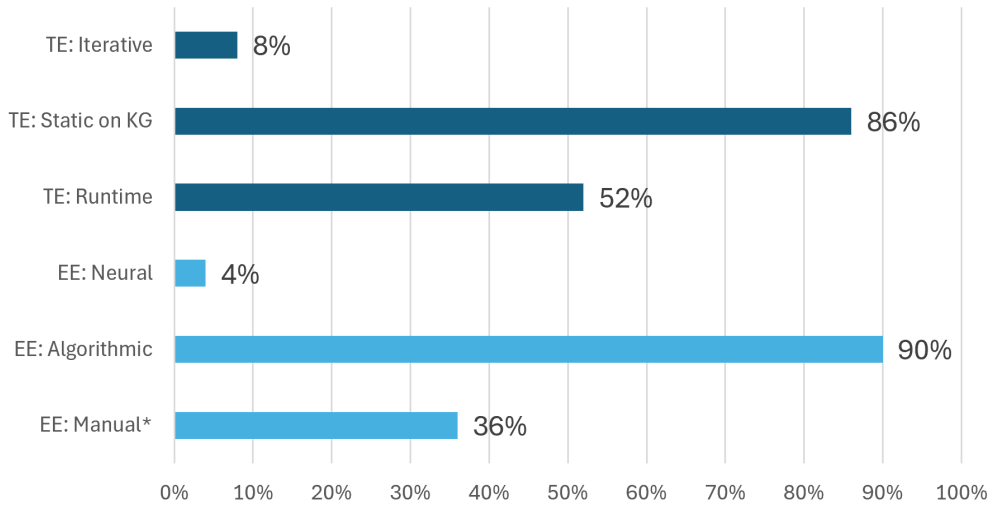


Figure 4: Case Study Results on Evaluation Executive (EE) and Temporal Execution (TE). Percentage (%) of studies reporting the respective V&V aspect. Manual*: This aspect is partially dependent on aspect *RB: Expert* (see Figure 3).

6.1. Limitations

In this part, we discuss the limitations of this work.

First of all, focussing on V&V of KG construction implies that some other aspects of KG construction are not in focus due to often inferior validity impact. In our case, these aspects include among others usability (user-friendliness), complexity of the framework, response times, and costs caused, e.g., by token count. Several of these items are hidden in the dimension *Evaluation Perspective* beneath aspect *Intrinsic*. Looking at KG construction from another angle, we expect some of these topics to be more prominent.

Second, we note that the dimensions are not fully orthogonal. Dimension *Evaluation Executive* with aspect *Manual* shows a partial dependence on dimension *Reference Basis* with aspect *Expert Review* as expert reviews constitute a manual step of V&V. We show this in Figure 4. Further, on a more detailed level, intrinsic evaluations using, e.g., accuracy metrics are in practice calculated on a dataset. The

ground truth is needed to calculate, e.g., for *precision*, the proportion of *positive predictions* from all *true values*, i.e., *positive* plus *false predictions*. Hence, there is a correlation between the aspects *Intrinsic* and *Dataset*. To not lose focus, we do not discuss correlations on lower than aspect level any further.

Third, as we conduct our case study in a manual manner with one assessor, the completeness of the report evaluations might show limitations. Overlooking a V&V step is possible due to human error. Additionally, as mentioned, there is a fine line between improving KG construction and beginning of V&V. We consider *Knowledge Refinement* as defined in [27] in general as part of V&V, as, e.g., an accuracy evaluation first is necessary before improvement steps can be taken.

Fourth, our taxonomy remains on an abstraction level which we believe is suitable for giving a holistic V&V view which covers the complexity of a KG construction process with neuro-symbolic methods. A comprehensive amount of literature is available for deep dives, e.g., on *ontology evaluation and validation* as discussed in Section 2.

Lastly, the case study to our taxonomy in the manufacturing domain primarily focusses on textual data which is due to the nature of the systematically gathered corpus. We have to be aware to not draw conclusions on audio, image or video too lightly from this as we may overlook important aspects.

6.2. V&V Reporting Gaps

Applying the V&V taxonomy proposed in this work to an excerpt of the review corpus on neuro-symbolic AI in manufacturing KG construction, supplies us with several insights, specifically gaps in reporting V&V, which we debate on in the following part.

Syntax Checks. A low number of reports shows explicit syntax checks. Pre-trained and, optionally, fine-tuned LLMs can generate triples, but may vary, caused by lacking robustness, in syntactical correctness. While assumably some of the syntax checks were integrated, yet not explicitly mentioned, we nevertheless recommend to explicitly mention them in future works to reveal insights on the quality of the underlying methods prior to syntactical correction.

Dataset Dependence. As scalability of KG adoption is an issue in industrial manufacturing, the high dependence of (well curated) datasets for V&V may point to a research gap. Identifying solid V&V mechanism which work with model- or rule-based reference bases can show potential.

Neural Methods. LLM-as-a-Judge or LLMs assisting in V&V steps is in our excerpt reported in *one* paper. We expect more exploration of LLMs in V&V in KG construction in the near future. However, caution is needed when assessing opaque methods with opaque methods.

Evaluation Perspective. Often, we observe reports with metrics targetting either (i) ontology, KG or its construction steps, or (2) an application in a task-based manner. We strongly recommend to include and report on both V&V aspects in studies whenever possible. The first target is an ideal basis to identify levers for improvement during ontology or KG construction. The latter target completes validation by demonstrating the performance and accuracy in an integrated or a downstream application of the KG.

Reproducibility. The low number of studies reporting explicit reproducibility checks is surprising. While tuning hyperparameters of a model is a valuable improvement, the re-iteration of result generation, here, KG construction, becomes increasingly important with neural methods to ensure robustness of the approach. With the increasing adoption of neural methods for KG construction, we expect more and more need for assessing the reproducibility of these opaque approaches for this specific task.

Structured V&V. In conclusion, when assessing the reports, we occasionally encountered readability challenges as V&V steps are not always prominently marked as such or organized and structured in an own subsection. Several reports feature a *Results*, *Experiments* or *Evaluation* section which is

already a good starting point. In some works, however, V&V steps are hidden within the *Methodology* or *Framework* section or among elaborate use case descriptions (which are relevant for validation, of course). We see a need for structured V&V, that stretches further than standard accuracy metrics, to ensure comprehensive assessment and comparability of approaches. This is necessary to achieve a common goal: to verify and validate approaches and make this comparable across the research landscape. We see our work as only one initial step in this endeavor.

7. Conclusion

In this work, we propose a V&V taxonomy for KG construction to act on the high KG quality requirements in industrial, safety-critical domains which are confronted with an increasing number of neuro-symbolic, and especially LLM-supported construction approaches. Our taxonomy is organized in *five* dimensions which structure and categorize the *Evaluation Artefact*, the *Reference Basis*, the *Evaluation Perspective*, the *Evaluation Executive*, and the *Temporal Execution*.

We put our taxonomy to practice on an excerpt of 50 reports from a systematic literature review on neuro-symbolic KG construction in the manufacturing domain. Semantic fact checks stand out as reported evaluation targets, while datasets are the most prominent reference basis for V&V. Static V&V on the final KG is dominantly reported as a temporal execution. As standard metrics such as *precision*, *recall*, and *f1-score* are often applied to verify the results, in a majority (88%) of reports there is at least one V&V step with algorithmic automation.

We further observe a V&V reporting gap on reproducibility and syntax which grow even more important with the increasing number of LLM-supported constructions. An additional gap is opened up by LLMs evaluating V&V steps. This area is expected to grow in the near future, however caution is needed regarding the traceability and responsibility of this *Evaluation Executive*. In future work, we aim to extend our taxonomy to multimodal data including audio, image and video data, as well as validating its generalizability to other industrial, safety-critical domains such as healthcare and autonomous driving.

Acknowledgments

This study was partially supported by the EU Project SMARTY (GA 101140087).

Declaration on Generative AI

During the preparation of this work, the authors used Gemini Enterprise in order to perform grammar, spelling checks, or rewording. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] Z. Min, C. J. Budnik, Verification and validation of llm-rag for industrial automation, in: 2025 IEEE International Conference on Artificial Intelligence Testing (AITest), IEEE, 2025, pp. 50–53.
- [2] Y. Wan, Y. Liu, Z. Chen, C. Chen, X. Li, F. Hu, M. Packianather, Making knowledge graphs work for smart manufacturing: Research topics, applications and prospects, *Journal of manufacturing systems* 76 (2024) 103–132.
- [3] G. Buchgeher, D. Gabauer, J. Martinez-Gil, L. Ehrlinger, Knowledge graphs in manufacturing and production: a systematic literature review, *Ieee Access* 9 (2021) 55537–55554.
- [4] T. Hubauer, S. Lamparter, P. Haase, D. M. Herzig, Use cases of the industrial knowledge graph at siemens., in: ISWC (P&D/Industry/BlueSky), 2018, pp. 107–108.

- [5] W. J. Schmidt, D. Rincon-Yanez, E. Kharlamov, A. Paschke, Llms on the rise: Neuro-symbolic ai for knowledge graph construction in manufacturing (systematic literature review), *IEEE access* (2026).
- [6] Y. Lu, W. Wu, X. Zhao, R. Peng, J. Wang, Karma: Leveraging multi-agent llms for automated knowledge graph enrichment, *Advances in Neural Information Processing Systems* 38 (2026) 56421–56453.
- [7] E. Huaman, E. Kärle, D. Fensel, Knowledge graph validation, *arXiv preprint arXiv:2005.01389* (2020).
- [8] J. Boylan, S. Mangla, D. Thorn, D. G. Ghalandari, P. Ghaffari, C. Hokamp, Kgvalidator: A framework for automatic validation of knowledge graph construction, *arXiv preprint arXiv:2404.15923* (2024).
- [9] N. Ibrahim, S. Aboulela, A. Ibrahim, R. Kashef, A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges, *Discover Artificial Intelligence* 4 (2024) 76.
- [10] C. Su, X. Tang, Q. Jiang, Y. Han, T. Wang, D. Jiang, Digital twin system for manufacturing processes based on a multi-layer knowledge graph model, *Scientific Reports* 15 (2025) 12835.
- [11] E. Zhu, Y. Hu, A multi-source fusion and dynamic verification framework for scientific knowledge graph construction: Methods, evaluation, and applications, *AI Magazine* 47 (2026) e70062.
- [12] J. Renkhoff, K. Feng, M. Meier-Doernberg, A. Velasquez, H. H. Song, A survey on verification and validation, testing and evaluations of neurosymbolic artificial intelligence, *IEEE Transactions on Artificial Intelligence* 5 (2024) 3765–3779.
- [13] M. McDaniel, V. C. Storey, Evaluating domain ontologies: clarification, classification, and challenges, *ACM Computing Surveys (CSUR)* 52 (2019) 1–44.
- [14] S. Lovrencic, M. Cubrilo, Ontology evaluation-comprising verification and validation, in: *Central European Conference on Information and Intelligent Systems, Faculty of Organization and Informatics Varazdin*, 2008, p. 1.
- [15] A. Gangemi, C. Catenacci, M. Ciaramita, J. Lehmann, Modelling ontology evaluation and validation, in: *European semantic web conference*, Springer, 2006, pp. 140–154.
- [16] A. Gómez-Pérez, *Ontology Evaluation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 251–273. URL: https://doi.org/10.1007/978-3-540-24750-0_13. doi:10.1007/978-3-540-24750-0_13.
- [17] P. Zabihi, D. Nawara, A. Ibrahim, R. Kashef, Analyzing bias in llm-augmented knowledge graph systems: Taxonomy, interaction mechanisms, and evaluation, *Applied Sciences* 16 (2026) 3410.
- [18] M. Samir, N. Fathy, W. Gad, Knowledge-augmented question answering: A taxonomy, critical review, and practical guidelines, *International Journal of Intelligent Computing and Information Sciences* 26 (2026) 39–50.
- [19] V. Tiwari, R. Othman, O. Suaifan, D. Wang, Rethinking llm evaluation: A comprehensive survey, taxonomy, and the evaluation trilemma, in: *2026 IEEE World AI IoT Congress (AIIoT)*, IEEE, 2026, pp. 0320–0329.
- [20] R. Sapkota, K. I. Roumeliotis, M. Karkee, Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges, *Information Fusion* (2025) 103599.
- [21] H. Kautz, The third ai summer: Aaai robert s. engelmore memorial lecture, *Ai magazine* 43 (2022) 105–125.
- [22] D. Yu, B. Yang, D. Liu, H. Wang, S. Pan, A survey on neural-symbolic learning systems, *Neural Networks* 166 (2023) 105–126.
- [23] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.
- [24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [25] R. G. Sargent, A tutorial on validation and verification of simulation models, in: *Proceedings of the 20th conference on Winter simulation*, 1988, pp. 33–39.
- [26] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods,

Semantic web 8 (2016) 489–508.

- [27] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, *ACM Computing Surveys* 56 (2023) 1–62.
- [28] Z. Sollenberger, J. Patel, C. Munley, A. Jarmusch, S. Chandrasekaran, Llm4vv: Exploring llm-as-a-judge for validation and verification testsuites, in: *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, 2024, pp. 1885–1893.
- [29] X. Liao, C. Chen, Z. Wang, Y. Liu, T. Wang, L. Cheng, Large language model assisted fine-grained knowledge graph construction for robotic fault diagnosis, *Advanced Engineering Informatics* 65 (2025) 103134.