Can LLMs Reason with Case-Based Structure? Model-Dependent Effects of CBR-Inspired Decomposition on AI Risk Assessment

Gareth McConomy[1,*], Raymond Bond[1], David Glass[1], Jun Liu[1] and Holly Toner[1]

[1]School of Computing, Ulster University, Belfast, UK

Abstract

The CBR+LLM literature identifies numerous opportunities for synergy between Case-Based Reasoning and Large Language Models, yet lacks empirical evidence about whether CBR-style reasoning decomposition consistently improves LLM performance—and whether such benefits transfer across architectures. We investigate this question through 360 experimental runs across three frontier LLMs, ten deployment cases, and four prompting conditions: unstructured pattern access, Chain-of-Thought (CoT), CBR-inspired structured decomposition (condition matching, causal tracing, contextual calibration), and multi-agent deliberation. All conditions use the same knowledge graph of 87 causal risk patterns. CBR-inspired decomposition achieves 83% expert recall for Grok (Cohen's $d = 1.74$, $p < .001$), 30 percentage points above unstructured access and 27 above CoT, while CoT provides negligible benefit ($d = 0.24$). The same decomposition produces no improvement for GPT-4o ($p = .18$) and is architecturally incompatible with DeepSeek R1. Multi-agent deliberation consistently degrades performance. While model sensitivity is itself unsurprising in LLM research, the implication for memory-based reasoning is specific and actionable: a case-based decomposition that transforms one model's performance can be inert for another, so CBR+LLM scaffolding must be validated per target model rather than assumed to transfer. The structured decomposition also yields a per-condition, evidence-grounded audit trail—each enabling condition classified against quoted case evidence—directly serving the provenance and traceability goals of responsible GenAI. We discuss the formalisations needed to advance from CBR-inspired prompting to a rigorous CBR system, including similarity measures, case representation, and retention mechanisms.

Keywords

Case-based reasoning, Large language models, Knowledge graphs, AI risk assessment, Model-dependency, Structured reasoning

1. Introduction

The deployment of AI systems increasingly outpaces the capacity for expert risk review. Industry surveys indicate that only 10% of organisations have formal AI governance policies [1], while the EU AI Act mandates human oversight for high-risk applications [2]. This tension motivates scalable approaches to risk assessment that preserve expert judgement.

Case-Based Reasoning offers a principled framework for this challenge: new deployments can be assessed by retrieving structurally similar risk precedents, tracing their causal mechanisms in context, and adapting assessments based on amplifying and mitigating factors. The CBR+LLM research manifesto [3] identifies 17 opportunities for synergy, and recent work has demonstrated CBR-based retrieval for legal question answering [4] and data science automation [5]. However, this literature currently lacks empirical evidence addressing two fundamental questions: (1) does explicit CBR-style decomposition improve LLM performance beyond generic Chain-of-Thought prompting, and (2) do such scaffolding benefits transfer across LLM architectures?

We investigate these questions through a controlled experiment using a knowledge graph of 87 causal risk patterns applied to ten AI deployment cases across three frontier LLMs. Our approach adopts CBR-inspired decomposition as a prompting strategy—structuring LLM reasoning into condition matching, causal tracing, and contextual calibration phases. Beyond performance, this decomposition produces an

ICCBR Workshop on Memory-Based Reasoning for Responsible GenAI at ICCBR2026, August 15, 2026, Bremen, Germany
*Corresponding author.
mcconomy-g@ulster.ac.uk (G. McConomy)

auditable reasoning trace—each enabling condition classified PRESENT/PARTIAL/ABSENT against quoted case evidence—rather than an opaque holistic judgement. This property is central to the responsible-GenAI motivation for memory-based reasoning: the value of grounding an LLM in an external case memory lies not only in improved recall but in the provenance and contestability of the resulting assessment. This is not yet a formal CBR system: it lacks defined similarity measures, structured case representations in the problem/solution/outcome sense, and automated retention mechanisms. However, the workflow shares significant structural parallels with the CBR cycle, and the empirical findings reported here are intended to motivate and inform the formalisations needed to build a proper CBR system—a direction we are actively pursuing (Section 5).

Our contributions to the CBR+LLM research agenda are: (1) controlled evidence that reasoning decomposition along CBR-inspired lines can produce very large improvements in LLM performance ($d = 1.74$) where generic CoT does not ($d = 0.24$), isolating reasoning structure from knowledge content; (2) evidence that these benefits are strongly model-dependent within the CBR+LLM setting, motivating per-model validation of any case-based scaffold rather than an assumption of transfer; (3) a concrete roadmap identifying the formalisations needed to advance from CBR-inspired prompting to a rigorous CBR system, informed by where the current approach succeeds and where it falls short.

2. Related Work

2.1. CBR+LLM Synergies

The intersection of CBR and LLMs is an active research area. Bach et al. [3] identify 17 research opportunities spanning case acquisition, retrieval, adaptation, and retention. Wiratunga et al. [4] developed CBR-RAG for legal question answering, demonstrating that case-based retrieval outperforms semantic similarity for precedent identification—a finding directly relevant to our use of structured pattern matching over embedding-based retrieval. Guo et al. [5] achieved strong results on data science tasks using the full 4R cycle, and Hatalis et al. [6] provide a theoretical framework for CBR-augmented LLM agents.

A foundational question is whether LLMs perform genuine analogical reasoning. Webb et al. [7] demonstrated surprisingly strong abstract pattern induction, while Hodel and West [8] showed failures on task variations suggesting memorisation. This tension motivates our empirical investigation: does making case-based structure explicit improve what LLMs do implicitly?

2.2. Structured Reasoning and Multi-Agent Approaches

Chain-of-Thought prompting improves multi-step reasoning by eliciting intermediate steps [9] but provides generic encouragement without domain-specific decomposition. Multi-agent deliberation has shown gains on mathematical reasoning [10], but Choi et al. [11] proved a martingale property showing debate does not improve expected correctness beyond ensembling, and Huang et al. [12] established that LLMs cannot self-correct without external feedback.

2.3. Structural Parallels and Formalisations Needed

Our approach shares significant structural parallels with CBR. The knowledge graph of 87 causal risk patterns functions as a case base of abstracted cases [13]—generalisations derived from multiple incidents that preserve causal structure. The LLM's condition-matching against deployment cases performs a function analogous to similarity-based retrieval, guided by what can be adapted rather than surface similarity [14]. Causal tracing applies retrieved knowledge to a new context, as in Reuse. Contextual calibration adapts assessments based on amplifying and mitigating factors, as in adaptation. Expert validation serves a Retain-like function.

However, we do not yet implement several components the CBR community considers essential for a formal CBR system. No similarity measure is defined over the case space—our condition-matching relies

on LLM assessment rather than a principled similarity function. Recent work on automatic adjustment of global similarity measures [15] demonstrates the level of rigour expected. Our patterns lack the explicit problem-descriptor/solution/outcome triple that defines a CBR case. Furthermore, Wilkerson et al. [16] identify hallucination risks in CBR+LLM systems that our evaluation does not specifically address, though our expert review provides partial coverage. These gaps define a concrete formalisation agenda (Section 5), and the empirical findings reported here are intended to inform which formalisations matter most.

3. Method

3.1. Knowledge Graph of Causal Risk Patterns

Our pattern library comprises 87 risk patterns constructed through systematic review of AI safety literature, including the MIT AI Risk Repository [17], AI Incident Database [18], and NIST AI RMF [19]. Each pattern encodes:

$$\text{Pattern } p = \langle EC(p), CM(p), H(p), Ev(p) \rangle \tag{1}$$

where $EC(p)$ is the set of enabling conditions, $CM(p)$ the causal mechanism, $H(p)$ potential harms, and $Ev(p)$ supporting evidence. The knowledge graph contains 1,310 nodes and 1,299 edges across seven risk categories.

To illustrate, Pattern P70 (*Proxy Disparate Impact*) is representative of the discrimination/bias category. Its enabling conditions are: (i) ML features correlated with protected characteristics; (ii) training data reflecting historical disparities; (iii) no fairness constraints applied. The causal mechanism encodes the chain: model learns statistical correlations → predictions systematically disadvantage groups → outcomes diverge along protected attributes. Harms include reduced approval rates for minorities and legal liability under fair-housing law; evidence cites Obermeyer et al. [20] on cost-as-proxy discrimination in healthcare. Patterns typically have 3–5 enabling conditions and 2–4 causal steps. As discussed in Section 2.3, these are domain-general abstracted cases [13] that preserve causal structure; formalising them with explicit problem-descriptor/solution/outcome attributes is planned.

3.2. Deployment Cases

We developed ten AI deployment cases in real estate, written as Harvard Business School-style narratives (1,000–1,500 words each) incorporating system specifications, stakeholder conflicts, regulatory context, and unresolved decision points. Real estate provides a high-stakes testbed with diverse AI applications and emerging regulatory scrutiny. The single-domain design follows Yin [21] on depth over breadth; cross-domain validation is needed.

3.3. Experimental Conditions

We compare four conditions, all receiving the same 87-pattern knowledge graph:

C1: Unstructured. The LLM receives the full pattern library and selects and ranks the top-10 patterns with no reasoning structure imposed.

C2: Chain-of-Thought. The same library with a CoT prompt: "Think step by step. For each pattern, reason about whether its conditions apply to this case. Trace the causal mechanism. Explain your ranking before producing the final output."

C3: CBR-Inspired Structured Decomposition. A two-stage pipeline using the same target model for both stages. Stage 1: the LLM assesses each pattern's enabling conditions against case features, outputting 15 candidates with match scores. Stage 2: the LLM produces structured analysis with three phases:

- *Condition matching*: every enabling condition of every candidate classified PRESENT/PARTIAL/ABSENT with quoted case evidence;

- *Causal tracing*: each causal chain step mapped from abstract form to the deployment context;
- *Contextual calibration*: amplifying and mitigating factors enumerated separately, producing a calibrated applicability score (0–1).

This enforced decomposition means the LLM cannot produce a holistic judgement without completing each phase. The same-model design ensures performance differences reflect decomposition structure, not model capability.

We use phase names (condition matching, causal tracing, contextual calibration) that describe what the LLM actually does, rather than mapping directly to the classical Retrieve-Reuse-Revise cycle. The relationship between these phases and CBR is discussed in Section 5.

C4: Multi-Agent Deliberation. Four sequential prompts: Risk Identifier nominates patterns; Sceptic challenges each; Identifier responds; Synthesiser produces the final ranking.

3.4. Models

Three frontier LLMs: GPT-4o (OpenAI), Grok (xAI), and DeepSeek R1 [22], a reasoning model with RL-trained chain-of-thought. All accessed via API at temperature 0.3.

3.5. Expert Evaluation

In applied risk assessment, no objective ground truth exists independent of expert judgement; the relevant standard is whether a qualified professional would act on the identified risks. The expert benchmark was established by the lead author (FRICS-qualified, 20+ years real estate advisory, PhD researcher in human-centric AI), who independently identified the top-5 most applicable patterns **before** viewing any LLM outputs. Pre-registration locked this benchmark prior to experiment execution. The expert subsequently reviewed all outputs, judging each pattern as Accept, Partial, or Reject with a written rationale.

The single-expert design is deliberate: the dual expertise required (professional domain qualification plus AI risk competence) is rare, and under-qualified additional raters would introduce a different validity threat rather than resolving one. We acknowledge directly that the expert is also a co-author of the workflow under evaluation; this dual role is a genuine limitation rather than an oversight. Three features constrain its effect on the reported claims. First, the benchmark was pre-registered before any LLM output was seen, so the expert could not tune the ground truth to favour a condition. Second, the design is within-subjects and the central claims are *relative* (C1 vs. C3 under an identical benchmark), so any systematic expert perspective cancels across conditions rather than inflating one. Third, all judgements carry documented per-pattern rationales, enabling independent audit. Absolute claims (e.g. the 81% accept rate) should be read as one qualified professional's assessment. A multi-expert study with cross-functional practitioners, designed to separate the system author from the evaluating experts, is planned for autumn 2026.

3.6. Metrics and Design

Expert Recall@K: |Expert Top-5 ∩ LLM Top-K|/5. *Accept Rate*: proportion judged Accept or Partial. *Discovery Rate*: patterns the expert adds upon review. We conducted 360 runs: 10 cases × 3 models × 4 conditions × 3 replications. We report means, standard deviations, Cohen's d, and Mann-Whitney U tests with Bonferroni correction ($\alpha = 0.0125$).

4. Results

Table 1 presents Expert Recall@10 disaggregated by model and condition.

Table 1
Expert Recall@10 by Model × Condition (mean ± SD, $n = 30$ per cell). Best result per model in bold.

Model	C1 Unstr	C2 CoT	C3 Struct	C4 Delib
GPT-4o	.50 ± .15	.44 ± .29	.45 ± .14	.47 ± .12
Grok	.53 ± .14	.57 ± .18	**.83 ± .20**	.51 ± .17
DeepSeek R1	.55 ± .19	.47 ± .23	FAIL	.48 ± .18

CBR-inspired decomposition produces large, model-dependent effects. For Grok, structured decomposition achieves 83% recall versus 57% for CoT, 53% for unstructured, and 51% for deliberation. The Grok C1→C3 improvement yields Cohen's $d = 1.74$ ($U = 121.5$, $p < .001$) and the C2→C3 improvement yields $d = 1.38$ ($U = 160.0$, $p < .001$). Both comparisons are significant after Bonferroni correction.

Generic CoT provides negligible benefit. CoT yields only +4pp over unstructured for Grok ($d = 0.24$) and *reduces* performance for GPT-4o (−6pp). This establishes that the C3 advantage is domain-specific decomposition, not a general scaffolding effect.

Model-dependency. GPT-4o shows no benefit from any structured prompting. Its highest mean (50%, C1) occurs without scaffolding ($U = 532.0$, $p = .18$ for C1→C3). DeepSeek R1 fails completely under C3: all 30 runs return zero extractable patterns because its verbose RL-trained reasoning [22] exhausts tokens before producing the required structured output. This is an output-format incompatibility rather than a reasoning failure—C1 succeeds because its simpler format fits within limits.

Deliberation degrades performance. C4 performs comparably to or worse than unstructured access across all models. Transcript analysis reveals the mechanism: the Sceptic agent demands explicit textual evidence, filtering structurally valid patterns whose enabling conditions require inference rather than direct textual support. This aligns with the martingale property [11].

Expert review confirms reasoning quality. Of 100 Grok C3 patterns assessed: 59% Accept, 22% Partial, 19% Reject. The 81% Accept+Partial rate indicates the large majority of identified patterns are at least directionally useful. The LLM also surfaced patterns beyond the expert's pre-registered top-5 that the expert subsequently judged valid—notably P76 (Endogenous Feedback) in 6/10 cases and P45 (Procurement Gap) in 4/10 cases—demonstrating complementary discovery.

5. Discussion: Implications for CBR+LLM Research

5.1. The Model-Dependency Question

The finding with the clearest implication for CBR+LLM system design is the non-transfer of decomposition benefits across models. The CBR+LLM literature [3, 6] currently lacks empirical evidence about whether scaffolding benefits generalise. Our results show that structured reasoning decomposition inspired by CBR principles does not transfer across architectures: identical decomposition produces $d = 1.74$ for Grok and $d = -0.38$ for GPT-4o. This suggests that formal CBR+LLM systems will face the same challenge and should be validated per target model.

We frame this cautiously. Model sensitivity is well established in broader LLM research—prompting effects, instruction following, and output-format compliance vary substantially across architectures. Our contribution is narrower but specific to the CBR+LLM intersection: if the community intends to build CBR-augmented LLM systems, scaffolding effects must be validated per target model. The theoretical question remains open: what properties of an LLM architecture determine whether explicit case-based decomposition adds value?

Several hypotheses merit investigation: (a) GPT-4o may already perform implicit CBR-like reasoning, making explicit structure redundant; (b) architectural differences in phase responsiveness; (c) differential instruction compliance. Hypothesis (c) alone is insufficient: if Grok simply followed instructions more faithfully, C4 should also improve, yet it degrades performance. The benefit is specific to case-based decomposition structure.

5.2. Knowledge Content versus Reasoning Structure

The +30pp C1→C3 improvement—both conditions using identical knowledge graph content—demonstrates that the value lies not just in *what* knowledge is provided but in *how* the model is required to reason with it. This separation of knowledge content from reasoning structure is directly relevant to memory-based reasoning for responsible GenAI: the pattern library functions as an external memory encoding documented risk precedents, and the structured decomposition protocol governs how an LLM retrieves from and reasons with that memory. The finding that protocol design matters as much as memory content has direct implications for any memory-augmented LLM system intended for responsible AI deployment.

A second consequence concerns auditability. Because C3 requires the model to classify each enabling condition against quoted case evidence before producing a judgement, its output is a structured trace linking every retrieved pattern to the specific case features that justify it. This is a form of case-level provenance: an assessment can be inspected, challenged, and revised at the level of individual conditions rather than accepted or rejected wholesale. Unstructured (C1) and CoT (C2) outputs do not expose this structure, and deliberation (C4) discards inference-supported patterns precisely because they lack direct textual evidence. For responsible GenAI, where contestability and traceability are first-order requirements, the reasoning structure is therefore not only a performance lever but a transparency mechanism—a property that the recall metric alone does not capture.

5.3. From CBR-Inspired Prompting to a CBR System

Our current approach adopts the conceptual structure of CBR without implementing its formal machinery. We identify three specific formalisations needed to bridge this gap:

Similarity measures. Stage 1 of our pipeline relies on LLM assessment of enabling-condition presence—an informal, unweighted matching process. A formal CBR system would define a similarity function over the case space, potentially incorporating weighted condition matching, partial-match aggregation, and learned similarity adjustments [15], with tools for comprehension and maintenance of the resulting similarity model [23]. The knowledge graph's typed enabling conditions provide a natural basis for this formalisation.

Case representation. Our 87 patterns are abstracted generalisations [13], not experiential episodes. A full CBR system would define cases with problem-descriptor attributes (deployment context features), solution components (applicable risk patterns with severity assessments), and outcome attributes (expert validation results). The 360 experimental runs themselves could seed such a case base, with each assessed deployment becoming a retained case.

Retain operationalisation. The Retain phase is examined in complementary work investigating expert-in-the-loop iterative calibration across successive assessment rounds. That study characterises how expert feedback propagates through the system using a five-type error taxonomy, and finds that overcorrection—the expert and system over-adjusting in response to each other—emerges as the dominant error category, overtaking initial retrieval errors by the second round. This indicates that a naive Retain loop can degrade rather than improve assessments, and that retention requires explicit guards against co-evolutionary drift. The present study focuses on the LLM-executed phases (condition matching, causal tracing, contextual calibration) that precede expert Retain. Integrating findings from both lines—connecting the structured retrieval and reasoning pipeline reported here with the iterative expert calibration studied separately—is a priority for the full CBR system.

5.4. Relationship to CBR: Current State and Direction of Travel

Our three decomposition phases (condition matching, causal tracing, contextual calibration) are structurally parallel to, but do not formally implement, the classical Retrieve-Reuse-Revise cycle [24]. Condition matching performs a function analogous to Retrieve, assessing structural similarity between patterns and the target case—but via LLM judgement rather than a defined similarity function. Causal tracing corresponds most closely to Reuse, applying the retrieved pattern's causal mechanism to the new context. Contextual calibration incorporates elements of both classical Reuse (adapting the solution) and Revise (evaluating the proposed solution against contextual constraints), as it simultaneously adapts severity assessments and evaluates their appropriateness.

We view the current work as occupying a specific position on a trajectory toward formal CBR. The empirical findings establish *that* CBR-like decomposition produces large benefits for LLM reasoning on this task. The next step is to establish *how* formalising the decomposition as proper CBR—with similarity measures that can be learned and adjusted [15], case representations that support systematic retrieval, and Retain mechanisms informed by our complementary work on expert calibration—changes those benefits. The model-dependency finding (Section 5.1) suggests this formalisation matters: if different LLMs respond differently to informal decomposition, formal CBR machinery may provide more consistent benefits by reducing dependence on the LLM's implicit reasoning capabilities. Whether the optimal decomposition aligns with the classical 4R cycle or requires a domain-specific variant is itself a research question we intend to address.

6. Limitations and Future Work

Single expert. Relative comparisons (C1 vs C3) are unaffected by systematic expert perspective due to pre-registration and within-subjects design. Absolute claims (81% accept rate) reflect one professional's assessment; documented rationales enable independent audit. A multi-expert study with cross-functional practitioners is planned for autumn 2026.

Single domain. All cases are in real estate. Cross-domain validation (healthcare, finance) is needed to assess generalisability.

No formal CBR components. As discussed in Section 5, the approach lacks defined similarity measures and structured case representations. Retain is examined in complementary work; integrating both lines of investigation into a formal CBR system is the primary direction for future work.

No vanilla baseline. All conditions receive knowledge graph patterns. We test reasoning structure given pattern access, not whether pattern access itself helps.

DeepSeek failure. The C3 failure reflects output-format incompatibility rather than a reasoning limitation. Future work should test with increased token limits or truncated reasoning traces to isolate the architectural constraint.

Ecological validity. Cases are expert-authored scenarios. While designed to capture real-world complexity, actual deployment documentation may be messier and less structured.

Data and Code Availability

The 87-pattern knowledge graph, case narratives, prompt templates, raw results (360 runs), expert ground truth, and analysis scripts are available from the authors on request and will be released in a public repository.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: assist with literature synthesis during knowledge graph construction, drafting and language editing of the manuscript, and case study development. Generative AI tools were excluded from the experiments themselves; the

evaluated systems were GPT-4o, Grok, and DeepSeek R1. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] ISACA, Generative AI Survey: Key Takeaways, Technical Report, ISACA, 2023.

[2] European Parliament, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act), Official Journal of the EU, L series, 2024.

[3] K. Bach, R. Bergmann, F. Brand, et al., Case-based reasoning meets large language models: A research manifesto, HAL preprint hal-05006761v1, 2025.

[4] N. Wiratunga, et al., CBR-RAG: Case-based reasoning for retrieval augmented generation, in: Proceedings of ICCBR, volume 14775 of *LNCS*, 2024, pp. 445–460.

[5] S. Guo, et al., DS-Agent: Automated data science by empowering LLMs with case-based reasoning, in: Proceedings of ICML, volume 235, 2024, pp. 16813–16848.

[6] K. Hatalis, D. Christou, V. Kondapalli, Review of case-based reasoning for LLM agents, arXiv:2504.06943, 2025.

[7] T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models, Nature Human Behaviour 7 (2023) 1526–1541.

[8] C. M. Hodel, J. West, Response: Emergent analogical reasoning in large language models, arXiv:2308.16118, 2023.

[9] J. Wei, et al., Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of NeurIPS, 2022.

[10] Y. Du, S. Li, A. Torralba, et al., Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of ICML, 2024.

[11] H. K. Choi, X. Zhu, S. Li, Debate or vote: Which yields better decisions in multi-agent LLMs?, in: Proceedings of NeurIPS, 2025.

[12] J. Huang, et al., Large language models cannot self-correct reasoning yet, in: Proceedings of ICLR, 2024.

[13] R. Bergmann, W. Wilke, Building and refining abstract planning cases by change of representation language, Journal of Artificial Intelligence Research 9 (1998) 1–64.

[14] B. Smyth, M. T. Keane, Adaptation-guided retrieval: Questioning the similarity assumption in reasoning, Artificial Intelligence 102 (1998) 249–293.

[15] A. Ottersen, K. Bach, Automatic adjusting global similarity measures in learning CBR systems, in: Proceedings of ICCBR, 2024.

[16] K. Wilkerson, et al., Case hallucinations and steps toward repair, in: ICCBR 2025 Workshop Proceedings, 2025.

[17] P. Slattery, et al., The AI risk repository, arXiv:2408.12622, 2024.

[18] S. McGregor, Preventing repeated real world AI failures by cataloging incidents, in: Proceedings of AAAI, volume 35, 2021, pp. 15458–15463.

[19] E. Tabassi, AI Risk Management Framework (AI RMF 1.0), Technical Report NIST AI 100-1, NIST, 2023.

[20] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453.

[21] R. K. Yin, Case Study Research: Design and Methods, 4th ed., Sage, 2009.

[22] DeepSeek-AI, et al., DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, arXiv:2501.12948, 2025.

[23] G. Jimenez-Diaz, B. Díaz-Agudo, Visualization of similarity models for CBR comprehension and maintenance, in: Proceedings of ICCBR, 2024.

[24] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI Communications 7 (1994) 39–59.