Information Extraction from Unlabeled Archival Index Cards with an Ensemble of Vision–Language Models

– Technical Report –

Arthur Matei[1,2], Benjamin Kram[3], Tobias Herrmann[3], and Gernot A. Fink[1,2]

[1] *TU Dortmund University, Department of Computer Science, Dortmund, Germany*
[2] *LAMARR, Institute of Machine Learning and Artificial Intelligence, Dortmund, Germany*
[3] *Landesarchiv Nordrhein–Westfalen, Duisburg/Münster, Germany*

September 16, 2026

Abstract

The Landesarchiv NRW holds 486,575 scanned index cards from Nazi-era restitution proceedings (*Wiedergutmachung*), digitised but not searchable: only 620 of them (0.13 %) are annotated, too few to train on. We extract four fields per card — Family Name, First Name, Birth Date and Reference Number — with an ensemble of five vision–language models used off the shelf and steered by prompt alone, chosen because each processes the card image differently, so their mistakes are less likely to coincide. Their outputs are merged field by field by rules measured on ground truth rather than a plain majority vote, and every published value carries a count of how many models agreed on it. Validated against the 620 annotated cards, that count orders values by trustworthiness: depending on the field, values on which all five models agree are correct between 97 and 99 % of the time, values on which no two agree at most 20 %. The count is a reliability signal for prioritising review, not a calibrated probability.

1 Introduction

Archives in Germany are legally mandated not only to secure and preserve the records in their custody, but in particular to make them available for use and to publish them (§§ 5(2) and 6(1) in conjunction with §§ 2(7) and 8 of the Archives Act of North Rhine-Westphalia, ArchivG NRW).[1] To enable use independent of time and place by a large, global user base, the Landesarchiv NRW (LAV NRW) digitises a substantial volume of records every year and publishes them online. Digitisation, however, makes a record viewable rather than findable. The rapid development of AI-supported infrastructure now makes it possible to close that gap: even very large holdings that have never been *described*, i.e. recorded item by item in a finding aid with the data that make each record searchable, can be read automatically and thereby made usable in the first place.

This applies above all to the records on redress for National Socialist injustice (*Wiedergutmachung*) in Germany and North Rhine-Westphalia. They date predominantly from the immediate post-war period, are heterogeneous, and are in large part not yet, or not sufficiently, described. Of particular importance are the index card holdings studied in this report, created in large numbers as finding aids in restitution and compensation proceedings. A card may record information

[1] `https://recht.nrw.de/lrgv/gesetz/04082026-archivgesetz-nordrhein-westfalen-archivg-nrw/`, last accessed 2026-09-15.

on persons, claims, assets, proceedings and administrative decisions, and it does so densely and at the level of the individual person: a handful of fields per card — family name, first name, birth date and reference number — is enough to make an entire holding searchable by name and to lead a user to the case file behind each card. Describing a volume such as the two holdings selected here, 486,575 cards in total, by hand would take several years depending on the number of staff assigned, and would come at the expense of other pressing archival tasks. Not least because of the sheer number of cards, evaluating the information in these sources, and in the as yet undescribed case files they refer to, is realistic in the near future only through AI-supported automated processing.

Running a vision–language model over such a collection is now straightforward. Knowing how far to trust its output is not. Without ground truth there is no estimate of correctness, and a model's answer reads equally fluent whether it is right or wrong. Even a small test set is costly to create, and the supervised approaches common in work on historical records [1, 2] would require a considerably larger annotated training set on top. For this corpus the LAV NRW annotated 620 cards, 0.13 % of the collection: enough to measure against, but not to train on. The problem this report addresses is therefore extraction that comes with a usable estimate of its own correctness, obtained without any training.

The work is part of a proof of concept for an automated process that takes index cards from digitisation through extraction to import into the archive's information system (*archivisches Fachinformationssystem*, AFIS). The LAV NRW's system, VERA (*Verwaltungs-, Erschließungs- und Recherchesystem für Archive*),[2] imports spreadsheets whose columns are mapped onto its VERA-XML attributes; the spreadsheets produced here contained all mandatory data — the four fields above — and were successfully imported, split across several finding aids (*Findbücher*), with negligible preparation. From VERA the data can be exported as EAD (Encoded Archival Description)[3] for the archive portal, so VERA is the only hurdle the data have to clear. Crucially, the import does not abort on erroneous values, and archivists can correct them before or after import. Imperfect extraction is therefore no obstacle to publication. What the LAV NRW needs is to know *which* values to check.

This report provides that, with the following contributions:

- We apply five vision–language models to the full corpus without fine-tuning, steered by prompt alone (Section 4). They differ in how they see the page, i.e. in the vision encoder that turns the card image into visual tokens, in the resolution at which the card is read, and in most cases in the language decoder that turns those tokens into text; one of them does not extract from the image at all but first transcribes the card and then parses the transcript.

- We merge their outputs field by field, following rules measured on ground truth rather than a plain majority count: when a blank counts as a vote, how names that differ only by a maiden name are matched, and which model to trust when no two models agree (Section 4.3).

- We ship an agreement count with every value and validate it against the annotated cards (Section 5.5): depending on the field, values on which all five models agree are correct between 97 and 99 % of the time, values on which no two agree at most 20 %. The count orders values by trustworthiness but is not a calibrated probability, since models can concur on a shared error [3].

- We score the family name in two parts — the base surname and any maiden or former name — so that a model which identifies the person but drops a maiden name is not counted as having misread the card (Section 5.3).

[2] `https://www.startext.de/projekte/v-e-r-a`, last accessed 2026-09-15.
[3] `https://www.loc.gov/ead/`, last accessed 2026-09-15.

Section 2 places the work among prior approaches, Section 3 describes the card collection, Section 4 the models and the consensus, and Section 5 their evaluation. Section 6 discusses what the agreement count is worth and Section 7 concludes.

2 Related Work

2.1 From layout-aware encoders to OCR-free extraction

Key information extraction from documents was long approached as a pipeline over recognised text. Layout-aware pre-trained encoders such as LayoutLM [4] made that pipeline substantially stronger by fusing token content with its position on the page, and a large body of work has since extended the idea across modalities and architectures [5]. All of these models, however, consume the output of an upstream OCR engine, and therefore inherit its errors.

Donut [6] broke that dependency by mapping the document image directly to a structured string, and Pix2Struct [7] showed that the same image-to-text formulation could be pre-trained at scale from web screenshots. The appeal for archival material is obvious: no OCR engine has to be tuned for a script it was never trained on. The cost is that the extraction schema is baked into a fine-tuned decoder, so a new field set means new training data.

2.2 Recognition and extraction on historical records

Historical documents put pressure on the pipeline assumption from a different direction. Because handwriting recognition on degraded material is itself error-prone, errors propagate into any downstream tagger, which motivated models that recognise text and label entities jointly [1, 8]. Applied at collection scale, such systems can turn whole series of registers into searchable databases [2]. These approaches are effective but supervised: they presuppose transcribed and labelled training material in the hand, language and layout of the target collection, which is precisely the resource an unindexed collection lacks.

2.3 Prompted multimodal models for archival extraction

Instruction-following vision–language models change that calculus, because the schema moves from the training set into the prompt. Recent work shows the practical consequences on historical material: multimodal LLMs outperform conventional OCR on nineteenth-century German city directories and can be used to post-correct their own transcriptions to below one percent character error rate [9]. On German index cards specifically — the document type studied here — Vafaie et al. [10] evaluate a range of multimodal LLMs end to end and report that prompt design and inference settings carry much of the performance, an observation our own experiments reproduce (Section 5.2).

Our own prior work sits in this line. The CM1 dataset [11] was built to measure few-shot information extraction with large vision–language models on archival forms, and a recent study additionally examines the *implicit normalisation* that fine-tuning introduces: a model trained towards a normalised target learns to silently regularise dates, name forms and identifiers rather than reproduce what the page says [12]. In other words, fine-tuning teaches a model the conventions of its target. This collection offers no training material, so the present work cannot rely on that. Conventions such as which parts of a name to include are instead stated in the prompt and enforced by postprocessing rules written by hand (Sections 5.2 and 5.3). Formatting variation that neither resolves carries over into the output (Section 6).

2.4 Agreement as a reliability signal

Where no ground truth exists, consistency is the obvious surrogate for confidence. Self-consistency [13] established the pattern for a single model: sample repeatedly, keep the answer that recurs. Test-

time augmentation varies the input instead of the sampling. Archibald and Martinez [14] apply it to extraction from historical records with a multimodal LLM: several augmented versions of each image are transcribed, and the readings are aligned into a consensus whose agreement also serves as a confidence score. Extending the idea across *different* models is attractive here, because independently trained systems should not share all of their failure modes.

Agreement, however, is not accuracy. Models can concur through shared bias rather than shared correctness, and confident errors recur systematically across prompts, runs and even vendors [3]. This is the reason the count shipped with our output is presented as a signal for prioritising review and never as a calibrated probability, and the reason we measure what each agreement level is actually worth against annotated data (Section 5.5) rather than assuming unanimity implies correctness.

2.5 Positioning

Relative to the work above, this report differs in three ways. First, the target is a full production collection of 486,575 cards rather than a benchmark split, so throughput and failure modes at scale are part of the result. Second, no model is fine-tuned, because no training material exists; the conventions a fine-tuned model would learn are instead encoded in the prompt and in a small set of postprocessing rules. Third, agreement between *heterogeneous* vision–language models serves two purposes: it produces the published value, which on the name fields is more accurate than any single model, and it attaches to every value a count that allows the review of an unlabelled corpus to be prioritised. Both are quantified per field against manually annotated ground truth, with the components the models share stated rather than assumed away.

3 The LAV Findmittel Card Collection

The description of the two card-index collections, BR 3012 and BR 3014, is being carried out as part of the nationwide project *Transformation der Wiedergutmachung* (Transformation of Compensation), funded by the Federal Ministry of Finance. Under this project, funds formerly used for the payment of compensation pensions are being redirected to support the preservation and accessibility of archival records documenting post-1945 West German compensation practices. In particular, the project aims to establish an online portal through which compensation files created under the Federal Compensation Act (Bundesentschädigungsgesetz) will be digitized, comprehensively described, and made accessible to the general public.

For the Landesarchiv Nordrhein-Westfalen, the index cards themselves constitute the key to these sources. The two largest collections of compensation files in the Rhineland Department, BR 3001 and BR 3002, have not yet been arranged and described. This is problematic for several reasons. First, they comprise approximately 387,700 individual case files in total, making them by far the largest group of individual case files held by any German archive in purely quantitative terms. The corresponding card-index collections comprise approximately 193,000 index cards for BR 3012 and approximately 294,000 index cards for BR 3014. Moreover, these files are increasingly being used not only by researchers but also, and on a very substantial scale, by descendants as evidence in nationality and citizenship proceedings under the new German citizenship law.

The two collections are also distinctive at the national level in that they relate to special responsibilities assigned to the state of North Rhine-Westphalia for supra-regional groups of persons persecuted under National Socialism, as well as for expellees from Eastern Europe, stateless persons, and refugees under the Geneva Convention. The Rhineland Department is therefore confronted with an almost unmanageable volume of inquiries from all over the world, most of which require access to precisely those two collections that have so far received the least archival description. Since the files have not yet been arranged and described, each inquiry

currently requires laborious research using the so-called Bundeszentralkartei (Federal Central Index of Compensation Cases).

The description of these index cards therefore provides the key to making the files accessible, and the cards are particularly well suited to this task. They are former internal finding aids of the competent authority, the Bezirksregierung Köln, and describe the files at item level. If these finding aids can be described, the archival arrangement and description of the files themselves will require little more than reconciling the holdings actually present with the information recorded on the index cards.

Describing the index cards manually is practically infeasible given the large volume of records: approximately 193,000 index cards for BR 3012 and approximately 294,000 index cards for BR 3014. Moreover, research would remain dependent on predefined indexing criteria and on the quality of human indexing, making implicit connections or unexpected links between individual entries difficult to identify. Handwritten and heterogeneously structured index cards also pose particular challenges for consistent and standardized description. By contrast, AI-assisted analysis can efficiently process large volumes of serial records, thereby opening up additional possibilities for research and analysis.

4 Method

4.1 Task definition

Each card is processed independently and yields a single JSON object with four fields: Family Name (JSON key `nn`), First Name (`vn`), Birth Date (`gd`) and Reference Number (`nr`) — the last being the archival identifier (*Aktennummer* / *Verzeichnungsnummer*) written on the card. These are the fields the LAV NRW's finding aids are keyed on.

Values are copied *as printed*: dates are not reformatted, abbreviations not expanded, misspellings not corrected. The deliverable is a transcription of the record, and a reference number that has been silently tidied is no longer the identifier on the card. Formatting variation therefore carries over into the output, where it stays visible (Section 6).

No model is fine-tuned, because there is no labelled training material for this collection. The models are used off the shelf and steered by prompt alone.

4.2 Five models

The consensus is only informative if the models fail differently, because agreement can only separate right from wrong readings if a wrong reading is unlikely to be repeated by the other models. Two checkpoints of one family would largely agree on their shared mistakes, and the agreement count would measure fluency rather than correctness. The five models were therefore chosen to differ in how they see the page, not only in their weights:

Qwen3.5-4B [15, 16] A promptable vision–language model whose visual token budget is tunable through `max_pixels`.

Gemma-3-4b [17] A fixed 896×896 SigLIP [18] vision tower producing 256 image tokens; its only resolution lever is pan-and-scan tiling.

GLM-OCR [19] + Qwen3.5-4B A two-stage composite, the *GLM-Qwen composite*: a CogViT encoder over a 0.5 B language decoder transcribes the card, and a text-only LLM parses that transcript into the schema. GLM-OCR's own extraction mode was tried first but confused the First Name with the birthplace or a death date, and every added clarifying instruction made it worse — a sign of limited decoder capacity rather than of prompt wording. Split into reading and parsing, the First Name rose to 87.1 % (Table 2), and every error can be attributed to one of the two stages.

Table 1: Inference settings and cost of the full-corpus runs (486,575 cards, one GPU per run, `max_new_tokens` = 96). The resolution settings were chosen on a 120-card tuning subset (Section 5.2).

Model	Image Input	Time/Card (s)	Wall Clock (h)
Qwen3.5-4B	`max_pixels` $= 768 \cdot 28^2$	0.66	89.4
Gemma-3-4b	fixed 896×896, no tiling	0.52	70.3
GLM-OCR + Qwen3.5-4B	native (two stages)	2.41	325.6
InternVL3.5-4B	up to 4 tiles of 448×448	0.78	106.0
Ministral-3-8B	longest edge 896 px	1.14	153.9

InternVL3.5-4B [20] An InternViT encoder over a Qwen3-4B language model. The image is cut into 448×448 tiles, and the number of tiles is its resolution lever.

Ministral-3-8B [21] A Mistral language model with a Pixtral-style vision encoder that reads the image at its native aspect ratio, resized to a configurable longest edge.

The models differ in effective resolution, in vision–language coupling, and — for the GLM-Qwen composite — in whether the extraction sees the image at all, so a transcription error and a parsing error do not occur on the same cards. The variety is nonetheless smaller than the list suggests. **Three of the five share a Qwen language model**: Qwen3.5-4B itself, the parser of the GLM-Qwen composite, and the language model inside InternVL3.5. Their vision encoders differ, so a misread character is not necessarily shared, but habits that live in the language model can be (Section 5.5). Ministral-3-8B is the only model with no Qwen, InternViT or SigLIP component. Table 1 lists the inference settings and the cost of each corpus run.

4.3 Consensus construction

The outputs are merged into one record per card, carrying for each field a published value and a count of how many models agreed on it. Four rules govern the merge, plus a fallback for cards where no two models agree; each was measured rather than assumed.

Agreement is counted per field, never per record. Only 21.3 % of annotated cards are unanimous on all four fields, because a single weak field drags down the whole record, yet the Birth Date alone is unanimous on 77.7 % of them and correct on 97.1 % of those. A record-level flag would discard that date because the card's reference number was contested, so a card may ship with a trusted birth date and a flagged family name (Section 5.5).

An abstention is not a dissent. A model that returns nothing for a field has declined to vote, and abstention is strongly model-specific: the GLM-Qwen composite omits the Birth Date on 11.8 % of annotated cards, InternVL3.5 the First Name on 10.2 % (Table 2). Counting blanks as disagreement would punish exactly those fields. Agreement is therefore counted over the models that answered, as `agree/n_answered`, which keeps "4 of 4" distinct from "4 of 5" — a silent fifth model is weaker evidence against the four than a contradicting one (Section 5.5).

Maiden names are excluded from the match but kept in the value. 151 of the 620 annotated cards carry a *geb./verw./gesch.* marker, and the models differ sharply in whether they reproduce it (Section 5.5). Most apparent Family Name disagreement is therefore one model reading *less* of the field — `Abeles` against `Abeles geb. Gruenberg` — rather than a conflict about who the person is. Agreement is decided on the base surname, and the published value is the longest variant in the winning group, so a maiden name that only one model transcribed still reaches the output. For evaluation, the Family Name is additionally split at its first marker

into a *base surname* and a *former name* (`Abonyi geb. Grosz` → `Abonyi`, `geb. Grosz`), in model output and annotation alike: the base surname answers whether the models identified the person, the former name whether they read the rest of the field.

The same postprocessing precedes every comparison. All models' outputs pass through the same postprocessing (Section 5.3) before they are compared, so formatting differences are not counted as disagreement.

When no two models agree, a last resort. On 1.3–3.8 % of cards per field (Table 4) no two models give the same value, including cards on which only one model answered. Rather than dropping the field, the consensus then publishes the reading of the model that is most accurate on that field in Table 2: the GLM-Qwen composite for the Reference Number and Family Name, Qwen for the Birth Date and First Name. If that model gave no answer, the next most accurate one is used. The field is emitted with an agreement count of one and a `no_majority` flag. The "No Two Agree" columns of Tables 3 and 5 score this reading; since the ranking comes from the same annotated cards and only 6–22 of them per field fall into this level, those figures are rough.

Matching rules

Agreement is decided per field by a matcher that tolerates variation known to be cosmetic and nothing beyond it. A reference number is a number of up to six digits, written with varying separators and sometimes preceded by a letter prefix or followed by a letter suffix (`123 456`, `123.456`, `II/47 419`, `A 291 732`, `740 296 a`). Separators are ignored, so `123 456` and `123.456` agree, but prefixes and suffixes are kept, because `II/47 419` and `47 419`, or `740 296 a` and `740 296 b`, refer to different files. Ignoring the prefix was tested and made more cards unanimous while making the unanimous values less often correct. Birth Dates written with a two-digit year agree with the same date written in full (`15.11.09` and `15.11.1909`), and Roman month numerals agree with their number (`5.III.1911` and `5.3.1911`); a missing century is never filled in. Family Names agree when their base surnames are equal, ignoring case, punctuation, whitespace and accents such as umlauts (*Glück* and *Gluck* agree).

These rules decide *agreement only*. The published value is the model's reading after the light cleanup of Section 5.3 and is never folded, reformatted or corrected beyond that. Dissenting readings are carried alongside it, so a reviewer can see the alternatives on any contested field.

5 Experiments

5.1 Evaluation data

The LAV NRW annotated 620 cards, split evenly between the two collections. The cards of BR 3014 name the claimant (*Anspruchsberechtigter*) at the top and have a second person block further down for the persecuted person (*Verfolgter*), which is filled when the two are not the same person. Only the claimant is extracted, so a model that reads the lower block returns the wrong person. 176 of the annotated cards (28.4 %) have this second block filled, so this possible failure is well represented.

620 cards is 0.13 % of the corpus, and measuring on so few cards leaves some uncertainty. Had a different set of 620 cards been annotated, a measured accuracy could just as well have come out 2 to 4 points higher or lower, depending on how high it is (95 % confidence interval). Two models whose accuracy differs by only a point or two can therefore not be ranked, and this report only draws conclusions from larger differences. The sample is also not random: the annotated cards come from a few contiguous letter ranges, so the frequency of compound and maiden names need not match the corpus. The numbers are exact for this set and indicative for the corpus.

Configuration decisions were tuned on a 120-card subset of the same annotations, so these cards are also part of the evaluation, which may make the reported figures slightly optimistic.

5.2 Preprocessing

Decisions made before the model runs — the output budget, the image resolution and above all the wording of the prompt — moved accuracy further than the choice between models did.

Output budget. Across 720 predictions on the 120 tuning cards (six resolution settings), the longest was 50 tokens. All corpus runs used `max_new_tokens` = 96 (for the GLM-Qwen composite, the parser stage), which leaves headroom while bounding the cost of a runaway repetition loop.

Input resolution. Each model exposes a different image lever, set on the tuning subset. Qwen's visual token budget was set to `max_pixels` $= 768 \cdot 28^2$, about 760 visual tokens. Gemma's pan-and-scan tiling was left disabled; since a landscape card squeezed into a square tower is distorted, this is a plausible contributor to Gemma's weak surname reading and worth revisiting. InternVL3.5 shows how much the lever can matter: its checkpoint ships with tiling disabled, which squeezes the card into one 448×448 tile and read the family name on 25 % and the first name on 42 % of tuning cards. Up to four tiles raised these to 51 and 76 %; twelve were not better and doubled the time per card. Ministral-3 has the opposite problem: its default longest edge of 1540 px yields about 3300 input tokens. Capping it at 896 px with a batch size of four cut its time from 2.48 to 1.15 s per card, with all field changes within the noise of 120 cards.

The prompt is preprocessing. Two properties of the annotation had to be stated in the prompt, and both were first mistaken for model errors. The Reference Number is preceded on the form by a printed reference (`V-56-II-Nr.`, `Art. V-56-II-Nr.`) which the models transcribed and the annotation excludes; naming its variants in the prompt was worth **+6.6 points on Gemma's Reference Number** at no cost to Qwen, so the end-to-end models share one prompt. And the Family Name field holds the entire surname, including Spanish compounds (*Abad Edo*) and maiden names (*Abonyi geb. Grosz*), whereas the models returned only the first token. Stating this in the prompt raised Qwen's family name accuracy from 55.0 to 70.8 % on the tuning cards. An accuracy number on archival data is thus a statement about the agreement between two conventions, not about recognition alone.

5.3 Postprocessing

Postprocessing is deliberately thin, since every rule that rewrites a value can also destroy one. It is applied identically to all five models.

Name spacing. The OCR stage splits letters on typewritten cards (*A a s* for *Aas*, *Abad - Segura* for *Abad-Segura*). A repair step rejoins runs of single letters without merging genuine multi-token names.

Surname overruns. Some models read the surname correctly and keep reading into the neighbouring text: the birthplace (*Aberfeld geb. in Rumänien*), the birth-date line, the next form label, the first name, or a marker with nothing after it (*Abad Sanchez verw.*). Ministral-3 does this most. These tails are cut. Scored against the annotation, no cut turns a correct surname into a wrong one for any model; they repair 24 of Ministral-3's surnames and at most three of any other model's. An instruction in the prompt barely helped, and a further rule — stripping a trailing `/ <token>` — was rejected because on these cards it is usually a name variant (*Jaacob/Jancu*).

Base surname and former name. After the overrun cut, the family name is split at its first marker (*geb.*, *verw.*, *gesch.*, *vorm.*, *verh.*) into a base surname and a former name, which keeps its marker. The annotation is split with the same function but not trimmed, and the whole family name is kept next to the two parts.

Reference Numbers and dates. Only a complete preprinted reference is stripped from the reference number, and separator spacing is unified. A bare leading `56` or an `II/A` prefix is left alone, because the annotation is itself inconsistent about them (Appendix A). Roman month numerals in the middle position of a date are resolved to Arabic; a leading Roman numeral is usually not a month (*II. Febr. 1919* is a misread day 11). Typewriter confusions (`o`→0, `l`→1) inside digit runs are folded for scoring only, so that `747 8o7` matches `747 807`; the published value is never rewritten this way.

How names are scored. The whole family name and the first name are compared casefolded with accents kept, since the LAV NRW treats *Fäe* and *Fae* as different strings. For the family name, whitespace is also ignored, because letter spacing splits words (*A bel*) and the annotation itself varies (*geb.Kalman*, *geb. Kalman*). This turns 118 predictions from wrong to right, and each of them is the annotated name with different spacing, never a different name. The base surname and the former name are compared on the agreement key — casefolded, without punctuation, whitespace or accents — so that for them agreement and correctness measure the same thing: whether the models identified the name. The former name is scored on the 151 cards whose annotation has one, with a separate rate for former names reported on the other 469.

5.4 Single-model baselines and the vote

Table 2 gives per-field accuracy on the 620 annotated cards. **No model wins every field**: the GLM-Qwen composite leads on the Reference Number and the Family Name, Qwen on the Birth Date and First Name. Ministral-3 leads on nothing but is within about three points of the leader on every field, inside the noise. This complementarity is the premise of the consensus.

The last row scores the value the five-model vote publishes as if it were one more model. On the names the vote is clearly better than the best single model: 91.6 against 86.0 % on the base surname and 93.4 against 89.4 % on the First Name. Against the best model per field, the

Table 2: Per-field accuracy on the 620 annotated cards (%). Best single model per column in bold. *Five-Model Vote* scores the published consensus value. *Family Name* is scored on the whole field. *Base Surname* and *Former Name* split it at its first maiden/former-name marker and are compared ignoring case, punctuation, whitespace and accents. *Former Name* is scored on the 151 cards whose annotation has one; *Former Added* is the share of the other 469 on which the model reports one (lower is better). *All Four* uses the whole Family Name. *Empty* is the share of cards on which the field is left empty (the annotation has no Birth Date on 3.1 % of cards and always has a First Name); on the other fields no model leaves more than 1.1 % empty.

									Empty	
Model	Reference Number	Family Name	Base Surname	Former Name	Former Added	Birth Date	First Name	All Four	Birth Date	First Name
Qwen3.5-4B	80.2	72.1	83.1	43.0	**0.9**	**92.9**	**89.4**	50.5	2.7	0.2
Gemma-3-4b	82.9	62.1	72.1	41.7	1.3	87.7	75.0	39.2	0.8	4.5
GLM-OCR + Qwen3.5-4B	**85.6**	**80.5**	**86.0**	**83.4**	1.1	87.1	87.1	**60.2**	11.8	4.7
InternVL3.5-4B	77.1	61.0	71.8	70.9	1.3	91.3	74.7	38.4	3.1	10.2
Ministral-3-8B	84.5	77.4	85.3	82.1	3.8	91.0	87.3	57.1	2.4	3.1
Five-Model Vote	84.8	80.8	91.6	89.4	4.9	91.8	93.4	62.1	0.5	0.0

vote fixes 42 base surnames and breaks 7, and fixes 27 first names and breaks 2 (both McNemar $p < 0.001$). On the Birth Date and Reference Number it is about a point below the best single model, within noise, and on the whole family name string and all four fields it is on par (80.8 and 62.1 %). The comparison favours the single models, whose best was picked on the same cards. The longest-variant rule cuts both ways: it recovers more former names than any single model (89.4 %) but also adds one to 4.9 % of cards whose annotation has none.

Splitting the family name shows where its errors come from. Every model identifies the base surname 5.5–11 points more often than it gets the whole field right, and the difference is the former name: Qwen and Gemma reproduce it on only 43 and 42 % of the cards that carry one, against 82–83 % for the GLM-Qwen composite and Ministral-3. Ministral-3 in turn reports a former name on 3.8 % of cards whose annotation has none. All four fields are right on at most 60.2 % of cards for any model, which is why agreement is tracked per field rather than per record.

5.5 Agreement against correctness

Table 3 is the main result: what the agreement count is worth. With the family name judged on the base surname, unanimity of all five models corresponds to between 96.8 % (Reference Number) and 98.6 % (base surname) correctness, while cards where no two models agree fall to at most 20 % (First Name), on only 6 to 22 cards per field. The ordering is monotonic on every field, which makes the count usable for prioritising review although it is not a calibrated probability. Unanimity is not common: the base surname is unanimous on 47.7 % of annotated cards, all four fields together on 21.3 % (21.5 % over the corpus). The whole family name string is right 88.5 % of the time at unanimity (Section 5.6). The former name is the one row that is not monotonic and rests on few cards. Figure 1 shows coverage and correctness together: the wrong cards concentrate in the thin low-agreement slices.

Occurrence and correctness are measured on different populations. Agreement needs no labels, so occurrence is exact over all 486,575 cards, while correctness is a 620-card estimate. Table 4 compares the two occurrences as a check on representativeness. On the Family Name and Birth Date they agree within 1.0 point. The annotated cards reach unanimity on the Reference Number 2.3 points less often than the corpus, which makes its figures conservative, and on the First Name 3.1 points more often, which makes them slightly optimistic. The former name is the least representative row.

Table 3: Agreement level against correctness on the 620 annotated cards (%). *Occurs* is the share of cards at that agreement level; *Correct* is accuracy within that subset. The level is the size of the largest group of models giving the same reading; abstaining models are not counted. Where two groups tie for largest (e.g. 2–2–1), the published value comes from the group containing the earlier model in the order Qwen, Gemma, GLM-Qwen composite, InternVL, Ministral. Where no two agree, it is the reading of the field's most accurate model that answered (Section 4.3). *Base Surname* and *Former Name* are grouped and scored ignoring case, punctuation, whitespace and accents; former-name shares are over the cards on which at least one model gives a former name (27.6 % of annotated cards).

	5 of 5		4 of 5		3 of 5		2 of 5		No Two Agree	
Field	Occurs	Correct	Occurs	Correct	Occurs	Correct	Occurs	Correct	Occurs	Correct
Family Name	47.7	88.5	29.2	83.4	14.8	73.9	5.6	51.4	2.6	12.5
Base Surname	47.7	98.6	29.2	95.0	14.8	87.0	5.6	62.9	2.6	12.5
Former Name	16.4	85.7	31.6	92.6	26.9	91.3	9.9	88.2	15.2	15.4
First Name	56.1	97.7	26.3	95.7	11.3	85.7	4.7	72.4	1.6	20.0
Birth Date	77.7	97.1	13.4	94.0	2.7	70.6	2.1	53.8	3.5	4.5
Reference Number	74.7	96.8	6.3	61.5	10.3	56.2	7.7	37.5	1.0	0.0

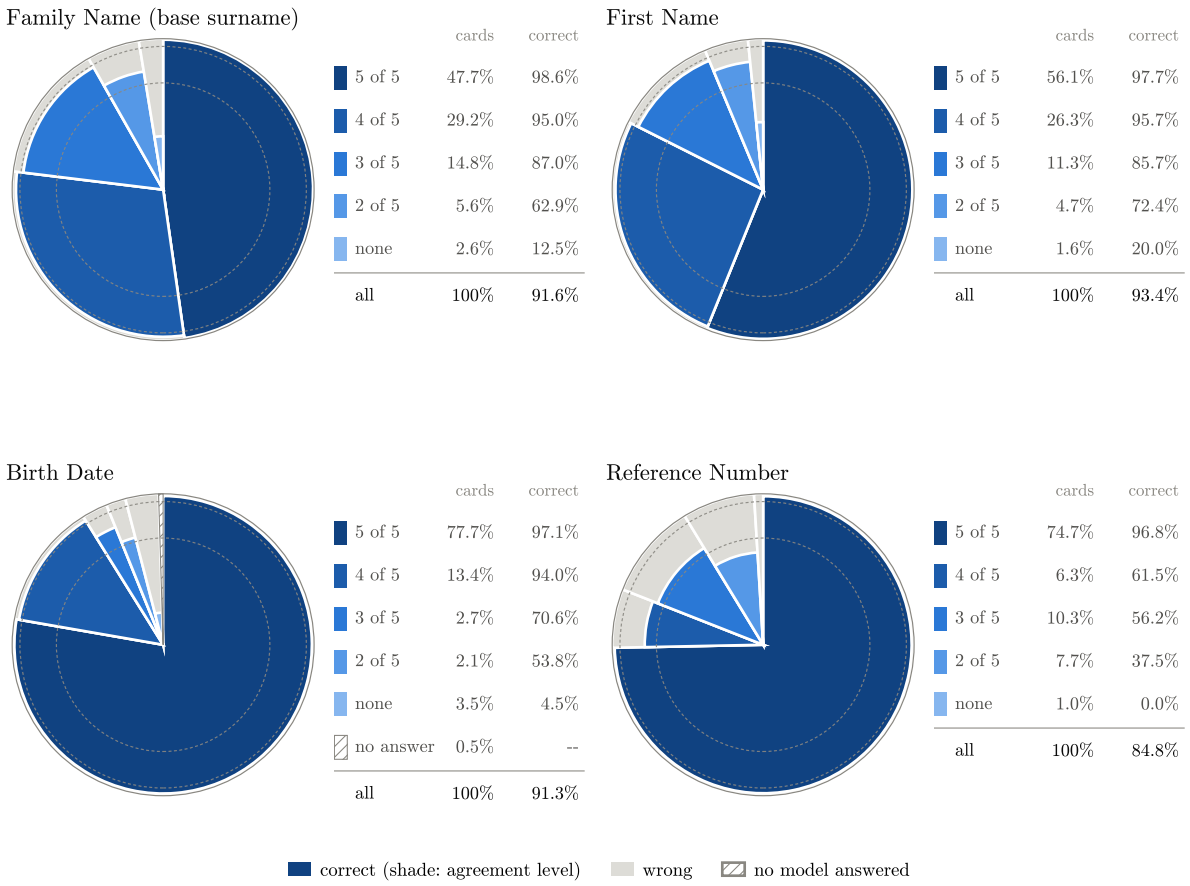

Figure 1: Coverage and correctness per agreement level on the 620 annotated cards, one pie per field. Each slice is an agreement level, starting at 12 o'clock with 5 of 5 and running clockwise; its *angle* is the share of cards at that level. The coloured part of a slice reaches the square root of the level's accuracy, so its *area* is the share of cards that are correct, and the grey part out to the rim the share that is wrong. The dashed rings mark 50 % and 90 % accuracy. The Family Name is judged on the base surname. Numbers as in Table 3.

The intermediate levels are field-specific. Four of five agreeing is almost as good as unanimity on the base surname, the First Name and the Birth Date (94.0–95.7 %), but only 61.5 % on the Reference Number. A single "4 of 5" reliability figure across fields would conflate a good bet with a poor one. Whether the fifth model abstains or contradicts can matter as well: on the Birth Date, four agreeing models were right on all 41 annotated cards on which the fifth was silent, against 88.1 % of the 42 on which it gave a different date. On the First Name the two cases do not differ.

Two model-specific habits show why a plain majority vote misreports. *Prefix conventions*: Qwen drops the `II/A` prefix of the reference number on 78 of the 88 annotated cards carrying one and InternVL3.5, whose language model is also a Qwen, on 80, while Gemma, the GLM-Qwen composite and Ministral-3 drop it on 38–44. This is the clearest case of agreement that is correlated error. The GLM-Qwen composite, whose parser is also a Qwen but works from a transcript under its own prompt, does not share it, so a shared language model makes such habits possible without making them inevitable. *Maiden names*: of the 151 cards that have one, Qwen reproduces it on 48 % and Gemma on 51 %, against 89–97 % for the other three. Matching on the base surname raises unanimity coverage from 38.2 to 47.7 % of annotated cards, and publishing the longest variant raises the maiden-name recall of the published value from 86 % to 95 %, at

Table 4: Occurrence of each agreement level over the full corpus (486,575 cards, exact) and over the 620 annotated cards (GT), in %. Cards on which all five models abstained are excluded from the corpus shares (Birth Date: 0.5 %, negligible elsewhere). For the former name, shares are over the cards on which at least one model gives one (corpus 31.7 %, annotated 27.6 %).

	5 of 5		4 of 5		3 of 5		2 of 5		No Two Agree	
Field	Corpus	GT	Corpus	GT	Corpus	GT	Corpus	GT	Corpus	GT
Family Name	46.7	47.7	29.4	29.2	14.4	14.8	6.1	5.6	3.4	2.6
Base Surname	46.7	47.7	29.4	29.2	14.4	14.8	6.1	5.6	3.4	2.6
Former Name	19.9	16.4	34.1	31.6	19.9	26.9	10.1	9.9	15.9	15.2
First Name	53.0	56.1	26.3	26.3	12.0	11.3	5.7	4.7	3.0	1.6
Birth Date	78.1	77.7	12.6	13.4	3.2	2.7	2.3	2.1	3.8	3.5
Reference Number	77.0	74.7	7.8	6.3	7.3	10.3	6.6	7.7	1.3	1.0

a cost of 8.1 points of whole-string accuracy at unanimity — a trade worth making, because a dropped maiden name is unrecoverable downstream whereas a flagged family name is merely reviewed.

How far off the published names are. Correctness is binary: a name off by one letter counts the same as an unrelated one. Table 5 gives the character error rate (CER) of the published base surname and first name per agreement level. Dates and reference numbers are left out for two reasons. A single wrong digit already gives a different date or a different file, so a reading that is close in characters is not nearly right. And the same date or number can be written in several ways (`15.11.09` and `15.11.1909`), which would count as errors although the value is correct. For both fields the CER rises from under 1 % at unanimity to 21 % (base surname) and 24 % (First Name) where two models agree. A published name that is wrong is rarely a near miss: depending on field and agreement level, between 21 and 66 % of its characters are wrong.

How far apart the disagreeing readings are. Figure 2 shows, over the whole corpus, how far each disagreeing reading is from the published value. The mean distance of 4.6–5.1 edits hides two groups: about a third of disagreeing readings are *one* edit away (*Abeles*/*Abelos*, or an accent on the first name), and another third are 5–8 edits away, roughly a whole name token — mostly a token added or missing (*Marianne* against *Marianne Henriette*) rather than a different name. The agreement level barely changes this distribution. A one-edit tolerance in the matcher would therefore turn about a third of the disagreeing readings into agreement without merging readings

Table 5: Character error rate (CER) of the published value per agreement level on the 620 annotated cards (%), five-model vote. Total edits over total reference characters; cards scored correct in Table 3 count as zero edits, the others are compared on the normalised strings used for scoring. *Wrong Cards Only* restricts the rate to cards scored wrong; at 5 of 5 these are only 4 (base surname) and 8 (First Name) cards. *No. of Cards* is the number of annotated cards at that agreement level, over which the CER in the first row of each field is computed.

Field	5 of 5	4 of 5	3 of 5	2 of 5	No Two Agree
Family Name (Base Surname)	0.7	3.4	8.3	20.8	71.5
Wrong Cards Only	39.4	65.6	55.1	44.5	84.5
No. of Cards	296	181	92	35	16
First Name	0.6	2.2	6.8	23.8	22.4
Wrong Cards Only	21.0	57.4	41.6	57.3	25.2
No. of Cards	348	163	70	29	10

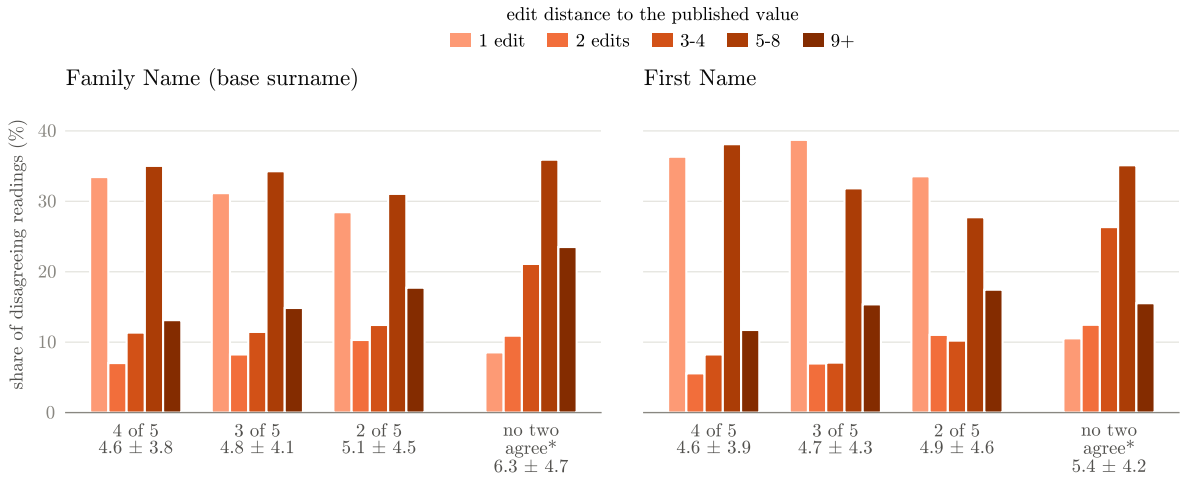

Figure 2: Edit distance of the disagreeing readings to the published value, per agreement level, over all 486,575 cards (five-model vote). Each bar is the share of disagreeing readings at that level in a distance bin; below each group, the mean ± standard deviation in edits. Distances are taken on the key the matcher compares; abstaining models are not counted. *Where no two models agree there is no majority, so the distance between every pair of readings is used instead.

that differ by a whole token; whether that helps depends on whether the one-edit dissent or the majority is right, so it is a threshold worth testing.

5.6 What unanimity still gets wrong

On the base surname, unanimity is wrong on only 4 of the 296 annotated cards where all five models agree, and all four look like annotation errors: a typo in the marker (*Abonyi, beg. Langyel*), twice *Papst* where all five read *PABST*, and *Abaddie geb. Miva* where all five read *Abadie geb. Riva*.

The whole family name string is wrong on 34 of those cards, and on 30 of them the base surname is right. On 20 of the 30, at least one model's reading matched the annotation exactly, but the rule published a longer one — with the first name appended, a former name the annotation does not record, or different punctuation. Judged by whether the person is identified, unanimity is as reliable on the family name as on any other field; for the published string to be usable unchanged, the rule would need to prefer the most common reading over the longest.

6 Discussion

What the agreement count is worth. The count does what the LAV NRW needs: it separates values that can be published unchecked from values that need review. Depending on the field, unanimous values are correct between 97 and 99 % of the time and cover half to three quarters of the cards, while the few cards where no two models agree are mostly wrong. The intermediate levels are field-specific — four of five agreeing is nearly as good as unanimity on names and dates but not on the reference number — so a review threshold should be set per field rather than once for the record.

Why the count is not a probability. Agreement can be shared error. Three of the five models share a Qwen language model, and the two end-to-end models among them both drop the `II/A` prefix of the reference number on about nine in ten cards that carry one. Five models are

therefore fewer than five independent readings. The correctness figures are also estimates from 620 cards taken from a few letter ranges; the lower agreement levels rest on a few dozen cards per field, and the fallback where no two agree on as few as six. Comparing agreement occurrence on the annotated cards with the corpus (Table 4) shows the sample to be representative within a few points, but not more.

A lone answer is not a majority. Because an abstention is not a dissent, a single model that answers where the others stay silent forms the only group, and its reading is published. On 12 annotated cards without a birth date, the other models correctly return nothing, while one model — usually Gemma — returns a stray token such as *in* or `?`, which the vote publishes. As a result the vote leaves the Birth Date empty on only 0.5 % of cards, although the annotation has none on 3.1 %. These cards are most of the reason the vote does not beat the best single model on the Birth Date (Table 2). A rule that withholds a lone answer against several abstentions, or flags it, is worth testing.

What is scored matters. For the family name, whether the person was identified and whether the published string matches the annotation are different questions. On the base surname, unanimity is as reliable as on any other field; on the whole string it is ten points lower, mostly because the rule publishes the longest reading in the winning group, which sometimes carries an appended first name or a former name the annotation leaves out. The unanimous errors that remain on the base surname look like annotation errors, so unanimous disagreement with the annotation is also a cheap way to audit the annotation itself.

Faithful output keeps variation. Because values are copied as written and never normalised, variation that exists on the cards reaches the output: two-digit years, differently grouped reference numbers, inconsistent spacing around maiden-name markers, and a few homoglyphs — Cyrillic or Greek letters inside Latin names that look identical but compare unequal. None of this blocks the import, but anyone joining or searching the data has to normalise for it. Appendix A gives the figures.

Conventions are handled by hand. Many of the remaining differences between models, and between models and annotation, are not misreadings but conventions: whether a maiden name is part of the family name, whether reading stops before the birthplace, whether a preprinted reference belongs to the reference number, how a date is written. Here they are addressed by prompt wording and hand-written postprocessing rules, each of which had to be found, written and checked against the annotation. These are exactly the regularities fine-tuning picks up from a consistently annotated target [12]. Once enough cards are verified, they could be learned rather than written by hand. Conventions are also where voting helps least. On the Reference Number, nearly all the cards on which the vote is wrong but the best single model is right carry the same digits in every reading; the majority merely drops the `II/A` prefix or keeps the preprinted form reference, while the GLM-Qwen composite happens to follow the annotation's convention. Agreement on a convention the annotation does not share cannot be outvoted.

7 Conclusion

We extracted four fields from 486,575 archival index cards with five off-the-shelf vision–language models, without fine-tuning and with only 620 annotated cards, and merged their readings field by field into a consensus that carries an agreement count for every value. On the annotated cards that count orders values by reliability: unanimous values are correct between 97 and 99 % of the time, depending on the field, and values on which no two models agree are mostly wrong. Review of an unlabelled corpus can thus be prioritised, and effort spent where the models disagree. The

count is a reliability signal, not a calibrated probability, since models that share a language model can share errors.

The chain also runs end to end. After postprocessing, the consensus was handed to the LAV NRW as one spreadsheet per collection, carrying the agreement count alongside every value, and imported into VERA — split across several finding aids and with negligible preparation. The extracted data therefore cleared the only hurdle between it and publication (Section 1).

Two lessons carry beyond this corpus. First, decisions before the model runs mattered more than the choice of model: one prompt sentence about multi-part surnames and the right image resolution moved a field by 16 and by up to 34 points respectively. Second, the merge rules have to be measured rather than assumed — an abstention is not a dissent, and a dropped maiden name is not a different person — or a plain majority vote misreports how reliable the output is.

Scored as a system of its own, the vote is also more accurate than the best single model on the names, and on par on dates and reference numbers. Further steps are to measure how correlated the models' errors are, in particular between models sharing a language model, to test an edit-distance tolerance in the matcher and a publishing rule that prefers the most common reading over the longest, and to feed the low-agreement cards into human verification. The verified cards would in turn provide the training material this work lacked, so that the extraction conventions now handled by prompt and postprocessing rules could be learned by fine-tuning.

References

[1] Manuel Carbonell et al. "Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model". In: *Proc. IAPR Int. Workshop on Document Analysis Systems (DAS)*. arXiv:1803.06252. 2018.

[2] Solène Tarride, Mélodie Boillet, and Christopher Kermorvant. "End-to-end Information Extraction in Handwritten Documents: Understanding Paris Marriage Records from 1880 to 1940". In: *Proc. Int. Conf. on Document Analysis and Recognition*. arXiv:2404.19329. 2023.

[3] Kevin Ding. "When LLMs Agree, Are They Right? Auditing Self-Consistency and Cross-Model Agreement as Confidence Signals". In: *arXiv preprint arXiv:2607.08065* (2026).

[4] Yiheng Xu et al. "LayoutLM: Pre-training of Text and Layout for Document Image Understanding". In: *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. 2020. DOI: 10.1145/3394486.3403172.

[5] Yihao Ding, Jean Lee, and Soyeon Caren Han. "Deep Learning based Visually Rich Document Content Understanding: A Survey". In: *arXiv preprint arXiv:2408.01287* (2024).

[6] Geewook Kim et al. "OCR-free Document Understanding Transformer". In: *Computer Vision – ECCV 2022*. Springer, 2022. DOI: 10.1007/978-3-031-19815-1_29.

[7] Kenton Lee et al. "Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding". In: *Proc. Int. Conf. on Machine Learning*. 2023.

[8] Ahmed Cheikh Rouhou et al. "Transformer-based Approach for Joint Handwriting and Named Entity Recognition in Historical Documents". In: *Pattern Recognition Letters* 155 (2022), pp. 128–134. DOI: 10.1016/j.patrec.2021.11.010.

[9] Gavin Greif, Niclas Griesshaber, and Robin Greif. "Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents". In: *arXiv preprint arXiv:2504.00414* (2025).

[10] Mahsa Vafaie, Sven Hertling, and Harald Sack. "End-to-end Information Extraction from Archival Records with Multimodal Large Language Models". In: *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*. Seoul, Republic of Korea, 2025. DOI: 10.1145/3746252.3761503.

[11] Fabian Wolf et al. “CM1 – A Dataset for Evaluating Few-Shot Information Extraction with Large Vision Language Models”. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Wuhan, China, 2025. URL: https://arxiv.org/abs/2505.04214.

[12] Arthur Matei et al. “Recent Advances in Information Extraction from Historical Archival Records”. In: *Proc. Int. Conf. on Document Analysis and Recognition*. to appear. Vienna, Austria, 2026.

[13] Xuezhi Wang et al. “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: *Proc. Int. Conf. on Learning Representations (ICLR)*. 2023.

[14] Taylor Archibald and Tony Martinez. “Improving MLLM Historical Record Extraction with Test-Time Image Augmentation”. In: *Proc. Int. Conf. on Document Analysis and Recognition*. Vienna, Austria, 2026. arXiv: 2509.09722.

[15] Qwen Team, Alibaba Cloud. *Qwen3.5: Towards Native Multimodal Agents*. https://qwen.ai/blog?id=qwen3.5. Model release announcement; technical report forthcoming. 2026.

[16] Qwen Team, Alibaba Cloud. “Qwen3-VL Technical Report”. In: *arXiv preprint arXiv:2511.21631* (2025).

[17] Gemma Team, Google DeepMind. “Gemma 3 Technical Report”. In: *arXiv preprint arXiv:2503.19786* (2025).

[18] Xiaohua Zhai et al. “Sigmoid Loss for Language Image Pre-Training”. In: *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*. 2023.

[19] Shuaiqi Duan, Yadong Xue, Weihan Wang, et al. “GLM-OCR Technical Report”. In: *arXiv preprint arXiv:2603.10910* (2026).

[20] Weiyun Wang et al. “InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency”. In: *arXiv preprint arXiv:2508.18265* (2025).

[21] Alexander H. Liu et al. “Ministral 3”. In: *arXiv preprint arXiv:2601.08584* (2026).

A Residual variation in the extracted data

Values are published as written, so formatting variation on the cards carries over into the output. Percentages are over all 486,575 cards.

Birth Dates. 68.9 % are written as `D.M.YYYY`; the other forms are listed in Table 6. Two-digit years (18.5 %) leave the century to whoever parses the date, and year-only dates (5.2 %) are usually all the card offers. 1.0 % contain no digit at all (*in*, `-`, `?`) and should be filtered downstream.

Reference Numbers. 63.9 % are space-grouped (`123 456`), 9.2 % ungrouped, 7.6 % dot-grouped and 4.1 % carry an `II`/`A` prefix; the rest mix separators or carry an `a`/`b` suffix. Grouping can be folded when matching; prefixes and suffixes cannot, because they distinguish different files. The annotation is itself inconsistent here: it resolves `56/II/47 419` to `II/47 419` on one card but `56/II/56 265` to `56 265` on another, which affects 51 of the 620 annotated cards.

Maiden-name markers. Of 129,266 surnames with a *geb./verw./gesch.* marker, 82.6 % read *geb. X*, 10.3 % have a comma before the marker and 6.8 % no space after it. A join or deduplication on surnames must normalise the marker.

Homoglyphs. 36 names contain Cyrillic or Greek letters inside Latin words (*Gеbriella*, *Militiζano*). They look identical on screen but compare unequal, so a lookup fails without visible cause. Non-Latin characters in Latin-script fields should be normalised or flagged before indexing.

Table 6: Written forms of the Birth Date field across all 486,575 cards (483,883 non-empty).

Form	Cards	Share (%)
Canonical `D.M.YYYY`	333,452	68.9
Two-Digit Year	89,361	18.5
Year Only	24,984	5.2
Month Spelled Out	10,252	2.1
Other	9028	1.9
Spaced Separators	8366	1.7
No Digits	4875	1.0
Roman Month	2790	0.6
Month and Year Only	628	0.1
Day and Month Only	147	0.0