Ontology Supported Quality Control in Laparoscopic Hemicolectomy

1st Bächle, Felix
Department of Mathematics
Hochschule für Technik Stuttgart
Stuttgart, Germany
felix.baechle@hft-stuttgart.de

2nd Hellig, Ortrun
Department for Visceral, Thoracic, and Vascular Surgery
Centre for Tactile Internet with Human-in-the-Loop (CeTI)
University Hospital Carl Gustav Carus at TUD Dresden University of Technology
Dresden, Germany

3rd Rayan, Younis
Department for Visceral, Thoracic, and Vascular Surgery
Centre for Tactile Internet with Human-in-the-Loop (CeTI)
University Hospital Carl Gustav Carus at
TUD Dresden University of Technology
Dresden, Germany

4th Speidel, Stefanie
Centre for Tactile Internet with Human-in-the-Loop (CeTI)
Department of Translational Surgical Oncology
NCT/UCC Dresden, a partnership between DKFZ,
Faculty of Medicine and
University Hospital Carl Gustav Carus
Dresden, Germany

5th Wagner, Martin
Department for Visceral, Thoracic, and Vascular Surgery
Centre for Tactile Internet with Human-in-the-Loop (CeTI)
University Hospital Carl Gustav Carus at
TUD Dresden University of Technology
Dresden, Germany

6th Katic, Darko
Department of Mathematics
Hochschule für Technik Stuttgart
Stuttgart, Germany
darko.katic@hft-stuttgart.de

***Abstract*—Decision-support in robot-assisted surgery requires reasoning that is both interpretable and formally verifiable. While large language models (LLMs) provide flexible interfaces, they lack guarantees of correctness and traceability, limiting their trustworthiness and applicability in such safety-critical domains. Symbolic knowledge representations such as ontologies offer a complementary approach by enabling explicit, logically verifiable reasoning over formally defined domain knowledge.**

We present a neuro-symbolic framework for ontology-grounded surgical reasoning, combining LLM-based query interpretation with Description Logic [1] inference over OWL 2 [12] ontologies. General workflows in surgery and quality criteria are encoded as formal TBox definitions, representing abstract classes of surgical procedures and their necessary and sufficient conditions. Actual intraoperative scenarios are represented as ABox assertions, i.e., concrete instantiations of surgical steps together with observed condition evidence for a specific case. Given a question, the system maps it to structured reasoning tasks, retrieves relevant ontological evidence, and returns answers grounded in a logical entailment.

We exemplarily evaluate the approach on right hemicolectomy (removal of large intestine) scenarios, covering both workflow progression and step-level quality assessment. First results show that the system produces logically valid answers aligned with formal ontology reasoning, while providing explicit evidence traces for each decision, enabling auditability.

Our findings suggest that combining symbolic reasoning with natural language interfaces enables auditable and verifiable surgical assistance as a step towards safe AI in surgery.

***Index Terms*—Neuro-symbolic, Ontology, Reasoner, surgical process model, LLM, AI,**

The authors thankfully acknowledge support by the state of Baden-Württemberg through bwHPC and the thank the Carl Zeiss Stiftung for funding this work under as part of the project "Ontology-driven Large Language Model for Surgery (OntoSPM-LLM)". And German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden

I. Introduction

Global demographic trends are widening the gap between the demand for surgical interventions and the availability of qualified surgeons, amplifying the need for surgical assistance to decrease the number of surgeons as well as the cognitive load on remaining ones [23]. Robotic surgical systems hold great promise to help reduce both manpower and cognitive as well as physical workload per procedure. However, shared surgical autonomy is safety-critical: errors are often irreversible, with exceptionally high failure costs. This makes transparency, reliability, and verifiability core requirements.

Recent advances in large language models (LLMs) have enabled their use in robotic decision-making, including high-level planning [6], [15]. While these models show great flexibility, their use in surgery is limited. A key concern is that LLMs produce plausible but unverified output without justification. In surgery, such behavior is not acceptable, as decisions must be transparent, reliable and verifiable.

Retrieval-augmented methods [16] (RAG), including Graph-RAG [21], partially satisfy these requirements by grounding with retrieved context. Yet, retrieval quality and evidence selection errors remain important failure modes [3], [18]. In high-risk surgical scenarios, these limitations motivate approaches that go beyond retrieval-only grounding and provide explicit, checkable reasoning constraints.

To address this, we investigate a neuro-symbolic approach in which surgical knowledge and reasoning processes are represented explicitly within a formal ontology and evaluated through Description Logic [1] inference.

In this work, we propose an ontology-grounded neuro-symbolic framework for surgical quality assessment, with potential applicability to high-level decision support. The current design focuses on quality control, as the ontology explicitly encodes step-level performance criteria; however, the framework itself is not limited to this scope and can be extended by enriching the ontology. An overview of our framework is depicted in Figure 1

Our approach maps natural-language questions to structured ontology queries, which are executed using a Description Logic reasoner [26]. The resulting entailed evidence is then used to generate grounded natural-language responses.

In contrast to retrieval-based approaches, the system relies on explicit ontological representations of surgical states and reasoning processes. This externalizes both intraoperative state representation and logical inference, reducing reliance on implicit model knowledge and fostering transparency.

We demonstrate the framework on quality assessment and workflow questions for the right hemicolectomy through representative ontology-based question-answering examples covering workflow relations (e.g., positioning of the patient precedes dissection) and step-level quality assessment, illustrating its ability to produce transparent, evidence-grounded answers.

II. Related Work

a) Ontology-based surgical knowledge modeling.: [9] introduce a Surgical Procedure Ontology based on the Basic Formal Ontology (BFO) [20], providing a standardized representation of surgical workflows and anatomical entities. This framework was later extended by [8] to support phase recognition in laparoscopic visceral surgeries by formalizing procedure-specific phases. Both approaches rely on OWL 2 [12] as a formal representation language, enabling structured and machine-interpretable surgical knowledge. Additionally, [8], [9] utilize the Foundational Model of Anatomy [10] (FMA) which presents a ontology for the human anatomy.

b) LLM-ontology integration for querying and reasoning.: More recently, ontologies have been integrated into LLM-based reasoning pipelines. Graph-based approaches such as Think-on-Graph [25] enable LLMs to traverse knowledge graphs in a query-aware manner, selecting relevant nodes and relations during reasoning. Alternative approaches include breadth-first traversal strategies [17], LLM-based generation of SPARQL queries [4], and ontology-guided verification of LLM outputs through iterative self-refinement [27]. [24] uses a multi-modal entity relationship extraction pipeline to enrich a predefined ontology, which is then applied to a hybrid RAG/Graph-RAG approach. These methods improve interpretability and grounding, but often lack explicit integration of formal reasoning with domain-specific quality criteria.

III. Methods

The proposed framework combines LLM-based query interpretation with symbolic reasoning over an OWL ontology. Natural-language questions are mapped to ontology-grounded reasoning tasks, followed by ontology traversal, query construction, and Description Logic inference. The LLM operates only on ontology-derived entities and relations during traversal and query generation, while all logical inference is delegated to an OWL reasoner. The resulting entailed evidence is subsequently used to generate grounded natural-language responses.

a) Ontology Design: While the proposed framework is generalizable across surgical procedures, we focus on right hemicolectomy as a representative visceral surgery procedure that offers sufficient procedural complexity while remaining non-trivial for quality assessment and tractable for proof-of-concept modeling. We selected phase 3 based of the competency assessment tool (CAT) described in [2], which details the dissection of the superior mesenteric vein and ligation of ileocolic vessels during right hemicolectomy.

We formalize this phase as an OWL 2 ontology [12], capturing intraoperative scenarios, surgical steps, and associated condition evidence. Due to the CAT-based design, the ontology primarily encodes quality assessment criteria in its current state, rather than a exhaustive relationship ontology. As a result, the current model focuses on ontology-grounded evaluation of surgical performance and workflow questions. However, the framework itself is not inherently limited to this scope and can be extended to support higher-level planning as more detailed workflow relations are incorporated.

To ensure interoperability with existing surgical knowledge representations, the ontology extends the ontology of surgical process models (OntoSPM) [9], which itself is grounded in the Basic Formal Ontology (BFO) [20].

Surgical steps, condition evidence, and quality outcomes are modeled as interrelated classes. Instead of explicitly annotating quality labels, the ontology derives the quality through logical class expressions. In particular, quality levels (e.g. *optimal*, *completed*) are defined using OWL equivalent class axioms that require the presence of specific condition evidence. This shifts quality assessment from manual annotation to formal inference, enabling consistency and verifiability.

We adopt a modeling pattern in which intraoperative observations are represented as instances of condition evidence, while quality assessments are inferred via description logic reasoning. This separation supports transparent, explainable reasoning and avoids hard-coded classifications.

The ontology was developed iteratively: procedural steps and quality criteria were extracted from the RIGHT study [2], refined through expert consultation, and formalized as OWL classes and object properties with corresponding axioms.

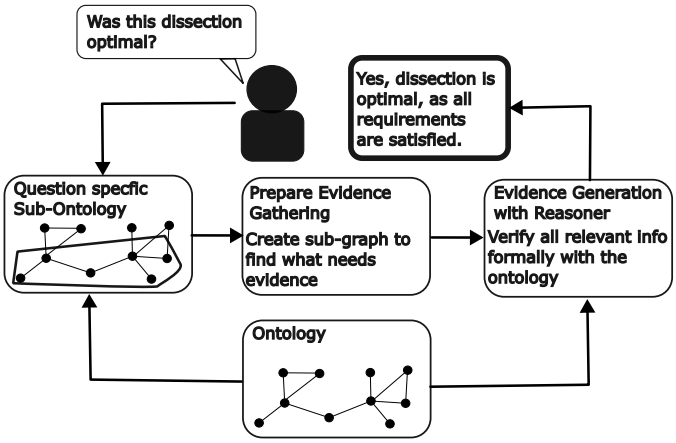

Fig. 1. architecture of our approach, the question is used to apply a semantic search on the ontology, the resulting subset of the ontology can be used as a context together with the clinical question to generate queries to the entire ontology. The results are then used to provide evidence supported answers to the original queries.

b) LLM-to-Ontology Querying Pipeline: At runtime, we assume an ABox representing a concrete intraoperative scenario, where individuals represent the real-world progress of a right-hemicolectomy. Such representation can be obtained from sensor-equipped operating rooms and situation recognition systems [14] and fed into the ontology by employing ABox generation approaches such as [7], [22].
Given a natural language question, an LLM agent first identifies a suitable anchor entity within the ontology.

Starting from the identified anchor entity, an LLM-guided traversal performs a bounded exploration of the ontology graph. At each traversal step, the LLM assigns relevance scores in the range $[0, 1]$ to ontology-derived neighboring entities and relations based on their relevance to the user query and the current reasoning context. This traversal strategy is conceptually inspired by query-aware graph exploration approaches such as Think-on-Graph [25]. Based on this relevance scoring, the system deterministically expands the top-k candidates, ensuring that traversal is restricted to valid ontology elements rather than freely generated entities or relations. While this ranking strategy may lead to suboptimal exploration paths, it prevents the generation of non-existent classes, properties, or individuals by constraining all traversal decisions to ontology-defined candidates.

The traversal acts as a relevance-filtering mechanism that extracts a compact, query-specific ontology fragment, enabling scalable reasoning without requiring traversal of the full ontology. Exploration continues either until the LLM determines that sufficient evidence has been collected or until a predefined traversal budget is exhausted. The corresponding hyperparameters were selected empirically.

The resulting sub-ontology represents a query-relevant subset of entities, relations, and axioms extracted from the ontology traversal. This fragment serves as contextual guidance for the subsequent query generation step, where the LLM constructs executable ontology queries. The generated queries are then evaluated by a Description Logic reasoner over the full ontology, which returns logically entailed results together with justification paths explaining the corresponding entailments.

The extracted entailment's form the basis for the final answer and provide an explicit reasoning trace, delegating formal logical inference to the reasoner.

We define a set of ontology-invariant query templates, including class membership, object property assertions, object property fillers, and class axiom relations. Each template corresponds to a specific OWL reasoning task, enabling a structured and verifiable mapping from natural language to formal queries. The templates are designed to be general-purpose query patterns that cover the supported reasoning capabilities of the ontology. The evaluation queries therefore represent a small but representative subset of possible questions that can be expressed within the semantic scope of the ontology, rather than a task-specific query set tailored to individual scenarios.

IV. Experiments

We construct an ontology-based evaluation setup to assess the ability of the proposed LLM-to-ontology pipeline to map natural language queries to logically entailed answers. the evaluation focuses on two complementary reasoning capabilities supported by the ontology: class-based reasoning over surgical quality criteria and relational reasoning over procedural workflow structure.

Due to the reasoning-centric nature of the task, relying on formal Description Logic entailment rather than probabilistic prediction, we do not employ quantitative performance metrics. Instead, we analyze representative questions to assess correctness, grounding, and interpretability.

a) Reasoning Backend: Ontology reasoning is implemented using the OWLAPY [5] framework in combination with the JFact [13] OWL DL reasoner. OWLAPY is used to interface with the ontology, construct class and property queries, and retrieve entailed axioms, while JFact performs the underlying Description Logic inference. We select this setup due to its compliance with OWL 2 DL semantics and its efficient support for classification and property reasoning tasks.

b) Large Language Model: We use the Gemma 4 `gemma4:latest` [11] model via the Ollama framework [19] for all language-based components, including query interpretation and natural language generation of answers from ontological evidence.

c) Scenarios: The evaluation is conducted on two handcrafted surgical scenarios derived from clinical literature and expert surgical knowledge. Each scenario encodes intraoperative condition evidence corresponding to clinically plausible states of the right hemicolectomy workflow.

V. Evaluation

Each case study consists of a natural language query, the system-generated answer, and the corresponding ontological evidence. All outputs are derived from formally entailed ABox and TBox assertions using the OWL 2 reasoner.

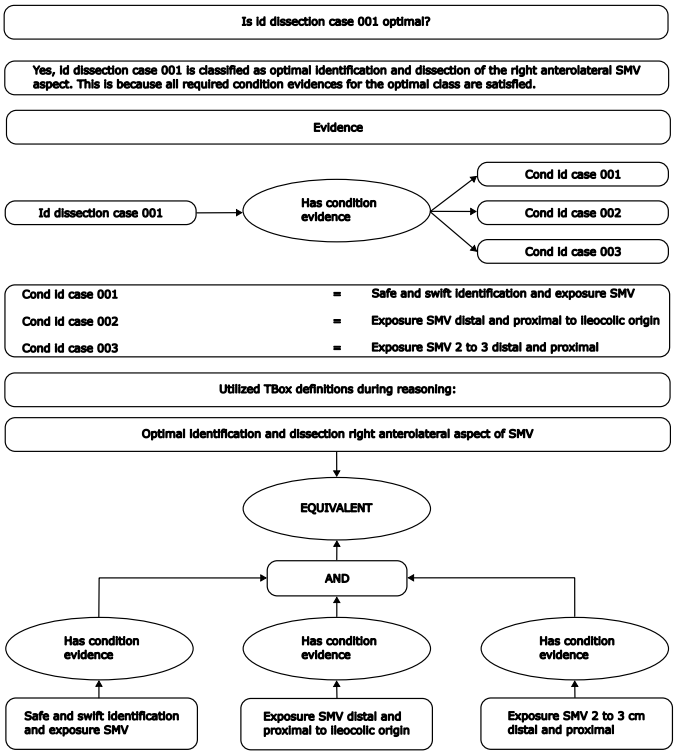

Fig. 2. Ontology-driven quality assessment query based on class expressions over `has_condition_evidence`.

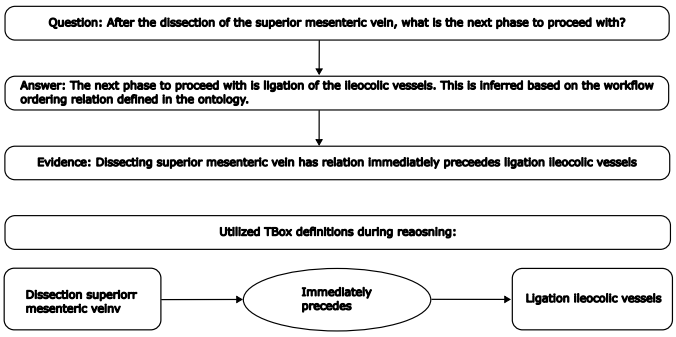

Fig. 3. Workflow reasoning query based on the `immediately_precedes` object property.

a) Q1: Step quality classification: This query (Figure 2) evaluates the system's ability to perform Description Logic-based classification over OWL class expressions. The correct decision is derived by verifying satisfaction of all existential restrictions in the TBox definition, demonstrating that surgical quality assessment is performed via formal logical entailment rather than heuristic matching or surface-level similarity.

b) Q2: Surgical workflow progression: This query (Figure 3) evaluates relational reasoning over procedural structure encoded in the ontology. The system retrieves the successor surgical phase by exploiting explicitly defined object property relations, ensuring that workflow transitions are derived through logical consequence rather than learned sequence prediction or retrieval heuristics.

c) Summary of findings: Both case studies demonstrate complementary reasoning capabilities of the proposed framework: Q1 highlights class-based semantic reasoning over surgical quality criteria, while Q2 demonstrates relational reasoning over procedural workflow structure. In both cases, the system produces answers that are fully grounded in OWL-based entailment, ensuring transparency and logical consistency of the results.

VI. Discussion

The presented framework demonstrates how ontology-grounded neuro-symbolic reasoning can enable auditable and formally verifiable surgical question answering. In contrast to retrieval-augmented generation, our approach delegates inference to a Description Logic [1] reasoner operating over explicitly defined ontological knowledge. Generated answers are therefore grounded in formally entailed facts rather than probabilistic associations. A key property of the framework is the separation between language interpretation and symbolic reasoning. The LLM is responsible for query interpretation, ontology traversal, and answer verbalization, while reasoning itself is performed symbolically through OWL inference. This reduces reliance on opaque internal model representations and enables explicit inspection of logical justifications. By representing intraoperative situations as ABox assertions and surgical workflow knowledge as TBox definitions, the framework provides auditable reasoning behavior, which is particularly important in safety-critical domains such as surgery. Furthermore, the ontology-invariant query formulation suggests that the approach can generalize beyond the current ontology and transfer to other structured knowledge bases. Several limitations remain. The framework depends on the quality and completeness of the ontology, whose construction requires substantial domain expertise and manual effort. Errors during the LLM-based semantic mapping stage may also lead to incomplete or incorrect ontology queries. Finally, the system can only reason over explicitly represented knowledge, which limits the scope of answerable questions but prevents hallucinated or unsupported outputs at the reasoning stage.

Future work will focus on extending the ontology to cover the full right hemicolectomy procedure, including richer anatomical and procedural knowledge. Another important direction is the automated population of the ABox. Integrating situation recognition methods [14], [28] would enable dynamic ontology instantiation from intraoperative data streams and allow real-time surgical state modeling. Finally, the framework should be evaluated across additional surgical procedures and heterogeneous ontologies to assess its generalizability. A systematic comparison with retrieval-based approaches such as Graph-RAG [21] would further clarify the strengths and limitations of explicit symbolic reasoning in neuro-symbolic systems.

References

[1] *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2 edition, 2007.

[2] Grüter AAJ, Toorenvliet BR, Belgers EHJ, Belt EJT, van Duijvendijk P, Hoff C, Hompes R, Smits AB, van de Ven AWH, van Westreenen HL, Bonjer HJ, Tanis PJ, and Tuynman JB. Nationwide standardization of minimally invasive right hemicolectomy for colon cancer and development and validation of a video-based competency assessment tool (the right study). *Br J Surg. 2024 Jan 3;111(1):znad404. doi: 10.1093/bjs/znad404. PMID: 38103184; PMCID: PMC10763527.*, 2024.

[3] Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Mindful-rag: A study of points of failure in retrieval augmented generation. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 607–611, 2024.

[4] Mohamed Ali, Zaki Taha, and Mohamed Mabrouk Morsey. Ontology-grounded knowledge graphs for mitigating hallucinations in large language models for clinical question answering. *Journal of Biomedical Informatics*, 175:104993, 2026.

[5] Alkid Baci, Luke Friedrichs, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. Owlapy: A pythonic framework for owl ontology engineering, 2025.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

[7] Giovanni Ciatto, Andrea Agiollo, Matteo Magnini, and Andrea Omicini. Large language models as oracles for instantiating ontologies with domain-specific knowledge. *Knowledge-Based Systems*, 310:112940, February 2025.

[8] Katić D, Julliard C, Wekerle AL, Kenngott H, Müller-Stich BP, Dillmann R, Speidel S, Jannin P, and Gibaud B. Lapontospm: an ontology for laparoscopic surgeries and its application to surgical phase recognition. *Int J Comput Assist Radiol Surg. 2015 Sep;10(9):1427-34. doi: 10.1007/s11548-015-1222-1. Epub 2015 Jun 11. Erratum in: Int J Comput Assist Radiol Surg. 2016 Apr;11(4):679. doi: 10.1007/s11548-015-1314-y. PMID: 26062794.*, 2016.

[9] Bernard Gibaud, Germain Forestier, Carolin Feldmann, Giancarlo Ferrigno, Paulo Gonçalves, Tamás Haidegger, Chantal Julliard, Darko Katić, Hannes Kenngott, Lena Maier-Hein, Keno März, Elena de Momi, Dénes Ákos Nagy, Hirenkumar Nakawala, Juliane Neumann, Thomas Neumuth, Javier Rojas Balderrama, Stefanie Speidel, Martin Wagner, and Pierre Jannin. Toward a standard ontology of surgical process models. *International Journal of Computer Assisted Radiology and Surgery*, 13(9):1397–1408, 2018.

[10] Christine Golbreich, Julien Grosjean, and Stefan Jacques Darmoni. The foundational model of anatomy in owl 2 and its use. *Artif. Intell. Med.*, 57(2):119–132, February 2013.

[11] Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2026. Accessed: 2026-07-07.

[12] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. Owl 2 web ontology language primer (second edition). W3C Recommendation, 2012. Accessed: 2026-04-30.

[13] Palmisano Ignazio. Jfact reasoner (release archive). SourceForge Project Files, 2013. Accessed: 2026-04-30.

[14] D. Junger, S. M. Frommer, and O. Burgert. State-of-the-art of situation recognition systems for intraoperative procedures. *Medical & Biological Engineering & Computing*, 60(4):921–939, 2022.

[15] Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*, 13:162467–162504, 2025.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[17] Heiner Ludwig, Thorsten Schmidt, and Mathias Kühn. An ontology-based retrieval augmented generation procedure for a voice-controlled maintenance assistant. *Computers in Industry*, 169:104289, 2025.

[18] Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions, 2025.

[19] Ollama. gemma4:latest. https://ollama.com/library/gemma4:latest, 2026. Accessed: 2026-05-13.

[20] J. Neil Otte, John Beverley, and Alan Ruttenberg. Bfo: Basic formal ontology. *Applied ontology*, 17(1):17–43, 2022.

[21] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52, 2025.

[22] Albert Sadowski and Jarosław A. Chudziak. Structured decomposition for llm reasoning: Cross-domain validation and semantic web integration, 2026.

[23] Alexander Schüßler, Christian Kunz, Rayan Younis, Benjamin Alt, Jamie Paik, Martin Wagner, and Franziska Mathis-Ullrich. Semi-autonomous robotic assistance for gallbladder retraction in surgery. *IEEE Robotics and Automation Letters*, PP:1–8, 01 2025.

[24] Chotanansub Sophaken, Kantapong Vongpanich, Wachirawit Intaphan, Tharathon Utasri, Chutamas Deepho, and Akkharawoot Takhom. Leveraging graph-rag for enhanced diagnostic and treatment strategies in dentistry. In *2024 8th International Conference on Information Technology (InCIT)*, pages 606–611, 2024.

[25] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, 2024.

[26] Dmitry Tsarkov and Ian Horrocks. Fact++ description logic reasoner: system description. In *Proceedings of the Third International Joint Conference on Automated Reasoning*, IJCAR'06, page 292–297, Berlin, Heidelberg, 2006. Springer-Verlag.

[27] Ruslan Idelfonso Magana Vsevolodovna and Marco Monti. Enhancing large language models through neuro-symbolic integration and ontological reasoning, 2025.

[28] Martin. Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M. Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Carissa Reid, Tong Yu, Armine Vardazaryan, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eung-Joo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng Ann Heng, Isabell Twick, Kadir Kirtac, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Long Zhao, Zhenxiao Ge, Haiming Sun, Di Xie, Mengqi Guo, Daochang Liu, Hannes G. Kenngott, Felix Nickel, Moritz v. Frankenberg, Franziska Mathis-Ullrich, Annette Kopp-Schneider, Lena Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical Image Analysis*, 86:102770, May 2023.