# Retrieval-Augmented Case-Based Reasoning for Trustworthy Document Understanding at Scale

Lasal Jayawardena[1,*,†]

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

**Abstract**

This doctoral research investigates how Case-Based Reasoning (CBR) principles can wrap modern Large Language Model (LLM) pipelines to deliver calibrated, grounded and revisable predictions for industrial document understanding at scale. The PhD began with CBR combined with Retrieval-Augmented Generation (RAG) and Genetic Algorithm optimisation for resource allocation in waste management, but quickly converged onto a more fundamental bottleneck observed at the industry partner WM Nicol: organisational insights are locked inside heterogeneous, unstructured legacy documents. To unlock them, a human-in-the-loop Intelligent Document Processing (IDP) platform was built, in which multi-label classification acts as the routing gate for downstream extraction. Three convergent research contributions follow. (1) An LLM-RAG clarification mechanism for the iSee XAI platform with a Reasonability metric and Markov-chain revision-knowledge discovery (ICCBR 2025). (2) *RAPT*, a retrieval-augmented post-hoc thresholding wrapper that calibrates label-set selection for metric-learning and transformer backbones without retraining. (3) A CBR adaptation and Self-Reflection Prompting extension that transfers *RAPT*'s calibration principles to prompt-elicited LLM scores. Together, these works establish CBR as a lightweight wrapper layer for trustworthy LLM-based document understanding across waste-management, energy and adjacent regulated sectors processing hundreds of thousands of documents.

**Keywords**

Case-Based Reasoning, Large Language Models, Retrieval-Augmented Generation, Multi-Label Classification, Industrial Document Processing, Explanation Revision

## 1. Introduction

Decision-support systems in regulated industrial sectors such as waste management and energy must integrate decades of operational knowledge with real-time data, yet most of that knowledge sits inside heterogeneous, unstructured documents (waste transfer notes, hazardous waste prenotifications, weighbridge tickets, purchase orders, cleanliness certificates, email correspondence). At the industry partner WM Nicol, this document backlog is the single largest bottleneck preventing downstream resource allocation, scheduling and compliance analytics from operating at scale.

This doctoral research began with the ambition of combining CBR, LLM-based RAG [1] and Genetic-Algorithm optimisation to schedule fleet and engineering resources at WM Nicol. In the first year, attempts to instantiate this framework collided with a data-readiness problem: no upstream pipeline could reliably turn the document estate into the structured attributes the downstream optimiser needed. The trajectory therefore converged onto building an *Intelligent Document Processing* (IDP) platform [2], with multi-label classification as the decision gate that routes each document to its appropriate downstream extractor. This pivot kept the original CBR-LLM thesis intact, but reframed it around a problem with much higher operational leverage and broader applicability across waste-management and energy sectors.

In parallel, work on the *iSee* XAI platform [3] contributed a complementary CBR-LLM strand: detecting when end-user explanation intents evolve during an interaction, so that the CBR revise step can update the recommended explanation strategy. Both strands share the same research question,

*ICCBR DC'26: Doctoral Consortium at ICCBR2026, August 16, 2026, Bremen, Germany*

*Corresponding author.

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

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

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

restated as: *How can CBR principles (retrieval, reuse, adaptation, revision) wrap modern LLM pipelines to deliver calibrated, grounded and revisable predictions for industrial document understanding at scale?*

The PhD answers this question through three convergent contributions (Section 3): an LLM-RAG clarification and revision-knowledge mechanism for iSee [4]; the *RAPT* post-hoc thresholding wrapper for multi-label classifiers [5]; and a CBR adaptation plus Self-Reflection Prompting extension that transfers these calibration principles to LLM-based multi-label classifiers. The methodological foundation across all three strands is the *CBR-RAG* [6] formulation, which established that the CBR vocabulary, case-base and similarity knowledge containers can be reused as the retrieval scaffolding of LLM systems.

## 2. Background and Related Work

**CBR and LLM integration.** The classical CBR cycle of retrieve, reuse, revise and retain [7] provides a natural scaffold for wrapping LLMs. A recent research manifesto [8] identifies prompt-based adaptation accuracy, output validation and the absence of robust external retrieval mechanisms as key open challenges for hybrid CBR-LLM systems. Wilkerson and Leake [9] report that direct LLM-based retrieval matches external $k$-NN selection only 20–28% of the time and that LLMs hallucinate case attributes, motivating designs in which CBR contributes verifiable retrieval and adaptation infrastructure rather than asking the LLM to do everything.

**Multi-label thresholding.** Converting per-label scores into a final label set is the operational decision in multi-label classification [10]. Static label-wise strategies such as SCut [11] and gap-based cutoffs such as MCut [12] improve on a fixed global threshold but remain static at test time, motivating instance-adaptive alternatives. LLMs in multi-label settings additionally exhibit token distributions concentrated on a single label as model scale increases [13], making prompt-elicited scores particularly hard to threshold.

**LLM calibration and revision.** LLM responses can be hallucinated, miscalibrated or inconsistent [14], and Self-Refine [15] shows iterative self-feedback can improve outputs. However, a critical survey of self-correction in LLMs [16] finds that purely internal feedback often introduces self-bias and can degrade outputs; methods such as CRITIC [17] are more reliable when feedback is grounded in external evidence. LLM-as-judge methodologies [18] provide one external signal, and CBR neighbourhoods provide another – which this PhD exploits.

## 3. Research Aims and Methodology

The three contributions below share a common design pattern: a labelled case base is constructed from existing training (and verified) data, neighbours of the query are retrieved through an embedding similarity, and their solutions or score residuals are reused and adapted to produce a calibrated or revised output for the query. Each contribution instantiates this pattern at a different layer of an LLM-driven document understanding stack.

### 3.1. Contribution 1 – LLM-RAG Clarification and Revision Knowledge in iSee

In the iSee XAI platform, explanation strategies are encoded as Behaviour Trees, and end-user feedback drives the CBR revise step. We introduced a *Clarification Node* that triggers an agentic LLM-RAG workflow when the user signals uncertainty, and a *Reasonability Score* that aggregates judgments from over 30 LLM judges to quantify how coherent and context-aligned a clarification is. A Markov chain over a set of 16 user intents (e.g., transparency, debugging, comprehensibility) is fitted from session traces and used to detect when the user's explanation intent has diverged from the originally designed strategy – a trigger for revision. This work was published at ICCBR 2025 [4] and validated through experiments with end users across four real-world use cases spanning telecommunications, medical imaging, manufacturing inspection and loan-approval explanation. The interpretability question it answers – *when does an explanation strategy need to be revised?* – is the same instance-adaptive question

that recurs in the document-understanding strand under the guise of *when does a label set need to be revised?*

### 3.2. Contribution 2 – RAPT: Retrieval-Augmented Post-hoc Thresholding

The IDP platform's bottleneck is not learning a good per-label ranking but turning that ranking into a label set. In multi-label document classification, per-class optimal thresholds can range from below 0.1 to above 0.7, yet a fixed global threshold of $\tau = 0.5$ is the most common default. *RAPT* addresses this through a model-agnostic CBR wrapper [5]: a case base is built from training-set predictions and ground-truth labels, the query's neighbours are retrieved through an embedding index, and the query's label set is computed via *dynamic thresholding*. Two strategies were introduced: *AvgCount*, which estimates the expected label cardinality from similarity-weighted neighbour counts, and *RankCal*, which infers a confidence threshold from the minimum true-label scores observed in the neighbourhood. A *score adaptation* step transfers per-neighbour residuals ($\Delta_1$ alignment, $\Delta_2$ residual correction) to the query's raw scores before thresholding. *RAPT* is model-agnostic: it wraps episodic metric learners (Matching Networks, Prototypical Networks, Relation Networks, SNAIL) and fine-tuned transformer encoders (BERT, DeBERTa-v3) without retraining. On the WM Nicol industrial dataset it outperfomed few-shot LLM baselines by approximately $\geq$ 115× less inference time and 13.5× less GPU memory. Across 175 model-dataset combinations on six public benchmarks plus the industrial dataset, *RAPT* beat the best static thresholding baseline in 74.3% of cases.

### 3.3. Contribution 3 – CBR Adaptation and Self-Reflection for LLM Multi-Label Classifiers

*RAPT*'s case-based score adaptation transfers naturally to LLMs that emit per-label scores through prompting, but LLM scores are not calibrated: fine-tuning is rarely cost-effective at scale and token probabilities are a poor confidence proxy. We therefore elicit scores through structured prompts (percentage, 5-point Likert, 10-point numerical, binary) and apply two complementary CBR. *Score Adaptation* adjusts prompt-elicited LLM scores using ground-truth residuals from retrieved neighbours, requiring no additional LLM calls. *Self-Reflection Prompting* feeds aggregated neighbourhood statistics – expected label cardinality $\bar{n}_q$, rank threshold $\tau_q^{\text{rank}}$, and similarity-weighted label-frequency distribution $f_q$ – back to the LLM as external evidence in an iterative generate-refine cycle. This instantiates the CBR *revise* step inside the LLM loop, addressing the self-bias problem identified in [16] by replacing internal critique with external neighbourhood evidence. On the WM Nicol dataset, the best Self-Reflection configuration with score adaptation reaches 0.803 Micro-F1 and 0.737 Macro-F1, surpassing strong few-shot baselines [19, 20, 21] at a much lower token budget. The robustness holds under label anonymisation, which removes semantic label names and therefore simulates proprietary industrial taxonomies.

## 4. Industrial Use Case: WM Nicol IDP Platform

The IDP platform deployed at WM Nicol is a human-in-the-loop system in which incoming PDFs are converted, OCR'd, classified across 10+ document types (Hazardous-waste notes, waste transfer notes, weighbridge tickets, cleanliness certificates, purchase orders, delivery notes, job tickets, invoices, email correspondence , SEPA notes and etc.) and routed to document-class-specific extractors. Reviewers verify or override the extracted fields through the interface shown in Figure 1. Verified documents become new cases for the *RAPT* and Self-Reflection case bases, closing the CBR retain loop in production.

The pain point at WM Nicol is shared by the wider waste-management and energy sectors, where compliance-driven workflows produce hundreds of thousands of documents annually in heterogeneous templates that evolve over time. The CBR wrapper layer is attractive in this setting precisely because it adapts to drift without retraining the underlying classifier or LLM: case-base growth is the adaptation mechanism.

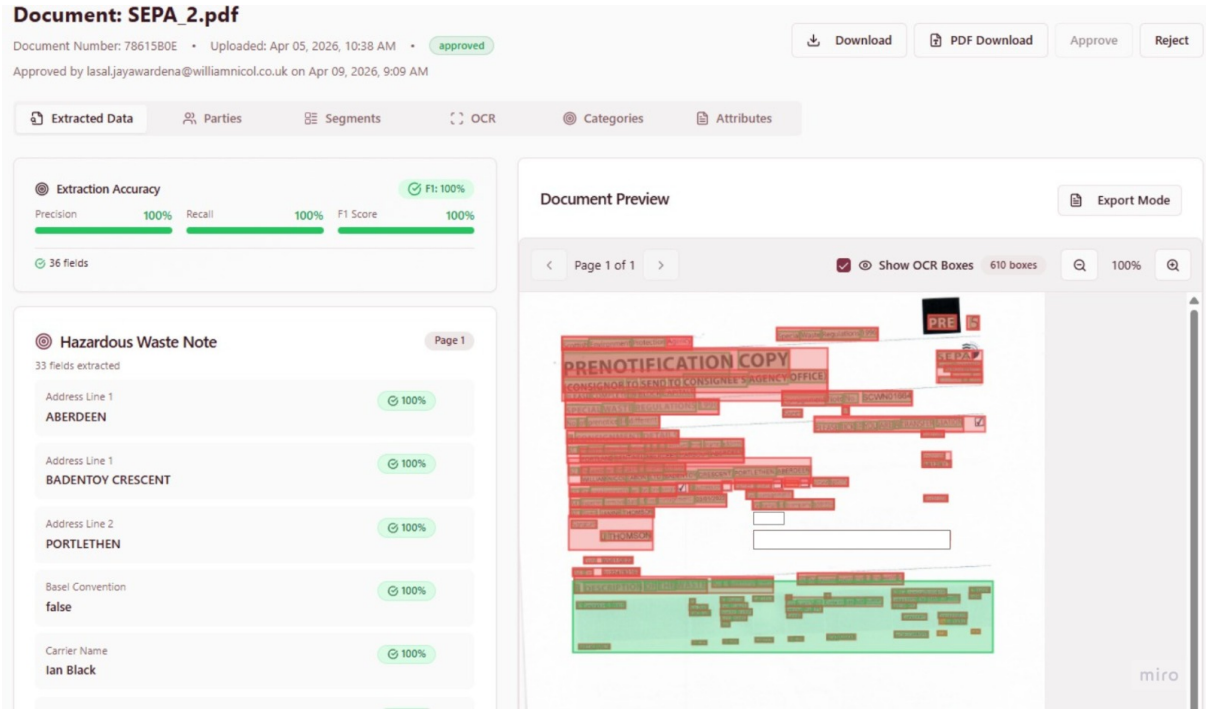


**Figure 1:** The WM Nicol HITL Intelligent Document Processing interface reviewing a SEPA hazardous-waste prenotification. Multi-label classification (top, document categories) is the routing gate for the downstream class-specific extractors; per-field extraction (left, with confidence) and the OCR overlay drive the reviewer's verify-or-override decision. Each approved document becomes a new case that grows the *RAPT* and Self-Reflection case bases.

## 5. Conclusion and Future Work

The PhD has converged on: CBR as a lightweight wrapper layer for calibrating, grounding and revising LLM outputs in industrial document understanding. The remaining work has three threads. First, the CBR *retain* step in production, as verified documents flow back from the IDP, the case base should grow online, and we will study case-base maintenance and active case-selection policies that bound the labelling budget while tracking template drift. Second, extreme multi-label settings: *RAPT* and Self-Reflection have been evaluated on label spaces of up to a hundred classes, but legal (EUR-LEX) and biomedical (MIMIC-III, OHSUMED) corpora will stress the neighbourhood-driven design at scale. Third, and arguably the most important for the IDP setting, is extending the wrapper from text-only LLMs to *Vision-Language Models* (VLMs). Industrial documents are inherently visual: stamps, signatures, table structure, hand-annotated overrides and form layout all carry information that current OCR-then-text pipelines flatten away. We will test whether the *RAPT* score-adaptation and Self-Reflection mechanisms carry over to VLM-elicited per-label scores, where the case-base embedding is now multimodal and neighbourhood evidence may include visual rather than purely textual cues, and use this as a vehicle to find better end-to-end methods for IDP that bypass brittle intermediate text extraction.

### Declaration on Generative AI

During the preparation of this manuscript, the authors used generative AI tools for assistance with writing refinements. All intellectual content, research design, methodology, experiments, data analysis, and interpretation were conducted solely by the authors, who reviewed and approved the final manuscript. The authors take full responsibility for the content of this publication.

## References


[1] P. Lewis, E. Perez, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proc. of the 34th Int. Conf. on Neural Information Processing Systems, Red Hook, NY, USA, 2020.

[2] M. Ahmadi Achachlouei, O. Patil, T. Joshi, et al., Document automation architectures: Updated survey in light of large language models, 2023. URL: https://arxiv.org/abs/2308.09341.

[3] M. Caro-Martínez, J. A. Recio-García, B. Díaz-Agudo, et al., iSee: A case-based reasoning platform for the design of explanation experiences, Knowledge-Based Systems 302 (2024) 112305. doi:10.1016/j.knosys.2024.112305.

[4] L. Jayawardena, A. Liret, et al., Context driven multi-query resolution using LLM-RAG to support the revision of explainability needs, in: Case-Based Reasoning Research and Development, Springer Nature Switzerland, Cham, 2025, pp. 111–125. doi:10.1007/978-3-031-96559-3_8.

[5] L. Jayawardena, N. Wiratunga, I. Nkisi-Orji, et al., RAPT: Retrieval-augmented post-hoc thresholding for multi-label classification, 2026. URL: https://arxiv.org/abs/2605.16535.

[6] N. Wiratunga, R. Abeyratne, et al., CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering, in: CBR and Development, volume 14775, Springer Nature Switzerland, Cham, 2024, pp. 445–460. doi:10.1007/978-3-031-63646-2_29.

[7] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI Communications 7 (1994) 39–59. doi:10.3233/AIC-1994-7104.

[8] K. Bach, R. Bergmann, et al., Case-based reasoning meets large language models: A research manifesto for open challenges and research directions, 2025. URL: https://hal.science/hal-05006761.

[9] K. Wilkerson, D. Leake, On implementing case-based reasoning with large language models, in: Case-Based Reasoning Research and Development (ICCBR 2024), volume 14775, Springer Nature Switzerland, Cham, 2024, pp. 404–417. doi:10.1007/978-3-031-63646-2_26.

[10] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining 3 (2007) 1–13. doi:10.4018/jdwm.2007070101.

[11] Y. Yang, A study on thresholding strategies for text categorization, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, ACM, 2001, pp. 137–145. doi:10.1145/383952.383975.

[12] C. Largeron, C. Moulin, M. Géry, MCut: A thresholding strategy for multi-label classification, in: Advances in Intelligent Data Analysis XI, volume 7619, Springer Berlin Heidelberg, 2012, pp. 172–183. doi:10.1007/978-3-642-34156-4_17.

[13] M. Ma, G. Chochlakis, et al., Large language models do multi-label classification differently, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2025, pp. 2472–2495. doi:10.18653/v1/2025.emnlp-main.126.

[14] L. Huang, W. Yu, W. Ma, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. URL: http://arxiv.org/abs/2311.05232.

[15] A. Madaan, N. Tandon, P. Gupta, et al., Self-Refine: Iterative refinement with self-feedback, in: Advances in Neural Inf. Processing Systems 36 (NeurIPS 2023), 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

[16] R. Kamoi, Y. Zhang, N. Zhang, et al., When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs, Transactions of the Association for Computational Linguistics 12 (2024) 1417–1440. doi:10.1162/tacl_a_00713.

[17] Z. Gou, Z. Shao, Y. Gong, et al., CRITIC: Large language models can self-correct with tool-interactive critiquing, in: International Conference on Learning Representations (ICLR), 2024.

[18] H. Li, Q. Dong, J. Chen, et al., LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods, 2024. URL: http://arxiv.org/abs/2412.05579.

[19] A. Grattafiori, A. Dubey, A. Jauhri, et al., The Llama 3 herd of models, 2024. URL: http://arxiv.org/abs/2407.21783.

[20] A. Yang, A. Li, B. Yang, et al., Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388.

[21] OpenAI, S. Agarwal, L. Ahmad, et al., gpt-oss-120b & gpt-oss-20b model card, 2025. URL: https://arxiv.org/abs/2508.10925.