Case-Based Reasoning Meets Large Language Models: A Research Manifesto For Open Challenges and Research Directions

Kerstin Bach, Ralph Bergmann, Florian Brand, Marta Caro-Martínez, Viktor Eisenstadt, Michael W. Floyd, Lasal Jayawardena, David Leake, Mirko Lenz, Lukas Malburg, et al.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Case-Based Reasoning Meets Large Language Models: A Research Manifesto For Open Challenges and Research Directions

Kerstin Bach[1] | Ralph Bergmann[2,3] | Florian Brand[2,3] | Marta Caro-Martínez[4] | Viktor Eisenstadt[5] | Michael W. Floyd[6] | Lasal Jayawardena[7] | David Leake[8] | Mirko Lenz[2,3] | Lukas Malburg[2,3] | David H. Ménager[9] | Mirjam Minor[10] | Brian Schack | Ian Watson[11] | Kaitlynne Wilkerson[8] | Nirmalie Wiratunga[7]

[1]Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

[2]Artificial Intelligence and Intelligent Information Systems, University of Trier, 54296 Trier, Germany

[3]German Research Center for Artificial Intelligence (DFKI), Branch University of Trier, Behringstraße 21, 54296 Trier, Germany

[4]Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, Madrid, Spain

[5]German Research Center for Artificial Intelligence (DFKI), Smart Data & Knowledge Services, Trippstadter Str. 122, 67663 Kaiserslautern, Germany

[6]Knexus Research LLC, National Harbor, MD, USA

[7]School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen, United Kingdom

[8]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

[9]AI and Autonomy Group, Parallax Advanced Research, Beavercreek, OH, USA

[10]Department of Computer Science, Goethe University, Frankfurt am Main, Germany

[11]School of Computer Science, University of Auckland, Auckland, New Zealand

Correspondence

Lukas Malburg, Artificial Intelligence and Intelligent Information Systems, University of Trier, 54296 Trier.

Email: malburgl@uni-trier.de

Abstract

In recent years, the surge of Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), has led to a significant increase in the use of hybrid systems, which combine the strengths of different Artificial Intelligence (AI) paradigms to achieve better performance and efficiency. Although LLMs demonstrate remarkable effectiveness across numerous tasks due to their flexibility and general knowledge, they often face challenges related to accuracy, explainability, and their limited memory. Case-Based Reasoning (CBR), on the other hand, excels by recalling past experiences and using them to solve new problems, making it particularly well suited for tasks that require contextual understanding and decision-making. However, CBR systems suffer from issues such as the acquisition of various kinds of knowledge and the application of methods during the 4R cycle. In this paper, we identify several challenges plaguing LLMs and CBR systems and propose opportunities to combine the strengths of both methodologies to address these challenges. In addition, we outline future research directions for the community to explore.

KEYWORDS

Large Language Models, Case-Based Reasoning, Large Reasoning Models, Retrieval Augmented Generation, Challenges and Opportunities.

1 | INTRODUCTION

Generative Artificial Intelligence (GenAI) has gained significant importance in recent years and has excited the public, industry, and academia with its capabilities. This is particularly due to the development of Large Language Models (LLMs), which have demonstrated their capabilities in natural language processing but are easily applicable to a wide range of tasks. LLMs can be integrated as part of broader systems, combined with other Artificial Intelligence (AI) approaches, to develop hybrid AI systems that combine the strengths of different methods. For example, in the domain of AI planning, LLMs have been used to generate plans[13,28] and to remedy the high knowledge acquisition costs for traditional planning systems[8]. In addition to this investigation, in which a symbolic AI approach in the form of model-based AI planning is combined with data-driven methods in the form of LLMs, the integration of LLMs with Case-Based Reasoning (CBR) can be considered as a promising connection as well. CBR is a problem-solving methodology[33] based on analogy with the premise that similar problems have similar solutions[1,15]. In this vein, a case base with previously solved problems is used to retrieve similar situations to solve new problems. This approach is inspired by the human ability to solve problems by drawing analogies to past experiences. However, Case-Based Reasoning (CBR) applications require a certain kind of knowledge in the form of cases for the case base, a vocabulary describing the cases, similarity measures required to compare cases, and perhaps adaptation knowledge to modify retrieved cases to better fit the user query. In this context, the CBR research community sees very promising opportunities for future synergies between CBR and LLMs, benefiting hybrid AI systems. Conversely, CBR can help LLMs due to its ability to store and reuse lessons such as valid and best practices in the form of cases. However, the integration of LLMs with CBR has challenges that must be explicitly addressed.

In this paper, we explore key open challenges that require solutions to beneficially integrate LLMs with CBR and vice versa. The challenges are based on a comprehensive literature review and are also based on the expertise of the authors in the field of CBR and LLMs. In addition to these challenges, we provide a set of opportunities as future research directions to address the challenges. This paper aims to guide researchers and practitioners in the field of AI in the development of hybrid AI systems that combine the strengths of LLMs and CBR to solve complex problems in a wide range of domains. In this vein, we want to elucidate that the use of CBR in combination with LLMs is not only beneficial for the CBR community but also for the LLM community, which might currently not be familiar with the CBR methodology. This paper is the first analysis of challenges and opportunities for the integration of LLMs with CBR and vice versa and thus can be seen as a first step toward bridging the gap between these two AI paradigms.

The paper is structured as follows: In Section 2, we present relevant foundations of CBR and LLMs. Subsequently, we discuss open challenges that arise when CBR and LLMs are synergistically combined and must be addressed in future research in Section 3. However, the combination of CBR and LLMs also opens up new research directions with associated opportunities that can be achieved by a hybrid system. We discuss these upcoming research directions and opportunities in Section 4. Finally, Section 5 concludes the paper by summarizing the main findings of our research work and by discussing possible future work for the CBR research community.

2 | FOUNDATIONS

This section discusses the synergy between CBR and LLMs, exploring how insights from cognitive science and classical CBR can be integrated with modern AI methods to enhance long-term memory, knowledge management, retrieval mechanisms, maintainability, and explainability. Inspired by cognitive science, Section 2.1 proposes episodic memory for enhanced LLM retention. Section 2.2 introduces structured CBR knowledge containers guiding LLM generation. Integrating with Retrieval-Augmented Generation (RAG), Section 2.3 discusses context-specific retrieval improvements. Section 2.4 highlights LLM-supported knowledge acquisition in CBR systems. Approaches to maintaining and refining CBR knowledge bases through LLMs are covered in Section 2.5. Finally, Section 2.6 illustrates LLM-enhanced, personalized explanations for improved model transparency.

2.1 | Case-Based Memory

Initial LLMs lack long-term memory, hindering tasks that require long-term context, personalization, and learning; episodic memory has been proposed as "the missing piece" for long-term LLM agents[24]. Cognitive science and CBR lessons offer an opportunity to develop AI memory systems. Human memory includes sensory, short-term (working), and long-term memory. Long-term memory has declarative (episodic and semantic) and procedural memory. Episodic memory is crucial for recalling past events, while procedural memory is vital for embodied AI.

Episodic memory is the primary focus of CBR and is most relevant to acquiring human-like memory in LLMs. The memory model inspiring many CBR systems is Schank's Dynamic Memory Theory[29]. In this theory, memory processing is tied to understanding processes. The same structures, called Memory Organization Packages (MOPs), guide understanding and organize memory. MOPs are organized hierarchically, and collections of MOPs are assembled dynamically for processing or retrieval. Because a MOP can index cases from multiple contexts, the model accounts for cross-contextual remindings; because learning can change a MOP used later to assemble a memory. Memory structures and the memories they reconstruct can change to reflect new understanding. Such a model, in LLMs, could enable recognition and application of useful information across contexts and automatically adjust retrieved information.

2.2 | Case-Based Knowledge Structures

The knowledge container view of CBR[27] describes the knowledge components used by CBR. The vocabulary container (1) provides the foundation for explicitly describing the knowledge elements used in the CBR system. The case base container (2) records experiences as cases. Representation formalisms for cases range from plain text or simple attribute-value pairs to structures like workflows, graphs, or plans. The similarity container (3) consists of similarity functions that are used for retrieval purposes. Embedding-based similarity functions are a special form complemented by a large body of additional similarity concepts. Typical examples are Euclidean distance, explicit value tables for symbolic attributes, or edit distances for graphs or texts. Some traditional CBR similarity functions derive values from additional knowledge structures, such as taxonomies or knowledge graphs. The adaptation container (4) comprises information on how to modify a solution, for instance, by a rule. The knowledge units from the four different containers help solve a problem in combination. Structured knowledge plays an important role in guiding LLM generation processes and for the purpose of de-hallucination.

2.3 | Case-Based Retrieval

Case-based retrieval is a fundamental component in integrating CBR with LLMs. In classical CBR, a case base stores past problem–solution episodes that are retrieved and adapted to solve new challenges. Within the context of RAG[18], the structured information from these cases can be leveraged to enrich the LLM's prompt with context-specific details. For example, the CBR-RAG model[35] employs both local and global embedding strategies to select legal cases that closely match the current query, thereby enhancing the precision of the retrieved information. Furthermore, by grounding the retrieval process in explicit case data, the system supports transparent decision-making and explainability, allowing users to trace the rationale behind retrieved results through similarity metrics and case comparisons. This integration not only improves the accuracy and relevance of generated responses, but also aids in mitigating issues such as hallucinations by providing verifiable evidence in retrieval processes. Additionally, the case container structure itself in the retrieved cases can aid the LLM in generating more robust and well-structured answers, an advantage that is not readily available with traditional RAG approaches.

2.4 | LLM-Based Knowledge Acquisition

LLMs offer great potential to support knowledge management activities in various scenarios in research and practice. One general application scenario is to help software developers in writing programming code or meaningful documentation afterwards. In addition to this, LLMs have recently been applied to remedy the high knowledge acquisition and modeling efforts that are typically required to use AI methods. For example, the AI planning community has been one of the first to use LLMs to support the demanding task of creating formal planning domain descriptions[8]. Thus, the application of LLMs for CBR can also be beneficial for knowledge acquisition to fill CBR knowledge containers. Brand et al.[4], demonstrate how LLMs can be used to create cases or domain knowledge in the form of the needed vocabulary based on given few-shot prompts. The results indicate that LLMs are helpful in the context of CBR to support knowledge acquisition and modeling.

2.5 | LLM-Based Maintainability

LLMs can be crucial in maintaining case-based reasoning systems by acquiring and curating knowledge for the knowledge containers. Besides creating cases, the refinement of similarity measures can be accelerated. They can assist in adjusting similarity measures by learning latent features from past cases, dynamically revising attribute weights based on contextual factors, and incorporating structured attributes and unstructured textual descriptions. For adaptation knowledge, LLMs can propose adaptation rules or adaptation cases. By continuously updating the case base, refining similarity assessments, and enhancing adaptation strategies, LLMs can help maintain a more accurate and efficient CBR system.

2.6 | LLM-Based Explainability

A few recent works in the literature illustrate how LLMs can enhance the explainability features in artificial intelligence models. One of these examples is the iSee project, where a CBR-driven platform has been developed to define, share, and reuse explanation experiences[5]. The iSee platform includes an LLM-based module that extends image-based explanations with natural language description to answer questions the user asks about those explanations. Another example by Chen and Zhang[6], builds a case base of questions that people may ask an AI system (seed cases). The questions and answers are not created by CBR, they are collected from Internet forums or studies. LLMs are used to generate variations of these cases. In the work by Feng et al.[7], LLMs are applied in the retrieval step and help find the relevant facts that need to be retrieved in an explanation, which is a fact weighting method. We have another example about LLMs employed in XAI in the work by Queipo-de-Llano et al.[25]. The paper describes a deep learning model (ResNet) to detect fractures in X-rays. It includes explanations generated by Grad-CAM, and a CBR module to generate explanations-by-examples and textual explanations modified by LLMs. LLMs are used as a module to personalize the explanations according to the profile of the user who will receive the explanation. Therefore, there are advantages of using LLMs to enhance XAI according to previous works, and based on them, there are many research lines of work that we could follow.

3 | OPEN CHALLENGES

In this section, we present the open challenges for the integration of CBR and LLMs based on a comprehensive literature review and our expertise in these fields. An overview of the open challenges and their thematic groups is shown in Figure 1. We begin with open challenges for LLMs that are categorized into four thematic groups: *Guided Generation* (see Sect. 3.1), *Accuracy & Dehallucination* (see Sect. 3.2), *Explainability & Trust* (see Sect. 3.3), and *Operational Efficiency* (see Sect. 3.4).

Guided Generation explores challenges in prompt engineering, memory management, adaptation, and reasoning. *Accuracy & Dehallucination* focuses on challenges for evidence-based retrieval, fine-granular RAG, and validation. *Explainability & Trust* addresses the challenges related to explainable reasoning, human alignment, and expertise. *Operational Efficiency* is a more general category that overspans the other groups as it deals with challenges related to the overall performance w.r.t. fine-tuning, case-based distillation, and Large Reasoning Models (LRMs).

3.1 | Guided Generation

C1 – Aligned Prompts. Creating good prompts is a knowledge-intensive task, which depends heavily on the user's experience[2,9]. In addition, the performance of LLMs depend on the diversity and similarity of the examples in the few-shot prompts to the current query[31]. Furthermore, LLMs need to re-generate the answer for every new query from the user, which is computationally expensive and inefficient.

C2 – Context Window Limitation. LLMs typically have a limited context window size in the form of possible tokens. Structured information, e.g. in the form of workflows, graphs or plans, need to be fed into the model in some way. Using regular

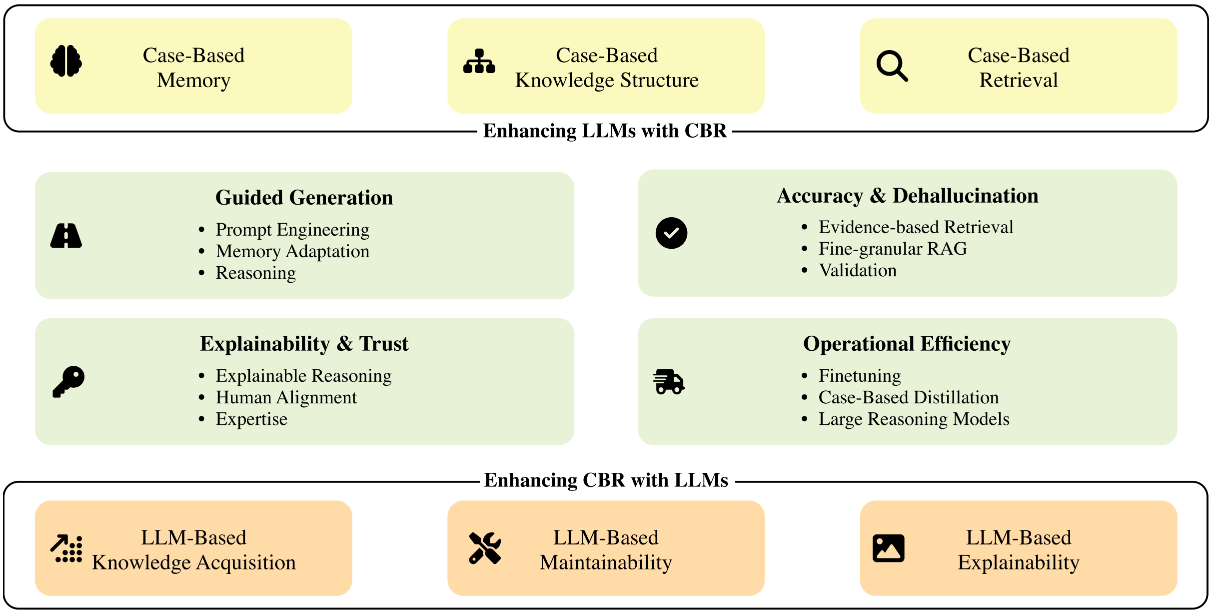

FIGURE 1 Overview of Open Challenges for CBR-LLM Integration

serialization formats (e.g., JSON, XML) can lead to a large input size, which the model may not support, reduces the response time, and increases the cost of the inference[21]. Textual representations may not be sufficient to capture the full context. Overly large contexts could negatively impact the generation even before hitting the strict limit of the context window size[21].
C3 – Domain and Adaptation Knowledge. LLMs re-generate the entire content instead of solving individual issues by actual refinement. Partial solutions get lost during iterations. New errors are introduced in previously correct solution parts[21]. Adaptation via plain prompt instructions led to low accuracy values in a classification task scenario[34].

3.2 | Accuracy & Dehallucination

C4 – Validation Requirements. The output of LLMs is not deterministic, leading to multiple issues[16,17]: The same input can lead to multiple outputs depending on the probability distribution of the model. Even if prompted to return some specific format, the response may still not conform to the expected structure. The model may return structurally correct responses, but with the wrong semantics, also known as hallucinations. This means that the output of the model must be validated to ensure that it meets the requirements of the task.
C5 – Enhancing Relevance and Precision during Generation. Often with RAGs or other LLM retrieval methods, the retrieved results are broad or generic, lacking the fine-grained specificity needed for critical domains such as legal or medical reasoning. This can result in less relevant or noisy data being processed, affecting the quality of the generated responses by the LLM. The absence of robust retrieval mechanisms that enable LLMs to access external, verifiable data to ground their responses is an omission that contributes to factual inaccuracies and increases the risk of hallucinations.

3.3 | Explainability & Trust

C6 – Enhancing Interpretability and Trust in Generation. The generation processes of LLMs can be opaque, making it difficult to understand or trust why certain information was generated. This lack of transparency hinders user trust and acceptance. Transparent retrieval mechanisms that provide explanations based on similarity metrics or other criteria are essential for justifying retrieval decisions and underpinning trust in LLM-generated responses.
C7 – Enriching Explanations. Explanations generated by Explainable Artificial Intelligence (XAI) techniques might not be as useful as expected. CBR-based explanations may lack details since they are instance-based explanations that do not usually provide descriptions of the examples shown as explanations.

C8 – Building XAI Strategies as a Combination of Explanations. Explanation units are usually shown to the target users in solitude. However, these users need to understand AI models from different points of view, so different types of explanations are required for specific situations.

3.4 Operational Efficiency

C9 – Computational Complexity and Size of LLMs. Typically, LLMs are large models trained on large corpora of training data to ensure high usability of the LLM in diverse application scenarios. However, this also leads to a high computational complexity and size of the LLM during inference. Furthermore, they are constrained by the context window, which limits the amount of information that can be processed at once.
C10 – Knowledge Acquisition Bottleneck. When developing CBR systems, the knowledge acquisition bottleneck[11], which originates from knowledge-based systems, is still one of the most prevalent challenges for the community. This touches all areas where knowledge needs to be captured and made accessible for the CBR system: Most comprehensively in the adaptation container, but it also drives the similarity measure(s) development and, to some degree, the case base and the vocabulary[26]. Furthermore, one of the biggest challenges for CBR systems is the "cold-start problem", i.e., the creation of the initial knowledge.
C11 – Limited Applicability of CBR Algorithms. Over the recent years, the CBR community has offered a pool of concrete methods and algorithms that can be applied to the phases of the 4R cycle[1]. However, some of these methods have to be adapted to the concrete application domain, which limits the transferability and applicability to these domains.

4 RESEARCH DIRECTIONS

A wide range of exciting research opportunities arise for future CBR-LLM systems. Today, we see the most promising strands in case-based *memory support* (see Sect. 4.1), *guided generation* (see Sect. 4.2), improved *accuracy and dehallucination* (see Sect. 4.3), increased *explainability and trust* (see Sect. 4.4), better *operational efficiency* (see Sect. 4.5), and the use of *LLMs for CBR* (see Sect. 4.6). In the following, we will sketch a couple of opportunities and ideas for each research direction. Some of them like case-based RAGs can be easily achieved, while others like reducing dehallucination pose more significant research challenges.

4.1 Memory Support

O1.1 – Case-Based Long-Term Memory (LTM) for Interaction and Knowledge Sharing. A long-term memory with cases for user interactions and system responses, functioning within a Dynamic Memory framework enabling the generation of useful abstractions and larger-scale knowledge structures, can provide the basis for new functionality for LLM users. For example, learned cases can provide a context-focused resource for RAG, can provide information for long-term personalization, can enable the generation of more consistent responses through reuse and adaptation, and can be used to anticipate errors. Cross-contextual reuse supported by dynamic memory can increase flexibility, and the resulting memory structure can be shared to enable the reuse of experiential knowledge for analogous tasks across systems and task domains. Case capture and reuse could also support speedup compared to solution generation from scratch.

4.2 Guided Generation

O2.1 – Episodic Memory for Prompts and Their Completions. A case base of previous prompts that resulted in successful, user-accepted responses can be leveraged to enhance prompt formulation. This includes rephrasing, clarifying, and completing user inputs. Novice users can benefit from guidance to improve their prompts, while collaborative knowledge can be integrated to refine and optimize prompt quality.
O2.2 – Prompt Engineering. Investigating prompt formulation mechanisms can lead to knowledge for streamlining generation workflows. Over time, this could lead to improved consistency, efficiency, and user satisfaction, particularly in complex or specialized use cases. Today, promising areas of focus include designing templates[30]—for instance, recommending prompt templates or best practices tailored to specific contexts and exploring ways to fill these templates with case data using methods

like episodic memory. This can help study how case selection affects the relevance and quality of responses, possibly combined with automated prompting techniques such as DSPy [14]. Potential applications include domain-specific workflows, such as generating code documentation, writing clinical summaries, or drafting legal contracts, with adaptable templates offering personalized support for varying expertise and stylistic preferences.

O2.3 – Reasoning for Filtering, Augmentation, and Refinement. Iterative generation approaches like RAG can benefit from novel CBR-LLM approaches. The optimal number of refinement iterations in RAG can be determined through CBR methods, for instance, via abort conditions or using similarity thresholds. Alternatively, case-based planning [10,3,23] can be used to guide the reasoning chain of the LLM. A solution idea is to decompose generation tasks and compose generated results following a case-based plan. LLMs can also serve as candidate plan generators [13] within a case-based planner taking the responsibility for reviewing, filtering and augmenting. Further, methods of process-oriented CBR [22] can be used to represent sequential calls to an LLM by a workflow and provide a way to formalize interactions between existing systems and LLMs. In the agent community, similar efforts are known as agentic workflows [32].

4.3 | Accuracy & Dehallucination

O3.1 – Case-Based Retrieval-Augmented Generation. Case-Based Reasoning (CBR) enhances retrieval precision by comparing cases at multiple levels—both locally (individual facts) and globally (overall context). It uses structured case knowledge to retrieve more relevant and context-specific information. This improves relevance and specificity in the information provided to LLMs, leading to higher-quality and more accurate responses.

O3.2 – Dehallucination through Case-Based Structured Knowledge. Structured knowledge from the CBR knowledge containers provides various opportunities to reduce hallucinations. Future work could investigate how to utilize the case structure for validation purposes. Hallucination in text-based outputs are hard to detect and fix, so structured outputs [20,19] provide a way of checking for correctness. Even supervisory agents for CBR-LLMs in different roles like output generators, reviewers, or adaptation advisors [12] are conceivable. Similarly, structured inputs can be used to inject additional knowledge (e.g., domain and/or adaptation knowledge) that have the potential of better guiding the LLM towards a correct response.

O3.3 – Evidence-Based and Fine-Grained Retrieval. CBR grounds LLM outputs in real cases and factual data, providing concrete evidence to support generated answers. It reduces hallucinations and increases the factual accuracy of LLM responses, thereby enhancing reliability and trustworthiness. Moreover, leveraging structured case knowledge to enable multi-level retrieval that compares cases both locally and globally can further enhance the relevance and specificity of the information provided to LLMs. Future research may explore optimised similarity metrics and multi-level retrieval strategies to refine prompt context and improve response quality.

4.4 | Explainability & Trust

O4.1 – Prompt Engineering Training for Human Experts. For users to build trust, we need to understand how they can be trained to follow the LLM's desired workflow: this includes prompt engineering as well as contesting the responses. Therefore, developing methodologies, tools, and best practices for equipping individuals to question and explore LLMs. Further, investigating the use of episodic memory or case-based reasoning systems to provide real-time feedback and guidance during training to enhance skill acquisition.

O4.2 – Explainable Reasoning with Domain Knowledge. CBR provides explanations based on similarities to past cases, making retrieval decisions understandable and justifiable. Developing transparent retrieval frameworks that explicitly explain case similarity-based decisions can significantly enhance interpretability and foster user trust. For example, a system could generate a visual or textual summary detailing which features or case attributes contributed most to the retrieval decision and how similar cases were weighted.

O4.3 – Human Alignment Opportunity. One of the main opportunities that we can carry out is to perform user studies. User studies can contribute to identifying and understanding common difficulties and misconceptions. As a preliminary work, we need to know how the explanations provided by CBR or LLMs (or both) can align users' mental models and how we can improve them. Furthermore, other opportunities for LLMs and CBR combined-based approaches exist in order to adapt and personalize XAI strategies to human alignment: getting different types of explanations obtaining them from transforming case-based explanations, explaining and personalising case-based explanations describing the cases, and their contexts for specific users, and creating case-based explanations from seed cases (users' opinions) that align users' mental models.

4.5 | Operational Efficiency

O5.1 – Case-Based Distillation of Language Models. CBR provides a framework to identify and focus on critical facts, improving the model's ability to reason about relationships between different pieces of information. This enhances the model's computational efficiency and ensures effective use of LLMs within processing constraints.
O5.2 – Fine-Tuning of LLMs with Structured Cases. The rich structure of typical case representations in CBR systems can be used to fine-tune LLMs. This may be especially useful for the trending topic of structured outputs, where LLMs are trained to generate responses that conform to a given JSON schema. This way, smaller models can be fine-tuned on structured cases to increase inference speed for certain domains.
O5.3 – Large Reasoning Models with Structured Cases. Compared to regular language models, the recently developed reasoning models use a built-in chain-of-thought reasoning mechanism to solve complex tasks. Via structured case information, this reasoning process can be augmented with additional contextual information, such as domain knowledge or adaptation knowledge.
O5.4 – Efficient Retrieval Mechanisms. Future research can explore efficient case selection strategies that focus on extracting critical facts from large case bases, especially when operating within limited context windows in LLMs. For example, one approach could include dynamic adjustment of similarity thresholds or adaptive retrieval pipelines that balance between explanation depth and computational efficiency.

4.6 | LLMs for CBR

O6.1 – LLMs as Knowledge Engineers. LLMs can be used to overcome the cold-start problem by generating initial knowledge bases or by improving existing ones[4]. They can also be used to generate new knowledge by analyzing large amounts of data and identifying patterns or relationships that may not be immediately apparent to humans. In addition, LLMs can be used to automate the process of creating and updating knowledge bases, reducing the time and effort required for manual curation.
O6.2 – LLMs in Various Parts of the CBR Cycle. They can be used in various parts of the CBR cycle during the run-time of the CBR system. This includes the retrieval step[34], as well as the adaptation of new cases. LLMs can also help to maintain the current knowledge base by identifying and correcting errors or inconsistencies, as well as by suggesting new knowledge to be added to the system.
O6.3 – LLMs to Improve CBR Outputs. LLMs can also be used to improve the quality of outputs by the CBR system itself by personalizing the output based on the user's preference or level of understanding[25]. It can also be used to turn structured cases or reasoning traces of the CBR system into human-readable, unstructured text or another format of data representation suitable to present the results.

5 | CONCLUDING REMARKS

The current central impact of LLMs is undeniable. The recency of their emergence also means that there is still much research to be done, with LLMs as the main focus. We have identified two main research directions: (1) the challenges and opportunities focused on improving the utility of LLMs, and (2) the challenges and opportunities that deal with how LLMs themselves can serve as a tool to support and advance CBR. CBR is a methodology fundamentally different from LLMs, because it works by storing and reusing past experiences and learning from experience to solve new problems. Due to their difference, both CBR and LLMs can complement each other, leveraging their respective strengths. In this work, we conducted an in-depth study of LLM and CBR synergies. Based on these two research directions, we have recognized four primary research areas: *Guided Generation*, *Accuracy & Dehallucination*, *Explainability & Trust*, and *Operational Efficiency*. From the literature review conducted and our own research experience, we have defined 13 main challenges and 17 research directions to overcome these challenges (see Table 1). Therefore, this work can serve as a guide for LLM researchers interested in using CBR as a complementary tool to overcome LLM weaknesses, and as a guide to the CBR community to employ LLMs to perform the 4R cycle, or for further application scenarios in the context of the CBR methodology.

TABLE 1 Open Challenges and Research Directions.

Open Challenges	Description	Research Directions and Opportunities
C1: Aligned Prompts	Creating good prompts is a highly knowledge-intensive task that depends on the user's experience. In addition, the quality of results depends on possible examples, which are provided to the model as part of few-shot prompts.	• O1.1: Case-Based LTM • O2.1: Episodic Memory for Prompts and Their Completions • O2.2: Prompt Engineering • O3.1: Case-Based Retrieval-Augmented Generation • O4.1: Prompt Engineering Training for Human Experts
C2: Context Window Limitation	LLMs have a limited context windows size in the form of possible tokens, which cannot handle more complex representations than text.	• O1.1: Case-Based LTM • O2.1: Episodic Memory for Prompts and Their Completions • O2.2: Prompt Engineering • O2.3: Filtering, Augmentation & Refinement • O3.1: Case-Based Retrieval-Augmented Generation • O5.2: Fine-Tuning of LLMs with Structured Cases • O5.4: Efficient Retrieval Mechanisms
C3: Domain and Adaptation Knowledge	LLMs do not solve individual issues, but re-generate new entire content that may introduce new errors in formerly correct solution parts.	• O2.3: Filtering, Augmentation & Refinement • O3.1: Case-Based RAG • O3.2: Dehallucination through Structured Knowledge • O5.3: Large Reasoning Models
C4: Validation Requirements	The output of LLMs is not deterministic, leading to issues such as retrieving different solutions for the same input or hallucinations.	• O2.3: Filtering, Augmentation & Refinement • O3.2: Dehallucination through Structured Knowledge
C5: Enhancing Relevance and Precision during Generation	LLMs do not retrieve the specific knowledge required when solving complex tasks. Instead, they retrieve general information.	• O1.1: Case-Based LTM • O3.1: Case-Based RAG
C6: Enhancing Interpretability and Trust in Generation and Retrieval	LLMs internal process and behavior is a black-box for their users, which limits trust and acceptance.	• O4.1: Prompt Engineering Training for Human Experts • O3.3: Evidence-Based and Fine-Grained Retrieval • O4.3: Human Alignment Opportunity
C7: Enriching Explanations	CBR-based explanation might not be as useful as expected since the examples provided might lack of details or descriptions.	• O4.2: Explainable Reasoning with Domain Knowledge • O4.3: Human Alignment Opportunity
C8: Building XAI Strategies as a Combination of Explanations	Users need to understand AI models from different points of view, which is not provided usually by XAI strategies, since they offer a unique explanation.	• O4.3: Human Alignment Opportunity • O6.3: LLMs to Improve CBR Outputs
C9: Computational Complexity and Size of LLMs	LLMs are large models that are training using a huge amount of data. This causes a high computational complexity.	• O1.1: Case-Based LTM • O5.1: Case-Based Distillation of Language Models • O5.4: Efficient Retrieval Mechanisms
C10: Knowledge Acquisition Bottleneck	One of the main challenges is acquiring knowledge in various forms, i.e., the adaptation knowledge, similarity measures, case base and vocabulary.	• O6.1: LLMs as Knowledge Engineers
C11: Limited Applicability of CBR Algorithms	The 4R cycle of CBR is tailored to a specific domain and use case, which causes the transfer to other domains a time-consuming and labor-intensive task.	• O6.2: LLMs in Various Parts of the CBR • O6.3: LLMs to Improve CBR Outputs

AUTHOR CONTRIBUTIONS

The authors are listed in alphabetical order, but all authors have contributed equally to this work.

References

1. Agnar Aamodt and Enric Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches . *AI Commun.*, 7(1):39–59, 1994.

2. Amirsiavosh Bashardoust, Yuanjun Feng, Dominique Geissler, et al. The Effect of Education in Prompt Engineering: Evidence from Journalists. *CoRR*, abs/2409.12320, 2024.

3. Ralph Bergmann, Hector Munoz-Avila, M. Veloso, et al. Case-based reasoning applied to planning tasks. In *Case-Based Reasoning Technology from Foundations to Applications*. Springer, 1998.

4. Florian Brand, Lukas Malburg, and Ralph Bergmann. Large Language Models as Knowledge Engineers. In *Proceedings of the Workshops at the 32nd International Conference on Case-Based Reasoning (ICCBR-WS 2024) co-located with the 32nd International Conference on Case-Based Reasoning (ICCBR 2024), Mérida, Mexico, July 1, 2024*, volume 3708 of *CEUR Workshop Proceedings*, pages 3–18. CEUR-WS.org, 2024.

5. Marta Caro-Martínez, Juan A. Recio-García, Belén Díaz-Agudo, et al. iSee: A case-based reasoning platform for the design of explanation experiences . *Knowl. Based Syst.*, 302:112305, 2024.

6. Quan Ze Chen and Amy X. Zhang. Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts . *CoRR*, abs/2310.07019, 2023.

7. Boda Feng, Hui Gao, Peng Zhang, et al. CBR-Ren: A Case-Based Reasoning Driven Retriever-Generator Model for Hybrid Long-Form Numerical Reasoning . In *Case-Based Reasoning Research and Development - 32nd International Conference, ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings*, volume 14775 of *Lecture Notes in Computer Science*, pages 111–126. Springer, 2024.

8. Lin Guan, Karthik Valmeekam, Sarath Sreedharan, et al. Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning . In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

9. Siyuan Guo, Cheng Deng, Ying Wen, et al. DS-Agent: Automated Data Science by Empowering Large Language Models with Case-Based Reasoning . In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

10. Kristian J. Hammond. Case-Based Planning: A Framework for Planning from Experience. *Cogn. Sci.*, 14(3):385–443, 1990.

11. Kathleen Hanney and Mark T. Keane. The Adaptation Knowledge Bottleneck: How to Ease It by Learning from Cases . In *Case-Based Reasoning Research and Development*, page 359370. Springer Berlin Heidelberg, 1997.

12. Sirui Hong, Mingchen Zhuge, Jonathan Chen, et al. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

13. Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, et al. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks . In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22895–22907. PMLR, 2024.

14. Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, et al. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *The Twelfth International Conference on Learning Representations*, oct 2023.

15. David Leake. CBR in context: The present and future. In *Case-Based Reasoning: Experiences, Lessons, and Future Directions*, pages 3 – 30. AAAI Press / MIT Press, 1996.

16. Mirko Lenz and Ralph Bergmann. Case-Based Adaptation of Argument Graphs with WordNet and Large Language Models . In *Case-Based Reasoning Research and Development - 31st International Conference, ICCBR 2023, Aberdeen, UK, July 17-20, 2023, Proceedings*, volume 14141 of *Lecture Notes in Computer Science*, pages 263–278. Springer, 2023.

17. Mirko Lenz, Lorik Dumani, Ralf Schenkel, et al. ArgServices: A Microservice-Based Architecture for Argumentation Machines . In *Robust Argumentation Machines - First International Conference, RATIO 2024, Bielefeld, Germany, June 5-7, 2024, Proceedings*, volume 14638 of *Lecture Notes in Computer Science*, pages 352–369. Springer, 2024.

18. Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

19. Yinghao Li, Rampi Ramprasad, and Chao Zhang. A Simple but Effective Approach to Improve Structured Language Model Output for Information Extraction . In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 5133–5148. Association for Computational Linguistics, 2024.

20. Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, et al. "We Need Structured Output": Towards User-centered Constraints on Large Language Model Output . In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9, Honolulu HI USA, may 2024. ACM.

21. Mirjam Minor and Eduard Kaucher. Retrieval Augmented Generation with LLMs for Explaining Business Process Models . In *Case-Based Reasoning Research and Development - 32nd International Conference, ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings*, volume 14775 of *Lecture Notes in Computer Science*, pages 175–190. Springer, 2024.

22. Mirjam Minor, Stefania Montani, and Juan A. Recio-García. Editorial: Process-oriented Case-based Reasoning. *Inf. Syst.*, 40:103–105, mar 2014.

23. Héctor Muñoz-Avila and Michael T. Cox. Case-Based Plan Adaptation: An Analysis and Review. *IEEE Intell. Syst.*, 23(4):75–81, 2008.

24. Mathis Pink, Qinyuan Wu, Vy Ai Vo, et al. Position: Episodic memory is the missing piece for long-term llm agents, 2025.

25. Enrique Queipo-de-Llano, Marius Ciurcau, Alejandro Paz-Olalla, et al. Explainable Classification System for Hip Fractures: A Hybrid CBR+LLM Surrogate Approach . In *Proceedings of the Workshops at the 32nd International Conference on Case-Based Reasoning (ICCBR-WS 2024) co-located with the 32nd International Conference on Case-Based Reasoning (ICCBR 2024), Mérida, Mexico, July 1, 2024*, volume 3708 of *CEUR Workshop Proceedings*, pages 65–80. CEUR-WS.org, 2024.

26. Michael M Richter. Knowledge Containers. In *Readings in Case-Based Reasoning*. Morgan Kaufmann Publishers, 2003.

27. Michael M. Richter and Rosina O. Weber. *Case-Based Reasoning - A Textbook*. Springer, 2013.

28. Nicholas Rossetti, Massimiliano Tummolo, Alfonso Emilio Gerevini, et al. Enhancing GPT-Based Planning Policies by Model-Based Plan Validation. In *Neural-Symbolic Learning and Reasoning - 18th International Conference, NeSy 2024, Barcelona, Spain, September 9-12, 2024, Proceedings, Part II*, volume 14980 of *Lecture Notes in Computer Science*, pages 328–337. Springer, 2024.

29. Roger C. Schank. *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge University Press, Cambridge, England, 1982.

30. Sander Schulhoff, Michael Ilie, Nishant Balepur, et al. The Prompt Report: A Systematic Survey of Prompting Techniques, jul 2024.

31. Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. On the Brittle Foundations of ReAct Prompting for Agentic Large Language Models . *CoRR*, abs/2405.13966, 2024.

32. Lei Wang, Chen Ma, Xueyang Feng, et al. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 18(6):186345, 2024.

33. Ian Watson. Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, 12(5):303–308, 1999.

34. Kaitlynne Wilkerson and David Leake. On Implementing Case-Based Reasoning with Large Language Models. In *Case-Based Reasoning Research and Development - 32nd International Conference, ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings*, volume 14775 of *Lecture Notes in Computer Science*, pages 404–417. Springer, 2024.

35. Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, et al. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering . In *Case-Based Reasoning Research and Development - 32nd International Conference, ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings*, volume 14775 of *Lecture Notes in Computer Science*, pages 445–460. Springer, 2024.