QUANTIFYING THE INVISIBLE: WOMEN DOCTORS IN THE ROSENWALD GUIDES (1887-1906)

LLM-BASED STRUCTURED DATA EXTRACTION FROM THE ROSENWALD GUIDES: METHODS AND HYBRID EVALUATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SEMESTER PROJECT

Ren Yi

Superviseuse et superviseurs :
Mikhaël Moreau
Dre Amélie Puche
Pr. Jérôme Baudry

8th January 2026

ABSTRACT

Historical medical directories are important sources for studying professional trajectories and social change, but their large scale and noisy OCR outputs make systematic analysis difficult. This report compares different pipelines to transform scanned pages of the *Rosenwald Guides* into structured data, supporting the MEDIF project goal of quantifying the presence of women doctors.

We first curate a focused corpus from the *Rosenwald Guides* (pilot scope: 20 editions, 1887–1906), identifying and extracting 4,116 relevant pages from sections listing physicians and related professions. We then propose a *Double Triangle* human–model workflow to create high-precision gold labels under limited annotation budgets: two independent multimodal LLM systems generate initial structured outputs, agreement is used to auto-accept easy cases, and disagreements are routed to a lightweight human review process. Applying this approach yields a benchmark with approximately 60 columns, where model agreement exceeds 85%; in the final review stage only 991 fields require manual correction out of 13,595 total fields (7.2%). Finally, we evaluate extraction strategies across different input modalities and models, and analyse the extraction performance from both quantitative and qualitative perspectives. Particularly, we analyse the extraction quality of female doctors and conduct initial analysis on the distribution of female doctors, based on our extraction results.

Declaration: ChatGPT is used in this paper to improve the language.

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Historical archives contain information that is critical for research in the Digital Humanities. However, many sources exist only as scanned images, and traditional OCR systems often produce noisy and inconsistent text. This lack of machine-readable structure makes large-scale analysis difficult and limits historians' ability to work with archival data efficiently.

1.2 PROJECT CONTEXT

This project was carried out within the framework of a partnership between **EPFL** and the **Institute of Humanities in Medicine (IHM)**, as part of the SNSF-funded **MEDIF project** (grant no. 215100[1]), which investigates the history of the first women physicians in French-speaking Switzerland and France.

The project was supervised by **Professor Jérôme Baudry (EPFL)**, **Mikhaël Moreau**—an SNSF PhD candidate whose doctoral research focuses on the professional strategies developed by early women physicians—and **Dr. Amélie Puche (IHM)**.

The initiative builds on a research proposal originally formulated by **Mikhaël Moreau**, with institutional and scientific support from the MEDIF project. Its objective is to contribute to the development of the MEDIF database, a resource dedicated to documenting women who practiced medicine during the Belle Époque and to addressing their long-standing invisibility in the history of medicine.

This work responds to a significant historiographical gap by enabling the identification and quantification of early women physicians, establishing their proportion within the overall medical workforce in France at the time, and mapping their professional trajectories. Beyond its methodological contribution to the annotation of digitized historical sources in the humanities and social sciences, the project pursues a concrete goal rooted in the MEDIF initiative: to support robust quantitative and longitudinal analysis of women's medical careers in the late nineteenth and early twentieth centuries.

1.3 DATASET

The Rosenwald Guides[2] is a series of medical directories listing doctors across France and its colonies from 1887 to 1949, and constitute a key source for identifying doctors, documenting their areas of

[1] https://data.snf.ch/grants/grant/215100
[2] https://gallica.bnf.fr/ark:/12148/cb344120051/date

practice and specializations, and mapping their professional trajectories. Although OCR transcriptions are available, the results remain noisy text without structured formatting, which makes systematic querying and data extraction difficult.

1.4 CHALLENGES

Recent progress in large language models provides a promising approach for extracting structured information from historical documents. In particular, multimodal LLMs, with proper prompts, can take scanned images as input and produce structured output. However, their reliability remains uncertain: hallucination can introduce information that did not exist in the original source.

A further challenge concerns evaluation. Historical documents rarely come with ground-truth annotations, so assessing extraction accuracy requires additional labeling effort. Traditional human annotation and inter-annotator agreement procedures demand substantial expert labor. Although LLMs can assist with benchmark creation, their outputs must themselves be validated before they can be considered reliable.

1.5 CONTRIBUTIONS

In this paper, we introduce a hybrid pipeline for extracting structured information from the Rosenwald Guides and evaluate it using a new benchmark and annotation paradigm.

Our contributions are threefold:

1. **Annotation methodology.** We propose a double-triangular annotation framework that combines LLMs with human labeling for scalable and trustworthy annotations.
2. **Benchmark construction.** We develop a benchmark dataset on the Rosenwald Guides to support the evaluation of structured information extraction from historical documents.
3. **Method evaluation.** We compare the 4 extraction methods on our benchmark in order to find the best method for our extraction.

CHAPTER 2

RELATED WORK

2.1 TEXT EXTRACTION METHODS

Extracting textual information from historical sources is a longstanding challenge in the Digital Humanities. Existing approaches can be broadly grouped into four categories: (1) traditional OCR pipelines, (2) multimodal LLMs that bypass classical OCR, and (3) hybrid approaches that combine OCR with post-correction or structuring by LLMs, (3) commercial OCR solutions.

2.1.1 TRADITIONAL OCR APPROACHES

We define traditional OCR systems as dedicated models that take scanned images or PDFs as input and generate textual output. In contrast to multimodal LLMs, which support a broader range of image understanding tasks (e.g., captioning, translation, and structured extraction), traditional OCR engines are designed specifically for recognizing text.

A large number of traditional OCR toolkits are available for direct use. Tesseract (Smith 2007), Calamari OCR (Wick, Reul and Puppe 2020), EasyOCR (*EasyOCR* 2020), Kraken (*kraken* n.d.), and olmOCR (Poznanski *et al.* 2025) provide open-source solutions. The research community also contributes to OCR-based extraction pipelines: for example, Löffler (2023) propose a three-stage architecture (layout detection, OCR, and named entity recognition) for extracting structured information from historical plans, and Fujitake (2024) present a decoder-only Transformer OCR model supported by a pre-trained language model.

While these approaches are efficient and easy to deploy, they are often limited in their adaptability to specialized tasks and domains. The extracted text may suffer from noise and formatting inconsistencies, which restricts their direct usability for downstream structured information extraction.

2.1.2 OCR-FREE MULTIMODAL LLMS

Multimodal LLMs can take multiple input modalities (such as images and text) and generate textual outputs based on a prompt. With appropriate prompting, these models can be used for OCR tasks. S. Kim *et al.* (2025) demonstrate that multimodal LLMs outperform traditional OCR and HTR approaches on metrics such as CER (Character Error Rate) and BLEU. Similarly, Nunes *et al.* (2025) compare GPT-4o with conventional models on table extraction and show that, while classical methods achieve higher performance in identifying table structure, GPT-4o substantially improves the accuracy of cell-level text extraction.

The language understanding capabilities of multimodal LLMs allow task-specific prompting and enable the generation of clean, structured results directly from the image input. However, despite their effectiveness, multimodal LLMs are susceptible to generating non-factual outputs, a phenomenon known as hallucination (Bai *et al.* 2024; Leng *et al.* 2024).

2.1.3 HYBRID OCR + LLM POST-CORRECTION

With the rapid development of LLMs, a natural strategy is to combine traditional OCR for efficient raw text extraction with LLMs for post-processing and structuring. In this hybrid setup, OCR produces the initial transcription and the LLM is responsible for correcting errors and formatting the text into a clean and structured form.

Nguyen *et al.* (2021) provide a comprehensive survey of post-OCR processing methods, typical system pipelines, and available datasets. Kanerva *et al.* (2025) show that LLM-based post-correction can substantially reduce CER on English texts, though the improvement does not generalize equally to Finnish. Similarly, O. M. Machidon and A. L. Machidon (2025) compare hybrid Tesseract + ChatGPT-4 with olmOCR on digitizing Slovenian folkloristic materials, and report that the hybrid approach achieves better accuracy on two out of three datasets. Their study also notes that, while LLMs improve readability and formatting, they must be applied carefully to avoid distorting historically or linguistically meaningful content, and that input scan quality remains a crucial factor for performance.

Thomas, Gaizauskas and Lu (2024) apply Llama 2 to post-OCR correction of English newspapers and observe improved CER reduction compared to the sequence-to-sequence model BART (Lewis *et al.* 2020). Do *et al.* (2025) perform post-correction for Vietnamese OCR and demonstrate that contextual references can be used to reduce hallucination. Bourne (2024) further show that some LLMs significantly reduce error rates and that providing socio-cultural context in prompts can improve performance. Boros *et al.* (2024) evaluate fourteen foundation models on several post-OCR correction benchmarks and, however, find that LLMs generally struggle when correcting historical transcription errors.

2.1.4 COMMERCIAL OCR SOLUTIONS

In addition to open-source and research-based toolkits, commercial OCR systems provide ready-to-use, no-code solutions. Examples include ABBYY FineReader (*ABBYY OCR* n.d.) and Adobe Acrobat OCR (*Adobe Acrobat OCR* n.d.), as well as cloud-based services such as Azure Cognitive Services (*Microsoft Azure OCR* n.d.), Google Cloud Vision (*Google Cloud Vision OCR* n.d.), and Amazon Textract (*Amazon Textract* n.d.). These tools are mature and practical to deploy, but the underlying models are proprietary and undocumented, so we do not further investigate them in this work.

2.2 HUMAN–LLM HYBRID ANNOTATION

With the rapid advancements in the text understanding capabilities of modern LLMs, using LLMs for data annotation has become an increasingly popular approach. Ding *et al.* (2023) show that GPT-3 can annotate data for a range of tasks at substantially lower cost than human annotation, although their results also indicate that annotation quality deteriorates for complex tasks involving subtle distinctions, domain-specific knowledge, or ambiguous labels.

Several studies have adopted a hybrid paradigm that combines LLM annotation with human verification. S. Wang *et al.* (2021) use GPT-3 as a low-cost labeler—reducing annotation expenses by 50–96%—and improve annotation accuracy by having humans re-label items with the lowest confidence scores. A similar strategy is used by H. Kim *et al.* (2024), where human only reevaluate the low-confidence cases. X. Wang *et al.* (2024) deploy two LLMs, one for labeling and the other for generating confidence score.

However, Zhang *et al.* (2025) caution that LLM-as-judge evaluation is not always reliable, casting doubt on confidence-based label quality control.

Pangakis and Wolken (2025) further integrate human-in-the-loop refinement by manually analyzing LLM labeling errors and incorporating misclassified examples back into the prompt. Their experiments reveal that LLM annotators tend to achieve high recall but comparatively low precision. Moreover, with sufficiently large training sets, supervised encoder-based models outperform GPT-4, whereas in low-resource settings GPT-4 yields superior performance.

The integration of LLM predictions into human workflows also presents risks. Schroeder, Roy and Kabbara (2025) find that providing annotators with LLM-generated labels increases their confidence but does not accelerate annotation, and introduces anchoring effects where annotators are influenced by the LLM outcome. This aligns with the concern of automation bias discussed in Gu *et al.* (2025).

Model capability and language coverage are additional considerations. Mohta *et al.* (2023) show that Vicuna-13B v1.5 performs significantly worse on non-English language data, underscoring the importance of considering language adaptability for French annotation. For imbalanced classification tasks, Chen *et al.* (2025) propose iteratively refining prompts based on recall for rare classes to improve performance.

CHAPTER 3

DATASET

3.1 DATASET: THE ROSENWALD GUIDES

3.1.1 HISTORICAL CONTEXT AND SIGNIFICANCE

The *Guides Rosenwald* are medical directories listing physicians and pharmacists in France and its colonies. Published annually between 1887 and 1949, the collection comprises 47 volumes. Originally conceived as a strategic reference work, the Guides aimed to provide a comprehensive overview of the medical profession across French territories to assist doctors and pharmacists in choosing their place of practice. Their stated mission was to regulate the spatial distribution of medical professionals so as to avoid overcrowding and ensure balanced access to care. As the preface to the 1887 edition states:

> "Éviter un encombrement également préjudiciable aux nouveaux venus et aux médecins et pharmaciens déjà fixés."
>
> — *Préface du Guide Rosenwald (1887)*

However, it is worth emphasizing the directory's highly practical utility for the general public: enabling users to find a specific physician in a given location. This utilitarian function is clearly reflected in the organization of the lists. Users could search for a doctor established in their neighborhood or even on their specific street (location-based lists), or locate the contact details of a known individual (alphabetical lists). Furthermore, specialized lists were provided for those seeking specific practitioners, such as surgeons, dentists, pharmacists, or doctors acting in thermal spas.

From a digital humanities perspective, this corpus constitutes a rich and still underexplored historical source for examining the geography of medical practice, the evolution of healthcare networks, and the social and institutional dynamics of French medicine in the late nineteenth and early twentieth centuries.

3.1.2 DOCUMENT LAYOUT AND EXTRACTION CHALLENGES

Figure 3.1 shows a double-page spread from the Rosenwald Guides. Our focus is on the doctor entries in red boxes; however, the pages also contain advertisements and other information. Figure 3.2 illustrates the structure of typical entries. Several challenges arise: some fields are missing for certain individuals, line breaks cause entries to overlap in layouts. These noises present difficulties to our task of extracting structured information.

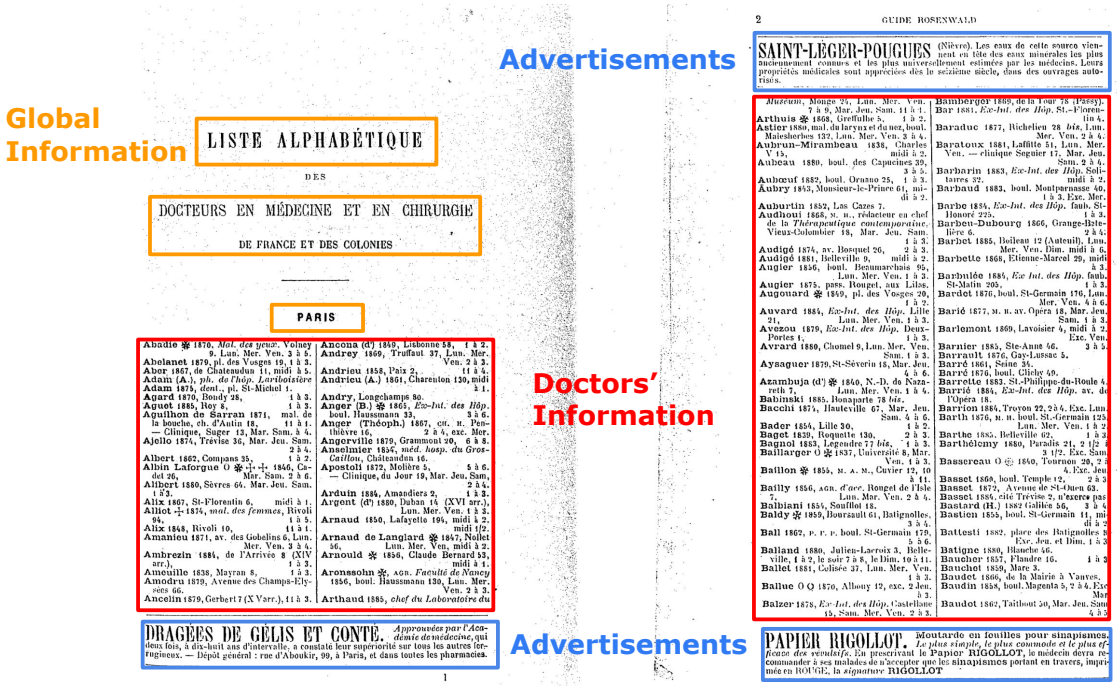


FIGURE 3.1
Page view of the Rosenwald Guide (1887, page 22–23)

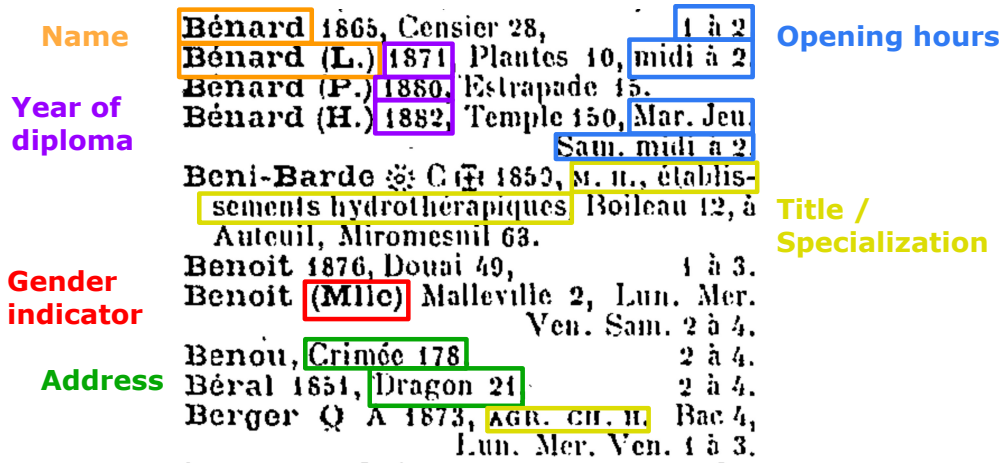


FIGURE 3.2
Close-up of annotated entries in the Rosenwald Guide (1887, page 24)

3.1.3 CORPUS CONSTRUCTION AND STATISTICS

In this paper, we focus on extracting doctor entries, with particular attention to female doctors, for example, their name, year of diploma, specialization. Table 3.1 shows an example extraction of Figure 3.2.

Nom	Année	Notes	Adresse	Horaires	Sexe
Bénard	1865		Censier 28	1 à 2	
Bénard (L.)	1871		Plantes 10	midi à 2	
Bénard (P.)	1880		Estrapade 15		
Bénard (H.)	1882		Temple 150	Mar. Jeu. Sam. midi à 2	
Beni-Barde	1853	M. H., établissements hydrothérapiques	Boileau 12 (Auteuil), Mir-omesnil 63		
Benoit	1876		Douai 49	1 à 3	
Benoit (Mlle)			Malleville 2	Lun. Mer. Ven. Sam. 2 à 4	Mlle
Benou			Crimée 178	2 à 4	
Béral	1851		Dragon 21	2 à 4	
Berger	1873	AGR. CH. H.	Bac 4	Lun. Mer. Ven. 1 à 3	

TABLE 3.1
Doctor Entries of Figure 3.2

To narrow the scope of our analysis, we process only the sections titled *Docteurs en médecine et en chirurgie* and *Officiers de santé* for Paris and for the départements and colonies, published between 1887 and 1906 (20 volumes). We manually identified the relevant page ranges for each volume and extracted the corresponding pages as images.

Table 3.2 presents the number of pages extracted for each edition, yielding a dataset of 4,116 pages in total. We can see that except for 1903 and 1904, each year has around 200 relevant pages.

TABLE 3.2
Number of pages per edition of the Rosenwald Guides (1887–1906)

Year	1887	1888	1889	1890	1891	1892	1893	1894	1895	1896
Pages	202	186	187	183	183	184	190	191	189	216
Year	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906
Pages	208	192	198	197	201	206	304	301	222	226

CHAPTER 4

DOUBLE TRIANGULAR EVALUATION FRAMEWORK

4.1 METHODOLOGY: THE DOUBLE TRIANGLE PARADIGM

4.1.1 MOTIVATION AND OBJECTIVE

Reliable evaluation of Optical Character Recognition (OCR) systems on complex historical documents requires high-precision "Golden Truth" data. Traditional double-blind human annotation is expensive, thus unscalable, while fully automated pipelines suffer from model bias and hallucination. To resolve this tension, we introduce the **Double Triangle Annotation** paradigm.

This framework synthesizes the computational power of Multimodal Large Language Models (MLLMs) with targeted human oversight. Rather than serving as primary annotators, humans act as "conflict resolvers" within a statistically rigorous filtering system. This design is motivated by three key insights:

1. **Statistical Improbability of Coincident Error:** While independent models may make mistakes, the probability of two architecturally distinct models making the exact same character-level error is statistically negligible. Consequently, model consensus serves as a high-confidence proxy for truth.
2. **Semantic Denoising:** MLLMs leverage linguistic priors to "denoise" visually degraded text (e.g., correcting "Bandry" to the valid surname "Baudry"), providing a semantic advantage over traditional OCR while managing the risk of hallucination.
3. **Mitigation of Human Fallibility:** To counter human fatigue and subjectivity, our "Double Triangle" structure compares two independent model-human systems against each other, serving as a meta-verification layer.

The objective is to produce a Golden Truth dataset with near-perfect accuracy ($> 99\%$ character match) at a fraction of the cost of manual annotation. Figure 4.1 illustrates this architecture.

4.1.2 THE FIRST LAYER: THE ANNOTATION TRIANGLE (MACHINE-HUMAN COLLABORATION)

The **First Layer Triangle** maximizes throughput by leveraging the consensus of distinct models to automate labeling, strictly reserving human labor for ambiguous cases.

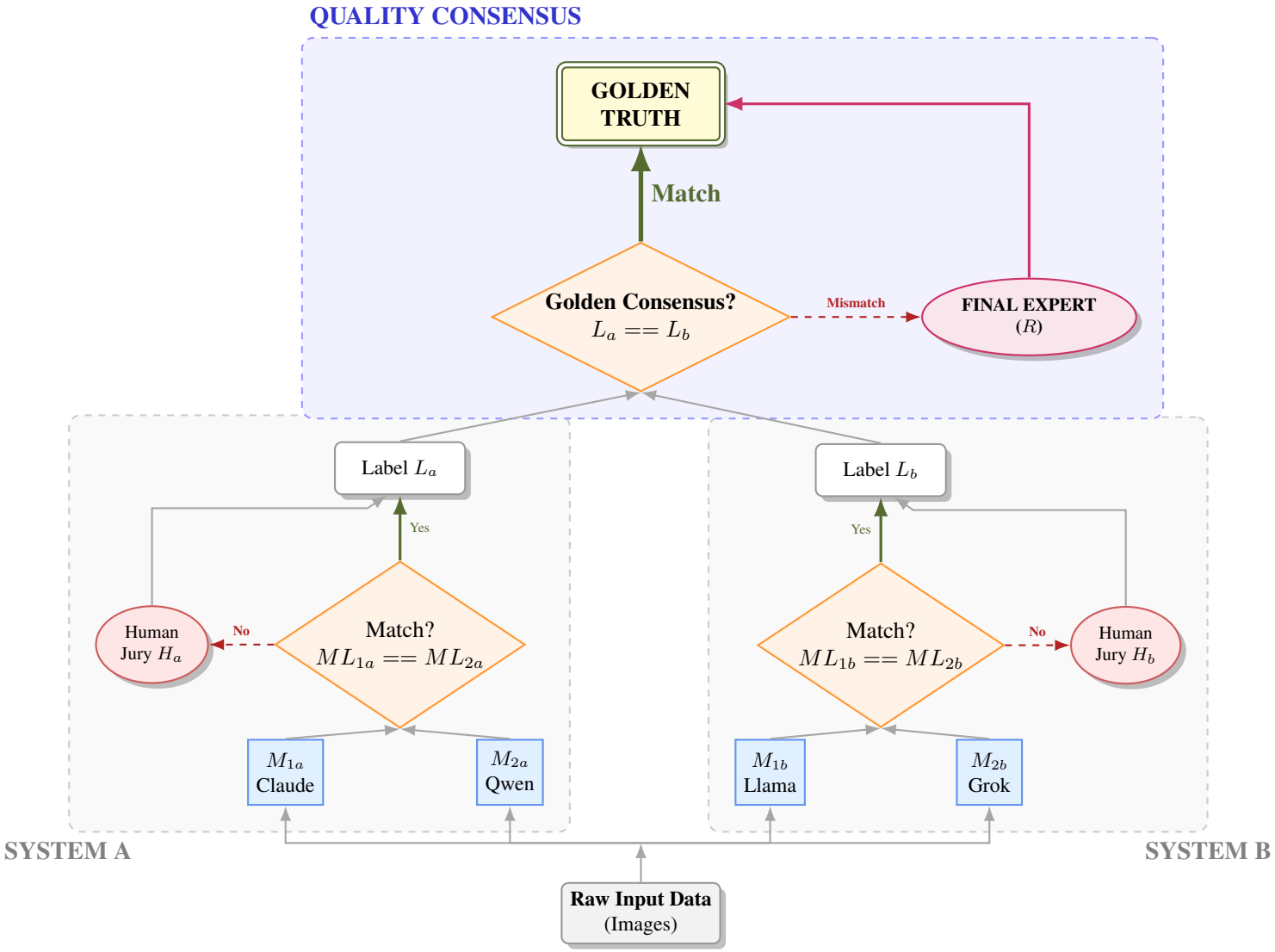


FIGURE 4.1
Overview of the Double Triangle Annotation Framework with Layer 2 Highlighted.

ARCHITECTURAL COMPONENTS

Each annotation unit, defined as a **System** (S), comprises three agents:

- **Two Machine Annotators** (M_1, M_2)**:** Independent models (e.g., MLLMs or OCR engines) that extract structured text in parallel.
- **One Human Jury** (H)**:** An annotator acting solely as a conflict resolver.

CONSENSUS-BASED WORKFLOW

The process follows a "Verification by Agreement" logic for every image I:

1. **Parallel Inference:** M_1 and M_2 independently generate provisional labels ML_1 and ML_2.
2. **Automated Filtering:**
   - **Consensus** ($ML_1 = ML_2$)**:** The label is automatically accepted as System Output (L) based

on the low probability of coincident error.

- **Divergence ($ML_1 \neq ML_2$):** The discrepancy is flagged, and the case is routed to the Human Jury.

3. **Human Adjudication:** The Jury (H) reviews image I and conflicting predictions to determine L, shifting the human role from "producer" to "verifier."

THE INDEPENDENCE REQUIREMENT

To validate the assumption that $P(\text{Joint Error}) \approx 0$, M_1 and M_2 must be **architecturally independent** to avoid "Model Collapse" (shared hallucinations). We employ "Orthogonal Independence" via two strategies:

- **Cross-Family:** Pairing models from different providers (e.g., Llama vs. Claude) to ensure distinct training distributions.
- **Cross-Modal:** Pairing a Vision model (OCR) with a Semantic model (MLLM, Multimodal LLM). This ensures error distributions are disjoint (topology vs. probability).

VERIFICATION SAMPLING

To monitor "Silent Error Rate," we implement a **Stochastic Quality Check**. A random subset of "Consensus" cases is routed to the Human Jury to verify the ongoing validity of the independence assumption.

4.1.3 THE SECOND LAYER: THE VALIDATION TRIANGLE (SYSTEM-SYSTEM COLLABORATION)

To mitigate the subjectivity and fatigue risks inherent in a single Human Jury, the **Second Layer** operates as a meta-verification step to guarantee “Golden Truth” standards.

META-ARCHITECTURAL COMPONENTS

This layer treats the First Layer System (S) as a single annotator unit. The structure comprises:

- **Two Independent Systems (S_1, S_2):** Distinct First Layer instances utilizing different underlying architectures (e.g., Llama-based vs. Claude-based) to maintain orthogonal independence.
- **One Final Reviewer (R):** A domain expert (e.g., historian) acting as the ultimate arbiter.

THE VALIDATION WORKFLOW

The process filters residual model and human errors through strict comparison:

1. **Dual Generation:** Systems S_1 and S_2 independently process the data to produce refined labels L_a and L_b.
2. **Meta-Comparison:**
   - **Golden Consensus ($L_a = L_b$):** If systems agree, the label is automatically designated as **Golden Truth** (G). This implies that disjoint models (and potentially distinct Human Juries) reached the same conclusion.
   - **Conflict ($L_a \neq L_b$):** Mismatches indicate complex edge cases or rare model divergence.

3. **Expert Resolution:** Conflicting entries are routed to the Final Reviewer (R) to determine the final Golden Truth.

ERROR PROBABILITY REDUCTION

By cross-checking decisions from disparate Human Juries, we induce a "Swiss Cheese" failure model. The probability of an error persisting in the Golden Truth is the product of independent system failures:

$$P(\text{Error}_G) \approx P(\text{Error}_{S1}) \times P(\text{Error}_{S2}) \tag{4.1}$$

Since S_1 and S_2 are high-accuracy systems, this product approaches zero. This ensures expensive expert intervention is allocated only to the most difficult small percentage of the dataset.

4.2 EXPERIMENTS AND EVALUATION

The experiments were designed to answer two key research questions:

1. **Quality:** Does the double layer effectively corrected the model and human errors?
2. **Efficiency:** Does the double layer successfully automate the majority of the workload?

4.2.1 EXPERIMENTAL SETUP

DATASET CHARACTERISTICS

To validate the efficacy of the Double Triangle Annotation framework, we conducted evaluation on the target dataset: page 32 of book 1887. This page is labelled independently by two annotators independently, manually (without LLM or OCR help). Since annotating one page is not that much workload (40 entries the left column and 36 entries ont he right), and the annotations of two annotators are compared with each other to get the final result, we regard this final result as golden standard.

However, out of convenience, these two annotators here are the same as H_a and H_b in Section 4.2.1, which could potentially harm the independence between the golden truth and the results to be checked.

MODEL CONFIGURATION

To satisfy the requirement of *Orthogonal Independence* (Section 4.1.2), we instantiate two distinct Systems using cross-company model pairs:

- **System 1 (S_1):** M_{1a} is `claude-sonnet-4-5` and M_{1b} is `qwen3-vl-235b-a22b-thinking`.
- **System 2 (S_2):** M_{2a} is `llama4-maverick` and M_{2b} is `grok-4-0709`.

By design, each System uses heterogeneous models to reduce correlated failure modes (e.g., shared blind spots on degraded typography or consistent formatting hallucinations), thereby lowering the risk of *Model Collapse* under agreement-based auto-labeling.

A stronger configuration would pair one MLLM (Multimodal LLM) with one traditional OCR engine, since classical OCR relies on fundamentally different recognition pipelines than end-to-end generative MLLMs and can therefore provide more complementary errors. In practice, however, our pilot experiments showed that the OCR outputs were too noisy for this setting: common failures included character-level misrecognitions and structural errors such as merging adjacent entries or mixing fields across different doctors. To keep the pipeline reliable and to reduce downstream human correction effort, we therefore use two MLLMs per System. In our data, MLLM outputs were consistently cleaner and more structured than

Doucet, Martyrs 74. 2 à 4.
Doury 1882, Blomet 73, 1 à 3.
Douvillé 1858, Rivoli 124, 1 à 2.
Dreyer-Dufer O ✠ 1873, CH. H., Richer 52, Jeu. Sam. 2 à 4.
Dreyfous 1870, *Ex-Int. des Hôp.*, Capucines 9, 1 à 3.
Dreyfus-Brissac 1878, M. H., Clichy 46, Lun. Mer. Ven. 1 à 2.
Dromain 1877, Bonaparte 45, 1 à 3.
Dromard 1853, boul. Magenta 46, 1 à 3.
Drouadaine 1867, Moines 18, 1 à 3.
Dubief 1886, Taylor.
Dubois 1859, *Ex-Int. des Hôp.*, Bausset 10, 1 à 2.
Dubois 1868, Fontaine-St-Georges 22, 1 à 3.
Dubois 1873, boul. Montparnasse 154.
Dubois 1880, Brézin 23, 1 à 3.
Dubouchet 1867, boul. des Capucines 8, 2 à 4.
Dubousquet-Laborderie 1883, Soubise 11, à St-Ouen. 1 à 2.
Duboys de Lavigerie 1880, Mogador 5. à 4.
Dubrisay ✠ 1860, Marengo 6, Lun. Mer. Ven. 2 à 4.
Dubroca 1879, Abesses 48. à 1.
Dubuc ✠ 1864, *Ex-Int. des Hôp.*, Taibout 83, midi à 3, Exc. Jeu.
Dubuisson 1874, boul. Montparnasse 46.
Ducamp 1876, av. de Wagram 53, Lun. Mar. Jeu. Sam. 1 à 2.
Duchastelet 1886, Denfert-Rochereau 18 *bis*.
Dulastel 1872, M. H., Bellechasse 14. Mar. Jeu. Sam. 2 à 3.
Ducat 1866, Compans 23, midi à 1
Duchaussoy ✠ 1854, AGR. Beaux-Arts 8, Mar. Jeu. Sam. midi à 3.
Duchesne 1864, *Ex-Int. des Hôp.* Sts-Pères 85, 1 à 2.
Ducor 1879, Jouffroy 68 *bis*, Lun. Mer. Ven. 1 à 3.
Ducoudray, la Victoire 60.
Duflocq 1881, *Ex-Int. des Hôp.*, boul. Malesherbes 19.
Duguet 1866, AGR., M. H., Londres 60, Lun. Mer. Ven. midi 1/2 à 2 1/2.
Duhamel 1885, Gde rue St-Mandé 106.
Duhomme 1859, *Ex-Int. des Hôp.*, pass. Saulnier 11, 2 à 3.
Dujardin-Beaumetz O ✠ 1862, M. H., M. A. M., boul. St-Germain 176, Lun. Mer. Ven. 2 à 3.
Dumesnil, *Ex-Int. des Hôp.*
Dumonteil-Grandpré 1875, à Aubervilliers (Seine).
Dumontpallier O ✠ Q 1857, M. H., Vignon 24, midi à 2.
Dunoyer ✠ 1849, Dragon 30, 5 à 6.
Du Périer 1883, boul. Arago 38.
Dupertuis 1878, Pergolèse 48, Mar. Jeu. Sam. 1 à 3.
Dupierris 1860, St-Florentin 4, 2 à 3.
Duplaix 1883, *Ex-Int. des Hôp.*, St-Lazare 107.
Duplantier 1877, Custine 1, 1 à 3.
Duplay (Simon). M. A. M. CH. H., Penthièvre 2.
Duplay ✠ 1865, St-Lazare 107, Lun. Mer. Ven. 2 à 4.
Dupont 1882, des Pyramides 17.
Dupouy 1869, boul. Sébastopol 81, 2 à 3.
Dupouy 1885, Roussin 79, 1 à 3.
Duprat, Monsigny 17.
Dupré 1850, boul. St-Germain 74, 2 à 5.
Dupré 1884, cours de Vincennes 37.
Dupré 1884, Montorgueil 67.
Dupuy 1855, boul. Sébastopol 76, midi à 2.
Dupuy Catullienne 5, à St-Denis.
Durand ✠ 1854, Rivoli 196, midi à 2.
Durand 1872, Laugier 84.
Durand 1854, Ponthieu 11, 1 à 3.
Durand 1870, à Arcueil, Mar. Sam. midi à 1.
Durand 1874, à Puteaux (Seine).
Dureau, Tour d'Auvergne 16.
Durand-Fardel ✠ 1840, *Ex-Int. des Hôp.*, Guénégaud 17, Mar. Jeu. Sam. midi à 2.
Durand-Fardel 1886, faub. St-Honoré 166.
Duroziez 1853, CH. H., St-Roch 10, 1 à 3.
Durut 1854, Chabanais 10, 2 à 5.
Dusart ✠ 1865, *Ex-Int. des Hôp.*, av. de Villiers 16, midi à 2.
Dusaussay ✠ Q 1877, M. H., faub. Poissonnière 9, 3 à 5.
Dutrieux 1877, *Ex-Int. des Hôp.*, Mogador 5, 2 à 4.
Dutrieux-Bey ✠ 1885, faub. Poissonnière 9, 3 à 5.
Duval 1849, Jacob 20, midi à 3.
Duval 1859, *hydrothérapie*, Dôme 3.
Duval 1869, AGR. cité Malesherbes 11.
Duvernet, Bac 1, Mar. Jeu. Sam. 3 à 5 1/2.
Duvivier O ✠ O ✠ 1844, Vignon 28, midi à 3.
Ecalle 1885, Bac 38.
Echerac 1865, Rivoli 74, midi à 2.
Ehrhardt 1863, Meslay 10, 1 à 2. Exc. Jeu. Dim.

FIGURE 4.2
Doctors of medicine and surgery (Paris), Rosenwald Guide (1887, page 32): first (left) and second (right) columns.

the traditional OCR baseline, which substantially reduced the amount of manual correction required in the final verification step.

HUMAN JURY AND FINAL REVIEWER CONFIGURATION

We have two human juries (H_a, H_b) and one final reviewer (R) in our design. Ideally, H_a, H_b and R should be three different and independent annotators, and R should be more experienced than H_a and H_b to solve their conflict.

In our actual experiment setup, H_a is an annotator from Université de Lausanne, who speaks French fluently. H_b is the researcher of the project, who has basic knowledge of French. R is the same person as H_b. This setup is based on the convenience of the experiment. However, H_b and R as the same person could harm the independence assumption, and the heavy workload could further improve the error rate.

4.2.2 EVALUATION METRICS

Our evaluation uses the manually produced reference transcription (Section 4.2) as the **gold standard** G. For any system output (a model prediction, a jury-corrected version, or the reviewer-corrected version), we denote the hypothesis text as X. Metrics are computed **against** G.

WORD ERROR RATE (WER). WER measures word-level edit distance between X and G:

$$\mathrm{WER}(X, G) = \frac{S_w + D_w + I_w}{N_w}, \tag{4.2}$$

where S_w, D_w, I_w are the numbers of word substitutions, deletions, and insertions required to transform X into G, and N_w is the number of words in G. We report WER **before** and **after** human correction at each layer (Table 4.1).

CHARACTER ERROR RATE (CER). CER measures character-level edit distance:

$$\mathrm{CER}(X, G) = \frac{S_c + D_c + I_c}{N_c}, \tag{4.3}$$

where S_c, D_c, I_c are the numbers of character substitutions, deletions, and insertions, and N_c is the number of characters in G. CER could capture more fine grained errors than WER.

FIELDS TO CORRECT. In addition to text-level error rates, we quantify human workload directly at the **field** level. Let the dataset contain N_{fields} atomic fields (e.g., `nom`, `adresse`, `horaires`, etc.). For a given human operator (jury or reviewer), we define:

$$N_{\text{correct}} = \sum_{j=1}^{N_{\text{fields}}} \mathbb{I}\left[X_j \neq Y_j\right], \tag{4.4}$$

where X_j is the pre-review value shown to the annotator for field j, Y_j is the final value after the annotator's intervention, and $\mathbb{I}(\cdot)$ is the indicator function. This yields the "Fields to correct" column in Table 4.1.

HUMAN EFFORT RATIO. We normalize the correction count to obtain a comparable workload measure:

$$\mathrm{EffortRatio} = \frac{N_{\text{correct}}}{N_{\text{fields}}} \times 100\%. \tag{4.5}$$

This metric directly captures the fraction of the dataset that required manual edits (first-layer juries) or manual re-edits (second-layer reviewer), matching the "Human effort ratio" reported in Table 4.1.

4.2.3 RESULTS

TABLE 4.1
Experiment results of double triangular annotation framework on page 32, 1887, Rosenwald Guide

Human	Input	WER		CER		Fields to correct	Human effort ratio
		Before	After	Before	After		
$\mathbf{H_a}$	claude qwen	0.0872 0.0436	0.0302	0.0528 0.0141	0.0179	70	18.4%
$\mathbf{H_b}$	llama grok	0.0688 0.1409	0.0084	0.0274 0.0499	0.0028	100	26.3%
R	H_a H_b	0.0302 0.0084	0.0034	0.0179 0.0028	0.0007	16	4.2%

The experiment results are reported in Table 4.1.

4.2.3.1 ACCURACY

FIRST TRIANGLE LAYER. In the first layer of the triangle, the two annotators (H_a and H_b) correct the model outputs. After correction, the Word Error Rate (WER) decreases for both annotators, and the Character Error Rate (CER) decreases for H_b but not for H_a. By inspecting the original annotations, we found one doctor entry where H_a left all fields blank. A plausible explanation is that H_a accidentally forgot to annotate this entry and saved blank values. This suggests that the validity checks in our annotation platform could be strengthened, and it also motivates the need for a second-level review (the final reviewer) to further reduce single-annotator subjectivity. Overall, the reduction in error rates indicates that the first-layer correction is effective.

SECOND TRIANGLE LAYER. In the second layer, the final reviewer (R) reviews and corrects the results produced by H_a and H_b. Both WER and CER further decrease after R's correction, demonstrating the effectiveness of the second-layer review. Nevertheless, the error rate is not zero. A detailed inspection reveals two remaining errors in the final output:

1. One character substitution. *Douvillé* is incorrectly recognized as *Douville*. The missing-accent error appears in both `llama` and `grok` (i.e., the same mistake on the same data point), so H_b did not have an opportunity to catch it through disagreement-based checking. `claude` does not extract this entry, `qwen` also omits the accent, and H_a fails to annotate this entry due to the aforementioned omission. As a result, the discrepancy is escalated to R, who accepts the *Douville* variant from H_b without noticing the accent.
2. One character insertion. *Taibout* is incorrectly recognized as *Taitbout*. This error occurs for both `claude` and `qwen`, leaving H_a no chance to detect it via cross-model disagreement. `llama` outputs *Taitbout* while `grok` outputs *Taihout*; H_b follows the `llama` output. Because H_a and H_b end up with the same surface form, the case is not flagged for review, and R does not re-check it.

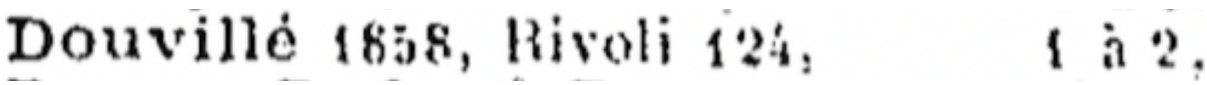


FIGURE 4.3
The entry *Douvillé*

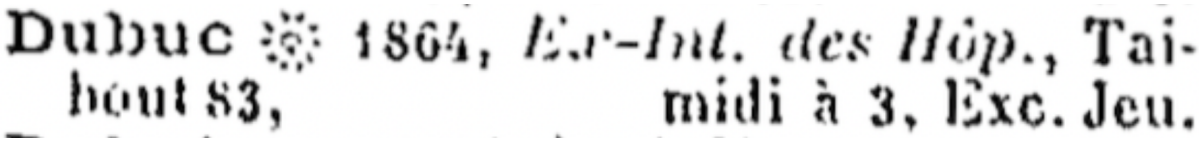


FIGURE 4.4
The entry *Taibout*

From the error analysis above, we identify 3 problems of the double triangular annotation framework:

1. **Correlated model errors.** Different models can converge to the same mistake on the same data point. For example, `llama`, `grok`, and `qwen` all omit the accent in *Douvillé*, and `claude`, `qwen`, and `llama` all misrecognize *Taibout* as *Taitbout*. Figure 4.3, 4.4 shows the corresponding snippets. In the *Douvillé* entry, the accent is visually close to a dot rather than a clear *accent aigu*. In the *Taibout* entry, the hyphen after *Tai-* (introduced by a line break) can plausibly be mistaken for a faded *t*. Such visual ambiguity can explain why multiple models fail in the same way. However, ambiguity is pervasive in historical documents, so we should be cautious about relying on independence assumptions across models.

2. **Semantic correction.** Due to their probabilistic nature, LLMs may output a *more common* or *more plausible* string rather than the exact string present in the image. For instance, *Douvillé* is a surname, whereas *Douville* is a commune name in France. A plausible explanation is frequency bias: *Douville* may occur more often than *Douvillé* in LLM training corpora, making it a more likely completion, even though the field is `nom` (not `adresse`). Similarly, *Taibout* is a surname, while *Rue Taitbout* is a street name in Paris. The model could prefer *Taitbout* due to the same probablity reason as *Douville*. Thus, although semantic priors can help resolve visually ambiguous characters, they can also introduce errors, especially for rare strings such as proper names. One way to mitigate this effect is to compare LLM outputs against a strong traditional OCR baseline: unlike LLMs, traditional OCR is typically more faithful to the observed glyphs and less influenced by semantic plausibility. However, the results of traditional OCR could be much more noisy than LLM, which increases manual effort.

3. **Residual human subjectivity.** Human decisions still influence the final output even with two layers of correction. In the *Douvillé* case, H_a accidentally leaves the entry blank, and R accepts H_b's *Douville* without noticing the missing accent. In the *Taibout* case, H_b follows `llama`'s *Taitbout* without detecting the subtle discrepancy. These examples indicate that human judgement remains a source of error: when presented with two candidate strings, the jury/reviewer may default to the option that *looks* more plausible, even when it is still incorrect. One possible mitigation is to modify the interface so that, when two candidates disagree, the system does not display them side-by-side; instead, it prompts the annotator to transcribe directly from the original image. This design could reduce anchoring effects and encourage closer inspection, at the cost of additional annotation time.

4.2.3.2 EFFICIENCY

H_a and H_b need to correct 70 and 100 fields, respectively. This corresponds to only 18.4% and 26.3% of the 380 fields in total, indicating a substantial gain in efficiency. The final reviewer R only needs to correct 16 fields, implying a relatively low review workload. Such reduced effort lowers human annotation costs and may also improve accuracy by alleviating fatigue.

4.3 CONCLUSION

This chapter proposed the **Double Triangular Evaluation Framework** to produce high-precision Golden Truth for noisy historical OCR while keeping human labor scalable. The first layer uses cross-model agreement to auto-accept easy cases and routes disagreements to a human jury; the second layer cross-checks two independent systems and escalates residual conflicts to a final reviewer.

On Rosenwald Guide (1887, p. 32), the framework delivers strong **efficiency gains**: only 18.4% and 26.3% of fields required first-layer correction, and the reviewer intervened on just 4.2%, while WER and CER generally decrease after each correction stage (Table 4.1). And the final output of our framework only has one character substitution error and one character insertion error, with 0.0034 WER and 0.0007 CER, which is close to perfect match with the golden truth. These results suggest that consensus filtering can automate most routine work, and that the two-stage verification effectively corrects many model and single-annotator errors.

However, the experiment also exposes key **limitations**. First, **model errors can be correlated**: multiple MLLMs may converge on the same wrong string under visual ambiguity, so agreement is not a guarantee of correctness. Second, **semantic priors can mislead** (frequency/plausibility bias), which is problematic for rare surnames and diacritics. Third, **human oversight remains fallible** (omissions, anchoring, fatigue), and our convenience setup (H_b also acting as R) weakens strict independence. Future work should strengthen independence (ideally pairing MLLMs with a strong OCR baseline), increase sampling of "consensus" cases to estimate silent error, and adjust the interface to encourage direct transcription in disputed cases.

CHAPTER 5

CREATE BENCHMARK WITH DOUBLE LAYER ANNOTATION FRAMEWORK

5.1 DATA SELECTION

As described in Section 3.1.3, we restrict our analysis to the sections titled *Docteurs en médecine et en chirurgie* and *Officiers de santé*, covering both Paris and the *départements* and colonies. Table 3.2 shows that most volumes contain roughly 200 target pages, with the exceptions of 1903 and 1904, which each include about 300 pages. In total, 4116 pages.

To construct a benchmark that remains representative of the full corpus, we apply stratified sampling with a target size of 100 pages. Specifically, for each year we allocate a number of sampled pages proportional to that year's share of the total target pages, and then sample pages within each year accordingly. This procedure yields a final benchmark of 105 pages.

To improve the annotation accuracy, we manually cropped the advertisements at the top and bottom of each page, and split the two columns into two images. Removing advertisements and splitting the columns could help the annotation accuracy, as is discussed in Section 5.4.

5.2 MODEL ANNOTATION

5.2.1 MODEL SELECTION

For two systems in the first layer, our model selection is the same as Section 4.2.1.

- **System 1 (S_1):** M_{1a} is `claude-sonnet-4-5` and M_{1b} is `qwen3-vl-235b-a22b-thinking`.
- **System 2 (S_2):** M_{2a} is `llama4-maverick` and M_{2b} is `grok-4-0709`.

These are the state-of-the-art models in current LLM competitions (december 2025), ensuring model capability. Models from different families ensure independency.

5.2.2 DATA CLEANING AND ALIGNMENT

The MLLM outputs do not always strictly follow the requested TSV format. We therefore applied an automatic post-processing step based on regular expressions to normalize the structure. After this correction, all TSV files are valid: each row contains the same number of fields as the header, ensuring consistent downstream parsing. Implementation details are provided in the codebase.

To measure agreement between the two models on the same page, we compute edit distance using the `jiwer`[1] library. This alignment step enables robust comparisons even when outputs differ in punctuation, spacing, or minor tokenization artifacts. Further details are available in the code.

5.2.3 EFFICIENCY AND QUALITY ANALYSIS

TABLE 5.1
Aggregated field-level matching rate by model pair.

Model pair	Matched fields	Total fields	% matched fields
claude_vs_qwen	37314	46830	79.7%
llama_vs_grok	34962	46550	75.1%

As is shown in Table 5.1, both model pairs achieved more than 75% matching rate over the fields. This shows that using MLLM as initial annotators could reduce over 75% of human labour in data labelling. In addition, based on our previous assumption, high level of matching rate also indicates high accuracy, which proves the quality of our labelling.

5.2.4 DATA POST-FILTERING

Figure 5.1 illustrates the distributions of field-level matching rate, character-level matching rate, number of distinct fields, and total number of fields across the 210-column dataset. We observe a long-tail phenomenon in the matching rates and the count of distinct fields; files located within these tails require significantly more time for human annotation.

To balance dataset quality and annotation feasibility within project time constraints, we filtered the dataset to retain only files meeting the following criteria: fewer than 70 distinct fields, a field matching rate greater than 0.7, and a character matching rate greater than 0.6.

This filtering strategy maximizes the utility of machine consensus and reduces the necessary human labor. We acknowledge a potential risk of favoring shorter columns or instances where the MLLM performed well, since we are deviating from the original stratified sampling.

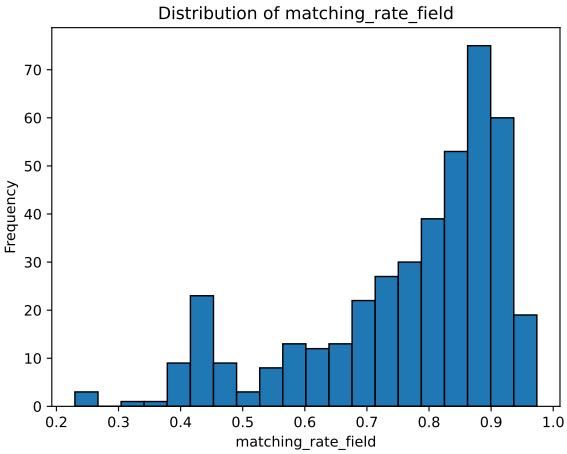
After filtering, we have the 58 columns to serve as the benchmark. The field matching statistics are shown in Figure 5.2.

On the second layer of our framework, we have 991 fields to correct, out of 13595 fields in total, which is only 7.2% human effort. We can still improve the efficiency of our framework by conducting more rule-based automatic review, for example resolve the conflict in address with only the difference like "Marseille" and "à Marseille".

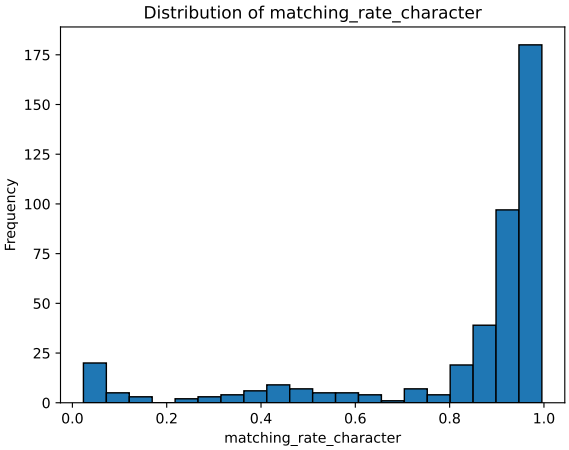
TABLE 5.2
Aggregated field-level matching rate by model pair in filtered data

Model pair	Matched fields	Total fields	% matched fields
claude_vs_qwen	11,294	12,895	87.6%
llama_vs_grok	11,048	12,905	85.6%

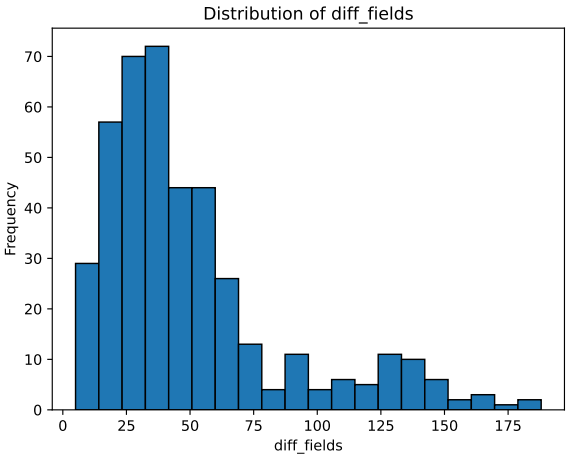
[1] https://github.com/jitsi/jiwer



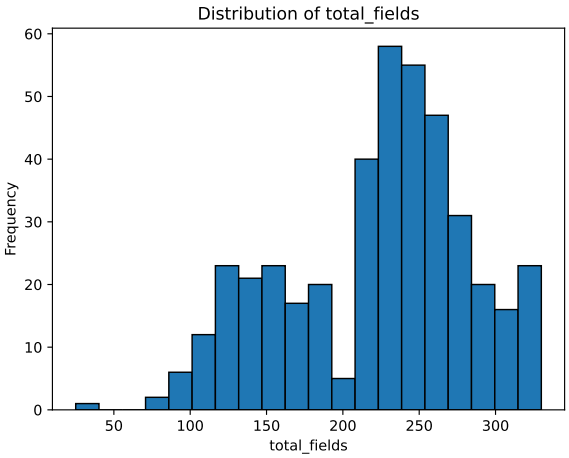
(A) Field-level matching rate



(B) Character-level matching rate



(C) Distribution of different fields



(D) Distribution of total fields

FIGURE 5.1
Distributions of matching rate and fields count over the data 210 columns

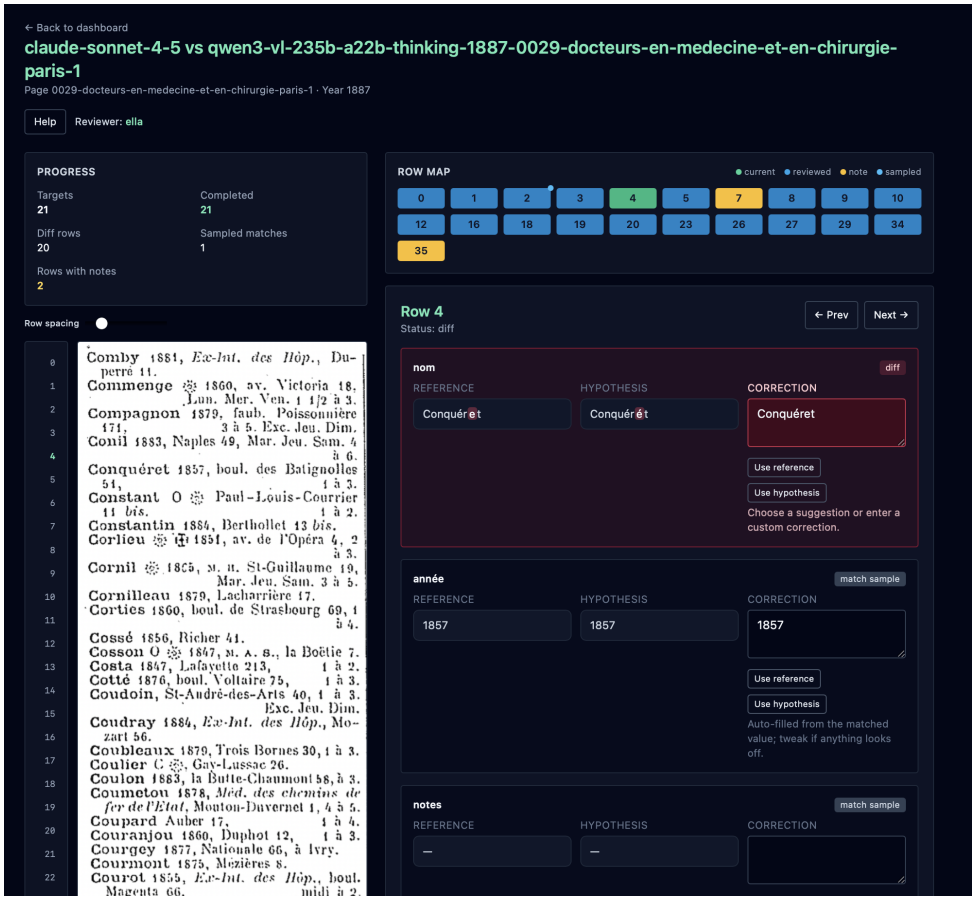


FIGURE 5.2
Human Correction Platform

5.3 CORRECTION PLATFORM

To facilitate the human correction of the data fields where machines don't reach consensus, we developed an annotation platform. With this platform, our annotator could focus on correcting the different results from machine annotations, as is shown in Figure 5.2.

We use red box to highlight the fields with different machine annotations, and we also highlight the exact different character in the field. In this way, we exploit the maximum utility of the machine consensus and improve the manual correction efficiency and accuracy.

We have our researcher Ren Yi and Ella Bischoff from UNIL to conduct the human correction. We appreciate their work.

5.4 IMAGE PREPROCESSING NECESSITY

Table 5.3 reports word error rate (WER; lower is better) under four image-input setups: *raw* (baseline), *no-ad* (removing auxiliary "ad/extra" content), *split* (splitting the page into columns before inference), and *concat* (concatenating split columns). Overall, removing auxiliary content is broadly beneficial: **12/13** models improve from *raw* to *no-ad*, with a median relative WER reduction of **20.03%** (mean: 22.27%). Splitting is often the most effective downstream strategy: **split** yields the best WER for **7/13** models (vs. **5/13** for *no-ad*), and achieves the best absolute result in our runs with `gemini-3-pro-preview` (**WER=0.0351**). In contrast, concatenation is unstable: only **4/13** models improve under *concat*, while

LLM	WER_{raw}	$WER_{no\text{-}ad}$	$\Delta_{raw \to no\text{-}ad}$ (%)	WER_{split}	$\Delta_{no\text{-}ad \to split}$ (%)	WER_{concat}	$\Delta_{no\text{-}ad \to concat}$ (%)
claude-haiku-4-5	0.1371	0.0987	28.01	0.0870	11.85	0.5987	-506.59
claude-opus-4-1	0.1338	0.1070	20.03	0.1104	-3.18	0.5151	-381.40
claude-sonnet-4-5	0.1438	0.0886	38.39	0.0836	5.64	0.5234	-490.74
gemini-2.5-flash	0.1371	0.0870	36.54	0.0669	23.10	0.0736	15.40
gemini-2.5-pro	0.0652	0.0619	5.06	0.0669	-8.08	0.0619	0.00
gemini-3-pro-preview	0.0585	0.0552	5.64	0.0351	36.41	0.0569	-3.08
gpt-5-mini	0.0803	0.0753	6.23	0.0870	-15.54	0.1187	-57.64
gpt-5.1	0.1221	0.1656	-35.63	0.0853	48.49	0.2408	-45.41
grok-4	0.1923	0.1756	8.68	0.1288	26.65	0.1589	9.51
grok-4-1-fast-reasoning	0.9532	0.3662	61.58	0.9097	-148.42	0.9448	-158.00
llama4-maverick	0.1421	0.0920	35.26	0.0686	25.43	0.0702	23.70
qwen3-vl-235b-a22b-thinking	0.0853	0.0702	17.70	0.0719	-2.42	0.0736	-4.84
qwen3-vl-8b-thinking	0.3478	0.1321	62.02	0.1405	-6.36	0.1120	15.22

TABLE 5.3
WER ablation across image-input strategies. Positive Δ indicates a relative WER reduction (improvement) versus the reference configuration.

several degrade severely (e.g., Claude models). Notably, `gpt-5.1` is an outlier where *no-ad* harms performance (WER 0.1221 $\rightarrow$ 0.1656), but *split* largely recovers accuracy (WER 0.0853).

BEST STRATEGY FOR THE SELECTED MODELS. For our selected models, we choose the image-input configuration that minimizes WER (Table 5.3). For `claude-sonnet-4-5`, **split** performs best (WER=0.0836), improving over *no-ad* (0.0886), whereas *concat* degrades substantially (0.5234). For `qwen3-vl-235b-a22b-thinking`, **no-ad** is marginally best (WER=0.0702); *split* (0.0719) and *concat* (0.0736) yield no further gains. For `llama4-maverick` and `grok-4`, **split** is clearly optimal (WER=0.0686 and 0.1288, respectively), suggesting that region-wise processing is beneficial for these models. In summary, **split** is optimal for **3/4** selected models and is within 0.0017 WER of the best setting for the remaining one; for consistency across systems, we therefore adopt **split** for all four models.

Particularly, the dramatic performance drop with the strategy concat, could result from the too long image after concatenation, Claude 4.5 Sonnet would be scaled down before converting to tokens [2], which could degrate the performance significantly.

5.5 CONCLUSION

This chapter constructs a representative benchmark from the Rosenwald Guides via stratified sampling and creates annotations using a double-layer framework: two independent MLLM systems produce initial TSV outputs. After post-filtering to prioritize high-consensus and feasible cases, we obtain a 60-column benchmark in which model agreement exceeds 85%. In the final review layer, we only need to manually correct 991 fields out of 13595, and this efficiency could be further improved. Finally, an ablation study shows that removing ads and splitting pages into columns consistently improves OCR quality, motivating the use of *split* preprocessing across our selected models.

[2] https://platform.claude.com/docs/en/build-with-claude/vision

CHAPTER 6

EXTRACTION PIPELINE

6.1 MOTIVATION AND OBJECTIVE

Extracting accurate, structured information from historical documents presents a distinct set of challenges. Primary difficulties include scanning noise and complex page layouts, such as fading ink, skew, and dense multi-column formatting. In this context, the central question is how to obtain *structured* outputs (e.g., TSV/JSON) while remaining faithful to the evidence present on the page.

To motivate our design choices, we compare four extraction paradigms that represent the main families of practical approaches for this task:

1. Traditional OCR
2. Multimodal LLM (image input)
3. Multimodal LLM (image + traditional OCR output)
4. Text-only LLM (traditional OCR input)

6.2 COMPARISON OF FOUR EXTRACTION METHODS

6.2.1 METHOD 1: TRADITIONAL OCR

Traditional OCR engines are designed for faithful character-level transcription from document images. In this setting, they offer a strong form of grounding: every output token is directly tied to observed glyphs. However, they lack semantic and layout reasoning. When glyphs are unclear or layouts are intricate, their performance degrades significantly, leading to typical failure modes such as:

- **Layout entanglement:** merging distinct entries or mixing details from adjacent columns.
- **Noise sensitivity:** misrecognizing stains or unwanted decorative symbols
- **Schema mismatch:** producing text that is not naturally aligned with the desired structured format.

As a result, OCR alone could be used for transcription yet unreliable for *record-level* extraction.

6.2.2 METHOD 2: MULTIMODAL LLM (IMAGE INPUT ONLY)

A multimodal LLM can directly map the document image to structured outputs, often producing clean and well-formatted results even in the presence of layout complexity. This approach is attractive because the

model can jointly reason about visual layout and semantic content. Nevertheless, it suffers from a crucial limitation: the lack of explicit grounding constraints. The model may **hallucinate** or **over-correct**, for example: replacing uncertain glyphs with plausible but incorrect names, dates, or addresses. In historical documents where uncertainty is frequent, such unconstrained generation can be costly.

6.2.3 METHOD 3: MULTIMODAL LLM (IMAGE + OCR OUTPUT)

A common mitigation is to provide the multimodal LLM with both the image and the traditional OCR transcription. Intuitively, the OCR text can serve as an additional evidence channel that encourage the model toward fidelity, while the image preserves access to layout cues and hard-to-recognize regions. This hybrid setup often improves robustness, but it raises cost as it is using both textual and image input.

6.2.4 METHOD 4: TEXT-ONLY LLM (OCR INPUT ONLY)

As is shown in Figure 6.1, an alternative is to restrict the LLM to *text-only* processing by feeding it the OCR transcription (and a target schema) while withholding the image. In this design, the OCR output becomes the sole evidence source, and the LLM is tasked with transforming an imperfect transcription into a structured representation. Concretely, the LLM can play two constrained roles:

1. **Noise filtering:** identifying and removing non-entry content (e.g., advertisements) and OCR artifacts.
2. **Schema alignment:** segmenting, labeling, and formatting the remaining text into TSV/JSON records.

Because the model cannot access visual signals, it has fewer opportunities to introduce visually-motivated "guesses." However, lack of visual information could also limit the accuracy of annotation, especially when the input OCR text is too noisy.

6.3 SUMMARY OF TRADE-OFFS

The four methods have different trade-offs for **fidelity to observed evidence** and **ability to resolve layout/semantic ambiguity**:

- **Traditional OCR** emphasizes character-level faithfulness but struggles with structure in complex layouts.
- **Multimodal LLM (image only)** emphasizes end-to-end structuring but risks hallucination and over-correction.
- **Multimodal LLM (image + OCR)** adds an evidence anchor yet cost more.
- **Text-only LLM (OCR only)** prioritizes evidence-constrained restructuring, but also constrained by the quality of OCR text.

Since each method has its own advantages and drawbacks, it's important to evaluate them in concrete application scenarios, in order to find out which works best in our task.

6.4 EVALUATION

In this section, we conduct a series of experiments to validate the effectiveness of our approach. Specifically, we aim to address the following research questions (RQs):

1. **RQ1:** Which data source should we use for the extraction?

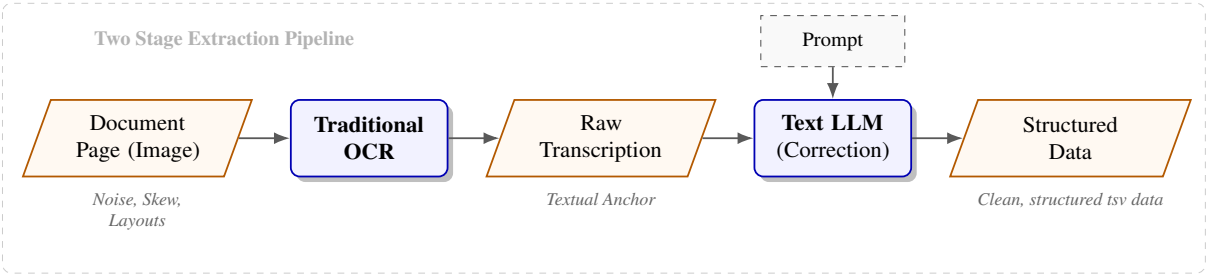


FIGURE 6.1
The two-stage extraction pipeline. Traditional OCR creates a textual anchor, which the LLM structures according to the prompt.

2. **RQ2:** Which model should we use for the extraction?

6.4.1 DATASET

Our evaluation benchmark consists of a curated set of historical documents (30 pages) created in the previous section.

6.4.2 EXPERIMENTAL SETUP

MODEL SELECTION For the generative components, we evaluate 4 state-of-the-art Large Language Models: **GPT-5.2** and **Gemini 3 Pro** and their light version **GPT-5 mini** and **Gemini 3 Flash**. For the initial text recognition layer, we utilize two sources: the open-source **Tesseract** engine and the legacy OCR metadata provided with the *Guide Rosenwald* digitization.

BASELINES AND COMPARISONS To validate the effectiveness of our extraction strategy, we compare four representative methods below:

1. **Traditional OCR:** Direct use of OCR transcriptions (Tesseract and the existing *Guide Rosenwald* OCR) without any LLM-based post-processing.
2. **Multimodal LLM (Image Only):** End-to-end extraction that maps an image crop directly to structured output using a vision-language model, without providing the OCR transcription.
3. **Multimodal LLM (Image + OCR):** A hybrid setup in which the vision-language model receives both the image crop and the OCR transcription, allowing it to leverage visual cues while being guided by textual evidence.
4. **Text-only LLM (OCR Only):** A two-stage configuration, where a text LLM receives only the OCR transcription (with a target schema) and is used solely for noise filtering and schema alignment, treating OCR as the sole evidence source.

6.5 QUANTITATIVE ANALYSIS

6.5.1 RQ1: WHICH DATA SOURCE SHOULD WE USE FOR THE EXTRACTION?

Table 6.1 shows the Word Error Rate (WER) and Character Error Rate (CER) across all the models and data sources. By comparing the best-performing (bolded) results across all data sources, we observe that *Image + Original OCR* consistently achieves the lowest error rate on all metrics, followed by *Image* alone. Both image-based settings show a substantial lead over *Original OCR* and *Tesseract OCR*. However, in

Source	Model	Avg WER	Mid WER	Avg CER	Mid CER
Image + Original OCR	gemini-3-pro-preview	**0.0360**	**0.0293**	**0.0139**	**0.0088**
	gemini-3-flash-preview	0.0435	0.0403	0.0177	0.0108
	gpt-5.2-2025-12-11	0.0581	0.0543	0.0178	0.0149
	gpt-5-mini-2025-08-07	0.1026	0.0645	0.0581	0.0192
Image	gemini-3-pro-preview	**0.0372**	**0.0347**	**0.0143**	**0.0092**
	gemini-3-flash-preview	0.0413	0.0355	0.0167	0.0106
	gpt-5.2-2025-12-11	0.2024	0.1133	0.1405	0.0384
	gpt-5-mini-2025-08-07	0.1274	0.0877	0.0650	0.0327
Original OCR	gemini-3-pro-preview	**0.0959**	**0.0751**	0.0373	**0.0264**
	gemini-3-flash-preview	0.0989	0.0910	**0.0314**	0.0292
	gpt-5.2-2025-12-11	0.1132	0.0794	0.0415	0.0317
	gpt-5-mini-2025-08-07	0.1580	0.1155	0.0701	0.0471
	raw_ocr	0.3948	0.3257	0.1862	0.0998
Tesseract OCR	gemini-3-pro-preview	0.2405	0.1681	0.1592	0.0667
	gemini-3-flash-preview	**0.1618**	**0.1267**	**0.0890**	**0.0475**
	gpt-5.2-2025-12-11	0.2644	0.1566	0.1838	0.0709
	gpt-5-mini-2025-08-07	0.3230	0.2378	0.2246	0.1068
	raw_ocr	0.8387	0.5956	0.4594	0.3611

TABLE 6.1
OCR and post-OCR correction performance across sources and models (30 documents). The best result of each source is bold, and the second best is underlined. (Avg is the macro avg over files)

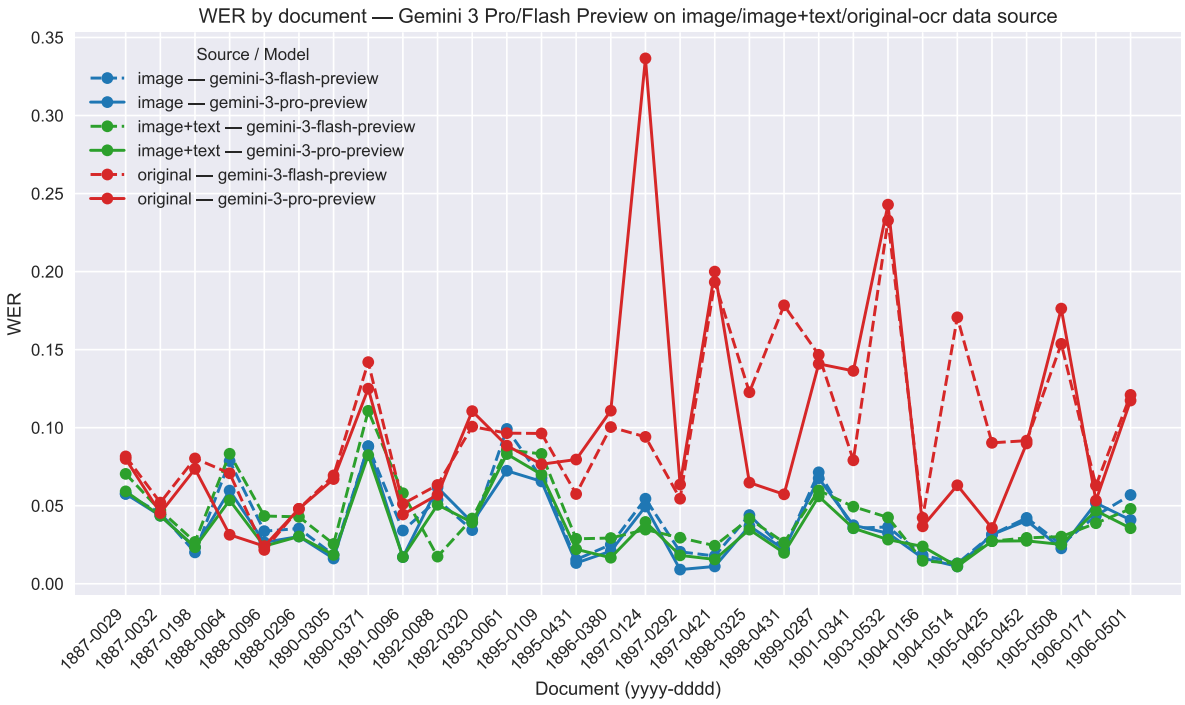


FIGURE 6.2
WER by document (Gemini 3 family)
('1899', '0287'), ('1903', '0532') these two years do not have original OCR, using Tesseract OCR instead

Figure 6.2, while the WER of Original (text only LLM with Original-OCR input) is usually higher than image and image+text source, there are exceptions in 1888-0064 and 1888-0096.

These results indicate that the quality of the input data source is the primary factor influencing extraction performance. The image modality contains the most complete and faithful information, and starting directly from images leads to strong results, especially given the capabilities of current multimodal LLMs. Augmenting images with the original OCR text can further improve accuracy, as the textual signal may help disambiguate certain fields and reduce model uncertainty.

In contrast, when relying solely on OCR outputs (either original or Tesseract) without access to images, the LLM functions mainly as a post-correction mechanism. In such cases, extraction quality is strongly constrained by the underlying OCR quality. For example, Tesseract OCR produces noisier text than the original OCR, its downstream extraction and post-correction performance degrades accordingly.

In summary, images should be included in the input for extraction. Adding the original OCR as auxiliary information provides modest additional gains. The final choice between *Image* and *Image + Original OCR* should be informed by further analysis using more fine-grained, field-level metrics rather than relying solely on aggregate performance measures, since the metric difference is small between these two sources.

6.5.2 RQ2: WHICH MODEL SHOULD WE USE FOR THE EXTRACTION?

Table 6.1 shows that across both the *Image* and *Image + Original OCR* settings, `gemini-3-pro-preview` consistently achieves the best performance, followed closely by `gemini-3-flash-preview`. This highlights the strong overall capability of the Gemini 3 family for our extraction task.

At the same time, the performance gap between the Pro and Flash variants remains relatively small, and the advantage of the Pro model is not always consistent across all the files (see Figure 6.2). This suggests that lighter models, which offer substantially lower latency and cost, are already capable of delivering competitive extraction quality in this scenario.

Consequently, while `gemini-3-pro-preview` represents overall higher accuracy, `gemini-3-flash-preview` constitutes a compelling alternative when efficiency considerations are prioritized. The final model choice should therefore balance marginal accuracy gains against computational cost and throughput requirements.

6.5.3 SUMMARY OF QUANTITATIVE ANALYSIS

From the quantitative analysis above, to pick the best strategy for our extraction, we can narrow down our choices to:

1. Model: Gemini 3 Pro Preview or Gemini 3 Flash Preview
2. Source: Image + Text or Image only

Since these four combinations exhibit comparable performance, we further rely on qualitative analysis to differentiate them and identify the most suitable strategy. In addition, the text-only LLM with original OCR input is also included in the qualitative analysis. This model served as our primary approach in the initial studies, and its inclusion provides a complementary perspective on text-only post-correction performance.

6.6 QUALITATIVE ANALYSIS

In this section, we conduct a qualitative, file-level analysis to examine representative cases in which the word and character error rates differ across model–input combinations. Rather than focusing on aggregate

metrics, we inspect individual files to understand the nature of errors, the conditions under which error rates increase or decrease, and whether these behaviors are consistent with the overall quantitative results.

1. **1897-0124**: In case of error rate, *Image+Text Gemini 3 Pro* < *Image Gemini 3 Pro* < *Original OCR + Gemini 3 Flash Preview* < *Original OCR + Gemini 3 Pro Preview*. The first inequality is consistent with the global evaluation results. However, the latter comparison is unexpected from a model-capacity perspective, as the Pro variant performs worse than the Flash model under identical original OCR input.
2. **1904-0514**: *Image+Text Gemini 3 Pro*, *Image Gemini 3 Pro* < *Original OCR + Gemini 3 Pro Preview*. A case in which only the text-only pipeline with original OCR input produces a high error rate, while all image-based or image+text configurations perform reliably. This file highlights the limitations of post-correction without visual grounding.
3. **1893-0061**: *Image Gemini 3 Pro* < *Image+Text Gemini 3 Pro*, contradicting the overall trend in which multimodal input generally yields lower error rates. This case motivates a closer inspection of how auxiliary textual input may introduce noise.

6.6.1 1897-0124

WHY IMAGE + TEXT WINS OVER IMAGE ONLY

Figure 6.3 shows the entries in page 1897-0124. Table 6.2 shows two examples where Image + Text input outperforms Image only input in this image. In the first example, with Image Only, the model recognizes the year as 1868, while in the image it shows 1868. In raw OCR text (original OCR), it shows 1868, and with Image + Text input, the model outputs the correct year. With Original OCR only, the model also outputs the correct result. This shows that Image Only input method could be unstable on certain points, and additional Text information could help the correction.

Source	Name	Year	Address
REF	viégarnier	1868	av des ternes 63
Image + Text	viégarnier	1868	av des ternes 63
Image Only	viégarnier	1866	av des ternes 63
Original OCR Only	viégarnier	1868	av des ternes 63
Raw OCR	viégarnier	1868	av dès ternes 63
REF	vigoureux	1868	vaugirard 33
Image + Text	vigoureux	1868	vaugirard 33
Image Only	vigouretix	1868	vaugirard 33
Original OCR Only	vigoureux	1868	vaugirard 33
Raw OCR	viçfoureux	186s	vaugirard 33

TABLE 6.2
Gemini 3 Pro Preview outputs compared to ground truth (REF). Errors are highlighted in red.

In addition, in the raw OCR, the "des" in the address is recognized as "dès". Although "dès" is a valid French word, it is not grammatically and semantically correct here. Image Only, Image + Text and Original OCR Only model corrected this error, showing the capability of semantic correction, or the usability of LLM prior knowledge in our correction task.

In the second example, with Image Only, the model has a wrong recognition on the name. The raw OCR also made mistakes on the name, which however reserves the correct suffix "reux", which could be a common suffix in French words. With Image + Text input, and Original OCR Only input, our model successfully corrected this error. One possible explanation is that vigoureux is a common French word and also a possible surname. This word is recognized by our model through its prior knowledge. This

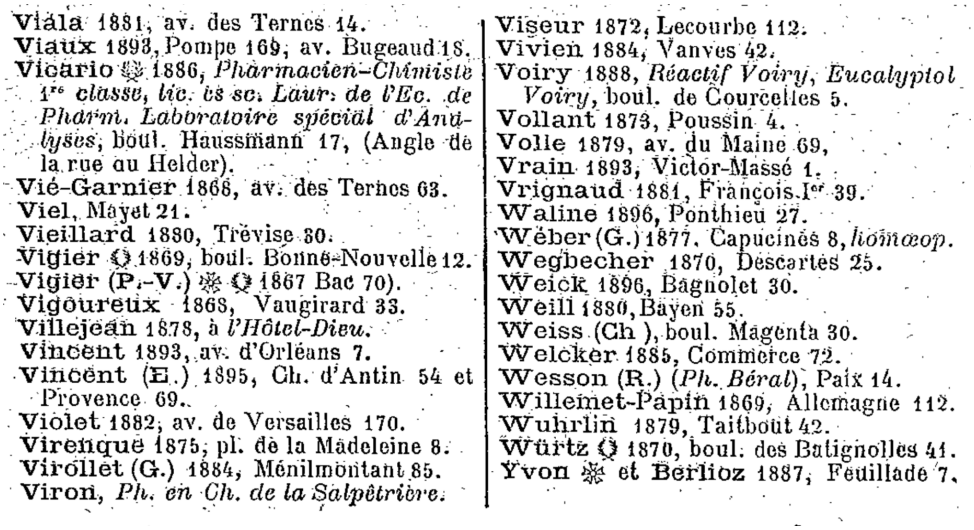

Viala 1881, av. des Ternes 14.
Viaux 1893, Pompe 169, av. Bugeaud 18.
Vicario ✠ 1886, *Pharmacien-Chimiste 1re classe, lic. ès sc. Laur. de l'Ec. de Pharm. Labbratoire spécial d'Analyses*, boul. Haussmann 17, (Angle de la rue du Helder).
Vié-Garnier 1868, av. des Ternes 63.
Viel, Mayet 21.
Vieillard 1880, Trévise 30.
Vigier ✠ 1869, boul. Bonne-Nouvelle 12.
Vigier (P.-V.) ✠ ✠ 1867 Bac 70).
Vigoureux 1868, Vaugirard 33.
Villejean 1878, à *l'Hôtel-Dieu.*
Vincent 1893, av. d'Orléans 7.
Vincent (E.) 1895, Ch. d'Antin 54 et Provence 69.
Violet 1882, av. de Versailles 170.
Virenque 1875, pl. de la Madeleine 8.
Virollet (G.) 1884, Ménilmontant 85.
Viron, *Ph. en Ch. de la Salpêtrière.*
Viseur 1872, Lecourbe 112.
Vivien 1884, Vanves 42.
Voiry 1888, *Réactif Voiry, Eucalyptol Voiry*, boul. de Courcelles 5.
Vollant 1873, Poussin 4.
Volle 1879, av. du Maine 69,
Vrain 1893, Victor-Massé 1.
Vrignaud 1881, François Ier 39.
Waline 1896, Ponthieu 27.
Wéber (G.) 1877, Capucines 8, *homœop.*
Wegbecher 1870, Descartes 25.
Weick 1896, Bagnolet 30.
Weill 1880, Bayen 55.
Weiss (Ch), boul. Magenta 30.
Welcker 1885, Commerce 72.
Wesson (R.) (*Ph. Béral*), Paix 14.
Willemet-Papin 1869, Allemagne 112.
Wuhrlin 1879, Taitbout 42.
Würtz ✠ 1870, boul. des Batignolles 41.
Yvon ✠ et Berlioz 1887, Feuillade 7.

FIGURE 6.3
1897-0124 image

once more shows the unstability of Image Only input method.

When it comes to the year, the raw OCR has the year 1868 recognized as 186s, which is not a valid year. It's acceptable for the OCR because it only follows the string as it is, rather than cooperating semantic knowledge. All of our input modalities, corrected this mistake, again showing the capability of semantic correction.

WHY PRO FAILS FLASH

In Table 6.3, we have several examples where Gemini 3 Flash Preview outperforms its pro version, Gemini 3 Pro Preview. In all these 5 examples, Gemini 3 Pro completely missed the entries and Gemini 3 Flash Preview faithfully corrected them. All these entries, except for the last one, have the specializations/titles related to the doctors. The last example has two names. One possible explanation is that in our input prompt, we didn't give examples of such complicated title/specializations and the examples of multiple names, and the model may not be familiar with this pattern. Gemini 3 Pro Preview may have strong prior knowledge of the common strings, but it does not recognize these rare strings about titles/specializations, so it chose to treat them as noises and ignore them. Gemini 3 Flash Preview, on the other hand, could have less prior knowledge and process the input string as it is, which in turn gives good results. There are no other results where Gemini 3 Pro Preview missed the entire entry other than these examples.

This explanation shows that the Pro model may not always outperforms the light model in our Text Only scenario, especially on not fully prompted and rare cases.

6.6.2 1904-0514

WHY TEXT ONLY ORIGINAL FAILS

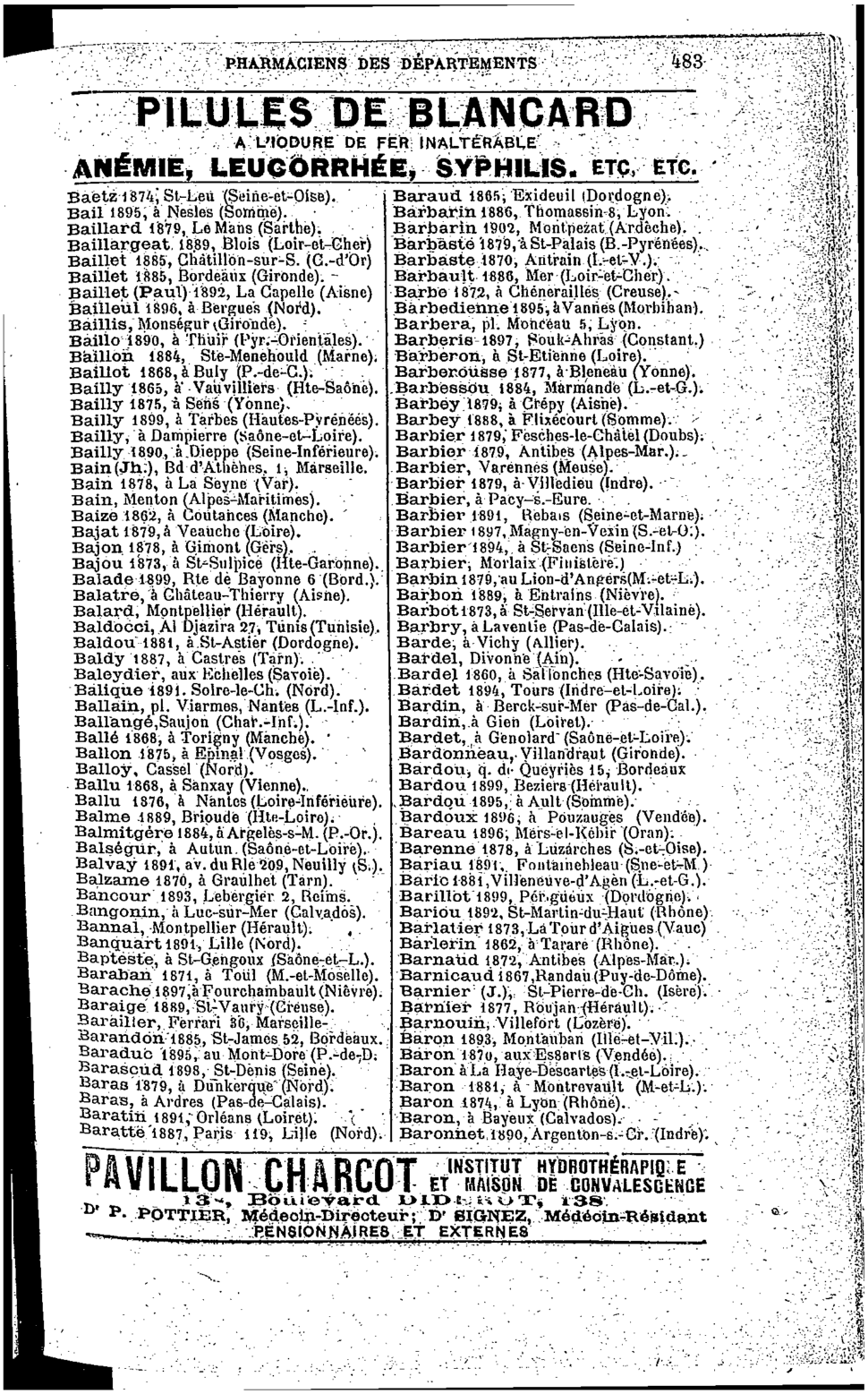
Table 6.4 shows the examples where text only input doesn't work. Example 1, 5 shows that text only input may inherit the errors from raw OCR. Since it doesn't have image input, it is unlikely to find the mistakes that are not so obvious, especially in proper nouns like names. Example 2 and 4 show the risk of over-correction of text-only input. "Bully", "rue" and "Bordeaux" are common words, but in our extraction task, we would prioritise faithfulness over textual prior knowledge. Example 3 shows that in noisy OCR cases, the model, without image, could make wrong corrections. In example 6, only the

Source	Transcription
REF	vicario 1886 pharmacien-chimiste 1re classe lic ès sc laur de l'ec de pharm laboratoire spécial d'analyses boul haussmann 17 angle de la rue au helder
Gemini 3 Flash Preview	vibario 1886 pharmacien-chimiste 1re classe lic ès sc laur de l'ec de pharm laboratoire spécial d'analyses boul haussmann 17 angle de la rue au helder
Gemini 3 Pro Preview	[entry missed completely]
Raw OCR	vibàrio sjï886 phàrmaciencmviislé classe lie bs se laur de vec de pharm laboratoire spécial dana lyses boul haussiïiânn 17 angle dé larue au helder
REF	viron phu en ch de la salpêtrière
Gemini 3 Flash Preview	viron ph en ch de la salpêtrière
Gemini 3 Pro Preview	[entry missed completely]
Raw OCR	viroh ph eh gh de lasàlpùtriere
REF	voiry 1888 réactif voiry eucalyptol voiry boul de courcelles 5
Gemini 3 Flash Preview	voiry 1888 réactif voiry eucalyptol voiry boul de courcelles 5
Gemini 3 Pro Preview	[entry missed completely]
Raw OCR	voiry 1888 réactif voiry èucalypiol voiry boul de courceiles 5
REF	wesson r ph béral paix 14
Gemini 3 Flash Preview	wesson r ph béral paix 14
Gemini 3 Pro Preview	[entry missed completely]
Raw OCR	wesson r phbéral paix 14
REF	yvon et berlioz 1887 feuillade 7
Gemini 3 Flash Preview	yvon ☆et berlioz 1887 feuillade 7
Gemini 3 Pro Preview	[entry missed completely]
Raw OCR	ïvon $fe et berlioz 18s7 feuiliade7

TABLE 6.3
Qualitative comparison of transcription with Original OCR Input. Errors are highlighted in red.

1	REF	baillo 1890 à thuir pyrorientales
	image + text	baillo 1890 à thuir pyrorientales
	image only	baillo 1890 à thuir pyrorientales
	text only	bàillo 1890 à thuir pyrorientales
	raw	bàillo 1890 à thùir pyrorientâles
2	REF	baillot 1868 à buly pdec
	image + text	baillot 1868 à buly pdec
	image only	baillot 1868 à buly pdec
	text only	baillot 1868 à bully pdec
	raw	baiuot 1868 à buly pdec
3	REF	bailly à dampierre saôneetloire
	image + text	bailly à dampierre saôneetloire
	image only	bailly à dampierre saôneetloire
	text only	bailly à dompierre saôneetloire
	raw	bailly à ûarhpierre saôneeuloire
4	REF	balade 1899 rte de bayonne 6 bord
	image + text	balade 1899 rte de bayonne 6 bord
	image only	balade 1899 rte de bayonne 6 bord
	text only	balade 1899 rue de bayonne 6 bordeaux
	raw	balade 1899 rie dé bayenne 6 bprd
5	REF	ballangé saujon charinf
	image + text	ballangé saujon charinf
	image only	ballangé saujon charinf
	text only	bairangé saujon charinf
	raw	bairangésaujpn charihf
6	REF	barbe 1872 à chénerailles creuse
	image + text	barbe 1872 à chénerailles creuse
	image only	barbe 1872 à chénérailles creuse
	text only	barbe 1872 à chénérailles creuse
	raw	barbe 1872 à chéneraillés creuse

TABLE 6.4
Qualitative transcription examples with Gemini 3 Pro Preview. The left label column aligns model sources; only character-level deviations from REF are highlighted in red. Horizontal lines separate examples.



PHARMACIENS DES DÉPARTEMENTS 483

PILULES DE BLANCARD

A L'IODURE DE FER INALTÉRABLE

ANÉMIE, LEUCORRHÉE, SYPHILIS. ETC. ETC.

Baetz 1874, St-Leu (Seine-et-Oise).
Bail 1895, à Nesles (Somme).
Baillard 1879, Le Mans (Sarthe).
Baillargeat 1889, Blois (Loir-et-Cher)
Baillet 1885, Châtillon-sur-S. (C.-d'Or)
Baillet 1885, Bordeaux (Gironde).
Baillet (Paul) 1892, La Capelle (Aisne)
Bailleul 1896, à Bergues (Nord).
Baillis, Monségur (Gironde).
Baillo 1890, à Thuir (Pyr.-Orientales).
Baillon 1884, Ste-Menehould (Marne).
Baillot 1868, à Buly (P.-de-C.).
Bailly 1865, à Vauvilliers (Hte-Saône).
Bailly 1875, à Sens (Yonne).
Bailly 1899, à Tarbes (Hautes-Pyrénées).
Bailly, à Dampierre (Saône-et-Loire).
Bailly 1890, à Dieppe (Seine-Inférieure).
Bain (Jh.), Bd d'Athènes, 1, Marseille.
Bain 1878, à La Seyne (Var).
Bain, Menton (Alpes-Maritimes).
Baize 1862, à Coutances (Manche).
Bajat 1879, à Veauche (Loire).
Bajon 1878, à Gimont (Gers).
Bajou 1873, à St-Sulpice (Hte-Garonne).
Balade 1899, Rte de Bayonne 6 (Bord.).
Balatre, à Château-Thierry (Aisne).
Balard, Montpellier (Hérault).
Baldocci, Al Djazira 27, Tunis (Tunisie).
Baldou 1881, à St-Astier (Dordogne).
Baldy 1887, à Castres (Tarn).
Baleydier, aux Echelles (Savoie).
Balique 1891, Solre-le-Ch. (Nord).
Ballain, pl. Viarmes, Nantes (L.-Inf.).
Ballangé, Saujon (Char.-Inf.).
Ballé 1868, à Torigny (Manche).
Ballon 1875, à Epinal (Vosges).
Balloy, Cassel (Nord).
Ballu 1868, à Sanxay (Vienne).
Ballu 1876, à Nantes (Loire-Inférieure).
Balme 1889, Brioude (Hte-Loire).
Balmitgère 1884, à Argelès-s-M. (P.-Or.).
Balségur, à Autun (Saône-et-Loire).
Balvay 1891, av. du Rle 209, Neuilly (S.).
Balzame 1870, à Graulhet (Tarn).
Bancour 1893, Lebergier 2, Reims.
Bangonin, à Luc-sur-Mer (Calvados).
Bannal, Montpellier (Hérault).
Banquart 1891, Lille (Nord).
Bapteste, à St-Gengoux (Saône-et-L.).
Baraban 1871, à Toul (M.-et-Moselle).
Barache 1897, à Fourchambault (Nièvre).
Baraige 1889, St-Vaury (Creuse).
Barailler, Ferrari 36, Marseille.
Barandon 1885, St-James 52, Bordeaux.
Baraduc 1895, au Mont-Dore (P.-de-D.
Barascud 1898, St-Denis (Seine).
Baras 1879, à Dunkerque (Nord).
Baras, à Ardres (Pas-de-Calais).
Baratin 1891, Orléans (Loiret).
Baratte 1887, Paris 119, Lille (Nord).
Baraud 1865, Exideuil (Dordogne).
Barbarin 1886, Thomassin 8, Lyon.
Barbarin 1902, Montpezat (Ardèche).
Barbaste 1879, à St-Palais (B.-Pyrénées).
Barbaste 1870, Antrain (I.-et-V.).
Barbault 1886, Mer (Loir-et-Cher).
Barbe 1872, à Chénerailles (Creuse).
Barbedienne 1895, à Vannes (Morbihan).
Barbera, pl. Monceau 5, Lyon.
Barberis 1897, Souk-Ahras (Constant.)
Barberon, à St-Etienne (Loire).
Barberousse 1877, à Bleneau (Yonne).
Barbessou 1884, Marmande (L.-et-G.).
Barbey 1879, à Crépy (Aisne).
Barbey 1888, à Flixécourt (Somme).
Barbier 1879, Fesches-le-Châtel (Doubs).
Barbier 1879, Antibes (Alpes-Mar.).
Barbier, Varennes (Meuse).
Barbier 1879, à Villedieu (Indre).
Barbier, à Pacy-s.-Eure.
Barbier 1891, Rebais (Seine-et-Marne).
Barbier 1897, Magny-en-Vexin (S.-et-O.).
Barbier 1894, à St-Saens (Seine-Inf.)
Barbier, Morlaix (Finistère.)
Barbin 1879, au Lion-d'Angers (M.-et-L.).
Barbon 1889, à Entrains (Nièvre).
Barbot 1873, à St-Servan (Ille-et-Vilaine).
Barbry, à Laventie (Pas-de-Calais).
Barde, à Vichy (Allier).
Bardel, Divonne (Ain).
Bardel 1860, à Sallonches (Hte-Savoie).
Bardet 1894, Tours (Indre-et-Loire).
Bardin, à Berck-sur-Mer (Pas-de-Cal.).
Bardin, à Gien (Loiret).
Bardet, à Genolard (Saône-et-Loire).
Bardonneau, Villandraut (Gironde).
Bardou, q. de Queyriès 15, Bordeaux
Bardou 1899, Beziers (Hérault).
Bardou 1895, à Ault (Somme).
Bardoux 1896, à Pouzauges (Vendée).
Bareau 1896, Mers-el-Kébir (Oran).
Barenne 1878, à Luzarches (S.-et-Oise).
Bariau 1891, Fontainebleau (Sne-et-M.)
Baric 1881, Villeneuve-d'Agen (L.-et-G.).
Barillot 1899, Périgueux (Dordogne).
Bariou 1892, St-Martin-du-Haut (Rhône)
Barlatier 1873, La Tour d'Aigues (Vauc)
Barlerin 1862, à Tarare (Rhône).
Barnaud 1872, Antibes (Alpes-Mar.).
Barnicaud 1867, Randau (Puy-de-Dôme).
Barnier (J.), St-Pierre-de-Ch. (Isère).
Barnier 1877, Roujan (Hérault).
Barnouin, Villefort (Lozère).
Baron 1893, Montauban (Ille-et-Vil.).
Baron 1870, aux Essarts (Vendée).
Baron à La Haye-Descartes (I.-et-Loire).
Baron 1881, à Montrevault (M-et-L.).
Baron 1874, à Lyon (Rhône).
Baron, à Bayeux (Calvados).
Baronnet 1890, Argenton-s.-Cr. (Indre).

PAVILLON CHARCOT INSTITUT HYDROTHÉRAPIQUE ET MAISON DE CONVALESCENCE
13 bis, Boulevard DIDEROT, 13 bis
Dr P. POTTIER, Médecin-Directeur; Dr SIGNEZ, Médecin-Résidant
PENSIONNAIRES ET EXTERNES

FIGURE 6.4
1904-0514 Image

iamge+text input gets the correct answer. While chénérailles is a commune in France, in the original picture it is chénerailles.

In conclusion, only OCR text input, without image input, may inherit errors from raw OCR, conduct over correction.

6.6.3 1893-0061 SOMETIMES ONLY IMAGE WORKS BETTER THAN IMAGE+TEXT?

Table 6.5 shows some examples from 1893-0061. Figure 6.5 shows the related entries. Although we asked the model to ignore the decorations in the image since they are non-formal characters, sometimes they still extract them in unexpected ways. Example 1 and 2 show that image+text extracts the decorations while image only didn't, that's why the image only received higher matching rate. This behaviour could be explained that with text input, the model is influenced by the raw OCR text. In these two examples, the raw OCR text input indicates that there is something between the name and the year, and the model tries to extract it, resulting in unnecessary information.

1	REF	aber 1867 lafayette 47 midi à 5
	image+text	aber ✠ ☆ 1867 lafayette 47 midi à 5
	image only	aber 1867 lafayette 47 midi à 5
	raw	aber iji ij 1867 lafayctto47 midi à 5
2	REF	allix 1867 stflorentin 6 midi à 1
	image+text	allix ✠ ☆ 1867 stflorentin 6 midi à 1
	image only	allix 1867 stflorentin 6 midi à 1
	raw	allix iji5£ 1867 stflorenlin 6 midi
3	REF	abadie 1870 mal des yeux volney 9 clin boul stgermain 172 lun mer ven 5
	image+text	abadie ☆ 1870 mal des yeux volney 9 lun mer ven 5
	image only	abadie 1870 ☆ mal des yeux volney 9 lun mer ven 5
	raw	abadieî 1870mal desyeux volney 9 lunmerven 5 clin boul stgermain 172

TABLE 6.5
Qualitative transcription examples showing spurious symbol insertion and OCR corruption with Gemini 3 Pro Preview. Only character-level deviations from the reference are highlighted in red.

In example 3, both image+text and image only input extract the decorations, which shows that image only input could also be influenced by the decoration markers.

However, these special characters could be filtered out easily with rule-based filtering in the next stages, and it won't be a problem for our extraction.

6.7 CONCLUSION AND LIMITATIONS

Based on the quantitative metrics, the optimal configurations are either *image-only* or *image+text* inputs combined with models from the Gemini 3 family. Further qualitative analysis shows that *image-only* inputs can be unstable in certain cases, while the addition of textual input helps correct ambiguities and improves robustness.

When comparing Gemini 3 Pro Preview and Gemini 3 Flash, the performance gap is minimal, with an average difference of about 0.01 in WER. Across different subsets of files, both models yield highly comparable results. As such, the choice between the two should primarily depend on budget constraints and processing-time requirements: when resources allow, Gemini 3 Pro Preview is preferable; otherwise, Gemini 3 Flash remains a strong and reliable alternative. **For the final extraction of the selected dataset**

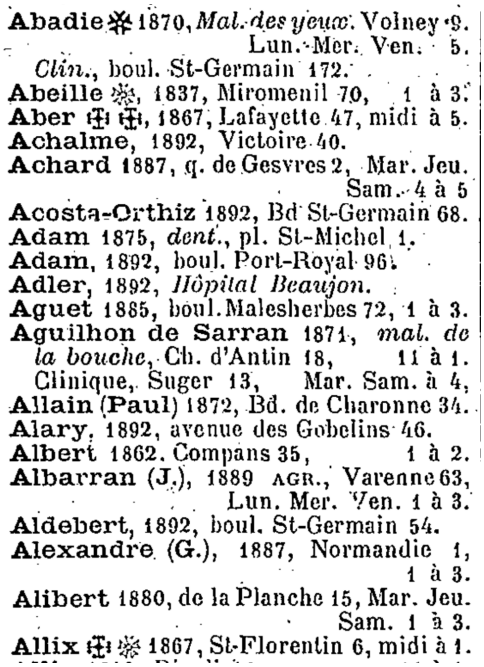

Abadie ✻ 1870, *Mal. des yeux*, Volney 9.
Lun. Mer. Ven. 5.
Clin., boul. St-Germain 172.
Abeille ✻, 1837, Miromenil 70, 1 à 3.
Aber ✠ ✠, 1867, Lafayette 47, midi à 5.
Achalme, 1892, Victoire 40.
Achard 1887, q. de Gesvres 2, Mar. Jeu.
Sam. 4 à 5
Acosta-Orthiz 1892, Bd St-Germain 68.
Adam 1875, *dent.*, pl. St-Michel 1.
Adam, 1892, boul. Port-Royal 96.
Adler, 1892, *Hôpital Beaujon.*
Aguet 1885, boul. Malesherbes 72, 1 à 3.
Aguilhon de Sarran 1871, *mal. de la bouche*, Ch. d'Antin 18, 11 à 1.
Clinique, Suger 13, Mar. Sam. à 4,
Allain (Paul) 1872, Bd. de Charonne 34.
Alary, 1892, avenue des Gobelins 46.
Albert 1862, Compans 35, 1 à 2.
Albarran (J.), 1889 AGR., Varenne 63,
Lun. Mer. Ven. 1 à 3.
Aldebert, 1892, boul. St-Germain 54.
Alexandre (G.), 1887, Normandie 1,
1 à 3.
Alibert 1880, de la Planche 15, Mar. Jeu.
Sam. 1 à 3.
Allix ✠ ✻ 1867, St-Florentin 6, midi à 1.

FIGURE 6.5
related part in 1893-0061

comprising 4,166 pages, we adopted Gemini 3 Pro Preview with an Image+Text (original OCR) input configuration. Section 7 presents an initial analysis of the extracted female physicians.

Several limitations should be noted. Global metrics tend to obscure more fine-grained behaviors, while qualitative analyses are inherently case-based and cannot exhaustively cover all error patterns within a limited analysis time. Moreover, the benchmark itself is not always error-free (e.g., the “phu” case in Table 6.3, Example 2), despite our efforts to verify and clean the reference data.

CHAPTER 7

FEMALE DOCTORS

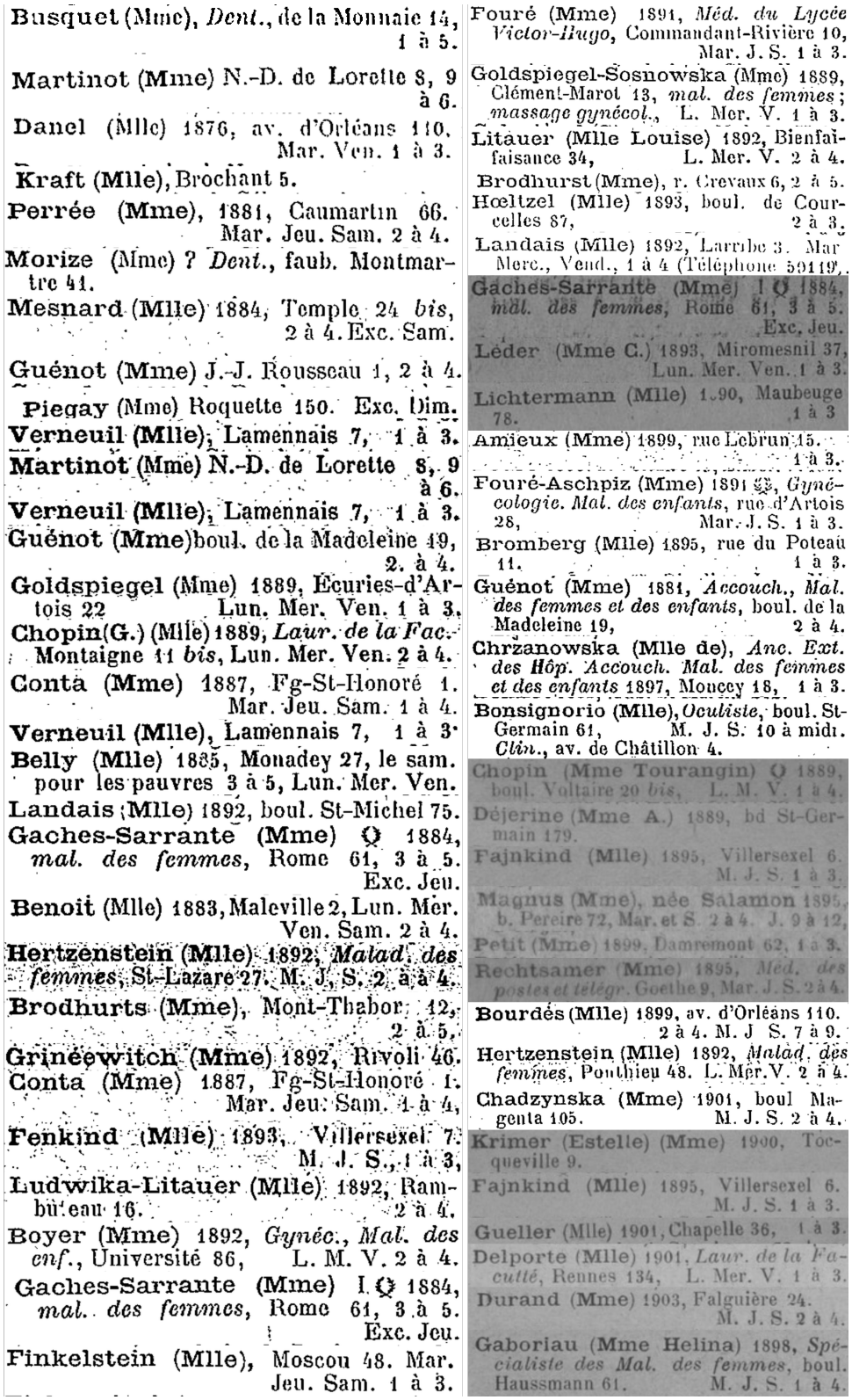
7.1 EXTRACTION QUALITY ANALYSIS

7.1.1 MANUALLY CREATE A PAGE OF FEMALE DOCTORS

Our final target is to extract female doctors from Rosenwald Guide. However, female doctors are sparsely distributed in the book, making it hard to evaluate the extraction accuracy and analyse the extraction quality on our core assets, female doctors.

In order to make the evaluation on female doctors possible, we searched the original-ocr text throughout our target years (1887-1906), find the doctor names suffix containing "Mme" or "Mlle" (3 entries per year, 60 in total) and put them into one page, as is shown in Figure 7.1. We can measure the extraction quality on this page to be aware of the extraction pipeline performance on our most valuable targets.

However, this manually created page also has its limitations. In reality, the female doctors are sparsely distributed, making it different from our concentrated page. In addition, here we only focus on the doctors with "Mlle" and "Mme" suffix, while in the dataset, the doctors' first name could also indicate if it indicates a female doctor. Finally, in our production scenario, we use the original OCR result to aid the image, while in our test here, we use tesseract OCR on our manually created page.



Busquet (Mme), *Dent.*, de la Monnaie 14, 1 à 5.

Martinot (Mme) N.-D. de Lorette 8, 9 à 6.

Danel (Mlle) 1876, av. d'Orléans 110. Mar. Ven. 1 à 3.

Kraft (Mlle), Brochant 5.

Perrée (Mme), 1881, Caumartin 66. Mar. Jeu. Sam. 2 à 4.

Morize (Mme) ? *Dent.*, faub. Montmartre 41.

Mesnard (Mlle) 1884, Temple 24 *bis*, 2 à 4. Exc. Sam.

Guénot (Mme) J.-J. Rousseau 1, 2 à 4.

Piegay (Mme) Roquette 150. Exc. Dim.

Verneuil (Mlle), Lamennais 7, 1 à 3.

Martinot (Mme) N.-D. de Lorette 8, 9 à 6.

Verneuil (Mlle), Lamennais 7, 1 à 3.

Guénot (Mme) boul. de la Madeleine 19, 2 à 4.

Goldspiegel (Mme) 1889, Écuries-d'Artois 22 Lun. Mer. Ven. 1 à 3.

Chopin (G.) (Mlle) 1889, *Laur. de la Fac.* Montaigne 11 *bis*, Lun. Mer. Ven. 2 à 4.

Conta (Mme) 1887, Fg-St-Honoré 1. Mar. Jeu. Sam. 1 à 4.

Verneuil (Mlle), Lamennais 7, 1 à 3.

Belly (Mlle) 1885, Monadey 27, le sam. pour les pauvres 3 à 5, Lun. Mer. Ven.

Landais (Mlle) 1892, boul. St-Michel 75.

Gaches-Sarrante (Mme) Q 1884, *mal. des femmes*, Rome 61, 3 à 5. Exc. Jeu.

Benoit (Mlle) 1883, Maleville 2, Lun. Mer. Ven. Sam. 2 à 4.

Hertzenstein (Mlle) 1892, *Malad. des femmes*, St-Lazare 27. M. J. S. 2 à 4.

Brodhurts (Mme), Mont-Thabor 12, 2 à 5.

Grinéewitch (Mme) 1892, Rivoli 46.

Conta (Mme) 1887, Fg-St-Honoré 1. Mar. Jeu. Sam. 1 à 4.

Fenkind (Mlle) 1893, Villersexel 7. M. J. S. 1 à 3.

Ludwika-Litauer (Mlle) 1892, Rambuteau 16. 2 à 4.

Boyer (Mme) 1892, *Gynéc., Mal. des enf.*, Université 86, L. M. V. 2 à 4.

Gaches-Sarrante (Mme) I Q 1884, *mal. des femmes*, Rome 61, 3 à 5. Exc. Jeu.

Finkelstein (Mlle), Moscou 48. Mar. Jeu. Sam. 1 à 3.

Fouré (Mme) 1891, *Méd. du Lycée Victor-Hugo*, Commandant-Rivière 10, Mar. J. S. 1 à 3.

Goldspiegel-Sosnowska (Mme) 1889, Clément-Marot 13, *mal. des femmes*; *massage gynécol.*, L. Mer. V. 1 à 3.

Litauer (Mlle Louise) 1892, Bienfaisance 34, L. Mer. V. 2 à 4.

Brodhurst (Mme), r. Crevaux 6, 2 à 5.

Hœltzel (Mlle) 1893, boul. de Courcelles 87, 2 à 3.

Landais (Mlle) 1892, Larribe 3. Mar Merc., Vend., 1 à 4 (Téléphone 59119).

Gaches-Sarrante (Mme) I Q 1884, *mal. des femmes*, Rome 61, 3 à 5. Exc. Jeu.

Léder (Mme C.) 1893, Miromesnil 37, Lun. Mer. Ven. 1 à 3.

Lichtermann (Mlle) 1890, Maubeuge 78. 1 à 3

Amieux (Mme) 1899, rue Lebrun 15. 1 à 3.

Fouré-Aschpiz (Mme) 1891 ✠, *Gynécologie. Mal. des enfants*, rue d'Artois 28, Mar. J. S. 1 à 3.

Bromberg (Mlle) 1895, rue du Poteau 11. 1 à 3.

Guénot (Mme) 1881, *Accouch., Mal. des femmes et des enfants*, boul. de la Madeleine 19, 2 à 4.

Chrzanowska (Mlle de), *Anc. Ext. des Hôp. Accouch. Mal. des femmes et des enfants* 1897, Moncey 18, 1 à 3.

Bonsignorio (Mlle), *Oculiste*, boul. St-Germain 61, M. J. S. 10 à midi. *Clin.*, av. de Châtillon 4.

Chopin (Mme Tourangin) Q 1889, boul. Voltaire 20 *bis*, L. M. V. 1 à 4.

Déjerine (Mme A.) 1889, bd St-Germain 179.

Fajnkind (Mlle) 1895, Villersexel 6. M. J. S. 1 à 3.

Magnus (Mme), née Salamon 1895, b. Pereire 72, Mar. et S. 2 à 4. J. 9 à 12.

Petit (Mme) 1899, Damremont 62, 1 à 3.

Rechtsamer (Mme) 1895, *Méd. des postes et télégr.* Goethe 9, Mar. J. S. 2 à 4.

Bourdès (Mlle) 1899, av. d'Orléans 110. 2 à 4. M. J. S. 7 à 9.

Hertzenstein (Mlle) 1892, *Malad. des femmes*, Ponthieu 48. L. Mer. V. 2 à 4.

Chadzynska (Mme) 1901, boul Magenta 105. M. J. S. 2 à 4.

Krimer (Estelle) (Mme) 1900, Tocqueville 9.

Fajnkind (Mlle) 1895, Villersexel 6. M. J. S. 1 à 3.

Gueller (Mlle) 1901, Chapelle 36, 1 à 3.

Delporte (Mlle) 1901, *Laur. de la Faculté*, Rennes 134, L. Mer. V. 1 à 3.

Durand (Mme) 1903, Falguière 24. M. J. S. 2 à 4.

Gaboriau (Mme Helina) 1898, *Spécialiste des Mal. des femmes*, boul. Haussmann 61. M. J. S. 1 à 4.

FIGURE 7.1
Manually Created Page of Female Doctors

7.1.2 QUANTITATIVE ANALYSIS

Table 7.1 reports the extraction results across different models and input settings. Overall, the Gemini 3 model family exhibits substantially lower error rates than the GPT-5 series. However, two observations appear counterintuitive with respect to our production configuration (Gemini 3 Pro with Image+Text input). First, Gemini 3 Flash achieves slightly lower error rates than Gemini 3 Pro. Second, the Image+Text setting combined with Tesseract yields higher error rates than the Image-only input. Although these differences are relatively small, a qualitative inspection of the error cases remains necessary. Given the limited number of female doctors in the corpus, even minor extraction errors may have a disproportionate impact on the final analysis.

TABLE 7.1
Word Error Rate (WER) and Character Error Rate (CER) by input source and model

Source	Model	WER	CER
Image	Gemini-3 Flash (preview)	0.0296	0.0152
	Gemini-3 Pro (preview)	0.0385	0.0205
	GPT-5 Mini (2025-08-07)	0.1627	0.1371
	GPT-5.2 (2025-12-11)	0.2544	0.2325
Image + Text Tesseract	Gemini-3 Flash (preview)	0.0340	0.0231
	Gemini-3 Pro (preview)	0.0429	0.0258
	GPT-5 Mini (2025-08-07)	0.2988	0.2708
	GPT-5.2 (2025-12-11)	0.2559	0.2328
Tesseract	Gemini-3 Flash (preview)	0.7589	0.6387
	Gemini-3 Pro (preview)	0.7825	0.6407
	GPT-5 Mini (2025-08-07)	0.8077	0.6688
	GPT-5.2 (2025-12-11)	0.7766	0.6493

7.1.3 ERROR CASE STUDY

After ignoring mismatches that are purely due to control characters (e.g., $\star$), Table 7.2 summarizes the remaining error cases produced by the Image+Text Gemini 3 Pro configuration. Importantly, none of the observed errors affects gender identification.

In Examples 2, 4, 5, and 7, the output differs from the reference by a single character, typically involving ambiguous glyphs that are difficult to distinguish reliably. In Example 1, the model produces the grammatically correct form *bienfaisance*, whereas the reference contains *bienfaifaisance*, which likely reflects a spelling error in the original record. Example 3 includes an additional field (a telephone number): although this information was not requested, it is extracted correctly and does not interfere with the target fields. Example 6 contains two addresses, which exceeds our predefined schema; the model only omits one title (*Clin*) associated with an address while correctly extracting the remaining information, though in a slightly different order than the reference.

Overall, these errors have no substantive impact on gender recognition. They mainly consist of minor character-level confusions on unclear glyphs and occasional overflow when the record contains information outside the predefined paradigm. We therefore conclude that our extraction pipeline (Image+Text with Gemini 3 Pro Preview) provides reliable performance for female doctor information extraction.

TABLE 7.2
Character-level differences between the reference line and the `Image+Text Gemini 3 Pro` output (differences highlighted in **red**).

ID	Text (Reference on top; Model output below)
1	**Ref:** litauer mlle louise 1892 bienfai**fai**sance 34 l mer v 2 à 4 **Gemini:** litauer mlle louise 1892 bienfaisance 34 l mer v 2 à 4
2	**Ref:** h**œ**ltzel mlle 1893 boul de courcelles 87 2 à 3 **Gemini:** h**oe**ltzel mlle 1893 boul de courcelles 87 2 à 3
3	**Ref:** landais mlle 1892 larr**i**be 3 mar merc vend 1 à 4 **Gemini:** landais mlle 1892 larr**a**be 3 mar merc vend 1 à 4 **téléphone 59119**
4	**Ref:** l**e**der mme c 1893 miromesnil 37 lun mer ven 1 à 3 **Gemini:** l**é**der mme c 1893 miromesnil 37 lun mer ven 1 à 3
5	**Ref:** chrzanowska mlle de 1897 anc ext des hôp accouch mal des femmes et des enfants mo**n**cey 18 1 à 3 **Gemini:** chrzanowska mlle de 1897 anc ext des hôp accouch mal des femmes et des enfants mo**u**cey 18 1 à 3
6	**Ref:** bonsignorio mlle oculiste **clin** boul stgermain 61 clin av de châtillon 4 **m j s 10 à midi** **Gemini:** bonsignorio mlle oculiste boul stgermain 61 **m j s 10 à midi** clin av de châtillon 4
7	**Ref:** petit mme 1899 damr**e**mont 62 1 à 3 **Gemini:** petit mme 1899 damr**é**mont 62 1 à 3

7.2 FEMALE DOCTORS QUANTITY AND DISTRIBUTION

First we analyse the count of Mme and Mlle throughout the pages, as is shown in Figure 7.2. We can see 3420 out of 4166 files (over 82%) have no Mme or Mlle at all. The file count decreases nearly exponentially if we increase the total count of Mme and Mlle, illustrating a long tail distribution. This distribution over count of Mme and Mlle shows that the female doctors are sparsely distributed throughout the pages.

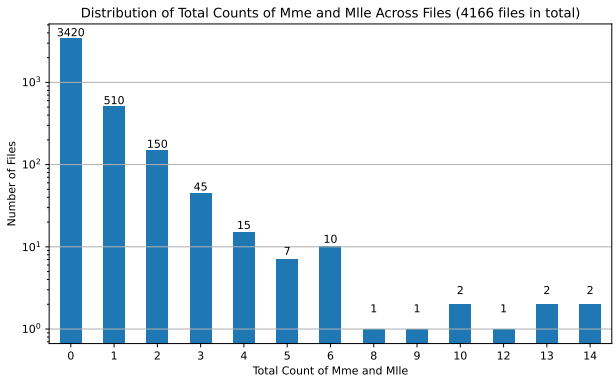


FIGURE 7.2
Distribution of File Count Over the Total Count of Mme and Mlle

Figure 7.3 shows the count of Mme and Mlle for each page. X axis is sorted by the year and page, in ascending order. For the first page of each year, we add the year label on x axis. We can see that as the years goes more recent, there are more files that have non zero Mme and Mlle count, and the peak value also has an increasing trend. It could be explained as more female doctors are noted in Rosenwald Guide in the later years.

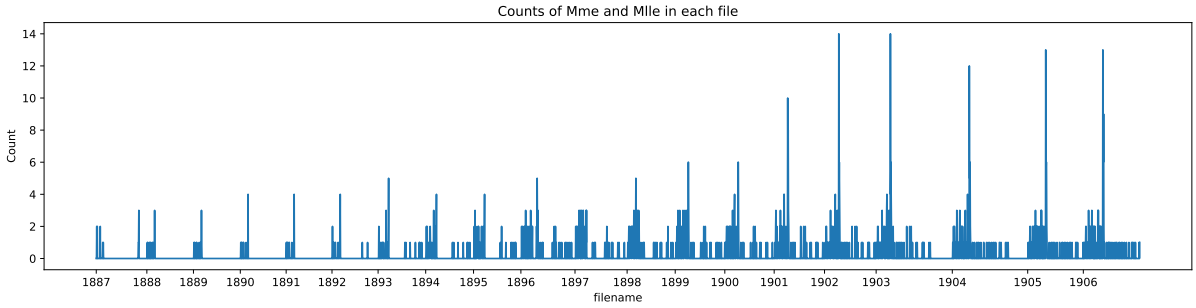


FIGURE 7.3
Count of Mme and Mlle for each page

Another interesting phenomenon is that the count of Mme and Mlle has a periodic appearance, especially after 1898. The count would quickly come to a peak and then drop. In order to look further into this pattern, we have Figure 7.4, showing the Count of Mme and Mlle, sorted descendingly by total Mme and Mlle count.

In the top five peaks, all pages come from the 1902–1906 volumes, and their page numbers cluster within a narrow range (129, 139, 152, 159, 169). Upon inspection, we found that these pages correspond to the opening page of the section *"LISTE ALPHABÉTIQUE DES OFFICIERS DE SANTÉ, CHIRURGIENS-DENTISTES DIPLÔMÉS ET DENTISTES, OFFICIERS DE SANTÉ, PARIS."* Table 7.3 presents the female doctors listed on page 139 in the 1902 volume. We then counted how often each name appears across

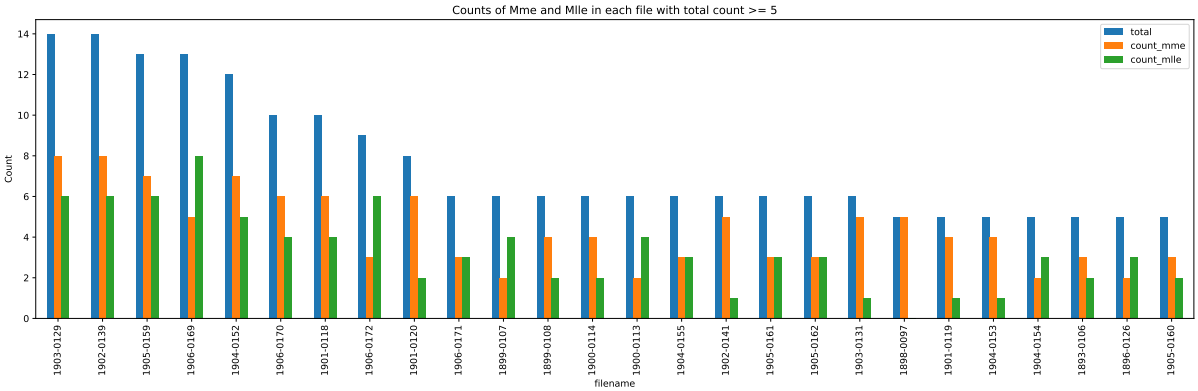
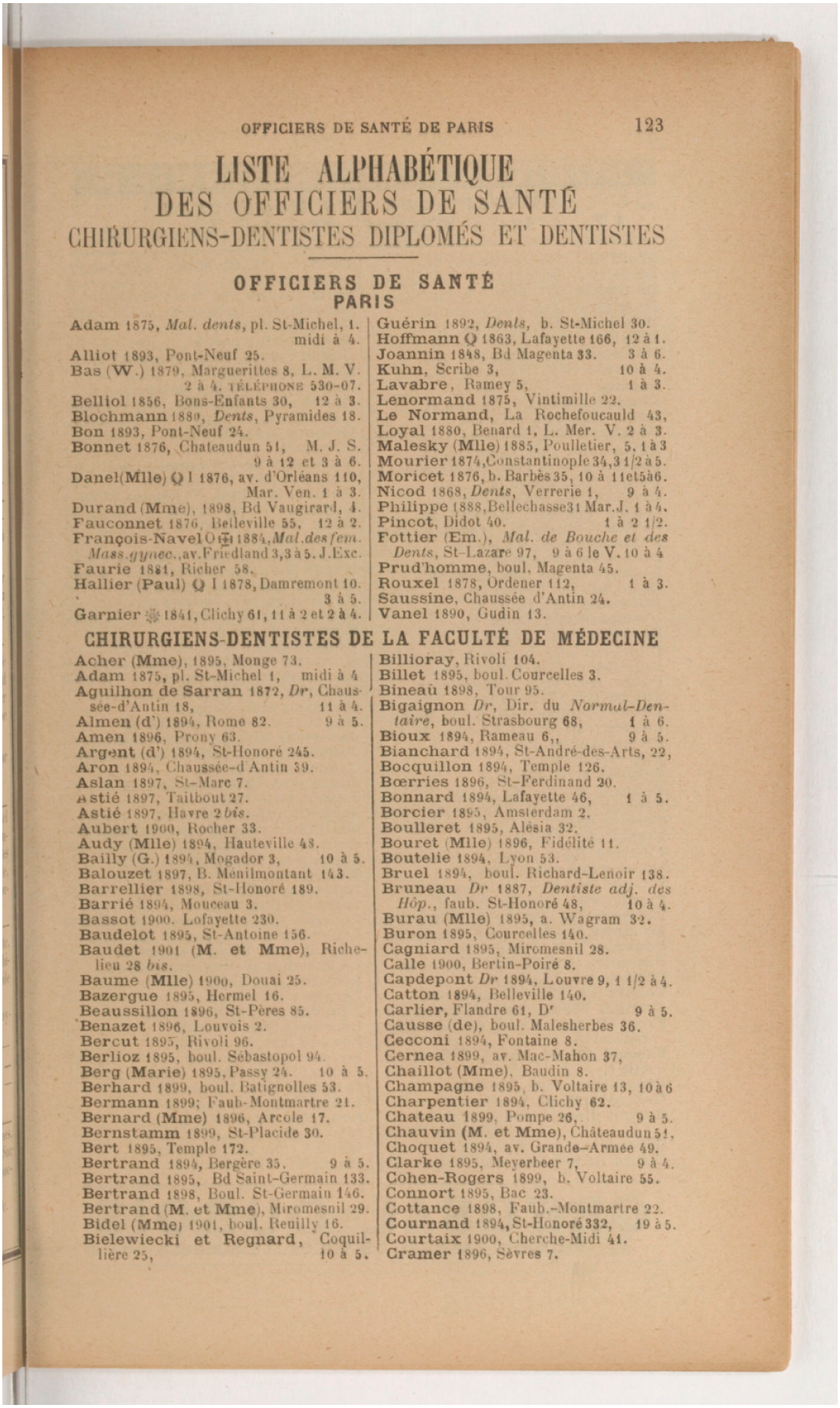


FIGURE 7.4
Counts of Mme and Mlle in each file with total count >= 5

these five pages: 12 of the 14 names occur four or five times, indicating that the top-five peaks in the total counts of *Mme* and *Mlle* are largely driven by the same set of female doctors. These high-frequency entries therefore provide valuable longitudinal samples for analyzing how individual records change over time.

Nom	#Appearance	Année	Notes	Adresse	Horaires	Sexe
Danel (Mlle)	5	1876	⋆ I	av. d'Orléans 110	Mar. Ven. 1 à 3	Mlle
Durand (Mme)	3	1898		Bd Vaugirard, 4		Mme
Acher (Mme)	5	1895		Monge 73		Mme
Audy (Mlle)	5	1894		Hauteville 43		Mlle
Baudet (M. et Mme)	4	1901		Richelieu 28 bis		Mme
Baume (Mlle)	2	1900		Douai 25		Mlle
Bernard (Mme)	5	1896		Arcole 17		Mme
Bertrand (M. et Mme)	5			Miromesnil 29		Mme
Bidel (Mme)	5	1901		boul. Reuilly 16		Mme
Malesky (Mlle)	5	1885		Poulletier, 5	1 à 3	Mlle
Bouret (Mlle)	5	1896		Fidélité 11		Mlle
Burau (Mlle)	5	1895		a. Wagram 32		Mlle
Chaillot (Mme)	4			Baudin 8		Mme
Chauvin (M. et Mme)	4			Châteaudun 51		Mme

TABLE 7.3
1902 page 139 female doctors (nom, #appearance, année, notes, adresse, horaires, sexe).



OFFICIERS DE SANTÉ DE PARIS 123

LISTE ALPHABÉTIQUE DES OFFICIERS DE SANTÉ CHIRURGIENS-DENTISTES DIPLOMÉS ET DENTISTES

OFFICIERS DE SANTÉ PARIS

Adam 1875, *Mal. dents*, pl. St-Michel, 1. midi à 4.
Alliot 1893, Pont-Neuf 25.
Bas (W.) 1879, Marguerittes 8, L. M. V. 2 à 4. TÉLÉPHONE 530-07.
Belliol 1856, Bons-Enfants 30, 12 à 3.
Blochmann 1889, *Dents*, Pyramides 18.
Bon 1893, Pont-Neuf 24.
Bonnet 1876, Chateaudun 51, M. J. S. 9 à 12 et 3 à 6.
Danel (Mlle) ✪ I 1876, av. d'Orléans 110, Mar. Ven. 1 à 3.
Durand (Mme), 1898, Bd Vaugirard, 4.
Fauconnet 1876, Belleville 55, 12 à 2.
François-Navel ✪ ✠ 1884, *Mal. des fem. Mass. gynec.*, av. Friedland 3, 3 à 5. J. Exc.
Faurie 1881, Richer 58.
Hallier (Paul) ✪ I 1878, Damremont 10. 3 à 5.
Garnier ✠ 1841, Clichy 61, 11 à 2 et 2 à 4.
Guérin 1892, *Dents*, b. St-Michel 30.
Hoffmann ✪ 1863, Lafayette 166, 12 à 1.
Joannin 1848, Bd Magenta 33. 3 à 6.
Kuhn, Scribe 3, 10 à 4.
Lavabre, Ramey 5, 1 à 3.
Lenormand 1875, Vintimille 22.
Le Normand, La Rochefoucauld 43,
Loyal 1880, Benard 1, L. Mer. V. 2 à 3.
Malesky (Mlle) 1885, Poulletier, 5, 1 à 3
Mourier 1874, Constantinople 34, 3 1/2 à 5.
Moricet 1876, b. Barbès 35, 10 à 11 et 5 à 6.
Nicod 1868, *Dents*, Verrerie 1, 9 à 4.
Philippe 1888, Bellechasse 31 Mar. J. 1 à 4.
Pincot, Didot 40. 1 à 2 1/2.
Fottier (Em.), *Mal. de Bouche et des Dents*, St-Lazare 97, 9 à 6 le V. 10 à 4
Prud'homme, boul. Magenta 45.
Rouxel 1878, Ordener 112, 1 à 3.
Saussine, Chaussée d'Antin 24.
Vanel 1890, Gudin 13.

CHIRURGIENS-DENTISTES DE LA FACULTÉ DE MÉDECINE

Acher (Mme), 1895, Monge 73.
Adam 1875, pl. St-Michel 1, midi à 4
Aguilhon de Sarran 1872, *Dr*, Chaussée-d'Antin 18, 11 à 4.
Almen (d') 1894, Rome 82. 9 à 5.
Amen 1896, Prony 63.
Argent (d') 1894, St-Honoré 245.
Aron 1894, Chaussée-d'Antin 39.
Aslan 1897, St-Marc 7.
Astié 1897, Taitbout 27.
Astié 1897, Havre 2 *bis*.
Aubert 1900, Rocher 33.
Audy (Mlle) 1894, Hauteville 48.
Bailly (G.) 1894, Mogador 3, 10 à 5.
Balouzet 1897, B. Ménilmontant 143.
Barrellier 1898, St-Honoré 189.
Barrié 1894, Mouceau 3.
Bassot 1900, Lafayette 230.
Baudelot 1895, St-Antoine 156.
Baudet 1901 (M. et Mme), Richelieu 28 *bis*.
Baume (Mlle) 1900, Douai 25.
Bazergue 1895, Hermel 16.
Beaussillon 1896, St-Pères 85.
Benazet 1896, Louvois 2.
Bercut 1895, Rivoli 96.
Berlioz 1895, boul. Sébastopol 94.
Berg (Marie) 1895, Passy 24. 10 à 5.
Berhard 1899, boul. Batignolles 53.
Bermann 1899; Faub-Montmartre 21.
Bernard (Mme) 1896, Arcole 17.
Bernstamm 1899, St-Placide 30.
Bert 1895, Temple 172.
Bertrand 1894, Bergère 35. 9 à 5.
Bertrand 1895, Bd Saint-Germain 133.
Bertrand 1898, Boul. St-Germain 146.
Bertrand (M. et Mme), Miromesnil 29.
Bidel (Mme) 1901, boul. Reuilly 16.
Bielewiecki et Regnard, Coquillière 25, 10 à 5.
Billioray, Rivoli 104.
Billet 1895, boul. Courcelles 3.
Bineau 1898, Tour 95.
Bigaignon *Dr*, Dir. du *Normal-Dentaire*, boul. Strasbourg 68, 1 à 6.
Bioux 1894, Rameau 6,, 9 à 5.
Bianchard 1894, St-André-des-Arts, 22,
Bocquillon 1894, Temple 126.
Bœrries 1896, St-Ferdinand 20.
Bonnard 1894, Lafayette 46, 1 à 5.
Borcier 1895, Amsterdam 2.
Boulleret 1895, Alésia 32.
Bouret (Mlle) 1896, Fidélité 11.
Boutelie 1894, Lyon 53.
Bruel 1894, boul. Richard-Lenoir 138.
Bruneau *Dr* 1887, *Dentiste adj. des Hôp.*, faub. St-Honoré 48, 10 à 4.
Burau (Mlle) 1895, a. Wagram 32.
Buron 1895, Courcelles 140.
Cagniard 1895, Miromesnil 28.
Calle 1900, Bertin-Poiré 8.
Capdepont *Dr* 1894, Louvre 9, 1 1/2 à 4.
Catton 1894, Belleville 140.
Carlier, Flandre 61, D' 9 à 5.
Causse (de), boul. Malesherbes 36.
Cecconi 1894, Fontaine 8.
Cernea 1899, av. Mac-Mahon 37,
Chaillot (Mme), Baudin 8.
Champagne 1895, b. Voltaire 13, 10 à 6
Charpentier 1894, Clichy 62.
Chateau 1899, Pompe 26, 9 à 5.
Chauvin (M. et Mme), Châteaudun 51,
Choquet 1894, av. Grande-Armée 49.
Clarke 1895, Meyerbeer 7, 9 à 4.
Cohen-Rogers 1899, b. Voltaire 55.
Connort 1895, Bac 23.
Cottance 1898, Faub.-Montmartre 22.
Cournand 1894, St-Honoré 332, 19 à 5.
Courtaix 1900, Cherche-Midi 41.
Cramer 1896, Sèvres 7.

FIGURE 7.5
1902 Page 139

7.3 LIMITATION

In the brief analysis of this part, we only recognize female doctors by the most significant label "Mme" and "Mlle", while they could also be identified by other information, for example, the first name, if any. However, we do not have enough time to conduct further anlaysis in this work, and it remains a valuable direction for future work.

CHAPTER 8

CONCLUSION

This report explored a practical workflow for turning scanned pages of the *Rosenwald Guides* into structured data, motivated by the MEDIF project's interest in making women doctors more visible in historical sources. Working with a pilot corpus (20 editions, 1887–1906; 4,116 pages), we focused on two main needs: creating reliable gold labels under limited resources, and assessing how well current OCR/MLLM-based extraction can support structured transcription.

We proposed the *Double Triangle* workflow, where two independent multimodal systems produce initial structured outputs, high-agreement cases are accepted with minimal intervention, and disagreements are routed to human review. Using this setup, we constructed a benchmark of 60 columns. Agreement on the benchmarked cases exceeds 85%, and in the final review stage 991 fields were corrected out of 13,595 total fields (7.2%). These numbers suggest that cross-model agreement can reduce the amount of human correction needed, although agreement does not guarantee correctness when errors are correlated.

In the extraction experiments, we compared several approaches and input modalities. Overall, the results support using image-based inputs (image-only or image+text) when OCR noise is substantial, while qualitative inspection also indicates that both modalities have characteristic failure modes (e.g., decorative symbols, OCR artifacts, and anchoring effects). Simple preprocessing, such as removing advertisements and splitting pages into columns, consistently improves OCR quality and is therefore a useful low-cost step.

We also conducted initial analysis on the extracted female doctors. However, due to time constraint, we only recognize the female doctors by "Mme" and "Mlle" in their name field, while their first name, if any, could also be used to infer the gender.

There are important limitations. Aggregate metrics can hide field-specific behavior, the benchmark may contain residual errors, and the pilot years do not capture the full variation across later volumes. Future work could focus on field-level evaluation, lightweight validation rules, improved procedures to reduce correlated errors, and scaling the pipeline to additional years in the 1887–1940 collection.

We have put all the code and extracted data on GitHub

- Double Triangular Annotation Framework: `https://github.com/nmrenyi/double-triangle-annotation`
- Extraction Pipeline: `https://github.com/nmrenyi/extraire-tesseract-openai`
- Rosenwald Benchmark (created with Double Triangular Annotation Framework): `https://github.com/nmrenyi/rosenwald-benchmark`

- Rosenwald Guide Extraction Result (created with Extraction Pipeline): `https://github.com/nmrenyi/rosenwald-extraction`

BIBLIOGRAPHY

Smith, R. (2007). 'An Overview of the Tesseract OCR Engine'. In: *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*. IEEE Computer Society, pp. 629–633. DOI: 10.1109/ICDAR.2007.4376991. URL: https://doi.org/10.1109/ICDAR.2007.4376991.

Wick, Christoph, Christian Reul and Frank Puppe (2020). 'Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition'. In: *Digit. Humanit. Q.* 14.2. URL: http://www.digitalhumanities.org/dhq/vol/14/2/000451/000451.html.

EasyOCR (2020). https://github.com/JaidedAI/EasyOCR.

kraken (n.d.). https://github.com/mittagessen/kraken. accessed: 2025-11-30.

Poznanski, Jake *et al.* (2025). *olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models*. arXiv: 2502.18443 [cs.CL]. URL: https://arxiv.org/abs/2502.18443.

Löffler, Kevin (2023). 'Digitize historic architectural plans with OCR and NER transformer models'. Other thesis. OST Ostschweizer Fachhochschule. URL: https://eprints.ost.ch/id/eprint/1189.

Fujitake, Masato (2024). 'DTrOCR: Decoder-only Transformer for Optical Character Recognition'. In: *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*. IEEE, pp. 8010–8020. DOI: 10.1109/WACV57701.2024.00784. URL: https://doi.org/10.1109/WACV57701.2024.00784.

Kim, Seorin *et al.* (2025). 'Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records'. In: *CoRR* abs/2501.11623. DOI: 10.48550/ARXIV.2501.11623. arXiv: 2501.11623. URL: https://doi.org/10.48550/arXiv.2501.11623.

Nunes, Guilherme *et al.* (Aug. 2025). 'Benchmarking Table Extraction: Multimodal LLMs vs Traditional OCR'. In: *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*. Ed. by Hao Fei *et al.* Vienna, Austria: Association for Computational Linguistics, pp. 8–15. ISBN: 979-8-89176-286-2. DOI: 10.18653/v1/2025.xllm-1.2. URL: https://aclanthology.org/2025.xllm-1.2/.

Bai, Zechen *et al.* (2024). 'Hallucination of Multimodal Large Language Models: A Survey'. In: *CoRR* abs/2404.18930. DOI: 10.48550/ARXIV.2404.18930. arXiv: 2404.18930. URL: https://doi.org/10.48550/arXiv.2404.18930.

Leng, Sicong *et al.* (2024). 'The Curse of Multi-Modalities: Evaluating Hallucinations of Large Multimodal Models across Language, Visual, and Audio'. In: *CoRR* abs/2410.12787. DOI: 10.48550/ARXIV.2410.12787. arXiv: 2410.12787. URL: https://doi.org/10.48550/arXiv.2410.12787.

Nguyen, Thi Tuyet Hai *et al.* (July 2021). 'Survey of Post-OCR Processing Approaches'. In: *ACM Comput. Surv.* 54.6. ISSN: 0360-0300. DOI: 10.1145/3453476. URL: https://doi.org/10.1145/3453476.

Kanerva, Jenna *et al.* (2025). 'OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches'. In: *CoRR* abs/2502.01205. DOI: 10.48550/ARXIV.2502.01205. arXiv: 2502.01205. URL: https://doi.org/10.48550/arXiv.2502.01205.

Machidon, Octavian Mihai and Alina L. Machidon (2025). 'Comparing OCR Pipelines for Folkloristic Text Digitization'. In: *CoRR* abs/2507.19092. DOI: 10.48550/ARXIV.2507.19092. arXiv: 2507.19092. URL: https://doi.org/10.48550/arXiv.2507.19092.

Thomas, Alan, Robert Gaizauskas and Haiping Lu (May 2024). 'Leveraging LLMs for Post-OCR Correction of Historical Newspapers'. In: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. Ed. by Rachele Sprugnoli and Marco Passarotti. Torino, Italia: ELRA and ICCL, pp. 116–121. URL: https://aclanthology.org/2024.lt4hala-1.14/.

Lewis, Mike *et al.* (July 2020). 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky *et al.* Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703/.

Do, Thao *et al.* (2025). 'Reference-Based Post-OCR Processing with LLM for Precise Diacritic Text in Historical Document Recognition'. In: *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*. Ed. by Toby Walsh, Julie Shah and Zico Kolter. AAAI Press, pp. 27951–27959. DOI: 10.1609/AAAI.V39I27.35012. URL: https://doi.org/10.1609/aaai.v39i27.35012.

Bourne, Jonathan (2024). 'CLOCR-C: Context Leveraging OCR Correction with Pre-trained Language Models'. In: *CoRR* abs/2408.17428. DOI: 10.48550/ARXIV.2408.17428. arXiv: 2408.17428. URL: https://doi.org/10.48550/arXiv.2408.17428.

Boros, Emanuela *et al.* (Mar. 2024). 'Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study'. In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*. Ed. by Yuri Bizzoni *et al.* St. Julians, Malta: Association for Computational Linguistics, pp. 133–159. URL: https://aclanthology.org/2024.latechclfl-1.14/.

ABBYY OCR (n.d.). URL: https://www.abbyy.com.

Adobe Acrobat OCR (n.d.). URL: https://www.adobe.com/acrobat/online/ocr-pdf.html.

Microsoft Azure OCR (n.d.). https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-ocr.

Google Cloud Vision OCR (n.d.). https://cloud.google.com/vision/docs/ocr.

Amazon Textract (n.d.). https://aws.amazon.com/textract/.

Ding, Bosheng *et al.* (2023). 'Is GPT-3 a Good Data Annotator?' In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber and Naoaki Okazaki. Association for Computational Linguistics, pp. 11173–11195. DOI: 10.18653/V1/2023.ACL-LONG.626. URL: https://doi.org/10.18653/v1/2023.acl-long.626.

Wang, Shuohang *et al.* (2021). 'Want To Reduce Labeling Cost? GPT-3 Can Help'. In: *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*. Ed. by Marie-Francine Moens *et al.* Association for Computational Linguistics, pp. 4195–4205. DOI: 10.18653/V1/2021.FINDINGS-EMNLP.354. URL: https://doi.org/10.18653/v1/2021.findings-emnlp.354.

Kim, Hannah *et al.* (2024). 'MEGAnno+: A Human-LLM Collaborative Annotation System'. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*. Ed. by

Nikolaos Aletras and Orphée De Clercq. Association for Computational Linguistics, pp. 168–176. URL: https://aclanthology.org/2024.eacl-demo.18.

Wang, Xinru *et al.* (2024). 'Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels'. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*. Ed. by Florian 'Floyd' Mueller *et al.* ACM, 303:1–303:21. DOI: 10.1145/3613904.3641960. URL: https://doi.org/10.1145/3613904.3641960.

Zhang, Yulong *et al.* (2025). 'Consensus Entropy: Harnessing Multi-VLM Agreement for Self-Verifying and Self-Improving OCR'. In: *CoRR* abs/2504.11101. DOI: 10.48550/ARXIV.2504.11101. arXiv: 2504.11101. URL: https://doi.org/10.48550/arXiv.2504.11101.

Pangakis, Nicholas and Samuel Wolken (2025). 'Keeping Humans in the Loop: Human-Centered Automated Annotation with Generative AI'. In: *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media, June 23-26, 2025, Copenhagen, Denmark*. Ed. by Jisun An *et al.* AAAI Press, pp. 1471–1492. DOI: 10.1609/ICWSM.V19I1.35883. URL: https://doi.org/10.1609/icwsm.v19i1.35883.

Schroeder, Hope, Deb Roy and Jad Kabbara (2025). 'Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks'. In: *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*. Ed. by Wanxiang Che *et al.* Association for Computational Linguistics, pp. 25771–25795. URL: https://aclanthology.org/2025.findings-acl.1323/.

Gu, Feng *et al.* (2025). 'Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators'. In: *CoRR* abs/2503.06778. DOI: 10.48550/ARXIV.2503.06778. arXiv: 2503.06778. URL: https://doi.org/10.48550/arXiv.2503.06778.

Mohta, Jay *et al.* (16 Dec 2023). 'Are large language models good annotators?' In: *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*. Ed. by Javier Antorán *et al.* Vol. 239. Proceedings of Machine Learning Research. PMLR, pp. 38–48. URL: https://proceedings.mlr.press/v239/mohta23a.html.

Chen, Hui *et al.* (2025). 'A human-LLM collaborative annotation approach for screening articles on precision oncology randomized controlled trials'. In: *BMC Medical Research Methodology* 25.1, pp. 1–8.