

CARE: Bounded Selective Evidence-to-Judgment Routing for Cost-Aware Document Security Classification

Yeseul Chang*

Chung-Ang University
yes961002@cau.ac.kr

Woosang Lim*[†]

KAIST
woosang_lim@kaist.ac.kr

Rahul Kailasa

San Jose State University
rahul.kailasa@sjsu.edu

Simon Shim

San Jose State University
simon.shim@sjsu.edu

Jaewoo Lee[†]

Chung-Ang University
jaewoollee@cau.ac.kr

Abstract

Enterprise document security classification is a selective-decision problem. Rare high-risk labels require evidence and auditability, yet judging every document with an LLM is costly and adds model-specific variance to cases that precedents already settle. We introduce CARE (Cost-Aware Resource-adaptive Evidence), a bounded, fixed commit–verify–adjudicate controller that routes each case by precedent trust. It forms a zero-LLM evidence vote, commits when a learned reliability gate predicts low vote-error risk, verifies uncertain cases through verifier–evidence agreement, and spends heterogeneous panel judgment only on the residual conflicts where evidence and verification disagree. Across two contrasting evidence regimes that defeat different fixed baselines, the same controller selects route depth document by document. On WikiLeaks, where contrastive retrieval already supports strong performance without LLM calls, CARE preserves the zero-call evidence endpoint within bootstrap uncertainty while staying above every always-LLM decoder. It reaches 0.9824 Macro-F1 and 0.9624 SECRET F1, with only 44 of 2,863 documents reaching any LLM stage. This requires 55 invocations in total, 1.9% of a one-call-per-document budget. On DISC, where the rare TOP SECRET tail destabilizes sparse and LLM decoders, it gives the strongest row in our controlled comparison (0.8785 Macro-F1, 0.8444 TOP SECRET F1) using 496 physical model invocations against 789 for a same-evidence always-LLM policy, 63% of that budget, with only 81 documents reaching the panel. CARE thus turns security classification into an auditable policy for deciding when evidence is enough and when judgment is worth spending.

1 Introduction

Document security classification is a practical requirement for enterprise information governance,

including data-loss prevention, access control, and policy enforcement (Obama, 2009; Hart et al., 2011; Alzhrani et al., 2017). Unlike ordinary topical classification, a security label is not determined only by what a document is about, but by the potential impact of disclosure and its relationship to prior security decisions (Obama, 2009; Hart et al., 2011). The important errors are rare, asymmetric, and governance-sensitive: under-classifying a sensitive document can create disproportionate operational risk, while over-classification can burden downstream review and access workflows. For deployment, the goal is therefore not only to improve average accuracy, but also to recover high-risk labels while producing auditable decisions on rare boundary cases (Rudin, 2019; Ribeiro et al., 2016).

We study this allocation problem on two public security-classification corpora with complementary failure geometry. In WikiLeaks, high-risk SECRET labels often depend on contextual disclosure risk rather than explicit lexical markers (Hart et al., 2011; Alzhrani et al., 2017); in DISC, administrative markers make sparse and fine-tuned models competitive, but the rare TOP SECRET tail remains sensitive to decoder family and evidence source (Bass et al., 2024). Together they show why document security labeling is not well served by a single static predictor: lexical and supervised models can miss context, retrieval-free LLMs vary by decoder, evidence-only retrieval-augmented classification (RAC) can over-trust misleading precedent neighborhoods, and same-evidence always-LLM verification spends judgment on cases already settled by precedents. The deployment question is therefore when to commit to precedent evidence, when to verify it, and when to escalate.

Prior security-labeling systems use rules, sparse features, and supervised classifiers (Alzhrani et al., 2017; Bass et al., 2024); RAC grounds predictions in labeled precedents (Lewis et al., 2020; Karpukhin et al., 2020). General selective pre-

* Equal contribution.

[†] Corresponding authors.

diction, learning-to-defer, and LLM-routing methods allocate computation by confidence, expected model quality, cost, preference, or query difficulty (Chow, 1970; Geifman and El-Yaniv, 2017; Madras et al., 2018; Mozannar and Sontag, 2020; Chen et al., 2023; Ding et al., 2024; Ong et al., 2025), but they are not designed around the security-specific question of whether retrieved precedents should be trusted. CARE fills this gap with a fixed, bounded selective-routing commit-verify-adjudicate controller whose auditable state is precedent trust: retrieved labels, reranker scores, inter-precedent agreement, verifier-evidence agreement, and unresolved evidence-verifier conflict. It first forms a zero-LLM evidence vote, commits when a learned reliability gate trusts the precedent neighborhood, verifies only uncertain cases, and sends remaining conflicts to a heterogeneous panel. It is bounded in a strict sense: predefined routing over a fixed action space with logged decisions, not open-ended tool use or online policy learning. The result is a single policy that exits cheaply when precedents are decisive, as in WikiLeaks, and spends targeted judgment when the high-risk tail is unstable, as in DISC.

CARE addresses this allocation problem as a bounded selective-routing controller for document-level security-tier classification. For each document it builds an auditable precedent state from retrieved and reranked labeled examples, and it records the stage at which the case was settled. CARE is therefore an evidence-to-judgment control policy that decides when precedent trust is enough, when one verifier is enough, and when residual boundary cases require model diversity. This framing yields three contributions:

1. **A failure-mode analysis of high-risk security classification.** We identify two complementary deployment regimes: WikiLeaks becomes largely precedent-saturated once contrastive evidence is retrieved, while DISC remains unstable in the rare TOP SECRET tail because sparse models, evidence votes, and LLM decoders fail differently. This reframes security classification as trust allocation rather than model selection.
2. **An auditable commit-verify-adjudicate controller.** We formulate classification as resource-adaptive allocation among precedent evidence, independent verification, and residual heterogeneous judgment. CARE implements this policy with zero-LLM evidence commitment, a reliability gate trained on evidence-vote error instead

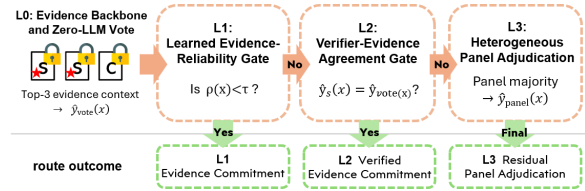


Figure 1: CARE routing policy. The controller commits reliable evidence-backed cases, verifies uncertain cases through verifier-evidence agreement, and adjudicates unresolved disagreements with heterogeneous verifier judgment.

of decoder confidence, verifier-evidence agreement, and panel escalation, producing both a label and an auditable route ledger.

3. **Evidence that one policy adapts across regimes.** On WikiLeaks, CARE mostly exits through reliable evidence and spends judgment only on the residual tail, where it is statistically at parity with the near-ceiling evidence endpoint; on DISC, it selectively verifies and adjudicates the cases where evidence trust breaks and lifts the high-risk tail. Paired bootstrap shows gains over evidence-only and same-evidence always-LLM baselines, the latter beaten at 63% of its one-call-per-document budget, and cost at parity against a uniform panel at 79% fewer invocations.

2 CARE Method

Let x denote a document and y its security label. Every document arrives with retrieved and reranked labeled precedents, and CARE decides how much of the decision those precedents can carry, how much needs a verifier, and how much needs a panel.

CARE is a bounded evidence-to-judgment allocation policy implemented as four stages. $L0$ produces a zero-LLM evidence vote from retrieved precedents. $L1$ estimates whether the evidence neighborhood is reliable enough to support direct commitment. $L2$ applies an independent verifier and commits only when verifier and evidence agree. $L3$ resolves the remaining disagreements through heterogeneous panel judgment. The output is both a security label and an auditable route record holding the stage at which the case exited, the precedents behind the evidence vote, and any verifier or panel labels the case collected on the way.

2.1 L0: Evidence Backbone and Zero-LLM Vote

For each document x , a shared backbone retrieves a candidate pool from the precedent store and reranks it, returning the top M as the ordered evidence context:

$$\mathcal{E}_{1:M}(x) = ((d^{(i)}, y^{(i)}, s^{(i)}))_{i=1}^M,$$

where M is the post-rerank depth, $d^{(i)}$ is the i -th precedent, $y^{(i)}$ its security label, and $s^{(i)}$ its reranker score. CARE forms a zero-LLM evidence candidate by majority vote over the top $T \leq M$ precedent labels,

$$\hat{y}_{\text{vote}}(x) = \text{Vote}(y^{(1)}, \dots, y^{(T)}),$$

falling back to the top-precedent label $y^{(1)}$ when no label holds a majority; the runs here use $T=3$ and $T=1$, where the vote is that label. The depth M and the vote width T are backbone hyperparameters, not controller ones. $L1$ routes on the reliability of whichever candidate they produce, so retuning them changes the evidence signal, not the route.

2.2 L1: Learned Evidence-Reliability Gate

After $L0$ produces an evidence vote, $L1$ asks whether the retrieved precedent neighborhood can serve as a decision authority rather than merely as context for a decoder. Its target is evidence-vote error risk, distinct from document difficulty, decoder confidence, and model-selection scores. Even a high-scoring, label-consistent neighborhood can be a misleading precedent cluster for a rare high-risk boundary case.

We train R_θ on evidence-only features $\phi(x)$, including the retrieved labels $y^{(1)}, \dots, y^{(M)}$, dense and reranker-score geometry, vote support, inter-precedent agreement, ordinal label patterns, and class-specific interactions. The supervised target is the evidence-vote error indicator:

$$z(x) = \mathbb{I}[\hat{y}_{\text{vote}}(x) \neq y],$$

$$\rho(x) = R_\theta(\phi(x)) \approx \Pr[z(x) = 1 \mid \phi(x)].$$

Thus $\rho(x)$ is calibrated evidence-commit risk, not model confidence. A held-out calibration split fixes the operating threshold τ before test replay. Documents with $\rho(x) < \tau$ commit $\hat{y}_{\text{vote}}(x)$ at zero LLM cost; documents with $\rho(x) \geq \tau$ enter the verifier queue $\mathcal{Q}_{\text{verify}}$. $L1$ therefore implements coverage-risk allocation: precedent-settled neighborhoods

exit cheaply, while suspicious neighborhoods become a compact, auditable queue for judgment rather than an indiscriminate LLM-call stream.

2.3 L2: Verifier-Evidence Agreement Gate

$L2$ turns the post- $L1$ queue into a reproducibility test. It asks whether the same label follows independently from the same document and evidence context $\mathcal{E}_{1:M}(x)$, so one small language model (sLLM) verifier confirms the evidence without the authority to overwrite it. Agreement converts a risky evidence vote into a verified evidence decision. Disagreement marks a genuine evidence-model conflict.

For each document in $\mathcal{Q}_{\text{verify}}$, CARE invokes an sLLM verifier

$$\hat{y}_s(x) = V_s(x, \mathcal{E}_{1:M}(x))$$

using the same document-and-evidence prompt format as the always-LLM baselines. The controller compares this verifier prediction with the evidence vote:

$$g_{\text{agree}}(x) = \mathbb{I}[\hat{y}_s(x) = \hat{y}_{\text{vote}}(x)].$$

If $g_{\text{agree}}(x) = 1$, CARE commits the shared label. If the sLLM and evidence vote disagree, the case enters the residual escalation queue $\mathcal{Q}_{\text{esc}}$. The role of $L2$ is confirmation and panel-exposure control. One targeted sLLM call settles uncertain but reproducible evidence cases, and unreproduced cases go to $L3$.

2.4 L3: Heterogeneous Panel Adjudication

The residual queue $\mathcal{Q}_{\text{esc}}$ contains cases where evidence alone is too risky and a single sLLM view cannot reproduce the evidence vote. These are the cases in which model-specific variance is most useful, as adjudication for unresolved evidence-model conflict.

CARE adjudicates $\mathcal{Q}_{\text{esc}}$ with a fixed panel of K verifier views $\{V_1, \dots, V_K\}$, where K counts adjudicating models rather than precedents. Each view receives the same document and evidence context $\mathcal{E}_{1:M}(x)$, and the final label is the panel majority:

$$\hat{y}_{\text{panel}}(x) = \text{Vote}_{1 \leq k \leq K} V_k(x, \mathcal{E}_{1:M}(x)).$$

When the panel has no strict majority, the implementation falls back to the highest-priority configured member, preserving deterministic replay. In

the main policy, all panel views share the same retrieved precedents, so diversity changes the judgment view rather than the evidence basis. K is a deployment parameter matched to the residual load: DISC leaves 81 unresolved conflicts and uses $K=3$ heterogeneous views, while WikiLeaks leaves only 11 and uses $K=1$, a single frontier adjudicator, because where evidence is already near the ceiling panel diversity has little to arbitrate.

3 Experimental Setup

3.1 Datasets and Leakage Control

We evaluate on two public security-classification corpora. WikiLeaks is a U.S. diplomatic-cables task (Hart et al., 2011; Alzhrani et al., 2017) with labels UNCLASSIFIED, CONFIDENTIAL, and SECRET; its locked split has 6,033 training precedents and $N = 2,863$ test documents, including 202 SECRET cases. DISC is a declassified U.S. national-security task (Bass et al., 2024) with labels UNCLASSIFIED, SECRET, and TOP SECRET; its locked split has 1,312 training precedents and $N = 789$ test documents, including 48 TOP SECRET cases. We treat SECRET in WikiLeaks and TOP SECRET in DISC as high-risk labels, covering a contextual SECRET regime and a smaller, marker-sensitive TOP SECRET tail.

Evidence stores contain only training precedents. We remove exact, normalized-exact, and high-overlap near-duplicate train-test matches, following the deduplication motivation of Lee et al. (2022). We set the operating cutoff ourselves at Jaccard similarity 0.85 or above over character 5-shingles, a conservative choice for templated government documents, and Appendix Table 5 shows that relocating it leaves the WikiLeaks evidence endpoint essentially unchanged. Test labels are used only for final scoring, never for retrieval, evidence construction, routing features, verifier prompts, threshold selection, or model selection.

3.2 Compared Systems and Shared Evidence

We compare TF-IDF sparse classifiers, BERT-base fine-tuning (Devlin et al., 2019), document-only LLMs, RAC/ConRAC evidence-only endpoints, RAG-style RAC/ConRAC always-LLM endpoints, and CARE. RAC predicts from the top precedent ($T=1$) or the top-three vote ($T=3$), the Top-1 and Top-3 rows below. ConRAC adds contrastive retrieval before the same evidence-only decision. Always-LLM endpoints pass every document and

its top- M precedents to a decoder. CARE receives the same context but routes each case to evidence commitment, verifier agreement, or adjudication by a K -model panel. Thus evidence-only, always-LLM, and CARE differ by decision policy over a shared evidence state, not by evidence access.

All evidence-based systems use BGE-M3 retrieval and reranking (Chen et al., 2024), retrieving 20 candidates as a high-recall pool and exposing the top $M=3$ as the shared auditable context. Three widths recur through the paper, M precedents, T voted labels and K panel models.

3.3 Implementation, Fine-Tuning, and Reporting

Verifier and panel candidates are open-weight Qwen, Gemma and Llama-family decoders with fixed decoding, deterministic label normalization, and a shared document-and-evidence prompt within each suite. Open-weight rows can be locally or enterprise hosted. Fixed CARE policies use the $L0$ – $L3$ controller from Sections 2.1–2.4 with the same structure on both corpora, differing only in width and endpoint. $L0$ keeps $M=3$ precedents after reranking on both, and commits at vote width $T=3$ on DISC and $T=1$ on WikiLeaks. $L1$ is a low-capacity MLP router, hidden widths 200–200–40 on DISC (a fixed 12-member seed ensemble) and 64–16 on WikiLeaks, trained on training and calibration evidence features and thresholded before test replay. $L2$ applies a fine-tuned Gemma-4-E4B (Google DeepMind, 2026) verifier on DISC and Llama-3.1-8B (Grattafiori and Llama Team, 2024) on WikiLeaks. $L3$ runs a $K=3$ panel on DISC with the roster Qwen3-8B (Qwen Team, 2025b,a) (fine-tuned), Qwen3.5-9B (Qwen Team, 2026) (base) and Gemma-4-E4B (fine-tuned), and a single frontier adjudicator on WikiLeaks (Gemini-3-Flash (Google DeepMind, 2025), $K=1$). WikiLeaks uses one of each width, $T=1$ and $K=1$. It tolerates $T=1$ because contrastive retrieval has already settled the neighborhoods the gate commits. The two widths disagree on 21 of the 2,863 documents, 20 of them already routed to judgment and one among the 2,819 commits.

Fine-tuning is not a test-time routing change. BERT-base FT is a supervised baseline, ConRAC FT rows are fine-tuned evidence endpoints, and FT-mixed rows mix fine-tuned and base endpoints inside the same fixed CARE route, leaving the evidence backbone, the $L1$ gate and its threshold, and the $L2/L3$ stopping rules untouched. Selected

CARE rows report frontier operating points. Fixed rows apply one preconfigured policy to the full split and support judge-call claims.

We report accuracy, Macro-F1, per-class F1, and high-risk precision/recall/F1, and we separate three cost quantities. *Unique LLM-exposed documents* counts documents that reach any LLM stage; *logical route-stage entries* count one entry for reaching $L2$ and one additional entry for reaching $L3$; *physical model invocations* count executed LLM calls, so a document escalated to the K -model panel costs K invocations rather than one. The two bases separate only where a panel runs. DISC multiplies its $L3$ documents by $K=3$, while $K=1$ on WikiLeaks makes its two bases coincide. Same-evidence always-LLM rows use N physical invocations, and zero-call rows use retrieval and reranking but no LLM verifier. Reporting both bases matters because they can disagree about which design is cheaper: a policy that visits fewer stages may still execute more models, and only the physical count is what a deployment pays. All efficiency claims here are stated in invocation counts.

4 Results and Analysis

4.1 DISC: Evidence-to-Judgment Routing Under High-Risk Differentiation

The primary DISC test is high-risk differentiation (Table 1). The comparison is deliberately strong. Supervised baselines are effective on aggregate, yet the rare TOP SECRET boundary stays brittle for every family, and same-evidence always-LLM decoding does not uniformly help. Several same-evidence LLM rows fall below the zero-LLM ConRAC Top-3 endpoint on key metrics, while the strongest same-evidence row still trails CARE. The vote width matters here. Widening it to $T=3$ lifts TOP SECRET F1 from 0.6250 to 0.6607 on ConRAC evidence and from 0.4272 to 0.5532 on weaker RAC evidence, so the vote is worth most where precedents are contested. LLM judgment helps in the same place, which is what the route is built to detect.

The fixed CARE policy reaches 0.8785 Macro-F1 and 0.8444 TOP SECRET F1 on the $N = 789$ split.¹ Of the 789 documents, 536 commit from

evidence and 253 reach an LLM. All 253 enter the $L2$ verifier, 172 stop there on verifier-evidence agreement, and 81 residual conflicts reach the $K=3$ panel. That is 334 logical route-stage entries and 496 physical model invocations ($253 + 81 \times 3$), which is 62.9% of the 789 invocations the strongest same-evidence always-LLM row spends, a 37.1% reduction. Paired bootstrap on the same split gives $+0.0211$ Macro-F1 [$+0.0050, +0.0403$] and $+0.0487$ TOP SECRET F1 [$+0.0100, +0.0968$] over that row, and $+0.0893$ Macro-F1 [$+0.0591, +0.1223$] over the zero-call ConRAC Top-3 evidence-vote control. All three intervals exclude zero. Under fully physical accounting CARE is therefore both cheaper and better than the strongest policy that sees the same evidence ($0.8785 > 0.8573$ Macro-F1, $0.8444 > 0.7957$ TOP SECRET F1).

Table 2 is a route-flow table. Every policy labels the same 789 documents, and the count columns show where model judgment is spent. Evidence-only prediction is free but weak on the TOP SECRET tail, and every uniform policy stays at or below Full CARE, whether verify-all at 789 invocations, panel-all at 2,367 or no- $L1$ at 1,107. Against the uniform panel, the deployment-relevant control, the same panel obtains 0.8702 applied to all 789 documents and 0.8785 applied only to the $L1$ -selected residuals ($+0.0082$ Macro-F1 [$-0.0001, +0.0207$]), so the deployment result is *cost at parity*: the same panel quality for 21% of the panel’s invocations, 79% fewer. Removing $L2$ is the complement. Sending all 253 residuals straight to the panel spends 759 invocations for the same quality, 0.8794 Macro-F1 against 0.8785 with identical TOP SECRET F1, a gap far below the across-seed SD of about 0.003. $L2$ therefore buys 263 panel invocations for one verifier call each, settling 172 of the 253 residuals.

Together, these controls give each stage a distinct deployment role. $L1$ supplies the main coverage-risk cut; $L2$ separates sLLM-confirmed evidence from residual conflict; and $L3$ is spent only where evidence and the single verifier do not agree. Appendix diagnostics show the safety profile of that placement. Among the 536 evidence commits, 16 are wrong, and the two TOP SECRET misses both land one level down on SECRET and none reaches UNCLASSIFIED, so none of the 16 crosses two label levels, while a broad commit-coverage plateau

¹These DISC numbers supersede the reviewed version. Re-verifying $L1$ after review, we found one gold-derived column in its training features with no inference-time counterpart. We removed it, re-ran the complete route under the primary roster and the same development-calibrated operating point, and report that rerun throughout: 0.8785 Macro-F1 against

0.8896 reviewed, TOP SECRET F1 unchanged at 0.8444, and 0.8740 / 0.8444 at the reviewed commit volume of 669.

Table 1: Full DISC test comparison ($N=789$). The main CARE row is the fixed primary policy after the $L1$ correction of Section 4. The judge-call column reports both cost bases; they differ only for CARE, whose $L3$ runs $K=3$ models per adjudicated document. The final column gives physical invocations as a share of the always-LLM budget of one per test document ($789 = 100\%$).

Method	Acc	Macro-F1	F1 (Unclass.)	F1 (Secret)	F1 (TOP SECRET)	Prec. (TOP SECRET)	Rec. (TOP SECRET)	Judge calls (logical/physical)	% phys. budget
TF-IDF + NB	0.8074	0.6002	0.8007	0.8460	0.1538	1.0000	0.0833	0 / 0	0
TF-IDF + SVM	0.8695	0.7828	0.8790	0.8861	0.5833	0.8750	0.4375	0 / 0	0
TF-IDF + KNN	0.8416	0.7712	0.8511	0.8604	0.6022	0.6222	0.5833	0 / 0	0
TF-IDF + GB	0.8758	0.8468	0.8724	0.8873	0.7805	0.9412	0.6667	0 / 0	0
BERT-base FT	0.8847	0.8448	0.8906	0.8962	0.7475	0.7255	0.7708	0 / 0	0
LLM-only Llama-3.1-8B	0.6160	0.5353	0.3712	0.7205	0.5143	0.8182	0.3750	789 / 789	100
LLM-only Qwen3-8B	0.7858	0.7830	0.7878	0.7848	0.7765	0.8919	0.6875	789 / 789	100
LLM-only Qwen3.5-9B	0.6781	0.6762	0.7258	0.6081	0.6947	0.7021	0.6875	789 / 789	100
LLM-only Gemma-4-E4B	0.7706	0.7651	0.7826	0.7594	0.7532	1.0000	0.6042	789 / 789	100
RAC Top-1	0.7440	0.6512	0.7479	0.7785	0.4272	0.4000	0.4583	0 / 0	0
RAC Top-3 vote	0.7909	0.7202	0.7914	0.8159	0.5532	0.5652	0.5417	0 / 0	0
ConRAC Top-1	0.8479	0.7843	0.8581	0.8698	0.6250	0.5469	0.7292	0 / 0	0
ConRAC Top-3	0.8555	0.7994	0.8600	0.8776	0.6607	0.5781	0.7708	0 / 0	0
ConRAC always-Llama-3.1-8B	0.7896	0.6808	0.7643	0.8264	0.4516	1.0000	0.2917	789 / 789	100
ConRAC always-Qwen3.5-9B	0.7668	0.7590	0.7878	0.7464	0.7429	0.6842	0.8125	789 / 789	100
ConRAC always-Gemma-4-E4B	0.8517	0.8317	0.8527	0.8575	0.7848	1.0000	0.6458	789 / 789	100
ConRAC always-Qwen3-8B	0.8834	0.8573	0.8824	0.8939	0.7957	0.8222	0.7708	789 / 789	100
CARE	0.8935	0.8785	0.8875	0.9035	0.8444	0.9048	0.7917	334 / 496	63

Table 2: DISC controller-policy controls under both endpoint rosters, with all three cost bases. Every policy labels the same 789 documents, and only the controller changes. The evidence-vote row is the zero-call vote re-executed from the frozen retrieval artifact the router consumes (0.7892); the ConRAC Top-3 row of Table 1 (0.7994) scores the same endpoint in the main comparison pipeline, and every paired delta in this paper uses the re-executed value so that the compared policies share one evidence state.

Policy	Controller change	Unique docs	Logical entries	Physical invocations	% phys. budget	Acc	Macro-F1	F1 (TOP SECRET)
<i>Open-weight endpoint stack (reviewed-version roster)</i>								
Evidence vote only	no verifier	0	0	0	0	.8568	.7892	.6195
Always $L2$ verifier	verify all once	789	789	789	100	.8517	.8317	.7848
Always $L3$ panel	panel all documents	789	789	2,367	300	.8580	.8474	.8222
No $L1$: verify all	remove reliability gate	789	916	1,170	148	.8580	.8474	.8222
CARE (open-weight)	$L1$ route + $L2$ agree + $L3$ panel	261	363	567	72	.8745	.8655	.8444
<i>Primary FT-mixed endpoint stack</i>								
Always $L2$ verifier	verify all once	789	789	789	100	.8771	.8530	.8000
Always $L3$ panel	panel all documents	789	789	2,367	300	.8872	.8702	.8315
No $L1$: verify all	remove reliability gate	789	895	1,107	140	.8859	.8693	.8315
Skip $L2$: $L1$ route to panel	remove agreement gate	253	253	759	96	.8948	.8794	.8444
Full CARE	$L1$ route + $L2$ agree + $L3$ panel	253	334	496	63	.8935	.8785	.8444

holds TOP SECRET F1 steady. What the route buys is placement. Each document stops at the cheapest stage that can settle it, and the panel is reserved for the cases that need it.

4.2 WikiLeaks: Saturated Semantic Evidence and Residual Routing

Where DISC stresses an unstable high-risk tail, WikiLeaks is the saturated-evidence regime, and Table 3 shows why. Surface and document-only methods struggle with contextual SECRET decisions. The best TF-IDF row reaches only 0.5350 SECRET F1, and Llama-3.1-8B obtains high recall only by collapsing SECRET precision to 0.1469. In contrast, ConRAC Top-3 finds high-quality prece-

dents and already reaches 0.9791 Macro-F1 and 0.9576 SECRET F1 with zero judge calls. But uniform LLM judgment is not a monotone remedy. Every same-evidence always-LLM row trails this zero-call endpoint while spending all 2,863 calls, so the useful question here is not how to judge better but when to judge at all.

CARE reaches the highest accuracy, Macro-F1 and SECRET F1 in the table (0.9913, 0.9824, 0.9624) while only 44 of 2,863 documents, about 1.5%, reach any LLM stage. The difference from the ConRAC Top-3 endpoint is within bootstrap uncertainty (+0.0033 Macro-F1 $[-0.0001, +0.0073]$; +0.0048 SECRET F1 $[-0.0048, +0.0144]$), so we read WikiLeaks as a

Table 3: Full WikiLeaks test comparison ($N=2,863$; minority class SECRET). The CARE route uses one sLLM verifier and a single frontier adjudicator ($K=1$), so the two cost bases coincide at 55 (44 verifier + 11 frontier), 1.9% of the always-LLM budget of 2,863.

Method	Acc	Macro-F1	F1 (Unclass.)	F1 (Conf.)	F1 (SECRET)	Prec. (SECRET)	Rec. (SECRET)	Judge calls (logical/physical)	% phys. budget
TF-IDF + NB	0.7859	0.6196	0.8544	0.7383	0.2661	1.0000	0.1535	0 / 0	0
TF-IDF + SVM	0.8861	0.7595	0.9429	0.8442	0.4912	0.8434	0.3465	0 / 0	0
TF-IDF + KNN	0.7894	0.6811	0.8613	0.7246	0.4575	0.5612	0.3861	0 / 0	0
TF-IDF + GB	0.8931	0.7788	0.9504	0.8511	0.5350	0.6929	0.4356	0 / 0	0
BERT-base FT	0.9773	0.9470	0.9944	0.9671	0.8796	0.9333	0.8317	0 / 0	0
LLM-only Llama-3.1-8B	0.5054	0.4420	0.7350	0.3363	0.2548	0.1469	0.9604	2,863 / 2,863	100
LLM-only Qwen3-8B	0.8295	0.7313	0.8901	0.7832	0.5205	0.8444	0.3762	2,863 / 2,863	100
LLM-only Qwen3.5-9B	0.8816	0.8427	0.9215	0.8253	0.7813	0.8242	0.7426	2,863 / 2,863	100
RAC Top-1	0.7761	0.6729	0.8602	0.7029	0.4556	0.4219	0.4950	0 / 0	0
RAC Top-3 vote	0.7855	0.6787	0.8620	0.7173	0.4567	0.4860	0.4307	0 / 0	0
ConRAC Top-1	0.9843	0.9698	0.9929	0.9789	0.9377	0.9447	0.9307	0 / 0	0
ConRAC Top-3 vote	0.9888	0.9791	0.9950	0.9846	0.9576	0.9648	0.9505	0 / 0	0
ConRAC always-Gemini-3-Flash-Preview	0.9742	0.9595	0.9844	0.9664	0.9277	0.9347	0.9208	2,863 / 2,863	100
ConRAC always-Llama-3.1-8B	0.9689	0.9378	0.9828	0.9686	0.8621	0.8578	0.8663	2,863 / 2,863	100
ConRAC always-Qwen3-8B	0.9511	0.8933	0.9776	0.9437	0.7586	0.8171	0.7079	2,863 / 2,863	100
ConRAC always-Qwen3.5-9B	0.9679	0.9333	0.9859	0.9567	0.8571	0.9871	0.7574	2,863 / 2,863	100
ConRAC always-Gemma-4-E4B	0.9012	0.8720	0.9274	0.8791	0.8097	0.8830	0.7475	2,863 / 2,863	100
CARE	0.9913	0.9824	0.9964	0.9882	0.9624	0.9746	0.9505	55 / 55	1.9

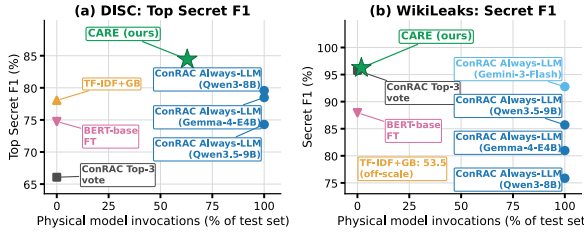


Figure 2: High-risk F1 versus physical model invocations. Panel width is counted: the DISC panel runs $K=3$ models per adjudicated document, WikiLeaks uses a single adjudicator ($K=1$), and uniform always-LLM rows cost one invocation per document.

retrieval-saturated quality-parity regime. The route reflects that. Of the 2,863 documents, 2,819 commit from evidence, 44 reach the verifier, 33 stop there on agreement, and 11 are adjudicated. $L3$ here is a single adjudicator ($K=1$), so those documents cost $44 + 11 = 55$ invocations under both accounting bases, 1.9% of the always-LLM budget.

Against the strongest same-evidence always-LLM row the margin is $+0.0229$ Macro-F1 and $+0.0347$ SECRET F1 at 98.1% fewer invocations. The cross-regime reading is stronger. The same ConRAC Top-3 evidence that saturates here leaves 0.6607 TOP SECRET F1 on DISC, where the route raises it to 0.8444.

4.3 Core Cost–Quality Comparison

Figure 2 puts the two regimes on one cost–quality axis. On DISC CARE beats evidence-only prediction and every same-evidence single-LLM baseline

while matching the uniform panel, and on WikiLeaks the zero-call endpoint is already near the ceiling and CARE preserves it. Efficiency is a property of placement, and one policy holds three things at once. It stays under a one-call-per-document budget in both regimes (63% on DISC, 1.9% on WikiLeaks), reaches the best high-risk F1 in both, and stays stable without retuning, since the primary gate is a fixed seed ensemble and a broad commitment-coverage plateau holds TOP SECRET F1 steady.

5 Conclusion

We presented CARE, an evidence-to-judgment routing policy for rare high-risk document security labels. Judgment is worth spending only where precedent trust runs out, and that one principle carries both regimes. The same fixed controller preserves a near-ceiling evidence endpoint on WikiLeaks and lifts the unstable TOP SECRET tail on DISC to 0.8444 F1, both under a one-call-per-document budget, beating the strongest same-evidence always-LLM policy and matching a uniform panel that spends 4.8 times more. Because CARE is a fixed policy over a shared evidence state rather than a new model, it can sit in an existing pipeline as a budget layer that decides where judgment is spent and makes the budget an operator’s parameter. We expect that property to carry to any agent pipeline where a cheap signal usually suffices and judgment must be reserved for what changes it.

Limitations

Our empirical claims cover two public document security corpora on one fixed evidence backbone. What the controller needs is not specific to documents. Given a labeled store of past decisions, a cheap candidate derived from it, held-out data for learning when that candidate fails, a costlier verifier and asymmetric error costs, the same commit–verify–adjudicate policy applies. Two settings meet those conditions and are where we would take it next. One is agent security, where prompts, queries and tool calls raise the same rare high-cost decisions under a per-call budget, and where the store of adjudicated cases is something a running system already accumulates. The other is multi-modal review, where the store holds images or audio rather than text, which is also the first step toward making the same budgeted decision in embodied systems. Neither is demonstrated here, because each needs its own store and its own calibration.

Two dependencies bound what the current evidence settles. The controller inherits the retrieval it conditions on. *L1* sends low-support or conflicting neighborhoods to verification, but a confidently wrong neighborhood can still be committed, and *L2* agreement bounds panel exposure without certifying correctness. The high-risk conclusions also rest on 48 TOP SECRET test documents, where *F1* carries a bootstrap interval of [0.75, 0.92]. Relocating the near-duplicate cutoff leaves the saturated regime unchanged (Table 5), so the tests we would run next are a chronological split and precedent stores for the two settings above.

Ethics Statement

This work studies automated security-level classification of documents using publicly available declassified and previously released corpora (WikiLeaks, DISC) that have been used in prior published research. The system is intended as a decision-support and human-review-routing tool for authorized information-governance workflows rather than as a fully autonomous classifier; the residual disagreement queue is explicitly designed as a human-review surface. We use no private or personally identifying data beyond what is already public in these benchmarks, and we report aggregate metrics only.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25597706).

GenAI Usage Disclosure

The authors disclose the following use of generative-AI tools. Generative AI was used as a writing aid for grammar and phrasing of the manuscript; all scientific content, claims, and analyses were written and verified by the authors. Within the system under study, open-weight LLM verifiers are evaluated experimental components rather than authoring tools, and their use is described in Section 3. No generative-AI tool was used to generate, label, or alter the WikiLeaks or DISC evaluation data or the reported numerical results. The authors take full responsibility for the content of this paper.

References

- Khudran Alzhrani, Ethan M. Rudd, C. Edward Chow, and Terrance E. Boult. 2017. Automated us diplomatic cables security classification: Topic model pruning vs. classification based on clusters. In *IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–6. IEEE.
- Elijah Bass, Massimiliano Albanese, and Marcos Zampieri. 2024. DISC: A dataset for information security classification. In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT)*, pages 175–185. SciTePress.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.
- C. K. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.
- Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks Lakshmanan, and Ahmed H Awadallah. 2024. Hybrid llm:

- Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, volume 2024, pages 41348–41366.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.
- Google DeepMind. 2025. Gemini 3 flash. Google DeepMind model card. Accessed: 2026-08-31.
- Google DeepMind. 2026. Gemma-4-e4b-it. Hugging Face model card. Accessed: 2026-06-14.
- Aaron Grattafiori and Llama Team. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Michael Hart, Pratyusa K. Manadhata, and Rob Johnson. 2011. Text classification for data loss prevention. In *Privacy Enhancing Technologies*, volume 6794 of *Lecture Notes in Computer Science*, pages 18–37. Springer.
- Vladimir Karpukhin, Barlas Oguz, 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)*, pages 6769–6781.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- 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. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31.
- Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087.
- Barack Obama. 2009. Executive order 13526: Classified national security information. In *United States. Office of the Federal Register*, Executive order 13526; EO 13526. United States. Office of the Federal Register.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, Mohammed Kadous, and Ion Stoica. 2025. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, volume 2025, pages 34433–34448.
- Qwen Team. 2025a. Qwen3-8b. Hugging Face model card. Accessed: 2026-06-14.
- Qwen Team. 2025b. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qwen Team. 2026. Qwen3.5-9b. Hugging Face model card. Accessed: 2026-06-14.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

A Additional Routing Diagnostics

Rare-class behavior and gate stability. Two questions decide whether the gate is safe on a rare class: whether it can be calibrated at all from few high-risk examples, and where its residual errors land. The $L1$ calibration split contains 327 documents (180 SECRET, 128 UNCLASSIFIED, and 19 TOP SECRET), and the operating point is fixed by five-fold out-of-fold cross-validation on that split, with no locked-test label used for feature construction, routing, threshold selection, or endpoint selection. Of the 536 evidence commits, 16 are incorrect: nine true SECRET documents committed as UNCLASSIFIED, four true UNCLASSIFIED committed as SECRET, one true SECRET committed as TOP SECRET, and two true TOP SECRET committed as SECRET. No TOP SECRET document is committed as UNCLASSIFIED, so the residual commit risk does not concentrate in the most damaging direction. At $L2$, 172 of the 253 residuals exit on verifier–evidence agreement and 38 of those exits are wrong, six of them on true TOP SECRET documents; agreement therefore bounds panel exposure without certifying correctness, and only evidence–verifier disagreements reach $L3$. End to end, CARE recovers 38 of the 48 TOP SECRET documents with four false positives (precision 0.9048, recall 0.7917, F1 0.8444 with bootstrap interval [0.75, 0.92]), at or above every single model in Table 1. For reference, the primary Macro-F1 of 0.8785 has a 95% bootstrap interval [0.843, 0.910].

The primary gate is a fixed 12-member seed ensemble, one router averaging twelve seed-initialized replicas, so the primary policy has no run-to-run training variance. In a separate study that retrains $L1$ from scratch under 31 independent seeds with the same corrected feature contract and development-only calibration, Macro-F1 is tight across seeds (standard deviation about 0.003) and TOP SECRET F1 stays within [0.822, 0.844], a band of roughly one to two documents on a 48-document class, with every retrained seed above the strongest single-model baseline (always-Qwen3-8B, 0.7957).

Table 4: DISC gate diagnostic. The learned $L1$ gate reduces unsafe evidence commits before verifier spending. Of the 16 residual commit errors, two involve true TOP SECRET documents and both are committed one level down to SECRET; none is committed as UNCLASSIFIED.

Gate	Evidence decisions	Errors	Sent to verifier
Deterministic agreement	740	86	49
Learned $L1$ reliability	536	16	253

Table 5: WikiLeaks decontamination sensitivity. The ConRAC Top-3 predictions are held fixed and only the evaluated subset changes, isolating the cutoff. The canonical Jaccard-0.85 row is the split used everywhere else; the protocol follows the motivation of Lee et al. (2022) while the 0.85 cutoff is our own. Macro-F1 stays within [0.9789, 0.9802] and SECRET F1 within [0.9567, 0.9606], so removing more near-duplicates does not lower the endpoint.

Evaluation subset	N	Macro-F1	F1 (SECRET)
Original held-out split	2,972	0.9802	0.9606
Remove exact / normalized matches	2,930	0.9798	0.9594
Canonical: add Jaccard-0.85 filter	2,863	0.9791	0.9576
Aggressive lexical/semantic union	2,783	0.9789	0.9567

Table 6: Paired-bootstrap support for the main DISC comparisons: 95% deltas (20,000 replicates, seed 928) on the locked $N=789$ split.

CARE vs.	Δ Macro-F1	Δ F1(TOP SECRET)
evidence vote	+ .0893 [.0591, .1223]	+ .2250 [.1510, .3099]
always-Qwen3-8B	+ .0211 [.0050, .0403]	+ .0487 [.0100, .0968]
uniform FT panel	+ .0082 [-.0001, .0207]	+ .0130 [.0000, .0447]
skip- $L2$ control	− .0009 [-.0029, .0000]	.0000 [.0000, .0000]

Table 7: DISC endpoint sensitivity under one fixed CARE route. Both rosters are open-weight; the primary CARE policy fine-tunes two of its three endpoints, the open-weight row uses base checkpoints throughout. Base endpoints cost aggregate quality (0.8785 to 0.8655 Macro-F1) and disagree more, so the gate routes 261/102 documents to $L2/L3$ instead of 253/81 and the bill rises from 496 to 567 invocations. The high-risk operating point does not move: TOP SECRET F1 stays 0.8444 at precision/recall .9048/.7917. The route, not the endpoint, is what holds it.

Policy	Acc	Macro-F1	F1 (TOP SECRET)	Phys. invoc.
CARE (primary)	0.8935	0.8785	0.8444	496
CARE (open-weight endpoints)	0.8745	0.8655	0.8444	567

Table 8: DISC commit-coverage sweep for the corrected primary policy. The $L1$ gate commits the lowest-risk documents first, so commit volume is the controller’s quality–invocation knob. The bold row is the primary CARE policy, fixed by five-fold out-of-fold calibration on the development split before test replay rather than chosen on test. A broad plateau between 400 and 669 commits holds TOP SECRET F1 at 0.8444 while invocations fall from 662 to 285; over-committing past 700 degrades both metrics.

$L1$ commits	To $L2$	To $L3$	Physical invoc.	% phys. budget	Macro-F1	F1 (TOP SECRET)
0	789	106	1,107	140	0.8693	0.8315
300	489	96	777	98	0.8758	0.8444
400	389	91	662	84	0.8776	0.8444
450	339	88	603	76	0.8794	0.8444
536	253	81	496	63	0.8785	0.8444
600	189	70	399	51	0.8739	0.8444
669	120	55	285	36	0.8740	0.8444
700	89	44	221	28	0.8694	0.8315
750	39	30	129	16	0.8608	0.8090
789	0	0	0	0	0.7892	0.6195

Table 9: DISC confusion matrices for the main error modes. Rows are true labels (U, S, TS) and columns are predictions in the order UNCLASSIFIED / SECRET / TOP SECRET. CARE removes most of the evidence-only TOP SECRET false-positive pattern while retaining the TOP SECRET recall of the strongest always-LLM row.

Policy	True U	True S	True TS
Evidence vote only	270 / 33 / 12	37 / 371 / 18	1 / 12 / 35
Always-Qwen3-8B	285 / 27 / 3	46 / 375 / 5	0 / 11 / 37
CARE (primary)	288 / 26 / 1	44 / 379 / 3	2 / 8 / 38

Table 10: DISC terminal-policy operating points after the learned $L1$ route, in physical invocations. Stopping CARE at $L2$ lets the verifier finalize every residual: it halves the bill to 253 invocations and lifts TOP SECRET precision to .9722, but recall falls to .7292. The primary policy, the bold row and the CARE row of Table 1, keeps the agreement gate and sends conflicts to $L3$, buying back the recall a rare high-risk class needs (.7917, F1 .8444). Both are the same route stopped at different stages, so the trade is a deployment choice rather than a different system.

Policy	Phys. invoc.	Prec. (TOP SECRET)	Rec. (TOP SECRET)	F1 (TOP SECRET)
CARE stopped at $L2$	253	.9722	.7292	.8333
CARE (primary)	496	.9048	.7917	.8444