Supplementary Information

Handwritten German Administrative Forms (HGAF): A Dataset for Handwriting Recognition with Fine-Tuned and Zero-Shot Baselines

Tobias Kalmbach[1*] and Andreas Wagner[1,2]

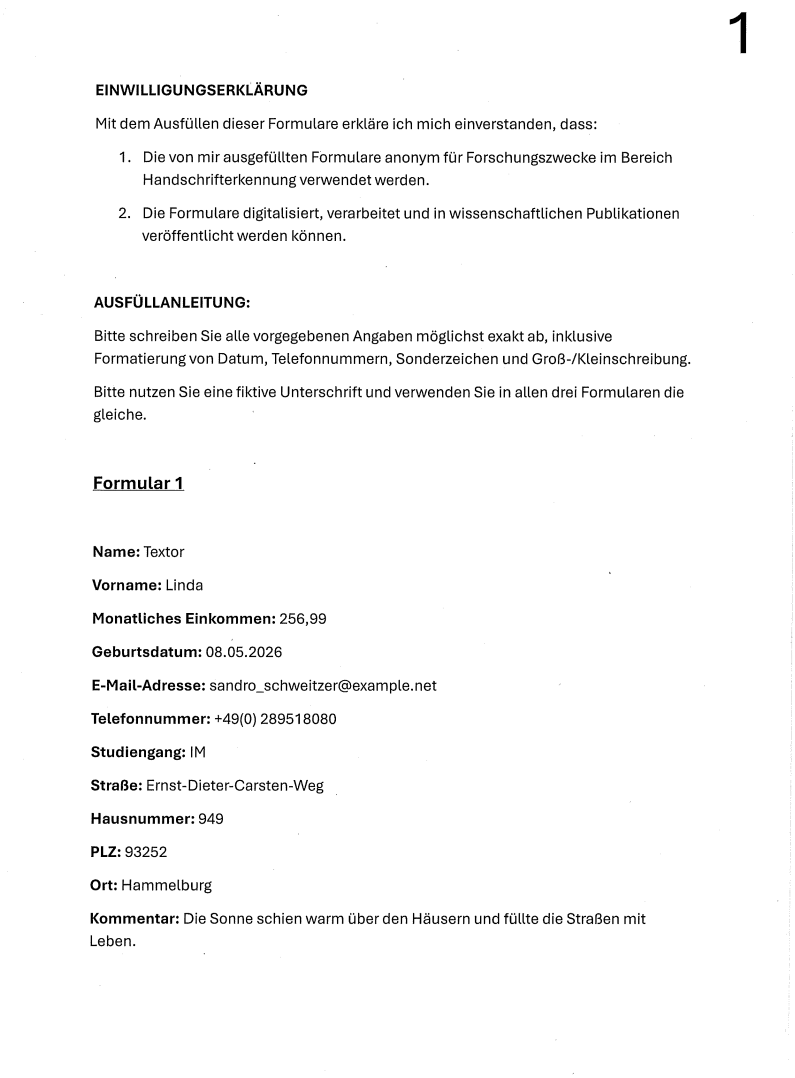
[1]Karlsruhe University of Applied Sciences, Faculty of Management Science and Engineering, Karlsruhe, Germany.
[2]Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany.

*Corresponding author(s). E-mail(s): tobias.kalmbach@proton.me;
Contributing authors: andreas.wagner@h-ka.de;

Supplementary information for the article submitted to the *International Journal on Document Analysis and Recognition (IJDAR)*. Section, figure and table numbers carry the prefix S and are cited from the article as, for example, "Supplementary Information, Fig. S7". Section S1 reproduces the three layouts and the instruction for completion at full size, Sects. S2 and S3 give the settings of the specialist pipeline and its post-processing together with the raw and post-processed results of every model, Sect. S4 collects all qualitative error examples, and Sect. S5 reproduces the prompts passed to the vision-language models. The field inventory of the forms and the stages of the specialist pipeline are given in the article itself.

S1 Form Layouts

The three layouts of the corpus are reproduced below at full size, together with the instruction for completion that each participant received alongside them. All four documents are shown for document ID 1; the corresponding pages exist for every document ID from 1 to 100. The instruction for completion carries the synthetically generated field values that the participant was asked to copy onto the three forms, and the consent statement under which the handwriting was collected. Scans are reproduced from the 300 dpi variant used throughout this work.



1

EINWILLIGUNGSERKLÄRUNG

Mit dem Ausfüllen dieser Formulare erkläre ich mich einverstanden, dass:

1. Die von mir ausgefüllten Formulare anonym für Forschungszwecke im Bereich Handschrifterkennung verwendet werden.
2. Die Formulare digitalisiert, verarbeitet und in wissenschaftlichen Publikationen veröffentlicht werden können.

AUSFÜLLANLEITUNG:

Bitte schreiben Sie alle vorgegebenen Angaben möglichst exakt ab, inklusive Formatierung von Datum, Telefonnummern, Sonderzeichen und Groß-/Kleinschreibung.

Bitte nutzen Sie eine fiktive Unterschrift und verwenden Sie in allen drei Formularen die gleiche.

Formular 1

Name: Textor

Vorname: Linda

Monatliches Einkommen: 256,99

Geburtsdatum: 08.05.2026

E-Mail-Adresse: sandro_schweitzer@example.net

Telefonnummer: +49(0) 289518080

Studiengang: IM

Straße: Ernst-Dieter-Carsten-Weg

Hausnummer: 949

PLZ: 93252

Ort: Hammelburg

Kommentar: Die Sonne schien warm über den Häusern und füllte die Straßen mit Leben.

Fig. S1 Instruction for completion, document ID 1. The upper block is the consent statement and the filling instructions, the lower block the generated field values to be copied onto the three forms

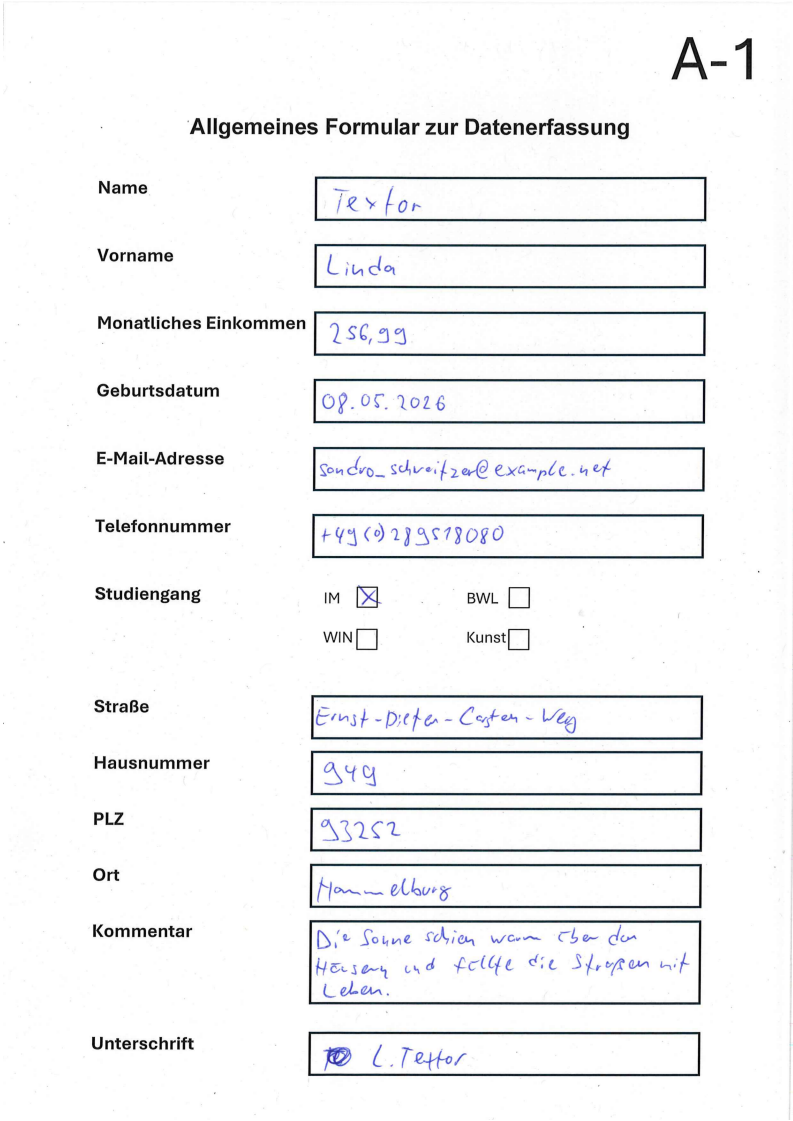
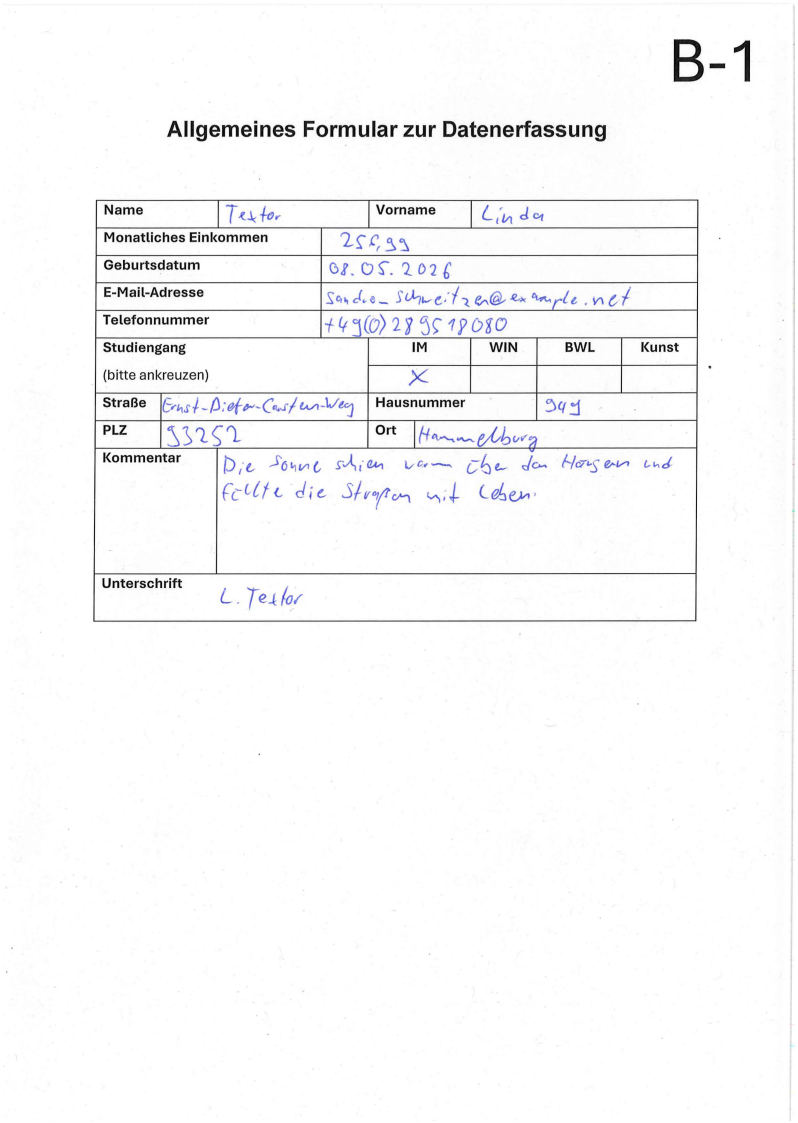


Fig. S2 Layout A, document ID 1. Every field has its own generously sized box, and the labels sit to the left of the boxes. Designed as the least constrained case for the writer and the most constrained case for field localisation



B-1

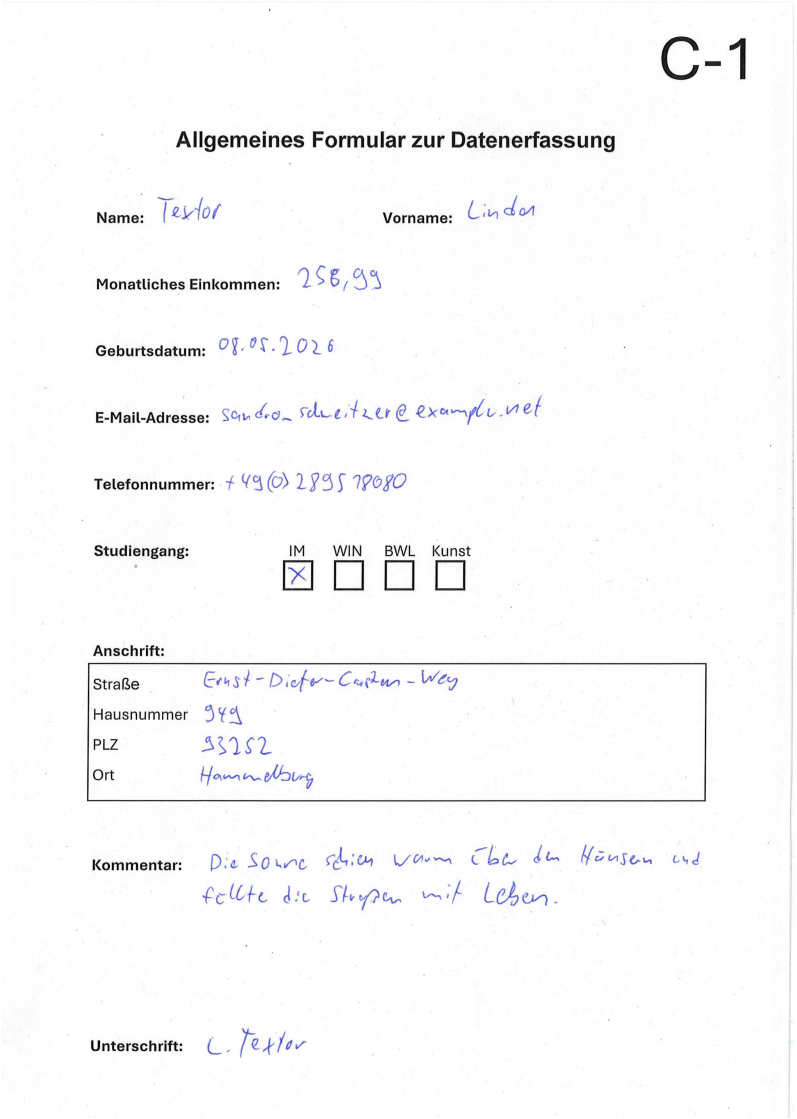
Allgemeines Formular zur Datenerfassung

Name	Testor	**Vorname**	Linda
Monatliches Einkommen	2500,99		
Geburtsdatum	08.05.2026		
E-Mail-Adresse	sandra_schreitzer@example.net		
Telefonnummer	+49(0)2895 18080		

Studiengang (bitte ankreuzen)	**IM**	**WIN**	**BWL**	**Kunst**
	X			

Straße	Ernst-Dieter-Carstens-Weg	**Hausnummer**	949
PLZ	33252	**Ort**	Hammelburg
Kommentar	Die Sonne schien warm über den Hügeln und füllte die Straßen mit Leben.		
Unterschrift	L. Testor		

Fig. S3 Layout B, document ID 1. A compact tabular arrangement with small cells, which forces the writer to write smaller and places neighbouring field values close together



C-1

Allgemeines Formular zur Datenerfassung

Name: Textor Vorname: Linda

Monatliches Einkommen: 258,99

Geburtsdatum: 08.05.2026

E-Mail-Adresse: sandra_schweitzer@example.net

Telefonnummer: +49 (0) 2895 78080

Studiengang: IM [X] WIN [ ] BWL [ ] Kunst [ ]

Anschrift:

Straße	Ernst-Dieter-Carstens-Weg
Hausnummer	949
PLZ	93252
Ort	Hammelburg

Kommentar: Die Sonne schien warm über den Häusern und
füllte die Straßen mit Leben.

Unterschrift: L. Textor

Fig. S4 Layout C, document ID 1. Most fields are not delimited by a box or a rule, so the writer chooses where and how to place the value relative to the printed label. The address block is the one exception

S2 Specialist Pipeline Settings

Table S1 lists the principal parameters of the fine-tuned specialist pipeline, whose stages are described in Table 4 of the article. All values were fixed on the validation split of layout A.

Table S1 Principal parameters of the fine-tuned specialist pipeline

Parameter	Value	Description
Page pre-processing		
`deskew_min_angle`	1.0°	Pages closer to upright than this are left unresampled.
Segment pre-processing		
`remove_lines`	enabled (Telea)	Printed lines and field borders are removed morphologically; the resulting gaps are filled by Telea inpainting.
`crop_to_text`	enabled	Each segment is cropped to the bounding box of its ink.
`crop_threshold`	200	Grey level (8-bit) separating ink from background.
`crop_padding`	5 px	Margin retained on each side of the crop.
`min_height`	32 px	Shorter crops are upscaled to the resolution the recogniser expects.
Segment alignment		
`orb_features`	5,000	ORB keypoints detected per page.
`max_matches`	500	Best matches kept for the homography.
`ransac_threshold`	5.0 px	Reprojection threshold of the RANSAC fit.
Line segmentation (long-text fields)		
`smooth_frac_of_height`	0.035	Smoothing window of the row-wise ink profile, as a fraction of crop height.
`peak_min_prominence_frac`	0.15	Minimum peak prominence, as a fraction of the profile maximum. Stable between 0.05 and 0.30.
`valley_max_frac`	0.35	Two peaks are merged when the valley between them exceeds this fraction of the smaller peak.
`peak_min_distance_frac`	0.30	Minimum spacing between peaks, as a fraction of the median peak spacing.
`join_separator`	single space	Separator concatenating the per-line predictions of a multi-line field.
Checkbox fields		
`min_ink`	0.02	If the highest ink ratio in a group stays below this, the group is recorded as `no_mark`.
`margin_threshold`	0.15	If the two highest options differ by less than this, the group is recorded as `ambiguous`.

S3 Post-Processing

Table S2 shows the specific post-processing methods applied to each field and Table S3 the rules and thresholds governing them. Post-processing was defined for TrOCR on

the validation data of layout A; Table S4 reports the effect of applying the same rules to the raw output of every model on the test splits.

Table S2 Overview of applied post-processing methodologies by field

Fields	**Post-processing**
First name Last name	Letter-number error matrix (e.g. correct a "6" to a "G").
Monthly income House number Postal code	Number-letter error matrix (e.g. correct an "I" to a "1").
Birthday	If the date consists only of digits and symbols, remove white spaces. If it contains text, correct the month word with a lexicon of German month names. Accepted only if the resulting date denotes a real calendar day.
Email	Remove white spaces. Domain correction with a lexicon of common domains.
City	Correction with a lexicon of all German cities.[a]
Street	Suffix correction with a lexicon of German street suffixes.

[a] Extracted from Wikipedia's list of cities and towns in Germany: https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Germany

Table S3 Thresholds governing the post-processing rules. Relative distances are expressed as a fraction of the length of the string being matched

Parameter	Value	Description
Numeric and date fields		
digit confusion map	17 entries	Hand-specified letter-to-digit substitutions. A correction is kept only if the result then matches the field's pattern.
`month_max_relative_distance`	0.34	Maximum distance to a German month name for the spelled-out date form.
calendar check	enabled	A corrected date is accepted only if it denotes a real calendar day, leap years included.
Lexicon fields (city)		
`lexicon_max_relative_distance`	0.34	Maximum distance to a lexicon entry. Distances are computed on diacritic-expanded lowercase forms; the lexicon's own spelling is returned.
Email		
`email_max_relative_distance`	0.34	Maximum distance to a known domain. The local part is never corrected.
Street		
`street_max_relative_distance`	0.34	Maximum distance to a street-type suffix, scaled per target so that longer suffixes tolerate more edits. The base name is never corrected.
Name fields		
reverse digit map	8 entries	Digits are mapped back to a letter where the confusion table justifies one, and deleted otherwise. No lexicon is applied.

Table S4 Raw against post-processed CER and FER. Lower is better throughout; a negative delta denotes an improvement. The better of the two values per model and metric is in bold

		Raw		Post-processed		Delta	
Layout	**Model**	**CER (%)**	**FER (%)**	**CER (%)**	**FER (%)**	**CER (%)**	**FER (%)**
A	TrOCR	7.4	39.9	**6.8**	**35.4**	−0.6	−4.5
	Qwen	4.6	27.5	**4.2**	**25.6**	−0.4	−1.9
	Chandra	4.5	27.0	**4.0**	**25.2**	−0.5	−1.8
	DeepSeek	16.1	57.1	**14.7**	**52.9**	−1.4	−4.2
B	TrOCR	12.9	53.1	**12.1**	**48.5**	−0.8	−4.6
	Qwen	7.2	37.3	**6.5**	**34.5**	−0.7	−2.8
	Chandra	7.2	35.2	**6.7**	**33.8**	−0.5	−1.4
	DeepSeek	20.7	64.6	**20.0**	**62.5**	−0.7	−2.1
C	TrOCR	16.1	57.3	**15.4**	**53.8**	−0.7	−3.5
	Qwen	6.5	36.8	**5.9**	**33.8**	−0.6	−3.0
	Chandra	5.8	33.3	**5.3**	**31.5**	−0.5	−1.8
	DeepSeek	26.1	63.4	**24.1**	**60.6**	−2.0	−2.8

S4 Qualitative Error Examples

This section collects all qualitative examples discussed in the article, including the instance shown in each case in the article itself, so that the complete set is available in one place.

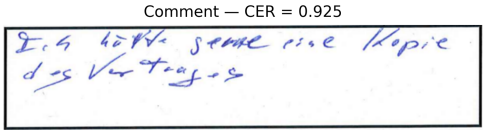
```
<|ref|>text<|/ref|><|det|>[[113, 480, 598,
     529]]<|/det|>
Studiengang: IM WIN BWL Kunst
X ☐ ☐ ☐
```

(a)

```
<|ref|>text<|/ref|><|det|>[[110, 475, 610,
     524]]<|/det|>
Studiengang: IM WIN BWL Kunst
☒ ☐ ☐ ☐
```

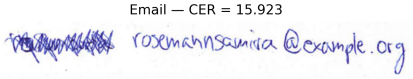
(b)

Fig. S5 DeepSeek's output for the *studies* field, with the two symbols it used for a marked checkbox: **a** form C-14, using a symbol that does not occur in the validation data, so that the extractor does not recover the selected option (Fig. 5 of the article); **b** form C-86, using a symbol that does occur in the validation data, so that the extractor recovers the selected option



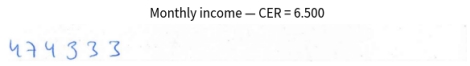

```
GT:   Ich hätte gerne eine Kopie des Vertrages
Pred: E. G. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4. 4.
```

(a)




```
GT:   rosemannsamira@example.org
Pred: [emailprotected][emailprotected][emailprotected][ema
```

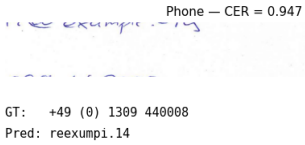
(b)



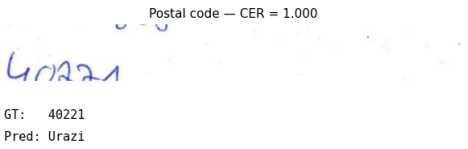

```
GT:   474333
Pred: ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¥ ¯
```

(c)

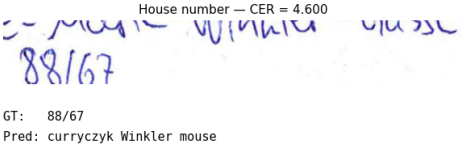
Fig. S6 Unexpected outputs by DeepSeek, which are recognition failures rather than extraction failures: **a** form A-78, *comment* field, a degenerate repetition (Fig. 7 of the article); **b** form C-12, *email* field, where the prediction repeats the string ** many times; **c** form C-81, *monthly income* field, a sequence of non-Latin characters



(a)

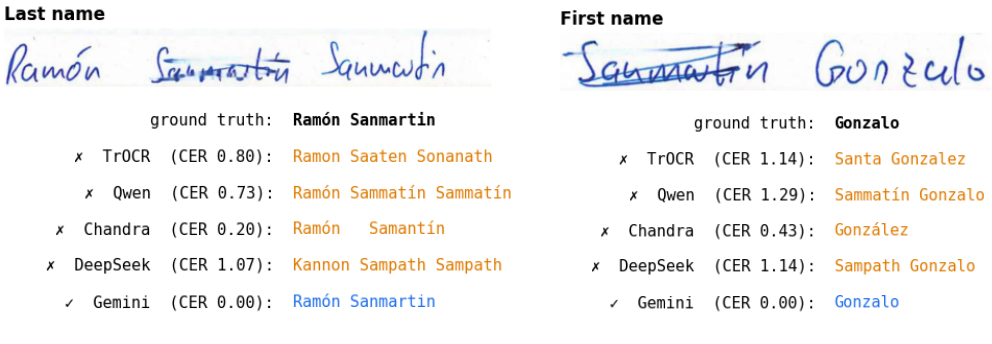


(b)



(c)

Fig. S7 Fields where TrOCR's localisation failed because of layout C's unconstrained fields: **a** form C-36, *phone* field (Fig. 8 of the article); **b** form C-3, *postal code* field; **c** form C-12, *house number* field



(a) **(b)**

Fig. S8 Struck-out text in form A-90, with the transcription of every model: **a** *last name* field (Fig. 9 of the article); **b** *first name* field

S5 Prompts

All prompts are reproduced exactly as passed to the model.

S5.1 Qwen3-VL-8B-Instruct

The generalist model receives a system prompt fixing its role and output format, and a user prompt specifying the field schema.
System prompt

```
You are a transcription system for handwritten German administrative forms.
You return JSON only.
```

User prompt

```
Transcribe the handwritten entries in this form. Fields (JSON key -> printed
label on the form):

  "last_name"      -> Feld "Name": handwritten entry, transcribed character by character
  "first_name"     -> Feld "Vorname": handwritten entry, transcribed character by character
  "monthly_salary" -> Feld "Monatliches Einkommen": handwritten entry, transcribed character by
  ↪ character
  "birthday"       -> Feld "Geburtsdatum": handwritten entry, transcribed character by character
  "email"          -> Feld "E-Mail-Adresse": handwritten entry, transcribed character by character
  "phone"          -> Feld "Telefonnummer": handwritten entry, transcribed character by character
  "studies"        -> Feld "Studiengang": checked option, exactly one of these values: IM | BWL |
  ↪ WIN | Kunst
  "street"         -> Feld "Straße": handwritten entry, transcribed character by character
  "house_number"   -> Feld "Hausnummer": handwritten entry, transcribed character by character
  "postal_code"    -> Feld "PLZ": handwritten entry, transcribed character by character
  "city"           -> Feld "Ort": handwritten entry, transcribed character by character
  "comment"        -> Feld "Kommentar": entire handwritten text; join multiple lines with " "

Rules:
1. Return exactly these keys, in this order: "last_name", "first_name",
   "monthly_salary", "birthday", "email", "phone", "studies", "street",
   "house_number", "postal_code", "city", "comment"
2. ALL values are strings in double quotes - including pure numbers. Write
   "01234", not 01234 and not 1234. Leading zeros are preserved.
3. Transcribe character by character. Correct NOTHING: no spelling, no date or
   number formats, no umlauts, no names, no place names. Reproduce spelling and
   punctuation exactly as written.
4. Transcribe only the HANDWRITING. Printed labels and the form number in the
   top right corner are not part of it.
5. If a field is empty or illegible, return "". Never guess.
6. Output only the JSON object. No prose, no explanation, no markdown code blocks.
```

S5.2 chandra-ocr-2

Chandra is used with one of its own predefined prompts, `ocr`, reproduced here as shipped with the model rather than written for this work.

```
OCR this image to HTML. Only use these tags [’math’, ’br’, ’i’, ’b’, ’u’,
’del’, ’sup’, ’sub’, ’table’, ’tr’, ’td’, ’p’, ’th’, ’div’, ’pre’, ’h1’, ’h2’,
’h3’, ’h4’, ’h5’, ’ul’, ’ol’, ’li’, ’input’, ’a’, ’span’, ’img’, ’hr’, ’tbody’,
’small’, ’caption’, ’strong’, ’thead’, ’big’, ’code’, ’chem’], and these
attributes [’class’, ’colspan’, ’rowspan’, ’display’, ’checked’, ’type’,
’border’, ’value’, ’style’, ’href’, ’alt’, ’align’, ’data-bbox’, ’data-label’].

Guidelines:
```

```
* Inline math: Surround math with <math>...</math> tags. Math expressions
  should be rendered in KaTeX-compatible LaTeX. Use display for block math.
* Tables: Use colspan and rowspan attributes to match table structure.
* Formatting: Maintain consistent formatting with the image, including spacing,
  indentation, subscripts/superscripts, and special characters.
* Images: Include a description of any images in the alt attribute of an <img>
  tag. Do not fill out the src property. Describe in detail inside the div tag.
  Also convert charts to high fidelity data, and convert diagrams to mermaid.
* Forms: Mark checkboxes and radio buttons properly.
* Text: join lines together properly into paragraphs using <p>...</p> tags. Use
  <br> tags for line breaks within paragraphs, but only when absolutely
  necessary to maintain meaning.
* Chemistry: Use <chem>...</chem> tags for chemical formulas with reactive SMILES.
* Lists: Preserve indents and proper list markers.
* Use the simplest possible HTML structure that accurately represents the
  content of the block.
* Make sure the text is accurate and easy for a human to read and interpret.
  Reading order should be correct and natural.
```

S5.3 DeepSeek-OCR-2

DeepSeek is used with its predefined `markdown` prompt in its grounded variant, so that the output carries bounding boxes alongside the transcription.

```
<image>
<|grounding|>Convert the document to markdown.
```