Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Read-Alike Test: Judging Whether Synthetic Degradation Can Stand In for Defocused Smartphone Document Captures by the Words a Reader Fails to Read

Hajime Hotta

hotta@hajime.institute

Hajime Institute

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11757258/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Read-Alike Test: Judging Whether Synthetic Degradation Can Stand In for Defocused Smartphone Document Captures by the Words a Reader Fails to Read

Hajime Hotta[1*]

[1*]Hajime Institute, Singapore.

Corresponding author(s). E-mail(s): hotta@hajime.institute;

Abstract

Purpose Reading robustness on phone-captured documents is measured with synthetic degradation validated by real-like appearance or error rate. Yet equal error rates can leave different words illegible, so synthetic word-level ground truth can contradict real captures. Real recaptures lose different words, and no yardstick says what overlap suffices. **Methods** We propose the read-alike test, which judges before building evaluation data whether a synthetic degradation (generator plus setting selection) can replace real degradation by whether a reader (a model reading text from document images) misses the same words. Per reader, synthetic-vs-real word-level agreement is compared with that between same-condition recaptures. A degradation passes if the confidence interval of its mean shortfall stays within the same-document recapture variation, fails if the interval lies entirely beyond it, and is undecided otherwise. Verdicts rest only on readers that tell real readings from random word dropping at real per-band rates. These were OCR engines. **Results** On public defocused captures of printed English documents, real captures put in place of a synthetic degradation almost never failed and often passed, and they did not fail on ten other documents of the same dataset. Randomizing more of the real readings moved verdicts toward failing. **Conclusion** The test screened out most of the synthetic degradations we tested. Choosing tilted-page defocus's setting by error rate alone gave word-level agreement below such random dropping. As ground truth for illegible words, each generator except band-selected tilted-page defocus, with or without the desk lamp, matched real misses no better than such random dropping.

Keywords: synthetic degradation, document image analysis, OCR robustness, defocus blur, error consistency, benchmark validation

1 Introduction

OCR engines [1], document parsers, and vision-language models [2] read document images,

including degraded smartphone captures. Recent benchmarks such as WildDoc [3] and Real5-OmniDocBench [4] center on camera-capture robustness. Evaluating or training robustness requires degraded images with ground truth. Because annotating real ones is laborious, researchers degrade clean documents synthetically.

Prior work has judged whether synthetic images can replace real ones by appearance [5] or error rate [6]. Some also compare error statistics [7]. Others compare character-image distributions [8]. Yet degradations sharing an error rate can still make different words illegible. In autonomous-driving simulation, too, close aggregate scores can hide errors in different places [9]. For example, uniform blur is equally strong across the page, whereas camera defocus grows away from the in-focus region.

In word-level evaluations, answers and ground truth depend on which words synthetic degradation makes illegible. Field extraction [10] and question answering [11] hinge on reading the answer words. OCR errors also change downstream language processing [12]. When readers are scored on reporting illegible text [13], the illegible words are the ground truth. When synthetic degradation makes different words illegible, the score of a reader that correctly reports real illegible words can differ from its score on real images, even at the same error rate. Yet rescanning a page shifts where OCR errs [14], and even same-condition real recaptures leave partly different words illegible. Medical measurement judges whether methods are interchangeable by within-method agreement [15]. So does a test of LLMs replacing human annotators [16]. But document degradation lacks a yardstick for how far synthetic illegible words must overlap real ones to count as a replacement. Without one, a degradation that looks right can quietly put the wrong words into an evaluation's ground truth.

We therefore make three contributions. The first is the read-alike test (Section 4, Fig. 1), which tells, before any evaluation data are built, whether a synthetic degradation (generator plus setting selection) can replace real degradation, by whether a reader fails on the same words. Its margin is same-document recapture variation, and it uses only readers that tell real readings from the null baseline, which drops words at random at the real rate in each of five bands along the word sequence. The second checks the test on known-answer inputs (Sections 5.2 and 5.5). On defocused captures in the public dataset, real captures put in place of a synthetic degradation almost never failed and often passed. The third shows that, on the same defocus, the test screened out most of the synthetic degradations we tested (Sections 5.6–5.9). What it screened out would have given poor ground truth. Tilted-page defocus set by error rate alone agreed with real captures word by word less

than the null baseline did, and as ground truth for illegible words every generator except band-selected tilted-page defocus, with or without the desk lamp, matched real misses no better than the null baseline (Sections 5.6 and 5.10). We release the code and verdicts.

2 Related Work

Degradation models and their validity

Parametric models describe document image defects [17], including whole-page and local ones [18]. Their synthetic images have replaced real ones in recognition experiments [19]. Baird [20] surveys this work. Li et al. [7] compared OCR error statistics on synthetic and real images. Nagy [21] compared the same classifier's confusion matrices. Kanungo et al. [8] tested by permutation whether synthetic and real character images share a distribution, and confirmed it rarely errs on real-vs-real comparisons. Kanungo and Zheng [22] estimated parameters by matching local pixel-pattern distributions. Barney Smith and Qiu [23] statistically compared populations of character images. Barney Smith and Andersen [24] used this similarity to choose training data. In Hale and Barney Smith [25], people preferred thick-stroked degraded characters to faithful ones. The read-alike test checks synthesis by whether the same words become illegible rather than by distributions or statistics.

Degradation synthesis and synthetic data

DocCreator [26] and Augraphy [5] combine effects such as ink bleeding and capture shadows into degraded images with ground truth. ShabbyPages [27] uses Augraphy-degraded pages to train and evaluate denoising and binarization. The albumentations library [28] also provides blur and noise. Synthetic training data sometimes underperform real data [29]. For generated images, realism is measured by feature-distribution distance [30]. HYPE measures how often people mistake generated images for real [31]. Real and synthetic blur have been compared on real blurred-sharp image pairs [32]. The read-alike test checks generators and settings against the words a reader misses in real images.

Robustness benchmarks using degradation

In image recognition, ImageNet-C measures robustness with synthetic noise and blur [33]. Yet robustness to synthetic shifts transfers little to natural ones [34]. Robustness methods helped against some kinds of real shift and not others [35]. Others test with score correlations or rankings whether synthetic data can replace real data. Semantic segmentation scores under synthetic and real degradations correlate on average but diverge for some types [36]. Structured synthetic rain and fog preserve the real-data ranking of monocular SLAM methods, but naive synthetic lighting does not [37]. Robot navigation research tested whether

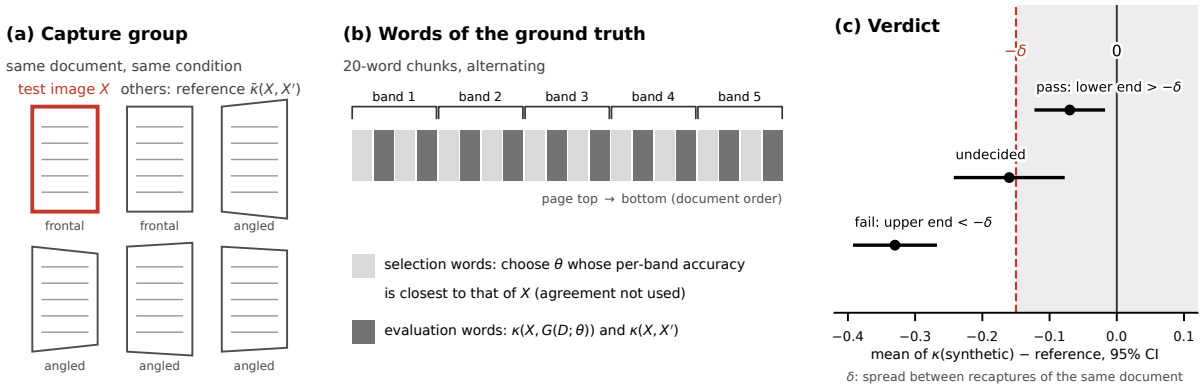


Fig. 1 Procedure of the read-alike test. (a) One capture in a capture group (one document, one condition) is the test image X, and the reference is the mean $\bar{\kappa}(X, X')$ of its κ (agreement on whether each word was read) with the remaining captures X'. (b) 20-word chunks of the ground truth alternate from the start between selection words, which choose the setting θ, and evaluation words, which measure agreement. (c) The mean of the synthetic image's κ minus the reference is judged with its 95% interval against $-\delta$ (same-document recapture variation).

simulation predicts real-robot performance [38]. These aggregate comparisons do not examine item-wise correctness. For documents, RoDLA measures layout-analysis robustness under synthetic degradation [39]. Another benchmark compares OCR engines and vision-language models under synthetic degradation [40]. PureDocBench [2] compares only overall parser scores across clean, synthetically degraded, and recaptured versions of a page. Real5-OmniDocBench [4], WildDoc [3], and DocPTBench [41] use recaptured documents. OHR-Bench [11] synthesizes OCR errors to study their effect on retrieval and answering. All these measure reader robustness, whereas the read-alike test judges whether synthetic degradation itself can replace real degradation.

Quality assessment and real data for captured documents

Document image quality assessment models predict OCR performance or human ratings [42]. Some are trained on synthetic text-line images [43]. SmartDoc-QA [44] and Kumar et al. [45] provide documents captured under varying blur and lighting with OCR performance. The Brno Mobile OCR Dataset [46] provides phone photographs of documents with ground-truth text. The Helsinki Deblur Challenge 2021 dataset [47] pairs in-focus and progressively defocused shots of the same targets. None judges whether a synthetic degradation can stand in for real ones.

Instance-level comparison of real and synthetic images

Lindström et al. [48] checked whether autonomous-driving detectors err alike on real and rendered images, using detections on one as

ground truth for the other. Chen and Negrut [49] matched objects across real and synthetic images and chose a synthesis method with a small detection gap. Neither uses real-vs-real agreement as the yardstick or gives pass/fail verdicts.

Item-wise error agreement

Deep networks matching primates in object-confusion patterns still failed to predict primates' per-image performance [50]. Geirhos et al. [51] measured trial-wise agreement in two decision makers' correctness with Cohen's κ, named it error consistency, and judged model-human against human-human agreement. For error consistency, Klein et al. [52] gave confidence intervals and showed that, at equal accuracy, copying another classifier's answers with probability q yields $\kappa = q$. In medical measurement, two methods are judged interchangeable by comparing their agreement with that of one method's repeated measurements [15]. Haber et al. [53] extended this comparison to binary observations. Imaging tests use the same idea [54]. Another study tested whether electronic images can replace film by whether radiologists' decisions change [55]. Calderon et al. [56] left each annotator out and asked whether an LLM or that annotator agreed better with the rest, with a one-sided test and a margin. He et al. [16] tested whether agreement is equivalent after an LLM replaces one human annotator. They set the margin from the agreement gap between two human groups and, as a control, found no equivalence when random labels replaced one group. The read-alike test borrows this yardstick and adds two parts specific to documents, a null baseline that drops words at the real rate in each band of the word sequence as the placebo arm, and a judged setting selection that keeps the words used for selection apart from those used for measurement. The three verdicts (pass, undecided, fail) follow the usual reading of non-inferiority trials [57]. Bioequivalence studies sometimes set the margin from the reference product's variability across repeated doses [58]. In three-arm non-inferiority trials, the reference beating placebo confirms assay sensitivity [59].

3 Problem Setting

Before synthetic degradation, meaning a generator and its setting selection, is used to build evaluation data from smartphone document captures, we want to know whether it can replace real degradation. The selection is judged too because it changes which words the same generator makes illegible. A source document D has the ground-truth word sequence $w_1, \ldots, w_n$. Let X be a real image of D captured under a degradation condition. A generator G creates a synthetic image $G(D;\theta)$ from D with parameters θ. A reader R, an OCR engine or vision-language model, outputs a string from an image.

If an evaluation counts only the error rate, approximating the real one suffices. Word-level evaluations instead need the reader to fail on the same words as on real images, yet even same-condition recaptures lose different words as camera angle varies. We therefore compare word-level agreement between synthetic and real images with that between real recaptures. If, averaged over the judged real images, synthetic agreement falls short by no more than recapture variation, the synthetic degradation stands in for that condition. With the readers to be evaluated unknown in advance, a single synthetic degradation must reach this agreement for every reader that can tell real misses from words dropped at random at the same per-position rates (qualified readers, Section 4).

The simplest approach chooses θ to match R's error rate (the share of words not read) between synthetic and real images and accepts the generator when the rates are close [6]. Making the upper half of the page illegible and making every other word illegible both give a word accuracy (the share of words read) of one half. Comparisons of distributions or error statistics (Section 2) pool real and synthetic samples without pairing words and so cannot show whether the same words become illegible.

The first form of setting selection picks, per real image, the best-matching setting tried. A synthetic degradation failing under it does not stand in for that condition at the chosen settings. The second chooses the setting on other documents (Section 5.9).

4 Read-Alike Test

Each capture in a capture group, the real images of one document under one condition, serves in turn as the test image that a synthetic image is made to match (Fig. 1).

Per-paragraph reading

So that differences in how readers segment paragraphs and lines do not enter the verdict, the same paragraph regions, taken from the source document, are cut from real and synthetic images and read one at a time. Real images are first rectified from the four page corners, detected as the corners of the page region against the red backdrop (the minimum-area rectangle when no quadrilateral is found), onto an upright A4-proportioned rectangle as wide as the page's mean top and bottom edges. Paragraph boxes are mapped by their position relative to the page size. Each paragraph box is widened by 1.2% of the page on every side.

Read vector

After normalization (case, ligatures, quotation marks and punctuation at word ends), Python's difflib aligns reader R's output for image I with the ground-truth word sequence in order, keeping all matching blocks. If the ith ground-truth word

falls within a matching block, $r_i(I) = 1$, and otherwise $r_i(I) = 0$. Because readers miss some words even without degradation, the target words M are only those with $r_i = 1$ on both the rendered source PDF and a sharp real capture. Target words a reader fails to read ($r_i = 0$) are its illegible words.

Selection words, evaluation words, and bands

Measuring agreement on the words used to choose the setting would inflate it [60]. We therefore divide the ground-truth word sequence into chunks of 20 words. Chunk by chunk from the first, the words of M alternate between selection and evaluation words. With several hundred words per document in the main set, both kinds span the page and so reflect where words become illegible. The sequence is also divided by position into five equal bands.

Agreement

Agreement between two images is Cohen's κ [61] of their read vectors over the evaluation words. Even with equal word accuracies, κ is near 0 if illegible words overlap only at chance level. The overlap, uncorrected for chance, is the number of words illegible in both images divided by the mean number illegible in each. For the same reader, it equals the F1 of the real image's illegible words with the synthetic image's as ground truth. Counted from the same readings as κ, it measures the quality of that ground truth.

Real-vs-real agreement

When almost no or almost all words are illegible, κ swings widely on whether a few words are read. Only real images with 3% to 97% word accuracy on the evaluation words (measurable images) are therefore used for κ. A test image's reference is its mean κ with the other measurable captures in its group. Test images whose group has no other measurable capture are dropped.

Margin

For each condition and reader, δ is how much the reference varies between recaptures of the same document, its standard deviation around document means, pooled over all documents. This is modeled on scaled average bioequivalence for highly variable drugs [58], with documents as subjects, recaptures as repeated administrations of the reference, and a capture's reference as its response, so δ is the within-subject standard deviation of the reference.

Verdict

The synthetic image's κ minus the reference gives one difference per test image. A document-level bootstrap (2,000 resamples, percentile interval) gives the 95% confidence interval of the mean difference. Standing in (Section 3) means that, averaged over test images, synthetic agreement is

at most δ below real-vs-real agreement. A non-inferiority test [62] decides this. It is the lower side of an equivalence test [63], because higher agreement is no obstacle to standing in. The synthetic degradation passes if the interval's lower end is above $-\delta$, fails if its upper end is below $-\delta$, and is undecided otherwise.

Qualified readers

Under each condition, a reader is a candidate if its measurable test images span at least 5 documents. For each candidate, the null baseline is a read vector reading each evaluation word with probability equal to its band's word accuracy on the test image's selection words. A band without selection words uses the image's word accuracy over all its selection words. Insofar as word accuracy differs between bands, even the null baseline gets κ above 0, and error rates alone cannot reject it. Only the candidates with which the null baseline fails are qualified readers. A reader that cannot tell real misses from this reading cannot judge word-level agreement. The null baseline acts as the placebo arm of a three-arm non-inferiority trial [59]. Its κ is the mean over 200 draws. The mean keeps a reader's qualification from depending on a single draw. Both δ and the qualified readers come from real captures and the null baseline alone.

Setting selection

On the selection words, the squared differences in per-band word accuracy between the synthetic and test images are summed, skipping bands without them. These sums are added over every qualified reader for which the test image is measurable. Each test image takes the θ with the smallest total from the settings in Table 1.

Per-condition verdict

Under each condition, a synthetic degradation passes if it passes with every qualified reader (an intersection–union test [64]), fails if it fails with any one, and is undecided otherwise. How often this combined verdict fails real captures is counted with real captures in place of the generator (Section 5.2). Verdicts with unqualified readers count toward neither pass nor fail. Conditions with no qualified reader receive no verdict.

Known-answer inputs

We check the test on two known-answer inputs. One replaces the generator with a measurable real capture taken out of its group. The difference is its κ with the test image minus the reference computed without it, with the condition's δ and qualified readers unchanged. The taken-out image is a real capture, so the known answer is that it does not fail. The other replaces a fraction p of a real capture's read vector with the null baseline. A larger p is farther from the real image, so we

check whether verdicts move toward failing as p rises (Section 5.2).

5 Experiments

Sections 5.2–5.5 check that the test rarely fails real captures, and Sections 5.6–5.10 use it to screen synthetic defocus.

5.1 Setup

Data

Real degraded images came from SmartDoc-QA [44] (Table 1), public captures of printed documents by Nokia and Samsung smartphones, with source PDFs and ground-truth text. We used defocused captures of 10 modern English documents (9 for Nokia 2). Two focus steps per model and three lightings, daylight with a ceiling lamp (L2), a desk lamp at night (L3), and a desk lamp with an object's shadow (L4), give 12 conditions such as "Nokia 1 L2" (model, focus step, lighting). Each document was captured from 5 angles per condition, labeled ax_by in the dataset with a0_b0 frontal, plus 1 defocused frontal capture under L2. Target words come from the sharp real capture, an in-focus frontal capture in daylight.

Readers

The four readers differ in robustness to degradation. Tesseract 5.5.1 [1] is an open-source OCR engine, and docTR 1.1.0 [65] a deep-learning one. Apple Vision, macOS 26.4's built-in text recognizer, ran at the "accurate" level. At temperature 0 the vision-language model Mistral-Small-3.2-24B-Instruct-2506 [66] transcribed paragraph images, prompted to transcribe all the text exactly as written in reading order and to output only the transcription.

Generators

Tilted-page defocus approximates defocus on a page tilted about its horizontal axis. With a slight tilt the blur-disk radius varies roughly linearly with vertical position, so we interpolate linearly between top- and bottom-edge radii and use the absolute value as the radius. This is the tilted-plane case of depth-varying blur generators [67]. Under L3 we also compared tilted-page defocus with desk-lamp lighting added. Illumination falls off linearly from the lamp at the top right (50% from top to bottom and 25% from right to left), and ink brightens to a fixed fraction of the paper brightness. Baselines are uniform disk blur (one radius over the whole page) and three existing generators. Every generator degrades the source PDF rendered at 208 dpi for Nokia and 285 dpi for Samsung, and its output is saved as JPEG at quality 90.

Table 1 Data, readers, generators, and candidate settings.

Item	Content
SmartDoc-QA	Defocus blur for 10 documents (9 for Nokia 2). 2 devices × 2 focus steps × 3 illuminations = 12 conditions. 5 images per condition and document (6 for L2, 5 for one Nokia 2 L2 document), 623 in total
Readers	Tesseract 5.5.1, docTR 1.1.0, Apple Vision (macOS 26.4), Mistral-Small-3.2-24B
Tilted-page defocus	Radii at the top and bottom edges. Nokia 120 pairs within −8 to 8 pixels, Samsung 152 pairs within −11 to 14 pixels (negative radii mean the focal plane crosses the page)
Tilted-page defocus + desk lamp	L3 only. Top and bottom radii (Nokia 64 pairs, Samsung 49 pairs) × 5 ink brightness levels
Uniform disk blur	Radius in 0.25-pixel steps, 1.5–10 pixels for Nokia and 1.5–14 pixels for Samsung (35 and 51 values)
Existing generators	Defocus (radius and edge blur, 30 and 54 settings) and GaussianBlur (σ, 16 and 26 settings) from albumentations 2.0.8, DepthSimulatedBlur (5 ellipse sizes × 3 random seeds) from Augraphy 8.2.6

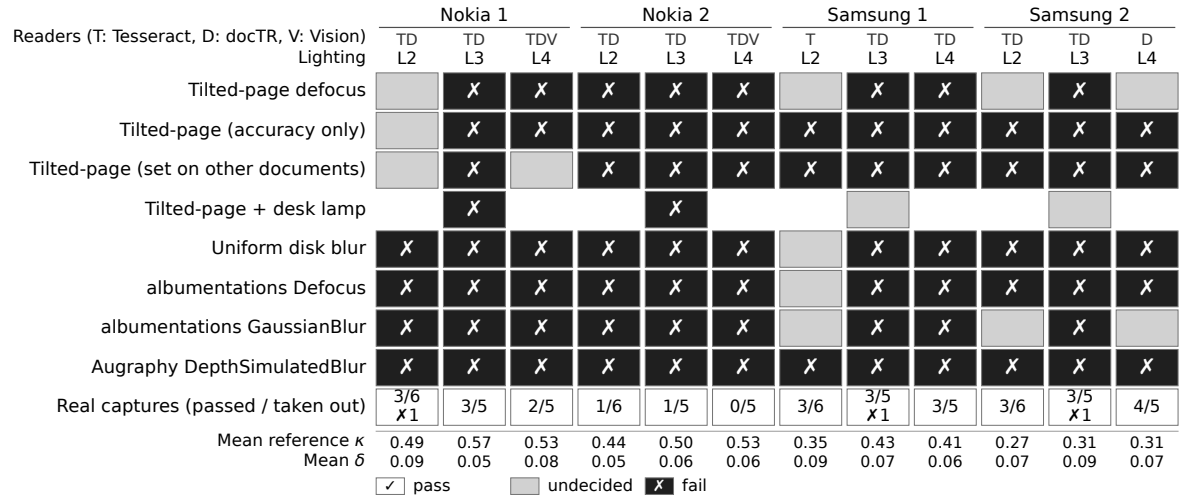


Fig. 2 Verdict per condition, combined over the qualified readers (white for pass, gray for undecided, black for fail, blank for combinations not compared). Rows without a selection rule are band-selected. Above each column are the readers qualified under that condition. Below it are the real captures passed/taken out when used in place of the generator (the number failed follows ✗) and the mean reference and mean δ.

5.2 Does the Test Work on Known-Answer Inputs?

We first check the test on known-answer inputs. Real captures put in place of a synthetic degradation often passed and almost never failed. Mixing more null-baseline readings into real ones moved verdicts from pass to fail.

With each angle taken out in turn for every document, the 12 conditions with a qualified reader give 64 combinations of condition and capture. With the qualified readers combined, the test passed 29 of these real captures and failed only 3, leaving 32 undecided (Table 2, Fig. 2). Of the 125 cases judged per qualified reader (Tesseract 59, docTR 56, Vision 10), 4 failed.

Following Klein et al. [52], the second input replaces a fraction p of the taken-out capture's read vector with the null baseline and is judged, like the null baseline, by the mean κ over 200 draws per p. At $p = 0$ the input is the real capture

Table 2 Number of verdicts per generator and setting selection, and overlap of illegible words. The left columns count verdicts with the qualified readers combined, per condition (per combination of condition and capture taken out in the bottom row, and over the 4 L3 conditions for the generator with the desk lamp added). The center columns count verdicts per reader, only under the conditions where that reader qualified. The rightmost column is the overlap averaged over combinations of qualified reader and condition (8 combinations for the generator with the desk lamp added). The bottom row gives verdicts for real captures used in place of the generator, with the overlap between real recaptures. Rows that name no selection use band-based selection.

	Combined verdicts			Per reader (pass/undecided/fail)			
Generator	Pass	Undecided	Fail	Tesseract	docTR	Vision	Overlap
Tilted-page defocus	0	4	8	0/4/7	0/5/6	0/1/1	0.45
Tilted-page defocus (accuracy-based)	0	1	11	0/2/9	0/1/10	0/0/2	0.38
Tilted-page defocus (set on other documents)	0	2	10	0/2/9	0/2/9	0/2/0	0.39
Tilted-page defocus + desk lamp	0	2	2	0/2/2	0/2/2	0/0/0	0.57
Uniform disk blur	0	1	11	0/2/9	0/0/11	0/0/2	0.36
albumentations Defocus	0	1	11	0/2/9	0/0/11	0/0/2	0.35
albumentations GaussianBlur	0	3	9	0/2/9	0/4/7	0/0/2	0.40
Augraphy DepthSimulatedBlur	0	0	12	0/0/11	0/0/11	0/0/2	0.07
Null baseline (band-matched random dropout)	0	0	12	0/0/11	0/0/11	0/0/2	0.41
Real capture (one at a time)	29	32	3	39/19/1	33/20/3	6/4/0	0.60

itself, and at $p = 1$ every word is replaced by the null baseline. Replacing 3 in 10 words lowered the overlap to 0.91 times that between real recaptures, and 27 of 64 failed (9 passed). As p rose from 0 to 1 in steps of 0.1, passes fell from 29 to 0 and failures rose from 3 to 58, both monotonically.

5.3 Do Synthetic Degradations Still Fail Under Other Design Values?

We check whether synthetic degradations still fail when δ and other design values change (Table 3). Most still did. At every value, 72–91% of the judged combinations of band-selected generator and condition failed.

The margin δ was multiplied by 0.9 to 1.8, redefined as the standard deviation between documents or the standard error of the document-mean reference (that deviation over $\sqrt{n}$ for n documents), or replaced by fixed margins of 0.05 to 0.20. Settings were not reselected. The synthetic images stayed as in the main text, only the margin changed, and readers that no longer failed the null baseline were dropped.

Table 3 Verdicts with one design value changed at a time. Cond. gives the number of conditions with a qualified reader. Real gives the real captures taken out that failed, with the number taken out in parentheses. Generator gives the verdicts (pass/undecided/fail) for combinations of band-selected generator and condition, with the number judged in parentheses.

Variant	Cond.	Real	Generator
Main (δ = SD between recaptures)	12	3 (64)	0/11/53 (64)
$\delta \times 0.9$	12	4 (64)	0/9/55 (64)
$\delta \times 1.2$	12	2 (64)	1/15/48 (64)
$\delta \times 1.5$	10	1 (53)	2/10/41 (53)
$\delta \times 1.8$	8	1 (42)	1/3/39 (43)
δ = SD between documents	10	2 (52)	0/7/47 (54)
δ = SE of the document-mean reference	12	6 (64)	0/6/58 (64)
Fixed margin 0.05	12	4 (64)	0/8/56 (64)
Fixed margin 0.08	12	2 (64)	1/11/52 (64)
Fixed margin 0.10	12	1 (64)	1/17/46 (64)
Fixed margin 0.12	9	1 (47)	1/8/39 (48)
Fixed margin 0.15	8	1 (42)	1/4/38 (43)
Fixed margin 0.20	7	1 (37)	3/6/28 (37)
No selection/evaluation split	11	6 (58)	0/9/50 (59)
10-word chunks	12	5 (64)	0/9/55 (64)
40-word chunks	11	3 (58)	0/8/51 (59)
3 bands	12	3 (64)	0/10/54 (64)
10 bands	10	2 (53)	0/7/46 (53)
Measurable range 1–99%	11	1 (58)	1/13/45 (59)
Measurable range 5–95%	11	1 (58)	0/9/50 (59)
Measurable range 10–90%	9	0 (46)	0/5/44 (49)

Other design values were varied one at a time, rerunning setting selection through verdicts. The 20-word chunks became 10 or 40 words or were dropped so the same words served both roles. The 5 bands became 3 or 10, and the measurable word-accuracy range went from 3–97% to 1–99%, 5–95%, and 10–90%. Of the judged combinations, 76–90% failed at every value (Table 3). Without the split between selection and evaluation words, real captures failed more than twice as often (6 of 58 against 3 of 64).

5.4 Which Readers Can Support a Verdict?

We ask which readers can support a verdict. Every verdict rests on readers shown to tell real misses from random ones, mainly the OCR engines Tesseract and docTR. Tesseract qualified under 11 conditions, docTR under 11 and Vision under 2, together covering 12, and Mistral under none (Fig. 2).

5.5 Do Real Captures Fail on Another Document Set?

With design values fixed, we check whether real captures fail on another document set. None did. SmartDoc-QA documents 11–20 are old business letters (40–390 words each). Their source images and paragraph regions were taken from the rectified in-focus frontal captures. Tesseract, docTR, and Vision qualified in 32 condition-reader combinations, covering 12 conditions (Vision under 8). The 64 real-capture cases gave 15 pass, 49 undecided, and 0 fail verdicts, and none of the 168 cases per qualified reader failed.

5.6 Does Tilted-Page Defocus Selected by Word Accuracy Alone Pass?

We ask whether tilted-page defocus passes with the setting closest to real images in word accuracy (or error rate) alone [6]. Tilted-page defocus selected by word accuracy alone passed under no condition with a qualified reader, and its per-word agreement with real images was on average no higher than the null baseline's.

This setting minimized the absolute word-accuracy difference from the test image, summed over all qualified readers, and on evaluation words its mean absolute difference was 0.115, below 0.133 under band-based selection. It still failed under 11 of the 12 conditions (Table 2). Its κ with real images averaged 0.11, below the null baseline (0.17) and real recaptures (0.44). Band-based selection raised the same generator's mean κ to 0.24.

In Fig. 3, illegible words concentrate at the bottom in the real capture (with some at the top) and under band-based selection. Accuracy-only selection comes close in word accuracy, yet



Fig. 3 Where words become illegible on one page (the a0_b0 capture of document 1 under Nokia 1 L2). Top: word positions, with Tesseract's misses in red. Bottom: top and bottom three lines, enlarged. Settings were chosen jointly for both qualified readers, Tesseract and docTR. Numbers give word accuracy on evaluation words and κ with the real capture (Tesseract).

its illegible words concentrate at the top. Uniform disk blur scatters them across the page. The three had κ of 0.60, 0.00, and -0.03 with the real capture (Tesseract).

5.7 Does Band-Selected Tilted-Page Defocus Pass?

We check whether tilted-page defocus passes with band-selected settings. It received 0 pass, 4 undecided, and 8 fail verdicts over the 12 conditions (Table 2, Fig. 2). Band-selected tilted-page defocus with the desk lamp (Section 5.1) gave 0 pass, 2 undecided, and 2 fail under the 4 L3 conditions.

5.8 Do Band-Selected Library Blurs Pass?

We ask whether three blur generators from two libraries pass with band-selected settings. None of the three passed under any condition with a qualified reader. DepthSimulatedBlur failed under all 12 conditions (Table 2, Fig. 2).

5.9 Does Tilted-Page Defocus Selected on Other Documents Pass?

We check whether tilted-page defocus passes with a setting chosen without seeing the target document. This defocus passed under none of the 12 conditions with a qualified reader (Table 2, Fig. 2).

Leaving each target document out in turn, one band-based setting was chosen on all captures of the other documents. Verdicts were 0 pass, 2 undecided, and 10 fail.

Excluding the null baseline, the test screened out 74 of the 88 combinations of generator, setting selection, and condition with a qualified reader (14 undecided, 0 passed), while failing only 3 of the 64 real-capture cases and passing 29 (Table 2). The verdicts also ranked the synthetic degradations. Among the generators judged under all 12 conditions, band-selected tilted-page defocus was undecided most often (4) and had the highest overlap (0.45), whereas DepthSimulatedBlur failed under all 12 with the lowest (0.07).

5.10 What Would Screened-Out Degradations Have Given as Ground Truth?

We ask what the screened-out degradations would have put into evaluation data, measured by how far their illegible words overlap real ones. We take the overlap (F1 against the reader's misses on the synthetic image) relative to that between real recaptures (Fig. 4, rightmost column of Table 2). Over the 128 combinations of qualified reader, condition, and generator (5 band-selected generators and the desk-lamp one), this ratio averaged 0.87 for undecided per-reader verdicts (24) and 0.48 for failed ones (104), below the null baseline's 0.64. With the 24 qualified reader-condition pairs resampled 2,000 times, the 95% interval of the failed combinations' ratio minus the null baseline's was −0.24 to −0.09, excluding 0. In raw overlap (Table 2), every generator and selection rule except band-selected tilted-page defocus, with or without the desk lamp, stayed below the null baseline (0.07 to 0.40 against 0.41), while real recaptures reached 0.60.

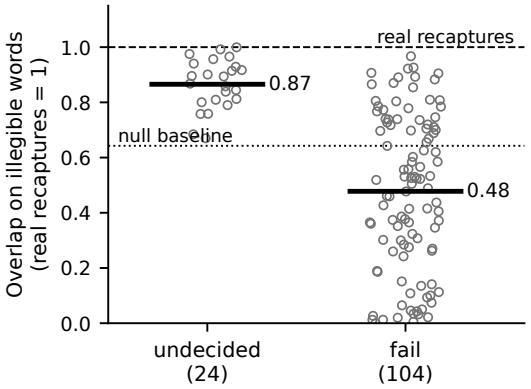


Fig. 4 Overlap of illegible words by per-reader verdict (real recaptures = 1 for each reader and condition). Points (128) are combinations of qualified reader, condition, and band-selected generator. Horizontal lines are means, and parentheses give the number of points.

6 Conclusion

For defocus in smartphone document captures, the read-alike test checks a synthetic degradation (generator plus setting selection) before it enters evaluation data, by whether a reader fails on the same words as on real captures, with real-vs-real agreement and same-document recapture variation as yardsticks. On known-answer inputs, verdicts moved toward failing as more real readings were replaced by null baseline readings, which drop words at random at the real per-band rates. Under defocus in the public dataset, real captures put in place of a synthetic degradation almost never failed and often passed, whereas the test screened out most of the synthetic degradations we tested. Among those screened out were tilted-page defocus set by error rate alone and the library blurs, whose ground truth for illegible words matched real misses no better than random dropping. Before synthesizing evaluation data for defocused captures of printed English documents, we recommend running the test with OCR readers to rule out candidate degradations that cannot stand in for real ones. We release the code and verdicts.

Statements and Declarations

- Funding: No funding was received for this work.
- Competing interests: The author has no competing interests to declare.
- Ethics approval and consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Data availability: Real degraded images come from the public SmartDoc-QA dataset. The verdicts and the statistics behind them are released with the code. The code regenerates the SmartDoc-QA transcriptions from the dataset, whose license does not allow re-distribution.
- Materials availability: Not applicable.
- Code availability: https://github.com/Hajime-Institute/read-alike-test
- Author contribution: The author confirms sole responsibility for the study conception and design, the experiments, the analysis and the manuscript.
- Use of large language models: Mistral-Small-3.2-24B is a reader under test (Section 5.1). Large language models also helped finalize and check the manuscript. The author reviewed it in full and takes responsibility for it.

References

[1] Smith, R.: An overview of the Tesseract OCR engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633 (2007). https://doi.org/10.1109/ICDAR.2007.4376991

[2] Li, Z., Ma, Z., Chen, J., Zhang, J., Su, Z., Zhang, Y., Yu, Z., Liu, R., Lv, X., Li, B., Gao, J., Zhang, Z., Yuan, C., Li, B., Hu, W.: How Far Is Document Parsing from Solved? PureDocBench: A Source-Traceable Benchmark across Clean, Degraded, and Real-World Settings (2026). https://doi.org/10.48550/arXiv.2605.07492

[3] Wang, A.-L., Tang, J., Liao, L., Feng, H., Liu, Q., Fei, X., Lu, J., Wang, H., Liu, H., Liu, Y., Bai, X., Huang, C.: WildDoc: How far are we from achieving comprehensive and robust document understanding in the wild? In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 23002–23012 (2025). https://doi.org/10.18653/v1/2025.emnlp-main.1172

[4] Cui, C., Zhou, C., Gao, Z., Wang, X., Gao, T., Tang, J., Liu, Y.: Real5-OmniDocBench: A full-scale physical reconstruction benchmark for robust document parsing in the wild. In: Computer Vision – ECCV 2026, pp. 208–222 (2026). https://doi.org/10.1007/978-3-032-37350-2_12

[5] Groleau, A., Chee, K.W., Larson, S., Maini, S., Boarman, J.: Augraphy: A data augmentation library for document images. In: Document Analysis and Recognition – ICDAR 2023, vol. 14189, pp. 384–401 (2023). https://doi.org/10.1007/978-3-031-41682-8_24

[6] Gupte, A., Romanov, A., Mantravadi, S., Banda, D., Liu, J., Khan, R., Meenal, L.R., Han, B., Srinivasan, S.: Lights, Camera, Action! A Framework to Improve NLP Accuracy over OCR Documents (2021). https://doi.org/10.48550/arXiv.2108.02899

[7] Li, Y., Lopresti, D., Nagy, G., Tomkins, A.: Validation of image defect models for optical character recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **18**(2), 99–107 (1996) https://doi.org/10.1109/34.481536

[8] Kanungo, T., Haralick, R.M., Baird, H.S., Stuezle, W., Madigan, D.: A statistical, nonparametric methodology for document degradation model validation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11), 1209–1223 (2000) https://doi.org/10.1109/34.888707

[9] Manivasagam, S., Bârsan, I.A., Wang, J., Yang, Z., Urtasun, R.: Towards zero domain gap: A comprehensive study of realistic LiDAR simulation for autonomy testing. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8238–8248. IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.00760

[10] He, Z., Zhang, C., Wu, Z., Chen, Z., Zhan, Y., Li, Y., Zhang, Z., Wang, X., Qiu, M.: Seeing is believing? mitigating OCR hallucinations in multimodal large language models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 38, pp. 82480–82498 (2025). https://doi.org/10.52202/085713-2487

[11] Zhang, J., Zhang, Q., Wang, B., Ouyang, L., Wen, Z., Li, Y., Chow, K.-H., He, C., Zhang, W.: OCR hinders RAG: Evaluating the cascading impact of OCR on retrieval-augmented generation. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 17443–17453 (2025). https://doi.org/10.1109/ICCV51701.2025.01620

[12] Lopresti, D.: Optical character recognition errors and their effects on natural language processing. International Journal on Document Analysis and Recognition **12**(3), 141–151 (2009) https://doi.org/10.1007/s10032-009-0094-8

[13] Guan, S., Lin, M., Xu, C., Zhao, J., Greene, D.: Teaching VLMs to admit uncertainty in OCR from lossy visual inputs. In: International Conference on Learning Representations (ICLR) (2026). https://mlanthology.org/iclr/2026/guan2026iclr-teaching/

[14] Lopresti, D., Zhou, J.: Using consensus sequence voting to correct OCR errors. Computer Vision and Image Understanding **67**(1), 39–47 (1997) https://doi.org/10.1006/cviu.1996.0502

[15] Barnhart, H.X., Kosinski, A.S., Haber, M.J.: Assessing individual agreement. Journal of Biopharmaceutical Statistics **17**(4), 697–719 (2007) https://doi.org/10.1080/10543400701329489

[16] He, J., Leng, Z., McKay, D., Spina, D., Trippas, J.R.: Can we hide machines in the crowd? quantifying equivalence in LLM-in-the-loop annotation tasks. In: Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, pp. 426–436. ACM, Xi'an, China (2025). https://doi.org/10.1145/3767695.3769508

[17] Baird, H.S.: Document image defect models. In: Structured Document Image Analysis, pp. 546–556. Springer, Berlin, Heidelberg (1992). https://doi.org/10.1007/978-3-642-77281-8_26

[18] Kanungo, T., Haralick, R.M., Phillips, I.: Nonlinear global and local document degradation models. International Journal of Imaging Systems and Technology **5**(3), 220–230 (1994) https://doi.org/10.1002/ima.1850050305

[19] Ho, T.K., Baird, H.S.: Large-scale simulation studies in image pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(10), 1067–1079 (1997) https://doi.org/10.1109/34.625107

[20] Baird, H.S.: The state of the art of document image degradation modelling. In: Digital Document Processing, pp. 261–279. Springer, London (2007). https://doi.org/10.1007/978-1-84628-726-8_12

[21] Nagy, G.: Validation of simulated OCR data sets. In: Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR '94), Las Vegas, NV, USA, pp. 127–135 (1994). https://www.osti.gov/biblio/68570

[22] Kanungo, T., Zheng, Q.: Estimating degradation model parameters using neighborhood pattern distributions: An optimization approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(4), 520–524 (2004) https://doi.org/10.1109/TPAMI.2004.1265867

[23] Barney Smith, E., Qiu, X.: Statistical image differences, degradation features, and character distance metrics. International Journal on Document Analysis and Recognition **6**(3), 146–153 (2003) https://doi.org/10.1007/s10032-003-0117-9

[24] Barney Smith, E.H., Andersen, T.: Text degradations and OCR training. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05), vol. 2, pp. 834–838 (2005). https://doi.org/10.1109/ICDAR.2005.226

[25] Hale, C., Barney Smith, E.: Human image preference and document degradation models. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 257–261 (2007). https://doi.org/10.1109/ICDAR.2007.4378715

[26] Journet, N., Visani, M., Mansencal, B., Kieu, V.-C., Billy, A.: DocCreator: A new software for creating synthetic ground-truthed document images. Journal of Imaging **3**(4), 62 (2017) https://doi.org/10.3390/jimaging3040062

[27] Groleau, A., Chee, K.W., Larson, S., Maini, S., Boarman, J.: ShabbyPages: A Reproducible Document Denoising and Binarization Dataset (2023). https://doi.org/10.48550/arXiv.2303.09339

[28] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2), 125 (2020) https://doi.org/10.3390/info11020125

[29] Etter, D., Rawls, S., Carpenter, C., Sell, G.: A synthetic recipe for OCR. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 864–869 (2019). https://doi.org/10.1109/ICDAR.2019.00143

[30] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 6626–6637 (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[31] Zhou, S., Gordon, M., Krishna, R., Narcomey, A., Fei-Fei, L., Bernstein, M.: HYPE: A benchmark for human eye perceptual evaluation of generative models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 3444–3456 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/65699726a3c601b9f31bf04019c8593c-Paper.pdf

[32] Rim, J., Kim, G., Kim, J., Lee, J., Lee, S., Cho, S.: Realistic blur synthesis for learning image deblurring. In: Computer Vision – ECCV 2022. Lecture Notes in Computer Science, pp. 487–503. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_29

[33] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019). https://doi.org/10.48550/arXiv.1903.12261

[34] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 18583–18599 (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

[35] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution

generalization. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8320–8329 (2021). https://doi.org/10.1109/ICCV48922.2021.00823

[36] Agnihotri, S., Schader, D., Sharei, N., Kaçar, M.E., Keuper, M.: Are Synthetic Corruptions A Reliable Proxy For Real-World Corruptions? CVPR 2025 Workshop on Synthetic Data for Computer Vision (2025). https://doi.org/10.48550/arXiv.2505.04835

[37] Thomas, A.S., Agnihotri, S., Keuper, M.: Failure or Drift? Evaluating Monocular SLAM under Synthetic and Real-World Corruptions (2026). https://doi.org/10.48550/arXiv.2608.30690

[38] Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., Batra, D.: Sim2Real predictivity: Does evaluation in simulation predict real-world performance? IEEE Robotics and Automation Letters **5**(4), 6670–6677 (2020) https://doi.org/10.1109/LRA.2020.3013848

[39] Chen, Y., Zhang, J., Peng, K., Zheng, J., Liu, R., Torr, P., Stiefelhagen, R.: RoDLA: Benchmarking the robustness of document layout analysis models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15556–15566 (2024). https://doi.org/10.1109/CVPR52733.2024.01473

[40] Yılmaz, Y., Hanoğlu, E.G., Özkan, A.G., Öztoprak, K.: Benchmarking OCR and vision-language models for Turkish text recognition: A comprehensive evaluation using synthetic data. International Journal on Document Analysis and Recognition (2026) https://doi.org/10.1007/s10032-026-00613-6

[41] Du, Y., Chen, P., Ying, X., Chen, Z.: DocPT-Bench: Benchmarking End-to-End Photographed Document Parsing and Translation (2025). https://doi.org/10.48550/arXiv.2511.18434

[42] Alaei, A., Bui, V., Doermann, D., Pal, U.: Document image quality assessment: A survey. ACM Computing Surveys **56**(2), 1–36 (2023) https://doi.org/10.1145/3606692

[43] Li, H., Zhu, F., Qiu, J.: Towards document image quality assessment: A text line based framework and a synthetic text line image dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 551–558 (2019). https://doi.org/10.1109/ICDAR.2019.00094

[44] Nayef, N., Luqman, M.M., Prum, S., Eskenazi, S., Chazalon, J., Ogier, J.-M.: SmartDoc-QA: A dataset for quality assessment of smartphone captured document images – single and multiple distortions. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1231–1235 (2015). https://doi.org/10.1109/ICDAR.2015.7333960

[45] Kumar, J., Ye, P., Doermann, D.: A dataset for quality assessment of camera captured document images. In: Camera-Based Document Analysis and Recognition (CBDAR 2013). Lecture Notes in Computer Science, pp. 113–125. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05167-3_9

[46] Kišš, M., Hradiš, M., Kodym, O.: Brno mobile OCR dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1352–1357 (2019). https://doi.org/10.1109/ICDAR.2019.00218

[47] Juvonen, M., Siltanen, S., Moura, F.S.: Helsinki deblur challenge 2021: Description of photographic data. Inverse Problems and Imaging **17**(5), 1008–1023 (2023) https://doi.org/10.3934/ipi.2022073

[48] Lindström, C., Hess, G., Lilja, A., Fatemi, M., Hammarstrand, L., Petersson, C., Svensson, L.: Are NeRFs ready for autonomous driving? towards closing the real-to-simulation gap. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4461–4471 (2024). https://doi.org/10.1109/CVPRW63382.2024.00449

[49] Chen, B.-H., Negrut, D.: Instance Performance Difference: A Metric to Measure the Sim-To-Real Gap in Camera Simulation (2024). https://doi.org/10.48550/arXiv.2411.07375

[50] Rajalingham, R., Issa, E.B., Bashivan, P., Kar, K., Schmidt, K., DiCarlo, J.J.: Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. The Journal of Neuroscience **38**(33), 7255–7269 (2018) https://doi.org/10.1523/JNEUROSCI.0388-18.2018

[51] Geirhos, R., Meding, K., Wichmann, F.A.: Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency.

In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 13890–13902 (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/9f6992966d4c363ea0162a056cb45fe5-Paper.pdf

[52] Klein, T., Meyen, S., Brendel, W., Wichmann, F.A., Meding, K.: Quantifying uncertainty in error consistency: Towards reliable behavioral comparison of classifiers. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 38, pp. 56807–56837 (2025). https://doi.org/10.52202/085713-1702

[53] Haber, M., Gao, J., Barnhart, H.X.: Assessing observer agreement in studies involving replicated binary observations. Journal of Biopharmaceutical Statistics **17**(4), 757–766 (2007) https://doi.org/10.1080/10543400701329547

[54] Obuchowski, N.A., Subhas, N., Schoenhagen, P.: Testing for interchangeability of imaging tests. Academic Radiology **21**(11), 1483–1489 (2014) https://doi.org/10.1016/j.acra.2014.07.004

[55] Obuchowski, N.A.: Can electronic medical images replace hard-copy film? defining and testing the equivalence of diagnostic tests. Statistics in Medicine **20**(19), 2845–2863 (2001) https://doi.org/10.1002/sim.929

[56] Calderon, N., Reichart, R., Dror, R.: The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16051–16081 (2025). https://doi.org/10.18653/v1/2025.acl-long.782

[57] Piaggio, G., Elbourne, D.R., Altman, D.G., Pocock, S.J., Evans, S.J.W., CONSORT Group: Reporting of noninferiority and equivalence randomized trials. JAMA **295**(10), 1152 (2006) https://doi.org/10.1001/jama.295.10.1152

[58] Tothfalusi, L., Endrenyi, L., Garcia Arieta, A.: Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. Clinical Pharmacokinetics **48**(11), 725–743 (2009) https://doi.org/10.2165/11318040-000000000-00000

[59] Pigeot, I., Schäfer, J., Röhmel, J., Hauschke, D.: Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. Statistics in Medicine **22**(6), 883–899 (2003) https://doi.org/10.1002/sim.1450

[60] Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I.: Circular analysis in systems neuroscience: The dangers of double dipping. Nature Neuroscience **12**(5), 535–540 (2009) https://doi.org/10.1038/nn.2303

[61] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960) https://doi.org/10.1177/001316446002000104

[62] Walker, E., Nowacki, A.S.: Understanding equivalence and noninferiority testing. Journal of General Internal Medicine **26**(2), 192–196 (2011) https://doi.org/10.1007/s11606-010-1513-8

[63] Lakens, D.: Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science **8**(4), 355–362 (2017) https://doi.org/10.1177/1948550617697177

[64] Berger, R.L., Hsu, J.C.: Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science **11**(4), 283–319 (1996) https://doi.org/10.1214/ss/1032280304

[65] Mindee: docTR: Document Text Recognition. GitHub (2021). https://github.com/mindee/doctr

[66] Mistral AI: Mistral-Small-3.2-24B-Instruct-2506 (2025). https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

[67] Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3D common corruptions and data augmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18941–18952 (2022). https://doi.org/10.1109/CVPR52688.2022.01839