Malafak: On-Device Arabic Document Vision, Reading and Handwriting with Privacy by Design

malafak.e-tony.com · Nov 14, 2026 · @amr

Abstract

Malafak («ملفك», *your file*) is an independent Arabic mobile web app. It helps Egyptian lawyers photograph, clean, assemble and size their case documents on their own phones, and then submit those documents themselves on official judicial portals. Every case document stays on the phone, encrypted; the server never receives, reads or crops a document.

This paper describes how we are teaching Malafak to *see* and *read* Arabic legal paperwork without breaking that rule. Three models do the work:

- **Malafak-Vision** runs on the phone. It finds the page, judges its quality and orientation, and detects stamps, signatures and handwriting.
- **Malafak-Reader** runs on the phone. It reads printed Arabic and handwriting.
- **Malafak-Teacher** runs only in a closed training environment. It labels data and distils its knowledge into the two small models.

Training data comes from three sources:

- a synthetic generator of Arabic court paperwork;
- open Arabic handwriting corpora;
- three narrow channels that lawyers may choose to join with one explicit consent: tiny blurred samples, a paragraph they copy in their own hand, and a whole page they pick and confirm themselves.

Anything that leaves a phone for training is encrypted *on the phone* with a public key. The server stores it but cannot open it.

As of November 2026, the following are built and tested: the app, the consent, the three channels, the encryption and the synthetic generator. All three channels stay closed until a legal review is complete. Accuracy figures in this paper are **targets** until we measure them on a hidden test set.

1. Problem and context

Egyptian lawyers increasingly file case documents through official online portals. They prepare those files on a phone, often in a courthouse corridor, with weak connectivity. Four obstacles recur:

- **Photos are hard to use.** Pages are photographed on desks, on top of other papers, tilted, shadowed or creased. A wrong crop can cut off a stamp, a case number or a signature, and that can get a filing rejected.
- **Files are too heavy.** Portals cap file size and page count. A lawyer usually cannot tell, before uploading, whether a file meets the limit or how to shrink it without making stamps unreadable.
- **Arabic is poorly served on the phone.** General OCR engines read clean printed Arabic tolerably. Court paperwork, however, mixes print, stamps, signatures and hurried handwriting, and on-device Arabic handwriting recognition is still weak.
- **The documents are confidential.** Client names, case numbers and personal data cannot be shipped to a cloud OCR service without legal and ethical risk.

Malafak's answer is to do the work on the phone itself, so the documents never need to leave it. This rule constrains everything that follows: the models must be small, fast and trained without collecting lawyers' documents.

2. System overview

Malafak is a Progressive Web App (React, TypeScript). It installs from the browser and keeps working without a connection. Its first load is small: 94.6 KB of JavaScript (gzip, measured in our build), against a budget of 150 KB. The app is independent, is not affiliated with the Ministry of Justice, and never touches government credentials: the lawyer opens the official portal and uploads the file themselves.

Where	What happens there	What it holds
The lawyer's phone	Capture, page finding and cropping, enhancement, page check, PDF assembly, size reduction against a portal's limits, and later reading	All case documents, encrypted in the browser's storage with a key that cannot be exported

Where	What happens there	What it holds
Our server	Accounts, the console's settings, the list of official portals and their file rules, and the consent records	No case document, client name, file name or case number, ever
The training environment (rented GPUs, by the hour)	Opening the locked training items with the private key, training, and evaluation	Only what lawyers chose to contribute, plus synthetic and open data; wiped after each run

Every request from the app is signed by a device key that is generated in the phone and cannot be copied. A stolen session cookie therefore does not work from another device. Cropping, reading and all other processing happen on the phone; no feature sends a page to a server to be processed.

3. The three models

We build on open models under licences that allow commercial use, and adapt them with transfer learning, low-rank adaptation (LoRA) and distillation. Training a model from scratch would need millions of images and months of compute without a better result at this size.

Model	Runs on	What it does	Base	Targets (not yet measured)
Malafak-Vision	The phone, after saving and in idle time	One backbone with several heads: page corners, page kind (sheet, ID card, stack, nothing), quality, orientation, stamps and signatures, presence of handwriting	A small open vision backbone, quantised to 8 bits	About 4 MB on the phone; finds at least as many pages as today's model on the fixed test set and never crops more harmfully

Model	Runs on	What it does	Base	Targets (not yet measured)
Malafak-Reader	The phone, after saving and in idle time	Detects and recognises printed Arabic and handwriting, orders the paragraphs, and will make searchable, editable files	PP-OCRv5 detection + the Arabic recognition model `arabic_PP-OCRv5_mobile_rec`	Character error rate (CER) of at most 4% on print, 15% on neat handwriting and 30% on quick notes
Malafak-Teacher	Only the closed training environment, never a phone or a public server	Labels images, teaches the two small models (distillation), and tests them	Qwen3-VL-2B-Instruct (Apache 2.0) with LoRA	Label quality checked by people on a sample

For the Arabic recogniser, its publishers report 81.27% accuracy on their own Arabic set of 2,676 text images. That is our starting point, not our result. Quick handwritten notes, read at a 30% CER, are used for search only, and the lawyer sees a «راجع النص» (*check the text*) warning. A legal name or number is never treated as confirmed without a person reviewing it.

Today the app crops with a research-licensed corner network (DocCornerNet via scanic-ml) and checks readability with Tesseract. Malafak-Vision and Malafak-Reader are meant to replace both.

4. Why the models run on the phone

We run the models on the phone for one reason above all: then the documents never have to leave it. Four reasons support the choice:

1. **Privacy.** A case file holds clients' names, national ID numbers and family and financial details. When reading and cropping happen on the device, the file never leaves the device, and nothing can leak from the server, its logs or its monitoring.
2. **Courts are offline places.** Lawyers prepare files in corridors and basements. A phone model works with no signal, and so does the whole app.

3. **Cost that does not grow with use.** Reading a page on the phone costs us nothing per page. Cloud OCR would charge per page, and a free tier for lawyers could not survive that.
4. **Speed.** There is no upload or round trip, and no wait behind other users.

The phone sets hard limits, and we design for them:

- **Size.** A model must download once and stay cached. We target about 4 MB for Malafak-Vision after 8-bit quantisation; today's corner network is 1.9 MB and its WebAssembly engine 1.5 MB.
- **Memory and heat.** Models run inside a Web Worker, one page at a time, never on the screen's thread.
- **Never in the lawyer's way.** Heavy models do not sit on the capture path. They run after the page is saved, when the phone is idle.
- **Graceful failure.** On an old phone without WebAssembly, or when a download is cut, the app falls back to its classic colour-based page finder and keeps working.

The runtime is ONNX Runtime Web (WebAssembly), served from Malafak's own origin; no third-party site is contacted.

5. Training data

Most of the training data never came from a lawyer. Real lawyer contributions are a small, narrow, consented supplement.

Source	What it is	Gives the models	Personal data
Synthetic generator (built and tested)	Invented Arabic court papers rendered as a phone would see them: on a desk or a stack, tilted, upside down, shadowed, creased, blurred, compressed. Invented names, numbers and dates; fake stamps; no ministry name or logo	Exact corners, page kind, orientation, stamps, signatures, every text line with its text (PaddleOCR format). Handwriting in Aref Ruqaa, word by word, with varied slant, size and baseline	None

Source	What it is	Gives the models	Personal data
Open corpora	KHATT: 1,000 writers, 1,000 forms, 9,327 lines (Pattern Recognition, 2014). Muharaf: 1,600+ expert-transcribed historic handwritten pages (NeurIPS 2024)	Real Arabic handwriting	Used under each corpus's terms
Channel 1: blurred samples	A copy of a few saved pages: the longest edge 384 px, *the whole picture blurred* until it cannot be read, plus the corners the model proposed and the corners the lawyer kept	Page finding and cropping	Unreadable; no id of any document, case, page, file, person or phone
Channel 2: «انسخ هذه الفقرة بخطك»	We show an invented legal paragraph; the person copies it by hand and photographs it, so the right answer is known in advance	Handwriting recognition	The photo of the copied paragraph only
Channel 3: a page the lawyer picks	The lawyer chooses a page that has no client data, ticks «راجعتُ الصفحة: لا توجد فيها بيانات لأي موكل أو شخص», and sends it	Real layouts and print	Only the page the lawyer chose and checked

Teachers. Channel 2 is also open to adult teachers who volunteer by private link, with their own consent text. Children's exercise books are excluded for two reasons: children's handwriting differs too much from lawyers', and children's data needs written parental consent.

The generator is deterministic: the same seed produces the same pictures. Every training run can therefore be repeated exactly.

6. Privacy and security by design

Nothing leaves a phone for training unless the lawyer explicitly agreed, and what does leave is unreadable to anyone but the training environment. Ten safeguards make this so:

1. **One explicit consent.** A single invitation lists the three ways of helping in a numbered list. It names one purpose, improving document processing in E-Tony apps, and says nothing is sold or given to anyone. The lawyer joins with one tap on «أوافق وأساعد», and nothing is pre-ticked. «ليس الآن» is remembered. Every answer is recorded with the text edition it answered.
2. **A new meaning means a new question.** The edition of the text lives in the code. When a release changes what the lawyer agrees to, everyone who agreed earlier is asked again automatically.
3. **The lawyer acts; consent only allows.** The copied paragraph is photographed by the lawyer. A whole page leaves only after the lawyer taps «أرسل هذه الصفحة» on that page and ticks the check themselves. The app never collects on its own beyond the blurred samples.
4. **Per-document control.** Every document has a switch. A document switched off is never sampled.
5. **A waiting time.** Each item waits on the phone, encrypted, for a period the owner sets (24 hours by default) before it may leave. A page added by mistake can be deleted, and a waiting page can be cancelled from «حسابي».
6. **Blurring.** Channel 1 samples are reduced to 384 px and blurred as a whole picture, so text cannot be read.
7. **Locked on the phone.** Each item is sealed on the phone before it leaves: ECDH on P-256 with a fresh key for every item, then HKDF-SHA-256, then AES-256-GCM. The header is authenticated too. The server stores the sealed file under a random name; the private key never touches the server.
8. **Strict, minimal records.** The server keeps a random id, the day, the size and the text edition. It never keeps a document, case, page or file id, a name, a phone number or a time of day.
9. **Withdrawal deletes.** «إيقاف المساعدة» stops everything at once and deletes what waits on the phone. A database trigger deletes the server rows in the same moment, and a background job removes their files.

10. **The owner's switches.** Training, sending and whole pages are separate switches in the console, all off by default. They stay off until the legal review is complete.

Legal basis. Egypt's Personal Data Protection Law (No. 151 of 2020) and its Executive Regulations (Ministerial Decision No. 816 of 2025) apply (summary). They require consent that is explicit, informed, specific, freely given and easy to withdraw, and they hold that consent is unsuitable when a service depends on it. For that reason helping is never a condition of using Malafak, and no prompt is disguised as anything else. The text, the storage and the transfers are under the owner's legal review before any sending opens. This paper is not legal advice.

What never leaves the phone: case documents (except a page the lawyer picks and confirms), client names, case numbers, file names, phone numbers, and the codes lawyers type.

7. Training and evaluation

All training runs on GPUs rented by the hour in the cloud. No training happens on a personal computer or on the app's server. The training environment is one Docker image with pinned dependencies. It is built once, started for a run, and wiped afterwards together with the decrypted items.

- **Hardware.** An RTX 4090 for the two small models and an A100 80 GB for the teacher. Generating synthetic data needs only many CPU cores.
- **Method.** The teacher (Qwen3-VL-2B with LoRA) labels real and synthetic images. The small models learn from those labels and from the synthetic ground truth by distillation, then are quantised for the phone.
- **Hidden test set.** A fixed set of real-world phone photos that no model trains on. Scores count only on this set. Today it records that the current corner model finds the page in 9 of 11 hard photos and never makes a harmful crop.
- **Metrics.** Pages found, and *harmful crops* (a crop that cuts off part of the page, which we treat as worse than a missed page). For reading, the character error rate (CER) on print, neat handwriting and quick notes.

A new model replaces the old one only after four steps:

1. **Shadow.** It runs silently on the phones of lawyers who agreed. Its results are compared and never shown.
2. **Hidden test.** It meets its targets on the hidden test set.

3. **No more harm.** It never crops more harmfully than the model it replaces, and finds at least as many pages. A rule in the code enforces this.
4. **Gradual rollout.** It reaches 5%, then 25%, then all lawyers, set from the console. One tap rolls it back.

8. Current status (November 2026)

The product is live with a closed group of 10 lawyers. Everything needed to collect training data safely is built and tested, but collection is switched off until the legal review is complete.

Part	State	Evidence
App: capture, cropping, enhancement, PDF, size reduction to a portal's limit, hand-off to the official site	In use by the closed group	125 browser tests and about 800 unit and integration tests pass on every change
One consent, per-document switches, withdrawal	Built; invitation switched off	Tests for each rule, including withdrawal deleting server rows by trigger
Channel 1: blurred samples	Built; sending off	Only numbers reach the server until sending opens
Channel 2: copied paragraph (lawyers and teachers)	Built; first real uploads verified end to end	The owner's own test uploads arrived sealed
Channel 3: a page the lawyer picks	Built; off by default	A browser test shows the tick is required, the page waits, can be cancelled, and leaves sealed
On-phone encryption	Built	A test vector written in JavaScript opens in Python on the training side
Synthetic generator	Built; tested in CI	Deterministic output per seed
Malafak-Vision, Malafak-Reader	Not yet trained	All accuracy figures are targets

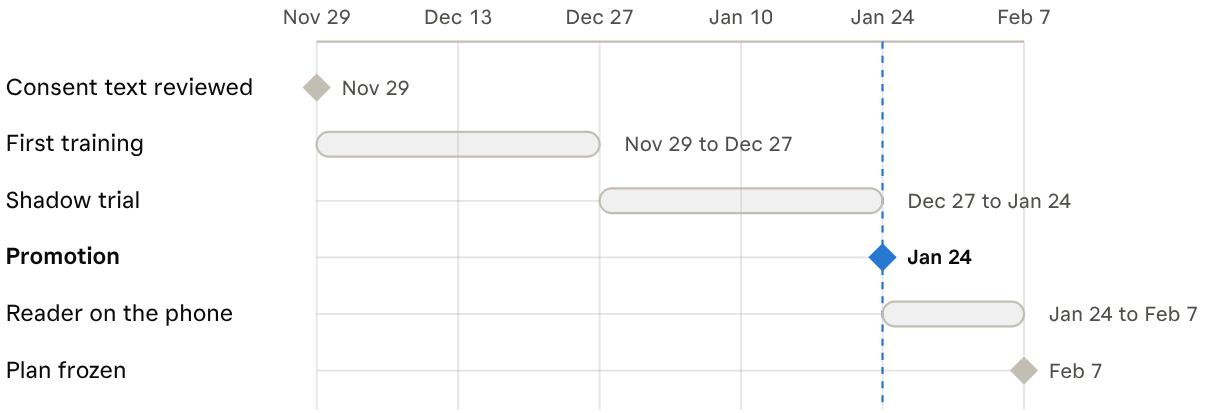
The searchable-PDF tool was removed from the app this month. Today's on-device Arabic reading missed too many words to make search useful. It will return with Malafak-Reader, once the reader's accuracy is measured.

9. Roadmap and future development

The first 12 weeks, from 16 November 2026 to 7 February 2027, take the models from no data to a promoted model on lawyers' phones.

New models reach lawyers on 24 January, after a month in the shadow

The 12-week plan, checkpoints by date, 2026 to 2027



12-week plan · 4 checkpoints

The dates are checkpoints, and a model is promoted only if it passes the rule in section 7. If it fails, the date moves; the rule does not.

After the plan, in order of value to lawyers:

1. **The searchable and editable PDF returns**, made by Malafak-Reader on the phone, with a «راجع النص» warning wherever the reading is unsure.
2. **Layout understanding.** Headings, parties, case number and dates will be found on the page, so files can be named and sorted automatically. The lawyer confirms each result.
3. **More document kinds.** Powers of attorney, ID cards, receipts and court forms, each with its own fixed test set.
4. **Hosting in Egypt** for training storage, as already decided.
5. **Research partnerships.** Arabic handwriting is under-resourced. We aim to share our synthetic generator, our evaluation method and aggregate results (never lawyers' data) with universities working on Arabic HTR.

10. Impact, grant fit and budget

Funding would buy compute, measurement and Arabic data. It would not buy access to anyone's documents.

Who benefits.

- **Lawyers.** Fewer rejected filings and less time spent fixing photos and file sizes, on a phone they already own, with no new privacy risk.
- **Courts and portals.** Cleaner, correctly sized uploads.
- **Arabic language technology.** An on-device method, a synthetic generator and an evaluation protocol for Arabic legal print and handwriting, an area with few public resources.

Why this approach is fundable.

- It shows that useful document AI can be built for a sensitive profession without centralising that profession's data.
- Every claim is checked against a hidden test set, and results are reported as targets until measured.
- The privacy controls are built and tested before any data flows, not promised afterwards.

Compute budget for the 12-week plan. Estimated at USD 560 to 720 of rented GPU time, reviewed every week. Synthetic generation runs on CPUs. Training uses one RTX 4090 for the small models and one A100 80 GB for the teacher, and each machine is switched off when its job ends.

What additional funding enables.

1. A larger and more varied hidden test set, built with lawyers' explicit permission.
2. More handwriting contributors, through volunteer teachers with their own consent.
3. Moving training storage to hosting in Egypt.
4. An independent privacy and security audit.
5. Publication of the synthetic generator and the evaluation protocol for researchers.

11. Limitations, risks and ethics

Risk	What we do
A blurred sample could still reveal a large stamp or logo	The blur radius grows with the picture, and blurring is measured again on trial photos before sending opens
A lawyer could tick «راجعتُ الصفحة» by mistake	A warning on the screen, a waiting time with «ألغِ الإرسال», deletion on withdrawal, and whole pages off by default
Synthetic pages differ from real phone photos	Augmentation (shadow, crease, blur, compression), real corpora for handwriting, and a hidden test set of real photos as the only judge
Handwriting accuracy may stay low	Notes read at the 30% CER target serve search only, always with a «راجع النص» warning; no legal fact is taken from them unchecked
Small models on older phones	Idle-time work, one page at a time, a 4 MB target, and a fallback that keeps the app working without the model
Pressure to collect more	One consent, one stated purpose, and helping is never a condition of service; prompts say plainly what they are for
Licences of base models and data	Only bases licensed for commercial use; each corpus used under its own terms; today's research-licensed corner model is the one being replaced

Malafak is independent of the Ministry of Justice and of any portal. It never stores government credentials or automates uploads, and it uses no ministry logo.

References

1. PaddleOCR. Introduction to PP-OCRv5 Multilingual Text Recognition, including `arabic_PP-OCRv5_mobile_rec`. See also the PaddleOCR 3.0 Technical Report (2025).

2. Qwen. Qwen3-VL-2B-Instruct, Apache 2.0 licence.
3. Mahmoud, S. A., et al. KHATT: An open Arabic offline handwritten text database. Pattern Recognition 47(3): 1096–1112, 2014.
4. Saeed, M., et al. Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition. NeurIPS 2024 Datasets and Benchmarks Track.
5. scanic, the source of the DocCornerNet model used by today's app.
6. CMS. Egypt's PDPL: Executive Regulations issued, one-year compliance countdown begins. Covers Law No. 151 of 2020 and Ministerial Decision No. 816 of 2025.

Contact and product: malafak.e-tony.com. An Arabic version follows in the next tab; a short public summary is in the last tab.