Double Triangle Annotation: A Scalable Human-in-the-Loop Framework for High-Precision Historical Document Annotation

Anonymous ACL submission

Abstract

Evaluating structured-information extraction from historical documents at scale requires high-precision ground-truth annotations, yet traditional manual labeling is expensive and fully automated pipelines built on large language models are prone to hallucination. We propose **Double Triangle Annotation**, a two-layer human-in-the-loop framework that leverages cross-model consensus to automate the majority of annotation work while ensuring high-precision outputs. In the first layer, two architecturally independent Multimodal Large Language Models annotate each document in parallel; when they agree, the label is auto-accepted, and disagreements are routed to a human jury. A second layer cross-checks two such systems against each other, escalating residual conflicts to a domain expert. The framework rests on a single assumption—error independence between models—requires no distributional priors or task-specific calibration, and becomes more autonomous as model capability improves. On the *Guides Rosenwald*, a corpus of French medical directories spanning 1887–1906, the framework achieves a final Word Error Rate of 0.003. Applied at scale, model consensus auto-accepts over 85% of 13,595 fields. We release the resulting benchmark—the first structured-extraction ground truth for the Rosenwald Guides—to support future work on historical document processing.

1 Introduction

Historical archives—medical directories, census records, notarial registers—hold structured information that is critical for research in the digital humanities. Extracting this information at scale requires converting degraded page images into machine-readable records, a task that demands high-precision ground-truth annotations for both training and evaluation. Yet producing such annotations remains a bottleneck: traditional independent manual annotation is expensive and difficult to scale. Recent Multimodal Large Language Models (MLLMs) have shown increasingly strong performance on document understanding tasks, but the reliability of any single model remains insufficient for gold-standard annotation due to hallucination. The fundamental challenge is therefore not whether models can assist annotation, but how to *verify* their outputs efficiently.

We propose the **Double Triangle Annotation** framework, a two-layer human-in-the-loop paradigm that addresses this challenge through *consensus-based filtering*. In the first layer, two architecturally independent models annotate each document in parallel; when they agree, the label is auto-accepted, **based on our assumption that independent strong models are unlikely to produce the same error on the same field**. Disagreements are routed to a human jury, shifting the annotator's role from producer to verifier. A second layer cross-checks two such systems against each other, **reapplying the same assumption**, and escalating residual conflicts to a domain expert. The design minimizes human effort while inducing a cascaded error reduction that drives the final error rate toward zero.

We evaluate the framework on the *Guides Rosenwald*, a corpus of French medical directories spanning 1887–1906. In a controlled evaluation on 380 fields, the first-layer juries correct 18.4% and 26.3% of fields respectively, and the final reviewer corrects a further 4.2%, yielding a final Word Error Rate of 0.003. Applied at scale to 105 sampled pages, model consensus auto-accepts over 85% of the 13,595 fields in the final benchmark.

Our contributions are as follows:

1. We introduce the **Double Triangle Annotation** framework (§3), a two-layer paradigm that leverages cross-model consensus to automate the majority of annotation work while

reserving human effort strictly for ambiguous cases. On our evaluation data, model consensus auto-accepts over 73% of fields while achieving a final WER of 0.003; applied at scale, the auto-accept rate rises to over 85%.

2. The framework rests on a single, intuitive assumption—that independent strong models are unlikely to produce the same error on the same field—requiring no task-specific calibration or prior assumption on annotation distribution. This **minimal-assumption design** makes it directly applicable to any structured extraction task where two or more models can be run in parallel.

3. The architecture is **future-proof**: as model capability improves, the framework will benefit from the resulting increase in consensus, further reducing human effort and producing even higher-precision outputs.

4. We apply the framework to construct the first **structured-extraction benchmark** for the Rosenwald Guides (§6), comprising 60 document columns and 13,595 annotated fields, and release it to support future work on historical document processing.

2 Related Work

2.1 Historical Document Annotation

Existing approaches on constructing annotated benchmarks for historical documents converge on a common two-phase paradigm: apply automated processing first, then route the output to human reviewers for correction. In the simplest variant, projects such as DIVA-HisDB (Simistira et al., 2016), SCUT-COUCH (Yan et al., 2010), and the Japanese historical document dataset of Shen et al. (2020) employ dedicated pipelines for binarisation, line segmentation, character segmentation, or layout analysis. Their outputs are then corrected in a single pass by human annotators.

More sophisticated approaches invest in multi-round quality control. M5HisDoc (Shi et al., 2023) applies three rounds of strict manual checking after machine-assisted extraction. CASIA-AHCDB (Xu et al., 2019) adds a classifier-based confidence filter that selectively routes low-confidence items to cross-validation—a form of selective review that reduces effort relative to exhaustive checking, though it relies on a single system's internal confidence rather than agreement between independent systems. At larger scales, ENP (Clausner et al., 2015) and IMPACT (Papadopoulos et al., 2013) outsource annotation to commercial service providers and enforce structured verification combining automated guideline checks with manual inspection.

Crucially, while methods such as CASIA-AHCDB reduce review volume through single-system confidence filtering, none of these approaches leverage *agreement between independent systems* as a quality signal.

2.2 Human–LLM Collaborative Annotation

LLMs can substantially reduce annotation cost. Ding et al. (2023) show that GPT-3 produces usable annotations across a range of tasks at a fraction of human cost, though quality deteriorates on complex tasks involving subtle distinctions or domain-specific knowledge. The dominant hybrid strategy is *confidence-based selective review*: the LLM labels all items and a confidence score determines which ones are routed to human annotators. Wang et al. (2021) report 50–96% cost reductions with this approach by having humans re-label only the lowest-confidence items; Kim et al. (2024) build a similar pipeline in MEGAnno+. Wang et al. (2024) extend the pattern by deploying two LLMs—one for labeling and one for generating confidence scores—though the quality signal remains single-system confidence rather than inter-model agreement.

However, these workflows introduce systematic risks. Schroeder et al. (2025) find that presenting annotators with LLM-generated labels increases their self-reported confidence but does not accelerate annotation, and introduces *anchoring bias* whereby annotators defer to the model's suggestion. Gu et al. (2025) reach a complementary conclusion: LLMs are effective annotation *assistants* but unreliable as independent annotators, and humans working alongside them are susceptible to automation bias. Pangakis and Wolken (2025) observe that LLM annotators tend toward high recall but low precision, and that iterative human-in-the-loop prompt refinement only partially mitigates this imbalance. Additionally, LLM annotation quality can degrade on non-English data (Mohta et al., 2023), a concern relevant to our French extraction task.

Our framework sidesteps the confidence-based paradigm entirely, using *agreement between independent models* rather than a single model's self-assessed uncertainty as the quality signal. This

design also mitigates anchoring bias, because the human jury is presented with two conflicting outputs rather than a single LLM suggestion to accept or reject.

2.3 Consensus-Based Verification

Recent work leverages cross-model consensus as a quality signal for LLM outputs. Zhang et al. (2025) propose Consensus Entropy, a metric that quantifies agreement among multiple vision-language models and routes low-agreement OCR outputs to a stronger expert model—achieving fully automated quality control but offering no human fallback when all models share the same blind spot. Davoudi et al. (2025) show that inter-model majority voting can validate LLM answers without ground truth, confirming that cross-model agreement is a reliable quality signal, though their framework remains purely model-based. Tseng et al. (2025) introduce a multi-agent discussion protocol in which disagreeing LLMs share their reasoning and re-annotate over multiple rounds; this improves agreement but incurs substantial computational cost, and cases that exceed model capability receive no human support. Most closely related to our work, Yuan et al. (2025) present MCHR, a framework where three LLMs independently annotate content and majority-vote disagreements are routed to human reviewers—the same consensus-then-escalate logic as our Layer 1, though our design uses strict pairwise agreement between two models rather than three-way majority voting, yielding a more conservative gate with a lower silent error rate at the cost of a higher escalation rate.

Our framework extends this line of work in three ways. First, we employ a *redundancy design*: two independent consensus-plus-human systems rather than one, enabling a second round of cross-system consensus that catches errors a single system would miss. Second, the resulting second layer specifically targets *human annotator error*—mistakes introduced by fatigue or subjectivity—which single-layer designs propagate to the final output unchecked. Third, when consensus fails, we escalate to a *domain expert* rather than a stronger model, which is essential for tasks where models exhibit systematic blind spots on domain-specific edge cases. The framework requires no distributional priors or iterative calibration—only a simple agreement gate—making it directly applicable wherever two or more models can be run in parallel.

3 Double Triangle Annotation Framework

3.1 Motivation and Overview

Reliable evaluation of structured extraction systems on complex historical documents requires high-precision ground-truth data. Traditional human annotation is expensive and difficult to scale, while fully automated pipelines suffer from hallucination. We propose the **Double Triangle Annotation** framework, which combines the computational capacity of Multimodal Large Language Models (MLLMs) with targeted human oversight.

Rather than serving as primary annotators, humans act as *conflict resolvers* within a statistically motivated filtering system. The design rests on two key insights:

1. **Low probability of coincident error.** When two architecturally independent models produce the same output for a given field, the probability that both are wrong in the same way is low. Model consensus therefore serves as a high-confidence proxy for correctness.

2. **Mitigation of human fallibility.** A second layer compares two independent model–human systems against each other, catching errors introduced by annotator fatigue or subjectivity.

The objective is to produce a gold-standard dataset with near-perfect accuracy at a fraction of the cost of manual annotation. Figure 1 illustrates the full architecture.

3.2 Layer 1: The Annotation Triangle

The first layer maximizes throughput by leveraging cross-model consensus to automate labeling, reserving human labor strictly for ambiguous cases.

3.2.1 Architectural Components

Each annotation unit, termed 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 from document images in parallel.

- **One human jury** (H): an annotator who acts solely as a conflict resolver.

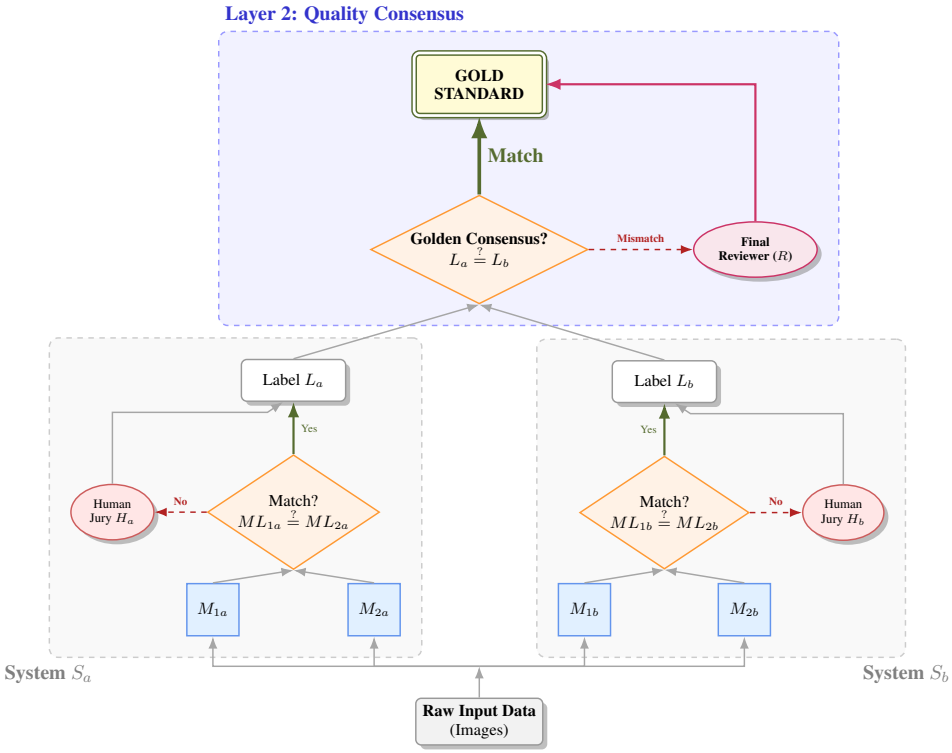


Figure 1: Overview of the Double Triangle Annotation framework. Layer 1 (gray boxes) pairs two independent models per system; consensus auto-accepts labels while conflicts are routed to a human jury. Layer 2 (blue box) cross-checks the two systems' outputs and escalates residual mismatches to a final reviewer.

3.2.2 Consensus-Based Workflow

For every input image I, the process follows a verification-by-agreement logic:

1. **Parallel inference.** M_1 and M_2 independently produce provisional labels ML_1 and ML_2.

2. **Automated filtering.**

   - **Consensus** ($ML_1 = ML_2$): the label is automatically accepted as the system output L, based on the low probability of coincident error (§3.2.3).
   - **Divergence** ($ML_1 \neq ML_2$): the case is flagged and routed to the human jury.

3. **Human adjudication.** The jury reviews the original image I alongside both model predictions and determines the final label L, shifting the human role from *producer* to *verifier*.

3.2.3 The Independence Requirement

The consensus mechanism rests on a simple probabilistic argument: if two models fail independently, the chance that both produce the *same* wrong output for the *same* field is negligible.

Let ϵ_1 and ϵ_2 denote the per-field error rates of M_1 and M_2. Under an independence assumption, the probability that both models err on a given field is $\epsilon_1 \cdot \epsilon_2$. A consensus error—a *silent* failure that passes the agreement gate—is even less likely, because it requires the two models to produce not merely wrong outputs, but the *same* wrong output. When two independently erring models produce *different* incorrect strings, the gate catches the disagreement. The silent error rate is therefore strictly bounded:

$$P(\text{consensus error}) \leq \epsilon_1 \cdot \epsilon_2 \qquad (1)$$

We define the per-system silent error rate $\epsilon_S = P(\text{consensus error})$; from Equation 1, $\epsilon_S \leq \epsilon_1 \cdot \epsilon_2$. Because the human jury corrects all detected disagreements, ϵ_S is the dominant source of residual error within each Layer 1 system—and it is the quantity that Layer 2 further reduces through cross-system comparison (Equation 2).

The bound in Equation 1 is conservative: the true rate is lower by a factor that depends on how many distinct wrong outputs a field admits. With empirically observed error rates of $\epsilon_1 \approx 0.087$ and $\epsilon_2 \approx 0.044$ (Section 5), we obtain $\epsilon_S \leq 0.004$—fewer than 1 in 250 consensus-accepted fields is

expected to contain an error, and the actual rate is likely lower still.

For this independence assumption to hold, M_1 and M_2 must be **architecturally independent** so that their failure modes do not correlate. We identify two strategies:

- **Cross-family**: pairing models from different providers (e.g., Llama vs. Claude) to ensure distinct training data distributions.
- **Cross-modal**: pairing a vision-based model (OCR) with a semantic model (MLLM), so that their error distributions are structurally disjoint.

3.2.4 Verification Sampling

To monitor the *silent error rate*—errors that pass consensus undetected—we implement a stochastic quality check: a random subset of consensus-accepted cases is routed to the human jury to verify the ongoing validity of the independence assumption.

3.3 Layer 2: The Validation Triangle

To mitigate the subjectivity and fatigue risks inherent in a single human jury, the second layer operates as a meta-verification step.

3.3.1 Meta-Architectural Components

This layer treats each Layer 1 system (S) as a single annotator unit. The structure comprises:

- **Two independent systems** (S_a, S_b): distinct Layer 1 instances with different underlying model pairs, maintaining orthogonal independence at the system level.
- **One final reviewer** (R): a domain expert (e.g., a historian) acting as the ultimate arbiter.

3.3.2 The Validation Workflow

The process filters residual model and human errors through cross-system comparison:

1. **Dual generation.** Systems S_a and S_b independently process the same data to produce refined labels L_a and L_b.
2. **Meta-comparison.**
   - **Golden consensus** ($L_a = L_b$): the label is automatically designated as gold standard G, since two disjoint model pairs and potentially distinct human juries reached the same conclusion.

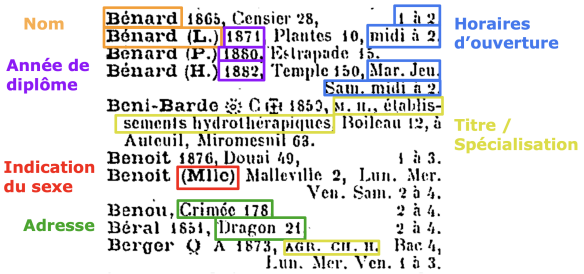


Figure 2: Annotated entries from the Rosenwald Guide (1887, p. 24). Each entry contains structured fields (name, year, notes, address, hours), but missing values and overlapping layouts introduce extraction noise.

   - **Conflict** ($L_a \neq L_b$): mismatches indicate complex edge cases or residual annotator error.
3. **Expert resolution.** Conflicting entries are routed to the final reviewer R, who determines the gold-standard label.

3.3.3 Error Probability Reduction

By cross-checking decisions from independent systems with different human juries, the framework induces a cascaded failure model. The probability of an error persisting in the final gold standard is bounded by the product of the two per-system error rates defined in Equation 1:

$$P(\text{error in } G) \leq \epsilon_{S_a} \times \epsilon_{S_b} \leq (\epsilon_{1a} \cdot \epsilon_{2a})(\epsilon_{1b} \cdot \epsilon_{2b}) \quad (2)$$

Since each ϵ_S is already the product of two small error rates, their product approaches zero rapidly. Expensive expert intervention is thereby allocated only to the small fraction of the dataset where the two systems disagree.

4 Dataset: The Rosenwald Guides

The *Guides Rosenwald* are French medical directories listing physicians and pharmacists across France and its colonies, published annually from 1887 to 1949 (47 volumes, each approximately one thousand pages). As illustrated in Figure 2, each page lists doctor entries in a semi-structured format: a typical entry contains a name (*nom*), diploma year (*année*), professional notes or specialization, address (*adresse*), and consultation hours (*horaires*). Extraction is complicated by missing fields, overlapping entries, interspersed advertisements, and degrading print quality.

In this work, we process the sections titled *Docteurs en médecine et en chirurgie* and *Officiers de santé* for Paris and the départements,

covering 20 volumes (1887–1906) with a total of 4,116 pages. This subset provides sufficient scale and diversity—in typography, layout variation, and print degradation—to evaluate the Double Triangle annotation framework under realistic conditions.

5 Experiments

We evaluate the Double Triangle framework along two axes: *quality*—whether the two-layer design effectively corrects model and human errors—and *efficiency*—whether consensus filtering successfully automates the majority of annotation work.

5.1 Experimental Setup

Dataset. We conduct a controlled evaluation on page 32 of the 1887 Rosenwald Guide, which lists doctors of medicine and surgery in Paris. The page contains 76 entries (40 in the left column, 36 in the right), yielding approximately 380 structured fields (name, address, consultation hours, etc.). Two annotators independently produced manual transcriptions of this page without any model assistance; their reconciled output serves as the gold standard G.

Model configuration. Following the independence requirement (§3.2.3), we instantiate two systems with cross-provider model pairs:

- **System S_a:** M_{1a} = `claude-sonnet-4-5` (Anthropic, 2025), M_{2a} = `qwen3-vl-235b` (Qwen Team, 2025).
- **System S_b:** M_{1b} = `llama4-maverick` (Meta AI, 2025), M_{2b} = `grok-4-0709` (xAI, 2025).

We use MLLMs rather than traditional OCR engines, as pilot experiments showed OCR outputs were too noisy for reliable consensus comparison.

Human annotators. Jury H_a is a research assistant fluent in French; jury H_b is the project researcher with B1-level French proficiency.[1] The final reviewer R is the same person as H_b. This represents a weaker configuration than the ideal three-independent-annotator design; the reported error rates therefore constitute an *upper bound* on the errors expected under full independence, since a dedicated reviewer would be strictly more effective than H_b serving dual roles.

[1] B1 ("intermediate") on the Common European Framework of Reference for Languages (CEFR).

5.2 Evaluation Metrics

We evaluate all system outputs against the gold standard G. For any hypothesis text X (a model prediction, jury-corrected version, or reviewer-corrected version), we report four metrics.

Word Error Rate (WER). WER measures word-level edit distance:

$$\mathrm{WER}(X, G) = \frac{S_w + D_w + I_w}{N_w}, \tag{3}$$

where S_w, D_w, I_w are word-level substitutions, deletions, and insertions, and N_w is the word count of G. We report WER **before** and **after** human correction at each layer.

Character Error Rate (CER). CER captures finer-grained errors:

$$\mathrm{CER}(X, G) = \frac{S_c + D_c + I_c}{N_c}, \tag{4}$$

where S_c, D_c, I_c and N_c are defined analogously at the character level.

Fields to correct. We quantify human workload at the field level. Let N_{fields} be the number of atomic fields in the dataset. For a given operator (jury or reviewer), we define:

$$N_{\text{correct}} = \sum_{j=1}^{N_{\text{fields}}} \mathbb{I}[X_j \neq Y_j]\,, \tag{5}$$

where X_j is the pre-review value for field j, Y_j is the value after intervention, and $\mathbb{I}(\cdot)$ is the indicator function.

Human effort ratio. We normalize the correction count:

$$\mathrm{EffortRatio} = \frac{N_{\text{correct}}}{N_{\text{fields}}} \times 100\%. \tag{6}$$

This metric captures the fraction of the dataset requiring manual edits at each layer.

5.3 Results and Analysis

Results are reported in Table 1.

5.3.1 Accuracy

First layer. After jury correction, WER decreases for both systems: from 0.087/0.044 to 0.030 for S_a, and from 0.069/0.141 to 0.008 for S_b. CER decreases for H_b but anomalously increases for H_a; inspection reveals that H_a accidentally left all fields blank for one entry, contributing errors that the models had partially captured. This incident underscores the value of the second-layer review for catching single-annotator lapses.

Human	Input	WER		CER		Fields to correct	Effort ratio
		Before	After	Before	After		
H_a	Claude Qwen	0.087 0.044	0.030	0.053 0.014	0.018	70	18.4%
H_b	Llama Grok	0.069 0.141	0.008	0.027 0.050	0.003	100	26.3%
R	H_a H_b	0.030 0.008	**0.003**	0.018 0.003	**0.001**	16	4.2%

Table 1: Results of the Double Triangle framework on page 32 of the 1887 Rosenwald Guide (380 fields). Layer 1 rows show jury corrections (H_a, H_b); the Layer 2 row shows the final reviewer (R). "Before" and "After" refer to error rates before and after human correction at the respective layer.

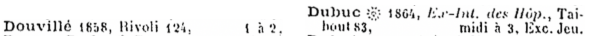
Second layer. The final reviewer R further reduces both WER (to 0.003) and CER (to 0.001). However, two errors persist in the final output, which we analyze below.

Layer-wise comparison. Table 1 can be read as an implicit ablation across increasingly strong configurations. A single model achieves WER between 0.044 and 0.141; a single annotation triangle (one model pair plus jury) reduces this to 0.008–0.030; the full double triangle achieves 0.003. Each layer yields a substantial improvement: Layer 1 consensus filtering eliminates the majority of model errors, and Layer 2 cross-system verification catches residual human and model mistakes that any single system would miss. This progressive reduction empirically validates the two-layer design and demonstrates that both layers contribute meaningfully to the final accuracy.

5.3.2 Error Analysis

Detailed inspection reveals two residual errors, both traceable to genuine visual ambiguity in the source images. Under such extreme ambiguity, the source itself induces correlated errors: when the printed character is objectively difficult to resolve, any reader—human or machine—is likely to arrive at the same incorrect interpretation, effectively breaching the independence assumption. These cases represent an irreducible failure mode of any consensus-based mechanism, since the quality signal collapses precisely when all annotators share the same degraded input.

1. **Character substitution.** The surname *Douvillé* appears with an accent aigu that is printed as a faint dot, visually indistinguishable from noise (Figure 3a). In S_b, Llama and Grok both recognize *Douville* (missing accent), so H_b has no disagreement signal and accepts the incorrect form. In S_a, Claude omits the entry entirely while Qwen also drops the accent; H_a accidentally left all fields blank for this entry (the annotator lapse noted in §5.3.1). The resulting discrepancy between L_a and L_b is escalated to R, who accepts the *Douville* variant from H_b without noticing the missing accent.



(a) *Douvillé*: accent resembles noise. (b) *Taibout*: hyphen misread as *t*.

Figure 3: The two residual errors both stem from genuine visual ambiguity that induces correlated errors across readers

2. **Character insertion.** The surname *Taibout* contains a line-break hyphen after *Tai-*, which models interpret as a faded character (Figure 3b). In S_a, Claude and Qwen both produce *Taitbout*, so the consensus gate auto-accepts the incorrect form without involving H_a. In S_b, Llama produces *Taitbout* while Grok produces *Taihout*—a different misreading that triggers the disagreement gate. However, H_b resolves the conflict in favor of *Taitbout*. Since $L_a = L_b$, the error passes Layer 2 undetected.

5.3.3 Efficiency

The consensus mechanism substantially reduces human workload. In Layer 1, H_a reviews 70 of 380 fields (18.4%) and H_b reviews 100 fields (26.3%), meaning 74–82% of fields are auto-accepted via model consensus. In Layer 2, R reviews only 16 fields (4.2%). The total human effort across both layers remains well below full manual annotation, and is distributed across three individuals (H_a, H_b, R), further reducing per-person workload and the risk of fatigue-induced errors.

6 Benchmark Construction

To demonstrate the scalability of the Double Triangle framework, we apply it to construct a benchmark from the Rosenwald Guides corpus described in §4.

Model pair	Raw (210 cols)		Filtered (60 cols)	
	Fields	%	Fields	%
Claude / Qwen	37,314 / 46,830	79.7	11,294 / 12,895	87.6
Llama / Grok	34,962 / 46,550	75.1	11,048 / 12,905	85.6

Table 2: Field-level agreement rates by model pair, before and after quality filtering. "Fields" shows matched / total.

6.1 Sampling and Preprocessing

We apply stratified sampling over the 4,116 target pages, allocating pages per year proportional to each volume's share of the corpus. This yields 105 sampled pages. To improve model accuracy, we manually remove advertisements from each page and split the two-column layout into separate images, producing 210 column images. The same cross-provider model pairs from §5.1 are used for annotation.

6.2 Agreement and Filtering

Table 2 reports field-level agreement rates before and after quality filtering. On the raw 210-column dataset, both model pairs achieve over 75% field-level consensus, indicating that the majority of fields can be auto-accepted without human review.

A key advantage of the Double Triangle design is that Layer 1 consensus is computed *before* any human correction, providing a cost-free signal of data quality. This enables selective filtering: we retain only columns with fewer than 70 distinct fields, a field-level matching rate above 0.7, and a character-level matching rate above 0.6, discarding outlier columns where models performed poorly or the layout was unusually complex. Because filtering occurs upstream of human effort, it further reduces annotation cost without compromising accuracy on the retained data. The tradeoff is a potential selection bias toward easier columns, though the filtered set preserves the typographic and temporal diversity of the original stratified sample.

After filtering, 60 columns remain. Agreement rates rise to 87.6% and 85.6% for the two model pairs.

6.3 Final Benchmark

The resulting benchmark comprises 60 columns containing 13,595 structured fields. In Layer 1, the consensus gate auto-accepts the majority of fields: jury H_a reviews 1,601 divergent fields (12.4%) and jury H_b reviews 1,857 (14.4%), while the remaining fields pass through without human involvement. In Layer 2, only 991 fields (7.2%) require correction by the final reviewer—the remaining 92.8% are resolved by cross-system consensus alone. Two annotators performed the corrections using a custom platform (Appendix A) that highlights disagreement fields and marks differing characters, further reducing review time. To monitor the silent error rate of consensus-accepted fields (§3.2.3), we applied the verification sampling procedure: one randomly selected consensus-accepted field per column was routed to the human jury for independent verification. Across the 60 sampled fields, no errors were detected.

This benchmark is, to our knowledge, the first structured-extraction ground truth for the Rosenwald Guides, and is made publicly available to support future work on historical document processing.

7 Conclusion

We presented the Double Triangle Annotation framework, a two-layer human-in-the-loop paradigm that uses cross-model consensus to automate the majority of annotation work while reserving human effort for genuinely ambiguous cases. The framework rests on a single assumption—partial error independence between models—and requires no task-specific calibration or distributional priors.

On the Rosenwald Guides, the framework achieves a final WER of 0.003 and CER of 0.001, with model consensus auto-accepting over 85% of the 13,595 fields in the at-scale benchmark. The two residual errors both stem from genuine visual ambiguity in the source images, where the degraded input induces correlated errors that breach the independence assumption. We release the resulting benchmark—the first structured-extraction ground truth for the Rosenwald Guides—to support future work on historical document processing.

Future directions include cross-modal model pairing (MLLM with OCR) to further reduce correlated failures, lightweight validation rules (e.g., address normalization) to catch systematic errors at the consensus stage, and progressive replacement of human juries with LLM-as-judge as model capability improves.

Limitations

Single-page controlled evaluation. The accuracy analysis in §5 is based on a single page (76 en-

tries, 380 fields). While the benchmark construction (§6) demonstrates scalability, the detailed error analysis and WER/CER measurements rely on a limited sample. Results may not generalize to pages with different layouts, denser entries, or greater print degradation.

Source ambiguity as an irreducible failure mode. In our error analysis (§5.3), the two residual errors reveal a distinct limitation: when the source image is genuinely ambiguous (e.g., faint diacritics, line-break artifacts), all readers—human and machine—converge on the same incorrect interpretation. Such cases represent situations where extreme source ambiguity induces correlated errors that breach the independence assumption—an irreducible failure mode that no consensus-based mechanism can detect, since the degraded input constrains all readers toward the same incorrect interpretation.

Annotator overlap. In our experimental setup, the final reviewer R is the same person as jury H_b, weakening the independence between Layers 1 and 2. Although we argue that this configuration provides an upper bound on error rates (§5.1), a fully independent three-annotator setup would provide stronger guarantees.

Benchmark filtering bias. The post-filtering step in benchmark construction (§6) removes columns with low model agreement, which may bias the benchmark toward easier cases. While filtering is a deliberate design choice that reduces human effort, the resulting benchmark may underrepresent the most challenging document layouts.

MLLM-only model pairs. All model pairs in our experiments use MLLMs. We do not evaluate the cross-modal configuration (MLLM paired with traditional OCR) discussed in §3.2.3, which could provide structurally more independent error distributions at the cost of noisier baseline outputs.

No external framework comparison. We do not compare the Double Triangle framework against alternative annotation methodologies (e.g., another LLM-consensus based annotation framework, or standard human annotation). Our evaluation relies on an inward comparison—single model vs. single triangle vs. double triangle—which validates the incremental value of each layer but does not directly benchmark against existing approaches. A controlled comparison would require applying multiple annotation frameworks to the same dataset under comparable cost constraints, which we leave to future work.

References

Anthropic. 2025. Claude sonnet 4.5 system card. Technical report, Anthropic.

Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. 2015. The ENP image and ground truth dataset of historical newspapers. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 931–935.

Seyed Pouyan Mousavi Davoudi, Amin Gholami Davodi, Alireza Amiri-Margavi, Alireza Shafiee Fard, and Mahdi Jafari. 2025. Collective reasoning among LLMs: A framework for answer validation without ground truth. *CoRR*, abs/2502.20758.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 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)*, pages 11173–11195. Association for Computational Linguistics.

Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *CoRR*, abs/2503.06778.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176. Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 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*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Nicholas Pangakis 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*, pages 1471–1492. AAAI Press.

Christos Papadopoulos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2013. The IMPACT dataset of historical document images. In *Proceedings of the 2nd International Workshop on*

Historical Document Imaging and Processing, HIP '13, pages 123–130. Association for Computing Machinery.

Qwen Team. 2025. Qwen3-VL technical report. *CoRR*, abs/2511.21631.

Hope Schroeder, 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*, pages 25771–25795. Association for Computational Linguistics.

Zejiang Shen, Kaixuan Zhang, and Melissa Dell. 2020. A large dataset of historical Japanese documents with complex layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. M5HisDoc: A large-scale multi-style Chinese historical document analysis benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pages 78483–78495. Curran Associates, Inc.

Foteini Simistira, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, and Rolf Ingold. 2016. DIVA-HisDB: A precisely annotated large dataset of challenging medieval manuscripts. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. Evaluating large language models as expert annotators. *CoRR*, abs/2508.07827.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 303:1–303:21. ACM.

xAI. 2025. Grok 4 model card. Technical report, xAI.

Yue Xu, Fei Yin, Da-Han Wang, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. 2019. CASIA-AHCDB: A large-scale Chinese ancient handwritten characters database. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 793–798.

Hanyu Yan, Lianwen Jin, Christian Viard-Gaudin, and Harold Mouchere. 2010. SCUT-COUCH Textline_NU: An unconstrained online handwritten Chinese text lines dataset. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 581–586.

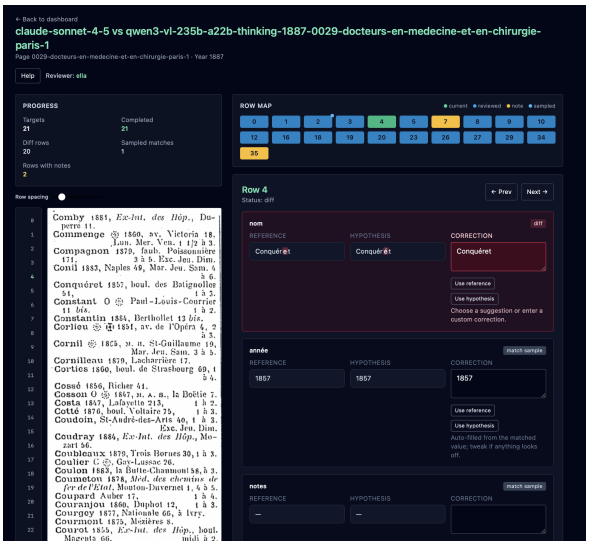


Figure 4: Screenshot of the annotation platform. The left panel shows the source document; the right panel presents model outputs side by side with a correction field. Character-level differences are highlighted to guide the reviewer.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Gelareh Mohammadi, and Aaron Quigley. 2025. A case study of scalable content annotation using multi-LLM consensus and human review. *CoRR*, abs/2503.17620.

Yulong Zhang, Tianyi Liang, Xinyue Huang, Erfei Cui, Xu Guo, Pei Chu, Chenhui Li, Ru Zhang, Wenhai Wang, and Gongshen Liu. 2025. Consensus entropy: Harnessing multi-VLM agreement for self-verifying and self-improving OCR. *CoRR*, abs/2504.11101.

A Annotation Platform

Figure 4 shows the annotation interface used by human juries and the final reviewer. For each row flagged as a disagreement, the platform displays the original document image alongside both model outputs ("Reference" and "Hypothesis") and a correction field. Differing characters are highlighted to direct the annotator's attention to the specific point of conflict.

B Extraction Prompt

All four models (Claude, Qwen, Llama, Grok) receive the same prompt, reproduced below in its original French. The prompt instructs the model to extract structured physician entries from a document image and return the results as tab-separated values.

```
Vous analysez directement une image d'annuaire
    médical français du 19e et du 20e siècle.
L'entrée est une IMAGE d'UNE page complète
    d'annuaire médical historique.
```

```

IMPORTANT:
- L'image peut contenir des publicités, des
    textes d'éditeur, ou du texte non-médical
    en haut ou en bas.
- Ignorez-les complètement. Extrayez SEULEMENT
    les entrées de médecins.
- Lisez attentivement tous les textes visibles
    dans l'image, même s'ils sont
    partiellement effacés ou difficiles à lire.
- Si l'image ne contient AUCUNE entrée de
    médecin (seulement des publicités ou du
    texte non-médical), retournez une ligne
    vide dans le TSV.

Chaque entrée de médecin contient généralement
    :
- nom de famille (parfois avec préfixe ou
    parenthèses),
- année (diplôme ou année de référence),
- notes optionnelles (ex: "Ex-Int. des Hôp.",
    "AGR.", "M. A. M.", "CH. H.", "M. H."),
- adresse (rue, boulevard, avenue, etc.),
- horaires d'ouverture (abrégés: Lun., Mar.,
    Mer., Jeu., Ven., Sam., Dim., "midi à 2",
    "2 à 4", etc.).

### TÂCHE
Analysez l'image et produisez un fichier TSV
    (Tab-Separated Values), où chaque ligne
    correspond à UNE entrée de médecin, avec
    ces colonnes séparées par des tabulations :

nom     année   notes   adresse horaires

### DESCRIPTION DES COLONNES
- nom: Nom de famille du médecin (lisez
    l'orthographe exacte de l'image)
- année: Année de diplôme ou de référence
    (lisez exactement, ex: "1881" ou "1870")
- notes: Titres/affiliations hospitalières
    (ex: "Ex-Int. des Hôp.", "AGR.", "M. A.
    M.", "CH. H.")
- adresse: Adresse de rue complète (conservez
    accents/abréviations visibles)
- horaires: Informations horaires (ex: "Lun.
    Mer. Ven. 3 à 5")

### RÈGLES DE LECTURE D'IMAGE
- Lisez l'orthographe EXACTE visible dans
    l'image, accents, symboles (, AGR., etc.).
- Si un champ n'est pas visible ou illisible,
    laissez-le vide.
- N'inventez pas d'informations non visibles
    dans l'image.
- Ne manquez AUCUN médecin : parcourez
    systématiquement toute l'image pour
    extraire toutes les entrées présentes.
- ORDRE DE LECTURE CRITIQUE : L'image contient
    généralement DEUX COLONNES d'entrées de
    médecins. Respectez l'ordre de lecture
    suivant :
  1. Lisez la colonne de GAUCHE de haut en bas
     COMPLÈTEMENT
  2. Puis lisez la colonne de DROITE de haut
     en bas COMPLÈTEMENT
  3. NE MÉLANGEZ PAS les colonnes - terminez
     entièrement la gauche avant la droite
  4. Cet ordre est ESSENTIEL pour l'évaluation
     ultérieure
- Ne produisez que les entrées de médecins
    (ignorez publicités, textes d'éditeur).
- Incluez toujours les titres de civilité
    (Mme, Mlle, etc.) dans le champ nom s'ils
    sont visibles.
- Si aucune entrée de médecin n'est trouvée
    dans l'image, retournez seulement
    l'en-tête TSV.
- Séparez les colonnes par des TABULATIONS
    (    ), pas des virgules.
- Aucun échappement nécessaire pour les
    virgules dans le texte.
- NE PAS entourer la sortie avec des balises
    markdown ou des blocs de code.
- Retournez UNIQUEMENT le contenu TSV brut,
    sans formatage markdown.

### CONSEILS POUR LA LECTURE D'IMAGE
- Examinez attentivement les zones de texte
    dense qui contiennent les listes de
    médecins.
- Les entrées de médecins sont généralement
    organisées en DEUX COLONNES distinctes sur
    la page.
- IMPORTANT : Suivez l'ordre de lecture
    colonne par colonne (gauche complète, puis
    droite complète).
- Faites attention aux abréviations typiques
    de l'époque (boul. pour boulevard, etc.).
- Les horaires peuvent être sur la même ligne
    ou sur la ligne suivante.
- Certains textes peuvent être en caractères
    plus petits ou partiellement effacés.
- Identifiez visuellement la séparation entre
    les deux colonnes pour éviter de mélanger
    l'ordre.

### EXEMPLE DE FORMAT ATTENDU:
nom     année   notes   adresse horaires
Vallois 1848            St-André-des-Arts
    50    2 à 4
Ravaux (Mme)    1883            Assomption
    75
Berline Hering (Mme)    1881    Halles
    13
Darcus Richardson       1861
    Poisson 3       3 à 4, Exc. Dim
Benoit (Mlle)                   Malleville
    2    Lun. Mer. Ven. Sam. 2 à 4
Rigal   1866    AGR. M. H.      Murillo
    6       Lun. Mer. Ven. 1 à 3
Redard  1879    Méd. en chef des ch. de fer,
    Ex-Chef de Clin.   Constantinople 2
    Lun. Mer. Ven. 5 à 6
Peltier (Henri) 1877            Marseille
    2     4 à 6

### IMAGE À TRAITER
Analysez l'image ci-jointe et extrayez les
    données médicales selon les instructions
    ci-dessus.
```