Enhancing Decision Making through Similarity-Driven Knowledge Integration in Resource Allocation and Content Matching

Lasal Jayawardena[1,2,*,†]

[1]Robert Gordon University, Garthdee House, Garthdee Road, Aberdeen AB10 7AQ, United Kingdom

Abstract

This research aims to build a novel framework that enhances decision-making through an integration of similarity-driven Case-Based Reasoning (CBR) with advanced Large Language Model (LLM) techniques via Retrieval-Augmented Generation (RAG) and Genetic Algorithm (GA) optimisation. Currently, experimental work focuses on refining the loss function components to tune angle-optimised embedding models using both semi-supervised and unsupervised approaches. In parallel, experiments are being conducted to fine-tune LLMs as baselines for evaluation and to determine the best way to use LLMs as evaluative judges. Preliminary data analysis and enrichment have been conducted on operational datasets (e.g., WM Nicol company records). The final goal is to advance the state-of-the-art in CBR methods while providing a robust foundation for adaptive, context-aware decision support across multiple domains.

Keywords

Case-Based Reasoning, Large Language Models, Retrieval Augmented Generation, Genetic Algorithms, Embedding Models

1. Introduction

Decision support systems play a crucial role in managing complex, dynamic environments where historical knowledge must be effectively integrated with real-time data. Natural Language Generation (NLG) has become a cornerstone of many modern applications, driven by advances in LLMs [1]. However, challenges remain in generating responses that are both accurate and contextually relevant—especially in knowledge-intensive domains where precision and reliability are critical [2, 3]. In response, the integration of RAG with similarity-driven self-supervised metric learning offers a promising solution. By retrieving relevant information from vast datasets before text generation, RAG systems ground the generated content in factual data [4], while self-supervised metric learning enables models to understand and leverage semantic similarities more effectively.

In practical applications such as resource allocation and planning, organisations struggle to efficiently match resources (e.g., vehicles, personnel, equipment) to tasks under constraints like maintenance schedules, certifications, legislative requirements, and unexpected disruptions. By integrating similarity measures and leveraging past-case knowledge, the research seeks to develop AI-driven planning tools that allocate resources optimally and potentially use Genetic Algorithms for further optimisation. Similarly, in the career development domain, it is crucial to accurately match candidate profiles with job requirements and offer personalised upskilling recommendations. Integrating CBR with LLMs and RAG, guided by sophisticated similarity metrics and knowledge exploitation, offers promising avenues for improving the retrieval and adaptation of unstructured content such as CVs and historical hiring data.

Integrating CBR with LLMs presents unique opportunities and challenges [5]. Although LLMs excel in language understanding, their responses often lack traceability and accountability, a gap that can be

ICCBR DC'25: Doctoral Consortium at ICCBR-2025, July, 2025, Biarritz, France

*Corresponding author.

l.jayawardena@rgu.ac.uk (L. Jayawardena)

https://solo.to/lasal (L. Jayawardena)

0009-0002-7100-6015 (L. Jayawardena)

bridged by embedding the structured retrieval processes of CBR into the RAG framework. Approaches such as Hybrid CBR-RAG [6] merge the strengths of both paradigms, organising the retrieval process to match cases to queries more effectively through various similarity metrics. This integration not only enhances the contextual relevance and factual accuracy of LLM outputs but also holds potential to improve performance on knowledge-intensive tasks.

In addition to enhancing retrieval for Retrieval-Augmented Generation, there is a fundamental issue of grounding, ensuring that abstract representations within the system correspond accurately to real-world cases and constraints. [7] distinguish five dimensions of grounding—sensorimotor, communicative, epistemic, relational, and, most critically, referential grounding, which links each symbol or embedding directly to its denoted concept. Referential grounding is essential for unifying similarity driven Case-Based Reasoning with LLM-based RAG and genetic-algorithm optimisation, as it guarantees consistent semantics across retrieved cases, generated text, and fitness evaluations. Moreover, [8] highlight several complementary techniques, such as constrained decoding to anchor outputs in verified information, automated guardrails and NLI-based checks, domain specific corpus tuning, and iterative revision loops that augment pure RAG, reducing hallucinations and improving accountability and interpretability. By combining RAG with these additional grounding strategies, this can provide transparent generation and optimisation in both resource allocation and content-matching domains.

Research Aim: This research aims to address these challenges by developing enhanced techniques for RAG and self-supervised metric learning and integrating similarity knowledge to optimisation algorithms like GAs, thereby improving the accuracy and reliability of NLG systems across various complex domains

2. Research Plan

2.1. Research Objectives

The primary aims are to:

- Develop a scalable CBR system that accurately retrieves and adapts historical cases using refined similarity metrics.
- Integrate LLM-based RAG to generate decision support outputs that are verifiably anchored in past cases.
- Experiment with novel loss function formulations for such as angle-optimised contrastive learning [9] and Matryoshka representation learning [10], leveraging both semi-supervised and unsupervised approaches.
- Incorporate Genetic Algorithms with similarity-informed fitness evaluations to optimise scheduling and resource allocation.
- Explore the use of LLMs as evaluative judges to continuously assess and improve the generated responses.

2.2. Approach / Methodology

The methodology is planned to be built around a multi-agent system composed of three core modules:

1. **CBR Module:**
    - *Case Repository:* Creation and curation of an annotated database of historical cases.
    - *Attribute Extraction and Similarity Embedding:* Leveraging advanced LLMs to extract relevant information from various data sources and leverage self-supervised learning to be able to create contextual embeddings on local(attribute-level) and global levels.
2. **LLM-RAG Module:**
    - *Retrieval Process:* Generating context-sensitive queries from extracted metadata, followed by retrieval of relevant cases.

- *Generation Process:* Synthesising candidate answers by augmenting LLM outputs with grounded data from the case repository.

3. **Optimisation and Evaluation Module:**
   - *Genetic Algorithm Optimisation:* Implementing GAs to select and refine candidate decisions under complex constraints.
   - *LLM Judges:* Deploying LLMs as evaluative agents that provide a score for each generated response, thereby capturing the possible need for further revision or improvement.

3. Initial Experiments on Embedding Fine-Tuning

One of our initial experimental focus is to identify the bottlenecks and potential opportunities for improving embedding models. As part of these considerations, we draw upon the contributions of [9], who introduced a composite loss function comprising three objectives:

$$L \;=\; w_1\, L_c(S_U, S_L) + w_2 \underbrace{\left(-\sum_{i=1}^{B} \log \frac{\exp\big(S_L^{(i)}/\tau\big)}{\exp\big(S_L^{(i)}/\tau\big) + \sum_{j\neq i} \exp\big(S_U^{(i,j)}/\tau\big)} \right)}_{L_{\text{inb}}} + w_3\, L_c\big(S'_U, S'_L\big) \quad (1)$$

where

$$L_c(S_U, S_L) \;=\; \log\Big(1 + \sum \exp\big((S_U - S_L)/\tau\big)\Big),$$

and $S_L = \cos(X_a, X_L)$, $S_U = \cos(X_a, X_U)$, while S'_L and S'_U denote the cosine similarities computed on each half of the split embedding vector that conceptualises the angle formulation.

Here, τ is a temperature scalar, and each weight w_i is selected via grid search under both semi-supervised and unsupervised regimes. In preliminary experiments, fine-tuning with this composite loss function demonstrated robust stability in weighted retrieval: the variation in performance across different weighted retrieval settings was substantially lower than for other embedding models. We evaluated the experiments against the AnglE-BERT model [9] and the Vanilla BERT model [11], where our approach achieved a notable improvement in Recall@K with much higher stability.

Parallel to this loss function experimentation, we are also evaluating alternative loss formulations in both semi-supervised and unsupervised modes to determine the optimal configuration for training similarity embedding. These experiments are critical for identifying the best methodology to harmonise the CBR components with modern LLM techniques across different domains [12].

4. Progress Summary

Notable progress has been made since the start of the PHD:

- **Industry Operational Data Gathering and Enrichment:** Gathered data from WM Nicol, which is a trucking company, and this data captures operational insights that could be used to predict potential new job requirements such as time constraints and resource constraints based on similar past jobs. For now, I have completed data cleaning and preliminary data analysis.
- **Contrastive Loss Experiments:** Currently testing various contrastive loss formulations, including multi-level approaches inspired by angle-optimised and Matryoshka learning techniques, to fine-tune embedding representations by extending prior work of Sri Lanka QA context [12].
- **LLM Fine-Tuning and Evaluation:** Establishing fine-tuned LLMs as baseline comparators for QA so that it could be compared with the LLM-RAG Module to see performance gains and factual correctness through LLM-as-a-Judge methods [13].
- **Paper Acceptance:** Work done on LLM-based evaluation and its role in revision knowledge capture has been accepted for presentation at ICCBR 2025, which effectively identifies when explanation strategies require revision in Isee [14].

5. Conclusion and Future Work

This research aims to introduce a comprehensive framework that integrates similarity-driven CBR with advanced LLMs to enhance decision support systems. Current progress includes successful data enrichment, contrastive loss experiments, the initial fine-tuning of LLMs as baselines for QA, and identifying methodologies for LLM-as-a-Judge to capture revision needs.

Looking ahead, future work will focus on constructing the case repository using WM Nicol's operational data, and extending this repository across additional domains to evaluate the framework's generalisability. Further exploration will focus on enhancing multi-level embedding approaches to capture nuanced, context-aware similarities more effectively. Additionally, the integration of Genetic Algorithms will be evaluated to better manage the optimisation of resource allocation under varying constraints. These efforts are expected to contribute to the development of more reliable and transparent decision support systems, ultimately fostering higher levels of user trust and accountability in real-world scenarios.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT for the purpose of: grammar and spelling check, paraphrase and reword. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] OpenAI, GPT-4 Technical Report (2023). URL: https://arxiv.org/abs/2303.08774. doi:10.48550/ARXIV.2303.08774, publisher: arXiv Version Number: 3.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023. URL: http://arxiv.org/abs/2311.05232, arXiv:2311.05232 [cs].

[3] J. K. Kim, M. Chua, M. Rickard, A. Lorenzo, ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine, Journal of Pediatric Urology 19 (2023) 598–604. URL: https://linkinghub.elsevier.com/retrieve/pii/S1477513123002243. doi:10.1016/j.jpurol.2023.05.018.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020. Event-place: Vancouver, BC, Canada.

[5] K. Bach, R. Bergmann, F. Brand, M. Caro-Martínez, V. Eisenstadt, M. W. Floyd, L. Jayawardena, D. Leake, M. Lenz, L. Malburg, D. H. Ménager, M. Minor, B. Schack, I. Watson, K. Wilkerson, N. Wiratunga, Case-Based Reasoning Meets Large Language Models: A Research Manifesto For Open Challenges and Research Directions, 2025. URL: https://hal.science/hal-05006761.

[6] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, B. Fleisch, CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering, in: J. A. Recio-Garcia, M. G. Orozco-del Castillo, D. Bridge (Eds.), Case-Based Reasoning Research and Development, volume 14775, Springer Nature Switzerland, Cham, 2024, pp. 445–460. URL: https://link.springer.com/10.1007/978-3-031-63646-2_29. doi:10.1007/978-3-031-63646-2_29, series Title: Lecture Notes in Computer Science.

[7] M. L. Maher, D. Ventura, B. Magerko, The grounding problem: An approach to the integration of cognitive and generative models, Proceedings of the AAAI Symposium Series 2 (2024) 320–325. URL: https://ojs.aaai.org/index.php/AAAI-SS/article/view/27695. doi:10.1609/aaaiss.v2i1.27695.

[8] K. Kenthapadi, M. Sameki, A. Taly, Grounding and evaluation for large language models: Practical challenges and lessons learned (survey), in: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, Kdd '24, Association for Computing Machinery, Barcelona, Spain and New York, NY, USA, 2024, pp. 6523–6533. URL: https://doi.org/10.1145/3637528.3671467. doi:10.1145/3637528.3671467, number of pages: 11.

[9] X. Li, J. Li, AnglE-optimized Text Embeddings, 2024. URL: http://arxiv.org/abs/2309.12871, arXiv:2309.12871 [cs].

[10] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, A. Farhadi, Matryoshka Representation Learning, 2024. URL: http://arxiv.org/abs/2205.13147. doi:10.48550/arXiv.2205.13147, arXiv:2205.13147.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: http://aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[12] L. Jayawardena, N. Wiratunga, R. Abeyratne, K. Martin, I. Nkisi-Orji, R. Weerasinghe, SCaLe-QA: Sri lankan Case Law Embeddings for Legal QA, in: SICSA REALLM, 2024, pp. 47–55. URL: https://ceur-ws.org/Vol-3822/short6.pdf.

[13] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods, 2024. URL: http://arxiv.org/abs/2412.05579. doi:10.48550/arXiv.2412.05579, arXiv:2412.05579 [cs].

[14] M. Caro-Martínez, J. A. Recio-García, B. Díaz-Agudo, J. M. Darias, N. Wiratunga, K. Martin, A. Wijekoon, I. Nkisi-Orji, D. Corsar, P. Pradeep, D. Bridge, A. Liret, iSee: A case-based reasoning platform for the design of explanation experiences, Knowledge-Based Systems 302 (2024) 112305. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705124009390. doi:10.1016/j.knosys.2024.112305.