Accurate and Unbiased?
Vision-Language Models for Historical Data

Mathias Bühler
Davide Cantoni
Pascal Henschke*

July 2026

Abstract

Vision-language models (VLMs) can transcribe historical records at a fraction of the cost of manual transcription. But can they be trusted? We evaluate the quality of VLM-based transcriptions using a large corpus of historical city directories for the city of Munich, comparing them to high-quality manual transcriptions (our "ground truth"). The VLM recovers 99.4% of records and errs on fewer than seven characters per thousand. Because a VLM leans on a modern-language prior, its errors in transcription of historical data may be biased. However, using three common estimands relevant for economic history research (the occupational status distribution, the gender status gap, and linkage rates to four external sources), we show that the estimates using VLM-transcribed data do not differ from those using manual transcriptions.

*Bühler: Ludwig-Maximilians-Universität Munich. Email: mathias.buehler@econ.lmu.de. Cantoni: Ludwig-Maximilians-Universität Munich, CEPR, and CESifo. Email: davide.cantoni@econ.lmu.de. Henschke: Ludwig-Maximilians-Universität Munich. Email: pascal.henschke@econ.lmu.de. We acknowledge financial support from the LMUexcellent Investment Fund and from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement n. 101044546, "CityRising").

1 Introduction

Economic history research often relies on records that survive only on paper; turning them into data has long been the field's bottleneck. Hand transcription is accurate but slow and costly, and optical character recognition (OCR), though cheaper, needs clean scans, orderly layouts, and heavy post-processing. Vision-language models (VLMs) promise the same transcription faster and at a fraction of the cost (Bäcker-Peral et al., 2025). How good are the transcriptions they produce?

We answer this question with the city directories of Munich, for which we hold one of the largest manual transcriptions and can compare a state-of-the-art VLM against it. Published for many European cities, directories record residents, their addresses, and their occupations (Albers and Kappner, 2023; Bühler et al., 2024). We transcribe 29 Munich directories published between 1845 and 1953 twice, once by a professional data entry company and once by a VLM, and align the two into 3.4 million matched record pairs.

We first examine *accuracy*: how closely does the transcription match a benchmark? We use the manual transcription as our benchmark ("ground truth"). The firm worked to a contractual error rate below 0.5%, which our own checks confirm. Measured against this benchmark, the VLM is accurate: it recovers 99.4% of records, and its character error rate is 0.7%, about seven mistakes in every thousand characters.

Whether those errors change a researcher's inference turns on one question: are they as good as random, or do they systematically concentrate on particular entries? Errors emerge where the image of a character is ambiguous; in such a case, the result of a transcription depends on whether, and how, the chosen method leans on a prior. Human transcribers bring a potentially strong prior based on their background and contextual knowledge (a data-entry clerk will differ from a professional historian). Traditional OCR methods carry only a weak prior, so their errors are visually driven (Karamolegkou et al., 2026). A VLM carries a far stronger one: the prior of its underlying large language model, trained on predominantly modern text (Lee et al., 2025; Luo et al., 2026). In the few evaluations to date, transcription errors lean toward modern forms (Vesalainen et al., 2026; Humphries et al., 2025). A modern prior is more likely to resolve an uncertain word toward the one common today: for example, an archaic occupational title such as *Verweser* (a steward, or administrator) may be read as *Vermesser* (surveyor). If such errors are systematic, they can influence downstream estimates produced by a researcher, even when the transcription is accurate on average. Whether the pull toward modern forms is strong enough to do so is an empirical question.

The presence and extent of such a *bias* will depend on the precise estimand that is recovered from the transcribed data. We call the VLM unbiased when, across different estimands, it returns the same value as the manual transcription, our benchmark. Our tests examine different moments of the data: the mean of the occupational-status distribution, its dispersion, the estimated gender gap in status, and the rate at which records link to four outside sources. In all cases, estimates recovered from the VLM transcription do not differ meaningfully from those using the manual transcription:

for instance, the gender gap in occupational status is 1.894 points on the manual transcription and 1.888 on the VLM, a difference indistinguishable from zero across more than three million records. The mean, the dispersion, and the linkage rates match as closely, never differing enough to change a conclusion.

Importantly, our comparison of VLM-based transcriptions and manual transcriptions holds the post-processing "pipeline," which cleans, standardizes, and codes the raw fields into variables such as occupational status and gender, fixed. The pipeline built and developed on the manual transcription (Bühler et al., 2024) is applied unchanged to the VLM transcription.

A recent literature has examined the performance of VLM transcription methods against OCR baselines on a variety of historical documents: city directories, tabular and handwritten records, and archival print (Greif et al., 2025; Kim et al., 2025; Humphries et al., 2025; Griesshaber and Streb, 2025). These evaluations rely on fewer than 50 pages of ground truth, typically transcribed by the researchers themselves. We align 3.4 million record pairs across 15,000 pages, transcribed by a professional data-entry company to a contractual error rate below 0.5%. At this scale, systematic biases can be uncovered.

We further examine the question whether VLM-transcribed data deliver not only similar or identical text, but also allow for the same inference on downstream estimates. When machine-generated variables enter a regression, their errors propagate, and a recent literature builds estimators that recover valid inference from imperfect predictions (Angelopoulos et al., 2023; Egami et al., 2023; Battaglia et al., 2024; Ludwig et al., 2025). Our results suggest that, in the context of German-language city directories, the result of VLM transcriptions is accurate and does not result in bias when estimating plausible downstream parameters of interest. VLM transcription reproduces the same conclusions a researcher would draw from manually transcribed data. At €0.08 per thousand entries against €39 in 2026, it brings within reach historical sources once too costly to digitize.

The remainder of the paper proceeds as follows. Section 2 presents the historical data of interest, the Munich city directories, and the variables recovered from these sources. Section 3 explains our transcription approaches (VLM and manual) and defines the metrics of interest. Section 4 studies the accuracy of transcriptions. Section 5 investigates the invariance of downstream estimates, and Section 6 examines specifically the performance in cross-source linking. Section 7 concludes.

2 Data

2.1 The directory volumes

The corpus comprises 29 Munich directories published between 1845 and 1953.[1] Together the 29 volumes span 15,179 pages and 3,442,200 manually transcribed records. We assemble the

[1]We sample volumes at five-year intervals, and annually across 1880–1890 (Bühler et al., 2024). Appendix Table A.1 lists all volumes and entry counts.

corpus from existing scans at the Bavarian State Library and from physical volumes that we digitize ourselves. Figure 1 shows exemplary pages. Directories of the era are printed in *Fraktur*, a blackletter typeface whose frequently near-identical letterforms represent a challenge for character recognition. A method that transcribes Fraktur accurately will handle cleaner scripts more easily, so establishing reliability on the German-language directories should also be informative for other sources.

A directory has two sections. The names section, present in every year, lists residents and businesses alphabetically. Each entry records a surname, first name, occupation, street, house number, floor, and building designation (*Vorderhaus*, main building, or *Hinterhaus*, courtyard lot). The houses section, introduced in 1875, organizes the same fields by location: a bold street name heads a block, and the entries below it record house numbers, residents, and the building owner. From 1875 onward we instructed the data entry company to transcribe the houses section rather than the names section, because its street names are consistent and unabbreviated and it adds information on building ownership.

The Munich police administration published the directories until 1914, drawing on the city's compulsory residence-registration system (*Meldewesen*), so each entry was sourced from a near-universal administrative register rather than from voluntary subscription. Private publishers continued the series after 1914 using the same registration data, with no visible break in coverage.

External sources. To validate external linkage in Section 6, we test whether the VLM and the manual transcription link external records to the same directory entry. We use four lists that vary in era and in the match keys they carry: the *Verlustlisten* (military casualty rolls) of the Franco-Prussian War and of WW1 record only names, whereas the 1909 *Auto-Adressbuch* (a register of early automobile owners) and the 1933 *Gedenkbuch* (a memorial register of Munich's persecuted Jewish community) record both names and addresses.[2]

2.2 Derived variables

Section 5 relies on two variables derived from the transcribed fields: an occupational status score and a gender indicator. Both are constructed by the same rule on the ground-truth and VLM transcriptions, so any downstream gap between sources reflects transcription, not classification.

Occupational status. We classify each occupation string into HISCAM (Lambert et al., 2013) and HISCLASS-12 (van Leeuwen and Maas, 2011). HISCAM is a continuous 0–100 occupational prestige score (unskilled laborer $\approx$ 30, craftsman $\approx$ 50, professional $\approx$ 80); HISCLASS-12 is a social class scheme.[3] We use the HISCAM-U1 variant, the universal scale estimated on the pooled cross-country

[2]The two Verlustlisten and the *Auto-Adressbuch* were transcribed by volunteers of the Verein für Computergenealogie (`des.genealogy.net`); the *Gedenkbuch* is published by the Munich Stadtarchiv (`gedenkbuch.muenchen.de`).

[3]The HISCLASS-12 scale runs natively from 1 (highest class) to 12 (lowest). We reverse it throughout the paper so that, as with HISCAM, a higher value denotes higher status; a positive coefficient on either scale therefore signals higher-status occupations.

sample with unconstrained occupational categories, so that scores are comparable across the long span of the corpus.

Raw occupation strings are heterogeneous, heavily abbreviated, and often record several roles per entry, so classification proceeds in three stages. First, an LLM normalizes each string: it expands abbreviations, modernizes spelling, splits multi-occupation entries, and extracts titles, retirement markers, and the gender signal implicit in the grammatical gender or dependency relationship (widow, daughter, wife). Second, two methods independently propose candidate 5-digit HISCO codes for each normalized occupation: embedding similarity against the HISCO codebook (van Leeuwen et al., 2002) and the OccCANINE neural classifier (Dahl et al., 2024). An LLM then reranks the candidates and selects one. The most common HISCO assignments, covering approximately 90% of directory entries, were additionally manually inspected and verified. Approximately 94% of entries receive a valid HISCAM score. Unmappable strings (blank occupations, institutional entries, occupational strings not covered by HISCO) are dropped from the analysis subsample.[4]

The resulting status distribution is right-skewed (Figure 2, Panel A): the mean HISCAM score (66) sits above the median (63), and the 10th percentile (53) lies closer to the median than the 90th (86). Occupation strings are also concentrated in a few common titles (Figure 2, Panel B): the three most frequent (*Kaufmann*, merchant, *Privatier*, rentier, and *Taglöhner*, day laborer) account for 10% of all entries, and the 1,000 most frequent cover four-fifths of the corpus, with a Zipf coefficient close to one over that range.

Gender. Gender is inferred from two complementary signals. First, we use the gender signal extracted from the occupational title during normalization. If the occupation indicates a widow, wife, or daughter, we code the entry as female. Second, we look up the first name in a canonical first-name dictionary. Approximately 94% of entries receive a gender assignment. The remainder are unresolvable on either signal. This includes business and institutional entries, and persons where both the first name and occupation field are gender-ambiguous due to abbreviations.[5]

3 Methodology

We first define the objects to be studied in the remainder of the paper. Let x denote a directory page image and y its correct transcription, the characters actually printed on the page, period spelling, any potential typos and abbreviations included. Two procedures aim to recover y from x: a manual transcription $m(x)$ and a VLM transcription $v(x)$. We treat the manual transcription as ground truth, $m(x) \approx y$. It is the conventional procedure for digitization, and where the two disagree it is correct more often than the VLM (Section 4.3).

[4]The largest group are entries recording only a widow marker (*We.*) and no own occupation; the widow information is retained in the gender variable.

[5]For instance, the abbreviated *Joh.* could be either *Johannes* (male) or *Johanna* (female). Abbreviations are common in the directory, and they are more common in the later years due to space constraints.

The two procedures may disagree whenever a character is ambiguous or uncertain. The prior used by a VLM to resolve such an ambiguity leans towards modern text (Vesalainen et al., 2026; Lee et al., 2025). Its errors are more likely to fall on entries a modern prior finds improbable: historical occupations, obsolete titles, and period spellings. Historical usage is not directly observable, so we proxy it by corpus frequency, calling an entry rare when its string appears few times in the manual transcription.[6]

We evaluate the VLM on two properties. Accuracy is fidelity to the page: the distance $d(v(x), m(x))$ under a given metric d defined below in Section 3.3. Unbiasedness, instead, is a property not of the transcription itself, but of an estimand F computed from it: for example, a status distribution, a regression coefficient, or a cross-source linkage rate. We call the VLM unbiased for F when $F(v(x)) = F(m(x))$. Bias is relative to the manual benchmark, not to the underlying population. Importantly, both procedures read the same scans, so any selection in the source enters $m(x)$ and $v(x)$ alike; what we test is whether the VLM adds bias beyond what the hand-coded corpus already carries.

Section 3.1 describes the manual transcription that supplies the ground truth. Section 3.2 describes the VLM pipeline. Section 3.3 defines the three metrics used throughout Sections 4–6: coverage, per-field match rates, and character error rate.

3.1 Manual Transcription

We commissioned a professional data entry company to manually transcribe the corpus. Transcribers entered the data into a spreadsheet, yielding more than 3 million entries. The average cost of manual transcription in the years 2024–2026 was around €39 per thousand entries. It worked under a contractual error rate of below 0.5% CER, which our checks confirm.

As in any manual transcription, some errors remain: transcribers occasionally misread visually similar Fraktur characters (such as 𝔅 (B) and 𝔙 (V), or ſ (s) and f (f)), or misclassify parts of an entry (for instance, recording personal titles in the occupation field rather than the name field). Weak print and page bleed-through during scanning can reduce legibility for any reader.

3.2 VLM-Based Extraction Pipeline

Classical OCR pipelines and their handwritten-text-recognition (HTR) counterparts first detect and segment text lines, then recognize individual characters, and finally parse the plain text into structured records in a separate step. We use a vision-language model (Gemini 3 Flash[7]) that takes the page image as a single input and produces structured fields in one pass. Image input lets the model use visual cues that classical OCR discards after text recognition: bold-printed surnames, indented sub-entries, and subscript floor numbers.

[6]The proxy is loose: rarity also captures uncommon but modern strings, which only enlarges the set of entries where error could appear and makes the tests in Sections 4–6 stricter than the historical set alone would require.

[7]Specifically *gemini-3-flash-preview*.

We submit each image separately and instruct the model to output entries in tab-separated format (TSV).[8] We disable the model's reasoning process, because it raises costs and introduces non-determinism without improving extraction quality.

Output format. The extraction prompt defines the required output format, matching the manual transcription: surname, first name, occupation, street, house number, floor, building, and owner indicator. The prompt instructs the model to resolve ditto marks for repeated surnames, distinguish personal titles (like *Dr.* or *von*) from occupational prefixes (like *kgl.* (royal) or *Wwe.* (widow)), parse fractional or suffixed house numbers, and handle institutional entries that lack a personal name. For the houses section, the prompt directs the model to output street headers and individual entries as distinct rows. Appendix D reproduces the full prompts.[9]

Post-processing. Entries inherit the street from the preceding header, so post-processing propagates each header forward through the column until the next header replaces it. The same step removes non-entry content interspersed in the volume: advertisements, business-registry numbers, opening hours, and directions.

Few-shot examples. We pair each prompt with three to five few-shot examples, real page scans matched to verified outputs, chosen during prompt development to cover edge cases and common errors.

Column-splitting at high density. Gemini 3 Flash represents each page image with 2,280 visual tokens regardless of how much text it contains. At three or more columns the model has too few tokens per entry to maintain reading order or to render entries near the page edges. From 1888 onward, the directory's layout features three or more columns per page, rising to four in 1905 and six in 1933 (Figure 1), so we adjust the process accordingly. Rather than submitting the whole page, we crop each dewarped page into its individual columns and submit them separately. Appendix C gives the procedure; Section 4 reports the accuracy gain.

The pipeline produces 3.44 million records across the 29 directories. The average processing cost of €10 per volume in 2026 is about €0.08 per thousand entries, a fraction of the manual transcription cost. Since the cost is dominated by the VLM inference, we expect it to fall as the technology advances.

[8] We choose TSV over the natively supported JSON because JSON's per-entry token overhead would inflate costs at scale.

[9] The prompt is the same for all volumes, with minor adjustments to reflect layout changes. For instance, the prompt for the names section instructs the model to read entries in alphabetical order, while the prompt for the houses section instructs it to read entries in directory order.

3.3 Evaluation

We score the VLM transcription against the manual transcription ("ground truth," hereafter: GT) on three metrics: coverage, per-field match rates, and character error rate (CER). Two preconditions make the comparison fair. The two corpora are not directly comparable as strings, because the GT transcribers apply editorial conventions that depart from a literal page transcription. In a normalization procedure, we instruct the VLM to follow the same conventions, so that a literal transcription does not register as error. The metrics also require pairs of corresponding entries rather than two unordered lists; row alignment supplies them.

Normalization. The GT applies editorial conventions that the VLM does not: transcribers split joint entries (so "*Müller Max und Anna, Doktor*" becomes two records), separate multi-floor residents, and strip metadata from street names. Without normalization, faithful VLM transcription registers as error. To match the GT conventions, we split entries with "*und*" in the first-name field and collapse duplicates that differ only in floor. Across all fields we lowercase, strip punctuation, and collapse whitespace. We apply the same rules to both corpora.

Row alignment. We align GT to VLM rows in two passes. The first pass identifies *exact anchors*: rows whose key appears exactly once in both transcriptions. The key is the person string concatenated with the house number in the houses section, and additionally with the street in the names section, where the same surname recurs across many addresses. Both transcriptions follow the same page order, so we enforce monotonicity among the anchors with a longest increasing subsequence on VLM indices, discarding any anchor that would break the order. The first pass anchors more than 90% of GT rows on exact matches alone, leaving short gaps for the second pass. In the second pass we match each remaining GT row to its nearest unmatched VLM row, where distance is the normalized Levenshtein distance averaged across fields; we accept a pair if its distance is at most 0.5, and we commit each match before moving on rather than searching for a globally optimal assignment. The procedure yields three categories: *aligned* pairs, *missed* entries (GT rows with no VLM counterpart), and *extra* entries (VLM rows with no GT counterpart).

3.3.1 Evaluation Metrics

We use three evaluation metrics to answer three distinct questions: which records did the alignment recover (coverage), how often is each field exactly right (per-field match rates), and among the imperfect pairs, how far off is the residual (CER).

Coverage. Coverage is the share of GT entries the alignment recovers: of every hundred records the manual transcriber wrote down, how many does the VLM also produce? Formally,

$$\text{Coverage} = \frac{|\text{aligned}|}{|\text{GT entries}|}. \tag{1}$$

The metric corresponds to recall in the OCR-evaluation literature (Levchenko, 2025).

Per-field match rates. A per-field match rate is the share of aligned pairs whose GT and VLM strings agree on a given field. We report two variants for each of the five fields (lastname, firstname, occupation, street, house number), corresponding to two definitions of agreement (Levchenko, 2025; Greif et al., 2025). *Exact match* requires the normalized strings to be identical, the standard for downstream uses that require exact strings (such as exact-match record linkage). *Fuzzy match* accepts a normalized Levenshtein distance up to $\tau = 0.1$, roughly one edit on a ten-character string, the standard for downstream uses that tolerate small character-level errors; L is the minimum number of character insertions, deletions, and substitutions.

Character error rate (CER). CER is the Levenshtein edit distance between the GT and VLM strings, normalized by GT string length:

$$\text{CER}(s_{\text{gt}}, s_{\text{ocr}}) = \frac{L(s_{\text{gt}}, s_{\text{ocr}})}{|s_{\text{gt}}|}. \tag{2}$$

Where match rates collapse the distance to a binary correct/incorrect, CER reports its size. By construction, an exact match yields CER $= 0$. A single substitution on a nine-character string yields CER $= 1/9 \approx 0.11$, for example the Fraktur misread of *n* as *m* that turns "Schreiner" (carpenter) into "Schreimer".

Person CER (headline metric). Person CER is computed on the concatenation of lastname, firstname, and occupation. We use it as the headline accuracy metric in Section 4 because it does not penalize correctly-transcribed tokens placed in the wrong field, a misattribution that inflates field-level CER without reflecting genuine character error. We also report *Entry CER*, computed on the full entry string in directory order, and *field-level CER* for each field; Appendix Table B.2 decomposes the field-level metric into the misattribution and genuine components.

4 Coverage and Accuracy

For a digitized corpus to be useful for research, the records must be recovered and the fields must be correct. Missing records bias every analysis if the absences correlate with traits like occupation, address, or status. Garbled names break linkage and corrupt any variable derived from the page.

We benchmark VLM extraction against manually transcribed ground truth on both: which records the model recovers (coverage) and what it writes in each field (accuracy). Section 3.3 defines the metrics; this section reports them.

4.1 Coverage and accuracy

VLM extraction recovers essentially every record present in the ground truth (Table 1). Row alignment matches 99.4% of GT records to a VLM counterpart and leaves 0.6% (20,075) unmatched. In the other direction, 0.5% (17,536) of VLM records have no GT counterpart, of the same order as the unmatched-GT rate. The miss rate does not concentrate in any single year: 28 of 29 directories clear 99% coverage, and even the worst year (1880) reaches 98.2%.

Across the 3.42 million aligned pairs, exact-match rates by field are 99.7% on house numbers, 97.2% on first names, 94.6% on last names, and 93.4% on occupations (Table 2). The relevant field differs by use: name fields drive linkage, occupations feed the status and gender analysis in Section 5, and house numbers carry every spatial join.

Person strings, the concatenation of lastname, firstname, and occupation, agree exactly between VLM and ground truth in 90.2% of aligned pairs. Allowing a normalized Levenshtein distance of 0.1, roughly one changed character in ten, lifts agreement to 98.4% (Table 2). This implies that 8% of the records that the VLM recovers differ from the GT by a single character. This tolerance helps the fields with more characters, so the fuzzy-match gain is 2.7 percentage points for occupational titles but only 1 percentage point for last names and zero for first names.[10] Exact-string matching treats one person record in ten as a non-match, while fuzzy matching recovers all but one in sixty.

Mean person-level CER is 0.007: seven substituted characters per thousand characters of name and occupation transcribed. These errors concentrate on a small set of common single-character Fraktur confusions ({r,n,u}, {s,f,t}, {V,B}, {r,x}) shown in Appendix Table B.1. Errors load on rare entries, as the model's language prior would predict: for occupation strings appearing exactly twice in the manual transcription, the VLM reproduces the string exactly in only 79% of cases, rising monotonically to 98% for occupations appearing more than 100 times (Appendix Figure G.1).

Figure 3 shows that CER is constant across directory-years: every year sits between 0.004 and 0.010, unweighted mean 0.0060. The two worst years, 1910 and 1938, both reach 1.0%, and they share a layout feature: four-column houses-section pages at high entries-per-image density. Name-ordered (1845–1870) and house-ordered (1875–1953) sections perform comparably, with unweighted means of 0.0058 and 0.0060, suggesting that the variation in CER is driven by density, not directory format.

[10] In our corpus, only 1.53% of first names are longer than ten characters, so a single-character error on most first names already exceeds the 0.1 threshold. As 13.33% of last names and 53.03% of occupational titles are longer than ten characters, the same error rate mechanically translates to a smaller normalized distance and thus a larger fuzzy-match gain.

4.2 The density channel

Density matters because of the model's fixed visual-token budget: Gemini 3 Flash represents each input image with 2,280 visual tokens regardless of how much text the page contains, so above a density threshold the model has too few tokens to render each entry reliably.[11] Common failures are omitted entries near page edges and broken reading order across columns, both first appearing in our corpus when directories adopted three-column layouts in 1888.

To identify the channel, we re-run the eight densest directory-years (1887–1905, houses section) in two modes: standard full-page submission, and a preprocessed variant that crops each page into its individual columns and submits the columns separately.[12] Holding the underlying images and the prompt fixed, any difference in measured accuracy is a density effect. Table 3 reports the comparison, with Δ defined as column-split minus full-page.

Column splitting raises person-level exact match by 1.9 percentage points (from 87.7% to 89.6%), cuts entry CER from 0.013 to 0.010, and lifts coverage by 0.4 percentage points. Field-level exact-match gains run from 0.7 to 1.4 percentage points across name and occupation fields, with the largest improvements in the densest years (Appendix Table E.1).[13] This suggests that page density is a binding constraint in VLM digitization. After we adjust the procedure and crop pages, the densest directories perform as well as the corpus mean.

4.3 Adjudicating the disagreements

The character error rate measures divergence between the two transcriptions, not how often the VLM is wrong. When the two disagree, neither is established as correct. We randomly selected 500 aligned pairs whose person strings (surname, first name, and occupation) disagree by at least one character, and adjudicated each by hand.[14] The manual transcription is correct in 71% of the cases, the VLM in 25%, and neither in 4%. The manual transcription is correct more often than the VLM, which is why we treat it as the reference. Still, the VLM is right in a quarter of the disagreements, so the 0.7% character error rate (Section 4.1) overstates VLM error: some of the divergence it records is the manual transcription's mistake.

Which source is right depends on the field, in the way the model's language prior predicts

[11] Consistent with the multi-column degradation documented by Greif et al. (2025) and the image-resolution thresholds for VLM-based OCR characterized by Inoue (2025).

[12] The crop uses a UVDoc dewarp followed by a peak-detection algorithm on the page's vertical projection. Full procedure in Appendix C.

[13] One tradeoff is that street headers spanning multiple columns are no longer visible to the model on a single image, so we re-attach them in post-processing using column-boundary geometry. We also tested stacking columns vertically into a single tall image, which preserves density but performs worse than either alternative, consistent with such aspect ratios being out-of-distribution for the VLM.

[14] The 500 are drawn at random from the aligned pairs that carry a valid HISCAM score on both sides, so the frame excludes disagreements on institutional or otherwise unclassifiable entries. Five of the 500 turned out during adjudication to be row-misalignments or field-misassignments rather than transcription disagreements and are excluded, leaving 495 adjudicated pairs.

(Table 4). A VLM resolves an unclear glyph toward the most probable string, which helps when the true reading is in common modern use and hurts when it is historical or otherwise rare. The true reading of a first name or an occupation is usually common, so the prior pulls toward it. The VLM is correct in a quarter to a third of these disagreements. Surnames are often rare or near-unique, so the prior pulls away from the true reading. The manual transcription is correct in 85% of last-name disagreements.

Sorting the same disagreements by character substitution shows the two sources erring in systematically different ways (Appendix Figure B.1). Some are VLM misreads of Fraktur type, such as *R* read as *N* (*Rabel* as *Nabel*), where the manual transcription is almost always correct (Appendix Table B.1). Others are manual misreads the VLM corrects, such as a final *l* recorded as *t* (*Stöckerl* written *Stöckert*).

5 Invariance of Downstream Estimates

Section 4 establishes character-level accuracy, but what matters is whether the downstream estimates are the same: does a regression on the manual transcription yield the same conclusion as one on the VLM transcription? A vision-language model transcribes by conditional language generation, so an ambiguous character resolves toward the most probable continuation under the model's modern language prior rather than toward the reading most faithful to the page (Vesalainen et al., 2026; Karamolegkou et al., 2026). That pull may bias the estimates a researcher reports.

For occupational status, the status distribution is right-skewed: a short dense floor near the common laborer *Taglöhner*, and a long thin tail of rarer elite titles above *Privatier* (Figure 2, Panel B). These elite titles are historical in usage and therefore rare in the corpus (Section 3), the readings a modern prior is least likely to reproduce. If the VLM pulls them toward their common modern forms, it moves entries inward, and disproportionately down from the long upper tail. The variance would fall, and the mean would fall with it.

We test this compression on the 3.10 million aligned pairs for which both transcriptions yield a valid HISCAM score. This represents 90.6% of all aligned pairs, of which roughly 24% are classified female.[15] At this sample size we can detect compression orders of magnitude smaller than any that would distort a downstream regression, so a null is informative, not underpowered.

Before testing whether downstream estimates yield different results, we describe where the two transcriptions disagree at the record level (Table 5). The model and the ground truth assign a different gender to 2,803 aligned pairs, 0.09% of all gender-classifiable pairs, and the disagreements are lopsided: 74.6% of them turn a woman into a man. A single Fraktur confusion drives almost all of that asymmetry. The abbreviated female name *Mar.* (Marie) can be read as the male *Max.* through the r/x substitution cataloged in Appendix Table B.1, and it alone accounts for three-fifths of the

[15] HISCAM, HISCLASS-12, and the female indicator are the variables defined in Section 2.2.

female-to-male flips. Net of it, the gender-flip rate falls to 0.05%, and the direction is close to even (54.5% female-to-male).

Status disagreements are rarer and more symmetric: the model assigns a different HISCAM score to only 0.45% of pairs, raising status in 52.8% of those cases and lowering it in 47.2%. On average, these record-level disagreements cancel out and show no systematic pull to either side. Whether these shifts still bias estimates or distributions is the test that follows.

5.1 Test

A natural test is the gender gap in occupational status, the difference in mean recorded status between women and men. In the directory this gap is positive: women appear at higher recorded status on average. The source under-records low-status women such as servants, lodgers, and the transient poor, while the women who do appear skew toward higher-status standalone entries: widows carrying a late husband's status, *Privatiere*, and *Rentnerinnen*. The VLM threatens the gap through two channels: it could compress the status scores, and it could mislabel women as men. Where the model turns a woman into a man (Table 5), it moves a higher-status entry into the male group and pulls the gap toward zero from above. Because both channels pull the gap toward zero, the gender gap is the ideal test for bias.

We estimate the difference between the gender-status gradients in the ground-truth and VLM transcriptions, stacked into one panel:

$$\text{HISCAM}_i = \alpha + \beta\,\text{Female}_i + \gamma\,\text{VLM}_i + \delta\,(\text{Female}_i \times \text{VLM}_i) + \lambda_t + \varepsilon_i, \tag{3}$$

where VLM_i marks the rows the VLM transcribed, λ_t are directory-year fixed effects, and we cluster standard errors by directory-year. The interaction $\hat{\delta}$ is the invariance test: it measures whether the gender-status gradient differs when the data come from the VLM rather than from manual transcription.

The downstream coefficient. Table 6 estimates the stacked regression of equation (3) and sets the ground-truth and VLM coefficients side by side. The estimated gender gap in HISCAM is 1.894 points on the manual data and 1.888 on the VLM data; in HISCLASS-12 it is 1.078 and 1.079.[16] The interaction $\hat{\delta}$ that separates the two is -0.006 for HISCAM and 0.000 for HISCLASS-12, neither distinguishable from zero and both below half a percent of the gender gap itself, a null the 6.2 million stacked observations make precise.[17] A researcher who estimated this regression on the

[16] We reverse the HISCLASS-12 scale here so that, as with HISCAM, a higher number denotes higher status; a positive Female coefficient on both measures therefore signals that women hold higher-status occupations on average.

[17] Estimating the same equation with entry fixed effects (one per aligned pair), which absorbs every entry-level characteristic and identifies $\hat{\delta}$ from the within-pair GT-vs-VLM comparison alone, returns an interaction of comparable magnitude: $+0.046$ HISCAM points and $+0.009$ HISCLASS-12 categories, 2.4% and 0.8% of the respective gender gaps. Appendix Table G.2 reports this specification alongside wild-cluster-bootstrap inference; both adjustments leave $\hat{\delta}$ within rounding distance of zero.

VLM transcription would report the same gradient, the same sign, and the same significance as one working from manual transcription.

The pattern is not specific to the gender-status gradient: Appendix Table G.1 repeats the test across four downstream regressions, including the white-collar elite share and the reverse status-to-gender direction, and finds $\hat{\delta}$ well under 1% of $\hat{\beta}_{\text{GT}}$ in each. It is also not specific to the full-sample average: Appendix Table G.4 re-runs the HISCAM specification on the bottom 5%, bottom 10%, top 5%, and top 10% of the manual-transcription HISCAM distribution, where any compression of the distribution would be largest, and finds that the only significant interaction on a tail with a nonzero gender gap is 1.2% of that gap. The standard error on $\hat{\delta}$ (HISCAM 0.009, HISCLASS 0.003) is small enough that the null reflects the absence of an effect, not the absence of power.

Mean status by year. Linear coefficients could still mask a compressed status distribution, if the compression happened to leave the gender gradient untouched. It does not. Figure 4 plots the mean of each status scale by directory-year, computed on the full transcribed corpora rather than the aligned subsample, so the comparison reflects the VLM data as a downstream user would receive it. The ground-truth and VLM lines coincide in every directory-year, on HISCAM (top) and on HISCLASS-12 (bottom), with each source's mean falling inside the other's 95% confidence interval throughout. The largest gap in mean HISCAM between the two sources is 0.07 points (1865), against the roughly fifty points that separate an unskilled laborer from a professional. A modern prior predicts a VLM mean below the ground-truth mean year after year; the figure shows none.

Dispersion by year. Equal means can still hide a compressed distribution, and compression is the sharpest form the prediction takes: pulling archaic occupations toward their modern forms lowers the dispersion of status. Figure 5 measures dispersion directly, with the Gini coefficient of HISCAM (top) and the Shannon entropy of the HISCLASS-12 distribution (bottom), each by directory-year. The ground truth and VLM lines again coincide in every year, and the widest HISCAM-Gini gap reaches only 0.0004 against a Gini near 0.12, two orders of magnitude smaller than the Gini itself. Because the VLM reproduces the status distribution itself, the matching coefficients in Table 6 cannot be an artifact of errors that happen to offset.

6 Cross-Corpus Record Linkage

Section 4 shows that VLM transcription reproduces the directory almost character by character, and Section 5 that it reproduces the status distributions and regression estimates drawn from a single corpus. Linking individuals across sources is one of the most common uses of manual transcriptions in economic history (Abramitzky et al., 2021), and it is the use where transcription errors are most severe: a single wrong character may break a match. In Table 7 we link the Munich directories to four independent lists and apply the identical matcher to the manual and the VLM transcription.

The four lists introduced in Section 2 span six decades and three administrative purposes: the casualty lists of the Franco-Prussian War and of WW1, the 1909 *Auto-Adressbuch*, and the 1933 *Gedenkbuch*. We match each list to the directory of its own year, and the matching rule follows the information each list carries. The two Verlustlisten record only a name: we match on an exact normalized surname and first name, and discard any external record whose name is not unique in the directory (Panel A). The *Gedenkbuch* and the *Auto-Adressbuch* also record a street address: we block on an exact street and house number, then accept the directory entry whose name is at most two character edits away from the external name (Panel B). Concordance is the share of records linked by at least one transcription that both transcriptions link to the same person.

The two Verlustlisten match on a name alone (Panel A). Concordance is 90.9% for the WW1 cohort, the largest and the only one precisely estimated.[18] A name-only match has no address to absorb a misread letter, so single-character Fraktur confusions (Appendix Table B.1) break the name key directly. The disagreement is close to undirected: GT-only and VLM-only links are of comparable magnitude in every cohort (223 against 183 in the WW1 cohort), with the manual transcription linking marginally more, in line with its slightly lower character-error rate.

The *Auto-Adressbuch* carries a street address, so the match blocks on address and tolerates a small name error (Panel B). Concordance is 97.9%, and no record is matched by the two transcriptions to different people: every disagreement is a record one transcription links and the other leaves unmatched, eight on the manual side and three on the VLM side.

The linked sample also reproduces the within-corpus invariance of Section 5 on a sample selected from outside the directory. Within their own buildings, the 1909 automobile owners outrank their neighbors: a regression of directory HISCAM on a car-owner indicator with building fixed effects places them 2.15 prestige points above the other residents of their address on the manual transcription (s.e. 0.80), and 2.11 on the VLM transcription (s.e. 0.79).[19] The status gradient survives the transition from a within-directory to an externally selected sample.

The *Gedenkbuch* reaches a concordance of 97.2%, close to the *Auto-Adressbuch* despite the denser 1933 directory. Here, too, no record is matched to two different people. The disagreements are again one-sided links, twenty-three on the manual side and ten on the VLM side, a count the larger external list and the denser directory inflate relative to the 1909 cohort.

The cross-corpus concordance gradient is informative about the matcher, not the transcription. The address block in Panel B absorbs the single-character Fraktur misreads of Panel A. The remaining 2.1 percentage-point gap from perfect concordance in the *Auto-Adressbuch* is one-sided links, records one transcription clears and the other leaves just outside the matching tolerance, split evenly between the two transcriptions. The four cohorts span 1870 to 1933 and both production regimes,

[18] The 1870/71 figure of 97.5% rests on 39 matched records and is too thin to weigh against the WW1 number.

[19] The regression runs on each transcription's own linked car owners with a valid HISCAM score (308 on the manual transcription, 301 on the VLM), across the roughly 270 buildings that contain a car owner and at least one other resident; standard errors are clustered by building.

full-page extraction before 1888 and column-split extraction from 1888 onward, so the result does not depend on a single year or pipeline configuration. A researcher applying either transcription to an external list would link the same records to the same people.

7 Conclusion

In this paper we establish that VLM transcription supports the three principal uses of digitized directories at rates indistinguishable from manual transcription: recovering the records and their fields, drawing distributional and regression-based inference from them, and linking them to outside sources. A vision-language model's modern prior, pulling historical titles toward their modern forms, would show up three ways in the status distribution: a lower mean, a narrower dispersion, or an attenuated gradient between groups. With 3.4 million aligned pairs, bias too small to shift inference would still be visible. None of the three appears. At €10 against €4,650 per volume, scale ceases to be the binding constraint.

The result rests on features that are not unique to Munich: printed text, tabular layout, a Latin-script language. Other features are specific to this corpus (Fraktur typeface, German, one particular VLM), and whether the invariance carries to handwritten records, non-Latin scripts, less structured layouts, or other VLM families is an open question. The methodological point generalizes more cleanly: character error rate is a necessary check but not a sufficient one, since a transcription can be accurate on average and still distort the inferences drawn from it. The field should validate VLM-transcribed corpora on estimate invariance directly, not on character accuracy alone.

References

Abramitzky, R., L. Boustan, K. Eriksson, J. Feigenbaum, and S. Pérez (2021). Automated Linking of Historical Data. *Journal of Economic Literature 59*(3), 865–918.

Albers, T. N. and K. Kappner (2023). Perks and Pitfalls of City Directories as a Micro-Geographic Data Source. *Explorations in Economic History 87*, 101476.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023). Prediction-powered inference. *Science 382*(6671), 669–674.

Bäcker-Peral, V., V. Meursault, and C. Severen (2025). Can LLMs Credibly Transform the Creation of Panel Data from Diverse Historical Tables? *arXiv preprint arXiv:2505.11599*.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024). Inference for Regression with Variables Generated by AI or Machine Learning. *arXiv preprint arXiv:2402.15585*. Also Cowles Foundation DP 2421 and CEPR DP 19115.

Bühler, M., D. Cantoni, and M. Weigand (2024). City Directories as a Source of Historical Microdata: Progress Report. Mimeo, LMU Munich.

Dahl, C. M., T. Johansen, and C. Vedel (2024). Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE. *arXiv preprint arXiv:2402.13604*.

Egami, N., M. Hinck, B. M. Stewart, and H. Wei (2023). Using Imperfect Surrogates for Downstream Inference: Design-Based Supervised Learning for Social Science Applications of Large Language Models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 68589–68601.

Greif, G., N. Griesshaber, and R. Greif (2025). Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. *arXiv preprint arXiv:2504.00414*.

Griesshaber, N. and J. Streb (2025). Multimodal LLMs for Historical Dataset Construction from Archival Image Scans: German Patents (1877–1918). *arXiv preprint arXiv:2512.19675*.

Humphries, M., L. C. Leddy, Q. Downton, M. Legacé, J. McConnell, I. Murray, and E. Spence (2025). Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents. *Historical Methods: A Journal of Quantitative and Interdisciplinary History 58*(3), 175–193.

Inoue, K. (2025). Context-Independent OCR with Multimodal LLMs: Effects of Image Resolution and Visual Complexity. *arXiv preprint arXiv:2503.23667*.

Karamolegkou, A., N. Angleraud, B. Sagot, and T. Clérice (2026). Reading or Guessing? Visual Grounding Failures of Vision-Language Models for OCR in Ancient Greek Editions. *arXiv preprint arXiv:2605.27750*.

Kim, S., J. Baudru, W. Ryckbosch, H. Bersini, and V. Ginis (2025). Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. *arXiv preprint arXiv:2501.11623*.

Lambert, P. S., R. L. Zijdeman, M. H. Van Leeuwen, I. Maas, and K. Prandy (2013). The Construction of HISCAM: A Stratification Scale Based on Social Interactions for Historical Comparative Research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History 46*(2), 77–89.

Lee, K.-i., M. Kim, S. Yoon, M. Kim, D. Lee, H. Koh, and K. Jung (2025). VLind-Bench: Measuring Language Priors in Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.

Levchenko, M. (2025). Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities. In *Proceedings of the First Workshop on NLP and Language Models for Digital Humanities (LM4DH 2025), associated with RANLP 2025*, pp. 75–85. INCOMA Ltd.

Ludwig, J., S. Mullainathan, and A. Rambachan (2025). Large Language Models: An Applied Econometric Framework. NBER Working Paper 33344, National Bureau of Economic Research. Preprint arXiv:2412.07031.

Luo, X., Z. Shinnick, N. Griesshaber, Y. Wang, J. Yu, F. Shi, P. Torr, and Y. Lu (2026). Pretraining Language Models on Historical Text. *arXiv preprint arXiv:2606.02991*.

van Leeuwen, M. H. and I. Maas (2011). *HISCLASS: A Historical International Social Class Scheme*. Leuven: Leuven University Press.

van Leeuwen, M. H., I. Maas, and A. Miles (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.

Verhoeven, F., T. Magne, and O. Sorkine-Hornung (2023). UVDoc: Neural Grid-based Document Unwarping. In *SIGGRAPH ASIA, Technical Papers*.

Vesalainen, A., E. Mäkelä, L. Ruotsalainen, and M. Tolonen (2026). Error Patterns in Historical OCR: A Comparative Analysis of TrOCR and a Vision-Language Model. *arXiv preprint arXiv:2602.14524*.

Figures and Tables

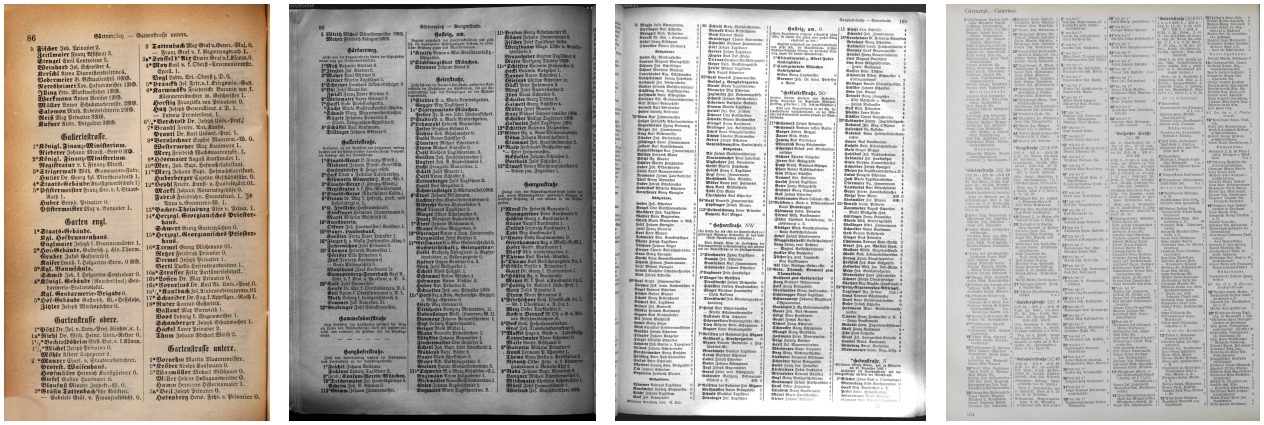


Figure 1: Evolution of directory page layouts: 1875 (2 columns), 1888 (3 columns), 1910 (4 columns), 1933 (6 columns).

Notes: Example pages from four directories in the corpus, drawn from the houses section. The four years span the layout transitions referenced in Section 4.2: two columns at the start of the houses-section series, three from 1888, four from the turn of the century, six from the 1930s. Density per page rises monotonically with column count.

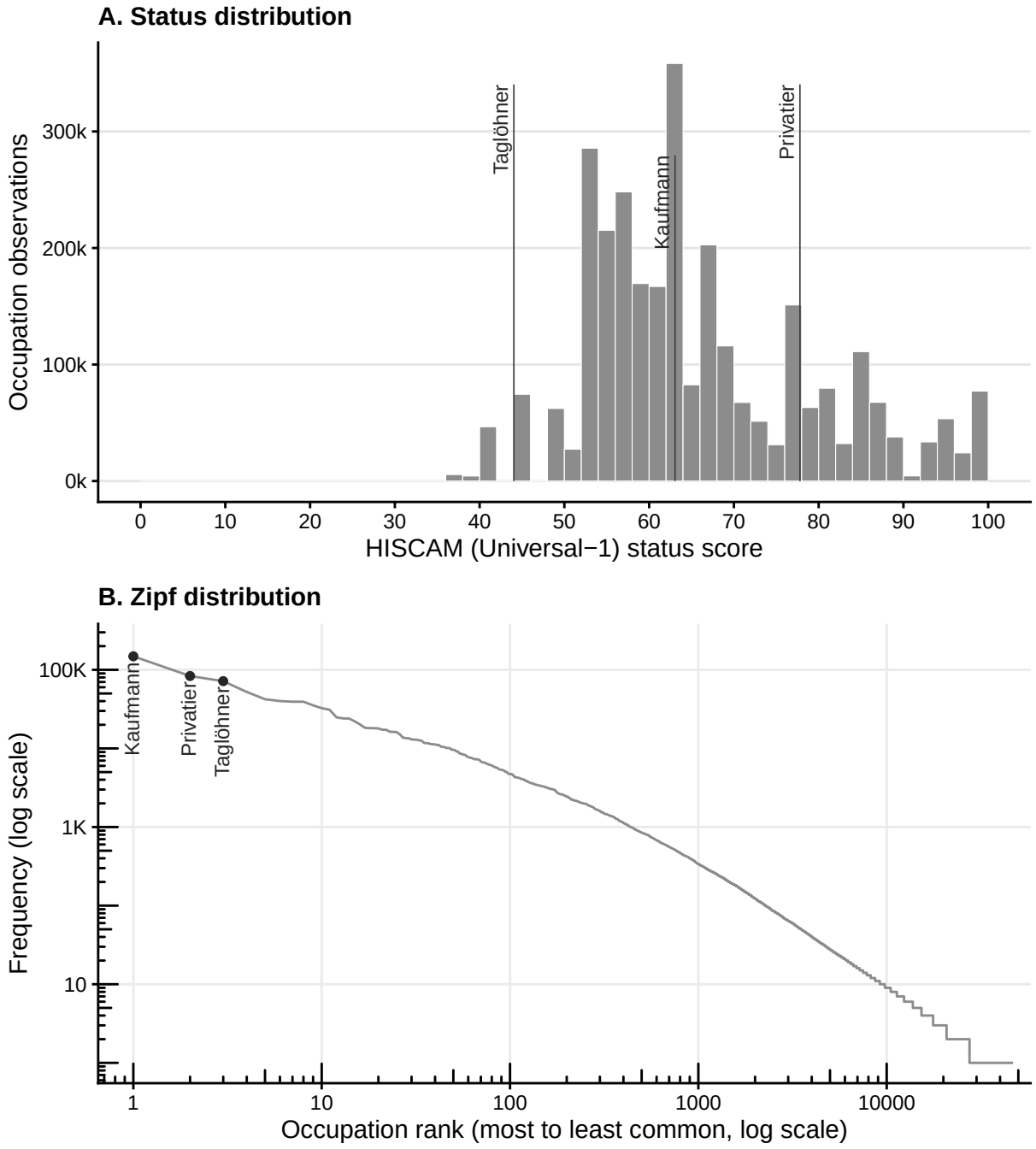
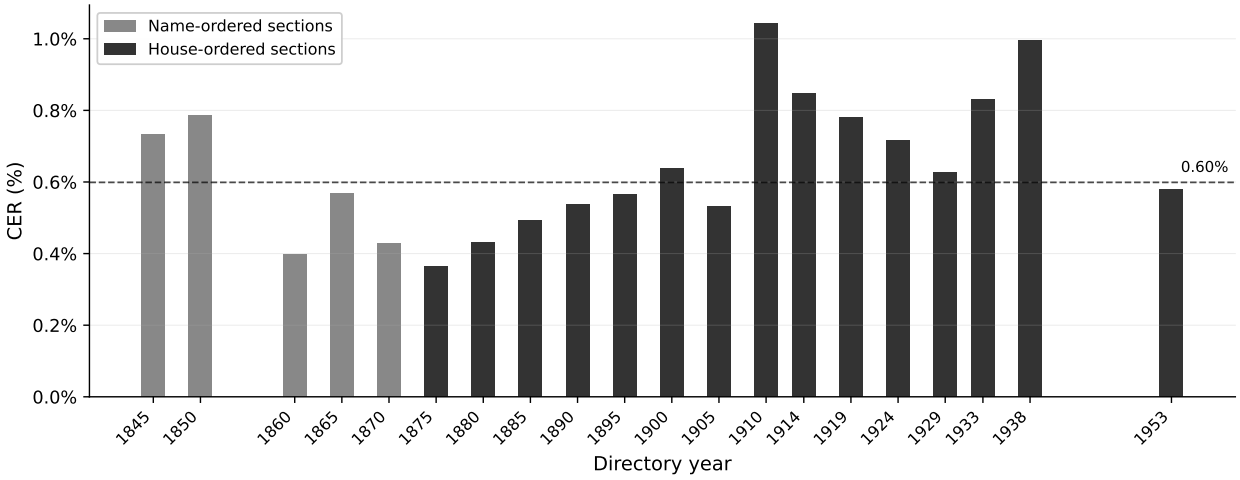


Figure 2: Occupational status distribution and occupation frequency

Notes: Panel A: distribution of HISCAM-U1 occupational status across all classifiable entries, weighted by occurrence; the three most frequent occupations (*Taglöhner*, *Kaufmann*, *Privatier*) are marked and span the low, middle, and high of the scale. Panel B: occupation frequency against rank on log–log axes; the same three occupations sit at ranks one to three. Munich directories, 1845–1953.

Figure 3: Transcription CER over time



Notes: Entry-level character error rate (CER) per directory-year, in percent, computed over all row-aligned entries. Dashed line indicates the unweighted mean across directory-years (0.60%). Dark bars: house-ordered sections (1875–1953); grey bars: name-ordered sections (1845–1870).

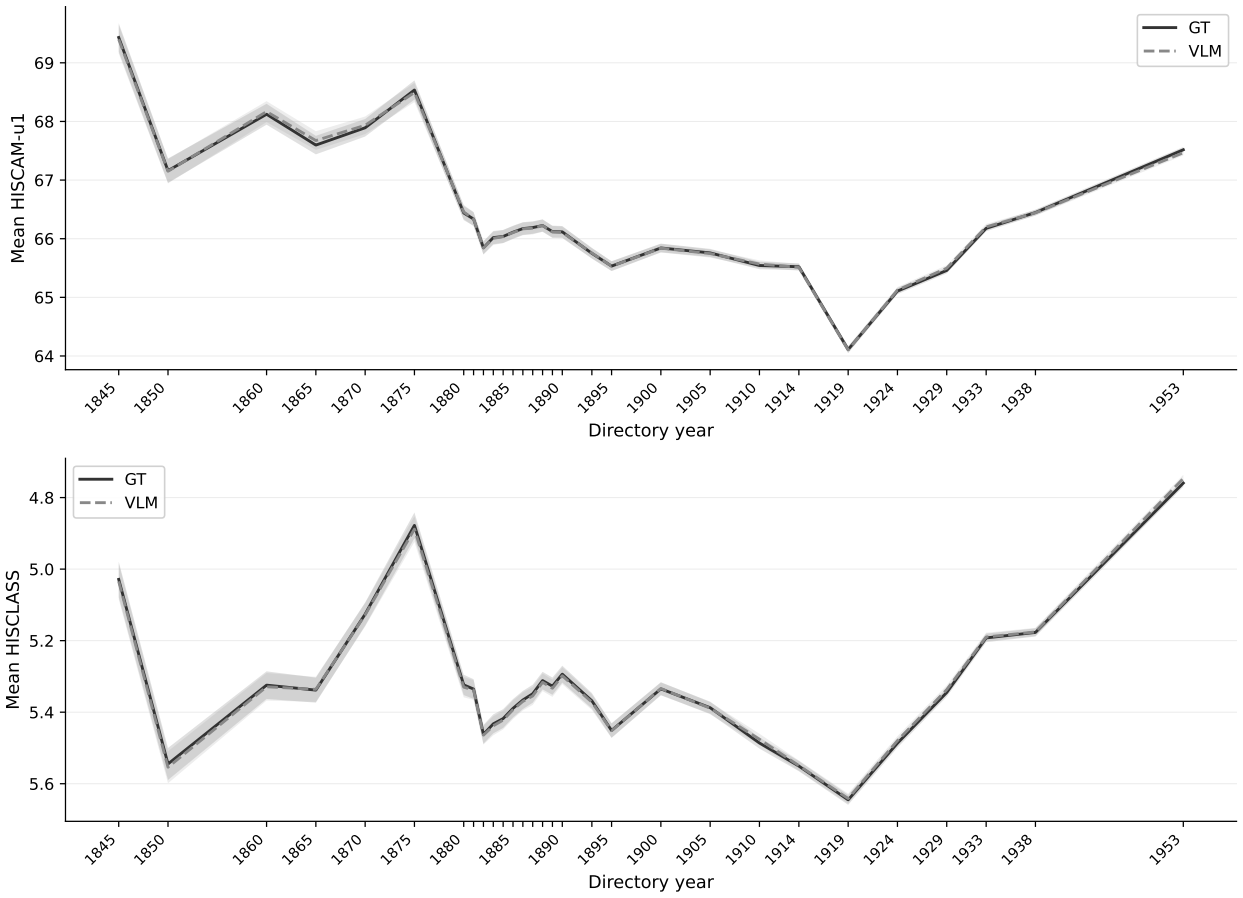


Figure 4: Average HISCAM and HISCLASS by Year: GT vs. VLM

Notes: Mean HISCAM (top) and mean HISCLASS (bottom, inverted so higher = higher status) per directory-year for manually transcribed (GT, solid) and VLM-transcribed (dashed) records, with 95% confidence intervals. Sample: all entries with a valid HISCAM or HISCLASS assignment, 1845–1953.

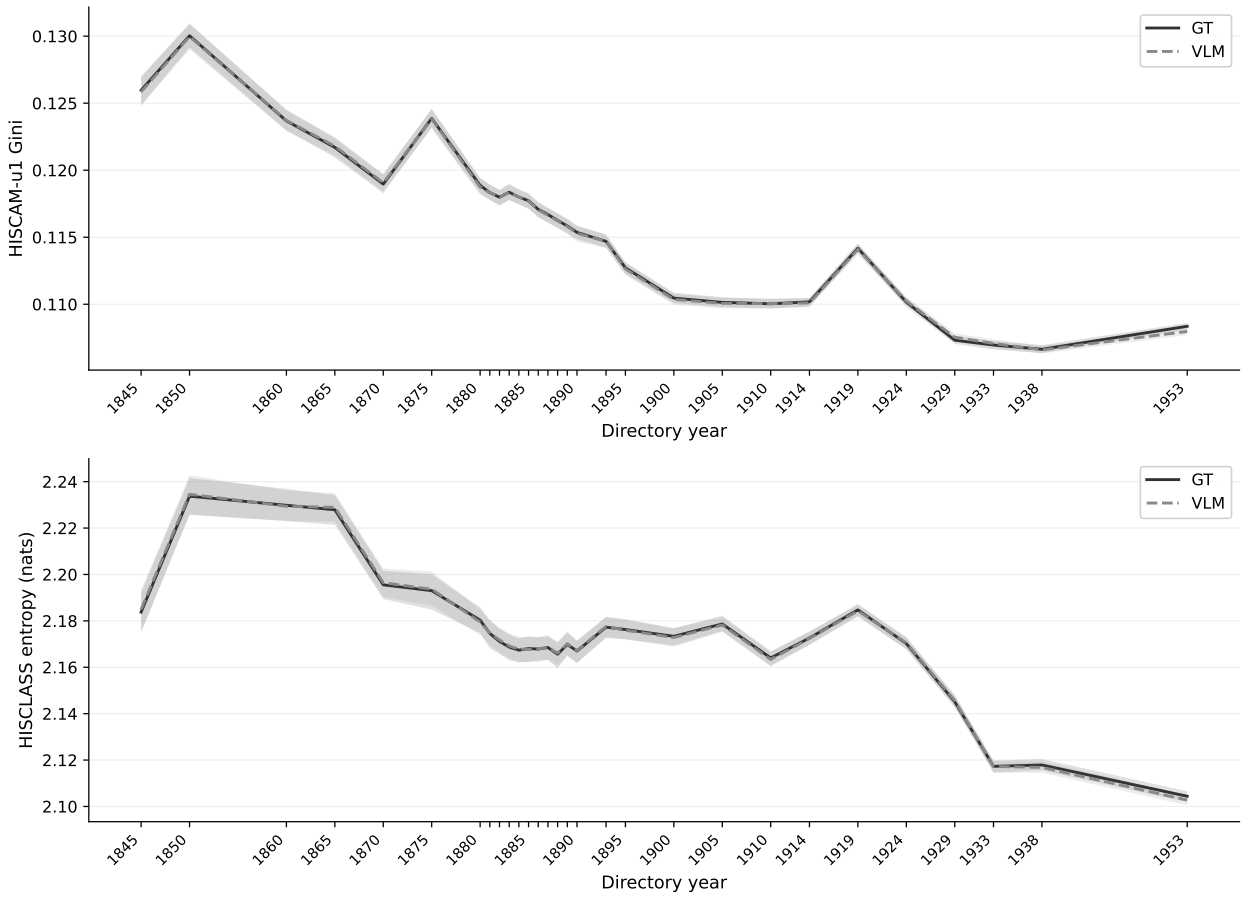


Figure 5: HISCAM Gini and HISCLASS entropy by year: GT vs. VLM

Notes: Per-directory-year HISCAM Gini coefficient (top) and Shannon entropy in nats over HISCLASS-12 categories (bottom) for manually transcribed (GT, solid) and VLM-transcribed (dashed) records, with bootstrap 95% confidence intervals ($B = 500$). Companion to Figure 4: where mean status is indistinguishable across sources, here we test whether the *distribution* of status is preserved. Sample: all entries with a valid HISCAM (top) or HISCLASS-12 (bottom) assignment, 1845–1953.

Table 1: Coverage: VLM vs. Ground Truth

# GT entries	3,442,200
# VLM entries	3,439,661
# Aligned	3,422,125 (99.4%)
# Missed (GT only)	20,075 (0.6%)
# Extra (VLM only)	17,536 (0.5%)

Notes: Row-aligned evaluation of VLM transcription against manual ground truth, pooled across 29 directory-years (1845–1953). Aligned: GT entries matched to a VLM counterpart. Missed: GT entries with no VLM match. Extra: VLM entries with no GT match. Percentages are relative to GT entries (Aligned, Missed) or VLM entries (Extra).

Table 2: Field Accuracy: VLM vs. Ground Truth

	Lastname	Firstname	Occupation	Number	Person	Entry
Exact match	94.6%	97.2%	93.4%	99.7%	90.2%	89.2%
Fuzzy match	95.6%	97.2%	96.1%	99.7%	98.4%	98.6%
CER	0.015	0.032	0.019	0.002	0.007	0.007

Notes: Per-field accuracy among the 3,422,125 aligned pairs, pooled across 29 directory-years (1845–1953). Field columns are Lastname, Firstname, Occupation, and House Number; aggregate columns are Person (concatenated lastname, firstname, and occupation) and Entry (full entry string in directory order). Exact match: share of pairs with identical normalized strings. Fuzzy match: share of pairs whose normalized Levenshtein distance does not exceed 0.1, exact matches included. CER: mean character error rate, the Levenshtein distance over the GT string length. Field rates are weighted by aligned pairs per directory-year.

Table 3: Per-field accuracy: Full-Page vs. Column-Split (average over 1887–1905)

	Exact Match (%)			CER		
Field	Full Pg.	Col. Split	Δ	Full Pg.	Col. Split	Δ
Coverage	99.2%	99.7%	+0.4%	—	—	—
Lastname	92.3%	93.6%	+1.4%	0.0199	0.0164	−0.3%
Firstname	97.8%	98.5%	+0.7%	0.0291	0.0171	−1.2%
Occupation	94.8%	95.6%	+0.8%	0.0135	0.0098	−0.4%
Street	97.4%	98.1%	+0.7%	0.0074	0.0059	−0.2%
House No.	99.6%	99.5%	−0.1%	0.0030	0.0040	+0.1%
Person	87.7%	89.6%	+1.9%	—	—	—
Entry CER	—	—	—	0.0132	0.0097	−0.4%

Notes: Averages over eight directories (houses section, 1887–1905). Exact Match: share of aligned pairs with identical normalized strings; the per-field rows require only that field to match, whereas *Person* requires every field of the entry to match. *Person* and *Entry CER* are the exact-match rate and the character error rate of the same full entry (all fields concatenated). Coverage: share of GT entries successfully aligned. CER: mean character error rate. Δ is column-split minus full-page; negative CER Δ indicates improvement.

Table 4: Adjudicated disagreements: who is correct, by field

		Adjudicated correct (%)		
Field	Disagreements	Manual	VLM	Neither
All disagreements	495	71.3	24.6	4.0
Last name	241	84.6	11.6	3.7
First name	54	64.8	25.9	9.3
Occupation	247	59.9	34.4	5.7

Notes: The 495 adjudicated GT–VLM person-string disagreements that carry a valid HISCAM score on both sides, split by the field in which the disagreement falls (a pair may disagree in more than one field). For each field, the columns report the share of disagreements adjudicated manual-correct, VLM-correct, and neither-correct. Last-name disagreements are resolved in the manual transcription's favor far more often than occupation disagreements, where the VLM is correct in a third of cases.

Table 5: Disagreement direction

	Gender		Status	
	All	Excl. Mar↔Max	HISCAM	HISCLASS
N (directional flips)	2,803	1,520	14,062	9,998
Share of total (%)	0.09	0.05	0.45	0.32
F→M / VLM lower[a]	2,090	828	6,637	5,112
Share (%)	74.6	54.5	47.2	51.1
M→F / VLM higher[a]	713	692	7,425	4,886
Share (%)	25.4	45.5	52.8	48.9

Notes: Sample: aligned GT–VLM person pairs. Two column blocks. Gender (cols. 1–2): pairs where the person string differs and both sides have a classifiable gender; column 1 covers all such pairs, column 2 excludes the Fraktur *r*→*x* confusion in abbreviated female first names (*Mar.*↔*Max.*). Status (cols. 3–4): pairs where the occupation string differs and both sides have a classified HISCAM or HISCLASS score. Rows report the count of directional flips, its share of all gender- or status-classifiable pairs, and the within-block split into VLM-lowers and VLM-raises (gender block: F→M vs. M→F). [a]For status columns: "lower" = HISCAM $\Delta < 0$ or HISCLASS $\Delta > 0$ (VLM assigns lower status); "higher" = the opposite direction.

Table 6: Coefficient invariance

	HISCAM			HISCLASS		
	GT	VLM	Stacked	GT	VLM	Stacked
Female	1.894^{***}	1.888^{***}	1.894^{***}	1.078^{***}	1.079^{***}	1.078^{***}
	(0.370)	(0.371)	(0.370)	(0.081)	(0.082)	(0.082)
VLM	—	—	0.001	—	—	0.000
			(0.002)			(0.000)
Female × VLM	—	—	−0.006	—	—	−0.000
			(0.009)			(0.003)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	3,099,366	3,099,366	6,198,732	3,099,413	3,099,413	6,198,826
Within R^2	0.004	0.004	0.004	0.021	0.021	0.021

Notes: Outcome: HISCAM (cols. 1–3) or HISCLASS (cols. 4–6). The HISCLASS-12 scale is reversed (higher = higher status) so that both outcomes share the HISCAM orientation; a positive Female coefficient indicates women hold higher-status occupations. Sample: aligned GT–VLM pairs where both sides have a non-missing SES score and a classifiable (male or female) gender; $N_{\text{stacked}} = 2 \times N_{\text{GT}}$. Each row carries its own source's Female indicator: GT rows use the GT-side classification, VLM rows use the VLM-side classification. The interaction $\hat{\delta}$ therefore absorbs both channels through which the VLM could bias a downstream gender gap — the status (HISCAM/HISCLASS) channel and the gender-misattribution channel — so that the GT and VLM columns describe what a researcher using each data source alone would estimate. SE clustered by year. Stars: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 7: Cross-corpus linkage concordance: VLM vs. Ground Truth

	N_{ext}	M_{GT}=1	M_{VLM}=1	Both	Diff. entry	GT-only	VLM-only	Neither	Concord.
Panel A: name-only match (last + first-token; cohorts lack a street address)									
1870/71 Verlustlisten (1870)	352	40 (11.4%)	39 (11.1%)	39	0	1	0	312	97.5%
WW1 Verlustlisten (1914)	20,441	4,292 (21.0%)	4,252 (20.8%)	4,069	0	223	183	15,966	90.9%
Panel B: address-blocked match (street + house no. + name Lev$\leq$2)									
Auto-Adressbuch (1910)	1,166	531 (45.5%)	526 (45.1%)	523	0	8	3	632	97.9%
Gedenkbuch (1933)	3,623	1,178 (32.5%)	1,165 (32.2%)	1,155	0	23	10	2,435	97.2%

Notes: Each row links one external list to the Munich directory of the stated year; Section 6 gives the matching rule, applied identically to the manually transcribed (GT) and the VLM-transcribed directory. Panel A holds the name-only cohorts, Panel B the address-blocked cohorts. N_{ext}: records in the external list. M_{GT}=1 and M_{VLM}=1: records the GT and the VLM directory link. Both: records both directories link, to the same person. Diff. entry: records both link, but to different entries. GT-only and VLM-only: records only one directory links. Neither: unlinked records. Concordance, the final column, is defined as Both/(Both + Diff. + GT-only + VLM-only). Samples: Munich-resident 1870/71 Bavarian and WW1 German Verlustlisten casualties, deduplicated per soldier; Munich-resident car owners in the 1909 *Auto-Adressbuch*, linked to the 1910 directory (the closest available); *Gedenkbuch* München victims with a Munich address active in 1933. In Panel A, external records whose name is not unique in the directory are counted under "Neither".

A Appendix — Corpus Inventory

Table A.1: Munich City Directories: Corpus Inventory

Year	Section	Pages	Cols	*N* (GT)	*N* (VLM)
1845	Names	178	2	15,780	15,724
1850	Names	338	2	22,062	21,982
1860	Names	297	2	27,232	27,192
1865	Names	361	2	33,100	32,928
1870	Names	381	2	36,006	35,875
1875	Houses	306	2	33,102	33,103
1880	Houses	456	2	52,744	52,784
1881	Houses	467	2	56,067	56,070
1882	Houses	492	2	61,572	61,564
1883	Houses	499	2	60,915	60,862
1884	Houses	521	2	62,311	62,263
1885	Houses	532	2	63,716	63,651
1886	Houses	544	2	65,415	65,314
1887	Houses	558	2	66,755	66,674
1888	Houses	336	3	69,506	69,405
1889	Houses	351	3	72,529	72,445
1890	Houses	368	3	75,658	75,594
1893	Houses	458	3	91,527	91,386
1895	Houses	527	3	103,630	103,596
1900	Houses	662	3	126,950	126,859
1905	Houses	516	4	148,154	148,097
1910	Houses	584	4	175,742	175,947
1914	Houses	705	4	208,259	207,868
1919	Houses	783	4	223,438	223,040
1924	Houses	900	4	254,892	254,419
1929	Houses	1,002	4	266,753	266,143
1933	Houses	603	6	289,562	288,958
1938	Houses	657	6	313,395	313,000
1953	Houses	797	6	365,428	366,918
Total	29 vols.	15,179		3,442,200	3,439,661

Notes: The 29 transcribed Munich directories, sampled at five-year intervals between 1845 and 1953, with annual sampling across 1880–1890 to assess matching quality across consecutive years. *Section*: names (alphabetical residents and businesses, used for 1845–1870) or houses (location-ordered, used for 1875 onward). *Pages*: page span of the transcribed section. *Cols*: number of text columns per page. *N* (GT): manually transcribed entries. *N* (VLM): VLM-extracted entries.

B Appendix — Error Mechanism

Table B.1: Most Frequent Character Substitutions

GT		OCR		GT→OCR	OCR→GT	% of subs
ℜ	R	𝔑	N	25,604	3,657	13.8%
ſ	f	ſ	s	9,166	7,255	7.7%
ß	tz	ß	ß	15,085	638	7.4%
𝔙	V	𝔅	B	9,880	4,038	6.6%
u	u	n	n	4,550	3,575	3.8%
ℜ	K	t	k	7,275	202	3.5%
r	r	x	x	4,409	2,288	3.2%
ß	ß	ſſ	ss	2,480	2,061	2.1%
ſ	s	ß	ß	3,040	1,295	2.0%
t	k	t	t	2,967	720	1.7%
t	t	l	l	1,731	1,630	1.6%
ff	ff	ſſ	ss	1,931	1,239	1.5%
ℜ	R	ℜ	K	2,387	696	1.5%
d	d	b	b	1,999	1,055	1.4%
e	e	c	c	1,753	1,218	1.4%

Notes: Top 15 substitution pairs by combined frequency, pooled across all 29 directory-years and all fields. Each side shows the Fraktur glyph beside its Latin transcription. The two count columns give the direction of the substitution: GT→VLM is the number of times the ground-truth character (first column) appears as the VLM character (second column), and VLM→GT is the reverse. % of subs is the pair's share of all substitution edits. Most pairs are symmetric Fraktur confusions; the glyphs make the visual similarity explicit (e.g. ℜ/𝔑 and 𝔙/𝔅 are near-identical in blackletter type). A few rows are ligature or orthographic substitutions rather than single-glyph look-alikes (e.g. the ß ligature read as ß, or ff/ff).

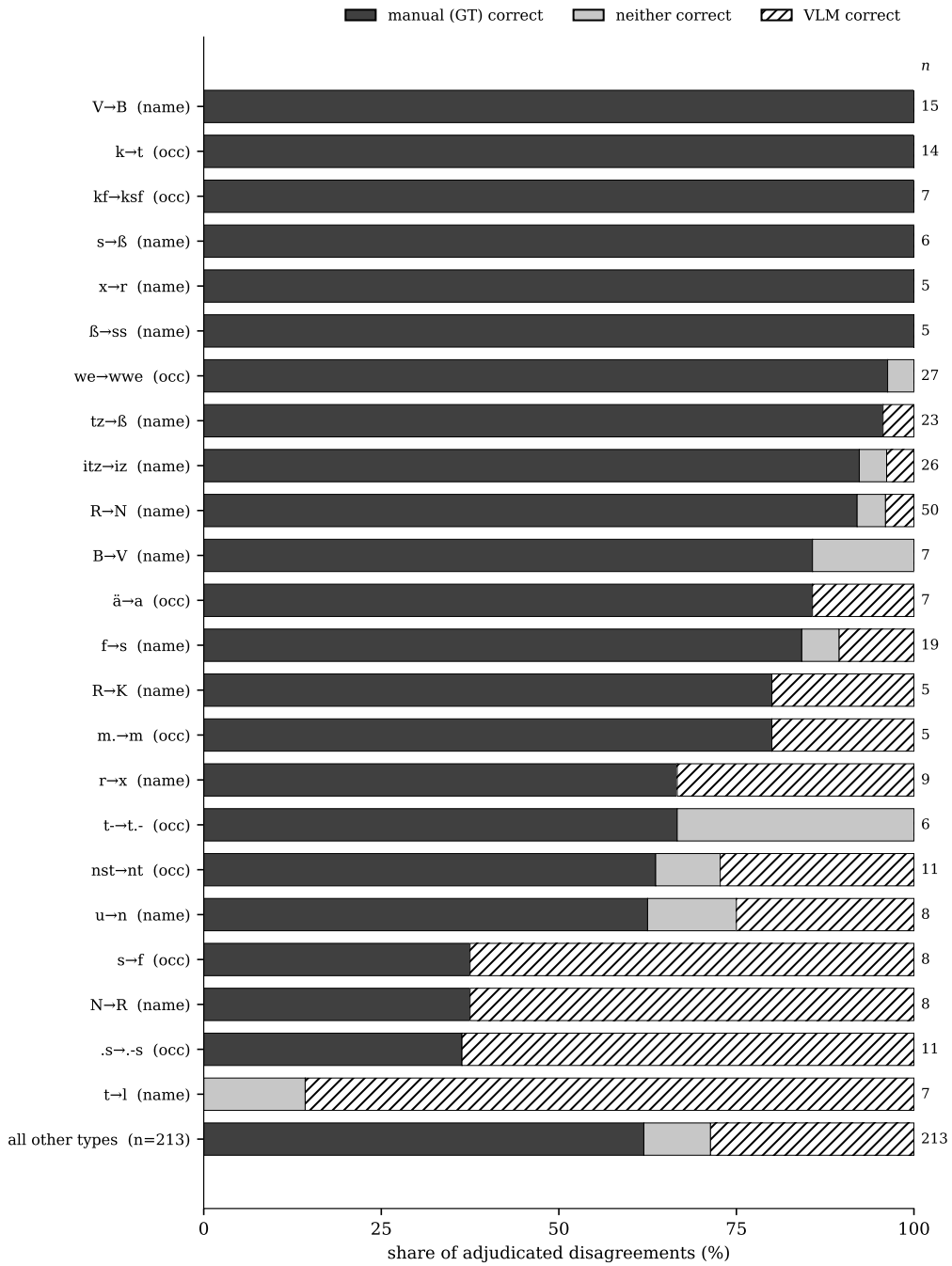


Figure B.1: Adjudicated disagreements: who is correct, by substitution type

Notes: Each bar is one character-substitution type among the 500 randomly sampled GT–VLM person-string disagreements, split into the share adjudicated manual-correct (dark), neither correct (grey), and VLM-correct (hatched). Rows are ordered by the manual-correct share; *n* is the number of adjudicated pairs of that type. "All other types" pools substitutions with fewer than five adjudicated pairs. See Table B.1 for the corpus-wide frequency of these substitutions.

Table B.2: Field-Level CER Decomposition: Misattribution vs. Genuine Error

	Lastname	Firstname	Occupation	Person
Character error rate (CER)				
Raw CER	0.015	0.032	0.019	0.007
Misattribution	0.004	0.022	0.008	—
Genuine	0.012	0.011	0.011	0.007
% from misattr.	24.0%	66.6%	40.2%	—
Exact match rate				
Raw	94.6%	97.2%	93.4%	90.2%
Adjusted	95.1%	98.8%	95.2%	—

Notes: Field-level CER mechanically penalizes tokens placed in the wrong field even when the page was read correctly (e.g., a personal title written in the occupation slot rather than the name slot); this table separates that misattribution component from genuine character error. A row is misattributed if the concatenated person string matches exactly but at least one individual field does not. Decomposition pooled across all 29 directory-years. Raw CER is the standard field-level mean; Misattribution is the CER contribution from misattributed rows; Genuine is the remainder. Adjusted exact match counts misattributed rows as correct. Person-level metrics shown for reference.

C Appendix — Column-Splitting Algorithm

For directories published from 1888 onward, we split each page image into individual columns before passing them to the VLM. The procedure has three stages: dewarping, separator detection, and column extraction.

Stage 1: Dewarping. Scanned book pages exhibit curvature from the binding. We straighten each image using *UVDoc* (Verhoeven et al., 2023), a neural grid-based dewarping model. A 50-pixel white border is added before dewarping to prevent the model from cropping text near the page edges.

Stage 2: Separator detection. We detect the $N-1$ vertical separator lines that divide a page into N columns. The algorithm operates in two phases. In the first phase, we isolate vertical structures via morphological filtering. The dewarped image is binarized using Otsu's method and then eroded and dilated with a tall, narrow kernel (1 px wide, $h/15$ px tall, where h is the image height). This suppresses text and horizontal rules while preserving structures that span a significant fraction of the page height. The resulting binary mask is projected onto the x-axis by summing pixel intensities along each column, producing a one-dimensional signal in which printed separator lines appear as sharp peaks. We identify candidate peaks using `scipy.signal.find_peaks` with permissive thresholds.

Each candidate peak receives a quality score that combines its height (intensity of the projection signal), its prominence (how much it stands out above the local baseline), and the inverse of its full width at half maximum. The width penalty is important because book-spine shadows produce broad dark bands that would otherwise score competitively with thin printed separator lines.

In the second phase, we evaluate all $\binom{K}{N-1}$ combinations of the top $K = 25$ candidates and score each combination on three weighted criteria: peak quality (50%), spacing regularity measured by the coefficient of variation of the resulting column widths (30%), and proximity of the separator positions to the uniformly spaced positions expected for N equal columns (20%). Hard constraints reject combinations in which adjacent separators are closer than a minimum fraction of the expected column width, or in which a separator falls too close to a page edge. If the strict constraints yield no valid combination, they are relaxed in two further tiers before a greedy fallback selects the highest-scoring peaks that satisfy the distance requirements.

Stage 3: Column extraction. Given the detected separator x-coordinates, the dewarped image is sliced into N column images. Each slice extends 20 pixels beyond its separator boundaries on both sides, so that characters straddling a separator line are not clipped.

Figure C.1 illustrates the procedure on a four-column page from the 1910 directory. Panel (a) shows the raw scan. Panel (b) displays the binary vertical mask after morphological filtering: the

three printed separator lines appear as bright vertical bands, while shorter fragments from the book spine and page edges are also visible but score lower in the subsequent selection step. The x-axis projection plotted below shows the corresponding one-dimensional signal; the three sharp peaks are clearly distinguishable from the lower, broader peaks produced by the book spine shadow and page edges. Panel (c) marks the three separator lines selected by the combinatorial scoring.

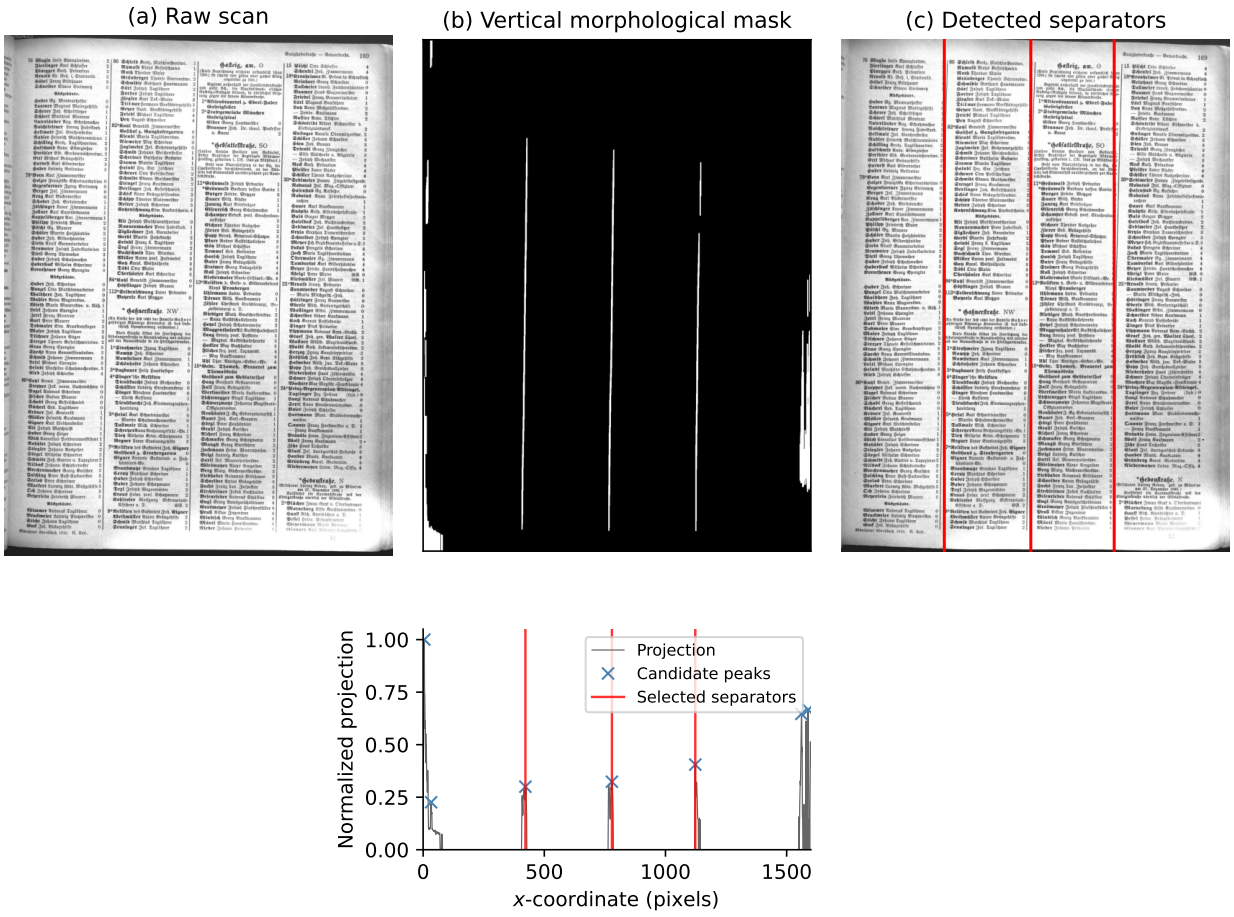


Figure C.1: Column-splitting pipeline on a four-column page (Munich 1910). (a) Raw scan. (b) Vertical morphological mask and its x-axis projection with candidate peaks and selected separators. (c) Detected separators.

D Appendix — Extraction Prompt

The prompts below are reproduced verbatim from `source/raw/directories/prompts.py`. We use two variants for the houses section: a full-page prompt for the two-column layouts (1875–1887) and a column-split prompt for the three- to six-column layouts (1888–1953). Both share the same TSV format and the same entry-logic block; they differ in how each handles the street-header context that the model must propagate through the page. Each prompt is followed by few-shot examples: two to four scans (full pages or single columns, matching the input format) paired with their gold-standard TSV transcriptions.

Houses, full-page (1875–1887)

```
Anbei ist eine Seite aus einem deutschen Adressbuch (19. Jh., Fraktur),
nach Straßen sortiert. Extrahiere *alle* Einträge als TSV.

**SCHEMA:** Strasse  Hausnummer  Besitzer  Nachname  Vorname  Beruf  Stock  Gebäude

**WICHTIG:** Wenn die Seite leer ist, keine Einträge enthält, oder nur
Überschriften/Seitenzahlen zeigt, gib NUR die Header-Zeile zurück ohne
Datenzeilen.

**Adressbuch-Format:**
Der Ausschnitt enthält (i) Straßenüberschriften, und (ii) Einträge
nach Hausnummer sortiert. Beide Arten von Einträgen müssen
extrahiert werden.
- (i) Straßenüberschriften.
    - Die Straßenüberschrift ist immer fett gedruckt.
    - Beispiel: "Musterstraße  -  -  -  -  -  -  -"
    - Beschreibungen der Lage einer Hausnummer sollen NICHT
      extrahiert werden (z.B "Zwischen 49 und 51 Hauptstraße"
      oder "58 an der Musterstraße")
    - Extrahiere den Seitenkopf als Straßenüberschrift in den
      ersten Eintrag der TSV.
- (ii) Hausnummer-Einträge.
  Beispiel: "-  5  1  Mustermann  Max  Musterberuf  1  RG"
    - Bei Hausnummer-Einträgen setzen wir das Feld "Straße"
      auf "-".

WICHTIG: Wenn der Ausschnitt leer ist, also keine
Straßenüberschriften oder Einträge enthält, gib eine leere
TSV-Datei zurück (nur den TSV header).

**Eintrags-Logik:**
- **Hausnummern:** Beginnen einen Block (z.B. "51", "5 1/2", "3a",
  "7d"). Einträge ohne Nummer gehören zur vorangestellten Nummer.
```

```
  Format: Hausnummer inklusive Brüche und Buchstabenzusätze
  (z.B: "5", "5a", "3c", "41b", "5 1/2", "9 1/3")
- **Nachname**: Bei einem Ditto-Zeichen bzw. Strich den Nachnamen
  des vorherigen Eintrags übernehmen. Bei "Derselbe" die Daten des
  vorherigen Eintrags übernehmen (außer Hausnummer).
- Im Adressbuch stehen oft **Institutionen/Geschäfte ohne
  Personennamen**. In diesem Fall sollte der Name in "Nachname"
  stehen und "Vorname" und "Beruf" mit "-" gefüllt werden, falls
  nicht vorhanden.
- **Besitzer:** Flag. "1" wenn "*" vor dem Nachnamen steht,
  sonst "0".
- **Titel-Zuordnung:** Manchmal ist nicht ganz klar, ob ein Titel
  zum Vornamen oder zum Beruf gehört. Dies hängt davon ab, ob der
  Titel die Person oder den Beruf modifiziert:
    - Titel die zu "Vorname" gehören: z. B. "Dr.", "v.", "von",
      "Frhr.", "Graf", etc.
    - Titel die zu "Beruf" gehören: z. B. "p.", "q.", "ehem.",
      "k.", "f.", "kgl.", "städt.", "prakt.", "Wwe.", etc.
- **Stock:** Die Zahl am Zeilenende ist das Stockwerk (z.B. "1",
  "2"). Falls nicht vorhanden, setze auf "-".
- **Gebäude:** RG, VG, HG, SG, r., etc., falls vorhanden. Sonst
  "-". In manchen Adressbüchern gibt es Zwischenüberschriften
  (z.B. "Rückgebäude", "I. Rückgebäude", "Eingang ...",
  "Aufgang X"). Diese gelten für alle folgenden Einträge dieser
  Hausnummer und gehören auch zu Spalte "Gebäude".
```

Houses, column-split (1888–1953)

The column-split prompt drops the Seitenkopf rule, since a single column has no page header, and instead describes the input as an excerpt (*Ausschnitt*). The entry-logic block is identical to the full-page variant above; the header reads:

```
Anbei ist ein Ausschnitt aus einem deutschen Adressbuch (19. Jh.,
Fraktur). Die Einträge sind nach Straße und Hausnummer sortiert.
Deine Aufgabe ist es, die Einträge als TSV zu extrahieren.

**TSV-Schema:** Strasse  Hausnummer  Besitzer  Nachname  Vorname
               Beruf    Stock       Gebäude



**Adressbuch-Format:**
Der Ausschnitt enthält (i) Straßenüberschriften, und (ii) Einträge
nach Hausnummer sortiert. Beide Arten von Einträgen müssen
extrahiert werden.
- (i) Straßenüberschriften.
    - Die Straßenüberschrift ist immer fett gedruckt.
```

```
    - Beispiel: "Musterstraße  -  -  -  -  -  -  -"
    - Beschreibungen der Lage einer Hausnummer sollen NICHT
      extrahiert werden (z.B "Zwischen 49 und 51 Hauptstraße"
      oder "58 an der Musterstraße")
- (ii) Hausnummer-Einträge.
  Beispiel: "-  5  1  Mustermann  Max  Musterberuf  1  RG"
    - Bei Hausnummer-Einträgen setzen wir das Feld "Straße"
      auf "-".

WICHTIG: Wenn der Ausschnitt leer ist, also keine
Straßenüberschriften oder Einträge enthält, gib eine leere
TSV-Datei zurück (nur den TSV header).
```

E Appendix — Page Density: Year-by-Year

Table E.1: Year-by-year comparison: Full-Page vs. Column-Split (1887–1905)

	Coverage (%)			Lastname Acc. (%)			Entry CER		
Year	Full Pg.	Col. Split	Δ	Full Pg.	Col. Split	Δ	Full Pg.	Col. Split	Δ
1887	99.6%	99.6%	±0.0%	93.1%	93.1%	±0.0%	0.0110	0.0110	±0.0%
1888	99.5%	99.7%	+0.2%	92.5%	93.8%	+1.3%	0.0104	0.0064	−0.4%
1889	99.2%	99.5%	+0.3%	92.1%	93.6%	+1.5%	0.0141	0.0079	−0.6%
1890	99.0%	99.7%	+0.6%	91.2%	92.9%	+1.7%	0.0169	0.0109	−0.6%
1893	99.3%	99.7%	+0.4%	92.2%	93.2%	+1.0%	0.0135	0.0159	+0.2%
1895	99.5%	99.8%	+0.3%	92.0%	93.4%	+1.4%	0.0132	0.0095	−0.4%
1900	99.1%	99.6%	+0.5%	91.3%	92.7%	+1.4%	0.0130	0.0089	−0.4%
1905	98.7%	99.7%	+0.9%	93.8%	96.3%	+2.5%	0.0137	0.0072	−0.7%
Overall	**99.2%**	**99.7%**	+0.4%	**92.3%**	**93.6%**	+1.4%	**0.0132**	**0.0097**	−0.4%

Notes: Row-aligned evaluation for the houses section. Coverage: share of GT entries successfully aligned. Lastname Acc.: exact match rate for the lastname field. Entry CER: character error rate of the full entry string (all fields concatenated in directory order, implicitly correcting for field-boundary misattributions), averaged across aligned pairs. Δ reports column-split minus full-page; negative CER Δ indicates improvement.

F Appendix — Per-Field Accuracy over Time

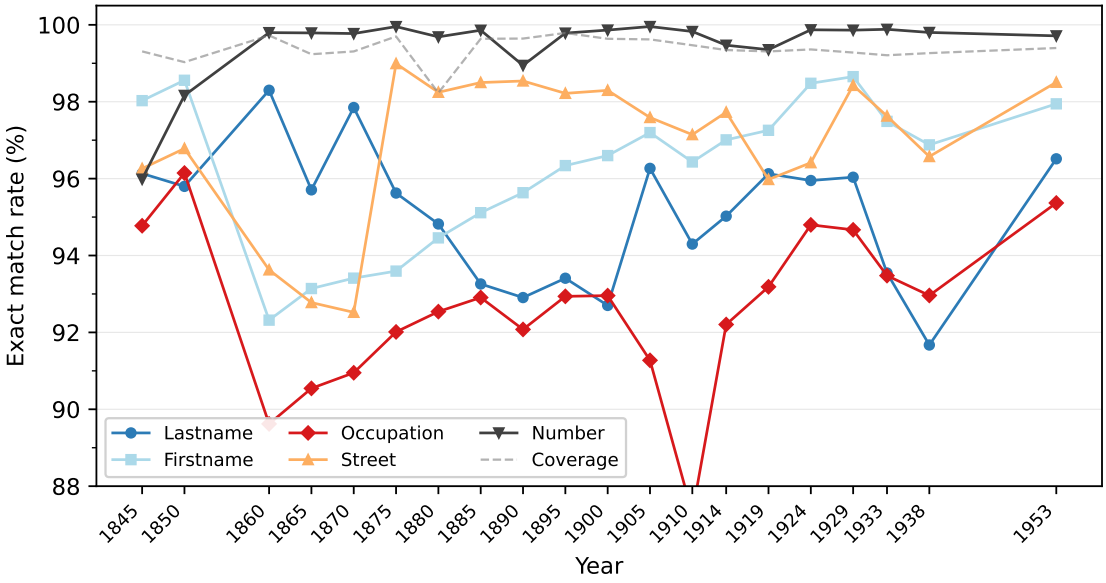


Figure F.1: Per-field exact match rates by year

Notes: Exact match rates for individual fields (Lastname, Firstname, Occupation, Street, Number) across all 29 directory-years, aggregated across name-ordered and house-ordered sections. Coverage (dashed line) is the share of GT entries successfully aligned to a VLM counterpart. Years with dense annual coverage (1880–1890) are thinned to five-year intervals for readability.

G Appendix — Downstream Coefficient Invariance

The headline result in Section 5 is that a researcher running the gender-status gradient on VLM transcription recovers the same coefficient as one running it on manual transcription. The headline test is conditional on a single downstream regression, on the full sample, with analytic year-clustered standard errors. This appendix subjects that test to four further checks, reported across Tables G.1–G.4 and Figure G.1. Each addresses a referee question the headline alone cannot answer.

Other downstream regressions. Table G.1 extends the test from the gender-status gradient to three further specifications a downstream user might run: HISCLASS instead of HISCAM, a white-collar elite-share indicator, and the reverse status-to-gender direction. In every case $\hat{\delta}$ is well under 1% of $\hat{\beta}_{\text{GT}}$, so the invariance is not specific to the headline regression.

Small-cluster inference. The stacked panel has 29 year clusters. A Wald t-statistic on year-clustered standard errors is asymptotically valid in the cluster dimension and can over-reject in small samples. Table G.2 replaces the analytic SE with a wild cluster bootstrap (Rademacher weights, null imposed, $B = 9{,}999$) and adds an entry-fixed-effects specification that identifies the interaction from within-pair GT-vs-VLM disagreement alone. Across both inference adjustments and both outcomes the interaction remains within rounding distance of zero.

Mass at the tails. A modern prior pulls archaic readings toward their common modern forms (Vesalainen et al., 2026; Karamolegkou et al., 2026), and these archaic occupations sit in the tails of the status distribution (rare in the corpus, Section 2.2). Any compression toward the center would thin the tails. The cleanest direct test asks whether the VLM produces the same mass at the tails of HISCAM as the manual transcription. Table G.3 reports the share of aligned pairs in five regions defined on absolute HISCAM-U1 thresholds: HISCAM < 40 (below unskilled laborer), HISCAM < 50 (working-class), HISCAM $\in [50, 80]$ (center), HISCAM > 80 (professional), and HISCAM > 90 (elite). For each region, shares are computed separately on the GT-side and the VLM-side score using the same threshold. The compression mechanism predicts GT share $>$ VLM share at the tails and GT share $<$ VLM share in the center. The data show no such pattern: the GT and VLM shares coincide to three decimal places in every region.

Tail-conditional regressions. Table G.4 re-runs the headline HISCAM specification on subsamples defined by the GT-side score, restricting to the bottom 5%, bottom 10%, top 5%, and top 10% of the manual-transcription HISCAM distribution. Within each tail the gender gap $\hat{\beta}$ differs from the full-sample value, and on two of the four tails $\hat{\delta}$ is statistically distinguishable from zero: the bottom 10% (0.017, $p < 0.001$) and the top 5% (0.073, $p < 0.05$). Neither shifts the conclusion a researcher would draw inside the tail. The bottom-10% estimate of 0.017 HISCAM points compares

with a within-tail $|\hat{\beta}_{\text{GT}}|$ of 1.35: the VLM-induced shift is 1.2% of the gender gap a researcher would estimate on that subsample. In the top 5% there is no gap to shift; both transcriptions place the within-tail gap at zero ($\hat{\beta}_{\text{GT}} = -0.04$, $\hat{\beta}_{\text{VLM}} = 0.02$, neither distinguishable from zero). On the remaining two tails $\hat{\delta}$ is itself indistinguishable from zero (bottom 5%: 0.009, top 10%: -0.036). Statistical significance on 155,000 to 337,000 pairs per tail is not the same as economic significance; on every tail the VLM transcription supports the same within-tail conclusion as the manual one.

Rare-occupation strings. The sharpest version of the same test bypasses HISCAM entirely and asks the question at the level of the raw occupation string. Figure G.1 plots, for each GT-frequency value N from 2 to 50, the share of GT entries with a VLM-side string that matches the GT-side string exactly (solid) or within a normalized Levenshtein distance of 0.1 (dashed). The exact-match rate rises from 79.4% at $N = 2$ to roughly 93% at $N = 50$, with the fuzzy-match line consistently 4–7 percentage points above. The reference rate for tokens appearing more than 100 times in GT is 97.8% exact and 98.8% fuzzy. So even at the rarest end, four GT occupation strings in five survive transcription exactly, and nine in ten survive within a single edit. The mechanism leaves a signature at the string level, but small enough that it propagates to none of the downstream tests above.

Taken together, the four checks support the invariance claim against the readings a referee can reasonably advance. A finding that holds across additional downstream specifications, under small-cluster inference, at the tails of the status distribution, and at the rarest end of the occupation distribution is not an artifact of any one specification choice.

Table G.1: Coefficient invariance across downstream regressions

	$\hat{\beta}_{\text{GT}}$	$\hat{\beta}_{\text{VLM}}$	$\hat{\delta} = \hat{\beta}_{\text{VLM}} - \hat{\beta}_{\text{GT}}$	$\lvert\hat{\delta}\rvert / \lvert\hat{\beta}_{\text{GT}}\rvert$ (%)	N pairs
HISCAM gender gap	1.8943^{***}	1.8883^{***}	-0.0056	0.30	3,099,366
	(0.3697)	(0.3711)	(0.0089)		
HISCLASS gender gap	1.0784^{***}	1.0786^{***}	-2.66×10^{-6}	0.00	3,099,413
	(0.0815)	(0.0823)	(0.0029)		
Elite-share gender gap	0.1577^{***}	0.1580^{***}	3.02×10^{-4}	0.19	3,099,413
	(0.0095)	(0.0097)	(5.89×10^{-4})		
Status→Female	0.0019^{***}	0.0019^{***}	-1.42×10^{-5}	0.74	3,099,366
	(3.51×10^{-4})	(3.51×10^{-4})	(8.19×10^{-6})		

Notes: Each row is one downstream regression $Y \sim X$ estimated three ways on the aligned-pair sample: (i) GT-only $Y_{\text{GT}} \sim X_{\text{GT}}$; (ii) VLM-only $Y_{\text{VLM}} \sim X_{\text{VLM}}$; (iii) stacked specification $Y_i = \alpha + \beta X_i + \gamma\, VLM_i + \delta\,(X_i \times VLM_i) + \lambda_t + \varepsilon_i$, from which $\hat{\delta}$ is read off. Year fixed effects, year-clustered SEs throughout. $\hat{\delta} = 0$ is the null of inferential equivalence: a downstream researcher draws the same coefficient on either dataset. Rows: (1) HISCAM ∼ Female; (2) HISCLASS ∼ Female (HISCLASS-12 reversed so higher = higher status, matching HISCAM); (3) $\mathbb{1}[\text{HISCLASS} \leq 3]$ ∼ Female (white-collar elite share, native top-3 classes); (4) Female ∼ HISCAM (status→gender direction). Gender is side-specific (GT-side on the GT leg, VLM-side on the VLM leg). Stars: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table G.2: Wild-cluster bootstrap robustness for the Female × VLM interaction

	HISCAM			HISCLASS		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	1.8941***	—	—	1.0785***	—	—
	[0.4188]***			[0.0938]***		
Female × VLM	−0.0056	0.0464	0.0464	−0.0000	0.0085	0.0085
	[0.0098]	(0.0530)	[0.0635]	—	(0.0163)	[0.0233]
Year FE	Yes	—	—	Yes	—	—
Entry FE	No	Yes	Yes	No	Yes	Yes
Cluster SE	—	Year	—	—	Year	—
Wild bootstrap SE	Year	—	Year	Year	—	Year
N (obs)	6,198,732	6,198,732	6,198,732	6,198,826	6,198,826	6,198,826
N (pairs)	—	3,099,366	3,099,366	—	3,099,413	3,099,413

Notes: Robustness for the Female × VLM interaction in Table 6, in three specifications per outcome. (A) Year fixed effects with wild cluster bootstrap (WCB) standard errors clustered by year; both the Female main effect and the Female × VLM interaction are identified and bootstrapped (Rademacher weights, null imposed, $B = 9{,}999$, year clusters). (B) Entry fixed effects, one per aligned GT–VLM pair, plus year fixed effects, with analytic year-clustered standard errors; the Female main effect is absorbed by the entry effect, and Female × VLM is identified off the within-pair contrast (pairs where GT and VLM disagree on gender). For computational tractability, columns (B) and (C) estimate the algebraically equivalent first-differenced specification, $\Delta y_e = \gamma + \delta\,\Delta\text{Female}_e + \nu_e$, on one row per aligned pair; the coefficient on ΔFemale equals the Female × VLM interaction in the stacked entry-FE regression. N (obs) is the equivalent number of stacked observations, 2× aligned pairs; N (pairs) is the number of aligned pairs that actually enter the diff regression. (C) Same identification as (B), with WCB standard errors on the within-pair contrast. Brackets $[\cdot]$ report WCB standard errors; parentheses $(\cdot)$ report analytic cluster standard errors. The WCB confidence-interval inversion for the HISCLASS Female × VLM cell in column (4) failed numerically: the point estimate rounds to 0.000, so the bootstrap test statistic is degenerate at the null and `fwildclusterboot`'s CI root-finder cannot bracket a finite interval. The bootstrap p-value itself is well-defined and the substantive conclusion is unaffected. Stars: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table G.3: Tail share invariance: HISCAM mass at the tails

	GT share	VLM share
HISCAM < 40	0.004	0.004
HISCAM < 50	0.053	0.053
$50 \leq$ HISCAM ≤ 80	0.772	0.772
HISCAM > 80	0.175	0.175
HISCAM > 90	0.065	0.065
N (pairs)	3,099,366	

Notes: Share of aligned pairs whose HISCAM falls in each region of the status distribution, computed separately on the GT-side and the VLM-side score using the same absolute HISCAM-U1 thresholds. The five regions anchor on the substantive HISCAM-U1 onboarding points: unskilled laborer $\approx$ 30 (so HISCAM $<$ 40 covers the bottom), craftsman $\approx$ 50, professional $\approx$ 80, and elite occupations above 90. The compression mechanism predicts GT share strictly larger than VLM share at the tails, and strictly smaller in the center. The shares are computed on 3,099,366 aligned pairs with non-missing HISCAM on both sides; we report rounded shares rather than within-pair differences because the latter are below 0.0001 in absolute value across every region, well below the precision the rounded shares can convey.

Table G.4: Coefficient invariance on tails of the HISCAM distribution

	Bottom tails		Top tails	
	Bottom 5%	Bottom 10%	Top 5%	Top 10%
$\hat{\beta}_{\text{GT}}$ (Female)	-0.7823^{***}	-1.3495^{***}	-0.0397	-0.4408^{***}
	(0.1075)	(0.2768)	(0.0672)	(0.1106)
$\hat{\beta}_{\text{VLM}}$ (Female)	-0.7726^{***}	-1.3317^{***}	0.0151	-0.4785^{***}
	(0.1109)	(0.2765)	(0.0585)	(0.0964)
$\hat{\delta}$ (Female $\times$ VLM)	0.0088	0.0167^{***}	0.0732^{*}	-0.0358
	(0.0045)	(0.0044)	(0.0307)	(0.0235)
N (obs)	310,576	674,316	325,068	628,698
N (pairs)	155,288	337,158	162,534	314,349

Notes: The headline HISCAM specification of Table 6 re-estimated on subsamples defined by the GT-side score. A pair enters a tail column if its GT-side HISCAM is in the indicated tail of the GT-distribution, and both its GT and VLM rows then flow into the stacked panel. The four columns are ordered Bottom 5%, Bottom 10%, Top 5%, Top 10%. $\hat{\beta}_{\text{GT}}$ is the gender gap estimated from $y \sim$ Female $|$ year on the GT half; $\hat{\beta}_{\text{VLM}}$ likewise on the VLM half; $\hat{\delta}$ is the Female $\times$ VLM interaction from the stacked panel. Year fixed effects throughout; year-clustered SEs in parentheses. Two of the four $\hat{\delta}$ estimates are statistically distinguishable from zero, reflecting the precision available on 155,000 to 337,000 pairs. In the bottom tails $|\hat{\delta}|$ is at most 1.2% of the within-tail $|\hat{\beta}_{\text{GT}}|$, the gender gap a researcher would report; in the top 5% neither transcription detects a within-tail gap for the interaction to move. Stars: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

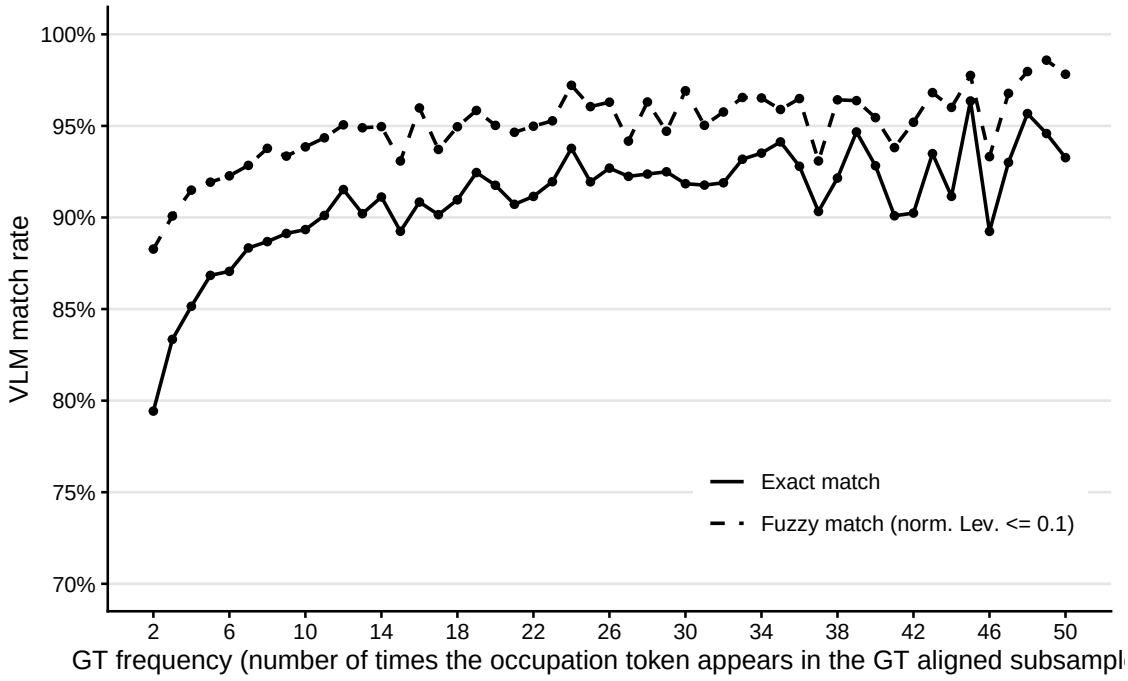


Figure G.1: Rare-occupation match: VLM string equality by GT frequency

Notes: For each GT-frequency value $N \in \{2, \ldots, 50\}$, the share of GT entries with that occupation token whose VLM-side string matches the GT-side string. Solid line: exact match. Dashed line: fuzzy match at a normalized Levenshtein distance of 0.1, the same tolerance used for the per-field accuracy metrics in Section 4.1. Sample: aligned pairs with non-missing and non-empty GT and VLM occupations; GT frequencies are computed within this aligned subsample. Reference rates for $N > 100$ are 97.8% exact and 98.8% fuzzy.