Applying Open-Weight Models and Case-Based Reasoning to Local Document Classification Tasks

Joseph Kendall-Morwick[1]

[1]Washburn University, 1700 SW College Ave, Topeka, KS 66621, USA

Abstract

Advances in machine learning, in particular embedding models and large language models, have been shown to be a natural fit within the case-based reasoning framework and this combination shows promise to assist with the complex task of document processing. However, document processing often involves the handling of private and confidential information while requiring significant computational resources, complicating matters for private individuals and small organizations wary of incorporating their data with compute resources controlled by third parties.

This study focuses on one important sub-problem in this task: document classification. In particular, the paper explores and evaluates existing classification techniques, especially as they've been applied to document classification, through the perspective of local document digitization projects. Public benchmark datasets as well as a novel private dataset are utilized, the latter providing new insight into the application of these techniques for personal use. Experiments are performed on consumer-grade hardware and the results validate the importance of combining embedding and large language models within the CBR cycle, particularly as these models mature and improve.

Keywords

Textual CBR, Document Classification, CBR and Large Language Models

1. Introduction

Document processing presents several distinct opportunities for applying artificial intelligence, from basic tasks like optical character recognition (OCR) to high-level tasks like question answering. Digitization projects involving manual scanning share a number of tasks in between, such as determining page membership in a document, determining document membership in a bundle or category, and detection of scanning errors.

Although deep learning models have been applied successfully to all of these tasks in modern document processing pipelines, deep learning alone may not satisfy the specific needs of those managing digitization projects, especially small groups or individuals. Cloud-based services present a threat to privacy, which is often a critical concern in these scenarios. Pre-trained models may not be adaptable to a specific distribution of document data and training from scratch, fine-tuning, or even simply running inference tasks on consumer-grade hardware can be cost prohibitive.

This paper provides the results of a study of contemporary techniques for incorporating embedding and large-language models into a case-based reasoning system for use with local document processing. These efforts align with Bach et al's "Case-Based Reasoning Meets Large Language Models" manifesto [1], most consequently toward addressing open challenge C9 (Computational Complexity and Size of LLMs) by leveraging the strengths of both case retrieval and adaptation through LLMs to address constraints of resource-limited computing in a meaningful application context. Specifically, this paper delivers the following contributions:

- Alignment of work both within and outside of the CBR community where there are common goals, especially in the domain of document classification
- Examination of methods in the context of a personal use case, including results from working with a novel, private dataset with properties that are not commonly found in public data

ICCBR Workshop on Memory-Based Reasoning for Responsible GenAI at ICCBR2026, August 15, 2026, Bremen, Germany
Joseph.Kendall-Morwick@washburn.edu (J. Kendall-Morwick)
0009-0000-2480-5901 (J. Kendall-Morwick)

- Novel validation of existing techniques for case-based classification, especially through the incorporation of mid-sized open-weight embedding and large-language models not previously reported on for this task
- An exploration of document classification adaptation strategies that reinforces the importance of LLMs in case-based classification
- An outline of important future directions for case-based document processing inspired by the work within this study and a larger document digitization project

All source code used in the experiments is publicly available at https://github.com/jmorwick/local-document-processing-scripts. Although results from the private dataset are not reproducible, results from the other two public datasets should be with this code.

The remainder of this paper is organized as follows: Section 2 elaborates on the document processing task in general with attention to local document processing constraints and introduces the document datasets used in this study while addressing related work. Section 3 addresses related work in document classification specifically and related CBR techniques while outlining the methodology of this study. Section 4 presents and examines the experimental results of this study, comparing them to prior work. Section 5 summarizes conclusions drawn from the study and section 6 provides a perspective on future work inspired by this study and the larger project it is a part of.

2. Document Processing

Although the contributions of this paper largely center around document classification specifically, and CBR more generally, they are derived from work in progress from an ongoing document digitization project first introduced at a prior ICCBR workshop [2]. The original goals of this project were to develop knowledge extraction techniques using CBR and open-weight models, but as the project progressed, smaller sub-problems arose relating to organizing and pre-processing the scanned documents before they would be ready for knowledge extraction. In particular, finding forms within the documents for knowledge extraction involved first classifying the documents to determine which were applicable to this task.

Many of these sub-problems (including document classification) were anticipated in document processing literature. In particular, in their paper "Beyond Document Page Classification", Van Landeghem et al. recognize additional classification tasks receiving less attention in document processing literature: classifying multi-page documents, recognizing document bundles, splitting apart "page streams" from document scans, and splitting smaller documents out of a single scanned page [3]. This paper primarily examines multi-page document classification but also briefly explores page stream segmentation and considers some of these other tasks in the future work section. Full document processing pipelines are considered out of scope for this paper – for a broader look at document processing pipelines, see Van Landeghem et al [3].

There are other smaller problems addressed in document classification literature that are not a focus of this paper, for example: Optical Character Recognition (OCR). Research on OCR is still ongoing but has been well addressed for local processing. Astier et al. even explore an application of CBR within the process of preparing a document for OCR [4].

All tasks in this study use a well-known and well studied tool, Tesseract (specifically version 5.5.0), for OCR [5]. It is important to note that the techniques studied later in this paper rely on the data from the OCR model and can vary depending on the quality of the OCR process. The choice of OCR model may lead to different results, especially with newer OCR models incorporating an LLM (such as DeepSeekOCR [6]). However, since the studies referenced in this paper all use Tesseract, the impact of OCR model choice is considered out of scope for this study and left to future work.

Many OCR models also solve the more simple problem of scan orientation (including Tesseract). Other issues, such as duplicate detection, can be more nuanced. While many of these document and dataset quality tasks can be impactful on the down-stream tasks such as document classification, these tasks

are considered outside the scope of this paper to maintain consistency with related studies. However, as a bridge to future work, the document stream segmentation task is briefly examined.

2.1. Local Document Processing Goals and Constraints

AI cloud-based services present privacy risks to those seeking to analyze documents containing sensitive data due to the necessity of sharing these documents, unencrypted, with servers under the control of the cloud service provider. These risks could be borne by third parties whose data is included in the documents or those entrusted with the data who may bear liability for its disclosure. Organizations with sufficient resources can establish controls over these risks through legal agreements with the cloud provider or by performing the analysis in-house on organization owned or controlled equipment.

Small organizations, families, and individuals more often lack such resources and yet have needs for managing personal documents that may contain sensitive information (such as medical records). In these cases, and in the absence of a better solution, the risk of working with cloud services may simply be accepted as the convenience they offer may be deemed to outweigh the severity of the risk.

Local document processing using efficient algorithms and open-weight models can eliminate these risks if the computing resources are sufficient. This project aims to analyze the feasibility of local document processing towards this use case. To this end, only open-weight models are analyzed and only moderate to low power and cost computing equipment are considered. All experiments for this study were performed on modest consumer hardware: a system with an AMD Ryzen 7 3700X 8-Core Processor, 64GB of RAM, and a single RTX 3090 with 24GB of VRAM. At the time of this writing, while RAM and GPU prices are unusually high, such a system would be estimated to cost $1700-2000 with used parts. The largest model used on this hardware in this study, gemma4-26b, is a "Mixture of Experts" model that performs efficiently for its size. The RTX 3090 has been benchmarked at 114.8 tokens/s with this model and completes the largest queries in this study under 20 seconds (or about 2.8 watt-hours of energy) [7].

For home users, simplicity of process is an additional aim. This is further complicated by the constraint that the semantic categories of datasets for individuals will vary significantly. For example, an architect may have a number of "Blueprints" in their document collection while a musician might have "Sheet Music". Interaction with the system should be minimized when possible, accepting that some degree of feedback may be needed from the user for the system to tailor its results to the user's preferences. Training and fine-tuning models can be both computationally expensive and complex, so only fully pre-trained models are considered for this project. This also exposes another benefit CBR brings to systems aimed at home users: new case retention provides a "lazy" learning mechanism that doesn't rely on complex and expensive batch processing.

Related to simplicity of process is the importance of explainability of automated actions. For a system aimed at home users, actions and decisions taken by the system should be explainable even to a novice. Both LLMs and CBR can be leveraged for this purpose, the latter going further to provide actionable explanations. Rather than a generated justification, users can be presented with the most similar case leading to the classification. This is helpful for two reasons: it is simple for the end-user to understand and is also actionable – if the resulting classification was an error, it is possible that the retrieved case was also errantly labeled, and this presents an opportunity for the user to improve the system's working case knowledge by correcting the mistake.

2.2. Document Datasets

Finding public data suitable for testing software to work with private data presents a difficult conflict of goals. Two public document datasets were chosen for this study based on their use for similar purposes in prior studies and one novel private dataset was developed. They are described through the remainder of this section.

2.2.1. RVL-CDIP-N

RVL-CDIP-N was developed by Larson et al. to test the ability of document classification models to generalize to novel distributions of data.[8] It contains 1002 documents over 12 classes. The 12 classes of documents this dataset uses come from the 16 classes of the related RVL-CDIP dataset. However the documents in RVL-CDIP-N do not, themselves, come from RVL-CDIP and contain otherwise unrelated content.

RVL-CDIP is a very popular dataset used to train and test document classification models [9]. RVL-CDIP contains many documents that would otherwise be private but were made public as part of legal discovery in litigation involving the American tobacco industry. Larson et al. provide many citations to related work on training classifiers with high accuracies over the RVL-CDIP dataset. However, in their study, they show that many of these models, once trained on RVL-CDIP, fail to generalize to the newer data unrelated to the tobacco industry in RVL-CDIP-N. This makes the results from their study (and others who use their dataset) a more suitable baseline for comparison with the pre-trained models relied on within this paper's study.

The model with the highest accuracy in the RVL-CDIP-N study is DiT-base with 78.6%, which also happens to be the only transformer model used in the study and was designed with resilience to distribution shift in mind [10]. This makes the accuracy score a good baseline for comparing other fully pre-trained or lazy-learning approaches on this dataset.

2.2.2. RVL-CDIP-MP

The next dataset, RVL-CDIP-MP [3] was derived mostly from RVL-CDIP-N but was incorporated into this paper's study to better align results between the two studies. The specific dataset used in this paper was called RVL-CDIP-MP-N in the prior study but will be referred to as RVL-CDIP-MP for the remainder of this paper. After downloading and incorporating the dataset, it was clear that there was a lot of overlap with RVL-CDIP-N, but also some differences with the files. For example, only 998 of the files were used in this dataset since four of the original 1002 caused a problem with Tesseract causing it to lock up (possibly a bug).

2.2.3. JBKM-PDF1

The original proposed study from the digitization project [2] was to incorporate data from the CIS department of Washburn University, but as that data is not yet authorized for use in a public study, a different dataset was developed from the author's own personal files to better represent the idiosyncrasies of a personal dataset. The dataset is comprised of multi-page, mostly (but not exclusively) scanned PDF files and was developed by mining personal files used on the author's personal computer over the last two decades. While it will remain private, statistics about the distribution of the data and how it may impact the classification task are examined in this paper. Through the remainder of this paper and any future publications, it will be referred to as JBKM-PDF1.

Table 1 provides class counts for the dataset. The data was classified by the folder it was organized into in the author's personal computer. Challenges inherent with this dataset include a large number of classes (23) and a clear class imbalance. Also, while RVL-CDIP uses very common and abstract categories, the categories in JBKM-PDF1 reflect the semantic ambiguities and complexities that can be expected in other personal datasets. For example, "Form" is only one category in RVL-CDIP but four in JBKM-PDF1. You can also see categories that are mostly unique to academics, such as "Course Evaluation" and "Rubric".

Table 1
Distribution of document classes in the dataset

Class	Count	Class	Count
Recipe	125	Rubric	11
Contract	97	Letter	10
Tax Form	57	Packing Slip	10
Book Chapter	43	Paystub	9
Bill or Statement	39	Course Evaluation	7
Receipt	39	Certificate	6
Insurance Form	34	Warranty	5
Academic Paper	22	Auto Form	4
Health Form	21	Lecture Notes	4
Documentation	21	Credit Report	3
Notes	14	Syllabus	2
Policy	11		

3. Document classification with open-weight models and Case-Based Reasoning

The following subsections present the methodology used in this study while continuing to explore related work. Leave-one-out testing was utilized for each experiment.

Only a few papers integrating LLMs into CBR are referenced in this section. These studies are the most relevant related work to the scope and specific methods of this study. However, many more papers have been published studying the relationship of LLMs with work in CBR more broadly. For a more thorough perspective on the past and future of incorporation of LLMs in CBR, see Bach et al [1].

3.1. Direct classification with pre-trained LLMs (Zero Shot Prompting)

Pre-trained Large Language Models are general purpose models capable of performing many tasks without the need for additional training data (zero shot prompting). For this approach, a recent medium-sized model (capable of running within 24GB of VRAM) was selected: Google's gemma4-26b released on April 2nd, 2026 [11]. The model was prompted with a request to classify a document according to one of the possible classes from a given dataset, followed by the OCR text of the document (either the first page or all pages).[1] The names of the classes are critical in this case since it is the only information the model receives to determine the meaning of the class. Default settings and model parameters for Ollama[2] were used for all LLM models.

This technique is neither new to CBR nor document classification research. For example, Scius-Bertrand et al. use this technique with the full RVL-CDIP dataset [12]. Wilkerson and Leake also apply this technique within the context of case-based medical triage [13]. Neither paper focuses exclusively on open-weight or small-to-medium sized models, but both papers involve at least one such model (Mistral-7B and Llama 2 70b-chat respectively).

3.2. K-NN classification with pre-trained embedding models

Embedding models transform sequential text input into a vector representing the semantic meaning of the text. These models are often much smaller than LLMs and can run on very modest hardware, even without a GPU. Embeddings of documents can be compared to assess the semantic similarity of the documents. Embedding models have been explored for the purpose of case retrieval with positive results (e.g. Wiratunga et al. [14]).

[1] prompt text is available in the source code repository
[2] https://ollama.com/

Case retrieval performance with embedding models is strong enough that, in a classification context, simply returning the solution from the most similar case will often be correct. This baseline is explored in Wilkerson and Leake's work [13]. It is also explored by Scius-Bertrand et al. who used it over the RVL-CDIP dataset[12]. Using 1600 training samples, 160 validation samples, and 160 test samples, they optimized a K value (not reported, but in the range 1-9). The highest accuracy achieved was 67.8% from OpenAI-large. They also used an open-weight model, mistral-embed, which achieved 66.8% accuracy.

For this study, several more recent pre-trained embedding models were explored to determine the extent that model choice and model size would play in the accuracy of K-NN. Several values of K were chosen for the highest performing model to explore the impact of that parameter as well. Similarity scores between two documents were determined by normalized dot products of the embedding vectors.

3.3. Adaptation using fusion over confidence values from multiple classifiers

The next two subsections will explore adaptation methods that combine the previous two. Multiple classifiers, if capable of providing a reliable confidence value, can be fused to produce a single classifier with potentially higher accuracy. Similar to ensemble methods, this technique has been explored as a means of adaptation in CBR in the past [15]. In order for classification accuracy to improve, confidence values should correlate with the probability that the classification is correct.

In this study, a simple approach to adaptation is taken through fusing 1-NN with zero shot prompting. In this approach, the LLM is not made aware of the most relevant case but is asked to provide a confidence value. In the case of competing answers, the classification with the higher confidence value is used. Final confidence values are boosted when both classifiers agree and reduced when they disagree. Specifically, if c_1 and c_2 are the confidence values for both classifiers and each are drawn from $[0, 1]$ with $c_1 \geq c_2$, then $1 - ((1 - c_1)(1 - c_2))$ is used as the confidence value when classifications match and $c_1(1 - c_2)$ is used when they do not. Despite their use as probabilities in these formulae, these confidence values were not calibrated to reflect actual probabilities so the final classification results could be considered a baseline for improvement in future iterations.

3.4. Adaptation using one / few shot prompting

A more recent approach to prompting LLMs is to retrieve and incorporate details of relevant documents into prompts to improve the results. This is popularly known as "Retrieval Augmented Generation" (RAG) and has been examined by Wiratunga et al. in their "CBR-RAG" study[14].

Scius-Bertrand et al., in their study on document classification, utilized one shot prompting by providing one example of each of the 16 different possible classifications as a static component of the prompt [12]. This improved results from their models (Mistral-7B, Mistral-8x7B, and GPT-3.5) by 1.7%, 23.2%, and 21.9% respectively, but accuracies from zero shot started fairly low (45.4%, 25.0%, 36.9%, respectively).

Wilkerson and Leake explore three more "case-based" approaches they dub 1-NN, 2-NN, and NUN (Nearest Unlike Neighbor)[13]. The first two strategies incorporate either the most similar or two most similar cases into the prompt, along with their classifications, as suggestions for the LLM to consider while leaving the LLM free to decide on an entirely different classification. The third, NUN, presents the most similar case as well as the next most similar case with a different classification. They observed an even greater increase in accuracy than Scius-Bertrand et al. but, again, from a low starting point. In the end, the more sophisticated prompting strategies performed similarly to K-NN.

This study incorporates the strategy taken by Wilkerson and Leake with some small changes to the prompts and a new model. gemma4-26b (context size 32k tokens) is used and prompts request a confidence value in addition to the classification. The confidence values provide an opportunity for users to fine-tune response rates for higher precision at the expense of lower recall – a useful quality for a document processing pipeline.

Table 2
Accuracies of embedding models used for retrieval in 1-NN over JBKM-PDF1

Model	First Page Only	Whole Document
jina-embeddings-v2-base-en	27.27%	26.94%
granite-embedding-97m-multilingual-r2	77.27%	78.45%
bge-small-en-v1.5	79.29%	80.47%
embeddinggemma-300m	79.63%	79.97%
qwen3-embedding-0.6b	79.46%	80.64%
qwen3-embedding-4b	80.13%	81.99%
qwen3-embedding-8b	81.48%	82.32%

3.5. Document stream segmentation via classification

In this study, an additional experiment was performed for the purpose of document stream segmentation by classifying "first pages" of a document. Classifiers were queried with every page of every document in the training set and asked to classify the page as either a first page or not a first page of a document. The highest performing 1-NN and zero shot models were tested as well as both adaptation approaches. As with other retrieval-based approaches, leave-one-out testing was used but instead of leaving out only the query page, every page from the query document was left out.

This task was not explored to the same extent as full document classification but included to investigate how broadly applicable the studied classification techniques may be to other aspects of document processing. Other page stream segmentation techniques that are more tailored to the task have been explored in other literature (e.g. [16], [17]).

4. Results

An important limitation of the results below is to consider that leave-one-out testing was used for each of this paper's experiments but some of the referenced experiments use smaller training sets, raising the possibility that this study's accuracies for retrieval-based techniques are higher than would be expected.

4.1. Embedding model evaluation for nearest neighbor tasks

A wide variety of embedding models were used to determine what aspects of embedding models may lead to better retrieval results for the purposes of 1-NN. The results are in Table 2. Notably, the accuracies of all models (other than Jina) were similar, but larger and newer models tended to perform better. Although all of the models performed well within the hardware constraints previously mentioned, the smaller models may be better choices for batching or for use with very old consumer hardware. However, only qwen3-embedding-8B will be used for retrieval in all further experiments in this study since it had the highest performance. It should also be noted that the Jina model was fine-tuned by Scius-Bertrand et al.[12] but wasn't here – this surely contributed to its relatively poor performance.

The last two columns of Table 2 compare the accuracies of 1-NN when only embeddings from the first page are used or the average of all page embeddings for the entire document. The findings show that, for the purpose of document classification, there isn't a large benefit to calculating embeddings for any pages beyond the first. This mostly comports with findings of Van Landeghem et al. [3] and shows that these insights haven't shifted much with the newer embedding models or the introduction of a novel document domain.

Finally, Table 3 explores several values of K for similarity-weighted K-NN. Classes were chosen by summing all similarity scores for each class from each retrieved document. The final confidence was determined by dividing this sum for the chosen class by K, thus results with mixed classes have lower confidence values. Accuracy clearly falls as K rises and for this reason only K=1 is used for K-NN in future experiments (with the exception of two RAG variants). Notably AUC-ROC does not strictly

Table 3
Performance for various K values with K-NN over qwen3-embedding-8b embeddings on JBKM-PDF1

K	Accuracy	AUC-ROC
1	82.32%	0.6932
3	79.97%	0.7773
5	77.78%	0.8180
10	74.41%	0.8331
20	70.54%	0.8167
40	67.34%	0.7960
80	61.78%	0.8093

fall as K increases, leaving open the possibility that some other values of K might make for better confidence-limited approaches or adaptation techniques.

4.2. Direct classification

Several LLMs were used to directly classify documents given their OCR text. These results are listed in Table 3. Mistral:7b was included as a baseline to compare results with Scius-Bertrand et al. Mistral:7b performed similarly to previous work on RVL-CDIP derived datasets but even more poorly on JBKM-PDF1. Model size again appears to play a large role, but the smaller (and newer) gemma3-4b outperformed mistral:7b by a wide margin, so recency is also likely a factor. The best performing model was gemma4-26b, but not by a large margin over gemma3-12b, meaning that older consumer equipment with less RAM than the test computer (24GB VRAM) may opt to use this smaller model. Using only the first page didn't consistently improve results (in fact, accuracy was reduced in 2 of the 3 datasets) so only the first pages of documents are used in remaining experiments. The accuracies from using all pages are indicated in the "All Acc" column in Table 4 and blank for experiments using first-page only.

The newest LLM tested, gemma4-26b, had lower accuracy relative to 1-NN on the JBKM-PDF1 dataset but remarkably higher accuracy on RVL-CDIP. This is understandable considering that the labels in JBKM-PDF1 are much more idiosyncratic and dependent on the interpretations held by the individual labeling the documents, increasing the importance of incorporating retrieved classification cases.

4.3. Adaptation

In Case-Based Reasoning, adaptation knowledge is applied towards adapting retrieved solutions to the particulars of a query problem. LLMs can serve both as the adaptation knowledge container and reasoning mechanism to perform this adaptation. However, in order for this process to work well, confidence values determined by all components of the system should correlate with the likelihood of a classification being correct. Furthermore, if confidence values have this property, users can set confidence thresholds to improve the accuracy of the system and inhibit it from substituting its judgment for the users' in cases where it is not confident in its response.

An important measurement to determine the quality of a confidence value is the Area Under the Curve for the Receiver Operating Characteristic (AUC-ROC). High values for AUC-ROC (close to 1) indicate strong predictive capability of a confidence metric. Middle values (close to 0.5) indicate that the confidence values aren't any more reliable than random guessing. The scores for the candidate component classifiers are listed in Table 4. For experiments involving adaptation of retrieved results, models with both high accuracy and high AUC-ROC were chosen. For all such experiments, qwen3-embedding-8b was used as the embedding model and gemma4-26b was used as the LLM.

Using confidence values to decide between classifications from two different classifiers can't lead to correct classifications unless at least one of the two classifiers is correct – fusion is not an open-ended form of adaptation. Thus, the fused 1-NN / Zero Shot classifier has a maximum accuracy potentially lower than 100% in which an oracle were to determine which classifier to use to maximize accuracy.

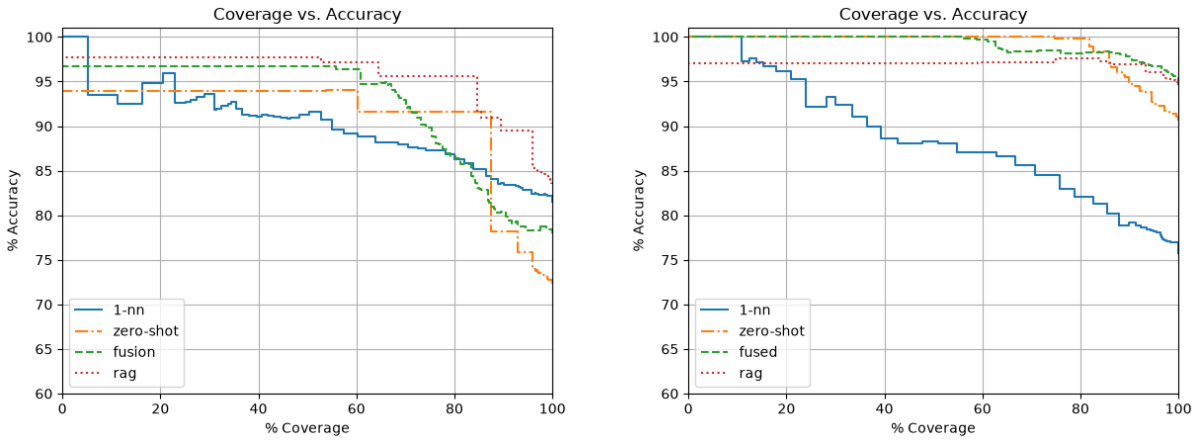
Table 4
Document Classification Performance

Technique	Model(s)	Dataset	Accuracy	All Acc	AUC-ROC	Coverage
1NN	qwen3-embedding-8b	JBKM-PDF1	81.48%	82.32%	0.7041	100.00%
1NN	qwen3-embedding-8b	RVL-CDIP-N	75.75%	74.55%	0.7399	100.00%
1NN	qwen3-embedding-8b	RVL-CDIP-MP	76.05%	74.95%	0.7423	100.00%
Zero Shot	mistral-7b	JBKM-PDF1	35.77%	33.33%	0.6226	45.02%
Zero Shot	mistral-7b	RVL-CDIP-N	33.95%	36.86%	0.7443	37.23%
Zero Shot	mistral-7b	RVL-CDIP-MP	33.73%	36.54%	0.7451	37.12%
Zero Shot	gemma3-4b	JBKM-PDF1	51.85%	46.62%	0.5861	77.78%
Zero Shot	gemma3-4b	RVL-CDIP-N	43.80%	25.05%	0.6300	68.30%
Zero Shot	gemma3-4b	RVL-CDIP-MP	43.88%	24.64%	0.6298	68.67%
Zero Shot	gemma3-12b	JBKM-PDF1	66.67%	69.36%	0.6772	96.80%
Zero Shot	gemma3-12b	RVL-CDIP-N	85.43%	81.92%	0.8412	98.20%
Zero Shot	gemma3-12b	RVL-CDIP-MP	84.77%	81.24%	0.8423	98.40%
Zero Shot	gemma4-26b	JBKM-PDF1	72.30%	77.16%	0.8254	99.49%
Zero Shot	gemma4-26b	RVL-CDIP-N	90.56%	89.55%	0.9393	99.50%
Zero Shot	gemma4-26b	RVL-CDIP-MP	89.92%	88.91%	0.9388	99.60%
Fusion	qwen3 / gemma4-26b	JBKM-PDF1	77.95%		0.8365	100.00%
Fusion	qwen3 / gemma4-26b	RVL-CDIP-N	94.63%		0.8899	100.00%
Fusion	qwen3 / gemma4-26b	RVL-CDIP-MP	94.18%		0.8796	100.00%
RAG 1NN	qwen3 / mistral-7b	JBKM-PDF1	46.21%		0.7450	53.79%
RAG 1NN	qwen3 / mistral-7b	RVL-CDIP-N	50.17%		0.6638	63.22%
RAG 1NN	qwen3 / mistral-7b	RVL-CDIP-MP	50.08%		0.6690	63.15%
RAG 1NN	qwen3 / gemma4-26b	JBKM-PDF1	83.50%		0.8545	98.32%
RAG 1NN	qwen3 / gemma4-26b	RVL-CDIP-N	94.81%		0.7138	100.00%
RAG 1NN	qwen3 / gemma4-26b	RVL-CDIP-MP	94.49%		0.7212	100.00%
RAG 2NN	qwen3 / mistral-7b	JBKM-PDF1	65.02%		0.6928	78.92%
RAG 2NN	qwen3 / mistral-7b	RVL-CDIP-N	68.42%		0.5941	89.60%
RAG 2NN	qwen3 / mistral-7b	RVL-CDIP-MP	68.34%		0.5854	89.24%
RAG 2NN	qwen3 / gemma4-26b	JBKM-PDF1	**85.52%**		0.7910	99.83%
RAG 2NN	qwen3 / gemma4-26b	RVL-CDIP-N	**95.99**%		0.7372	99.80%
RAG 2NN	qwen3 / gemma4-26b	RVL-CDIP-MP	**95.48**%		0.7446	99.80%
RAG NUN	qwen3 / mistral-7b	JBKM-PDF1	50.13%		0.7452	79.05%
RAG NUN	qwen3 / mistral-7b	RVL-CDIP-N	49.11%		0.6946	75.93%
RAG NUN	qwen3 / mistral-7b	RVL-CDIP-MP	49.17%		0.6928	75.97%
RAG NUN	qwen3 / gemma4-26b	JBKM-PDF1	84.34%		0.7895	98.65%
RAG NUN	qwen3 / gemma4-26b	RVL-CDIP-N	94.10%		0.7499	100.00%
RAG NUN	qwen3 / gemma4-26b	RVL-CDIP-MP	93.88%		0.7656	100.00%

These maximum accuracies are 90.57%, 96.56%, and 96.28% for JBKM-PDF1, RVL-CDIP-N, and RVL-CDIP-MP, respectively. The maximum for the JBKM-PDF1 dataset leaves some room for improvement through refinement of voting strategy, but for the RVL-CDIP datasets the accuracy is already so close to the maximum that refinements wouldn't provide a substantial improvement in accuracy.

The three methods of RAG-based adaptation for classification suggested by Wilkerson and Leake produced the best results. These results can be found in Table 4 with a result added for mistral:7b for comparison. Notably mistral:7b performed similarly over the RVL-CDIP derived dataset as it did for Scius-Bertrand et al. with their one shot technique, further emphasizing the improvement from changing the model to gemma4-26b. Much of the reason for the drop in performance from mistral:7b is due to its relatively frequent failure to answer the prompt as specified. Any malformed response is counted as an incorrect classification. Notably gemma4-26b very rarely generates malformed responses. The percentage of documents provided well-formed responses is indicated as "coverage" in Table 4.

It should be noted that Scius-Bertrand et al. did observe a significant accuracy boost with minimal



(a) JBKM-PDF1 (b) RVL-CDIP-N

Figure 1: Coverage vs Accuracy for Document Classification

fine-tuning of mistral:7b: 72.5% from 160 training samples. This rose even further to 83.4% with 1600 training samples. However, recall that the fused classifier is limited by the capabilities of its components, but in the RAG approach the LLM is free to make any classification it wishes, regardless of the retrieved case. Without any fine-tuning the gemma4-26b model achieves accuracies remarkably close to the oracle maximum of the fused classifier, and for the RVL-CDIP datasets, not very far from 100%. This suggests that the LLM is performing important reasoning over the prompt it is presented with leading to better outcomes than simply fusing classifiers. It's notable that there isn't a lot of difference between the performance of the different prompting strategies, but given the importance of the reasoning process, it seems likely that better prompt engineering could improve results further in domains like JBKM-PDF1.

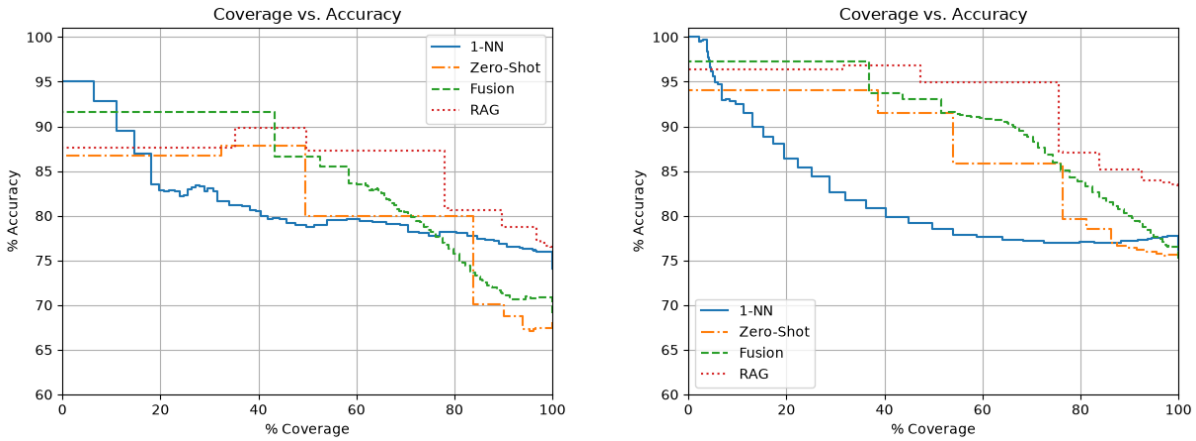
Raising a minimum confidence threshold on a classifier has the effect of reducing coverage while (usually) increasing accuracy. Figure 1 examines accuracies of classifiers when confidence thresholds are used (1-NN is used for the RAG approach in both Figure 1 and Figure 2). If an individual desired higher than 95% accuracy, the RAG method would achieve this through 80% coverage on the JBKM-PDF1 dataset. Fusion doesn't catch up until coverage is reduced to 60% and has similar accuracy from that point forward. On the RVL-CDIP dataset, both Fusion and Zero Shot outperform RAG through most of the plot with Fusion achieving 100% accuracy at 60% coverage and Zero Shot achieving 100% accuracy at 80% coverage. RAG also achieves over 95% accuracy through 80% coverage. Higher performance for zero shot could be expected on the more abstract labels in RVL-CDIP datasets and the results suggest that high-confidence zero shot classifications could be useful for bootstrapping a dataset. Overall, the results suggest that RAG may be the best choice once enough labels have been accepted.

4.4. Stream segmentation classifiers

The results from Table 5 show that there is more room for improvement for this technique when applied to first-page prediction. Overall results were poor relative to document classification for the JBKM-PDF1 dataset. RAG (the most consistently top-performing strategy) hit a ceiling of 90% accuracy with only 50% coverage (see Figure 2) over this data. The top-performing technique for the RVL-CDIP-N dataset (again RAG with 83.42% accuracy) is comparable to several baselines used in a separate study from Wiedemann and Heyer on a similar dataset (Tobacco800) [16], but also approaches 95% accuracy with close to 75% coverage when adjusting the minimum confidence threshold. Wiedemann and Heyer also achieved 91.1% on Tobacco800 with a model incorporating visual elements with text, emphasizing the importance of visual and structural elements of the document for this task and making vision models an intriguing opportunity to better apply the classification techniques from this study. gemma4-26b is also a vision-capable model, but in this study it was only tested with text input in order to compare it directly to other text-based models.

Table 5
First Page Classification Performance

Technique	Model	Dataset	Accuracy	AUC-ROC	Coverage
1NN	Qwen3-Embedding-8B	JBKM-PDF1	74.11%	0.6223	100.00%
Zero Shot	gemma4-26b	JBKM-PDF1	68.10%	0.6556	100.00%
Fusion	qwen3 / gemma4-26b	JBKM-PDF1	69.05%	0.7483	100.00%
RAG 1NN	gemma4-26b	JBKM-PDF1	76.56%	0.6747	100.00%
1NN	Qwen3-Embedding-8B	RVL-CDIP-N	76.23%	0.5649	100.00%
Zero Shot	gemma4-26b	RVL-CDIP-N	75.96%	0.6787	100.00%
Fusion	qwen3 / gemma4-26b	RVL-CDIP-N	75.22%	0.8123	100.00%
RAG 1NN	gemma4-26b	RVL-CDIP-N	83.42%	0.7086	100.00%



(a) JBKM-PDF1 (b) RVL-CDIP-N

Figure 2: Coverage vs Accuracy for First-Page Classification

5. Conclusions

Several important conclusions can be derived from the results of this study:

- CBR, using pre-trained models for similarity assessment and adaptation, is an effective approach to document classification, particularly in regard to the unique constraints of local document processing.
- Modern embedding models are powerful enough to yield substantial results on their own through 1-NN classification, all while having a light enough compute footprint to achieve these results on any budget for home computing equipment.
- Modern mid-sized LLMs have improved and are now also quite capable of local document classification without the need for fine-tuning and are additionally capable of providing usable confidence estimates.
- Adaptation of retrieved results using LLMs (including Fusion and RAG) increases performance and explainability of results, but RAG-based approaches appear uniquely capable in this regard.

6. Future Directions

There are two major categories of future directions for the work mentioned in this paper: future directions for the document digitization project and future directions for document classification specifically. First, for the work on document classification, although high accuracy was achieved on a challenging public dataset, the reduced performance on the private dataset shows there is still work to

be done. Testing more models (and models not yet released) will be important, as it was in this study, but this study did not experiment much with prompt engineering, parameter optimization, similarity metric formulation, confidence estimation, or calibration strategies. Re-ranking was also not explored in this study and should be an important avenue for continued research with newer models (both LLMs and re-ranking embedding models), as well as other similar MAC/FAC approaches in CBR.

Vision / multi-modal models are another very important area for further study (e.g. [12]). Most of the incorrect responses from the most accurate classifier on RVL-CDIP-N were for the "handwritten" category, which is very difficult to discern from OCR results, and there are many other visual structural elements of the document (headings, tables) that have important semantic value in document analysis, especially leading into knowledge extraction.

While knowledge extraction is an important goal in and of itself, it also can be a means of improving document classification tasks beyond those considered in this paper. For example, in form understanding, questions (the title of a field on a form) and answers (the value entered into the field) provide two very different semantic dimensions of the document. The questions can aid determination of the type of form, but the answers can help determine if the form belongs in a document bundle (such as a patient chart where each document relates to the same patient). It's expected that as work advances in knowledge extraction, it can also be leveraged to improve results on tasks studied in this paper.

Document bundling (as well as the rest of these document classification tasks) can also be viewed at various levels of abstraction. Consider documents relating to an entire "criminal record" or an individual criminal case. Document labels themselves, such as those explored in this paper, can be ambiguous or belong to complex ontologies. Although these complications are often mentioned in the literature, they are challenging to address and seldom explored, leaving room for important work in this area.

Finally, a very important issue not addressed at all in this paper but very relevant to this study is the identification of documents and the collection and management of digital provenance. Reasoning about data is aided by content-based identification of immutable data. These Content-based Identifiers (CID's) can be used to associate provenance and other metadata relating to transformations, information extraction, and inference performed over documents. This can facilitate long-term archiving of information, distributed analysis, privacy-preserving analysis (using CID's in the place of the actual data), and caching of results from expensive computational processes, especially in regards to scientific studies on data such as this study itself. CID's can also be used to address retrieval concerns outlined in Bach et al's manifesto (O5.4) [1]. Some of this work has already been started and can be seen within the source code accompanying this project.

Declaration on Generative AI

The author used the OpenAI Deep Research tool to assist with the literature search for this paper. The author used GPT-5.5 to assist with the development of some of the source code used to run experiments linked to in the paper. All references and source code incorporated in this study from generative AI tools were evaluated by the author who takes full responsibility for their validity. GPT-5.5 was also used to identify spelling and grammatical errors in the final draft of this paper which were subsequently hand-corrected. No part of the text of this paper was generated using generative AI tools.

References

[1] K. Bach, R. Bergmann, F. Brand, M. Caro-Martínez, V. Eisenstadt, M. W. Floyd, L. Jayawardena, D. Leake, M. Lenz, L. Malburg, D. H. Ménager, M. Minor, B. Schack, I. Watson, K. Wilkerson, N. Wiratunga, Case-based reasoning meets large language models: A research manifesto for open challenges and research directions, 2025.

[2] J. Kendall-Morwick, Working with ambiguous case representations, in: L. Malburg, D. Verma (Eds.), Proceedings of the Workshops at the 31st International Conference on Case-Based Reasoning,

Aberdeen, Scotland, UK, July 17, 2023, volume 3438 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 71–80.

[3] J. Van Landeghem, S. Biswas, M. Blaschko, M.-F. Moens, Beyond document page classification: Design, datasets, and challenges, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2962–2972.

[4] E. Astier, H. Iopeti, J. Lieber, H. Mathieu Steinbach, L. Yvoz, Case-based cleaning of text images, in: Case-Based Reasoning Research and Development: 31st International Conference, ICCBR 2023, Aberdeen, UK, July 17–20, 2023, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 344–358.

[5] R. Smith, An overview of the tesseract OCR engine, in: Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR '07, IEEE Computer Society, USA, 2007, pp. 629–633.

[6] H. Wei, Y. Sun, Y. Li, DeepSeek-OCR: Contexts optical compression, 2025. URL: https://arxiv.org/abs/2510.18234. arXiv:2510.18234.

[7] Gemma 4 on old GPUs: Why a $700 used card beats $20,000 of professional hardware, https://ai.gopubby.com/gemma-4-on-old-gpus-why-a-700-used-card-beats-20-000-of-professional-hardware-ba81680a2fc4, 2026. Accessed: 2026-05-28.

[8] S. Larson, Y. Y. G. Lim, Y. Ai, D. Kuang, K. Leach, Evaluating out-of-distribution performance on document image classifiers, volume 35, 2022, pp. 11673–11685.

[9] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: International Conference on Document Analysis and Recognition (ICDAR), 2015.

[10] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, DiT: Self-supervised pre-training for document image transformer, in: Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 3530–3539.

[11] Google DeepMind, Gemma 4 model card, 2026. URL: https://ai.google.dev/gemma/docs/core/model_card_4.

[12] A. Scius-Bertrand, M. Jungo, L. Vögtlin, J.-M. Spat, A. Fischer, Zero-shot prompting and few-shot fine-tuning: Revisiting document image classification using large language models, in: Pattern Recognition, Springer Nature Switzerland, 2024, p. 152–166.

[13] K. Wilkerson, D. Leake, On implementing case-based reasoning with large language models, in: J. A. Recio-Garcia, M. G. Orozco-del Castillo, D. Bridge (Eds.), Case-Based Reasoning Research and Development, Springer Nature Switzerland, Cham, 2024, pp. 404–417.

[14] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, B. Fleisch, CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering, in: J. A. Recio-Garcia, M. G. Orozco-del Castillo, D. Bridge (Eds.), Case-Based Reasoning Research and Development, Springer Nature Switzerland, 2024, pp. 445–460.

[15] D. Leake, J. Kendall-Morwick, Four heads are better than one: Combining suggestions for case adaptation, in: L. McGinty, D. C. Wilson (Eds.), Case-Based Reasoning Research and Development, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 165–179.

[16] G. Wiedemann, G. Heyer, Page stream segmentation with convolutional neural nets combining textual and visual features, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[17] M. M. Islam, M. S. Salekin, N. Balakrishnan, V. Bishop, N. Jain, S. Romo, B. Strahan, B. Xie, D. A. Socolinsky, Docsplit: A comprehensive benchmark dataset and evaluation approach for document packet recognition and splitting, ArXiv abs/2602.15958 (2026).