HIRA: A Human-in-the-Loop Retrieval-Augmented Cascade for Document Classification in Regulated Industries

Shangxuan Tian*
Standard Chartered Bank
Guangzhou, China
tianshangxuan@u.nus.edu

Yanhui Chen
OCBC
Singapore
yanhui79@gmail.com

Carlos Queiroz
OCBC
Singapore
caxqueiroz@gmail.com

Abstract

Document classification in regulated industries is constrained by data residency, limited cold-start labels, scarce review capacity, and costly model-governance procedures. In this setting, the practical goal is not only to obtain high accuracy, but also to reduce LLM calls, avoid frequent model retraining, and send only the most ambiguous cases to human review. We present HIRA, a training-free, on-premises retrieval-augmented cascade for document classification in regulated deployments. HIRA first combines BM25 over OCR text, dense text embeddings, and image-level representations through validation-calibrated weighted reciprocal-rank fusion. Confident documents are classified directly by retrieval. Uncertain or visually confusable documents are passed to a locally hosted LLM verifier, which receives the OCR text, retrieved exemplars, label descriptions, and confusion-specific terms. When the verifier remains uncertain, the document is sent to human review. Each correction is then stored as a margin-weighted retrieval exemplar and also updates a Dirichlet-smoothed confusion graph, allowing the system to improve without updating model weights.

On a private 80-class trade-finance corpus, HIRA processes the full 30,233-document production stream while requesting human correction for only 1,945 documents (6.4%), improving Macro-F1 from 0.6218 to **0.8548**. On the corrected Tobacco-3482 benchmark, HIRA reaches **0.9423** Macro-F1 with a locally hosted DeepSeek-R1-Distill-Qwen-32B verifier, **17.4 percentage points above** the zero-shot LLM baseline, while invoking the verifier for only about **40%** of documents and therefore reducing LLM calls by approximately **60%**. With 518 human corrections, corresponding to 24.8% of the pool, HIRA matches the fully labelled pool oracle in which all 2,086 pool documents are indexed with their ground-truth labels. These results show that selective human feedback and retrieval-memory adaptation can provide a practical alternative to repeated model retraining for long-tail document classification in regulated deployments.

CCS Concepts

• **Computing methodologies** → **Classification and regression trees**; • **Information systems** → **Information retrieval**.

Keywords

retrieval-augmented classification, human-in-the-loop, data residency, training-free adaptation, large language models

ACM Reference Format:
Shangxuan Tian, Yanhui Chen, and Carlos Queiroz. 2026. HIRA: A Human-in-the-Loop Retrieval-Augmented Cascade for Document Classification in Regulated Industries. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3799682.3840098

1 Introduction

Fine-grained document classification underpins information systems across regulated industries, including financial-services back offices, healthcare records, legal discovery, and government archiving. The literature has largely converged on a single recipe: pretrain a layout-aware backbone, fine-tune on a labelled training set, and deploy. The emergence of large language models (LLMs) and vision–language models (VLMs) has further encouraged a zero-shot default that delegates classification to a frontier API.

Neither approach is straightforward to apply in a regulated deployment. Documents in finance and healthcare cannot leave the deployment environment, ruling out cloud-hosted LLM APIs. At cold start the customer's taxonomy is typically bespoke and the documents are confidential, so a labelled training set for fine-tuning is unavailable. After deployment, document templates drift, new subcategories appear, and human annotation is the dominant marginal cost. This setting is not hypothetical: the primary motivation for this work is a production trade-finance system classifying 30,233 documents across 80 categories, including many near-synonym categories, under strict data-residency requirements.

We propose **HIRA**, a training-free retrieval-augmented classification cascade designed for this regime. The pipeline is shown in Figure 1. HIRA fuses three retrieval signals (BM25 over OCR text, dense sentence embeddings, and image-level representations) using validation-calibrated weighted reciprocal-rank fusion, and routes only uncertain or visually confusable cases to a verifier LLM. The verifier receives the OCR text, the top retrieved exemplars, label descriptions, and distinguishing terms mined from past misclassifications, and produces a structured prediction with an ordinal confidence bucket. Low-confidence verifications enter a human-in-the-loop (HITL) queue, and each correction updates the system in two ways: as a margin-weighted entry in the retrieval index, and as an update to a Dirichlet-smoothed confusion graph that augments later verification prompts. No model parameters are updated at any point; only the retrieval memory and a lightweight confusion graph are mutable.

*Part of this work was conducted while the author was with OCBC.

CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840098

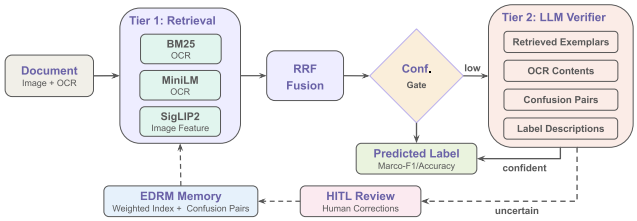

Figure 1: The proposed HIRA cascade framework.

On a private 80-class trade-finance corpus (Financial-80), HIRA processes the full 30,233-document production stream and requests human correction for only 1,945 documents (6.4%), improving Macro-F1 from 0.6218 to 0.8548. On the corrected Tobacco-3482 benchmark [12, 17], HIRA obtains Macro-F1 0.9423 with a locally hosted DeepSeek-R1-Distill-Qwen-32B verifier, 17.4 percentage points above the LLM zero-shot baseline at a 40% verifier call rate, and matches the fully labelled pool oracle (Section 6.3) after labelling 24.8% of the incoming stream.

Contributions. The paper makes three contributions:

- Cold-start adaptation under a small annotation budget. On Financial-80 (80 categories, 30,233 production documents), HIRA reaches Macro-F1 0.8548 with only 1,945 corrections (6.4% of the stream), and the queue rate falls to 6.2% as the index grows. On Tobacco-3482, 24.8% of the pool labelled is enough to match the fully labelled oracle.
- A training-free retrieval-augmented generation (RAG) cascade that exceeds LLM zero-shot. HIRA reaches Macro-F1 0.9423 on corrected Tobacco-3482, +17.4 points over a locally hosted DeepSeek-R1-Distill-Qwen-32B zero-shot baseline, at a verifier call rate of approximately 40%.
- EDRM (Evolving Document Relevance Memory), a dual-signal HITL memory. Each correction updates a margin-weighted retrieval index and a Dirichlet-smoothed confusion graph; the graph feeds distinguishing terms into the Tier-2 verifier prompt for confusion pairs.

2 Related Work

Existing research on document classification can be classified into two categories: (i) fine-tuned document-image classification methods, and (ii) retrieval-augmented and cascade inference approaches.

Fine-tuned document-image classification. Layout-aware and OCR-free document models such as LayoutLM [25], LayoutLMv2 [26], LayoutLMv3 [8], DiT [16], Donut [11], and UDOP [20] have become standard supervised baselines for document understanding. Two structural limitations apply to our setting: (i) they require a labelled training set, unavailable at cold start, and (ii) adapting to taxonomy changes requires a re-fine-tuning cycle that adds a GPU pass and a compliance review. HIRA differs in regime — it does not update weights, only the retrieval index — which avoids both.

Retrieval-augmented classification. Retrieval-Augmented Generation [14] and dense passage retrieval [9] showed that a non-parametric retrieval component can supplement learned parameters for knowledge-intensive tasks. For classification, kNN-LM [10] aggregates labels from nearest neighbours in a frozen-corpus index, kNN-Prompting [24] extends this to LLM in-context learning, Class-RAG [2] couples retrieval with an LLM verifier for content-moderation classification, and SRA [18] applies retrieval-augmentation to long-tail legal text classification. Our EDRM updates this index online from each correction: a margin-weighted exemplar plus a Dirichlet-smoothed confusion graph that feeds distinguishing terms back into the verifier's prompt for known confusion pairs.

Cascade inference and cost-aware routing. Cascading models of increasing capacity to balance cost and quality is a classical technique [6]. FrugalGPT [3] frames cascade routing as an optimisation problem and reduces LLM calls by deferring only uncertain queries to a stronger model. CALM [19] achieves similar cost reductions via early-exit confidence signals within a single model. These approaches assume either no retrieval or a fixed retrieval stage. In HIRA, the Tier-1 and Tier-2 thresholds are calibrated on a validation set, but the underlying retrieval index evolves as corrections are incorporated.

3 Problem Setting

Let $\mathcal{X}$ denote the document space and $\mathcal{Y} = \{1, \ldots, C\}$ a customer-specific label taxonomy. Given a document $x \in \mathcal{X}$ (scanned or digital PDF) in a regulated deployment — financial services, healthcare, legal discovery — the goal is to predict a class $y \in \mathcal{Y}$. Four constraints jointly preclude both supervised fine-tuning and cloud-LLM defaults:

(1) Cold start under drift. At deployment, only tens of labelled documents are available per class. Each customer has a custom taxonomy, and the data are long-tailed. As document templates change, a fixed classifier can gradually lose accuracy without being immediately noticed.

(2) Residency. Documents must not leave the deployment environment, excluding cloud LLM APIs as a runtime dependency. This constraint follows from data-residency and bank-secrecy regulations common to financial services and healthcare, and it applies to every stage of the pipeline, not only the final classifier: OCR, embedding, and verification must all run on infrastructure the customer controls.

(3) Compliance-bound retraining. In a regulated bank, retraining and redeploying a model triggers a model-governance review that adds significant process latency and nontrivial cost. Weight-update-based adaptation is rarely permissible at the cadence that drift requires. A governance review typically involves revalidating the model against a held-out compliance suite and obtaining sign-off from risk and legal stakeholders, a cycle that can take weeks even for a minor weight update.

(4) Audit trail and bounded annotation budget. Every decision must be explainable as a reference to similar documents already on file, which favours retrieval-style architectures over opaque end-to-end classifiers whose predictions cannot be traced back to a concrete precedent. HITL review is the slowest and most expensive step in the pipeline, and review capacity is fixed regardless of document

volume; the appropriate metric is therefore accuracy per human label rather than raw accuracy.

Retrieval-augmented classification fits naturally: only the retrieval memory and a lightweight confusion graph are updated from HITL corrections, without touching model weights.

4 The HIRA System

HIRA is a two-stage retrieval-augmented cascade with an HITL feedback loop. Figure 1 shows the architecture. All components are training-free post-deployment; the only state that changes over a deployment's lifetime is the retrieval index and a Dirichlet-smoothed confusion graph, both maintained by the Evolving Document Relevance Memory (EDRM) described in Section 4.3.

Figure 1 traces one document through the cascade: Tier 1 (Section 4.1) fuses three retrievers via weighted RRF and short-circuits when the top class clears τ_{t_1}; otherwise Tier 2 (Section 4.2) escalates to a locally hosted LLM verifier, which accepts a label when its confidence clears τ_{t_2} or else routes the document to the HITL queue. Every human correction feeds back into EDRM (Section 4.3), updating the retrieval index and confusion graph that inform future decisions, without ever updating model weights.

4.1 Tier 1: Multi-Modal Retrieval with RRF Fusion

Tier 1 runs three retrievers in parallel over the current index, each returning a top-k list:

- **BM25** over raw OCR text (PaddleOCR [15]).
- **MiniLM** embeddings over OCR (all-MiniLM-L6-v2 [22]).
- **SigLIP2** image embeddings [21].

The three ranked lists are fused via weighted reciprocal-rank fusion (RRF) [4]:

$$\text{score}(d) = \sum_{r \in \mathcal{R}} w_r \cdot \frac{1}{k_{\text{rrf}} + \text{rank}_r(d)} \tag{1}$$

where $\mathcal{R} = \{\text{BM25}, \text{MiniLM}, \text{SigLIP2}\}$ is the set of retrievers, $k_{\text{rrf}} = 10$ is the rank constant that dampens the contribution of lower-ranked results, and w_r are per-retriever weights calibrated per deployment by sweeping single-feature Macro-F1 on the validation split. RRF fusion improves on any single retriever, as reported in Section 6.

Per-class aggregation and the Tier-1 gate. Per-document scores are aggregated into per-class scores, with EDRM entry weights modulating each indexed exemplar's contribution. For class c with indexed exemplars $\mathcal{I}_c$,

$$S(c) = \sum_{d \in \mathcal{I}_c} w(d) \cdot \text{score}(d), \qquad s(c) = \frac{S(c)}{\sum_{c'} S(c')} \tag{2}$$

where $w(d)$ is the EDRM entry weight (Equation 5, $w(d) = 1$ for seed exemplars never corrected). Let $\hat{c}_1, \hat{c}_2$ denote the top-two classes ordered by $s(c)$; $\hat{c}_1$ is the top prediction.

Tier 1 accepts $\hat{c}_1$ when $s(\hat{c}_1) \geq \tau_{t_1}$ (Section 6.4); the cascade then short-circuits and returns $\hat{c}_1$. Cases that do not satisfy this condition are escalated to Tier 2, which supplies the LLM with additional context for verification.

4.2 Tier 2: LLM Contrastive Verification

Tier 2 invokes a verifier LLM on the document's OCR text, Tier-1's top-k retrieved exemplars $\mathcal{E}_1(d)$ (k=3 for Tobacco-3482, k=5 for Financial-80), and the category-description block. Let $\mathcal{L}_1(d)$ denote the set of labels appearing in $\mathcal{E}_1(d)$. When $(\hat{c}_1, \hat{c}_2)$ forms a high-confusion pair in $\mathcal{H}$ (Section 4.3), the prompt additionally includes distinguishing terms mined from prior corrections of that pair. The deployed verifier is a locally hosted DeepSeek-R1-Distill-Qwen-32B [5], constrained via guided generation to output a predicted label, an ordinal confidence bucket, and a free-text justification (Section 6.4).

The Tier-2 gate. Let $\tilde{c}$ and $\tilde{\text{conf}}$ denote the verifier's outputs. Tier 2 accepts when

$$\tilde{c} \in \mathcal{L}_1(d) \quad \text{and} \quad \tilde{\text{conf}} \geq \tau_{t_2}, \tag{3}$$

where τ_{t_2} is the verifier accept threshold (Section 6.4). Cases that fail either condition are routed to the HITL queue.

4.3 EDRM: Evolving Document Relevance Memory

Each HITL correction triggers two updates to EDRM.

(a) Time-decayed difficulty-aware exemplar insertion. For each correction (d, c_{true}), the true-class margin

$$m_{\text{true}} = s(c_{\text{true}}) - \max_{c \neq c_{\text{true}}} s(c) \tag{4}$$

measures how confidently Tier-1 ranked the correct class (negative when Tier-1 misclassified d). The correction is inserted into the retrieval index with weight

$$w(d) = \text{clip}(\exp(-m_{\text{true}}/\gamma),\ w_{\min},\ w_{\max}), \tag{5}$$

($\gamma = 0.1$, $w_{\min} = 0.1$, $w_{\max} = 10$), so misclassified and low-margin corrections receive larger weights than easy confirmations; clipping prevents a single confidently wrong correction from dominating the retrieval vote. Index entries decay multiplicatively as corrections accumulate; entries below $w_{\min}$ after decay are pruned. The decay time constant τ_{index} is a deployment parameter set against the customer's drift profile (Section 6.4).

(b) Dirichlet-smoothed confusion graph. EDRM also maintains a $K \times K$ confusion-count matrix C, where C_{ij} counts predictions of class i for true-label j. The Dirichlet-smoothed posterior

$$\hat{P}(c_{\text{true}} = j \mid \hat{c}_1 = i) = \frac{C_{ij} + \alpha}{\sum_k C_{ik} + \alpha K} \tag{6}$$

($\alpha = 0.1$) identifies high-confusion pairs (posterior $> \theta_{\text{conf}} = 0.15$), added to $\mathcal{H}$. For each such pair, the top distinguishing OCR tokens from past corrections are stored and injected into the Tier-2 prompt. Graph counts decay with half-life τ_{graph} to discard outdated confusion patterns.

4.4 HITL Loop and Frozen-State Evaluation

The protocol separates three concerns often conflated in HITL evaluation. (i) *Shared stream order*: all compared methods process the same documents in the same fixed order and mini-batch size. (ii) *Per-method HITL selection*: each method decides independently which documents require a human label, using its own uncertainty gate, so the methods consume the same arrival stream but build

different correction sequences — precisely what the annotation-cost comparison is intended to measure. (iii) *Shared frozen-state evaluation*: at each checkpoint of N corrections the resulting EDRM-augmented index is frozen and the full cascade (Tier 1 plus the real-LLM Tier 2) is re-run on the same held-out test set across methods.

5 Experimental Setup

5.1 Datasets

Financial-80, primary deployment corpus. A private 80-category trade-finance document corpus drawn from a production financial-services pipeline. The distribution is strongly skewed and a large fraction of categories are near-synonyms that require fine-grained disambiguation (e.g., multiple invoice and certificate subtypes). Under the data-sharing agreement we report only aggregate metrics (overall Macro-F1, head/medium/tail group F1, verifier call rate, HITL queue rate); per-class F1 tables and category names cannot be released. The class-count distribution, anonymised as C1…C80, is shown in Section 7.

Tobacco-3482 (corrected), public benchmark. The corrected Tobacco-3482 subset [17] re-annotates the original corpus [12] and removes 21% of multi-label or empty rows, yielding 2,737 documents (2,186 train / 272 val / 279 test). For the HITL-curve experiments, 100 seed documents (10 per class) initialise the index, 2,086 pool documents are streamed in a fixed random order, and 279 test documents are held out. Tier-1 RRF weights and τ_{t_1} are calibrated on the validation split. The ten categories (Email, Letter, Memo, Form, Report, Scientific, ADVE, Note, News, Resume) exhibit a roughly 5× head-to-tail ratio. OCR is produced with PaddleOCR [15].

5.2 Baselines

All baselines share HIRA's retrieval index, pool, and test split.

- bm25_only, minilm_only, siglip2_only: single-feature retrieval with top-1 voting.
- RRF fusion: Reciprocal Rank Fusion of the three retrievers above.
- kNN + Naive HITL: RRF with corrections injected at uniform weight 1.0, without EDRM.
- LLM zero-shot (DeepSeek): locally hosted DeepSeek-R1-Distill-Qwen-32B, prompted on OCR text with our system prompt and guided_json output.
- DiT fine-tuned (seed / HITL-augmented): Facebook Document Image Transformer [16] fine-tuned on the 100 seed documents (DiT-100), and separately on the same 100 seed documents plus the 518 HIRA-selected corrections (DiT-618, equal annotation budget to HIRA at N=518).

LLM verifier selection. We evaluated three locally hosted LLMs as Tier-2 verifier candidates on corrected Tobacco-3482 at N=0 (seed-only index): DeepSeek-R1-Distill-Qwen-32B [5] (Macro-F1 0.8819), Qwen-VL2.5 [1] (Macro-F1 0.8452), and Llama-3.1-8B [7] (Macro-F1 0.8233). DeepSeek outperformed both alternatives and was also faster per document on the same A100, so we use it as the deployed verifier in all remaining experiments.

Table 1: Static baselines on corrected Tobacco-3482 (seed-only index, 10 docs/class). HIRA at HITL convergence (N=518) is reported in Table 4.

Method	Macro-F1	Acc.	Head-F1	Med-F1	Tail-F1
BM25 only	0.4758	0.5161	0.623	0.365	0.477
MiniLM only	0.4418	0.4588	0.518	0.333	0.511
SigLIP2 only	0.7718	0.7670	0.731	0.789	0.790
RRF fusion	0.7736	0.7778	0.778	0.798	0.737
DS 0-shot	0.7683	0.8172	0.928	0.730	0.660
DiT-100†	0.8194	0.8065	0.839	0.841	0.971

†DiT fine-tuned on the same 100 seed documents.

Supervised fine-tuning scope. We include DiT fine-tuned at two label counts (DiT-100 and DiT-618) matched to HIRA's seed-only and N=518 checkpoints. Fully supervised fine-tuning on a training-set-sized corpus (e.g., LayoutLMv3 on the full training split) requires a labelled training set unavailable at cold start and triggers a model-governance review on each update, so we do not include it as a primary baseline.

5.3 Metrics and Protocol

The primary metric is Macro-F1, which weights all categories equally and is therefore appropriate for the imbalanced setting. We additionally report accuracy, head/medium/tail F1, HITL queue rate, verifier call rate, and Tier-1/Tier-2 invocation fractions as deployment-cost diagnostics. All HITL-curve numbers use the frozen-state, full-cascade protocol of Section 4.4.

Variance and seed protocol. HIRA's stochasticity is confined to the Tier-2 verifier. A three-run check on Financial-80 at N=1,945 yields Macro-F1 = 0.8523 / 0.8535 / 0.8586 (mean 0.8548, range 0.63 points), an order of magnitude below the smallest method gap; given this and the compute cost of full HITL trajectories, all main results use a single fixed seed.

6 Results

We report results across two corpora. Section 7 reports production deployment results on Financial-80 (80 categories, 30,233 documents); the present section reports controlled benchmark results on Tobacco-3482 (10 categories, 2,086 pool documents), which calibrate the V3 configuration transferred to Financial-80.

6.1 Static baselines (no HITL)

Table 1 reports baseline results on the corrected Tobacco-3482 split with a seed-only retrieval index (10 documents per class, 100 in total). With only 100 seed documents, RRF fusion (0.7736) already matches the DeepSeek zero-shot baseline (0.7683). DS zero-shot's 26.8-point Head–Tail F1 gap (versus 4.1 for RRF) shows why Macro-F1, not accuracy, is the appropriate metric here. DiT fine-tuned on the same 100 seed documents reaches 0.8194 Macro-F1, ahead of the retrieval-only baselines but −6.3 points behind the HIRA cascade at N=0 (0.8819).

Table 2: HIRA cascade vs. LLM zero-shot on corrected Tobacco-3482.

System	Macro-F1	Verifier call rate
DS 0-shot	0.7683	100%
RRF fusion	0.7736	0%
HIRA	0.9423	~40%
Δ(HIRA − DS 0-shot)	+17.4 pp	−60%

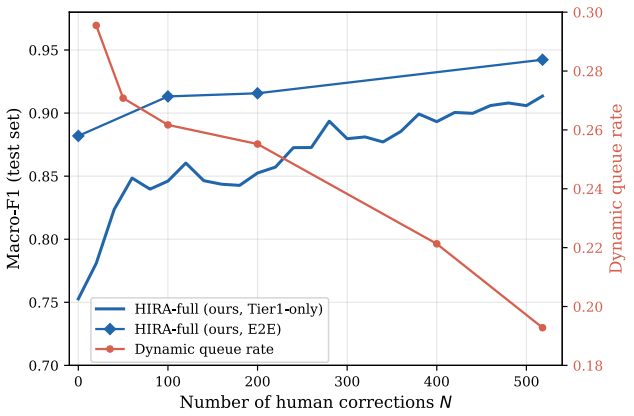

Figure 2: HIRA HITL improvement curve on corrected Tobacco-3482 (primary axis, green). Secondary axis (red): HITL dynamic queue rate.

6.2 HIRA cascade vs. LLM zero-shot

Table 2 summarises the cascade-versus-zero-shot comparison. With a locally hosted DeepSeek-R1-Distill-Qwen-32B verifier under the `enum5` configuration (Section 6.4), HIRA reaches Macro-F1 0.9423 after HITL convergence on a 2,086-document pool. This is +17.4 points above the DeepSeek zero-shot baseline (0.7683), at a verifier call rate of approximately 40% versus 100% for zero-shot. RRF fusion alone, with the 100-document seed index, already matches LLM zero-shot (+0.53 points); HITL enrichment and EDRM account for the remainder of the +17.4-point gain.

6.3 HITL gains and label cost

Figure 2 plots the test-set Macro-F1 of the frozen-state full cascade as a function of the number of HITL corrections N, under the default configuration ($\tau_{t_1} = 0.675$, $\tau_{t_2} = 0.95$, $\tau_{\text{index}} = 10^4$). Table 4 reports the same trajectory in tabular form, and Table 5 compares alternative configurations.

Annotation savings. At N=518, HIRA has labelled 24.8% of the 2,086-document incoming stream and reaches Macro-F1 0.9423, matching the fully labelled pool oracle, in which every pool document is inserted into the retrieval index with its ground-truth label at uniform weight 1.0. In other words, labelling less than a quarter of the incoming stream is sufficient to recover oracle performance.

Comparison with limited-label supervised fine-tuning. At equal annotation budget (618 labels total), DiT fine-tuned on the same documents reaches Macro-F1 0.9193, compared with HIRA's 0.9423 — a +2.3-point advantage with no weight update, training loop, or checkpoint management. DiT achieves higher tail-F1 (0.978 vs. 0.970), consistent with its visual pre-training aligning well with the distinctive layouts of tail categories (Note, News, Resume). HIRA leads on medium classes (0.926 vs. 0.915), where the confusion-graph mechanism provides gains that visual features alone cannot capture.

Table 3: EDRM contribution ablation on corrected Tobacco-3482 (Tier-1 only).

System	EDRM	Macro-F1	Acc.	Head	Med	Tail
kNN+Naive HITL	✗	0.8857	0.8823	0.875	0.854	0.936
HIRA	✓	0.9135	0.8961	0.871	0.907	0.964

EDRM versus naive correction injection. Table 3 isolates the EDRM contribution at Tier-1 (N=518, no Tier-2 verifier). Margin-weighted exemplar injection raises Macro-F1 from 0.8857 (naive, uniform weight) to 0.9135 (+2.78 points), with gains concentrated in medium classes (0.854⟶0.907). When the Tier-2 verifier is added, the full HIRA cascade reaches 0.9423 versus kNN+Naive HITL's 0.9135 (+2.88 points): the confusion graph populated by EDRM augments the verifier's distinguishing-term prompt, so the retrieval-level gain propagates and compounds through the verification stage.

Configuration ablation. Table 5 compares four operating points: V0 (aggressive decay, $\tau_{\text{index}} = 300$), V1 (low decay, $\tau_{t_2} = 0.9$), V2 (raises τ_{t_1} to 0.85), and V3 (the deployed configuration). V3 achieves the highest cascade-final Macro-F1 (0.9423) with 40% fewer corrections than V0 (518 versus 862).

Cascade efficiency. At the deployed V3 operating point, about 60% of test documents are short-circuited at Tier 1 and about 40% reach Tier 2. The dynamic queue rate, namely the fraction of the incoming document stream that ultimately enters the HITL queue, is 18.6% at the V3 checkpoint; the remaining cases are accepted by the verifier without human review. Figure 2 (secondary axis) shows that the queue rate decreases as the index matures, a self-reinforcing dynamic also observed on Financial-80 (Section 7). The verifier's net contribution shrinks correspondingly (+9.67 under V0, +2.88 under V3 in Table 5), so LLM dependence decreases with deployment age.

6.4 Cascade configuration: three key parameters

Three configuration parameters govern the quality-cost trade-off of the cascade, and each carries a non-obvious lesson.

τ_{t_1}: calibrate on validation rather than from a prior. The Tier-1 accept threshold was calibrated by sweeping $\tau_{t_1} \in [0.50, 0.90]$ on the corrected validation split and selecting the Pareto plateau on (Macro-F1, Tier-2 invocation rate); we obtain $\tau_{t_1} = 0.675$. Above $\tau_{t_1} = 0.85$ a false cliff appears: over-routing to Tier 2 incurs LLM cost without quality gain, because the verifier rarely overturns a borderline-confident Tier-1 prediction.

τ_{t_2} rather than τ_{t_1} is the queue-rate lever. Comparing V1 and V2 shows that tightening the Tier-1 gate from the default setting

Table 4: HITL learning trajectory with end-to-end evaluation on corrected Tobacco-3482.

N	Macro-F1	Acc.	Head	Med	Tail
DS 0-shot	0.7683	0.8172	0.928	0.730	0.660
0	0.8819	0.8925	0.935	0.859	0.859
100	0.9132	0.9176	0.952	0.875	0.953
200	0.9157	0.9211	0.936	0.892	0.948
518	0.9423	0.9355	0.937	0.926	0.970
DiT-618†	0.9193	0.9211	0.956	0.915	0.978
Oracle-2086*	0.9423	0.9462	0.957	0.930	0.944

*Oracle-2086 denotes a fully labelled retrieval oracle in which all 2,086 pool documents with ground-truth labels are inserted into the retrieval index with uniform weight 1.0.
†DiT fine-tuned on 100 seed + 518 HIRA-selected corrections.

Table 5: Four-way HITL configuration comparison on corrected Tobacco-3482.

Config (τ_{gate}, τ_{index})	Corr.	Ann. rate	T1-only	C-peak	C-final
V0: aggressive decay (0.9, 300)	862	41%	0.8337	0.9479	0.9304
V1: lower Tier-2 gate (0.9, 10k)	265	13%	0.8689	0.9248	0.9069
V2: tighter Tier-1 gate (0.85, 10k)	298	14.3%	0.8667	0.9080	0.9056
V3: deployed (0.95, 10k)	**518**	24.8%	0.9135	0.9423	0.9423

"T1-only": retrieval ceiling without verifier. "C-final": frozen-state full cascade at the listed N. "C-peak": maximum Macro-F1 across the trajectory (peak N: 400/200/200/518 for V0–V3). "Ann. rate": fraction of pool corrected (corrections/2,086). τ_{gate}: V0, V1, V3 use τ_{t_2}; V2 uses τ_{t_1} ($\tau_{t_2} = 0.9$).

changes the annotation rate only marginally (13% to 14.3%) and reduces cascade-peak Macro-F1. Raising τ_{t_2} from 0.9 to 0.95 (V1→V3) increases the queue rate to 18.6% and improves cascade-final Macro-F1 by 3.54 points: in a dense-index regime, it is the Tier-2 accept threshold rather than the Tier-1 gate that controls human-review volume. One practical prerequisite is that small open-weight LLMs collapse numeric confidence to the visible gate value (0.950); replacing it with an ordinal `enum5` enum [23] restores a usable spread, and the fix applies to any small-LLM cascade.

Index decay τ_{index} is deployment-specific. On a static benchmark a long decay ($\tau_{\text{index}} = 10^4$, config V3) dominates: the index is denser, Tier-1 short-circuits more often, and fewer corrections are needed. In production, where documents genuinely drift, a finite decay is necessary in order to flush stale exemplars. There is no universal best value; operators should set τ_{index} from the drift profile of their deployment. An alternative is calendar-time decay: entries older than a fixed period (e.g. six months) are expired regardless of correction volume, which is preferable when document templates change on a known schedule such as regulatory form revisions.

6.5 On-premises deployment configuration

The result in Table 2 is obtained with a locally hosted DeepSeek-R1-Distill-Qwen-32B verifier served by vLLM [13] on a single A100, accessed over an internal network. No document, retrieval snippet, or prompt leaves the deployment environment. A single A100 running vLLM is sufficient; the cascade reaches Macro-F1 above 0.94 with no document, snippet, or prompt leaving the environment, and at roughly 40% of the per-document LLM cost of zero-shot inference.

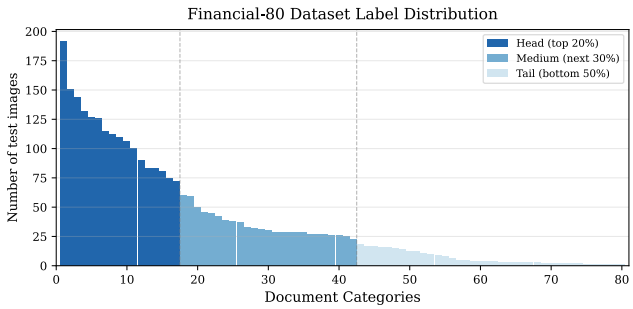

Figure 3: Class-count distribution of Financial-80 dataset.

7 Financial-80: Production Deployment Results

We validate HIRA on Financial-80, a private 80-category trade-finance document corpus drawn from a production financial-services pipeline. The setting is appreciably harder than Tobacco-3482: the category count is eight times larger, the distribution is more imbalanced, and many categories are near-synonyms (e.g., multiple invoice subtypes, multiple certificate variants) that require fine-grained disambiguation. Two documents from different near-synonym categories can share the majority of their visible text and differ only in a small set of boilerplate phrases or a single stamped clause, which is exactly the regime where retrieval alone is prone to voting for the wrong class.

7.1 Dataset and class-count distribution

Financial-80 contains 32,911 documents across 80 categories. For the HITL-curve experiments, 648 seed documents initialise the index, 316 documents form the validation split and 1,714 the test split, and the remaining 30,233 unlabelled documents form the pool. The taxonomy is confidential; under the data-sharing agreement, per-class F1 tables and category names cannot be released. We report aggregate metrics only (overall Macro-F1 and accuracy, head/medium/tail group F1, verifier call rate, and HITL queue rate).

Figure 3 shows the class-count distribution: strongly imbalanced with a steeper slope than Tobacco-3482. Head (ranks 1–17, 17 classes), Medium (ranks 18–42, 25 classes), and Tail (ranks 43–80, 38 classes) groups account for approximately 60%/30%/10% of volume.

7.2 Static baselines

Table 6 reports baselines (no HITL, seed-only index) on Financial-80 dataset. The same retrieval features and zero-shot LLM as Table 1 are used; the seed index is initialised with 648 seed documents (some long-tail categories have fewer than 10 samples).

Table 6 confirms that Financial-80 is harder than Tobacco-3482 (best baseline Macro-F1 0.622 vs. 0.774). RRF fusion improves over the best single retriever (Macro-F1 0.502 vs. 0.391), and the LLM head-bias is sharper on 80 categories: DS zero-shot achieves Head-F1 0.945 but Tail-F1 0.557, a 38.8-point Head–Tail gap (versus 26.8 on Tobacco). This gap matters operationally: a classifier that is excellent on the head categories but unreliable on the rest still

Table 6: Static baselines on Financial-80 dataset (seed-only index, 648 seed documents, ~8 docs/class).

Method	Macro-F1	Acc.	Head-F1	Med-F1	Tail-F1
BM25 only	0.351	0.484	0.517	0.435	0.307
MiniLM only	0.332	0.476	0.463	0.505	0.242
SigLIP2 only	0.391	0.554	0.681	0.514	0.316
RRF fusion	0.502	0.663	0.749	0.623	0.438
DS 0-shot	0.622	0.824	0.945	0.740	0.557

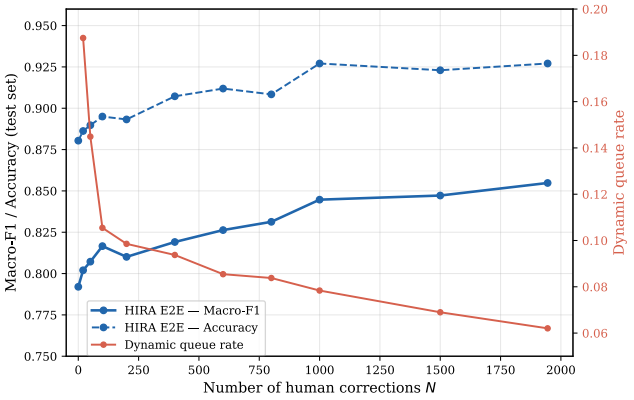

Figure 4: HIRA HITL improvement curve on Financial-80 (primary axis, green). Secondary axis (red): HITL dynamic queue rate.

Table 7: HIRA HITL trajectory on Financial-80 with end-to-end evaluation.

N	Macro-F1	Acc.	Head	Med	Tail
DS 0-shot	0.6218	0.8238	0.945	0.740	0.557
0	0.7920	0.8804	0.950	0.836	0.695
200	0.8101	0.8932	0.965	0.848	0.687
600	0.8263	0.9119	0.969	0.878	0.693
1000	0.8447	0.9271	0.982	0.895	0.667
1945	0.8548	0.9271	0.976	0.904	0.701

N denotes the number of human corrections. By the final checkpoint, the system has scanned the full 30,233-document pool and requested labels for 1,945 documents.

forces a human to check most tail-category documents, which is exactly the workload HIRA is designed to shrink.

7.3 HIRA cascade results

Figure 4 shows the HITL improvement trajectory on Financial-80 under the same V3 configuration ($\tau_{t_1} = 0.675$, $\tau_{t_2} = 0.95$, $\tau_{\text{index}} = 10^4$) used on Tobacco-3482. V3 transferred without any dataset-specific recalibration. Table 7 reports the trajectory in tabular form.

HITL efficiency. After HIRA processes the entire 30,233-document stream, only 1,945 documents (6.4%) are routed to human correction; at that final checkpoint HIRA reaches cascade-final Macro-F1 0.8548, an improvement of +23.3 points over the DS zero-shot baseline (0.6218). In absolute terms, 1,945 corrections took a single reviewer roughly six working days, well within the review team's existing capacity. The trajectory in Table 7 is still rising at the final evaluated checkpoint, suggesting further headroom as the index continues to mature under the more severe class imbalance.

EDRM and cascade contribution. Even at the modest checkpoint N=1,945, Tier-1-only retrieval (RRF fusion, no LLM verifier) reaches Macro-F1 0.7920 at N=0 while the full HIRA cascade reaches 0.8548 at N=1,945 — a +6.3-point gain from the Tier-2 verifier and EDRM exemplar injection. Near-synonym categories (multiple invoice and certificate subtypes) dominate the confusion clusters here, which the distinguishing-term mechanism is designed to address.

Dynamic queue rate. At N=1,945, **61.6%** of documents are accepted at Tier 1, **38.4%** reach the LLM verifier at Tier 2, and **6.2%** enter the HITL queue. This confirms that the verifier absorbs the majority of uncertain cases without human escalation. The secondary axis of Figure 4 shows that the dynamic queue rate, namely the fraction of incoming documents entering the HITL queue, decreases sharply as the system ingests more corrections, from approximately 18.8% at the start to about 6.2% at N=1,945. In practice, this mattered because review capacity — not GPU capacity — was the main operational bottleneck.

8 Limitations

Most baselines are training-free; the DiT baselines are included only as limited-label supervised references on Tobacco-3482. We do not include a fully supervised LayoutLMv3 baseline on Financial-80 because collecting a sufficiently labelled training set is not feasible under the deployment constraints.

9 Conclusion

We presented HIRA, a training-free, on-premises retrieval-augmented cascade for document classification in regulated industries, combining multimodal RRF retrieval, a locally hosted LLM verifier, and selective HITL correction stored as retrieval memory and confusion-graph updates. On Financial-80, HIRA reaches Macro-F1 0.8548 (from 0.6218) while routing only 6.4% of the 30,233-document stream to human correction; on corrected Tobacco-3482, it reaches Macro-F1 0.9423, +17.4 points above LLM zero-shot, matching the fully labelled pool oracle with 24.8% of the pool labelled. This combination of retrieval, a local verifier, and selective HITL was sufficient to ship without retraining the classifier.

GenAI Usage Disclosure

In this work, Generative AI software tools are utilized to edit and improve the quality of existing text, such as spelling, grammar, punctuation, clarity, and engagement, like a typing assistant.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] doi:10.48550/arXiv.2502.13923

[2] Jianfa Chen, Emily Shen, Trupti Bavalatti, Xiaowen Lin, Yongkai Wang, Shuming Hu, Harihar Subramanyam, Ksheeraj Sai Vepuri, Ming Jiang, Ji Qi, Li Chen, Nan Jiang, and Ankit Jain. 2024. Class-RAG: Real-Time Content Moderation with Retrieval Augmented Generation. *arXiv preprint arXiv:2410.14881* (2024).

[3] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176

[4] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 758–759. doi:10.1145/1571941.1572114

[5] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948

[6] David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontier, Henryk Michalewski, Maarten Bosma, Yuhuai Wu, Subhajit Roy, Charles Sutton, and Augustus Odena. 2022. Language Model Cascades. arXiv:2207.10342

[7] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783

[8] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*. ACM, 4083–4091. doi:10.1145/3503161.3548112

[9] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550

[10] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*.

[11] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In *Computer Vision – ECCV 2022*. Springer, 498–517. doi:10.1007/978-3-031-19815-1_29

[12] Jayant Kumar, Peng Ye, and David Doermann. 2012. Learning Document Structure for Retrieval and Classification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*. IEEE, 1558–1561.

[13] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. 611–626. doi:10.1145/3600006.3613165

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 9459–9474.

[15] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2022. PP-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System. arXiv:2206.03001

[16] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised Pre-training for Document Image Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*. ACM, 3530–3539. doi:10.1145/3503161.3547911

[17] Gordon Lim, Stefan Larson, and Kevin Leach. 2024. Label Errors in the Tobacco3482 Dataset. arXiv:2412.13140 [cs.CV] doi:10.48550/arXiv.2412.13140 WACV VisionDocs Workshop 2025.

[18] Boheng Mao. 2025. Exploring Selective Retrieval-Augmentation for Long-Tail Legal Text Classification. *arXiv preprint arXiv:2508.19997* (2025).

[19] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident Adaptive Language Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 17456–17472.

[20] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying Vision, Text, and Layout for Universal Document Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2212.02623.

[21] Michael Tschannen et al. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786

[22] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 5776–5788.

[23] Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. arXiv:2307.09702

[24] Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023. kNN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference. In *International Conference on Learning Representations (ICLR)*.

[25] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 1192–1200. doi:10.1145/3394486.3403172

[26] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*. Association for Computational Linguistics, 2579–2591. doi:10.18653/v1/2021.acl-long.201