

Anchor-Calibrated, Label-Free Screening of Non-Hand-Signed Signatures in Large-Scale Audit Reports

(Authors removed for double-blind review)

Abstract

Audit reports must carry each certifying accountant’s signature as the mark of an individual act of endorsement, yet once reports are produced and stored digitally a saved image of that signature can be pasted onto many reports instead — by manual stamping or by an automated signing system — producing what we term non-hand-signed signatures. The signer is genuine; the open question is whether an act of signing occurred for each report, and at archive scale this question carries no ground-truth labels. We present a label-free screening system for it and apply it to 86,071 Taiwanese statutory audit reports (2013–2023), within which the four largest audit firms contribute 150,442 analyzable signatures. The system finds the signature page, detects each signature, extracts deep features, and computes two similarities against the same accountant’s other signatures: a cosine similarity reflecting style and a perceptual-hash (dHash) distance reflecting pixel-level structure, on the logic that a consistent hand keeps style high while structure varies, whereas a reused image keeps both extreme. Because the archive has no labels and the data contain no natural gap (a unimodality test gives median $p = 0.35$ once firm effects and the hash’s integer steps are removed), no cutoff can be learned; instead we calibrate a five-way rule by how often it fires by chance between unrelated accountants in a clean reference group (the non-Firm-A firms, 2013–2019), where the strict high-confidence rule fires on about 1.2% of reports and a looser advisory band on about 17.5%. Held out from calibration as a known-positive benchmark — one firm independently described by interviews as a stamping firm, making this a confirmatory check rather than a blinded test — that firm fires the strict rule on 82% of its own signatures against 24–35% at the others, while its cross-firm rate sits at the clean floor, so the signal is entirely within the firm; the contrast survives stratification by comparison-pool size and resampling clustered at the accountant level, and 262 byte-identical signatures are direct evidence of reuse. Operationally, the screen locates where reuse concentrates without being told where to look and confines human review to exceptions. We are deliberate about what is and is not claimed: we report a between-accountant specificity proxy, not a true error rate — the within-accountant false-positive rate the question would require is not estimable without labels, and our coincidence rate is not even a bound on it — we cannot separate signing practice from a firm’s imaging pipeline, and we label no single signature. Calibrated on a large Chinese-signature corpus with script-agnostic descriptors, the rule serves as an operator-set reference point for comparable Chinese-signature pipelines.

Keywords: signature analysis, document forensics, perceptual hashing, deep features, unsupervised calibration, audit reports, anchor-based screening.

I. Introduction

An audit report is one of the main ways a company is held accountable to investors, and the certifying accountant’s signature is the visible sign that a named professional takes responsibility for it. In Taiwan, the Certified Public Accountant Act and the attestation rules of the Financial Supervisory Commission require certifying CPAs to put their signature or seal on each audit report [1]. The law accepts either a handwritten signature or a seal, but the point of the requirement is the same in both cases: the mark on each report should stand for a deliberate, individual act of endorsement for that particular engagement [2].

Going digital makes that harder to guarantee. Because reports are now created, sent, and stored as electronic files, it is easy to copy an accountant’s saved signature image onto many reports instead of signing each one. This can happen in two ways: a staff member can overlay a scanned signature onto the finished report (a stamping workflow), or a firm-wide electronic-signing system can do the same step automatically. We call signatures produced either way non-hand-signed. To fix the term operationally before any method is introduced: a signature is non-hand-signed when the mark on the report is a reproduction of a stored signature image rather than a fresh signing act for that engagement. This spans manual overlay (stamping),

automated firm-wide e-signing that pastes a saved image, and proxy application of a stored image by another person. It excludes a freshly handwritten signature or a hand-applied seal made for that specific report (the in-scope “hand-signed” case), and it is distinct from a cryptographic digital signature, which binds a document mathematically rather than reproducing an image. The criterion is therefore the visible outcome — image reuse, the same stored image recurring across reports — not the intent, the actor, or the legal status, and it is this outcome that our two measures and five categories track. The worry is not about legality; it is about meaning. A single image pasted onto hundreds of reports may not carry the individual endorsement the rule assumes — a concern the literature on signatures connects to behavior, and, in auditing specifically, to rules that name and identify the engagement partner [31], [32], [33]. This is also why the problem is not forgery: a non-hand-signed signature reuses the real signer’s own image, and at scale no reader can see the difference.

That difference matters for the method too. Almost all work on offline signature analysis is about forgery — deciding whether a questioned signature was really written by the person it claims to be [3]–[8]. In our setting the identity is not in doubt; the accountant is genuine. What we want to know is whether the person actually signed each report, or whether one signing was copied as an image. This removes the need to model clever forgers, but it adds a new difficulty: we must separate a person who signs consistently from a reused image. Someone who signs in a very steady hand will produce signatures that look alike year after year; a process that reuses one stored file will produce signatures that are structurally identical. The method has to tell these two cases apart.

Two facts make the obvious approach — pick a similarity cutoff and call everything above it a copy — unworkable, and they shape our design. First, archives like ours have no labels at the level of individual signatures: no signature is marked as “definitely hand-signed” or “definitely reused.” Without such labels, any cutoff we choose has unknown error rates; we cannot measure how often it would wrongly flag a genuine signature or miss a reused one. Second, even setting labels aside, the data themselves do not contain a natural cutoff. As we show in Section V, the raw numbers look at first as if they split into two groups, but that appearance comes from differences between firms and from the fact that the hash takes only whole-number values; once we remove those two effects, the distribution is a single smooth spread, not two clusters. You cannot read a dividing line off a distribution that has no gap, and you cannot test a line against labels that do not exist. So the method must get its cutoff some other way.

Our two similarity measures are chosen precisely to expose the distinction the problem turns on. For each signature we compute two numbers against the same accountant’s other signatures: a cosine similarity on deep ResNet-50 features, and an independent perceptual hash (dHash) distance. They carry different information. Cosine similarity measures overall style, and it is high both when an image is reused and when a person signs consistently. The dHash distance measures structure almost pixel by pixel, and a very small distance is the sign most specific to a reused image. But neither measure is enough on its own. Cosine alone over-flags a steady hand, because consistent signing also keeps it high. dHash alone has the opposite weakness: it is brittle to how an image is captured — a reused signature that has been re-scaled, re-cropped, or re-compressed can show a larger dHash distance and slip past a structure-only test — and a small dHash distance carries no meaning between two signatures whose styles do not match in the first place. The two are complementary precisely because they fail in different directions: cosine first establishes that the styles match, which catches reuse even when the image has been mildly altered, and dHash then asks whether the match is also near-identical in structure, which is what separates a reused image from a merely steady hand. A single similarity number blurs these two cases; two measures keep them apart. The implication between them runs one way only: a near-identical structure (a tiny dHash) forces a high cosine, but a high cosine in no way implies a near-identical structure — which is why the two-measure plane cannot be collapsed onto either single axis. This complementarity also shapes the rule (Section III-D): because a small dHash distance is only meaningful once cosine is already high, the structural cut subdivides the high-cosine cases rather than the low-cosine ones. This is the heart of the design.

On this basis we build and study a complete screening system. The pipeline takes raw PDF reports through four steps — find the signature page, detect each signature, turn it into features, and compute the two similarities — and sorts each signature into one of five categories. Because there is no natural cutoff to read off the data and no labels to learn one from, we instead measure how often the rule fires by chance

between unrelated accountants in a clean reference group. That chance rate is a between-accountant coincidence rate, which we treat as a proxy for the rule's specificity: it gives us a principled way to choose an operating point, and — just as important — it tells us what each category's flag is worth among unrelated accountants. It is not the within-accountant false-positive rate (how often a genuine consistent hand-signer would fire the rule), which the reuse question would ideally use but which no labels let us estimate (Section III-E).

What is the screen for? Two things. Run over a large archive, it discovers where reuse concentrates — which firms, which periods — without being told where to look. And it keeps human review at the scale of exceptions. In a reuse-dominated population (a stamping firm, a firm with an electronic-signing system), the high-confidence tier routes most signatures directly to a high-specificity candidate list, and the small residual goes through a defined review protocol (specified in Section IV-B) — side-by-side overlay inspection, secondary image-artifact checks, and bounded per-accountant sampling — that also accumulates labels for later calibration. In a mixed population, where hand-signing and informal stamping coexist, the ambiguous middle is larger, and the same disposition machinery delivers the same promise one level up, at the accountant: the low-specificity advisory band is demoted rather than worked, accountant-level scores concentrate attention on the few high-ranked or mixed cases, and byte-identity hits supply proof where proof exists, confirming that an accountant's stored image is in circulation. What the screen does not deliver there — and we say so plainly when we report the category proportions (Section IV-B) — is a per-signature verdict for the ambiguous middle. In every case the output is bounded triage, not a verdict on any single signature.

The Taiwan setting suits this study well. The Market Observation Post System offers a large, standardized, public collection of statutory audit reports, each with the same two-signature format, which makes large-scale extraction practical. In addition, anonymized interviews with certifying partners and signing-system staff at all four firms give us institutional facts about how each firm signs and about when each firm adopted a formal electronic-signing system — adoptions that were staggered from 2020 onward. This gives the study a natural before-and-after structure in time, and outside information against which to read the firm-level results (Section III-A).

We make four contributions:

1. An end-to-end screening pipeline that turns raw audit-report PDFs into operational risk strata for hundreds of thousands of signatures.
2. A dual descriptor that separates style consistency from image reproduction — a distinction a single similarity measure blurs.
3. The methodological core: a label-free way to construct and characterize a screening operating point when no signature-level labels exist — the question this paper is really organized around. With neither a natural cutoff in the data nor labels to learn one from, we set a tunable rule by measuring how often it fires by chance in a clean reference group, and we say plainly what that measure can and cannot support. This is the part we expect to transfer beyond the present setting; we demonstrate and stress-test it at scale on audit signatures rather than claiming it as a finished, fully general framework. Concretely the result is not only a calibration method but a concrete operating point — the high-confidence rule and its measured specificity proxy — that practitioners working with comparable Chinese-signature image pipelines can use as a starting reference (not transplant unchanged, since the proxy is conditional on a similar preprocessing and reference-group setup), together with a defined disposition path for the ambiguous middle (calibrated demotion of the low-specificity band, aggregation to the accountant level, byte-identity escalation, and a bounded manual protocol) that keeps human review at the scale of exceptions.
4. A demonstration on Chinese signatures, a structurally complex and comparatively under-served script for signature analysis. Because our descriptors work on the image rather than on script-specific strokes, the approach does not depend on Latin-script assumptions and is a candidate for other scripts.

The paper is organized to move from the problem to the evidence. Section II reviews related work and states the gap. Section III describes the study design — the data split, the pipeline, the five-way rule, and the calibration logic — and explains why each piece is built the way it is. Section IV reports the results: the calibration baseline, which category needs human review, and the held-out benchmark on Firm A. Section

V collects supporting analyses, including the diagnostic showing that no natural cutoff exists. Section VI concludes.

II. Related Work and Research Gap

Why reproduction matters: signatures carry symbolic weight. A signature is valuable mainly as a symbol — it stands for the signer’s identity and intent. Recent experiments show that this symbolism does not survive a change in how one signs. In studies that take the reader’s point of view, Chou [41] finds that electronic signatures give a weaker sense of the signer’s presence than handwritten ones, and that readers therefore judge an e-signed document as less valid and expect more non-compliance; across five kinds of e-signature (a checked box, a PIN, an avatar, a typed name, and a software-generated signature), the software-generated kind felt the most “present” of the electronic options but still less than a handwritten signature. In studies that take the signer’s point of view, Chou [42] finds that electronic signatures give a weaker sense of self-presence — the signer’s felt attachment to the mark — and that this, in turn, makes people more willing to cheat; the work singles out signing by proxy (an autopen) as cutting the tie between the document and the signer. These results matter for us because the practice we detect — a stored signature image laid onto a report by staff or by software — is, in this scheme, one of the lowest-presence modes: it looks like a software-generated signature and is executed like a proxy signature, because the accountant performs no signing act for the report. These effects are robust rather than one-off: in a pre-registered, multi-study replication with meta-analysis, Tzelios and Williams [43] reproduce Chou’s reader-side result — an avatar e-signature lowers the sense of the signer’s presence and raises the expectation that the contract will be breached. In their general discussion the same authors point to accounting as a next setting — noting the spread of online tax filing and asking how digital signatures affect an evaluator’s assessment of the legitimacy of claims, while cautioning that accounting documents may prove less sensitive to signature form than legal ones. We read that call precisely: their “auditors” are the readers of digitally signed filings — those who evaluate the claims — not the certifying accountants who sign. The signer-side question in auditing — what it means when the certifying professional’s own signature is reproduced rather than performed — is not addressed in that literature. Both questions, reader-side and signer-side, presuppose the same missing capability: a way to measure non-hand-signing at scale. The lesson we draw is not that non-hand-signing harms audit quality — that is a separate question we leave to a companion study (Section VI) — but that whether it matters is a real question, and one nobody can study without first being able to measure non-hand-signing at scale.

Signature analysis to date is about forgery, not reuse. The obvious toolkit for that measurement is signature analysis, but its main concern is the wrong one for us. Bromley et al. [3] introduced the Siamese network that still anchors the field; SigNet [4] extended it to compare writers it had never seen; Kao and Wen [5] worked from a single genuine sample; TransOSV [6] brought in a Vision Transformer; and meta-learning has been used to cut the effort of enrolling new signers [16]. All of this targets imitation by another hand, so it learns to tell different people apart. Our task is the opposite: spotting reuse of the genuine signer’s own image, which lives in the most-similar tail of one person’s signatures. The closest idea uses reference examples to set a sensible cutoff [8], but on benchmark data with known genuine references — whereas our archive has no signature-level labels at all. This body of work is also overwhelmingly built on Western, Latin-script signatures; non-Latin scripts such as Chinese are comparatively under-served, and reported accuracies for them are lower [44]. Chinese signatures are structurally distinctive — many strokes, with wide variation between writers — and the forensic literature on them is thin; the closest precedent, Chen [45], analyzes Chinese signatures with a maximum-similarity-to-same-class statistic that directly parallels our use of the maximum cosine to the same accountant. Our descriptors, however, work on the image rather than on script-specific strokes, so the method itself does not depend on the script.

Image-duplication and document forensics: useful parts, different setting. A second line of work looks directly at duplicated images. Copy-move detection finds regions copied within an image [11], and Abramova and Böhme [10] adapted it to scanned documents, noting that ordinary repeated characters confuse the standard methods. Self-supervised copy detection on everyday photos [13] shows that pretrained CNN features with cosine similarity make a strong baseline for spotting near-duplicates. Closest in pipeline terms, Woodruff et al. [9] pull signatures from corporate filings for anti-money-laundering work — but to group

signatures by who signed them, not to detect one signer’s image being reused across documents. The building blocks exist; the specific setting — one signer’s image reused across many scanned financial reports — does not seem to have been addressed.

Deep features and perceptual hashing as ready-made parts. Features from a pretrained CNN transfer well to document images without any retraining [20], [21], and perceptual hashes are built to survive the print–scan–rasterize cycle [27]. Jakhar and Borah [12] show that combining a perceptual hash with deep features beats either one alone for near-duplicate detection — a direct precedent for our two-measure design, though they work on natural images rather than signatures.

The recurring obstacle is the missing label. None of these lines solves the problem we face, because real archives carry no signature-level ground truth, and a similarity screen without it falls back on a hand-chosen cutoff whose error behavior is unknown. (The statistical tools we use to test for a natural cutoff and to describe the rule once we find none are introduced where they are used, in Section III and Section V, since they are part of our method rather than prior work on this problem.)

The gap, and our contribution. Two gaps follow. First, large-scale screening for non-hand-signed auditor signatures has not been done, even though there is good reason (above) to think it matters. Second, and more broadly, similarity-based screening has no principled way to set and describe an operating point when labels are missing. Our contribution sits exactly here: a label-free calibration that replaces both the arbitrary cutoff and the unavailable labeled validation with a chance-rate measured in a clean reference group, together with the pipeline and dual descriptor that make the screening possible (contributions listed in Section I).

It is worth being explicit about a design choice this implies, because it is easily mistaken for a missing component. A natural reflex would be to learn the discriminator — to fine-tune a Siamese or contrastive network to separate reused from hand-signed signatures. We deliberately do not, and the reason is not expedience but the defining constraint of the setting: supervised metric learning requires labeled pairs (genuine-vs-reused), which is exactly the ground truth the archive does not contain. Training such a network would require either fabricating labels or importing them from a different distribution (e.g., forgery datasets), reintroducing the unverifiable assumptions our calibration is designed to avoid; the resulting boundary would again have unknown error behavior on the real archive. Label-free operation is therefore not a weaker version of a supervised method but the only honest option when no labels exist, and the contribution is correspondingly methodological — a way to set and characterize an operating point by measured chance behavior — rather than a new network architecture. Off-the-shelf pretrained features are used precisely because they introduce no task-specific supervision; supervised fine-tuning is the right tool once a labeled sample exists, which is why we frame the review protocol’s first run (Section IV-B, Section V) as the route to that sample and to any future supervised validation.

III. Research Background and Study Design

This section explains how the study is built and why. We report no computed numbers here; all results appear in Section IV.

A. Institutional Background

To pin down the signing practices that we need in order to interpret the results, we held semi-structured interviews with certifying partners and signing-system staff at all four firms in the study.¹ Three points do real work later. First, all four firms allow handwritten signing but none require it. Second, formal firm-wide electronic signing or sealing systems were adopted on staggered dates from 2020 onward. Third, one firm — which we call Firm A throughout — has used scanned-image overlay stamping as its usual practice since at least 2013. We use these facts only as background, not as labels for individual signatures: they guide how we split the data below and how we read the firm-level results in Section IV-C, but they do not tell us the status of any single signature. A further caution applies to how the interviews are used as corroboration. They are self-reported, anonymized, and not independently reproducible, so when the screen’s firm-level output agrees with them (Section IV-C) that agreement is evidence of consistency with domain knowledge,

not a measurement of the screen’s accuracy or recall — quantifying those would require signature-level labels, which the archive does not provide. To be unambiguous about their role: the interviews are used only to contextualize the firm-level findings and are not treated as validation. They are corroborative, not confirmatory, and not independently reproducible; the empirical claims of this paper rest on the calibration and the byte-identical evidence, which stand without them. Their one load-bearing use is to motivate why Firm A is read as a known-positive benchmark rather than a blinded test (Section IV-C) — a framing that, if anything, lowers the evidentiary status we claim for that firm. The practical implication is that the years before the formal systems (before 2020) are the right “normal” period to use for calibration.

¹ Footnote — institutional detail. The interviews were conducted under institutional research-ethics approval and are reported in anonymized, aggregated form; firms are labeled A–D and no individual can be identified. The formal systems were reported to have been adopted at roughly one firm in early 2020, one in 2021, and one in late 2022 (exact firm-level dates are withheld for anonymity; see supplementary materials). Interviewees attributed this timing partly to the COVID-19 pandemic, which forced remote review and signing, and to firm-wide paperless and environmental (ESG) initiatives — both of which accelerated the move to formal electronic signing at Firms B/C/D. For Firm A, the reported workflow is that the certifying accountant approves the finished report electronically, after which the print room overlays the accountant’s stored seal or signature image onto the PDF and prints it; the stored image is rarely changed, and although handwritten signing is allowed it is reported to be very rare, and rarer over time. Before the formal systems, the other firms’ practice varied: some used informal scan- or photocopy-based stamping alongside handwritten signing, and at least one reported mostly handwritten signing before its system. The property the calibration relies on (Section III-E) is that, in the pre-2020 baseline firms, different accountants did not share a common template — not that every signature was handwritten.

B. Data and Analysis Design

The corpus is all retrievable Taiwan statutory audit reports for fiscal years 2013–2023; the four largest firms (A–D) form the primary analysis sample, and non-Big-4 firms enter only in the crossover-scope robustness check (Section V-C), never in the calibration or the headline rates. Signatures are extracted as described in Section III-C. To be precise about the headline denominator, since it recurs throughout: “150,442 analyzable signatures” means exactly those Big-4 signatures that are valid and have both similarity measures computed, assigned by each accountant’s registered firm (Firm A 60,448, Firm B 34,248, Firm C 38,613, Firm D 17,133). We then split the Big-4 corpus by firm and by period, giving each part a distinct job (Fig. 1):

- Calibration (the clean reference group): Firms B/C/D, 2013–2019.
- Held-out benchmark 1: Firm A, 2013–2023 (a known positive, not a blinded test).
- Held-out test 2 (secondary): Firms B/C/D, 2020–2023.

We explain the reason for each part in Section III-E. The key idea is simple: we calibrate only on the clean cell — the non-Firm-A firms in the years before formal systems — and test everything else against it. No numbers appear here; the calibration results start in Section IV-A.

Figure 1. Data split: calibrate on the clean cell, test everything else

Firm A	Held-out test 1 (Firm A, full record)	Held-out test 1 (Firm A, full record)
Firm B	Calibration (clean reference)	Held-out test 2 (secondary)
Firm C	Calibration (clean reference)	Held-out test 2 (secondary)
Firm D	Calibration (clean reference)	Held-out test 2 (secondary)
	2013–2019	2020–2023

Figure 1. The data split. Rows are Firms A–D; columns are 2013–2019 and 2020–2023. The B/C/D $\times$ 2013–2019 cells are the clean calibration group; Firm A (both periods) is held-out benchmark 1 (a known positive); B/C/D $\times$ 2020–2023 is the secondary held-out test. We calibrate only on the clean cell and test everything else against it.

C. Pipeline

The pipeline turns a raw PDF report into labeled signatures in five steps (Fig. 2).

Finding the signature page. A vision-language model [24], [35] scans only the first quarter of each document — where the auditor’s report page reliably sits — and stops as soon as it finds the page.

Detecting signatures. A YOLOv11n detector [25], [34], trained on 500 hand-labeled signature pages (425 for training, 75 for validation; 100 epochs; started from COCO weights), draws a box around each signature. A region counts as a signature if it holds handwritten content that belongs to a personal signature, even where it overlaps an official stamp. A red-stamp removal step (filtering in HSV color space) then strips away overlapping red seals, leaving the handwritten part.

Turning signatures into features. Each detected signature is passed through an ImageNet-pretrained ResNet-50 [26] used as a fixed feature extractor — we take the 2,048-number output of its global-average-pooling layer and drop the classification head. We resize each image to 224 $\times$ 224 while keeping its aspect ratio (padding with white), apply the standard ImageNet normalization, and scale the feature vector to unit length, so that cosine similarity is just the dot product. We use these off-the-shelf features rather than fine-tuning the network, for three reasons: the task is comparing similarity, not classifying; ImageNet features are known to transfer well to document images [20], [21]; and not fine-tuning avoids the risk of learning quirks of our particular dataset. The backbone choice is checked in Section V-C.

Assigning each signature to an accountant. Each signature is matched to a registered accountant by its position on the page (first or second) against the official registry. Signatures we cannot match are left out of the same-accountant comparisons, because the “most similar signature by the same accountant” measure has no meaning without an assigned accountant.

(Detection accuracy, signature counts, match rates, and the resulting analysis sample are reported in Section IV-A.)

Figure 2. The screening pipeline

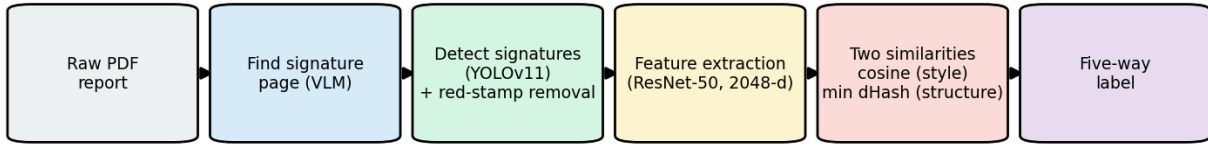


Figure 2. The screening pipeline. A raw PDF passes through page-finding (a vision-language model), signature detection (YOLOv11) with red-stamp removal, feature extraction (ResNet-50), the two per-signature similarities (cosine for style; the smallest dHash to the same accountant for structure), and a five-way label.

D. The Two Similarity Measures and the Five-Way Rule

For each signature we compute two numbers, both against the same accountant’s other signatures: *cos*, its highest cosine similarity to another of that accountant’s signatures, and *dHash*, its smallest perceptual-hash distance to another of them. As explained in Section I, the point of using two measures is to separate two things that one measure blurs. A high *cos* means the signatures look alike in style, which happens both when an image is reused and when a person signs consistently. A small *dHash* means the signatures are alike almost pixel for pixel, which is the sign most specific to a reused image. Together they are far more telling than either alone: a steady hand gives a high *cos* but a *dHash* that still varies, while a reused image gives a high *cos* and a tiny *dHash*.

The rule places each signature in one of five categories, with cosine acting as the primary gate and the structural (*dHash*) distance refining only the cases where cosine is already high. Each name states the screening hypothesis its region suggests — a candidate reading, not a confirmed determination:

- HC — high-confidence reuse candidate: cosine above the high cut and structure at or below the near-identical cut. Both measures point to a reused image.
- MC — moderate-confidence, advisory: cosine above the high cut and structure between the two structural cuts. Style is very similar, but structure is below the strict bar.
- HSC — high style-consistency: cosine above the high cut and structure above the upper structural cut. Style is similar with no structural support.
- UN — uncertain: cosine between the low cut (the same-vs-different-accountant crossover) and the high cut.
- LH — low reuse-similarity: cosine at or below the low cut.

A report takes the strongest label among its signatures ($HC > MC > HSC > UN > LH$). Table I-a summarizes the five categories, the thresholds that define them, and the notation used throughout; the high cut is cosine 0.95, the low cut is 0.8547 (the same-vs-different-accountant crossover; Section IV-A), and the two structural cuts are *dHash* 5 and 15.

Table I-a — Category definitions, thresholds, and notation.

Label	Name	Condition (cosine <i>c</i> , structure <i>dHash</i> <i>d</i>)	Role
HC	high-confidence reuse candidate	$c > 0.95$ and $d \leq 5$	self-certifying flag
MC	moderate-confidence	$c > 0.95$ and $5 < d \leq 15$	advisory
HSC	high style-consistency	$c > 0.95$ and $d > 15$	no structural support; no weight

Label	Name	Condition (cosine c , structure dHash d)	Role
UN	uncertain	$0.8547 < c \leq 0.95$	ambiguous middle
LH	low reuse-similarity	$c \leq 0.8547$	likely hand-signed

Other abbreviations used throughout: ICCR — inter-CPA coincidence rate, the between-accountant chance-firing rate that calibrates the rule (Section III-E); c — cosine similarity to the same accountant's other signatures (style); d — smallest dHash distance (structure).

Why the partition has this shape (five categories, not nine). As explained in Section I, a near-identical structure is decision-relevant only once the styles already match, so the two cosine cuts come first — splitting signatures into three style bands (low, uncertain, high) — and the two structural cuts subdivide only the high band. Three facts pin this shape down. First, structure carries little standalone decision weight in the two lower bands: between signatures whose styles do not clearly match, a moderate structural distance is hash noise, not evidence of reproduction — and even the near-identical structural matches that do appear below the style cut (quantified next) are not assigned HC; their structural information re-enters only through accountant-level aggregation and byte-identity review (Section IV-B), not through a separate cell. Second, the cells of the full 3×3 grid that pair a lower style band with a near-identical structure are sparsely populated rather than ignored — and the empirical reading is more precise than a simple “they are empty.” An explicit count makes this exact: of the 150,442 Big-4 signatures, 7,681 (5.1%) combine a near-identical structural match ($d\text{Hash} \leq 5$) with a sub-0.95 cosine, so the one-way implication of Section I (a tiny dHash forces a high cosine) holds approximately, not strictly. But the residents’ mass sits immediately below the high-cosine cut — 7,311 of them (95.2%) fall in cosine 0.90–0.95, and only 370 signatures (0.25% of the corpus) reach the genuinely low-cosine bands, of which just 38 lie below the LH/UN crossover (cosine ≤ 0.8547). These residents are not degenerate crops: their image size (mean 33k px) and detection confidence (0.875) match the rest of the corpus (28k px, 0.877). Under the coherent same-pair definition — style and structure satisfied on the same partner signature — the count falls further to 874 (0.58%). The point is therefore not that these cells are empty but that subdividing the lower style bands by structure changes no disposition: because cosine is the primary gate, a near-identical structural match beneath the style cut is already handled as UN, and the residual structural information re-enters through the accountant-level aggregation and byte-identity escalation of Section IV-B rather than through a separate cell. Third, a partition should cut only where the resulting actions differ: subdividing the two lower bands by structure would create cells whose dispositions (Section IV-B) are identical — all demoted or aggregated the same way — adding calibration burden without operational consequence, whereas the three structural cells inside the high band exist precisely because their dispositions differ. (Count from the deployed-rule descriptor columns; any-pair definition, full Big-4 corpus.)

The cuts are operator-tunable operating points, not learned boundaries: there is no natural gap to read off the data (Section V-A) and no signature-level labels to learn one from, so the cuts are chosen and their specificity is measured, not learned. The four cut values, and where each one comes from — two are read directly from this study’s data — are given in Section IV-A, alongside the chance-rate calibration that characterizes them and the figure of the two-measure plane (Fig. 3).

Any-pair versus same-pair: how the two extrema combine. One construction detail deserves to be explicit, because a careful reader will ask. The two per-signature values are independent extrema over the same accountant’s other signatures — the highest cosine and the smallest dHash, each taken on its own — so the two values may come from different partner signatures. We call this the any-pair rule, and the choice is deliberate, for three reasons. First, the two descriptors have different invariances: cosine survives re-scaling and re-compression; dHash does not. For a genuinely reused image that crossed different scan or compression pipelines, the style-nearest copy and the pixel-nearest copy can therefore legitimately be different reports — forcing both extrema onto one pair would miss exactly that most realistic positive case. Second, dHash takes whole-number values and ties are massive in duplicate-heavy pools: which tied copy wins the minimum is essentially arbitrary, so whether the two extrema land on the same file is largely tie-breaking noise — both point into the same duplicate cluster. Third, the chance-rate calibration of Section

IV-A applies the same any-pair rule to the clean reference group, so the high-specificity claim rests on the absolute clean-group rate (the HC rule fires by chance on only ~1.2% of clean-group reports), not on any firm-versus-floor ratio; the same rule is applied to every firm and to the reference group alike. The stricter same-pair variant, in which a single partner signature must satisfy both inequalities at once, is reported as a robustness check (Section V-C) and leaves every conclusion unchanged — the within-firm concentration of cross-accountant matches is in fact higher under same-pair (97.0–99.96% across the four firms) than under the deployed any-pair rule (76.7–98.8%) — because in the high-confidence region the two rules nearly coincide: a partner within the near-identical structural cut is pixel-near-identical and therefore clears the high style cut by itself.

Limitations, stated up front. Three follow directly from the design. (i) Because the cutoffs are chosen rather than learned, the system has a tunable operating point, not an optimal one. The dial moves in only one direction: a reviewer who wants more conservatism can tighten it for higher specificity, but cannot trade in the other direction toward higher recall, because recall is unobservable without labels — there is no precision–recall frontier to navigate (Section V). (ii) The chance rate we report is a between-accountant coincidence rate — a proxy for specificity, not a true false-acceptance rate — because we have no labeled negatives; it speaks to how often unrelated accountants collide, not to how often a genuinely consistent hand-signer would fire the rule, which is the quantity the reuse question needs and which we cannot estimate (Section III-E). (iii) For any single signature, the two measures cannot tell us why it is so similar to another: reuse of an image, a shared scanning pipeline, and a very uniform house style all push the numbers the same way, and we do not try to choose between them at the level of one signature. These limits apply to every claim that follows.

E. The Clean Reference Group and the Chance Rate

With no labeled negatives to learn from, the calibration uses a stand-in: a group in which the rule should fire only by chance — unrelated accountants whose signatures happen to look alike now and then. Choosing this group well is the central design decision, and two requirements force the choice.

Why not all four firms. As Section IV-C will show, almost all of one firm’s between-accountant matches fall on other accountants of the same firm, and we have byte-level proof of image reuse across about fifty of that firm’s partners. If we put Firm A into the reference group, we would be filling the “by chance” rate with exactly the within-firm matches the rule is supposed to catch — a circular calibration. So we use Firms B/C/D as the clean reference group and keep Firm A as a test case; we report the all-four-firm number only to show how much Firm A contaminates it.

Why 2013–2019. We further limit the reference group to the years before formal firm-wide electronic-signing systems (adopted from 2020 onward; Section III-A). What this buys us is the absence of a shared template across accountants — not a guarantee that every signature was handwritten. The interviews say some baseline firms used informal individual stamping before 2020, but each accountant’s stored image was their own, so different accountants’ signatures still match only by chance; the chance rate is about matches between accountants, which individual stamping does not inflate. One further channel deserves to be named, because it is not the template and we cannot fully exclude it: accountants at the same firm pass through a shared imaging pipeline — common scanners, PDF-assembly software, and the red-stamp-removal step (Section III-C, Section V-B) — and a shared pipeline can imprint correlated artifacts on otherwise-unrelated signatures, which would lift the inter-CPA rate above true chance. The pipeline audit of Section V-B confirms that such shared production paths exist and change over time. This is a reason to read the ICCR as a specificity proxy rather than a literal coincidence rate; its bias, like reference contamination, runs toward a higher floor, which makes the Firm-A contrast more conservative rather than less. After 2020, formal systems standardize how reports are assembled, so that period is not a clean reference — and indeed the chance rate rises after 2020 (Section V-B). We therefore calibrate on the Firms-B/C/D 2013–2019 cell and score every held-out cell against it.

We report the rule’s chance rate at three levels, because the rule takes the best match over a pool and so the per-signature rate is not the same as the per-pair rate: per comparison (sampled pairs of different accountants), per signature, and per report, each with a confidence interval. We call this the inter-CPA co-

incidence rate (ICCR) rather than a “false-acceptance rate,” which we reserve for settings that have labeled negatives. The ICCR is a between-accountant coincidence rate: how often the rule fires on the signatures of two different accountants. It is therefore at best a proxy for specificity, and only under the stated assumption (no shared template across accountants). It is important to be exact about what it is not. The quantity the reuse question actually needs is the within-accountant false-positive rate — how often the rule would fire on a genuinely consistent hand-signer’s own signatures — and that rate is not estimable here, because no accountant in the corpus is labeled as a known hand-signer. We considered benchmarking it against an external corpus of genuine repeated signatures (a public signature dataset supplies many authentic samples per writer), but such corpora are a different population and script acquired under a different pipeline, so the resulting rate would not transfer to this setting; importing it would reintroduce exactly the kind of unverifiable cross-distribution assumption our label-free calibration is built to avoid. We therefore report the limitation rather than a misleading proxy. The ICCR is not even a bound on it: a uniform individual hand keeps cosine high by design, so a true hand-signer’s within-accountant fire rate can sit far above the between-accountant coincidence rate. Any statement that divides a firm’s within-accountant fire rate by this between-accountant floor (an “X× the floor” comparison) therefore overstates the gap — the bias runs in the anti-conservative direction — and we do not report such ratios as effect sizes. Read as a between-accountant specificity proxy under the stated assumption, the ICCR is faithful to the evidence; read as a true error rate for the reuse question, it would claim more than we can show.

One further assumption deserves to be stated rather than buried, because it concerns how the clean group was chosen. The floor is conditional on the reference group actually being clean — it is a coincidence rate among accountants we take to be independent hand-signers, and the group (non-Firm-A firms, pre-2020) was selected partly because its rates are low and its practices, by the interviews, are not stamping-dominated. That selection is mild but not innocent: if some baseline accountants in fact reuse images undetected, the reference is contaminated. The direction of that error, however, is reassuring for the Firm-A contrast. Undetected reuse inside the baseline would only raise the between-accountant coincidence floor, which makes Firm A’s gap above it smaller, not larger — so contamination of the clean group biases the headline contrast conservatively, against our conclusion rather than toward it. Two pieces of evidence bound the concern empirically. First, the three baseline firms are mutually consistent and uniformly low (Firms B and C within about 3.5× of each other, none close to Firm A; Section IV-A), so the floor does not hinge on any single firm and a leave-one-baseline-firm-out reading does not move it materially. Second, the one data-derived threshold, the low cosine cut, is stable when the group composition is changed — 0.8547 on the calibration cell, 0.8302 with the non-Big-4 firms folded in, a shift of at most 0.025 (Section V-C) — so widening or narrowing the reference at its boundary does not move the operating point. We therefore treat the clean-group assumption as a stated limitation with a known-safe error direction, not as a hidden premise.

F. What HC Means and Does Not Mean

One sentence prevents the most common misreading of everything that follows. HC is not a reuse label. HC denotes an extreme within-accountant repetition pattern — a signature whose closest match among the same accountant’s own signatures is both stylistically near-identical (cosine > 0.95) and structurally near-identical (dHash ≤ 5) — that is statistically rare between unrelated accountants, by the ICCR calibration of Section III-E. That is the whole of what the rule, on its own, establishes. Reuse of a stored image is one interpretation of an HC pattern, and the most economical one, but the rule does not imply it: a very steady hand, a fixed scanning-and-assembly pipeline, or a uniform house style can each raise within-accountant repetition (Section V-A, Section V-B). Where we go further than “extreme repetition” — as we do for Firm A — the additional weight comes from outside the rule: byte-identical signatures, which independent hand-signing cannot produce, and the institutional context, neither of which is implied by HC alone. For Firms B/C/D we make no reuse claim at all; their HC signatures are reported as a within-accountant repetition rate, not as detected reuse. Read this way, HC is a calibrated, reproducible screening category, and “reuse” is a conclusion that has to be earned separately — firm by firm, or signature by signature — rather than read off the label.

IV. Findings

This section reports the numbers. It starts with the calibration baseline (Firms B/C/D, 2013–2019), then says which category needs human review, then presents the held-out benchmark on Firm A.

A. Detection Sample (Whole Corpus) and the Calibration Baseline (Firms B/C/D, 2013–2019)

Detection and the analysis sample (whole corpus). Two scopes appear in this section and must not be confused: detection and the analysis sample here are computed on the whole corpus, whereas both data-derived calibration quantities — the chance-rate ICCR and the low cosine cut (Section IV-C) — are computed only on the clean Firms-B/C/D 2013–2019 cell. Of the 90,282 reports, the page-finder flagged 86,084 as having a signature page (the other 4,198, or 4.6%, had none); 13 of those 86,084 could not be rendered, leaving 86,071 documents processed. On the validation set, the YOLOv11n detector reached precision 0.97–0.98, recall 0.95–0.98, mAP@0.50 0.98–0.99, and mAP@0.50:0.95 0.85–0.90. Across the corpus it extracted 182,328 signatures — 2.14 per document with detections, where two certifying accountants per report implies 2.00. The $\approx 6.7\%$ excess is explained by extra detections rather than missed accountants: of the 13,573 detections (7.4%) that could not be matched to a registered accountant and were excluded, 8,901 (66%) are third-or-later detections on a page — boxes beyond the two certifying signatures — and the unmatched set as a whole carries lower detection confidence than the matched set (mean 0.826 vs 0.874), consistent with these being extra boxes and low-confidence noise; the remaining 4,672 are first/second-position detections that failed registry matching. Throughput was 43.1 documents per second, and the detector agreed with the vision-language model on 98.8% of documents. Matching by position assigned 92.6% of signatures (168,755 of 182,328) to a registered accountant; of these, 168,740 have both similarity measures computed (the 15-signature difference is accountants with a single signature in the corpus, for whom no same-accountant comparison exists, so the full-corpus distributional statistics in the Appendix are reported on 168,740). The four-firm analysis sample is 150,442 signatures with both measures computed, from 457 accountants (Table I); the accountant-level partition (Table III, Section V-C) is fit on the 437 of these with at least ten signatures (171/112/102/52 across Firms A–D).

Table I — Detection and extraction summary.

Quantity	Value
Documents with a signature page	86,071
Detector precision / recall	0.97–0.98 / 0.95–0.98
Detector mAP@0.50 / mAP@0.50:0.95	0.98–0.99 / 0.85–0.90
Signatures extracted	182,328 (2.14 per document)
VLM–detector agreement	98.8%
Signatures matched to an accountant	168,755 (92.6%)
Four-firm analysis sample	457 accountants; 150,442 signatures

The calibrated operating point: the four cut values and their bases. The five-way rule of Section III-D uses four cut values; we state them here because two are read directly from this study’s data. The low cosine cut, 0.8547, is the crossover of the same-accountant and different-accountant cosine distributions computed on the calibration cell alone (Firms B/C/D, 2013–2019, closed-world: both the source signatures and their comparison set drawn from that cell; Section IV-C). We use this closed-world value as the primary cut rather than the corpus-wide crossover, so that the one data-derived threshold in the rule is estimated only on the calibration-only Firms-B/C/D 2013–2019 cell, held out from Firm A and from post-2020 scoring. The cut is stable across scopes — 0.8547 (calibration closed-world), 0.8367 corpus-wide, 0.8489 on the all-period baseline firms, 0.8302 with the non-Big-4 firms added; it moves by at most 0.025 across all four scopes (0.018 from the corpus-wide value), so the choice of scope is immaterial and the broader-scope values stand as robustness checks (Section V-C). The high cosine cut, 0.95, is the high-similarity operating point: it sits in the region where genuine reuse concentrates — the byte-identical anchor (Section IV-C) lies at cosine 1 — and a recalibration cannot move it onto a distributional antimode because none exists (no within-population bimodality, Section V-A). The near-identical structural cut, dHash ≤ 5 , is the perceptual-hash

distance below which two rasters are pixel-equivalent up to mild recompression, and $dHash \leq 15$ bounds the looser “structurally similar” band; both follow the standard 64-bit dHash distance scale [27]. We therefore do not re-derive these three as optimal cutoffs but characterize their chance-of-firing behavior directly (the full prior-calibration provenance is in the supplementary materials), and we make them operator-tunable in one direction: their specificity proxy at these values is read off the chance-rate calibration below, and an operator can tighten the floor by inverting the ICCR curve (for example, $dHash \leq 3$). This is a conservativeness dial, not a precision–recall control: tightening raises the specificity proxy and lowers the flag count, but there is no observable recall to trade back, so loosening cannot be calibrated against a known cost. We deliver these as a concrete, calibrated operating point — in particular the high-confidence (HC) rule, $\cosine > 0.95$ and $dHash \leq 5$ — whose between-accountant coincidence behavior the calibration below makes explicit. Because the rule is calibrated on a large Chinese-signature corpus, the HC values double as a practical starting reference for practitioners working with comparable Chinese-signature image pipelines, rather than a setting to transplant unchanged.

Figure 3. Two-measure plane: real density over the five regions (Big-4, $n=150,442$)

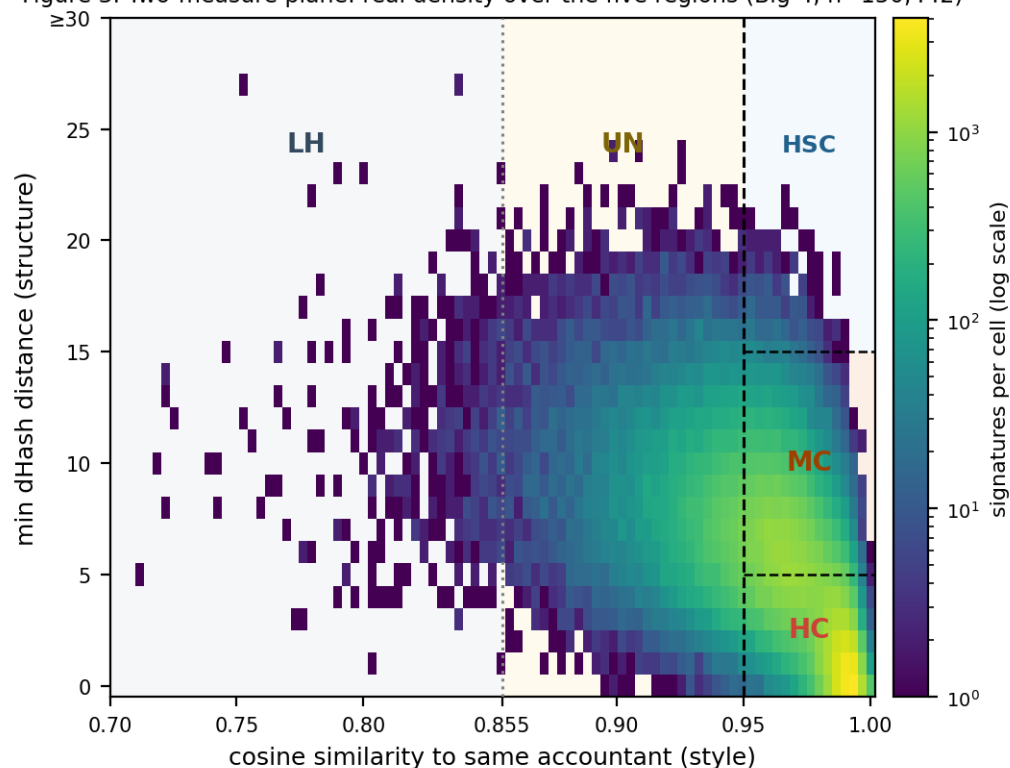


Figure 3. The two measures and the five regions, drawn as the real 2D density of all Big-4 signatures ($n = 150,442$; log color scale, integer dHash bins). The cosine axis is split at the low cut 0.8547 (the calibration-cell same-vs-different-accountant crossover) and the high cut 0.95; within the high-cosine band the dHash axis is split at 5 and 15. The mass concentrates in the bottom-right HC corner — high cosine with near-identical structure — and thins out as a single continuum toward lower cosine and higher dHash, with no gap separating a “reuse” cluster from a “hand-signed” one (Section V-A); note also that essentially all signatures sit above cosine ≈ 0.85 , the compressed high-similarity range discussed in Section V-A.

The calibration sample itself (Firms B/C/D, 2013–2019). The chance-rate calibration that follows is computed on the clean cell only, and the reader should be able to see the calibration base directly rather than infer it from the full-period totals above. The Firms-B/C/D 2013–2019 cell contains 226 accountants, 52,071 signatures with both measures computed, and 26,042 reports; the per-comparison ICCR below is estimated from 5×10^5 inter-CPA signature pairs sampled uniformly from this cell. Every ICCR source signature is restricted to this cell — the headline per-signature and per-document rates reproduce on the 52,071-signature 2013–

2019 cell, not on the full-period BCD record (~90,000 signatures), which is used only where a robustness figure is explicitly quoted — so no post-2020 or Firm-A signature enters the calibration.

How often the strict rule fires by chance (pooled). In the Firms-B/C/D 2013–2019 group, the strict (HC) rule fires by chance very rarely at every level (Table II): about 1 in 100,000 per comparison (Wilson 95% CI [4×10^{-6} , 2.3×10^{-5}]), 0.59% per signature ([0.45%, 0.73%]), and 1.2% per report. These are roughly ten times lower than the contaminated all-four-firm figures (1.4×10^{-4} , 11.0%, 18.0%); the difference is exactly the within-firm matching that the clean group leaves out. So a clean group of unrelated accountants almost never produces an HC report, which makes HC a high-specificity operating point. (The per-comparison figure rests on a small number of chance hits — 5 of 5×10^5 pairs — and is best read as an order-of-magnitude value; the per-signature and per-report figures, which are well powered, carry the weight.) A low rate is not a small number at archive scale, and we state the absolute consequence plainly: applied blindly across all 150,442 analyzable Big-4 signatures, the clean per-signature rate alone would be expected to yield about 888 HC flags by chance (95% CI [677, 1,098]), scaling further if the screen is run over the full archive. This is exactly why a single HC flag is never read in isolation: the evidential weight is carried by the firm-level contrast (Section IV-C) and accountant-level aggregation (Section IV-B), not by a raw archive-wide HC count.

Table II — Chance-firing rates (ICCR) by level and group: the strict HC rule (top two rows), with the looser MC band's per-report rate shown for contrast (bottom row).

Group / rule	Per comparison	Per signature	Per report
HC — B/C/D 2013–2019 (calibration)	1.0×10^{-5} [4×10^{-6} , 2.3×10^{-5}]	0.59% [0.45%, 0.73%]	1.2%
HC — all four firms (contaminated)	1.4×10^{-4}	11.0%	18.0%
MC band (HC+MC) — B/C/D 2013–2019	—	—	≈17.5%

Each baseline firm on its own (B, C, D). Reported separately, the three baseline firms are alike and uniformly low. A logistic regression of the per-signature HC flag on firm (with Firm D as the reference) over the baseline cell puts Firms B and C within about $3.5 \times$ of each other (odds ratios 1.73 and 0.49), and none of them comes close to the high rates we see for Firm A in Section IV-C. The 2013–2019 five-way breakdown for each of Firms B/C/D (counts and within-firm percentages) is reported in Table II-b; the full-period (2013–2023) breakdown is in Table IV for reference.

Table II-b — Five-way breakdown for each baseline firm, calibration period (B/C/D, 2013–2019).

Firm	HC	MC	HSC	UN	LH	signatures
Firm B	29.04%	39.31%	0.39%	30.91%	0.35%	19,677
Firm C	21.59%	42.09%	0.37%	35.53%	0.43%	22,449
Firm D	22.01%	29.67%	0.20%	47.35%	0.76%	9,945

One point in Table II-b needs to be made explicit, because at first glance it looks like a contradiction: the within-firm HC percentages here (Firm B 29.0%, Firm C 21.6%, Firm D 22.0%) are an order of magnitude above the 0.59% chance rate of Table II, even though both are computed on the same clean calibration cell. They are not in tension, because they measure different things. The 0.59% is a between-accountant rate — how often the HC rule fires on the signatures of two different accountants — and that is the quantity the calibration uses and that stays tiny. The 21–29% figures are within-accountant rates — how often an accountant's own signatures fire the rule against each other — and a substantial within-accountant rate is exactly what one expects from anyone with a consistent hand or a uniform house style, before any reuse is invoked; at these firms it also carries a genuine but smaller component of image reuse that grows after

2020 (Section V-B). “Clean,” in this paper, means the between-accountant coincidence is rare, not that within-accountant similarity is low. This is also why the within-accountant rate cannot be read as a false-positive rate (Section III-E), and why the contrast that isolates Firm A in Section IV-C is not “HC fires at all” but that Firm A’s within-accountant rate (82%) stands far above the 21–29% of three otherwise-alike firms.

B. From Categories to Actions: Review as Exception Management

The proportions first, stated plainly. Before saying what to do with each category, we show how much of the data falls into each (Table IV, full corpus): HC 49.6%, MC 26.5%, UN 23.4%, HSC 0.2%, LH 0.3%. The ambiguous middle (MC + UN) is therefore not a fringe: about half of all four-firm signatures, and 65–76% at Firms B/C/D individually, against 18.1% at Firm A. Read against the institutional background (Section III-A), this is exactly the expected shape. Firm A behaves in the screen as a reuse-dominated population — a reading consistent with the interviews and with the byte-identical evidence, though it rests on those rather than on per-signature ground truth — and the screen settles most of its signatures outright; Firms B/C/D in this period are mixed populations in which hand-signing and informal individual stamping coexist, so per-signature similarity is genuinely ambiguous there. The right response to a large middle is not to hide it but to give it a disposition path that does not require a per-signature verdict. Four moves do this. To be clear about what is established versus proposed here: the category proportions above, the per-band chance rates, and the byte-identity counts are empirical results, whereas the four moves are a designed operating procedure derived from the calibration — they are an argument that the workload is tractable, not a validated workflow. The protocol’s end-to-end first run on a bounded, human-labeled sample, which is what would actually measure its discriminating behavior, is left to future work (Section V); we therefore present the moves as the intended use of the calibrated rule, not as evidence in their own right.

Move 1 — calibrate each band’s evidential weight, and demote what fails. The calibration tells us what each flag is worth. The HC band fires by chance on only about 1.2% of reports in the clean reference group, so an HC flag is close to self-certifying: it needs essentially no verification effort, and it goes straight onto the action list — findings to count, report, or investigate — rather than onto a list of flags still to be checked. The MC band fires by chance on about 17.5% of reports in the clean reference group — roughly one clean-group report in six — and, unlike HC, this rate does not drop when Firm A’s accountants are excluded from the cross-accountant comparison pool (it edges up, because removing Firm A’s distinctive template leaves a pool whose members resemble one another a little more at the coarse dHash ≤ 15 scale); the boundary at dHash = 15 also sits in a flat region of the sensitivity sweep, adding flagged cases without adding specificity (Section V-C). An MC flag on its own therefore carries almost no information and does not justify verification effort; it matters only in combination with other evidence. The UN band is ambiguous in the same spirit and is treated alongside MC; on the clean baseline the UN cosine band is reached by chance about 88% of the time per signature (98.2% per report), confirming that a UN flag is essentially uninformative about reuse on its own, whereas the HSC band is reached by chance only about 0.13% of the time per signature (0.25% per report) and in any case points away from reuse (style match without structural support). The HSC band is tiny (0.2%), so it warrants only a light spot-check. The LH band needs no action. Demotion, however, only says what an MC or UN flag is not — standalone evidence; what becomes of these signatures is the business of the next three moves: their information flows into the accountant-level scores (Move 2), which byte-identity hits then sharpen by proving that an accountant’s stored image is in circulation (Move 3); the residual’s data needs are named rather than guessed at (Move 4); and where a human does look at individual cases, the bounded protocol specified below applies.

Move 2 — lift the unit of decision from the signature to the accountant. The middle categories rarely need to be resolved one signature at a time, because the operational question is almost always about an accountant or a firm, and the ambiguous signatures still carry information at that level. Three accountant-level scores — a mixture-model position score on the two-measure plane, a percentile relative to an external non-Big-4 reference population, and the accountant’s own rate of replication-consistent labels — rank the 437 accountants in close agreement (Spearman $\rho \geq 0.879$; reported as internal consistency among scores built on the same descriptors, not as external validation). A signature that is individually undecidable still moves its accountant’s position; several hundred per-signature questions collapse into one per-accountant judgment.

Move 3 — anchor with byte-identity, the one check that yields certainty. An exact byte-level comparison costs little, and what it finds is proof rather than evidence: independent hand-signing cannot produce byte-identical images, so every byte-identical pair is confirmed reuse with no human judgment required (the corpus contains 262 such signatures; Section IV-C). To be precise about where this bites: a byte-identical pair has cosine 1 and dHash 0, so these signatures sit in HC by construction — byte-identity rescues no case from the ambiguous middle. Its role is twofold. Within HC, it upgrades a subset from high-confidence candidate to logical certainty, removing even the pipeline-and-house-style caveat of Section III-D for those cases — the difference between a statistical screen and an exhibit one can act on without qualification. And at the accountant level, a byte-identical hit proves that a stored image of that accountant is in circulation, which raises the prior on the rest of that accountant’s near-identical cluster — including its MC and UN members — and thereby sharpens the per-accountant judgment of Move 2. (A byte-identical pair has cosine = 1 and dHash = 0, so it falls in HC by construction; that the rule “captures” the whole byte-identical set is therefore tautological, and we do not read it as a recall measure.)

Move 4 — state what the residual needs, instead of classifying it anyway. After the three moves, a residual middle remains whose mechanism the two measures genuinely cannot identify: reuse through a noisy pipeline, a very steady hand, and a homogeneous scanning infrastructure can occupy the same spot on the plane. We name the data that would resolve it — a proposed resolution path, not one executed in this study. Image-acquisition metadata is machine-readable provenance that could be extracted automatically rather than judged by eye: scanner identifiers and PDF-generator strings recorded in the files themselves, and compression markers such as JPEG quantization tables, which encode the processing history an image has been through. This adds the axis the two similarity measures lack — two near-identical images that arrived through different production pipelines are hard to explain except by reuse, while two that shared one pipeline may owe their similarity to the pipeline itself. (Whether this provenance survives the upload platform is itself an empirical question, and we checked: we verified across a stratified sample of MOPS reports (all four firms, 2014–2022) that producer/creator strings, PDF versions, and image encodings are heterogeneous report-to-report — distinct scanner models (Fuji Xerox D125, ApeosPort-III/IV/V), born-digital producers (Microsoft Word, Adobe, Acrobat Distiller), and a mix of CCITT-grayscale and JPEG-RGB encodings at differing resolutions — so the platform does not flatten uploads to a uniform template and the acquisition history is recoverable here; firms’ own internal archives would retain at least as much.) A small labeled set of known hand-signed examples — certified by the firms, or accumulated case by case as a by-product of the review protocol below — would turn the chance-rate calibration into directly estimated error rates. Naming these is the honest alternative to pretending the residual can be classified from similarity alone.

Where a human does look, the review follows a defined and bounded protocol. We specify the protocol here as a design deliverable of the method: the discriminating behaviors stated below are design expectations, following from the artifact properties of reused versus independently signed images, and the protocol’s first execution, on a bounded sample, is listed as future work (Section VI). (1) Side-by-side overlay inspection: the reviewer is shown the flagged signature next to the same-accountant signature(s) that produced its score, with a pixel-difference overlay and an edge-aligned superposition; a reused image is expected to overlay almost exactly, whereas two independent signings show natural variation in pressure, ink, and baseline. (2) Secondary artifact checks not used by the rule — exact registration, JPEG and scan-noise fingerprints (the compression and anti-aliasing traces a reused raster carries with it), and scaling traces — are designed to separate a reused raster from a re-scanned genuine signature at low cost. (3) Document and time context: the reviewer checks whether the matched signatures come from reports of different dates or engagements (reuse across time is more telling than within a single filing) and whether the surrounding layout shows a standard template or stamp. (4) Bounded per-accountant sampling: because the operational question is usually at the accountant or firm level, the reviewer judges a bounded random sample per accountant rather than every flagged signature, keeping the effort proportional to the number of accountants, not the number of signatures. (5) Feedback into calibration: each adjudicated case yields a label — reuse, hand-signed, or undetermined — and these accumulate into the small ground-truth set the setting otherwise lacks, which can later tighten the operating point or support supervised validation. The protocol’s relation to Move 4 is one of scale: steps 1–3 apply per-case versions of the same artifact evidence that Move 4 would collect corpus-wide, step 4 bounds how many cases a human ever sees, and step 5 accumulates the labeled set Move 4 asks for. What the protocol cannot do — and is not claimed to do — is resolve the residual at scale;

that is exactly what the corpus-wide metadata collection of Move 4 would add.

Why this is exception management rather than caseload. Where a firm's output is dominated by reuse, the high-confidence tier settles most signatures directly (82% at Firm A), and the four moves reduce the remaining 18% to per-accountant judgments and a review queue bounded by the number of accountants rather than the number of signatures — exception management at the signature level. In a mixed population the same machinery delivers the same promise one level up, at the accountant. Move 1 removes the bulk of the apparent caseload outright: an MC flag alone does not justify verification, which at Firms B/C/D takes 29–41% of signatures off the worklist before anyone is assigned to anything (the UN band carries no flag in the first place, so the demotion bites on MC alone). Move 2 positions every accountant on the replication-dominance spectrum, so attention concentrates on the few high-ranked or mixed cases rather than on tens of thousands of signatures. Move 3 supplies proof where proof exists: 117 of the 262 byte-identical signatures sit at Firms B/C/D, demonstrating that stored-image reuse is a real practice at the mixed firms too, and anchoring the accountant-level judgments there. And the staggered post-2020 adoption of formal systems gives the mixed firms a readable time axis (Section V-B). What is not delivered in a mixed population is a per-signature verdict for the ambiguous middle — a limit of identification, not of workload. Exception management therefore holds in both settings; what changes is only the level at which exceptions are defined — the signature where reuse dominates, the accountant where practices are mixed. Because the cutoffs are tunable, a reviewer who wants higher specificity can tighten them (for example, cosine > 0.98 and dHash ≤ 3), trading a lighter caseload against the risk of missing noisier reused signatures — a trade-off we cannot tune against recall, since recall is unobservable here.

C. Held-Out Benchmark: Firm A (a Known Positive)

Firm A — described by the interviews as a mainly-stamping firm, and kept out of the calibration — is our main benchmark. Because the interviews already identify it as a stamping firm, it is best read as a quasi-positive institutional benchmark: held out from calibration, but a known positive rather than a blinded out-of-sample test. What it can confirm is that the screen's measures move as expected on a firm independently believed to reuse images; what it cannot do is stand in for a blinded evaluation against ground-truth labels, which the corpus does not provide.

- (1) Firm A's two measures against the baseline. Comparing Firm A's within-accountant similarities to those of Firms B/C/D (full record, 2013–2023²), Firm A's cos values are shifted toward 1.0 and its dHash distances toward 0 — the direction we would expect if a stored image is reused rather than re-signed. Concretely, Firm A's within-accountant cosine is centered at a median of 0.986 (mean 0.980) versus 0.959 (mean 0.954) for Firms B/C/D, and its smallest-dHash distance at a median of 2 (mean 2.7) versus 7 (mean 7.0); both shifts are in the reuse direction and overwhelmingly significant (Mann–Whitney U, $p < 10^{-300}$ for each; two-sample Kolmogorov–Smirnov $D = 0.60$ for cosine and 0.57 for dHash). The decisive number is this: scored as a held-out (but not blinded) case — Firm A's signatures matched against unrelated accountants drawn from the clean 2013–2019 group — Firm A's per-signature cross-firm HC rate is 0.42% (154/36,552; Wilson 95% CI [0.36%, 0.49%]), at or below the clean reference ICCR of 0.59%. In other words, Firm A's cross-firm match rate sits at the level a clean inter-CPA comparison produces by chance — it is not elevated relative to the reference, and it is negligible beside the within-firm rate below — so the entire rise in Firm A's rate comes from matches with other Firm-A signatures, not from resemblance to other firms. The signal is inside the firm, not across firms. (Against the full-period BCD pool the same across-firm rate is 1.0%; the small difference reflects the post-2020 rise in baseline similarity of Section V-B. Both lie at the clean floor, two orders of magnitude below the within-firm rate that follows.)

² Restricting both groups to 2013–2019 gives essentially the same picture (Firm A cosine median 0.986, dHash 2; Firms B/C/D 0.957 and 7; Mann–Whitney $p < 10^{-300}$ for each), confirming the contrast is not a post-2020 artifact.

Firm A's within-firm repeatability, against the other firms. On their own signatures, the HC rule fires on 82% of Firm A's, versus 24–35% for Firms B/C/D. We deliberately report these as raw within-accountant fire rates and do not divide them by the between-accountant clean floor: as Section III-E explains, that floor is the

wrong null for a within-accountant question, so an “X× the floor” multiplier would overstate the gap. The firm-to-firm contrast in raw rates is what carries the result. A logistic regression of the per-signature HC flag on firm and pool size, with Firm A as the reference, gives odds ratios of 0.053, 0.010, and 0.027 for Firms B/C/D — one to two orders of magnitude lower (the odds ratio for log pool size is 4.01). Firm A stands alone, against a baseline of three firms that look alike.

Four further checks confirm the contrast is not an artifact of how the comparison pools are built, of the imaging-pipeline trend, or of any single year. First, pool size. Stratifying accountants by how many signatures they contribute and comparing within each stratum, Firm A’s HC rate exceeds the other firms’ at every level — 66% versus 20% for the smallest pools (under 50 signatures), rising to 76–84% versus 21–29% for larger pools. Even Firm-A accountants with few signatures to match against fire the rule far more often than B/C/D accountants with the same pool size; pool size raises the rate within every firm (the log-pool-size odds ratio of 4.01), but the firm gap dwarfs it and survives at fixed pool size, which rules out the “more signatures, more chances for an extreme match” explanation. Second, dependence among an accountant’s own signatures. Re-estimating the gap with the bootstrap resampled at the accountant level (179 Firm-A accountants, 280 at Firms B/C/D) rather than treating signatures as independent, the Firm-A-minus-B/C/D difference in HC rate is 53.7 percentage points with a 95% interval of [49.5, 57.5] — accountant-level clustering widens the intervals the per-signature Wilson bounds give, but leaves the contrast far too large to be explained away. Third, the time trend and pipeline shift (Section V-B). Adding year fixed effects to the logistic regression — so the firm effect is identified within year, net of the 2020–2021 imaging-pipeline transition — leaves Firms B/C/D at 0.06–0.12 times Firm A’s odds of an HC flag (odds ratios 0.116, 0.061, 0.070), still an order of magnitude lower once the common time trend is absorbed. Fourth, single-year dependence. Leaving out each calendar year in turn and recomputing, the Firm-A-minus-B/C/D gap stays within 53.1–54.9 percentage points (full-sample 53.7), so neither the high-reuse digital-native years (2022–2023) nor any earlier year drives it.

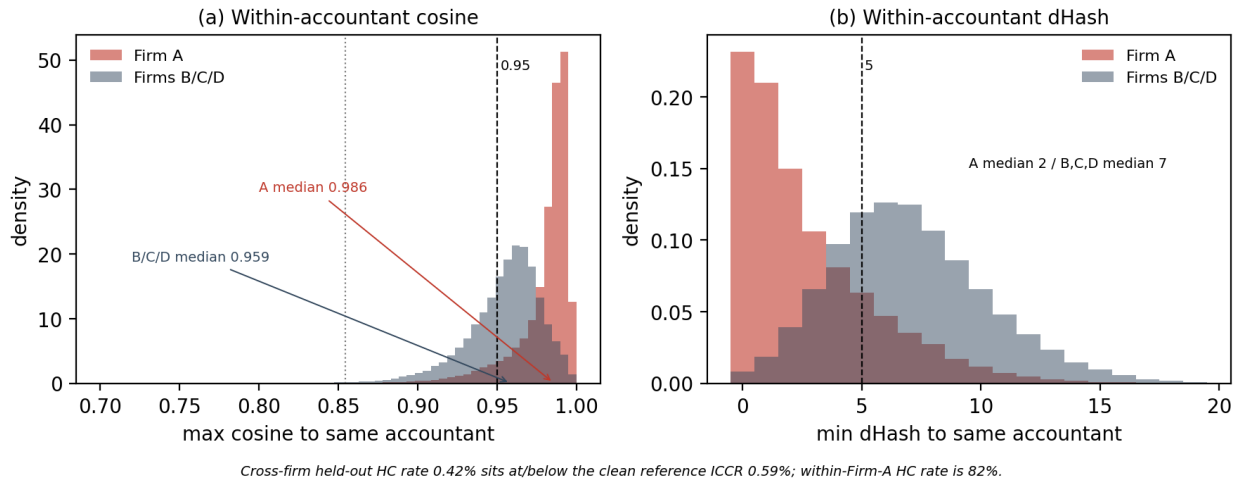


Figure 4. Within-accountant similarities, Firm A vs Firms B/C/D: (a) cosine, (b) dHash. Firm A’s mass sits near cosine = 1 and dHash = 0 (medians 0.986 / 2) against Firms B/C/D’s 0.959 / 7; dashed lines mark the cuts (cosine 0.95; dHash 5), the dotted line the LH/UN crossover (0.8547). The held-out cross-firm HC rate (0.42%) sits at/below the clean reference ICCR (0.59%), while the within-Firm-A HC rate is 82% — the signal is inside the firm (annotation below panels).

- (2) Ranking accountants by similarity, in each period. Ranking every accountant in Firms A–D by a single within-accountant similarity score, separately for 2013–2019 and for 2020–2023, Firm A’s accountants sit at the high-similarity (templated) end. A descriptive three-group summary of the two-measure space tells the same story: its high-cosine/low-dHash group holds 82.5% of Firm A’s accountants and almost none of the others’ (Table III). The period split confirms the expected pattern: Firm A’s per-signature HC rate is at the top in both periods (80.3% in 2013–2019, 83.8% in 2020–2023), while Firms B/C/D move upward after 2020 as the formal systems came in — Firm B from 29.0% to 42.0%, Firm C from

21.6% to 26.7%, Firm D from 22.0% to 28.0% (see Section V-B).

Table III — Firm by descriptive-group membership (whole corpus). The “high-cosine/low-dHash group” is the templated-end cluster of the three-group ($K = 3$) descriptive Gaussian-mixture partition of the accountant-level two-measure plane (Section V-C); membership is the cluster of maximum posterior probability for each accountant with at least ten signatures (437 accountants: 171/112/102/52 across Firms A–D). The groups are used for description only, never as operational labels.

Firm	Accountants	Share in the high-cosine/low-dHash group
Firm A	171	82.5%
Firm B	112	0.0%
Firm C	102	1.0%
Firm D	52	1.9%

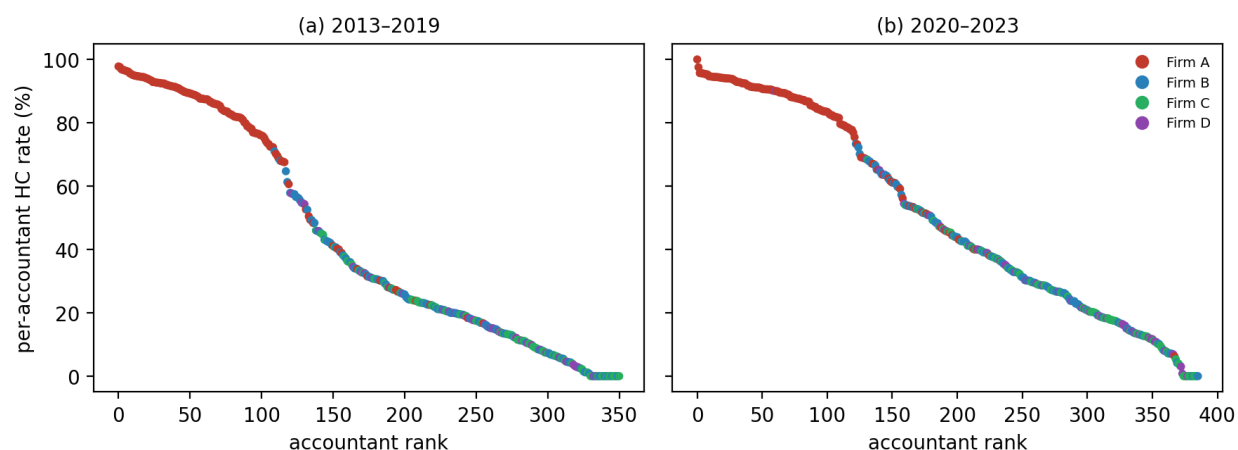


Figure 5. Per-accountant HC rate, ranked, one panel per period (2013–2019; 2020–2023), points colored by firm (accountants with ≥ 5 signatures in the period). Firm A (red) occupies the templated top of the ranking in both periods; Firms B/C/D rise after 2020 (HC rate B 29.0→42.0%, C 21.6→26.7%, D 22.0→28.0%; Firm A 80.3→83.8%), consistent with staggered formal-system adoption (Section V-B).

(3) Applying the calibrated rule to Firm A, 2013–2023. Taking the operating point calibrated on Firms B/C/D in 2013–2019 and applying it across Firm A’s full record, 81.70% of Firm A’s signatures (82% rounded) land in HC (per signature; the full five-way breakdown is in Table IV). Read together with the interview fact that Firm A mainly uses overlay stamping, the system’s firm-level output matches the practice the firm itself describes. We say this carefully: it is a match at the firm level, not a label on any single signature. We do not classify the individual signatures as non-hand-signed, because for any one signature the two measures cannot separate reuse from a shared scanning pipeline or a uniform house style (Section III-D).

Table IV — Five-way breakdown by firm (whole corpus, 2013–2023; for reference; $n = 150,442$).

Firm	HC	MC	HSC	UN	LH	signatures
Firm A	81.70%	10.76%	0.05%	7.35%	0.14%	60,448
Firm B	34.56%	35.88%	0.29%	28.95%	0.32%	34,248
Firm C	23.75%	41.44%	0.38%	33.97%	0.47%	38,613
Firm D	24.51%	29.33%	0.22%	45.28%	0.66%	17,133
Overall	49.58%	26.47%	0.21%	23.42%	0.32%	150,442

Reading the five-way mix across firms. Table IV is also the quantitative basis for the positioning in Section IV-B. At Firm A the ambiguous middle (MC + UN) is 18.1% — the screen reads this population almost cleanly, with four signatures in five settled outright. At Firms B/C/D the middle is 65–76% — the signature of a mixed population in which hand-signing and informal stamping coexist (Section III-A), where per-signature similarity is genuinely ambiguous. There the screen’s deliverables move up one level (Section IV-B): the MC share (29–41% of these firms’ signatures, against the 26.5% corpus-wide MC share) is demoted off the worklist, the accountant-level scores rank these firms’ accountants alongside everyone else’s, and the byte-identical signatures at these firms (117 of the 262) are threshold-free proof that reuse occurs there too. The per-signature mix stays ambiguous; the disposition does not.

Byte-identical signatures: direct evidence of reuse. Beyond the screening numbers, 262 signatures across the four firms are byte-for-byte identical to another signature — 145 of them at Firm A, spread across about fifty partners. Identical files cannot come from independent hand-signing, so their existence is direct, hard evidence that image reuse happens and that it concentrates at Firm A. These pairs are not a bookkeeping artifact: every one of the 262 matches a signature in a different report PDF (none is the same file double-counted), and 170 of the 262 fall in different filing months, so duplicate filings or corrected re-submissions of one report cannot explain them. One caveat belongs with this count, developed in Section V-B: most of the 262 (232) occur in the post-2020 digital-native era, where exact reuse is both easier and perfectly preserved, so the raw count is not a clean prevalence trend; the pipeline-independent core is the 30 in the pre-2021 pure-scan era (18 at Firm A), which scanning noise alone cannot produce. Because a byte-identical pair has cosine = 1 and dHash = 0, it lands in HC by definition; the rule’s “100% capture” of this set is therefore tautological, and we do not read it as a sanity check or a lower bound on recall. We use byte-identity only for what it can show directly — that reuse occurs and where it concentrates — as a prevalence signal, not a measure of detector performance.

V. Other Analyses

This section gathers analyses that support the design and test its robustness: (a) the diagnostic showing the data contain no natural cutoff — the premise the whole calibration rests on; (b) how the baseline behaves after 2020; and (c) sensitivity checks.

A. Why the Data Contain No Natural Cutoff

This diagnostic backs the design choice announced in Section III-D and Section III-E: that no cutoff can be read off the data, so the operating point has to be set from an outside reference. The Hartigan dip test [37] rejects a single-peak shape for both measures at the Big-4-pooled accountant level ($p < 5 \times 10^{-4}$), which might look like a clean split into two groups. But that rejection comes from two side-effects. Once we remove the differences between firms (by centering each firm on its own mean) and the effect of the hash taking only whole-number values (by adding a small jitter to dHash), the single-peak shape comes back (median $p = 0.35$ over jitter seeds). Tested firm by firm, each Big-4 firm is already unimodal on both axes (Firm A $p_{\text{cos}} = 0.99$, $p_{\text{dHash}} = 0.92$; B/C/D pooled $p_{\text{cos}} = 0.998$, $p_{\text{dHash}} = 0.91$), so the pooled rejection is a between-firm location-shift artifact, not within-population bimodality. A density-smoothness test in the Burgstahler–Dichev / McCrary style [38], [39] finds no real break in either measure at the Big-4 scope (Appendix A.1 shows the apparent signature-level breaks drift with histogram bin width and sit inside the high-similarity region — a resolution artifact, not an antimode). So the data hold no real gap; per-signature similarity is best read as one continuous spread of quality, not two separate classes. This is exactly why the operating point is set from an outside reference (Section III-E) rather than read off the data, and why the three groups used for description in Section IV-C are treated as a summary of composition, not as real mechanisms.

A property of the cosine measure reinforces this and explains why the rule never leans on cosine alone. On these fixed-size, white-padded, ImageNet-normalized crops the within-accountant cosine is compressed into a narrow band at the top of its range: 97.7% of signatures score above 0.90, the median is 0.969, and only 0.3% fall below 0.85 (Appendix). Two signatures of the same accountant are highly cosine-similar whether the hand is steady or the image is reused, because the shared crop geometry and normalization

contribute a common-structure baseline before any signature content is compared. The high cosine cut (0.95) therefore sits inside this saturated region — about three-quarters of signatures lie above it — so cosine on its own separates almost nothing; the structural dHash measure does the discriminating, which is why HC requires both and the cosine-only HSC band carries no evidential weight (Section III-D). We do not try to decompose this cosine baseline into its preprocessing and genuine-style parts here: doing so cleanly would mean re-extracting features with the padding and normalization ablated, which we flag as the way to quantify the preprocessing contribution and as a construct-validity check for future work.

B. Time Trend and the Firm–Pipeline Confound (Secondary)

Looking only at Firms B/C/D, the strict rule’s chance rate rises after 2020 (per comparison from 1.0×10^{-5} to 3.6×10^{-5} ; per signature from 0.59% to 1.05%), and the deployed HC rate rises in parallel (Firm B 29.0→42.0%, Firm C 21.6→26.7%, Firm D 22.0→28.0% across the two periods, Section IV-C). The rise is heterogeneous in timing rather than a common step. Tracing the yearly HC rate, Firm C’s increase is concentrated in 2022 (about 18% through 2021, then ~30% in 2022 and ~40% in 2023) and Firm B’s mainly in 2023 (about 33% in 2022, ~54% in 2023), while Firm D rises gradually with no visible step; Firm A, by contrast, is already high throughout the decade (80.3→83.8%) with no adoption-like jump — consistent with the interviews’ account of long-standing stamping. This firm-by-firm staggering is what one would expect from progressive, independent adoption of formal signing systems (Section III-A), and it is why we limit the calibration to the pre-2020 years. Table II-c gives the full five-way breakdown by firm for the 2020–2023 deployment period, as a companion to the calibration-period Table II-b and for direct cross-checking against the proportions quoted here and in Section IV-C.

Table II-c — Five-way breakdown by firm, deployment period (Firms A–D, 2020–2023).

Firm	HC	MC	HSC	UN	LH	signatures
Firm A	83.84%	9.13%	0.04%	6.88%	0.11%	23,898
Firm B	42.01%	31.24%	0.16%	26.31%	0.28%	14,571
Firm C	26.74%	40.55%	0.40%	31.80%	0.51%	16,164
Firm D	27.98%	28.85%	0.24%	42.42%	0.51%	7,188

We deliberately stop short of reading this as a detected e-signing effect, because of a confound these data cannot break: firm identity — and period within a firm — bundles signing practice together with the entire imaging pipeline, and that pipeline demonstrably changes across the decade. We audited the production provenance of a stratified sample of 880 report PDFs (20 per firm-year) from their embedded metadata and page structure. The shift is stark (Table V): through 2020, reports are overwhelmingly plain scanned rasters — 70–85% in the early years carry no text layer at all, and their PDF metadata names the scanning hardware directly (for example “Fuji Xerox D125” and “ApeosPort-IV 7080”) — whereas from 2021 plain scans collapse to about 1–2% as firms move to OCR’d and digital-native production. The two similarity measures are therefore computed on a substrate that itself transforms around 2020–2021, exactly when the baseline firms’ similarity rises; firms also differ from one another in this respect (Firm A adopts digital-native output earliest, Firm C latest), though the cross-firm gap is much smaller than the temporal one. A post-2020 rise in similarity could thus come from this coincident pipeline change just as easily as from a change in how signatures are applied (Section III-D), and with no labels and no externally-dated adoption events the two are not separable here.

Table V — Imaging-pipeline audit: production type by year (stratified sample, 880 PDFs, 20 per firm-year). “Scanned” = no extractable text layer; “digital-native” = text-based PDF with embedded image objects; the remainder are scanned-then-OCR’d.

Year	Scanned %	Digital-native %	Year	Scanned %	Digital-native %
2013	82	0	2019	56	0
2014	76	0	2020	52	0
2015	85	0	2021	1	7
2016	70	2	2022	2	16
2017	50	0	2023	2	30
2018	55	0			

This same transition qualifies the byte-identity evidence (Section IV-C), which we flag rather than let the raw count mislead. Of the 262 byte-identical signatures, 232 fall in the digital-native era (2021–2023), where embedding a discrete signature image makes exact reuse both easy to do and perfectly preserved — so the post-2020 surge in byte-identical pairs is inflated by detectability and should not be read as a purely behavioral increase. The pipeline-robust core is the 30 byte-identical signatures in the pre-2021 pure-scan era, 18 of them at Firm A: two independently hand-signed pages, separately scanned, cannot yield byte-identical crops, so these are direct evidence of digital image reuse that predates the digital-native transition and concentrates at Firm A. This is also why Firm A’s elevation, present throughout the scanned years, cannot be an artifact of digital-native embedding.

The clean way to separate them would be an event study that aligns each firm to its own externally-documented e-signing adoption date and absorbs static firm differences and common-year shocks with firm and year fixed effects; a within-firm jump locked to each firm’s own adoption month would be evidence for signing practice over static pipeline. We do not possess those adoption dates — and inferring them from the very HC series we would then test would be circular (Section III-E) — so we flag this event study as the natural next step rather than a result we can claim here. For this paper the trend serves only its narrower purpose: it justifies the pre-2020 calibration window and stands as a robustness check, not as a causal finding.

C. Sensitivity and Robustness

We summarize the robustness checks here; full detail is in the supplementary materials.

How sensitive the operating point is. Right around the HC cutoff the per-signature firing rate changes quickly — its local slope is about 25× the median across a cosine sweep and about 3.8× across a dHash sweep — which confirms that the HC point is a chosen, specificity-anchored operating point rather than a natural gap.

A single slope understates how the rule behaves, so we map the full surface rather than defend one cut. Figure 6 plots, over the entire (cosine cut × dHash cut) plane, the clean-group flag rate (panel a) and the Firm A – B/C/D flag-rate contrast (panel b), and neither view favors the chosen cut by construction. First, the surfaces are smooth: there is no cliff at (0.95, dHash ≤ 5), so the operating point is a readable choice on a continuous trade-off rather than a discovered boundary (Section V-A), and an operator who wants a tighter floor can move toward higher cosine and lower dHash and read the consequence off the surface. Second, the firm contrast is not an artifact of the threshold: it exceeds 45 percentage points across a broad region of low-dHash, high-cosine cuts and in fact grows as the cut tightens (for example 58 pp at cosine 0.97, dHash ≤ 3), so the deliberately looser HC point trades a few points of contrast for catching more reuse, not the reverse. The same surface makes the weakness of the cosine-only direction explicit: extending the structural cut to the MC bound (dHash ≤ 15) roughly halves the contrast (to about 27 pp) while sharply inflating the clean-group flag rate. That is precisely why the MC band is only advisory and the cosine-only HSC band carries no weight (Section III-D): the partition is not drawn to flatter the narrative, and the surface shows directly where each band earns its keep and where it does not.

Figure 6. Sensitivity surface of the deployed rule over the two-measure threshold plane (Big-4, $n=150442$).

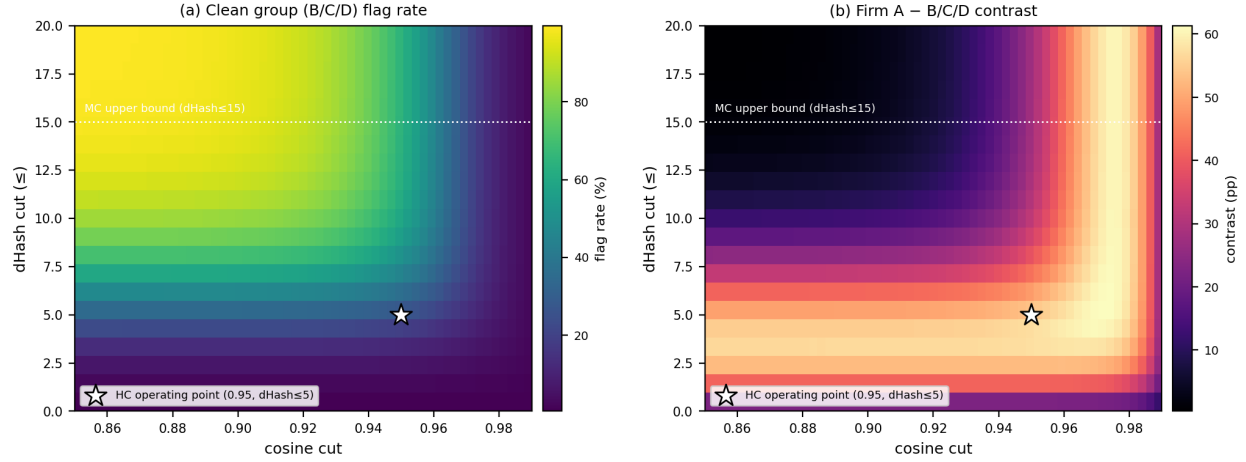


Figure 6. Sensitivity surface of the deployed rule over the two-measure threshold plane (Big-4, $n = 150,442$). (a) Clean-group (B/C/D) flag rate at each (cosine cut, dHash cut); the chosen HC operating point (star) sits in a low-rate, high-specificity region with no cliff. (b) Firm A minus B/C/D flag-rate contrast (percentage points); the contrast exceeds 45 pp across a broad low-dHash, high-cosine band and weakens toward the MC bound (dHash ≤ 15 , dotted), so the operating point is not a cherry-picked threshold and the MC band is visibly the less discriminating region. The MC/HSC boundary at dHash = 15 sits in a flat (saturating) region, where moving the line adds flagged cases without adding specificity; this is a further reason to treat the MC band as advisory (Section IV-B).

Leaving out one firm at a time. A two-group fit is unstable across firms — its boundary is basically a “Firm A versus the rest” divider — while a three-group fit keeps a stable shape (its low-cosine/high-dHash group drifts by at most 0.005 in cosine) but a membership that shifts with the mix of firms (by up to 12.8 percentage points). So we use the groups only as descriptions, never as operational labels.

Crossover scope. The low cosine cut is the same-vs-different-accountant cosine crossover; recomputing it across scopes moves it by at most 0.025 — 0.8547 on the calibration cell (the primary value; Section IV-A), 0.8367 corpus-wide, 0.8489 on the all-period baseline firms, 0.8302 with the non-Big-4 firms added — and because the cut affects only the UN/LH boundary, switching among these scopes changes no HC/MC/HSC result and shifts the UN/LH split by at most 0.4 percentage points per firm. We use the calibration-cell value as primary for held-out discipline and report the others as robustness.

The same-pair variant. A reader may worry that the deployed rule is a derived statistic rather than an observation: the cosine maximum and the dHash minimum are each taken over the accountant’s pool and can originate from different partner signatures, so the high-confidence region might in principle be assembled from two unrelated extrema. We therefore recompute the rule under the strict same-pair construction, where a single partner signature must satisfy both inequalities at once (Section III-D), and report it in the main text rather than the supplement. Two views agree. First, the within-firm concentration of cross-accountant matches is higher under same-pair (97.0–99.96% across the four firms) than under the deployed any-pair rule (76.7–98.8%). Second, and more directly, the per-signature HC flag rate — the quantity the any-pair concern targets — behaves the same way (Table VI): requiring one partner to satisfy both inequalities lowers every firm’s rate, as expected, but it widens the firm gap rather than narrowing it. Firm A still fires on a majority of its own signatures (57.3%) while the baseline firms fall to 5–9%, so the Firm-A-to-baseline ratio rises from about 2.4–3.4 $\times$ under any-pair to about 6.4–10.8 $\times$ under same-pair. The high-confidence region is therefore not an artifact of combining extrema from different partner signatures; pushed to the stricter event, the structure gets stronger.

Table VI — HC flag rate by firm under the deployed any-pair rule and the strict same-pair rule.

Firm	Signatures	Any-pair HC	Same-pair HC
Firm A	60,448	81.7%	57.3%
Firm B	34,248	34.6%	9.0%
Firm C	38,613	23.7%	5.3%
Firm D	17,133	24.5%	7.7%
All Big-4	150,442	49.6%	27.3%

Each gate adds specificity. On the all-four-firm pool the cosine gate alone fires per comparison at 6.0×10^{-4} ; adding the structural gate multiplies this by 0.234 (the conditional ICCR of $\text{dHash} \leq 5$ given $\cos > 0.95$), giving the joint 1.4×10^{-4} . Each axis contributes specificity beyond the other — quantitative support for the two-gate design over either measure alone (Section I, Section III-D).

Which network we use. We compare ResNet-50 against VGG-16 and EfficientNet-B0 under the same preprocessing and L2 normalization (Appendix A; supplementary backbone-ablation table). EfficientNet-B0 gives the largest intra/inter separation (Cohen’s $d = 0.707$) but also the widest descriptor spread (intra std 0.123 vs ResNet-50’s 0.098); VGG-16 is worst on every key metric despite its larger 4096-dim features. ResNet-50 is the best overall balance: its Cohen’s d (0.669) is competitive, its tighter distributions give more stable per-signature behavior, it yields the highest Firm A all-pairs 1st-percentile similarity (0.543), and its 2048-dim features are a practical compromise for processing 182K+ signatures. The comparison supports ResNet-50.

D. Threats to Validity

For the reader’s convenience we collect the main threats to validity in one place, each with a pointer to where it is treated and, where relevant, the direction of its bias. They are consequences of working without labels, and we state them as limitations rather than dissolve them.

1. No signature-level ground truth. The archive labels no signature as hand-signed or reused, so we report no recall, precision, ROC-AUC, or false-rejection rate, and every rate is a chance rate, not an error rate (Section III-D, Section VI).
2. Wrong null for the reuse question. The ICCR is a between-accountant coincidence rate; the within-accountant false-positive rate the question needs is not estimable and the ICCR is not even a bound on it, so “ $X \times$ the floor” comparisons are avoided as anti-conservative (Section III-E, Section IV-C).
3. Reference-group contamination / circular selection. The clean floor is conditional on the reference group truly being clean; undetected reuse there would only raise the floor, biasing the Firm-A contrast conservatively, and the floor does not hinge on any single baseline firm (Section III-E).
4. Pool-size and extremal dependence. The rule takes a maximum over a pool, so larger pools mechanically raise fire rates; the firm contrast nonetheless holds within every pool-size stratum and under accountant-clustered resampling (Section IV-C).
5. Firm–pipeline confound. Firm identity bundles signing practice with the imaging pipeline (crop geometry and reuse rates differ by firm), and internal timing cannot separate the two without externally-dated adoption events; a fixed-effects event study is the natural next step (Section V-B).
6. Preprocessing and construct validity. Padding and ImageNet normalization compress cosine into a narrow high band, so cosine alone discriminates little and the rule relies on the structural measure; a padding/normalization ablation is needed to quantify the preprocessing contribution (Section V-A).
7. Generalizability. Calibration is on a Chinese-signature corpus from one jurisdiction with a specific pipeline; the operating point is a starting reference for comparable pipelines, not a transplantable constant, and requires recalibration elsewhere (Section III-D, Section VI).
8. Non-reproducible corroboration and an unrun protocol. The interviews are self-reported and not reproducible, so agreement with them shows consistency with domain knowledge, not measured accuracy (Section III-A); and the review protocol of Section IV-B is a designed procedure whose validating first run remains future work.

VI. Conclusion

We have presented a label-free, anchor-calibrated way to screen for non-hand-signed signatures in large numbers of audit reports. It has three working parts — a pipeline that takes raw PDFs through page-finding, detection, feature extraction, and a two-measure similarity step; a pair of measures that separate style consistency from image reproduction; and, in place of a natural cutoff we do not have and labeled data we cannot get, a calibration based on how often the rule fires by chance in a clean reference group. That calibration yields both a between-accountant specificity proxy and a concrete operating point: the high-confidence rule almost never fires by chance on the clean group, so it is a usable, highly specific screen, with a defined, bounded human-review protocol (Section IV-B) for the advisory and uncertain cases. Operationally the screen earns its keep in two ways: run over an archive, it discovers where reuse concentrates; and it keeps human review at the scale of exceptions in both kinds of population — settling most signatures directly where reuse dominates, and, where practices are mixed, demoting the low-specificity band, ranking accountants, and confirming the byte-identical cases, withholding only the per-signature verdict for the ambiguous middle. We report the category proportions that make that distinction concrete. Because it is calibrated on a large Chinese-signature corpus and uses script-agnostic image descriptors, the rule offers a practical starting reference for comparable Chinese-signature pipelines and, in principle, an approach portable to other scripts — subject in each case to recalibration on the new setting. Held out as a known-positive benchmark, one firm stands alone in how alike its own signatures are, its output matches the stamping practice the firm itself describes, and byte-identical signatures give direct evidence that reuse happens and concentrates there.

The limits are built into working without labels, and we have stated them alongside the design. There is no signature-level ground truth, so we report no false-rejection rate, recall, ROC-AUC, or precision; every rate we give is a between-accountant chance rate read as a proxy for specificity, not a true false-acceptance rate, and not even a bound on the within-accountant false-positive rate the reuse question would need (Section III-E). The contrast between firms is something the method can see, not a finding about why the signatures look alike: for any single signature, the two measures cannot separate reuse from a shared scanning pipeline or a uniform house style, and for Firms B/C/D we make no claim about firm practice at all. Whether firm-level signing patterns matter for audit quality is a question for a dedicated companion study — one this screening points toward, together with the low-presence character of proxy-executed stamping shown in the behavioral literature, but one that similarity alone cannot settle.

Four directions follow. First, a set-level reading of each accountant: judging the shape of an accountant's whole signature set — a tight cluster that recurs near-identically across reports and years (the signature of a stored image) versus a dispersed cloud (the signature of a hand) — instead of per-signature extrema. This would collapse much of the remaining middle into a few per-accountant cluster decisions, and it is the natural tool for separating the mixed signers of the baseline firms, whose sets may contain both a tight recurring sub-cluster and a dispersed remainder if both practices are present. We view this as the highest-value methodological extension, while noting honestly that it narrows but does not remove the fundamental ambiguity: a very steady hand and a noisy reused image can still meet in the middle of any set-level statistic. A first-pass probe on the calibration cell is consistent with this caution — across the 206 of the 226 Firms-B/C/D 2013–2019 accountants with enough signatures for a set-level shape to be estimated, the within-accountant similarity forms a continuum that piles up just below the high-similarity cut rather than splitting into a tight reused cluster and a dispersed hand-signed cloud (no accountant shows a tight-versus-remainder cosine gap above 0.10), so the no-natural-cutoff structure of Section V-A recurs at the accountant level; we therefore treat set-level adjudication as a research direction rather than a ready robustness result. Second, executing the review protocol of Section IV-B on a bounded sample — its first run — would both test the protocol's expected discriminating behavior and accumulate the small human-labeled set that permits supervised validation and direct error rates. Third, image-acquisition metadata (scanner identifiers, PDF-generator fingerprints, compression markers) adds a provenance axis that could help resolve the pipeline-versus-reuse ambiguity similarity alone cannot; we confirmed this metadata survives in the present corpus rather than being flattened by the platform, though its discriminative power remains to be validated (Section IV-B, Move 4). Fourth, the audit-quality question itself: whether firm-level signing patterns correlate with audit outcomes, for which this screening supplies the measurement layer.

Appendix A. Supplementary Diagnostic Detail

A.1. BD/McCrory Bin-Width Sensitivity (Signature Level)

The main text (Section III-D, Section V-A) treats the Burgstahler–Dichev / McCrory discontinuity procedure [38], [39] as a density-smoothness diagnostic rather than as a threshold estimator. This subsection documents the empirical basis for that framing by sweeping the bin width across four (variant, bin-width) panels: Firm A and full-sample, each in the cosine and dHash direction.

Table A.I. BD/McCrory bin-width sensitivity (two-sided $\alpha = 0.05$, $|Z| > 1.96$).

Variant	n	Bin width	Best transition	z_below	z_above
Firm A cosine (sig-level)	60,448	0.003	0.9870	-2.81	+9.42
Firm A cosine (sig-level)	60,448	0.005	0.9850	-9.57	+19.07
Firm A cosine (sig-level)	60,448	0.010	0.9800	-54.64	+69.96
Firm A cosine (sig-level)	60,448	0.015	0.9750	-85.86	+106.17
Firm A dHash (sig-level)	60,448	1	2.0	-4.69	+10.01
Firm A dHash (sig-level)	60,448	2	no transition	—	—
Firm A dHash (sig-level)	60,448	3	no transition	—	—
Full-sample cosine (sig-level)	168,740	0.003	0.9870	-3.21	+8.17
Full-sample cosine (sig-level)	168,740	0.005	0.9850	-8.80	+14.32
Full-sample cosine (sig-level)	168,740	0.010	0.9800	-29.69	+44.91
Full-sample cosine (sig-level)	168,740	0.015	0.9450	-11.35	+14.85
Full-sample dHash (sig-level)	168,740	1	2.0	-6.22	+4.89
Full-sample dHash (sig-level)	168,740	2	10.0	-7.35	+3.83
Full-sample dHash (sig-level)	168,740	3	9.0	-11.05	+45.39

Two patterns are visible. First, the procedure consistently identifies a “transition” under every bin width, but the location drifts monotonically with bin width (Firm A cosine: $0.987 \rightarrow 0.985 \rightarrow 0.980 \rightarrow 0.975$ as bin width grows from 0.003 to 0.015; full-sample dHash: $2 \rightarrow 10 \rightarrow 9$ as bin width grows from 1 to 3), and the Z statistics inflate superlinearly with bin width because wider bins aggregate more mass and shrink the per-bin standard error on a very large sample. Both features are characteristic of a histogram-resolution artifact rather than a genuine density discontinuity. Second, the candidate transitions all locate inside the high-similarity region

(cosine ≥ 0.975 , dHash ≤ 10) rather than at a between-mode boundary. Taken together, the signature-level BD/McCrory transitions are not a threshold in the usual sense — they are histogram-resolution-dependent local density anomalies inside the high-similarity descriptor region rather than between modes — which supports using BD/McCrory as a density-smoothness diagnostic, not a threshold estimator (Section V-A).

A.2. Diagnostic Summary

The unsupervised-diagnostic strategy is a set of complementary checks, each addressing one specific failure mode of an unsupervised screening classifier under an explicitly disclosed untested assumption.

Table A.II. Diagnostics, failure mode addressed, and disclosed untested assumption (abridged).

Diagnostic	Failure mode addressed	Disclosed untested assumption
Composition decomposition (Section V-A)	Whether descriptor multimodality is within-population (mechanism) or between-group (composition + integer artifact); $p_{\text{median}} = 0.35$ under joint firm-mean centering + integer-tie jitter	Integer-tie jitter and firm-mean centering are unbiased over the descriptor support
Per-comparison ICCR (Section IV-A)	Pair-level specificity proxy under a random-pair negative anchor, on the BCD baseline	Inter-CPA pairs are negative; addressed by anchoring on B/C/D and holding Firm A out
Pool-normalised per-signature ICCR (Section IV-A)	Deployed-rule specificity proxy at per-signature unit, accounting for pool size	As above + pool replacement preserves the negative-anchor property
Document-level ICCR (Section IV-A)	Operational alarm-rate proxy at per-document unit (HC and HC+MC)	As above
Firm-heterogeneity logistic regression (Section IV-C)	Multiplicative effect of firm membership on per-signature rate, controlling for pool size	Observations clustered by CPA/firm; cluster-robust SEs are a future check
Cross-firm hit matrix (Section IV-C, Section V-C)	Concentration of inter-CPA collisions within source firm	Concentration depends on deployed-rule semantics (same-pair 97.0–99.96% vs any-pair 76.7–98.8%)
Alert-rate sensitivity sweep (Section V-C)	Local sensitivity of the deployed rule to threshold perturbation	Gradient comparison is descriptive, not a formal plateau test
Convergent score Spearman ranking (Section IV-B)	Internal consistency of three feature-derived per-CPA scores	Scores share inputs; not statistically independent
Pixel-identical positive capture (Section IV-C)	Prevalence evidence that reuse occurs and where it concentrates	Anchor is tautologically captured by any reasonable threshold; not read as a recall or performance measure

Appendix B. Reproducibility Materials

The full table-to-script provenance mapping, script source code, and report artifacts for every numerical table and figure in this paper are provided in the supplementary materials. Scripts run deterministically under fixed random seeds documented there (the inter-CPA candidate sampler uses seed 42 and a retry-loop matching the canonical samplers; CPA-block bootstraps use 1,000 replicates); reviewer reproduction should re-emit artifacts from the listed scripts rather than rely on any local path layout. The calibration baseline (BCD 2013–2019), the contamination-comparison scope (all-Big-4), the Firm-A out-of-sample scoring, and the five-way

classification are all emitted by the same canonical pipeline so that the headline numbers in Tables I, II, II-b, and IV reproduce bit-for-bit.

References

References follow IEEE numeric style; entries [41]–[45] are the behavioral-science and Chinese-script works added in this draft.

- [1] Taiwan Certified Public Accountant Act (會計師法), Art. 4; FSC Attestation Regulations (查核簽證核准準則), Art. 6.
- [2] S.-H. Yen, Y.-S. Chang, and H.-L. Chen, “Does the signature of a CPA matter? Evidence from Taiwan,” *Res. Account. Regul.*, vol. 25, no. 2, pp. 230–235, 2013.
- [3] J. Bromley et al., “Signature verification using a Siamese time delay neural network,” in *Proc. NeurIPS*, 1993.
- [4] S. Dey et al., “SigNet: Convolutional Siamese network for writer independent offline signature verification,” *arXiv:1707.02131*, 2017.
- [5] H.-H. Kao and C.-Y. Wen, “An offline signature verification and forgery detection method based on a single known sample and an explainable deep learning approach,” *Appl. Sci.*, vol. 10, no. 11, p. 3716, 2020.
- [6] H. Li et al., “TransOSV: Offline signature verification with transformers,” *Pattern Recognit.*, vol. 145, p. 109882, 2024.
- [7] S. Tehsin et al., “Enhancing signature verification using triplet Siamese similarity networks in digital documents,” *Mathematics*, vol. 12, no. 17, p. 2757, 2024.
- [8] P. Brimoh and C. C. Olisah, “Consensus-threshold criterion for offline signature verification using CNN learned representations,” *arXiv:2401.03085*, 2024.
- [9] N. Woodruff et al., “Fully-automatic pipeline for document signature analysis to detect money laundering activities,” *arXiv:2107.14091*, 2021.
- [10] S. Abramova and R. Böhme, “Detecting copy-move forgeries in scanned text documents,” in *Proc. Electronic Imaging*, 2016.
- [11] Y. Li et al., “Copy-move forgery detection in digital image forensics: A survey,” *Multimedia Tools Appl.*, 2024.
- [12] Y. Jakhar and M. D. Borah, “Effective near-duplicate image detection using perceptual hashing and deep learning,” *Inf. Process. Manage.*, p. 104086, 2025.
- [13] E. Pizzi et al., “A self-supervised descriptor for image copy detection,” in *Proc. CVPR*, 2022.
- [14] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Learning features for offline handwritten signature verification using deep convolutional neural networks,” *Pattern Recognit.*, vol. 70, pp. 163–176, 2017.
- [15] E. N. Zois, D. Tsourounis, and D. Kalivas, “Similarity distance learning on SPD manifold for writer independent offline signature verification,” *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1342–1356, 2024.
- [16] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Meta-learning for fast classifier adaptation to new users of signature verification systems,” *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1735–1745, 2020.
- [17] H. Farid, “Image forgery detection,” *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, 2009.
- [18] F. Z. Mehrjardi, A. M. Latif, M. S. Zarchi, and R. Sheikhpour, “A survey on deep learning-based image forgery detection,” *Pattern Recognit.*, vol. 144, art. 109778, 2023.

- [19] J. Luo et al., "A survey of perceptual hashing for multimedia," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 21, no. 7, 2025.
- [20] D. Engin et al., "Offline signature verification on real-world documents," in *Proc. CVPRW*, 2020.
- [21] D. Tsourounis et al., "From text to signatures: Knowledge transfer for efficient deep feature learning in offline signature verification," *Expert Syst. Appl.*, vol. 189, art. 116136, 2022.
- [22] B. Chamakh and O. Bounouh, "A unified ResNet18-based approach for offline signature classification and verification across multilingual datasets," *Procedia Comput. Sci.*, vol. 270, pp. 4024–4033, 2025.
- [23] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. ECCV*, 2014, pp. 584–599.
- [24] S. Bai et al., "Qwen2.5-VL technical report," *arXiv:2502.13923*, 2025.
- [25] Ultralytics, "YOLO11 documentation," 2024. [Online]. Available: <https://docs.ultralytics.com>
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.
- [27] N. Krawetz, "Kind of like that," *The Hacker Factor Blog*, Jan. 21, 2013. [Online]. Available: <https://www.hackerfactor.com/blog/index.php?archives/529-Kind-of-Like-That.html>
- [28] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.
- [29] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988.
- [30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [31] J. V. Carcello and C. Li, "Costs and benefits of requiring an engagement partner signature: Recent experience in the United Kingdom," *The Accounting Review*, vol. 88, no. 5, pp. 1511–1546, 2013.
- [32] A. D. Blay, M. Notbohm, C. Schelleman, and A. Valencia, "Audit quality effects of an individual audit engagement partner signature mandate," *Int. J. Auditing*, vol. 18, no. 3, pp. 172–192, 2014.
- [33] W. Chi, H. Huang, Y. Liao, and H. Xie, "Mandatory audit partner rotation, audit quality, and market perception: Evidence from Taiwan," *Contemp. Account. Res.*, vol. 26, no. 2, pp. 359–391, 2009.
- [34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016, pp. 779–788.
- [35] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, 2024.
- [36] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.
- [37] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *Ann. Statist.*, vol. 13, no. 1, pp. 70–84, 1985.
- [38] D. Burgstahler and I. Dichev, "Earnings management to avoid earnings decreases and losses," *J. Account. Econ.*, vol. 24, no. 1, pp. 99–126, 1997.
- [39] J. McCrary, "Manipulation of the running variable in the regression discontinuity design: A density test," *J. Econometrics*, vol. 142, no. 2, pp. 698–714, 2008.
- [40] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.
- [41] E. Y. Chou, "Paperless and soulless: E-signatures diminish the signer's presence and decrease acceptance," *Social Psychological and Personality Science*, vol. 6, no. 3, pp. 343–351, 2015.

- [42] E. Y. Chou, “What’s in a name? The toll e-signatures take on individual honesty,” *Journal of Experimental Social Psychology*, vol. 61, pp. 84–95, 2015.
- [43] K. Tzelios and L. A. Williams, “The psychological impact of digital signatures: A multistudy replication,” *Technology, Mind, and Behavior*, vol. 1, no. 2, 2020.
- [44] S. Pal, M. Blumenstein, and U. Pal, “Non-English and non-Latin signature verification systems: A survey,” in *Proc. 1st Int. Workshop on Automated Forensic Handwriting Analysis (AFHA)*, 2011, pp. 1–5.
- [45] X. Chen, “Extraction and analysis of the width, gray scale and radian in Chinese signature handwriting,” *Forensic Science International*, vol. 255, pp. 123–132, 2015.

Declarations

Conflict of interest. The authors declare no conflict of interest with Firm A, Firm B, Firm C, or Firm D, or with any other entity referenced in this work.

Data availability. All audit reports analyzed in this study were obtained from the Market Observation Post System (MOPS) operated by the Taiwan Stock Exchange Corporation, a publicly accessible regulatory disclosure platform. The CPA registry used to map signatures to certifying CPAs is publicly available. The reproducibility scripts and trained model weights are provided in the supplementary materials; signature-image release is subject to the firm-anonymization constraints of Section III-A (a de-identified subset and the per-table provenance mapping are included, with the full image set available to reviewers under the platform’s public-data terms).