

From Observed Model Assets to Tiny Local Document Intelligence

Chen, Ho Yiing (norika)
Independent Researcher, Taiwan
ORCID: 0009-0006-6816-9891
norika at charenix.com
Zenodo DOI: 10.5281/zenodo.20728655

June 17, 2026

Abstract

I report a practical method for turning a messy collection of observed language and vision systems into a small local document-intelligence bundle. The target is not a general chatbot and not a cloud OCR service. The target is an artifact below 50 MB when possible, below 100 MB during alpha, that converts guided-capture business cards or receipts into structured JSON. The method starts from a Mercury fullgrid archive: 19 trials are scored for coverage, formal validity, family, and deployment role. Qwen-family runs dominate the CJK OCR-to-JSON teacher list; Phi-2 and InternLM2 become schema-repair and normalization teachers; BLOOM/BLOOMZ become multilingual controls. A separate vision inventory provides the image front end: Qwen2.5-VL as a teacher and PP-OCrv5 mobile weights as the deployable OCR candidate. The resulting receipt student bundle is 31.716 MB and reaches 91.8% exact match and 99.2% field accuracy on a 500-example holdout. The card runtime stays under the 50 MB payload line, but realistic app tests expose a rollback case: a compact candidate scored 0% exact match on one app-side holdout while the older fusion pipeline scored 84.6% exact match and 97.4% average field accuracy. The lesson is not glamorous, but it is the business line: small bundles need observed teacher selection, explicit image/OCR boundaries, and ship gates that are allowed to say no.

1 Why This Note Exists

It is easy to confuse three different things. A language-only system can repair JSON, but it cannot see a business card. A vision-language system can label hard cases, but it is not a 50 MB phone artifact. A tiny local runtime can be useful, but only inside a narrow capture and page boundary.

This note writes down the bridge between those things. I do not claim that a 31 MB or 50 MB runtime replaces a 7B vision-language system. I claim something more useful: an observed archive can be mined into a teacher inventory, and that inventory can drive small, local, task-specific file-understanding bundles.

The shipping law is deliberately narrow:

Tiny bundle. Local. Fast. One painful file type. Clean structured output.

For the first pass, the painful file types are Japanese receipts and business cards. The raw user value is simple. Take a guided photo; get JSON. Month-end receipt work and contact-entry work should become smaller, not more ceremonial.

2 Observed Asset Selection

Let each observed trial be indexed by i . The first paper defined a fullgrid completion gate. I reuse only its consequence here: a trial is usable for teacher mining only if it crosses the formal observation line, unless it is explicitly marked as a control.

For trial i , define:

$$F_i = \mathbf{1}[\text{actual_cells}_i \geq 10^8] \cdot \mathbf{1}[\text{coverage}_i \geq 90].$$

This is not a quality score. It is a permission bit. A trial with $F_i = 0$ can still be historically interesting, but it should not silently become a teacher teacher.

I assign a deployment score:

$$S_i = 40F_i + 25CJK_i + 15J_i + 10M_i + 10R_i,$$

where CJK_i marks strong Japanese/Chinese usefulness, J_i marks structured JSON or instruction-following value, M_i marks multilingual robustness, and R_i marks repair or control-flow usefulness. The coefficients are not a law. They are a readable decision surface. If someone disagrees, they can change the weights and recompute the ranking.

Rank	Role	Coverage	Score
1	Qwen2.5 3B CJK OCR-to-JSON teacher	94.95%	97
2	Qwen3B CJK extractor backbone	94.88%	97
3	Qwen3B q6s15 compact CJK candidate	90.64%	97
4	Qwen7B CJK teacher	94.15%	95
5	Phi-2 schema repair teacher	97.05%	64
7	InternLM2 CJK normalization teacher	90.87%	56
8–10	BLOOMZ multilingual robustness controls	94.18–95.85%	54

Table 1: Observed teacher inventory extracted from the Mercury asset table. The table is a teacher-mining list, not a semantic claim about the systems.

The important decision is negative as much as positive. GPT-2 and GPT-2 Medium remain legacy tiny baselines. Gemma 2B in the local archive did not cross the formal line. Language-only Llama, Mistral, BLOOM, OPT, Pythia, Phi-2, and Qwen cannot be presented as image readers. If they are used, they operate after OCR or after a vision teacher has produced blocks.

3 The Image Boundary

File intelligence has at least three layers:

$$I \xrightarrow{O} B \xrightarrow{E} Y \xrightarrow{V} \hat{Y}.$$

Here I is the image, O is OCR or visual parsing, B is a set of OCR blocks with geometry, E is the tiny extractor, Y is candidate JSON, and V is validation and repair. A language-only system can help in E or V . It does not implement O .

The local inventory therefore uses two classes of assets.

- Teacher assets: Qwen2.5-VL-7B and larger multimodal systems can label hard images, produce structured supervision, and expose failure modes.
- Deployable assets: PP-OCRv5 mobile detector, recognizer, and orientation weights are small enough to be credible on device.

The size equation for a 50 MB bundle is:

$$P = O_d + O_r + O_\ell + E + V + A,$$

where O_d is line detection, O_r is character recognition, O_ℓ is orientation/layout support, E is the extractor, V is validation logic, and A is auxiliary metadata. The discovered mobile OCR core is:

$$O_d + O_r = 4.8 + 17.0 = 21.8 \text{ MB.}$$

Adding page and line-orientation support gives:

$$21.8 + 13.3 = 35.1 \text{ MB.}$$

That leaves roughly:

$$50 - 35.1 = 14.9 \text{ MB}$$

for extractor weights, validation tables, and metadata if the hard target is kept. UVDoc-style unwarping can help, but at 31 MB it moves the bundle toward the 100 MB alpha ceiling:

$$35.1 + 31.0 \approx 66.1 \text{ MB.}$$

4 Receipt Student Result

The receipt student is the cleanest current evidence that this route is not only a diagram. The receipt manifest reports:

$$P_{\text{receipt}} = 31.716 \text{ MB.}$$

On the 500-example holdout:

$$A_{\text{json}} = 1.000, \quad A_{\text{exact}} = 0.918, \quad A_{\text{field}} = 0.992.$$

The ship gate for this artifact can therefore be written:

$$G_{\text{receipt}} = \mathbf{1}[A_{\text{json}} = 1] \cdot \mathbf{1}[A_{\text{field}} \geq 0.99] \cdot \mathbf{1}[P \leq 50].$$

For the current manifest, $G_{\text{receipt}} = 1$ on those three measured conditions. That still does not mean the artifact is public-ready. The manifest keeps remaining blockers, including hostile real pages and OCR-front-end limits. This is the right behavior. The runtime should pass the narrow gate and still be honest about the wider public gate.

5 Card Runtime: The Failure That Matters

The business-card side is the cautionary result. Synthetic and source-ghost sets reached high scores, including 10,000-row hostile layout checks in which the local student and fusion layer reported 100% exact match. Those numbers are useful for regression. They are not enough for production.

The app-side rollback test gives the important contrast:

The failure has a clean formula. Let D_t be the training distribution and D_a the app distribution. A public claim is invalid when:

$$A(D_t) - A(D_a) > \epsilon.$$

For the v11 compact candidate, the exact-match gap is effectively:

$$1.00 - 0.00 = 1.00.$$

No marketing language can repair that. The correct action was rollback and handoff. The previous production path remained:

$$\text{v4 fusion} = 84.6\% \text{ exact, } 97.4\% \text{ field.}$$

Pipeline	Exact match	Average field accuracy
v3-DR alone	67.4%	81.66%
v4 fusion, older bundle + VLM	84.6%	97.4%
v11 compact, no VLM	0.0%	69.7%

Table 2: App-side card holdout result. The compact candidate is not allowed to replace the older fusion path until the distribution gap is repaired.

6 Privacy-Safe Learning Object

The runtime cannot require raw private pages. That would turn a local tool into another cloud data vacuum. The learning object should be a layout ghost:

$$L = (\text{geometry, field roles, confidence, failure type}) - \text{private content}.$$

Names, phone numbers, email addresses, addresses, raw receipt OCR, and full page images are not needed for the default improvement loop. What matters is the shape of the failure:

$$e = \hat{y} - y,$$

and the location of that error in the layout:

$$\ell(e) = (x_1, y_1, x_2, y_2, \text{field, file family}).$$

This is the business moat if it is done correctly. The system learns from weird layouts without needing to own the user’s private life.

7 Ship Gates

I use four gates.

$$G_{\text{size}} = \mathbf{1}[P \leq 50]$$

for public target size, with an alpha exception:

$$G_{\alpha} = \mathbf{1}[P \leq 100].$$

$$G_{\text{json}} = \mathbf{1}[\text{Pr}(\text{valid JSON}) = 1]$$

for structured output.

$$G_{\text{field}} = \mathbf{1}[A_{\text{field}} \geq \tau_f]$$

where τ_f should be at least 0.99 for accounting-like receipt fields.

$$G_{\text{photo}} = \mathbf{1}[A_{\text{guided photo}} \geq \tau_p]$$

for guided-capture photo tests. An artifact can pass synthetic regression and still fail G_{photo} .

The shipping bit is:

$$G_{\text{ship}} = G_{\text{size}} G_{\text{json}} G_{\text{field}} G_{\text{photo}}.$$

If any term is zero, the bundle should not be sold as finished. It may still be an alpha, an internal teacher, or a regression artifact.

8 Relation to Prior Work

This work borrows from three neighboring ideas but sits in a different place. Document systems such as LayoutLM jointly model OCR tokens and layout [1]. OCR-free systems such as Donut try to remove the external OCR step [2]. Qwen2.5-VL and similar VLMs are strong document teachers [3]. PP-OCR shows that lightweight OCR is not fantasy [4]. Distillation work shows that large-system behavior can be moved into smaller students under constraints [5, 6, 7].

The difference here is operational. I am not proposing one universal compact network. I am proposing a gated factory:

observed teachers → privacy safe supervision → tiny specialist artifact → app holdout gate.

9 Limitations

The evidence is still early. The card artifact has a real distribution-gap failure. The receipt artifact has strong 500-holdout numbers but still needs broader real receipt, invoice, and guided-photo tests. Some source-ghost data is useful for layout coverage but cannot replace private real-photo holdouts. The OCR front end remains a separate bottleneck. An extractor can only structure what the image layer sees.

Those limits are not footnotes. They are the reason to write the gates as mathematics. A gate line that can say “not yet” is more honest than a demo that pretends to be done.

10 Conclusion

The useful result is not that one small runtime suddenly became a 7B VLM. The useful result is a path: observe large systems, select teachers with a formal gate, manufacture privacy-safe supervision, distill into a tiny file artifact, and let app-side holdouts decide whether it is allowed to ship.

For Iseeu, the immediate direction is clear. Receipt JP has the stronger measured 50 MB story today. Card has a promising guided-capture path but must repair the app distribution gap before it gets public language. Manuals, warranty cards, forms, and contracts should enter the same factory only when their OCR boundary, schema, and ship gate are written down first.

Data and Artifacts

The local evidence used in this note is stored under:

```
docs/ISEEU_OBSERVED_MODEL_ASSETS.md
docs/ISEEU_VISION_MODEL_INVENTORY.md
dist/iseeu/
var/product/iseeu-llm/
```

Acknowledgements

This work was carried out as independent research using personally maintained local machines and opportunistic off-hours compute access. Draft preparation and code editing were assisted by AI coding tools under the author’s direction.

References

- [1] Xu, Yiheng; Li, Minghao; Cui, Lei; Huang, Shaohan; Wei, Furu; and Zhou, Ming. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. KDD, 2020. <https://arxiv.org/abs/1912.13318>.
- [2] Kim, Geewook; Hong, Teakgyu; Yim, Moonbin; Nam, Jeongyeon; Park, Jinyoung; Yim, Jinyeong; Hwang, Wonseok; Yun, Sangdoo; Han, Dongyoon; and Park, Seunghyun. OCR-free Document Understanding Transformer. ECCV, 2022. <https://arxiv.org/abs/2111.15664>.
- [3] Qwen Team. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025. <https://arxiv.org/abs/2502.13923>.
- [4] Du, Yuning; Li, Chenxia; Guo, Ruoyu; Yin, Xiaoting; Liu, Weiwei; Zhou, Jun; Bai, Yifan; Yu, Zilin; Yang, Yehua; Dang, Qingqing; and Wang, Haixiang. PP-OCR: A Practical Ultra Lightweight OCR System. arXiv preprint arXiv:2009.09941, 2020. <https://arxiv.org/abs/2009.09941>.
- [5] Sanh, Victor; Debut, Lysandre; Chaumond, Julien; and Wolf, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019. <https://arxiv.org/abs/1910.01108>.
- [6] Jiao, Xiaoqi; Yin, Yichun; Shang, Lifeng; Jiang, Xin; Chen, Xiao; Li, Linlin; Wang, Fang; and Liu, Qun. TinyBERT: Distilling BERT for Natural Language Understanding. EMNLP Findings, 2020. <https://arxiv.org/abs/1909.10351>.
- [7] Sun, Zhiqing; Yu, Hongkun; Song, Xiaodan; Liu, Renjie; Yang, Yiming; and Zhou, Denny. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. ACL, 2020. <https://arxiv.org/abs/2004.02984>.