Continual Learning for Generative AI Systems: Retrieval-Augmentation, Graph Reasoning, and Multimodal Integration

Lucian Gruia
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
luciancgruia@gmail.com

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—This report provides an overview of recent research activities exploring Generative Artificial Intelligence (GenAI) within the context of natural language processing, computer vision, and multi-modal reasoning. The work reviews investigations into novel architectures that extend the capabilities of large language models (LLMs) through Retrieval-Augmented Generation (RAG), Graph-Augmented RAG (GraphRAG), and multi-modal semantic search frameworks. These systems combine dense vector retrieval, graph-based reasoning, and multi-modal fusion techniques to enable context-aware generation and high-fidelity information retrieval. In addition, a survey of continual learning methods highlights state-of-the-art strategies for mitigating catastrophic forgetting and supporting adaptive, lifelong AI agents. This summary reflects ongoing efforts to develop scalable, interpretable, and adaptive AI systems, providing a cohesive perspective on current research directions and findings.**

***Index Terms*—Generative Artificial Intelligence, Large Language Models, Retrieval-Augmented Generation, Graph Neural Networks, Multi-Modal Systems, Continual Learning, Semantic Search, Knowledge Graphs**

I. Introduction

Recent advances in Artificial Intelligence (AI) have catalyzed transformative changes across a broad spectrum of domains, from natural language processing (NLP) and computer vision to knowledge management and decision support systems. Of particular significance are developments in Generative AI (GenAI), which leverages large-scale language models (LLMs) and transformer-based architectures to perform complex tasks such as text generation, summarization, and contextual reasoning. These innovations offer unprecedented opportunities for designing intelligent systems capable of interacting with humans in a context-aware and adaptive manner.

However, the integration of such models into enterprise and academic environments introduces a range of challenges, including domain adaptation, scalability, data privacy, and ethical considerations. Moreover, traditional retrieval and generation systems often lack the ability to handle multi-modal data streams or perform higher-order reasoning over complex knowledge structures. Addressing these gaps necessitates a concerted effort to investigate novel architectures and frameworks that combine state-of-the-art generative models with advanced retrieval and reasoning mechanisms.

A. Literature Review

The advent of large language models (LLMs), exemplified by GPT-3, GPT-4, and open-weight alternatives such as LLaMA, has catalyzed a paradigm shift in natural language processing (NLP), enabling unprecedented capabilities in few-shot and zero-shot generalization. However, leveraging these models effectively within domain-specific contexts often necessitates parameter-efficient fine-tuning strategies, notably Low-Rank Adaptation (LoRA) and adapter-based techniques. Complementary approaches, including prompt engineering and Retrieval-Augmented Generation (RAG) [1], have emerged as pivotal mechanisms for infusing external knowledge into LLMs, thereby enhancing their contextual relevance and factuality.

RAG architectures integrate dense or sparse retrieval mechanisms with generative transformers, enabling grounded response synthesis over large-scale corpora. Recent advancements, such as GraphRAG, augment traditional RAG pipelines by incorporating structured knowledge graphs, facilitating multi-hop reasoning and superior interpretability. Surveys on knowledge-oriented RAG [2] and multimodal RAG frameworks [3] further emphasize the potential of these architectures in addressing complex reasoning tasks across diverse modalities.

The integration of multi-modal systems, exemplified by BLIP-2 [4] and MMGraphRAG [5], underscores the growing demand for models capable of synthesizing heterogeneous data streams, including textual, visual, and auditory information. Such systems hold particular promise in domains like high-fidelity semantic video retrieval and cross-modal question answering.

Concurrently, continual learning paradigms have been proposed to mitigate catastrophic forgetting when incrementally updating models with streaming data. Foundational techniques

2nd Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 15 July, 2025

such as Elastic Weight Consolidation (EWC), replay-based strategies, and dynamic architectural modulation have shown promise in sustaining model performance across sequential tasks. Recent surveys [6] highlight the synergies between continual learning and retrieval-augmented approaches for constructing adaptive, lifelong AI agents.

These technological advances are complemented by interdisciplinary research initiatives, including the SOL projects, which explore applications in vehicular identification, NLP for low-resource languages, and semantic data integration. Such collaborative efforts underscore the imperative for scalable, robust, and ethically aligned AI systems.

B. Research Gap and Motivation

Despite notable advancements in the development of LLMs and allied technologies, several critical challenges persist:

- Engineering secure, scalable, and computationally efficient frameworks for enterprise-grade deployment of generative AI systems.
- Enhancing RAG architectures with graph-based reasoning capabilities to enable complex, multi-hop information retrieval and explainability [7], [8].
- Designing multi-modal systems capable of high-fidelity semantic video search, cross-modal retrieval, and multimodal grounding [3].
- Addressing continual learning constraints to realize adaptive and resilient lifelong AI agents [9].

This research endeavors to address these gaps by investigating novel architectures and methodologies, emphasizing practical deployment in enterprise and academic contexts.

II. Research Objectives

The overarching aim of this doctoral research is to advance the state of the art in integrating Generative Artificial Intelligence (GenAI) methodologies within enterprise and academic environments. This central goal is operationalized through the following specific objectives:

1) **To investigate and formalize methodologies for the design and deployment of large language models (LLMs) in enterprise-scale applications,** with a particular focus on parameter-efficient fine-tuning paradigms, prompt engineering, and secure, scalable deployment frameworks [10].
2) **To develop and evaluate Retrieval-Augmented Generation (RAG) systems** tailored to enterprise knowledge dissemination and laboratory workflows, enabling context-aware conversational agents capable of autonomously engaging with diverse user groups [1], [6].
3) **To design a multimodal video search framework leveraging Generative AI,** integrating textual, auditory, and visual representations for high-fidelity semantic retrieval over large-scale video corpora [3], [4].
4) **To augment RAG architectures with graph-based knowledge representations (GraphRAG),** facilitating multi-hop reasoning, improved context propagation, and enhanced interpretability of generative outputs [7].
5) **To conduct a comprehensive survey and critical analysis of continual learning methodologies,** addressing challenges such as catastrophic forgetting and lifelong adaptation in dynamic AI ecosystems [9].
6) **To contribute to collaborative SOL research projects,** advancing AI applications in vehicular identification, natural language processing for under-resourced languages, and semantic data integration, thereby underscoring the importance of interdisciplinary and socially responsible AI research.
7) **To establish a cohesive experimental framework for evaluating these systems,** incorporating quantitative performance metrics and human-in-the-loop assessment protocols to ensure robustness, scalability, and ethical alignment.

Collectively, these objectives aim to bridge the gap between foundational AI research and its translation into practical, enterprise-ready, and societally beneficial systems, thereby fostering innovation and the responsible adoption of GenAI technologies.

III. Methodology and Project Activities

A. Enterprise Trainer for Generative AI at a Multinational Company: A Strategic Partnership with the AIM Lab

In collaboration with a leading multinational corporation, the AIM Laboratory conceptualized and delivered an advanced training program on Generative Artificial Intelligence (GenAI), strategically tailored for enterprise-scale applications. This initiative sought to empower corporate professionals with a rigorous understanding of the theoretical underpinnings and practical methodologies for deploying large language models (LLMs) within complex organizational ecosystems [11].

1) Curriculum Design and Pedagogical Foundations: The curriculum was systematically structured to integrate foundational knowledge with advanced, application-driven techniques:

- **Foundational Modules:** An in-depth introduction to generative modelling paradigms, transformer-based architectures, and the principles of self-supervised learning. The ethical and legal considerations inherent in deploying generative models within enterprise contexts were critically analyzed [12].
- **Advanced Topics:** Exploration of cutting-edge LLMs, including GPT-4 and open-weight alternatives such as LLaMA 2. Modules incorporated parameter-efficient fine-tuning strategies (e.g., Low-Rank Adaptation (LoRA), adapter-based methods), prompt engineering paradigms, and Retrieval-Augmented Generation (RAG) frameworks [1] for domain-specific knowledge augmentation.
- **Hands-on Workshops:** Experiential learning sessions enabled participants to engage with API integrations (e.g., OpenAI, Hugging Face), dataset curation and annotation pipelines, secure data governance protocols, and containerized deployment practices leveraging Docker and Kubernetes [13].

2) Exploration of Cutting-Edge Technologies: Participants engaged with contemporary frameworks and tools designed to address real-world enterprise challenges:

- Fine-tuning and deployment of LLMs using `PyTorch` and Hugging Face's `transformers` library [14].
- Implementation of secure RAG pipelines integrating vector databases such as `Pinecone` and `Weaviate` for efficient, domain-aware information retrieval [6].
- Deployment of Proof-of-Concept (PoC) systems within enterprise sandboxes, ensuring scalability and compliance with stringent corporate security standards.

3) Proof-of-Concept Development and Evaluation: To validate the practical impact of the training, collaborative PoCs were co-developed with enterprise stakeholders, targeting high-value use cases such as intelligent document summarization, enterprise knowledge assistants, and interactive technical support agents.

The implementation pipeline comprised:

1) **Problem Definition:** Identification and formalization of mission-critical enterprise workflows amenable to LLM integration.
2) **System Design:** Engineering of data ingestion pipelines and integration of LLMs with retrieval mechanisms for context-aware response generation.
3) **Evaluation:** Rigorous assessment employing both quantitative metrics (e.g., ROUGE-L, BLEU) and human-in-the-loop evaluations to quantify relevance, fluency, and usability.
4) **Optimization:** Application of model compression techniques, including quantization and knowledge distillation, to reduce inference latency and resource overhead without degrading performance [15].

This initiative exemplifies the transformative potential of academia–industry partnerships in catalyzing innovation through the strategic deployment of generative AI systems at scale.

B. Development of a Retrieval-Augmented Generation (RAG) System for the AIM Laboratory

As part of the ongoing research and development initiatives within the AIM Laboratory, an advanced Retrieval-Augmented Generation (RAG) system was conceptualized and deployed to serve as an intelligent, context-aware conversational agent. This system autonomously engages with visitors, delivering semantically grounded responses to queries regarding the laboratory's research projects, scientific publications, and collaborative initiatives. Its design aligns with recent advancements in knowledge-intensive NLP architectures that integrate retrieval mechanisms with generative transformers [1], [6].

1) System Architecture and Design Principles: The RAG system architecture embodies a tightly coupled integration of large language models (LLMs) and a semantic retrieval layer to ensure factual consistency and response grounding. The key components include:

- **Document Ingestion and Indexing Pipeline:** Laboratory assets—such as research papers, project descriptions, and internal technical reports—are systematically preprocessed, vectorized using state-of-the-art embedding models (e.g., Sentence-BERT [16]), and indexed within a high-performance vector database (`FAISS` [17]) to support efficient approximate nearest neighbor (ANN) retrieval.
- **Contextual Retrieval Layer:** Upon receiving a user query, the system retrieves the top-k semantically relevant document embeddings, leveraging hybrid dense-sparse retrieval strategies [18] to maximize recall and relevance. This retrieved context is subsequently injected into the LLM to facilitate informed response generation.
- **Generative Component:** The generative module comprises cutting-edge LLMs (e.g., GPT-4 or open-weight alternatives such as LLaMA 3.2) conditioned on retrieved passages, enabling coherent, contextually aligned, and domain-specific output synthesis. Recent architectural enhancements such as Stochastic RAG [9] and Domain-Adapted RAG frameworks [10] inform the system's design to balance computational efficiency with factual fidelity.
- **Agentic Behaviors:** To foster user engagement and improve interaction quality, the system incorporates agentic capabilities including dynamic clarification prompts for ambiguous queries and proactive recommendations highlighting salient research outputs, inspired by contemporary autonomous LLM agent frameworks [19].

C. Advanced Video Search Leveraging Generative AI

Building upon prior research in semantic video retrieval, this work introduces an advanced algorithmic framework that integrates multi-modal analysis techniques within a Generative AI paradigm. The system facilitates efficient and context-aware retrieval of video content from large-scale repositories based on natural language queries. By transcending traditional text-to-video similarity matching, the framework incorporates audio signal processing, visual object recognition, and cross-modal reasoning capabilities, thereby supporting semantically rich and high-fidelity retrieval workflows [3], [4].

1) System Architecture and Multi-Modal Integration: The architecture synergistically combines a large-scale generative model for semantic understanding with specialized modules for multi-modal feature extraction and fusion:

- **Textual Query Embedding:** User queries are encoded using transformer-based language models (e.g., BERT, GPT-derived embeddings, or BLIP-2 [4]), generating dense vector representations aligned with multi-modal feature spaces.
- **Visual Feature Extraction:** Representative video frames are sampled and analyzed through advanced vision encoders, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Object-level semantics are captured via pretrained detectors such as YOLOv8 and DETR, while temporal coherence is modeled using recurrent architectures and temporal self-attention mechanisms [20].

- **Audio Feature Integration:** Audio streams are processed through spectrogram-based encodings and transformer architectures (e.g., Wav2Vec 2.0 [21]). Speech content is transcribed and temporally aligned with video frames, whereas non-speech acoustic signals (e.g., environmental sounds, background music) are vectorized to enrich semantic retrieval fidelity.
- **Multi-Modal Fusion:** A shared latent representation is constructed using cross-attention-based fusion layers [22], enabling the seamless integration of textual, visual, and auditory modalities within a unified embedding space. Recent multimodal GraphRAG frameworks [5] inform the design, facilitating cross-modal reasoning over structured knowledge graphs.

2) Retrieval Workflow and Optimization: Upon receiving a natural language query, the system executes a multi-stage retrieval pipeline:

1) **Similarity Computation:** Query embeddings are compared against multi-modal segment embeddings using hybrid scoring techniques, including cosine similarity and cross-modal attention mechanisms.
2) **Relevance Ranking:** Video segments are ranked based on aggregated semantic relevance scores, emphasizing alignment across all modalities and temporal consistency.
3) **Interactive Refinement:** The system supports user-driven iterative refinement, incorporating relevance feedback for adaptive query optimization and personalized retrieval results.

3) Evaluation and Enhancements: Quantitative evaluation was performed on benchmark datasets (e.g., ActivityNet, YouCook2, and MSR-VTT), supplemented by custom-curated video corpora relevant to enterprise domains. Retrieval performance was assessed using metrics such as Mean Average Precision (mAP), Recall@K, and Normalized Discounted Cumulative Gain (nDCG). Incorporating audio

D. Survey and Continued Research on Graph-Augmented Retrieval-Augmented Generation (GraphRAG)

As part of the ongoing research endeavors within the AIM Laboratory, a comprehensive survey and experimental investigation have been undertaken on Graph-Augmented Retrieval-Augmented Generation (GraphRAG) architectures. The primary objective is to explore the incorporation of graph-based knowledge representations within RAG systems to enhance multi-hop reasoning, contextual grounding, and interpretability in knowledge-intensive domains [7].

1) Motivation and Research Objectives: Conventional RAG architectures, which combine dense retrievers with generative transformers, often lack the capacity to model intricate semantic relationships and higher-order dependencies encoded in knowledge graphs [6]. By augmenting RAG pipelines with graph traversal and reasoning mechanisms, it becomes feasible to:

- Exploit structured knowledge graphs for improved context retrieval and entity disambiguation [23].
- Enable multi-hop reasoning pathways to address complex queries requiring chained inferences and relational understanding [24].
- Enhance the transparency and interpretability of generated outputs through explicit graph traversal paths and subgraph visualizations.

The survey systematically reviews state-of-the-art graph neural networks (GNNs), knowledge graph embedding models (e.g., TransE, RotatE), and their integration within retrieval-augmented generative pipelines.

2) Methodological Approach: The research methodology is bifurcated into two complementary phases:

1) **Literature Review:** A systematic analysis of cutting-edge GraphRAG frameworks, including KG-RAG [23], GraphRetriever [24], and GNN-enhanced LLMs, focusing on their architectural design, reasoning capabilities, and scalability.
2) **Experimental Framework:** Design and implementation of prototype GraphRAG systems incorporating:
   - Knowledge graph embedding techniques for dense vector space encoding of relational structures.
   - Graph traversal algorithms to augment retrievers with multi-hop reasoning capabilities.
   - Fine-tuned LLMs conditioned on retrieved subgraphs to generate contextually coherent responses.

3) Preliminary Findings and Future Directions: Initial experimental results indicate that GraphRAG systems outperform baseline RAG models in tasks requiring entity-relation reasoning and long-range dependency modeling. Future research will focus on optimizing the trade-off between retrieval efficiency and reasoning depth, devising scalable deployment strategies for enterprise applications, and formalizing evaluation metrics tailored to graph-augmented generative systems [5].

E. Survey on Continual Learning: Techniques and Challenges

A parallel line of inquiry involves a comprehensive survey on *continual learning* (CL) paradigms in artificial intelligence. The primary aim is to synthesize methods that enable AI systems to adaptively learn from sequential data streams while mitigating catastrophic forgetting—the tendency of models to overwrite previously acquired knowledge.

1) Scope and Motivation: This survey explores the state of the art in continual learning across multiple domains, with particular focus on:

- **Architectural Approaches:** Dynamic neural architectures, progressive networks, and expandable modules for incremental task acquisition [25].
- **Regularization Strategies:** Elastic Weight Consolidation (EWC), Synaptic Intelligence, and memory-aware synapses designed to preserve prior knowledge.
- **Replay Mechanisms:** Experience replay buffers, generative replay, and pseudo-rehearsal techniques to reinforce memory retention [26].
- **Application Domains:** Implementations in NLP, computer vision, and reinforcement learning, with an emphasis on scaling CL systems for real-world deployment [27].

2) Methodology and Expected Contributions: Employing a systematic literature review methodology, the study analyzes an extensive corpus of publications from premier conferences (e.g., NeurIPS, ICML, ACL). The anticipated contributions include:

1) A structured taxonomy of continual learning paradigms organized by architectural and algorithmic design principles.
2) Comparative evaluation of existing techniques based on empirical results and theoretical guarantees.
3) Identification of unresolved challenges such as scalability, task interference, and robust evaluation metrics for lifelong learning systems [28].

This survey aims to serve as a foundational reference for researchers and practitioners seeking to design adaptive, resilient, and lifelong learning AI systems.

F. Participation in SOL Research Projects

In parallel with the core doctoral research activities, significant contributions have been made to three *SOL* (Scalable, Open, and Linked) research initiatives. These projects advance cutting-edge artificial intelligence applications across diverse domains and exemplify interdisciplinary collaboration between academia and industry.

- **SOL5/2024:** *Advanced Integrated System for Vehicle Identification Using Multiple Recognition and Confirmation Elements Based on Artificial Intelligence.* This project focuses on designing an AI-powered framework for multi-factor vehicle identification, leveraging computer vision, sensor fusion, and probabilistic reasoning techniques to enhance system robustness and accuracy [29], [30]. Deep learning-based object detection (e.g., YOLOv8), license plate recognition algorithms, and multi-modal verification methods underpin the core system architecture.
- **SOL10/2024:** *Toolset for Processing and Linguistic Analysis for the Romanian Language (RoNLP).* This initiative aims to develop a comprehensive natural language processing (NLP) toolkit for the Romanian language, encompassing syntactic parsing, semantic analysis, and domain-adapted transformer-based language models [31]. The project also addresses challenges inherent in under-resourced languages, such as data sparsity and morphological complexity, by leveraging transfer learning and multilingual embeddings [32].
- **SOL12/2024:** *Detection of Relationships Between Entities in Unstructured and Structured Data Sets (DeteRel).* This project investigates methodologies for extracting relationships across heterogeneous datasets, combining knowledge graph construction with machine learning-based information extraction techniques [33]. By employing graph neural networks (GNNs) and ontology alignment methods, the system promotes semantic interoperability and facilitates scalable data integration workflows.

Collectively, these research contributions amplify the broader impact of the doctoral work by delivering innovative solutions in scalable AI architectures, language resource development for underrepresented linguistic communities, and semantic data processing for knowledge-centric applications.

ACKNOWLEDGMENTS

Special thanks are extended to the AI Multimedia Laboratory, CAMPUS Research Institute, and the Doctoral School of Electronics, Telecommunications, and Information Technology at the National University of Science and Technology Politehnica Bucharest (formerly University Politehnica of Bucharest) for their invaluable support and collaboration.

I am profoundly grateful to my PhD supervisor, Prof. Dr. Eng. Bogdan Ionescu, for his exceptional guidance, mentorship, and continuous encouragement throughout my research journey.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[2] H. Cheng, L. Zhang, and X. Wang, "A survey on knowledge-oriented retrieval-augmented generation: Models, applications, and challenges," *ArXiv preprint arXiv:2503.10677*, 2025.
[3] Y. Mei, R. Sun, and J. Zhao, "A survey of multimodal retrieval-augmented generation: Towards cross-modal reasoning," *ArXiv preprint arXiv:2504.08748*, 2025.
[4] J. Li, D. Hu, J. Wu *et al.*, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of CVPR*, 2023.
[5] Q. Li, Y. Xu, and W. Chen, "Mmgraphrag: Multimodal graph-augmented retrieval for generative models," in *Proceedings of ICLR*, 2025.
[6] S. Ding, B. Chen, and K. Yang, "A survey on retrieval-augmented large language models: Trends and open challenges," *ArXiv preprint arXiv:2403.09544*, 2024.
[7] X. Hu, R. Zhang, and K. Lin, "Grag: Graph retrieval-augmented generation for knowledge reasoning," in *Findings of NAACL*, 2025.
[8] S. Roll and L. Thomas, "Graph-enhanced retrieval-augmented generation for industrial knowledge systems," *Sensors*, vol. 25, no. 11, p. 3352, 2025.
[9] H. Zamani and M. Bendersky, "Stochastic retrieval-augmented generation: Towards efficient knowledge-intensive nlp," in *Proceedings of ACL*, 2024.
[10] M. Su, C. Li, and T. Wang, "Dragin: Domain-adapted retrieval-augmented generation for industrial knowledge," in *Proceedings of ACL*, 2024.
[11] L. Zhao, M. Chen, and A. Patel, "Deploying large language models in enterprise systems: Opportunities and challenges," *IEEE Intelligent Systems*, vol. 39, no. 2, pp. 34–45, 2024.
[12] L. Weidinger, J. Mellor, and M. Rauh, "Ethical and social risks of large language models: A survey," *ArXiv preprint arXiv:2112.04359*, 2021.
[13] B. Burns, J. Grant, and D. Oppenheimer, *Kubernetes: Up and Running*. O'Reilly Media, 2019.
[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of EMNLP: System Demonstrations*, 2020, pp. 38–45.
[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning Workshop*, 2015.
[16] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of EMNLP-IJCNLP*, 2019.
[17] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019.
[18] V. Karpukhin, B. Oguz, S. Min, P. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of EMNLP*, 2020.
[19] Z. Shen, S. Liang, X. Yu *et al.*, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *ArXiv preprint arXiv:2303.17580*, 2023.
[20] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of ECCV*, 2018.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[22] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of ACL*, 2019.
[23] J. Chen, W. Liu, and X. Sun, "Kg-rag: Knowledge graph enhanced retrieval-augmented generation," *ArXiv preprint arXiv:2402.08123*, 2024.
[24] H. Jiang, Z. Wang, and H. Xu, "Graphretriever: Multi-hop reasoning for open-domain question answering," in *Proceedings of ACL*, 2023.
[25] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," in *Proceedings of ICML*, 2016.
[26] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[27] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, R. Pless, and C. Kanan, "Lifelong machine learning with deep neural networks: A review," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
[28] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[29] W. Zhang, M. Chen, and A. Patel, "A survey on vehicle identification techniques: Computer vision, deep learning, and multimodal approaches," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
[30] Z. Liu, Q. Zhang, and Y. Wang, "Multimodal vehicle re-identification: A survey and benchmark," *Pattern Recognition*, vol. 114, p. 107865, 2021.
[31] M. A. Hedderich, L. Lange, D. I. Adelani, Q. Zhu, P. Nakov, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," *ACL Transactions*, vol. 9, pp. 1–43, 2021.
[32] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of ACL*, 2020.
[33] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.