2025 NDIA MICHIGAN CHAPTER
GROUND VEHICLE SYSTEMS ENGINEERING
AND TECHNOLOGY SYMPOSIUM
OPEN SESSION TECHNICAL SESSION
AUG. 12-14, 2025 - NOVI, MICHIGAN

ENABLING GROUNDED ANSWERS THROUGH KNOWLEDGE GRAPHS AND RETRIEVAL AUGMENTED GENERATION

Danny Hoang[1], David Gorsich, PhD[2], Matthew Castanier, PhD[2], and Farhad Imani, PhD[1]

[1]School of Mechanical, Aerospace, and Manufacturing Engineering, University of Connecticut, Storrs, Connecticut, USA
[2]US Army DEVCOM GVSC, Warren, Michigan, USA

ABSTRACT

In modern defense manufacturing, achieving technological superiority hinges on both rapid decision-making and unparalleled precision engineering. Advanced machining systems, such as 5-axis CNC machines, play a pivotal role by enabling the production of intricate, free-form geometries with micron-level accuracy. However, these advances often necessitate deep domain expertise for optimal tool selection and machining parameter configuration. This paper introduces GraphLLM*, a model-agnostic approach that integrates structured knowledge graphs with large language models (LLMs) to enhance the accuracy and reliability of technical responses. By automatically extracting domain-specific entities and relationships from documents,* GraphLLM *mitigates LLM hallucinations and improves performance, especially in technically challenging or out-of-distribution queries. Experimental evaluations across various LLaMA models demonstrate significant uplifts of 25%, highlighting the framework's potential to provide grounded answers for decision-making in advanced manufacturing.*

Citation: D. Hoang, D. Gorsich, M. Castanier, F. Imani, "Enabling Grounded Answers Through Knowledge Graphs and Retrieval Augmented Generation," In *Proceedings of the 2025 Ground Vehicle Systems Engineering and Technology Symposium (GVSETS)*, NDIA, Novi, MI, Aug. 12-14, 2025.

1. INTRODUCTION

The evolution of manufacturing systems, including the adoption of 5-axis CNC machines, enables the production of critical structural components, such as suspension brackets and armored panels, with consistent material properties for defense applications [1]. These technologies are increasingly becoming more vital as increased precision and tighter tolerances are required to achieve the complex designs [2]. However, as manufacturing systems continue to evolve, the increased reliance on domain expertise becomes essential for translating intricate designs into manufacturable components. Challenges

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited. OPSEC9504 such as selecting the optimal cutting tool for specific materials and configuring the right machining parameters require expertise to provide accurate, efficient, and reliable solutions [3–5].

Large language models (LLMs) have emerged as a transformative tool by leveraging their ability to analyze vast amounts of technical data and generate contextually relevant insights. Pre-trained on diverse corpora drawn from sources such as the internet, LLMs have demonstrated applications in code completion [6], document summarization, and language translation [7]. However, despite their broad training, LLMs face a significant challenge of providing factually accurate information with a major concern being “hallucinations”, or plausible-sounding but incorrect or misleading responses. These hallucinations, coupled with the lack of domain-specific specialization, have prevented wider adoption in industries such as manufacturing, where reliable grounded solutions are essential for ensuring safety and operational efficiency.

Knowledge graphs (KGs) are structured ontologies in which entities, attributes, and relations are represented as directed nodes and edges, typically encoded as triples (subject, relationship, object). The information in these graphs is grounded in verified data and directly reflects relationships between concepts, ensuring a high level of accuracy and coherence. KGs have been implemented across various industries such as search engines [8], recommendation systems [9], and supply chain management [10]. Despite the information in these graphs being grounded in verified data, conveying this information to end-users in an accessible and intuitive way presents a challenge. KGs in isolation are inherently complex, requiring specialized query techniques to effectively retrieve and interpret the connected information. Furthermore, they often lack the depth of contextual information necessary to fully capture the nuances of complex relationships. The typically short nature of the triples may fail to convey the broader context in which entities exist and how they relate to one another. Therefore, it is crucial to develop models that can effectively incorporate the grounded nature of KGs while providing greater contextual information for end-users.

In this paper, we present GraphLLM, a model-agnostic framework that enhances large language models with a knowledge graph. By integrating structured, domain-specific information from an automated knowledge graph built on relevant text corpora, GraphLLM delivers fact-based answers to end users. Leveraging the grounded nature of the KG, the model addresses key challenges such as risk of hallucinations and lack of domain expertise, thereby enhancing reliability and precision for decision making in manufacturing.

2. METHODOLOGY

2.1. Automatic Knowledge Graph Generation

The generation of a KG is generally a time-consuming and arduous process, where domain experts determine the specific entities and relationships from amassed information and data. This is especially true for technical documentation or workflows. LLMs provide the unique opportunity to automate this process due to their ability to process large amounts of text information and extract relevant entities and relationships. The creation of a KG from text information is developed into three primary steps: 1) Gathering input files, 2) Extracting text information, and 3) Processing the text information using an LLM with prompt engineering. Figure 1 provides an illustration of the process.

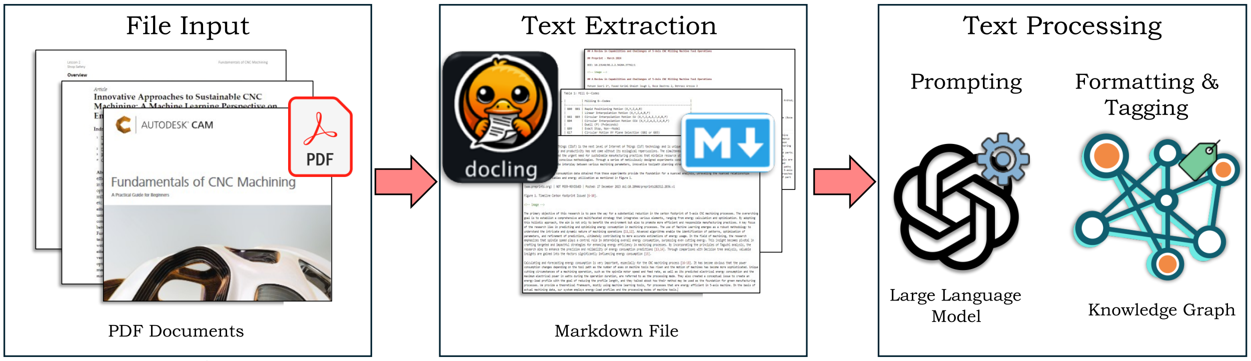

Figure 1: Overview of automatic knowledge graph generation with triplet tagging.

For the first step of generating the knowledge graph, PDF files based on the domain of interest were acquired. The files can be any material deemed relevant to the domain for which answers are required. Due to the heterogeneous nature of these PDF files, specifically in terms of formatting, Python was then used to preprocess the documents to the same formatting for all documents. More specifically, the object character recognition package Docling [11] was incorporated to extract all the text information, including those formatted in tables, and output the extracted data as general markdown files.

The last step of the automated KG pipeline involved utilizing ChatGPT-4o, along with predetermined instructions, to read through the text and extract all key entities, relationships, and the corresponding source information. By including the corresponding source information, this provides the model with additional context that is often removed when creating triplet information. This process was again implemented in Python, where the final output was a text file of generated triples per line, along with the corresponding sentence from which the entities and relationship were generated. The format of each line is as follows: subject, relationship, object, source information. The KG is subsequently stored in a PostgreSQL database for retrieval later on.

2.2. GraphLLM Model

Figure 2 provides the general overview of the KG enhanced LLM pipeline to provide more accurate and precise answers grounded in truth. Similar to the aforementioned knowledge graph generation pipeline, this model is implemented in Python.

The first step of the model involves the user providing a query for a question they wish to answer. From this query, it is given to a sentence-transformers model, specifically multi-qa-MiniLM-L6-cos-v1 (MiniLM), to perform semantic search between the question and all triples within the created knowledge graph. This sentence-transformer model was chosen due to its relatively low computational size and high performance in semantic question-answering matching, allowing users to deploy it on common desktop computers, such as those found on manufacturing floors. The MiniLM model first performs a tokenization operation where the input question is broken down into segments (tokens) and then maps these text tokens into a 384-dimensional dense vector space. This tokenization step is repeated for each of the triples in the graph as well.

Next, a cosine similarity check is performed between the embedded question vector and each of the corresponding triplet vectors from the knowledge graph. From here, the triplets with the highest semantic similarity to the question are collected based on a chosen top number of triplets.

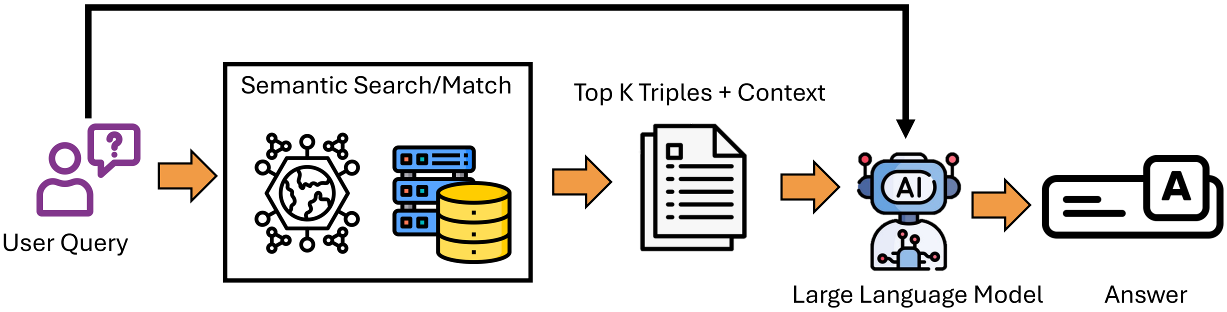

Figure 2: Overview of knowledge graph augmented large language model for question and answering.

The corresponding tagged information of the triplet is subsequently retrieved from the database and appended to the end of the triplet to be given to the LLM model answering the question. Finally, the LLM is prompted to integrate the retrieved KG information when answering a question, ensuring full fidelity to the sourced data.

3. EXPERIMENTAL DESIGN AND RESULTS

Testing of this model was conducted involving nine documents in CNC manufacturing, spanning the fundamentals of CNC machining to the capabilities and challenges facing 5-axis CNC milling machines. After processing the documents and generating the knowledge graph, a total of 2560 entities and 4329 triplets were extracted from the documents. Multiple-choice and open-ended questions were then generated using ChatGPT-o1, pertaining to the documents, into four different categories: 1) General Domain: refers to general knowledge regarding CNC machining, such as common problems and operations; 2) Intermediate Complexity: involves more domain expertise, such as common techniques frequently employed for real-time collision detection between the cutter and workpiece; 3) Technical Domain: requires greater domain expertise, such as the specific G-code required to produce a thread that requires a gradual approach rather than a single pass; and 4) Out-of-Distribution: questions regarding a recent paper published in 2025 on the optimum error design of a precision RG-cam system using a 5-axis CNC machine [12] that was not used in the knowledge graph or in LLM training for the models tested.

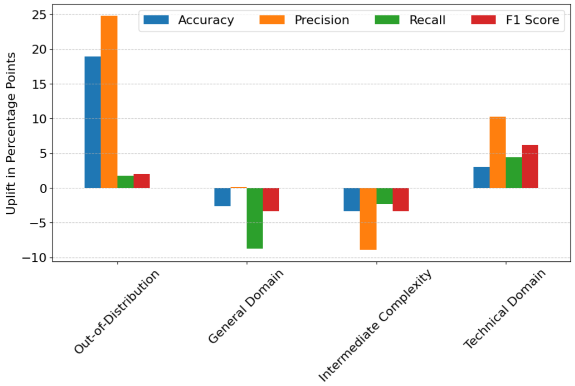

Figure 3: Uplift in the performance of LLaMA 3.2-1B across different multiple-choice questions for metrics accuracy, precision, recall, and F1-score.

The model was implemented using three open-source LLaMA variants: LLaMA 3.2-1B-Instruct, LLaMA 3.2-3B-Instruct, and LLaMA 3.1-8B-Instruct, as originally proposed by Touvron et al. [13]. Evaluating models of different sizes enabled us to assess performance across varying computational requirements; here, 1B, 3B, and 8B denote the approximate number of parameters (in billions) for each model.

Figure 3 shows the performance of LLaMA 3.2-1B on multiple-choice questions in each category for accuracy, precision, recall, and F1-score. As shown in the bar charts, the absolute uplifts (in percentage points) for the General Domain and Intermediate Complexity Domain were mostly negative. These negative uplifts can be explained by the model's limited capacity to incorporate and use the information retrieved from the knowledge graph. In these categories, the model performs relatively well on its own, but adding KG information may introduce noise or redundant details that degrade its performance. In contrast, the uplifts for the Out-of-Distribution and Technical Domain categories were all positive, with some metrics improving by up to approximately 25%. This greater improvement can be attributed to the questions being more challenging, here, the KG provides critical domain-specific insights that boost performance.

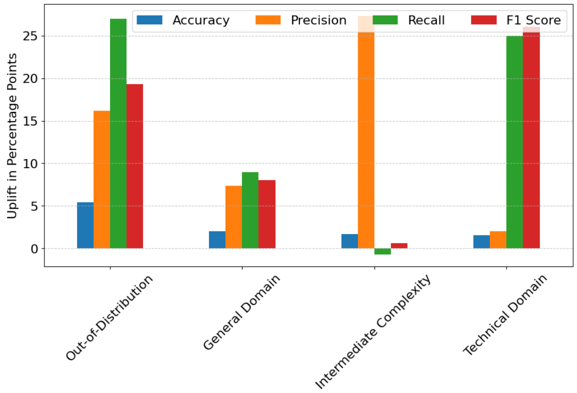

Figure 4: Uplift in the performance of LLaMA 3.2-3B across different multiple-choice questions for metrics accuracy, precision, recall, and F1-score.

Figure 4 shows the performance of LLaMA 3.2-3B. Compared to LLaMA 3.2-1B, the 3-billion-parameter model achieves mostly positive uplifts in performance. This can be attributed to the larger model's greater capacity to incorporate more data and its robust ability to integrate supplementary information into its answers. This is especially evident, as the 3B model achieved uplifts greater than 25% for metrics such as recall, precision, and F1-score.

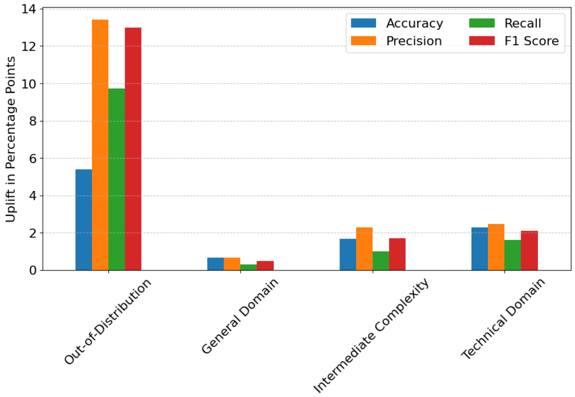

Figure 5: Uplift in the performance of LLaMA 3.1-8B across different multiple-choice questions for metrics accuracy, precision, recall, and F1-score.

Figure 5 illustrates performance of LLaMA 3.1-8B across the tested question categories and metrics. As with LLaMA 3.2-3B, performance in all question categories was positive. This suggests that as model size increases, incorporating information from the knowledge graph consistently helps the model answer questions. The maximum uplift achieved by incorporating the KG was approximately 13%. It should be noted that there appear to be diminishing returns as the largest model did not achieve a greater overall uplift compared to the smaller LLaMA 3.2-3B model.

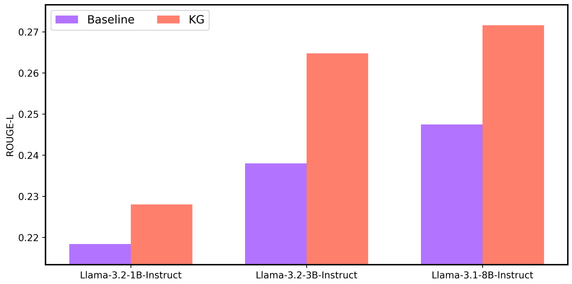

Figure 6: ROUGE-L scores comparing generated open-ended answers for GraphLLM model and baseline without knowledge graph.

Figure 6 shows the ROUGE-L of the models for open-ended questions. ROUGE-L measures the similarity between the generated text and the ground-truth answers by identifying the longest common

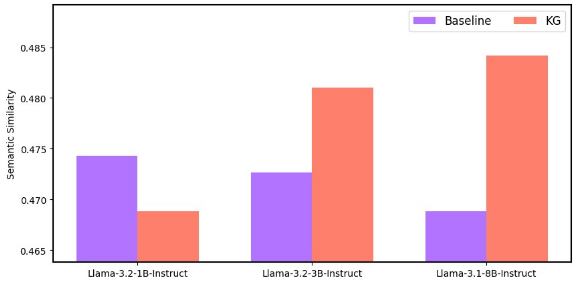

Figure 7: Semantic similarity comparing generated open-ended answers from the GraphLLM model and the baseline (without knowledge graph).

subsequence of words shared between them. Here, higher scores indicate greater agreement with the ground truth. As shown in the bar graph, LLMs with KG, compared to the baseline without, achieve higher ROUGE-L scores across all models. Furthermore, as the size of the model increases, the ROUGE-L score increases for both methods. This can be attributed to larger models being able to capture longer-range dependencies between the question and the relevant context, whether that context is inherent in their training data or supplemented via the knowledge graph. These improvements result in more coherent and contextually accurate responses that align better with the ground truth.

Figure 7 shows the open-ended semantic similarity between the generated LLM answers and the baseline. As shown, all models with the KG achieved higher semantic similarity scores than the baseline. This indicates that each model was able to incorporate the retrieved information from the knowledge graph to answer the questions effectively. Another observation is that as model size increases, similarity also increases. This is again attributed to the larger models (LLaMA 3.2-3B and LLaMA 3.1-8B) having a greater capacity to incorporate the KG information effectively compared to LLaMA 3.2-1B.

4. CONCLUSIONS

This paper introduces GraphLLM, a novel approach that incorporates a knowledge graph to tackle the challenges of domain-specific knowledge and hallucinations generated by large language models. It also demonstrates the ability to automatically generate a knowledge graph, a previously time-consuming process, from text corpora deemed relevant to questions requiring insight from technical documents. Three different sizes of large language models were integrated with this method to test its efficacy. The results show that incorporating the knowledge graph achieved uplifts of approximately 25% across different question categories. Future research will focus on incorporating additional modalities, such as image and video information, into knowledge graph.

5. ACKNOWLEDGMENT

This work was supported under Cooperative Agreement W56HZV-21-2-0001 with the US Army DEVCOM Ground Vehicle Systems Center (GVSC), through the Virtual Prototyping of Autonomy Enabled Ground Systems (VIPR-GS) program.

References

[1] D. Hoang, N. Mannan, R. ElKharboutly, R. Chen, and F. Imani, “Edge cognitive data fusion: From in-situ sensing to quality characterization in hybrid manufacturing process,” in *International Manufacturing Science and Engineering Conference*, vol. 87240. American Society of Mechanical Engineers, 2023, p. V002T06A029.

[2] D. Hoang, H. Chen, M. Imani, R. Chen, and F. Imani, “Brief paper:

Multi-task brain-inspired learning for interlinking machining dynamics with parts geometrical deviations," in *International Manufacturing Science and Engineering Conference*, vol. 88117. American Society of Mechanical Engineers, 2024, p. V002T05A012.

[3] D. Hoang, H. Errahmouni, H. Chen, S. Rachuri, N. Mannan, R. ElKharboutly, M. Imani, R. Chen, and F. Imani, "Hierarchical representation and interpretable learning for accelerated quality monitoring in machining process," *CIRP Journal of Manufacturing Science and Technology*, vol. 50, pp. 198–212, 2024.

[4] Z. Chen, D. Hoang, F. J. Piran, R. Chen, and F. Imani, "Federated hyperdimensional computing for hierarchical and distributed quality monitoring in smart manufacturing," *Internet of Things*, vol. 31, p. 101568, 2025.

[5] M. Soori, F. K. G. Jough, R. Dastres, and B. Arezoo, "Sustainable cnc machining operations, a review," *Sustainable Operations and Computers*, vol. 5, pp. 73–87, 2024.

[6] R. A. Husein, H. Aburajouh, and C. Catal, "Large language models for code completion: A systematic literature review," *Computer Standards & Interfaces*, p. 103917, 2024.

[7] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea Preprints*, vol. 1, pp. 1–26, 2023.

[8] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done," *Queue*, vol. 17, no. 2, pp. 48–75, 2019.

[9] B. Shao, X. Li, and G. Bian, "A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph," *Expert Systems with Applications*, vol. 165, p. 113764, 2021.

[10] M. Yahya, J. G. Breslin, and M. I. Ali, "Semantic web and knowledge graphs for industry 4.0," *Applied Sciences*, vol. 11, no. 11, p. 5110, 2021.

[11] C. Auer, M. Lysak, A. Nassar, M. Dolfi, N. Livathinos, P. Vagenas, C. B. Ramis, M. Omenetti, F. Lindlbauer, K. Dinkla *et al.*, "Docling technical report," *arXiv preprint arXiv:2408.09869*, 2024.

[12] C.-Y. Tsai, "Optimum error design and 5-axis cnc machining of preloaded roller-gear-cam in roller-drive system," *Measurement*, vol. 241, p. 115715, 2025.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.