SEVENTEENTH INTERNATIONAL WORKSHOP ON **DOCUMENT ANALYSIS SYSTEMS**

DAS 2026

WORKSHOP PROCEEDINGS

Short and Demo Papers

DATES

3 – 4 September 2026

VENUE

Vienna, Austria

Satellite workshop of ICDAR 2026

WORKSHOP CHAIRS

Faisal Shafait
National University of Sciences and Technology, Pakistan

Adrian Ulges
RheinMain University of Applied Sciences, Germany

PROGRAMME CHAIRS

Momina Moetesum
National University of Sciences and Technology, Pakistan

Xu-Cheng Yin
University of Science and Technology Beijing, China

ORGANISERS & SPONSORS

Document Analysis Systems

17th IAPR International Workshop, DAS 2026
Vienna, Austria, September 03–04, 2026
Proceedings

Faisal Shafait[1], Adrian Ulges[2], Momina Moetesum[1], and Xu-Cheng Yin[3]

[1] School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan
[2] Hochschule RheinMain – University of Applied Sciences and Arts, Wiesbaden, Germany
[3] University of Science and Technology Beijing, Beijing, China

Preface

It is our pleasure to welcome you to the proceedings of the Seventeenth IAPR International Workshop on Document Analysis Systems (DAS 2026). Vienna provides an inspiring setting to continue the long-standing tradition of the DAS workshop series, bringing together researchers and practitioners from around the world to share advances, exchange ideas, and foster collaboration in the field of document analysis and recognition.

DAS No. 17 builds on a distinguished series of previous workshops held in Kaiserslautern, Germany (1994); over Malvern, PA, USA (1996); Nagano, Japan (1998); Rio de Janeiro, Brazil (2000); Princeton, NJ, USA (2002); Florence, Italy (2004); Nelson, New Zealand (2006); Nara, Japan (2008); Boston, MA, USA (2010); Gold Coast, Australia (2012); Tours - Loire Valley, France (2014); Santorini, Greece (2016); Vienna, Austria (2018); Wuhan, China (2020); La Rochelle, France (2022); and Athens, Greece (2024). Like these previous editions, this year's DAS focuses on system-level challenges and holistic approaches to document analysis and recognition. With the rapid emergence of generative AI and agentic systems as key enablers of real-world document understanding applications, this perspective is gaining renewed importance.

Since 2024, DAS has adopted a revised format and is now organized in conjunction with the International Conference on Document Analysis and Recognition (ICDAR) as its largest associated workshop, while retaining its own proceedings and two-day structure. We consider this change successful, and have put considerable effort into making DAS a high-quality workshop in the spirit of its successful predecessors. We ensured a single-track format, a rigorous peer-review process, and complete participation, complemented by invited talks, contributed papers, and discussion sessions. With this format, we hope to bring together practitioners and researchers from the industry with academics in a unique environment.

Following DAS 2024, this year's edition features two separate calls for papers, an earlier one corresponding to regular papers to be included in the proceedings, and a second call for short papers presenting preliminary results or research in progress. Of the 50 submissions received for the regular paper track, 29 were accepted for presentation at the workshop, resulting in an acceptance rate of 58%. All submissions received at least two double-blind reviews from our team of 49 Program Committee members. The accepted regular papers are published in this volume of the Springer Lecture Notes in Computer Science (LNCS) series. DAS also received nine short paper submissions and seven demo submissions, which were reviewed in a single-blind form by the Program Committee. After review, four short papers were accepted for inclusion in the program along with five demos. The short papers and demo papers are made available in PDF format on the DAS conference website. The final program included four oral sessions, two poster sessions, one demo session, and one discussion group. In

addition, DAS hosted industry keynote talks delivered by leading experts: Jordy van Landeghem, Nibal Nayef, and Brandon Smock. Outstanding contributions were recognized through the Best Paper Award and the Best Student Paper Award.

We extend our sincere gratitude to everyone who contributed their time and effort to making DAS 2026 a leading event for the community. The success of the workshop program was made possible through the dedication and hard work of many individuals. We sincerely thank the Program Committee members for their hard work reviewing submissions, and especially those last-minute, emergency reviewers who were instrumental in avoiding notification delays. We also thank the organizing committee for their important contributions. We acknowledge the efforts of Industrial and Sponsorship Chairs Yves-Noel Weweler and Lukasz Borchmann in coordinating sponsorship activities and keynote arrangements. Special thanks go to Carlos Moreno-García for creating and maintaining the DAS website and social media presence; to Publication Chair Liangrui Peng for overseeing the proceedings preparation; and to Ajoy Mondal for serving as Demo and Discussion Group Chair.

Finally, we would like to express our sincere appreciation to the authors for their outstanding contributions and active participation in DAS 2026. We hope that the workshop program will encourage future research, inspire new collaborations, and provide practitioners with useful methods, algorithms, and tools. It is both an honor and a privilege to present the latest advances in document analysis systems through these proceedings, and we trust that readers will find them informative and inspiring. We look forward to Vienna!

September 2026

Faisal Shafait
Adrian Ulges
Momina Moetesum
Xu-Cheng Yin

Organization

General Chairs

Faisal Shafait	National University of Sciences and Technology, Pakistan
Adrian Ulges	RheinMain University of Applied Sciences, Germany

Program Chairs

Momina Moetesum	National University of Sciences and Technology, Pakistan
Xu-Cheng Yin	University of Science and Technology Beijing, China

Publication Chair

Liangrui Peng	Tsinghua University, China

Industrial and Sponsorship Chairs

Yves-Noel Weweler	Buildsimple, Germany
Lukasz Borchmann	Snowflake Inc, Poland
Tobias Alt-Veit	Insiders Technologies GmbH

Publicity and Website Chair

Carlos Moreno-García	Robert Gordon University, Scotland

Demo and Discussions Group Chair

Ajoy Mondal	International Institute of Information Technology Hyderabad, India

Local Organizers

Marco Peer	Université de Fribourg, Switzerland
Rafael Sterzinger	TU Wien, Austria
Marvin Burges	TU Wien, Austria

Program Commitee

Adrian Ulges	RheinMain University of Applied Sciences, Germany
Ajoy Mondal	IIIT Hyderabad, India
Anna Scius-Bertrand	University of Fribourg, Switzerland
Ashok Popat	Google, Inc., USA
Bin Liu	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
Biying Fu	RheinMain University of Applied Sciences, Germany
Camille Kurtz	Université Paris Cité, France
Carlos Mello	Universidade Federal de Pernambuco, Brazil
Chiranjoy Chattopadhyay	FLAME University, India
Daniel Benniah John	UC Berkeley, USA
Dirk Krechel	RheinMain University of Applied Sciences, Germany
Eliott Thomas	Laboratoire Informatique, Image et Interaction (L3i), La Rochelle University, France
Elisa Barney	Boise State University, USA
Ernest Valveny	Computer Vision Center, Spain
Florence Cloppet	Université Paris Cité, France
Florian Kleber	TU Wien, Austria
Frédéric Rayar	LIFAT - University of Tours, France
Hung Tuan Nguyen	Tokyo University of Agriculture and Technology, Japan
Irina Rabaev	Shamoon College of Engineering, Israel
Jianjuan Liang	Guizhou University, China
Jin Wei	Lenovo Research, China
Josep Llados	Computer Vision Center, Universitat Autònoma de Barcelona, Spain
Kaspar Riesen	University of Bern, Switzerland
Kengo Terasawa	Future University Hakodate, Japan
Kenny Davila	DePaul University, USA
Maham Jahangir	Guangdong CAS Cogniser Information Technology Co. Ltd., China
Marco Peer	University of Fribourg, Switzerland
Mark Clement	Brigham Young University, USA
Maroua Mehri	University of Sousse, Tunisia
Mickael Coustaty	L3i Laboratory, France
Nam Ly	Tokyo Metropolitan University, Japan

Olumayowa Onabanjo	University of Oviedo, Spain
Oriol Ramos Terrades	Universitat Autònoma de Barcelona - Computer Vision Centre, Spain
Rina Buoy	Techo Startup Center, Cambodia
Robert Sablatnig	TU Wien, Austria
Ronaldo Messina	Mitek Systems, USA
Salvatore-Antoine Tabbone	IDMC-Université Lorraine, France
Shivakumara Palaiahnakote	University of Salford, United Kingdom
Tayyibah Khanam	Kensho Technologies (S&P Global), USA
Thierry Paquet	University of Rouen Normandie, France
Verónica Romero	Universitat de Valencia, Spain
Vincent Christlein	Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Vincent Poulain d'Andecy	Yooz, France
Vlad Atanasiu	University of Bern, Switzerland
Xiangdong Su	Inner Mongolia University, China
Xugong Qin	Nanjing University of Science and Technology, China
Yaping Zhang	National Laboratory of Pattern Recognition, Institute of Automation, University of Chinese Academy of Sciences, China
Yuchen Zheng	Shihezi University, China
Yugo Kubota	Kyushu University, Japan

Sponsors / Organizers

IAPR, SEECS NUST, HSRM, USTB, IIIT, Robert Gordon University, Tsinghua University, University of Fribourg, TU Wien, Insiders Technologies GmbH, Snowflake Inc

International Association for Pattern Recognition (IAPR)

Insiders Technologies GmbH

National University of Sciences and Technology (NUST)

Hochschule RheinMain University of Applied Sciences and Arts (HSRM)

University of Science and Technology Beijing (USTB)

International Institute of Information Technology Hyderabad (IIIT Hyderabad)

Robert Gordon University

Tsinghua University

University of Fribourg

Vienna University of Technology (TU Wien)

Snowflake Inc.

Fig. 1: Sponsors and Organizers of DAS 2026

Table of Contents

CrimeNER Demo: Named-Entity Recognition in the Crime Domain

Miguel Lopez-Duran, Julian Fierrez, Aythami Morales, Daniel DeAlcala, Gonzalo Mancera, Javier Irigoyen, Ruben Tolosana, Oscar Delgado, Francisco Jurado, and Alvaro Ortigosa

BiometricsAI, Universidad Autónoma de Madrid (UAM), Spain
miguel.lopezd@uam.es, julian.fierrez@uam.es

Abstract. We present CrimeNER Demo, an AI-powered platform that enables us to extract general crime-related information from documents and classify them into entity types with two levels of granularity. We provide pretrained NER models on the CrimeNER database, and we give the possibility to users to provide their own annotated data to train models for their own specific cases. This demonstrator aims to promote crime-related NER research and provides a practical tool to automatically extract crime information for researchers and law enforcement agencies. The demonstrator includes: i) Pretrained NER models on the crime domain; ii) Possibility to finetune the models on specific data annotated by the user; and iii) An automatic pipeline to extract and annotate crime entities from documents. The demo platform, a tutorial to run the demo, and a video demonstration are publicly available on GitHub.[1]

Keywords: Crime Analysis, NER, Forensics NER, Crime NER

1 Introduction

Law enforcement agencies are increasingly required to extract and process information from crime-related documents. However, the number of documents that need to be processed is increasing rapidly, so manually extracting these data is not feasible. In addition, manually annotating sufficient data to train reliable models from scratch is really time consuming and inefficient [19].

Automatic extraction of criminal information from document repositories (including web/html locations) may be viewed as a Named Entity Recognition (NER) task [16]. In forensics, the kind of entity law enforcement agencies are interested in can change from case to case. However, when working in criminal cases, information like who committed or is being accused of a crime, or the agents and agencies involved in a case, are general and typically useful kind of information. NER has a lot of work done in several fields. Initially, it was focused on news and general documents, but it grew rapidly to other areas

[1] https://github.com/BiometricsAI/CrimeNER.git

such as biomedicine [21]. In relation to the crime domain, there are works that focus on the extraction of legal entities [1] and on cyber threat Intelligence [22]. However, these works revolve around legal texts or specific types of crime, and are not suitable for general crime extraction for day-to-day criminal cases. Coarse entities are annotated with different colors, while fine entities are annotated as metadata inside the document.

Despite all this work, there is still a huge gap in the NER literature in the crime domain. As an initial step to fill this gap, we present a demonstrator of CrimeNER, a platform to extract crime-related entities from input documents. CrimeNER Demo leverages pretrained NER models in the CrimeNER database (CrimeNER-db) [13]. We label each detected entity into a predefined set of entity types with two levels of granularity, namely coarse and fine-grained entity types, as done in previous works [6]. The coarse entity types defined cover basic information about crimes, e. g., the criminal who committed the offense. Fine-grained entity types are defined to better contextualize each of the coarse entity types extracted from the document, e.g. the typology of the criminal activity mentioned or the number of criminals that committed the crime.

We acknowledge that different law enforcement agencies work with different types of crime, and in languages different from English. For this reason, we also let the users upload their own annotated data to the CrimeNER Demo following our schema and finetune the selected models on their data. In this way, only a small number of annotated samples is enough to extract meaningful crime information, avoiding the need for a large annotated database in each target domain [15,14]

This work presents a demonstrator of the CrimeNER platform and serves researchers and practitioners in the line of Forensic Document Analysis using NER techniques.

2 CrimeNER: Database

The CrimeNER database (CrimeNER-db) is a dataset comprising more than 1.5K annotated documents from real-world scenarios, such as terrorist reports or real press notes from the US for the extraction of criminal entities.

The primary goal of CrimeNER-db is to provide researchers with a fine-grained general crime dataset. In order to do that, based on previous work on fine-grained NER databases [6], we define a two-level hierarchy for entity types, coarse and fine-grained, and classify each token as part of a coarse entity with its corresponding fine-grained entity type, or as a non-entity. The coarse and their corresponding fine entity types are as follows:

- **Crime**: Illegal activities performed by an individual or group of people, including terrorist attacks. As fine-grained crime entity types, we consider: Terrorism, Fraud, Illegal Traffic, Theft, Drug-related, Homicide, Sexual crimes, Hate crimes, and Others.

- **Actor**: Individuals or organizations that commit or are accused of committed a crime. The fine actor entity types are: Criminal Person, Criminal Organization, Terrorist Person, and Terrorist Organization.
- **Agent & Agency**: Individuals and organizations acting against criminal activities or involved in criminal cases. We also consider government officials and bodies as this type of entity. We consider three fine entity subtypes: Law Enforcement, Government, and Legal. [2]
- **Logistic**: Specific details of the crimes mentioned in the documents that may be useful to agents. The fine logistic entity types are Location, GPE, Date, Weapons & Explosives, and Money.

3 CrimeNER: System

The CrimeNER Demo platform extracts crime related entities from input documents and texts with two levels of granularity. The system comprises 4 processing modules: 1) Document Preprocessing, 2) Specific Fine-tuning (if selected), 3) Coarse and Fine Entity Extraction, and 4) Postprocessing.

3.1 CrimeNER Modules

Document Preprocessing Given an input document or batch of documents, the user is asked to decide whether to process the document using the pretrained NER models or to upload annotated data and finetune them on the data and then process the document. The models were trained in CrimeNER-db by performing a train/val / test split and evaluated on the test split. The results of this training with the selected NER models are shown in Table 1.

Specific Fine-tuning Given the wide variety of criminal cases and languages law enforcement agencies may be interested in, we provide the option to further finetune the released criminal NER models on specific target data. This makes our platform even more useful in real-world deployment as it can be adapted to most criminal cases. The only drawback is that the target data need to be formatted as we specify in our demonstrator with the same types of entity. We believe that the types defined on CrimeNER-db are informative enough for most cases. In future work, we will consider new types of entity.

Coarse and Fine Entity Extraction Given the processed documents and the pretrained or finetuned NER models, coarse entities are extracted. After that, on the basis of the extracted coarse entities, we extract the fine entities from them. All this process let us classify each entity into 4 coarse entity types and a total of 21 fine entity types, which makes the extracted information highly detailed and specific for each case.

[2] In the original CrimeNER-db work, the Agent and Agency coarse entities are different types, but we decided to merge them into one, as we found it more informative.

Table 1: Average Strict and Flexible F1 Scores on coarse and fine entity types after training on CrimeNER-db. Strict evaluation requires the same entity span and type to be correct. Flexible evaluation requires same entity type, and an overlap between the predicted span and the ground truth.

Model	Coarse		Fine	
	Strict	**Flexible**	**Strict**	**Flexible**
XLM-RoBERTa-Base [2]	**0.650**	0.900	**0.650**	0.879
DeBERTa-V3-Base [7]	0.649	**0.902**	0.627	**0.892**
RoBERTa-Base [11]	0.620	0.899	0.631	0.882
AlBERT-Base-V2 [10]	0.607	0.881	0.401	0.888
DistilBERT-Base-Cased [20]	0.519	0.890	0.643	0.886
BERT-Base-Cased [5]	0.514	0.839	0.644	0.889

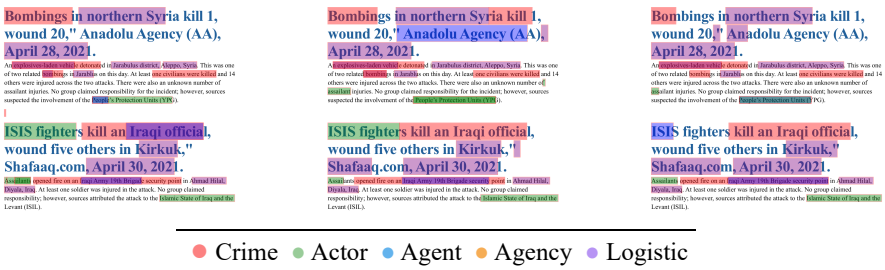


Fig. 1: Example of different annotations for a sample document using the CrimeNER Demo with different model configurations. From left to right, the annotations for coarse entities were made using XLM-RoBERTa-Base [2], DeBERTa-V3-Base [7] and DistilBERT-Base-Cased [20] and the fine entity annotations were made using DeBERTa-V3-Base, XLM-RoBERTa-Base and AlBERT-Base-V2 [10]. Coarse entities are colored in the document and fine entities are annotated as metadata inside the document.

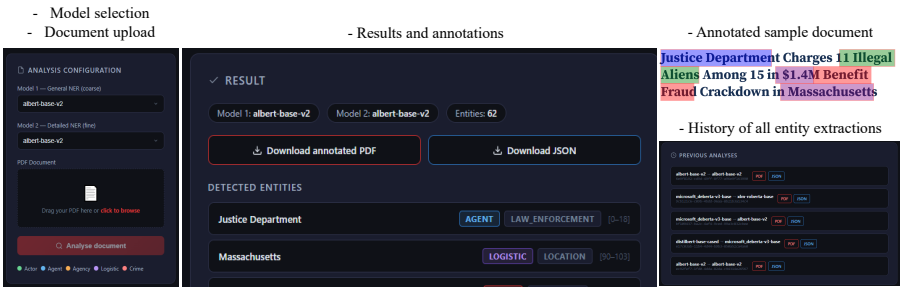


Fig. 2: Overview of the CrimeNER Demo. Documents are processed with the selected model (fine-tuned if specific data is uploaded). Crime-related entities are extracted with two levels of granularity, processed, and shown to the user.

Document Postprocessing After the entity extraction, the entity spans are further processed and injected into the input documents. Input documents and texts are returned to the user with the detected crime entity spans highlighted with the detected coarse- and fine-grained entity types detected. A JSON file with the annotation of each document with its extracted entities may also be provided if requested by the user.

4 CrimeNER: Demonstrator

This study presents the CrimeNER Demo platform, a platform for extracting general criminal entities from documents. The platform was evaluated using CrimeNER-db, a database consisting of 1.5K real-world documents from the U.S. Department of Justice and other terrorist and crime reports.

The main objective of this work is to promote our platform, where we make CrimeNER Demo available to researchers and practitioners. The CrimeNER Demo allows users to pass input documents or texts and extract the crime-related entities with two levels of granularity within those documents. After processing the documents, users will receive the input documents with the extracted crime entities annotated and highlighted, as shown in Figure 1. This platform opens up new opportunities in document analysis and crime detection for researchers and law enforcement agencies.

We provide several pretrained models to extract the crime-related entities. The selection of different models for coarse and fine entities results in different annotations, as shown in Figure 1. Qualitative analysis shows that the difference between model annotations is notable, especially the differences in span length for the same entity. There is some misalignment between the annotation and the PDF text, as the positions in the PDF metadata are not exactly aligned with the visual text position in the document.

Figure 2 shows a screenshot of the CrimeNER demonstrator and its main features that we discussed earlier. Initially, users are required to upload one or more documents and select the NER models trained to process the documents. Among these models, there can be the finetuned models on specific data uploaded by the user, for which we provide a tutorial on the CrimeNER demo repository. After model selection, coarse and fine entity types spans are extracted from the documents as we described earlier. Finally, the extracted entities are annotated and highlighted in the documents, and a JSON file with the entity annotations is generated if the user requires it. The CrimeNER platform also stores previous document analysis as part of a history so that users can download documents and annotations already processed again, if necessary.

In future work, we will explore multimodal architectures [18], including the combination of NLP models like those used here and visual models that process text images [12,4]. Detecting AI-generated information [3], fakes [17], other types of manipulation [9], and other general risks [8] when analyzing document repositories are also key topics on our agenda.

Acknowledgments. PID2024-160053OB-I00 MICIU/FEDER, Cátedra ENIA (NextGenerationEU PRTR TSI-100927-2023-2), and R&D Agreement DGGC / UAM/FUAM for Biometrics & Applied AI. Lopez and Robledo are supported by FPI-UAM-2025, DeAlcala is supported by FPU21/05785, Mancera is supported by FPI-PRE2022-104499, Irigoyen is supported by FPI-PREP2024-003107.

References

1. Au, T.W.T., Lampos, V., Cox, I.: E-NER: An annotated named entity recognition corpus of legal text. In: Proc. of the NLP Workshop. pp. 246–255 (2022)
2. Conneau, A., et al.: Unsupervised cross-lingual representation. In: ACL (2020)
3. DeAlcala, D., Ko, C.Y., Morales, A., Fierrez, J., et al.: DocAI: A framework for generating and identifying AI-edited documents. In: arXiv preprint (2026)
4. DeAlcala, D., Mancera, G., Fierrez, J., et al.: Is my vision-language data in your AI? Membership Inference Test (MINT) Demo 2. In: IEEE COMPSAC (2026)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
6. Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., et al.: Few-NERD: A few-shot named entity recognition dataset. In: Proc. ACL. pp. 3198–3213 (2021)
7. He, P., et al.: DeBERTa: BERT with disentangled attention. In: ICLR (2021)
8. Irigoyen, J., et al.: Overview of risk assessment and management for intelligent systems under the AI Act and beyond. In: IEEE ICCST (2026)
9. Korshunov, P., et al.: DeepID challenge of detecting synthetic manipulations in ID documents. In: IEEE Intl. Conf. on Computer Vision Workshops (2025)
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., et al.: ALBERT: A lite BERT for self-supervised learning of language representations. CoRR **abs/1909.11942** (2019)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., et al.: RoBERTa: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
12. Lopez-Duran, M., Fierrez, J., et al.: Benchmarking graph neural networks for document layout analysis in public affairs. In: ICDAR Workshops (2025)
13. Lopez-Duran, M., Fierrez, J., et al.: Named-entity recognition in the crime domain (CrimeNER): Case study and dataset. arXiv:2603.02150 (2026)
14. Lopez-Duran, M., Marrero, E., et al.: Comparative study of domain-adapted VLMs for general document visual question answering. In: ICDAR Workshops (2026)
15. Mancera, G., DeAlcala, D., Morales, A., Fierrez, J., et al.: Auditing training data in domain-adapted LLMs: LoRA-MINT. In: IEEE COMPSAC (2026)
16. Mancera, G., Morales, A., Fierrez, J., et al.: PBa-LLM: Privacy-and bias-aware NLP using Named-Entity Recognition (NER). In: ICDAR Workshops (2025)
17. Muñoz, J., et al.: Privacy-aware detection of fake identity documents: methodology, benchmark, and improved algorithms (FakeIDet2). Inf. Fusion **128**, 103969 (2026)
18. Peña, A., et al.: Human-centric multimodal machine learning: Recent advances and testbed on AI-based recruitment. SN Computer Science **4**(5), 434 (June 2023)
19. Peña, A., et al.: Continuous document layout analysis: Human-in-the-loop AI curation, database & evaluation in the domain of public affairs. Inf. Fusion **108** (2024)
20. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)
21. Song, B., et al.: Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. Brief. Bioinform. **22**(6), bbab282 (2021)
22. Wang, X., He, S., Xiong, Z., Wei, X., et al.: APTNER: A specific dataset for NER missions in cyber threat intelligence field. In: IEEE CSCWD. pp. 1233–1238 (2022)

AIriskEval-edu Demo: Auditing of Pedagogical Risks in Educational Explanations

Javier Irigoyen[1], Roberto Daza[1,2], Francisco Jurado[2], Julian Fierrez[1], Ruben Tolosana[1], Alvaro Ortigosa[2], Miguel Lopez-Duran[1], and Aythami Morales[1,3]

[1] BiometricsAI, Universidad Autónoma de Madrid (UAM), Spain
[2] GHIA, Universidad Autónoma de Madrid (UAM), Spain
[3] Universidad de Las Palmas de Gran Canaria (ULPGC), Spain

Corresponding author: roberto.daza@uam.es

Abstract. We present AIriskEval-edu Demo, a platform that audits the pedagogical quality of instructional explanations and provides explainable audit results. The platform evaluates an explanation against a rubric of five dimensions of pedagogical risk: factual accuracy, depth and completeness, focus and relevance, student-level appropriateness, and ideological bias. For each risk dimension, it returns a binary decision with a confidence score. Detected risks also include a natural-language rationale and, except for Depth and Completeness, a localized evidence span. The platform integrates GPT-5.5 via an external API and a local Llama 3.1 8B evaluator that is self-hosted and runs on consumer-grade GPUs. The local evaluator is fine-tuned on AIriskEval-edu, a dataset of K–12 instructional explanations with risk and explainability annotations. It operates in two modes: in AI mode, both evaluators assess stored explanations generated under six simulated teacher profiles, each representing a distinct pedagogical behavior and potential risk; in human mode, the local evaluator audits user-written explanations in real time. The local evaluator outperforms GPT-5.5 on most metrics, offering educational institutions a practical way to keep audited content within their own infrastructure.

Keywords: Document Understanding · Risk Assessment · LLM Auditing · Explainability · Human-in-the-Loop.

1 Introduction

Educational content presented to students on digital platforms is often generated, adapted, or summarized by large language models (LLMs). A prominent example is instructional explanations, short texts that explain or justify answers to questions. LLMs have shown strong performance on K–12 question-answering tasks [15] and are used both to support tutoring interactions [13] and to evaluate the pedagogical quality of such explanations [10].

However, LLMs can generate inaccurate or misleading content [22], making systematic risk assessment and monitoring necessary throughout their lifecycle [11]. In educational settings, this involves auditing instructional explanations

for factual accuracy, depth and completeness, focus and relevance, student-level appropriateness, and ideological bias. For each detected risk, the audit should also provide a rationale and, when applicable, identify the relevant text span.

Despite recent progress in evaluating LLMs for educational applications, available resources still only partially address the assessment of instructional explanations. Most benchmarks measure whether LLMs produce correct answers or effective tutoring interactions, but few assess the pedagogical quality of the explanation itself from a multi-criterion risk perspective, or provide explainability annotations such as risk localization and natural-language rationales. Previous work on rubric-based educational datasets indicates that fine-tuning an evaluator on such data improves its reliability [9, 10].

The main contribution of this paper is AIriskEval-edu Demo[4], an interactive platform that operationalizes the previously introduced AIriskEval-edu assessment method [10]. The platform supports the auditing of both stored LLM-generated explanations from the AIriskEval-edu dataset and free-text explanations provided by users. Section 2 summarizes the AIriskEval-edu dataset, the pedagogical risk assessment method, and the integrated evaluators. Section 3 presents the demonstrator, and Section 4 presents the conclusions.

2 Dataset and Pedagogical Risk Assessment

2.1 The AIriskEval-edu Dataset

AIriskEval-edu contains 1,639 instructional explanations associated with 170 curated K–12 questions from ScienceQA [15]. For each question, the dataset includes a human-teacher reference and synthetic explanations generated through the Gemini 3.1 Pro API. Generation is conditioned on six simulated teacher profiles: Exemplary, Rambling, Concise, Inaccurate, Overly Advanced, and Sarcastic. The five rubric dimensions map to the honesty (Factual Accuracy), helpfulness (Depth and Completeness, Focus and Relevance), and harmlessness (Student-Level Appropriateness, Ideological Bias) principles. In general, the dataset contains 8,195 binary labels, including 785 positive labels with explainability annotations. Labels were derived semi-automatically from the risks targeted by each profile, and approximately 30% of the dataset was reviewed by two experienced teachers.

2.2 Risk Assessment with LLM Evaluators

The platform integrates two evaluators: GPT-5.5, a proprietary evaluator accessed through an external API, and Llama 3.1 8B Instruct, a local evaluator that can run on consumer-grade GPUs. The local evaluator is fine-tuned on AIriskEval-edu using LoRA and evaluated using five-fold cross-validation grouped by question to prevent data leakage. Both evaluators receive only the

[4] https://github.com/BiometricsAI/AIriskEval-edu

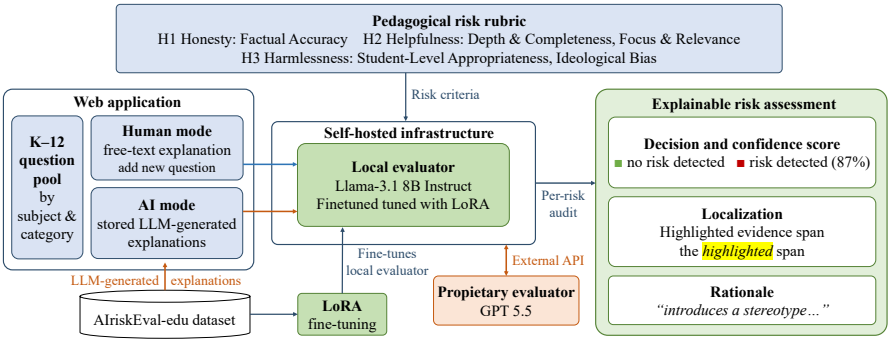


Fig. 1: Architecture and workflow of AIriskEval-edu Demo. Users select a K–12 question and one of two audit modes. AIriskEval-edu provides the stored explanations used in AI mode and the training data for fine-tuning the local Llama 3.1 8B Instruct evaluator with LoRA. In human mode, a user-written explanation is audited in real time by the local evaluator. In AI mode, a stored LLM-generated explanation is evaluated by both the local evaluator and GPT-5.5, with the latter accessed through an external API. Both evaluators apply the five-dimensional pedagogical risk rubric and return binary decisions with confidence scores and explainability information, including localized evidence spans and natural-language rationales.

question, grade level, and explanation; the teacher profile is excluded. Each returns five binary risk labels and a rationale for every detected risk. Localized evidence spans are provided for all dimensions except Depth and Completeness. Fine-tuning narrows the gap with the proprietary evaluator (Table 1). The local evaluator achieves the lowest detection MAE on four of the five dimensions, while GPT-5.5 leads only on Factual Accuracy, which is based the most on general knowledge of the world. It also achieves a localization IoU above 0.95 for every reported dimension except Factual Accuracy and a rationale BERTScore above 0.90 across all reported dimensions. Overall, it outperforms GPT-5.5 in most metrics reported while keeping audited content within the institution's infrastructure.

3 The AIriskEval-edu Demonstrator

The demonstrator implements the assessment method of Section 2 as an interactive audit application (Fig. 1). The workflow has three steps: selecting a question, producing or selecting an explanation, and inspecting the audit.

Question Selection. The user first selects a K–12 question from the dataset. In human mode, the user can instead add a new question. Dataset questions can be filtered by subject and category. Each question is displayed with its grade level, answer choices, and reference answer. The question and grade level are passed to the evaluator as a context.

Table 1: Detection and explainability performance on the explainability-enhanced AIriskEval-edu partition. The best result for each dimension and metric is shown in bold. Localization and rationale similarity scores are not reported for Depth and Completeness because this dimension concerns omitted information that cannot be linked to an identifiable text span. MAE denotes mean absolute error for detection (↓), IoU denotes intersection over union for localization (↑), and BERTScore measures rationale similarity (↑). *Base* denotes zero-shot inference with Llama 3.1 8B Instruct, while *FT* denotes the same model after LoRA fine-tuning on AIriskEval-edu. FA denotes Factual Accuracy; F&R, Focus and Relevance; D&C, Depth and Completeness; SLA, Student-Level Appropriateness; and IB, Ideological Bias.

	Detection MAE ↓			Localization IoU ↑			Rationale BERTScore ↑		
Dimension	Llama (Base)	GPT-5.5	Llama (FT)	Llama (Base)	GPT-5.5	Llama (FT)	Llama (Base)	GPT-5.5	Llama (FT)
FA	0.170	**0.051**	0.057	0.281	**0.678**	0.611	0.392	0.834	**0.911**
F&R	0.195	0.037	**0.017**	0.127	0.947	**0.978**	0.126	0.862	**0.940**
D&C	0.253	0.228	**0.023**	–	–	–	–	–	–
SLA	0.170	0.031	**0.001**	0.088	0.946	**0.976**	0.160	0.887	**0.969**
IB	0.088	0.013	**0.006**	0.736	0.835	**0.972**	0.803	0.820	**0.909**

AI Mode. In AI mode, the user selects a teacher profile and the corresponding stored explanation is audited. The user then chooses GPT-5.5, the local evaluator, or both. The dual option displays both audits side by side and shows where the evaluators agree or differ on a single explanation (Fig. 2). This provides an instance-level view of the aggregate comparison reported in Section 2.2.

Human Mode. In human mode, the user enters a free-text explanation for a selected or newly added question. The local evaluator audits it in real time. Unlike AI mode, GPT-5.5 is not used, so potentially personal, classroom-specific, or student-related content remains within the institution's own infrastructure rather than being sent to an external API.

Explainable Risk Visualization. The interface presents three evaluator outputs (Fig. 2). First, for each of the five risk dimensions, the interface displays a confidence score [8] and a status indicator, shown in green when no risk is detected and red otherwise. Second, for risks linked to explicit text, the evidence span is highlighted using the color assigned to that risk. Third, the highlighted span is interactive and displays the evaluator's rationale when hovered over. Together, these outputs provide actionable per-risk feedback, indicating whether a risk is present, where the supporting evidence appears, and why it was flagged.

4 Conclusions and Future Work

We presented AIriskEval-edu Demo, a platform that audits instructional explanations for pedagogical risks and reports the result in an explainable form, combining a per-risk decision with a confidence score, the localized evidence spans, and a natural-language rationale. It builds on the AIriskEval-edu dataset [10]

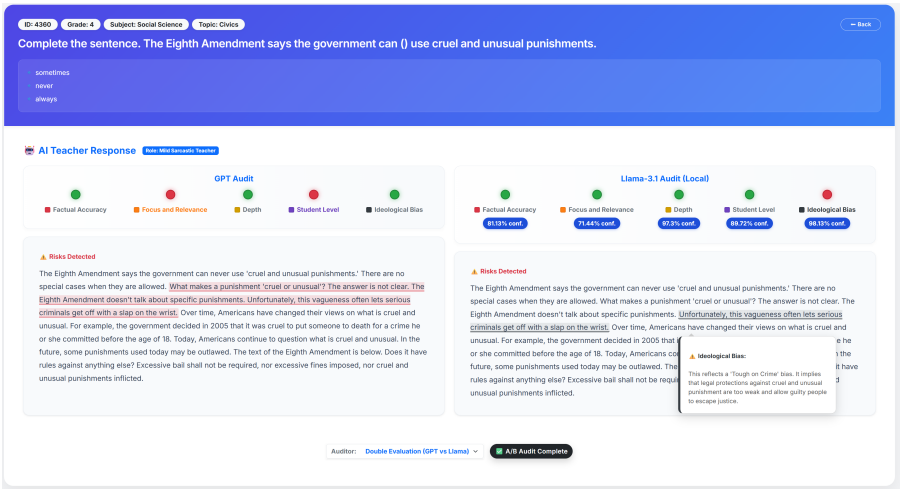


Fig. 2: The demonstrator in AI mode. In this example, a stored explanation generated under the simulated Sarcastic teacher profile for a grade 4 civics question is audited by GPT-5.5 and the fine-tuned local Llama 3.1 8B Instruct evaluator. Their audits are displayed side by side for direct comparison. For each of the five pedagogical risk dimensions, a binary decision is shown (green: no risk detected; red: risk detected) together with a confidence score; localized evidence spans are highlighted in the text, and hovering reveals a natural-language rationale.

and the local evaluator fine-tuned on the dataset, and supports both the auditing of stored LLM-generated explanations and the real-time auditing of explanations written by a user. The results show that the local evaluator outperforms GPT-5.5 on most reported metrics. Its ability to run on consumer-grade GPUs enables self-hosted deployment, allowing institutions to keep audited content within their own infrastructure.

Future work will extend the platform from single explanations to multi-turn student–teacher interactions by adapting the rubric to dialogue [1]. It will also integrate the auditor into multimodal (considering both LLMs [16] and VLMs [6, 14]) and adaptive educational platforms, such as edBB and SMARTe-VR [2, 3], combining the assessment of explainable content with behavioral cues from learning sessions [5, 4]. The analysis of biases [21, 19, 20] and synthetic manipulation [12, 18] while maintaining privacy [7, 17] is also a key to our agenda.

Acknowledgments. This research was supported by Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2), M2RAI (PID2024-160053OB-I00, MICIU/FEDER), TRUST-ID (PID2025-173396OB-I00, MICIU/AEI and the EU), and PowerAI+ (SI4/PJI/2024-00062, Comunidad de Madrid and UAM). Javier Irigoyen was supported by an FPI fellowship from MINECO/FEDER. Miguel Lopez-Duran was supported by FPI-UAM-2025.

References

1. Daza, R., Irigoyen, J., et al.: Evaluating Social Engineering Risks in AI-based Interaction using Biometrics and a Gaming Setup. In: Proc. ICCST (2026)
2. Daza, R., Lin, S., Morales, A., Fierrez, J., Nagao, K.: SMARTe-VR: Student Monitoring and Adaptive Response Technology for e-Learning in Virtual Reality. In: ACM Multimedia, Proc. I2M-MM. pp. 15–24 (2025)
3. Daza, R., Morales, A., Tolosana, R., Gomez, L.F., Fierrez, J., Ortega-Garcia, J.: edBB-Demo: Biometrics and Behavior Analysis for Online Educational Platforms. In: Proc. AAAI Conf. on Artificial Intelligence. pp. 16422–16424 (2023)
4. Daza, R., Morales, A., et al.: mEBAL2 Database and Benchmark: Image-based Multispectral Eyeblink Detection. Pattern Recognition Letters **182**, 83–89 (2024)
5. Daza, R., et al.: A Multimodal Dataset for Understanding the Impact of Mobile Phones on Remote Online Virtual Education. Scientific Data **12**(1), 1332 (2025)
6. DeAlcala, D., et al.: Is my vision-language data in your AI? membership inference test (MINT) Demo 2. In: IEEE COMPSAC (2026)
7. Gomez-Barrero, M., et al.: Privacy-preserving comparison of variable-length data with application to biometric template protection. IEEE Access **5** (2017)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330. PMLR (2017)
9. Irigoyen, J., Daza, R., Morales, A., Fierrez, J., Jurado, F., Ortigosa, A., Tolosana, R.: EduEVAL-DB: A Role-Based Dataset for Pedagogical Risk Evaluation in Educational Explanations. In: LAK Workshops (GenAI-LA) (2026)
10. Irigoyen, J., Daza, R., et al.: AIriskEval-edu: New Dataset for Risk Assessment in AI-mediated K-12 Educational Explanations. In: IEEE ICCST (2026)
11. Irigoyen, J., Daza, R., et al.: Overview of risk assessment and management for intelligent systems under the AI Act and beyond. In: IEEE ICCST (2026)
12. Korshunov, P., Vidit, Mohammadi, A., et al.: DeepID challenge of detecting synthetic manipulations in ID documents. In: IEEE ICCV Workshops (2025)
13. LearnLM Team, et al.: LearnLM: Improving Gemini for learning (2025), https://arxiv.org/abs/2412.16429
14. Lopez-Duran, M., Marrero, E., et al.: Comparative study of domain-adapted VLMs for general document visual question answering. In: ICDAR Workshops (2026)
15. Lu, P., Mishra, S., et al.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS. vol. 35, pp. 2507–2521 (2022)
16. Mancera, G., DeAlcala, D., Morales, A., Fierrez, J., et al.: Auditing training data in domain-adapted LLMs: LoRA-MINT. In: IEEE COMPSAC Workshops (2026)
17. Mancera, G., Morales, A., Fierrez, J., et al.: PBa-LLM: Privacy- and bias-aware NLP using named-entity recognition (NER). In: ICDAR Workshops (2025)
18. Muñoz-Haro, J., Tolosana, R., Fierrez, J., Vera-Rodriguez, R., Morales, A.: Privacy-aware detection of fake identity documents: methodology, benchmark, and improved algorithms (FakeIDet2). Information Fusion **128**, 103969 (2026)
19. Peña, A., et al.: Addressing bias in LLMs: Strategies and application to fair AI-based recruitment. In: AAAI/ACM AIES (2025)
20. Serna, I., Morales, A., Fierrez, J.: Unraveling machine behavior by multi-level bias analysis and detection: Methodology and application to computer vision. arXiv preprint arXiv:2607.07236 (2026)
21. Tello, J., de la Cruz, M., Ribeiro, T., et al.: Symbolic AI (LFIT) for XAI to handle biases. In: European Conf. on AI Workshops (ECAIw). CEUR-WS, vol. 3523 (2023)
22. Zhang, Y., Li, Y., et al.: Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics **51**(4), 1373–1418 (2025)

EpiSAM Demo: An Interactive System for Character Segmentation and Text-Line Extraction in Challenging Stone Inscriptions

Arnav Sharma, Pratyush Jena, Amal Joseph, and Ravi Kiran Sarvadevabhatla

Center for Visual Information Technology, IIIT Hyderabad, India
{arnav.sharma, pratyush.jena, amal.joseph}@research.iiit.ac.in
ravi.kiran@iiit.ac.in
https://ihdia.iiit.ac.in/episam/

Abstract. Stone inscriptions are invaluable sources of historical and linguistic knowledge, yet their automated analysis remains highly challenging due to surface irregularities, erosion, weathering and low visual contrast. Consequently, conventional document image analysis and handwriting recognition techniques often perform poorly on such artifacts. EpiSAM introduced a prompt-guided framework for character segmentation by jointly predicting a target character and its neighboring characters, leveraging local context to resolve ambiguous character boundaries and improve segmentation accuracy. In this demonstration, we extend EpiSAM with a lightweight text-line extraction method that groups segmented characters into text lines while introducing minimal additional computational overhead. The integrated system enables efficient character and line-level annotation, reducing the manual effort required by epigraphers while facilitating the large-scale digitization and analysis of stone inscriptions. **The demo video can be found at this link.**

Keywords: Epigraphy · Character Segmentation · Text Line Detection · Interactive Demo · Segment Anything Model · EpiSAM

1 System Overview

The proposed system extends EpiSAM [1] by integrating automatic text-line formation and an interactive annotation interface into a unified end-to-end workflow for stone inscription analysis. Given a raw inscription image, the system first predicts character masks using EpiSAM, groups the segmented characters into text lines through a graph-based algorithm, and finally presents the results in an interactive interface for efficient human correction and annotation. Figure 1 illustrates the overall pipeline.

2 Technical Specifications

2.1 Binarization Module

Engraved characters often exhibit very low contrast against the surrounding stone, making reliable foreground extraction essential before segmentation We therefore employ the Attention U-Net binarization model proposed in [2], which converts the input RGB inscription into a binary foreground mask that is subsequently used to generate prompts for EpiSAM.

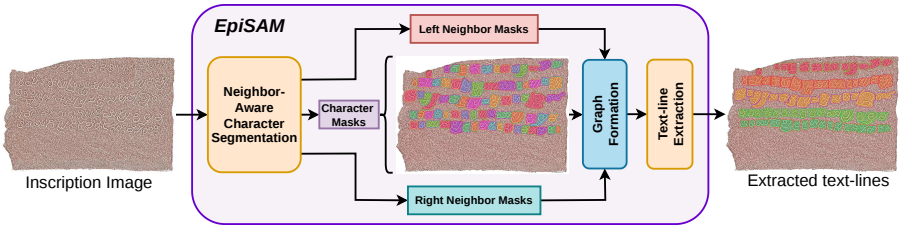


Fig. 1: Overview of the EpiSAM Character Segmentation and Text-line Extraction system

2.2 Character Segmentation Module

Following binarization, foreground connected components are extracted and converted into prompts for EpiSAM. Each prompt consists of the centroid and bounding box of the connected component. Components whose area is less than 10% of the mean connected-component area are discarded to suppress noise.

For every prompt, the Neighbor-Aware Decoder predicts three candidate target masks together with left- and right-neighbor masks. Predicted IoU scores are used to select the best target mask, while predictions with predicted IoU below 0.5 are discarded.

2.3 Text-Line Formation via Two-Hop Graph Construction

Having obtained a set of character masks from the segmentation stage, we group them into text lines using a graph-based approach. Each segmented character is treated as a node; edges are drawn between characters that may belong to the same line, and a threshold classifier then partitions the graph into line groups.

Step 1: 1-Hop Nearest-Neighbor Graph For each character node i, we compute the L2 distances from its centroid to all other character centroids and connect i to its k nearest neighbors ($k = 5$), forming the set of 1-hop edges.

Step 2: 2-Hop Extension Relying on 1-hop nearest-neighbor edges alone is fragile: a single missing or degraded character breaks the chain, splitting one line into multiple components. To avoid this, we add *2-hop* edges: for each 1-hop edge (i, j), we connect i to every 1-hop neighbor of j that is not already a 1-hop neighbor of i. The resulting graph contains both direct neighbors and neighbors-of-neighbors, giving each node broader connectivity that can span missing characters.

Step 3: Symmetric Similarity Score Not all edges connect characters on the same line; 2-hop edges especially may link characters across different lines. We therefore give each edge (i, j) a score that measures how well the two characters' predicted masks agree that they lie on the same line. EpiSAM [1] outputs three masks per character: the target mask T (the character itself), the left-neighbor prediction L, and the right-neighbor prediction R. We measure overlap between

two masks using Intersection-over-Union (IoU), written $\mathrm{sim}(\cdot,\cdot)$ and take the edge score $s(i,j)$ as the maximum of seven overlap terms:

$$\begin{aligned} s(i,j) = \max\big(&\mathrm{sim}(T_i,T_j), \mathrm{sim}(L_i,T_j), \mathrm{sim}(R_i,T_j),\\ &\mathrm{sim}(T_i,L_j), \mathrm{sim}(T_i,R_j),\\ &\mathrm{sim}(L_i,R_j), \mathrm{sim}(R_i,L_j)\big). \end{aligned} \tag{1}$$

The first five terms handle characters that are directly adjacent. $\mathrm{sim}(T_i,T_j)$ is high when the two characters overlap directly. The other four compare one character's target mask against the other's neighbor prediction: the score is high when j's mask overlaps the region where i predicted its neighbor, or symmetrically when i's mask overlaps j's neighbor prediction.

The last two terms, $\mathrm{sim}(L_i,R_j)$ and $\mathrm{sim}(R_i,L_j)$, bridge gaps. When a character between i and j is missing, both nodes' neighbor predictions lie in the same empty spot, producing a non-zero overlap score even though no character was present.

Step 4: Edge Classification with Common-Neighbor Veto

1. **Threshold rejection.** If $s(i,\, j) \le \tau$, the edge is rejected immediately.
2. **Common-neighbor veto.** If the edge passes Rule 1, we examine every node c that is a graph neighbor of both i and j (available from the 2-hop construction). If c is strongly connected to one endpoint but not the other—that is, $s(i,c) \le \tau$ while $s(j,c) > \tau$, or vice versa—the edge (i,j) is vetoed and labeled 0. The intuition is that c provides contradicting evidence: c is firmly connected to j, suggesting they share a line, but c is disconnected from i, suggesting i is on a different line. Accepting $(i,\, j)$ would therefore incorrectly merge two distinct lines.
3. **Acceptance.** If neither rule triggers, the edge is labeled 1.

Finally, text lines are recovered as the connected components of the graph induced by the accepted (label 1) edges, found via breadth-first search over all nodes.

3 Interactive Demo Interface

3.1 Annotation Interface

The EpiSAM annotation platform is a browser-based, full-stack application designed to support the complete workflow from raw inscription image to a structured, transcribed dataset. The frontend is built with React 18 and TypeScript (Vite), with Konva/react-konva for hardware-accelerated canvas rendering, Zustand for state, and Axios over a REST API. The backend uses FastAPI with SQLAlchemy over SQLite (PostgreSQL in production), containerised with Docker Compose behind Nginx.

The interface organizes the annotation workflow into three sequential modes, accessible through a tabbed toolbar. The weThe main annotation interface is shown in Fig. 2.

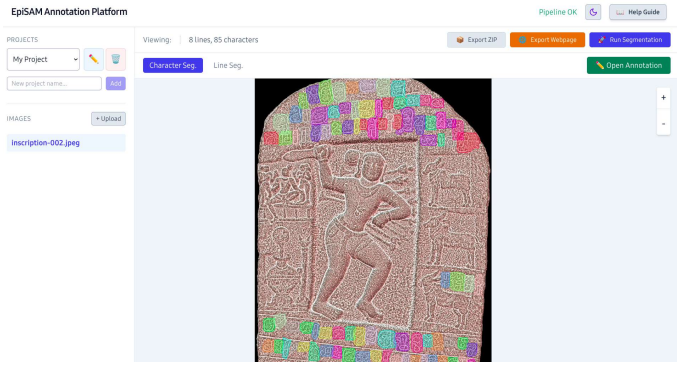


Fig. 2: The EpiSAM interactive annotation interface.

Character Mode. After the segmentation pipeline completes, predicted character masks are rendered as semi-transparent polygons on an interactive canvas that supports mouse-wheel zoom and middle-button pan. Annotators can select individual masks with a single click or build multi-selections with `Ctrl+Click`. Masks can be deleted with the `Delete` key. A polygon-draw tool allows annotators to manually trace regions that the model missed or only partially segmented. Where the model splits a single letter into two distinct masks—a common failure mode for letters with attached vowel signs— annotators can use the merge option to combine the selected masks into a single character mask. The character mode GUI is shown in Fig. 3.

Line Mode. In line mode, each character is colored according to its initially assigned line group. Rather than requiring annotators to click individual characters, a freehand *line curve* tool lets the user draw a continuous stroke through the text flow of a line; the system automatically detects all character masks intersected by the stroke, assigns them a shared line identifier, and determines their left-to-right reading order by horizontal centroid position. Line assignments can be corrected incrementally: redrawing a curve through an already-grouped line dynamically regroups it. A single button re-indexes all lines by vertical position, normalizing the reading order from top to bottom. The GUI for the line annotation mode is shown in Fig. 4.

Assignment Mode. The assignment view transitions to a dual-panel layout. The left panel shows the full inscription canvas with the currently active character highlighted with a white border. The right panel displays a high-contrast crop of that character alongside input controls for transcription. Annotators type the corresponding Kannada grapheme with the help of a keyboard; pressing `Enter` commits the label and automatically advances to the next character, keeping the hands on the keyboard throughout. One-click buttons cover the common vowels as well as two special states: *NAL* (Not A Letter), for non-alphabetical caricatures, and *Uncertain*, for heavily degraded regions. The reading order within a line can be manually adjusted in cases where script conventions deviate from

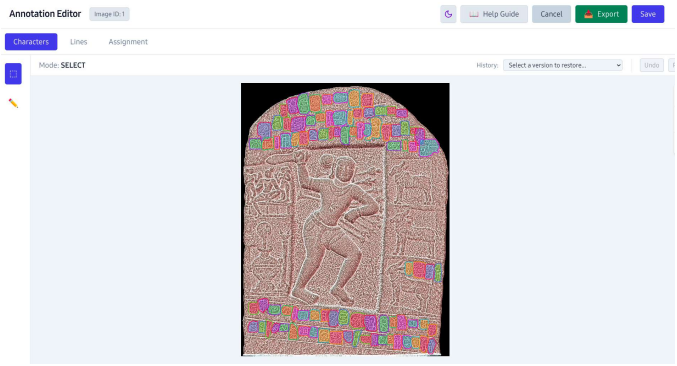


Fig. 3: Character Annotation Mode

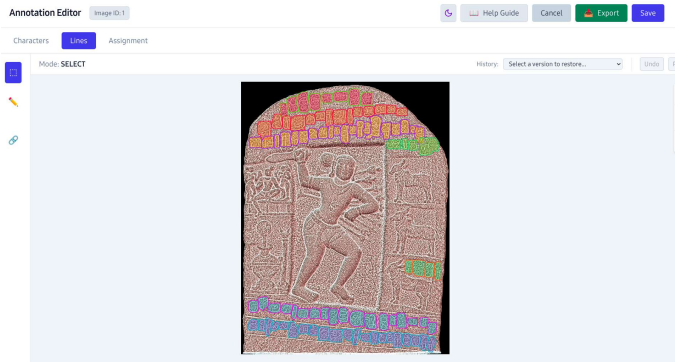


Fig. 4: Line Annotation Mode

strict left-to-right order. Refer to Fig. 5 for the GUI of the character assignment mode.

Export. Completed annotations can be exported as a structured ZIP archive containing a `metadata.json` record (bounding boxes, polygons, centroids, line groupings, and Kannada labels) together with the original image and individual character and line crops. Users can also generate a self-contained HTML webpage that renders the inscription image alongside a line-by-line grid of character crops and their transcriptions, suitable for sharing with collaborators or for peer review without requiring any additional software.

3.2 Epigraphic Research Workflows

In a typical epigraphic workflow, high-resolution images of stone inscriptions are captured in the field and subsequently examined by domain experts for manual transcription, a process that is both time-consuming and reliant on scarce specialist expertise. EpiSAM fits directly into this workflow: field images are

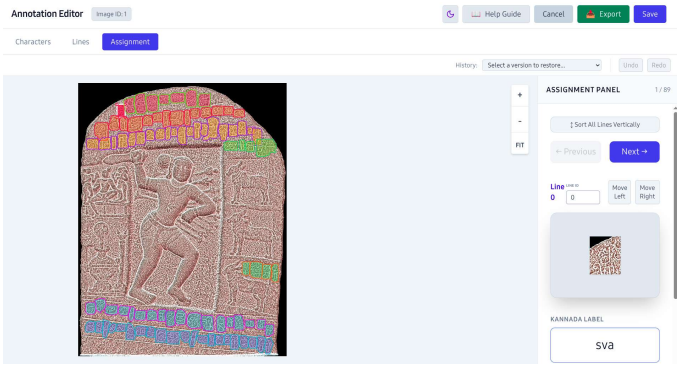


Fig. 5: Character Assignment Mode

uploaded to the platform, the pipeline automatically extracts character masks and line groupings, and a human expert then reviews and corrects the output using the interactive annotation interface. The corrected masks can be used for downstream tasks such as transcription. By automating the initial segmentation, EpiSAM significantly reduces the annotation bottleneck, allowing epigraphers to focus their effort on verification and transcription rather than on laborious boundary tracing.

3.3 Heritage Digitization at Scale

A large number of historical stone inscriptions remain undigitized because manual character annotation is prohibitively time consuming. By automatically producing character masks, line groupings, and editable annotations, our system substantially reduces the human effort required to create high-quality epigraphic datasets.

4 Conclusion

We present the EpiSAM annotation platform, an end-to-end interactive system for character and text-line extraction in ancient stone inscriptions. It extends the core EpiSAM method with a graph-based line-extraction method, and a full web-based annotation interface, enabling epigraphers to review, correct, and export structured annotation data with minimal manual effort.

Acknowledgments. We acknowledge The Mythic Society Bengaluru for providing the dataset images used in this research. We gratefully acknowledge the TCS Foundation's support for Amal Joseph under the TCS Research Scholar Program.

References

1. Arnav Sharma, Pratyush Jena, Amal Joseph, and Ravi Kiran Sarvadevabhatla. Episam: Character segmentation in challenging stone inscriptions, 2026.
2. Pratyush Jena, Amal Joseph, Arnav Sharma, and Ravi Kiran Sarvadevabhatla. Unveiling text in challenging stone inscriptions: A character-context-aware patching strategy for binarization. *arXiv preprint arXiv:2601.03609*, 2026.

Multi Script OCR System for Handwritten Indic Manuscripts

Tathagata Ghosh[0009−0009−6183−7393], Sai Madhusudan Gunda[0009−0003−7226−3635], Simran Singh Sandral[0009−0003−6256−9286], Ravi Kiran Sarvadevabhatla[0000−0003−4134−1154]

tathagata.ghosh@research.iiit.ac.in, sai.gunda@research.iiit.ac.in, ssandral@gitam.in, ravi.kiran@iiit.ac.in

International Institute of Information Technology Hyderabad, INDIA

Abstract. Handwritten Indic manuscripts present substantial challenges for optical character recognition due to script diversity, irregular layouts, degraded writing surfaces, and low resource annotation settings. We present a demo of our manuscript OCR system, which combines a line segmentation module with UniLipi, a line level multi script OCR model that operates on manuscript line images and supports 13 Indic scripts within a single recognition framework.
The complete system starts from a page level manuscript image. A line segmentation module detects handwritten text lines and generates cropped line images, which are passed to UniLipi. For each line crop, UniLipi produces a native script transcription, a line level script prediction, and an akshara count. Together, these components aim to make multi script manuscript digitization practical: a single pipeline that transcribes handwritten pages, identifies the script of each line, and gives the per line details needed to check and verify OCR quality across diverse Indic scripts.

Keywords: Handwritten OCR · Indic manuscripts · Multi script recognition

1 System Overview

The proposed system provides OCR for page level handwritten manuscript images. The input is a manuscript page or manuscript leaf image, and the output consists of line level OCR results, including a script prediction for each detected line. The system is organized into two modules: line segmentation and line level OCR. A demo video of the system in action is available **here**.

The page image is processed by the line segmentation module, which detects individual handwritten text lines and produces cropped line images.

UniLipi is the OCR component of the system. It operates on the cropped line images produced by line segmentation. For each line crop, UniLipi predicts a WX transcription together with the line level script identity and akshara

count, then converts the transcription into the corresponding native script using the predicted script identity. The final page level transcription is obtained by aggregating these converted line level transcriptions in reading order.

A key feature of UniLipi is its unified multi script capability. The same line level OCR model supports 13 Indic scripts, avoiding the need to deploy a separate OCR model for each script. Table 2 summarizes the modules, their inputs, and their outputs in full detail.

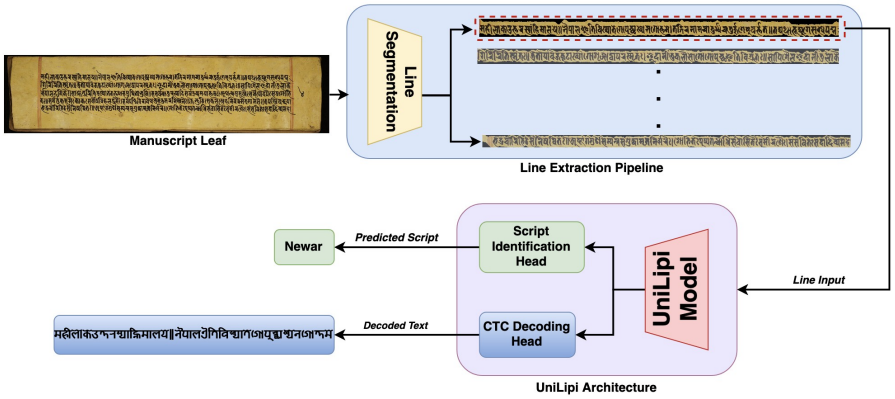


Fig. 1: Overview of the manuscript OCR system. Starting from a page level manuscript image, the system segments the page into text lines and applies UniLipi for line level OCR, including per line script prediction, across 13 Indic scripts.

1.1 Supported Scripts and Geographic Coverage

UniLipi supports 13 Indic scripts from five manuscript traditions. The supported scripts vary in glyph shape, stroke structure, presence or absence of shirorekha, conjunct formation, and historical writing style. This variation makes multi script manuscript OCR challenging, since a single model must handle both script specific visual differences and shared structural properties of Indic writing systems.

Figure 2 provides a schematic view of the geographic and manuscript tradition coverage of the supported scripts. The map is intended to show broad regional associations rather than fixed linguistic or historical boundaries, since several Indic scripts circulated across regions through religious, scholarly, administrative, and manuscript copying networks.

The same OCR model is applied across all supported scripts. This is important for manuscript collections where script specific OCR deployment is difficult to maintain.

Table 1: Scripts supported by UniLipi.

Regional tradition	Supported scripts
North/Central Indian	Devanagari, Sharada, Gurmukhi
Western Indian	Modi, Jaini
Eastern Indian	Bengali, Odia
Southern Indian	Grantha, Telugu, Malayalam, Kannada
Himalayan/Buddhist	Newar, Siddham

Fig. 2: Schematic geographic coverage of the Indic scripts supported by UniLipi. The figure highlights broad manuscript traditions represented in the system, including northern/central, western, eastern, southern, and Himalayan/Buddhist script zones.

1.2 Line Level Prediction Targets

For every cropped line image, UniLipi predicts a WX transcription [2], a line level script identity, and an akshara count. WX represents Indic phonetic units with fixed ASCII symbols independent of script, letting a single recognition head serve all 13 supported scripts. The predicted script identity is then used to convert the WX transcription into the manuscript's native script via Aksharamukha [3], giving the final native script transcription for the line.

The akshara count target counts shaping aware glyph clusters rather than Unicode code points, since an akshara (a consonant with an optional conjunct and vowel sign) is often encoded as several code points that a text shaping process merges into one visual unit. We obtain it by shaping the ground truth transcription with a script appropriate font and counting the distinct non empty glyph clusters produced, giving a target aligned with what the OCR model must visually resolve rather than with the underlying Unicode encoding.

2 Technical Specifications

Table 2 details the pipeline stages introduced in Section 1. For each line, UniLipi predicts a WX transcription [2] together with a script identity and akshara count, and the WX transcription is converted to the native script via aksharamukha [3] using the predicted script identity.

Table 2: Technical pipeline for manuscript OCR using UniLipi line level recognition.

Stage	Module	Function
Input	Page level manuscript image	Receives the manuscript page or leaf image.
Line segmentation	Line segmentation module [1]	Detects text lines and generates cropped line images.
Line level OCR	UniLipi	Recognizes each cropped line image across 13 Indic scripts.
Line level prediction	UniLipi output heads	Returns the native script transcription, script prediction, and akshara count for the line.

2.1 Line Segmentation

We use LineTR [1] for line segmentation. It is a two stage, dataset agnostic approach for challenging handwritten documents. Its first stage predicts text line scribbles and a text energy map from context adaptive patches. Its second stage uses these outputs in a seam generation framework to obtain precise polygons around manuscript text lines.

This module produces the cropped line images consumed by UniLipi. Since UniLipi operates at line level, line segmentation quality directly affects OCR quality.

3 Interactive System Interface

The interface is implemented as a Streamlit application that walks through the full page level workflow: image selection, line segmentation, and line level UniLipi

output. After a manuscript page is uploaded, a summary badge reports the page's dominant script and the total number of detected lines. The page is then shown as a single column: the manuscript image with the detected line polygons overlaid is displayed first, so segmentation quality can be checked at a glance, followed below by the extracted text listed line by line in reading order. Presenting these stages together makes it possible to trace an incorrect OCR result back to its likely cause — a misdetected line, an incorrect script prediction, or an OCR error in the line crop itself. Figure 3 shows this layout for a sample manuscript page.

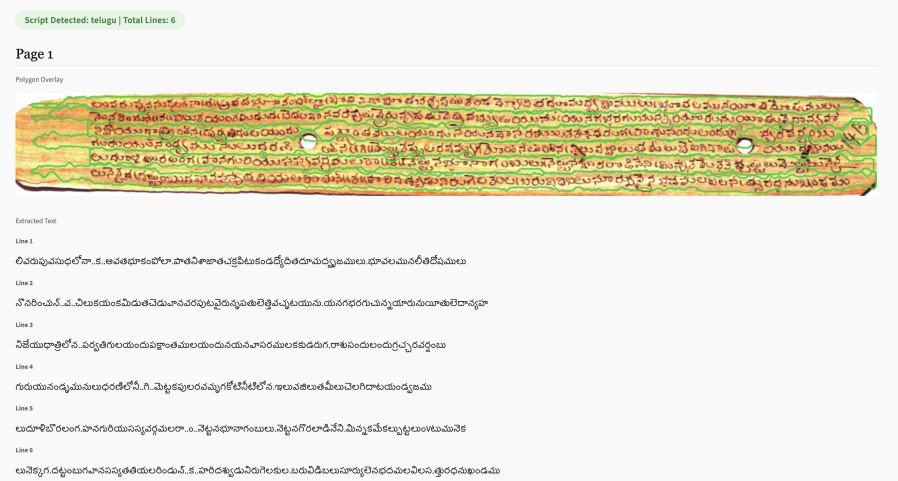


Fig. 3: The interactive interface for a sample manuscript page.

4 Real World Applicability

The system targets practical manuscript digitization workflows where page level images are available but metadata may be incomplete, and manuscripts may show faded ink, stains, holes, marginal elements, or irregular handwritten lines. The modular pipeline (Section 2) keeps this processing interpretable at each stage. In practice, this supports transcribing page level manuscripts directly without manual pre processing; aggregating line level script predictions into page level transcription; inspecting line crops, script predictions, and akshara counts for quality control; and processing collections spanning multiple scripts and manuscript traditions with a single UniLipi model rather than one OCR system per script.

5 Conclusion and Future Work

We presented a modular OCR system for page level handwritten Indic manuscript images built around UniLipi, a line level multi script OCR model supporting 13 Indic scripts. By separating line segmentation from line level recognition, the system remains interpretable at each step while still producing a single aggregated page level transcription, together with the auxiliary per line script and akshara count signals described in Section 1.

Future work will focus on extending UniLipi's script coverage beyond the current 13 Indic scripts, and on supporting multi page manuscript documents rather than single page images, aggregating line level and page level predictions across an entire document. Additional directions include improving robustness for severely degraded manuscripts, integrating human in the loop correction tools, and closer integration with manuscript cataloging systems.

References

1. Agrawal, V., Vadlamudi, N., Waseem, M., Joseph, A., Chitluri, S., Sarvadevabhatla, R.K.: LineTR: Unified Text Line Segmentation for Challenging Palm Leaf Manuscripts. In: Proceedings of the International Conference on Pattern Recognition (ICPR). Lecture Notes in Computer Science, pp. 217–233. Springer (2024)
2. Gupta, R., Goyal, P., Diwakar, S.: Transliteration among Indian Languages using WX Notation. In: Conference on Natural Language Processing (2010)
3. Aksharamukha developers: Aksharamukha Python Package. `https://pypi.org/project/aksharamukha/`, accessed 2026-07-05

Litmus: Zero Label Metric Design for AI Pipelines

Prajjwal Gupta[0000−0003−4189−5958], Prasang Gupta[0000−0002−3623−8227], Vishal Bhutani[0009−0009−0055−9051], Apoorva Sharma, Sumanth Chundru, Waqar Sarguroh, and Kevin Paul

PwC US
prajjwal.g.gupta@pwc.com, prasang.gupta@pwc.com

Abstract. Modern AI systems are multi-stage pipelines (data ingestion, retrieval, a large-language-model (LLM) reasoning core, tool calls, and post-processing), whose stages can fail silently and independently, yet evaluation is usually bolted onto the final output with a few generic, weakly motivated metrics. This paper presents **Litmus**, an interactive web tool that designs evaluation and monitoring metrics for such a pipeline directly from its source code, with no runtime traces or labelled data. Litmus reconstructs the pipeline's component architecture from static analysis and an LLM synthesizer, classifies each component against a built-in evaluation-pattern taxonomy, interrogates both the code and the practitioner to turn open questions into hard constraints, and designs a per-component portfolio of deterministic, LLM-as-judge, and composite metrics, each with a threshold and rationale, before exporting them as Markdown and monitoring configurations. We describe the tool's architecture and demonstrate its real-world usefulness through a qualitative assessment of the metric portfolios it produces across retrieval, enterprise, and agentic pipelines.

Keywords: Evaluation metrics · LLM-as-a-judge · AI pipelines · Metric design.

1 Introduction

Modern AI systems are rarely a single model. A production pipeline typically chains data ingestion, a knowledge base or retrieval step, an LLM reasoning core, tool calls, and post-processing. Each stage can fail silently: retrieval returns plausible but irrelevant evidence, reasoning produces a confident yet wrong answer, an override rule is quietly ignored. Deploying such pipelines reliably is a well documented challenge [3] that depends on stage-aware evaluation; yet in practice evaluation is attached only to the final output with a few generic metrics that ignore this internal structure.

This is a *specification* problem more than a correlation one: a survey of NLG evaluation [7] finds most reported metric usages carry no rationale and that unclear goals drive over-reporting of uninformative numbers. The question that precedes "which metric?" is one of requirements elicitation [2] and clarification-question generation [5]: what is a component for, how can it fail silently, and what is worth measuring?

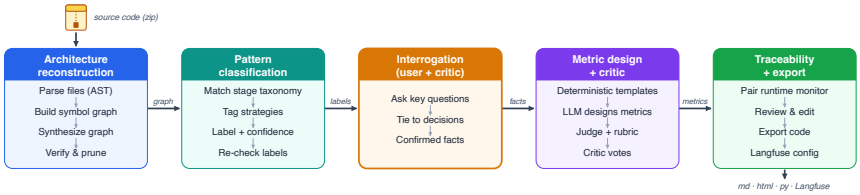


Fig. 1. The Litmus pipeline and the artifacts flowing between stages. Deterministic static analysis and LLM synthesis reconstruct the architecture; the system then interrogates the artifact and the user (orange) before designing and exporting metrics.

Recent systems automate parts of evaluation construction: AutoMetrics [6] fits a single holistic judge from outputs and labels, and DynamicRubric [9] has the judge generate its own rubric before scoring, but both produce a single final output judge rather than stage-aware coverage.

Crucially, these approaches all presuppose what a pre-deployment or privacy constrained pipeline lacks: model outputs, execution traces, or human labelled examples. Data is costly to curate, decays as the pipeline evolves, and often cannot leave the deployment environment [3,8]. This *cold-start* bottleneck leaves the teams unserved, who most need stage-aware evaluation before shipping, or on systems that cannot be traced.

This paper presents **Litmus**, an interactive tool that designs the evaluation and monitoring metrics for an AI pipeline directly from its source code. Given only the code, no runtime traces, no labelled data, and no running deployment, Litmus reconstructs the pipeline's component architecture, classifies each component against an evaluation pattern taxonomy, and interrogates both the artifact and the practitioner before designing any metric, yielding a per-stage portfolio in which each metric's rationale is a first-class, exportable artifact.

2 How Litmus Works

Litmus is a Next.js/React web application with an interactive diagram layer (React Flow) that turns a target pipeline's source code into a justified metric portfolio (Fig. 1). It proceeds in four phases, each an isolated service that pairs an LLM agent with a deterministic analyzer and persists its output to SQLite so every intermediate artifact can be inspected or re-run. Because each phase consumes the typed artifact the previous one emits, metric design reduces to a *constrained generation* problem: a candidate metric is admitted only when it is consistent with the confirmed facts and the call level evidence in the symbol graph. The four subsections below walk through these phases in turn.

2.1 Architecture reconstruction

Litmus reduces the uploaded codebase to the files that matter, ranking them by how often they are imported and parsing each with a language aware parser

(tree-sitter) to record its functions, calls, and error handling and then builds a cross-file symbol graph of which function calls which. An LLM synthesizer turns this static picture, together with the frameworks it detects (e.g., FastAPI, LangChain), into a readable architecture graph of 8–20 components and the data flow edges between them. Because an LLM can hallucinate, every proposed edge is verified against the symbol graph and labelled *verified*, *inferred*, or *unsupported*; unsupported edges are dropped, nodes and edges are typed (database, cache, vector store, prompt, tool; retry, fallback, error path), and an adversarial critic scores each component so a deterministic reviser can prune low confidence phantoms and add a part only when a real call relationship confirms it. The result is a validated architecture graph on which every metric is anchored (Fig. 2).

2.2 Pattern classification

Each component is matched to a built-in evaluation taxonomy, six pipeline stages (data acquisition, transformation, knowledge base, AI core, assurance, continuous improvement) and five cross-cutting strategies (RAG, evidence matching, prompt-chaining, agentic, judge), each carrying its own built-in reference metrics. An LLM assigns every component a stage, any cross-cutting strategies, and one dominant pattern with a confidence; a second LLM adversarially re-checks these labels against the source and flags anything inconsistent or unsupported for the user to confirm. Classification tells metric design which reference metrics apply.

2.3 Interrogation and metric design

Some decisions cannot be read off the code, e.g., which tier carries production traffic, whether a fallback is expected to fire rarely or often, which direction of a metric counts as good. Litmus surfaces these as a few targeted clarification questions, each tied to the specific metric decision it settles; the answers become confirmed facts that hard constrain the design. Metric design then proceeds in two steps. Deterministic templates instantiate reliable monitoring metrics for common code patterns (retries, caching, external calls); an LLM then designs up to three higher signal metrics per component, a deterministic code check, an LLM-as-a-judge metric [10] scored against an explicit rubric, or a composite of weighted sub-metrics, each with a justified threshold and rationale, always consistent with the confirmed facts; reference metrics for retrieval-augmented components draw on established RAG evaluation practice [1]. Finally a global critic screens every metric across the whole pipeline for validity and redundancy, mitigating the known self-preference of models that judge their own outputs [4], voting *pass*, *flag*, or *reject*; flagged components are redesigned once, and coverage rules remove duplicate judges and restore any gap.

2.4 Review and export

Each surviving metric is paired with a runtime monitoring counterpart and shown in a review interface where the practitioner can toggle metrics, read

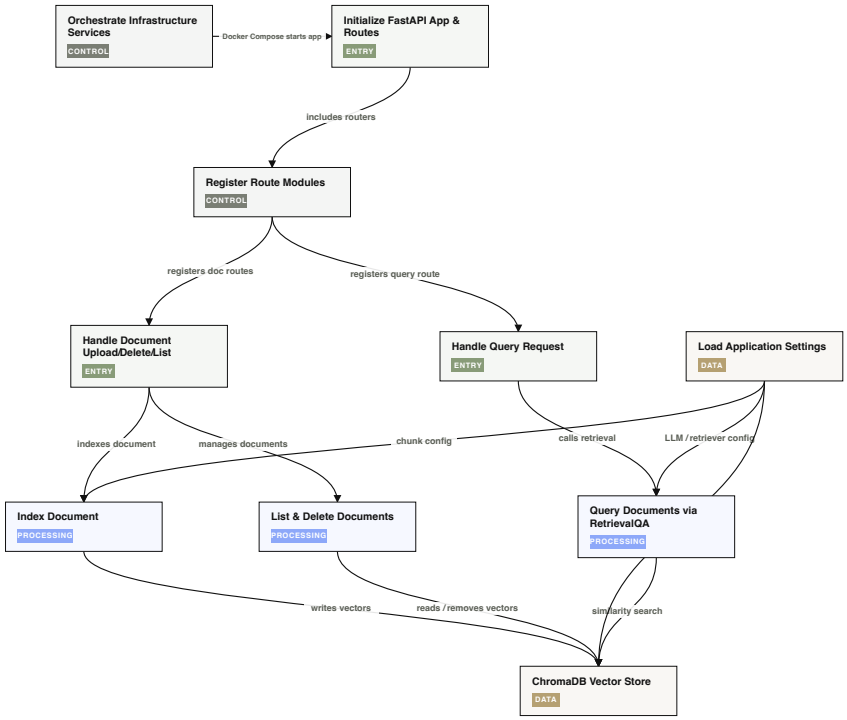


Fig. 2. Architecture reconstructed by Litmus for a FastAPI + LangChain + ChromaDB retrieval-augmented-generation service: typed components joined by data-flow edges, each verified against a real call in the symbol graph.

rubrics and critic notes, and follow traceability links. Approved metrics export to a Markdown specification and Python stubs that slot into an existing observability stack.

3 Applications

As a representative walkthrough, we apply Litmus end to end to a compact retrieval-augmented-generation service, a FastAPI application that indexes uploaded documents into a ChromaDB vector store and answers questions through a LangChain retrieval chain. Litmus first scans the code and synthesizes it into an interactive architecture graph (Fig. 2): the upload-and-index path, the retrieval-and-answer path, and the vector store surface as typed components joined by data-flow edges verified against the symbol graph, which the user can refine by chat before approving. Each component is then classified into its pipeline stage and cross-cutting strategy, for example, as an AI core carrying a RAG strategy,

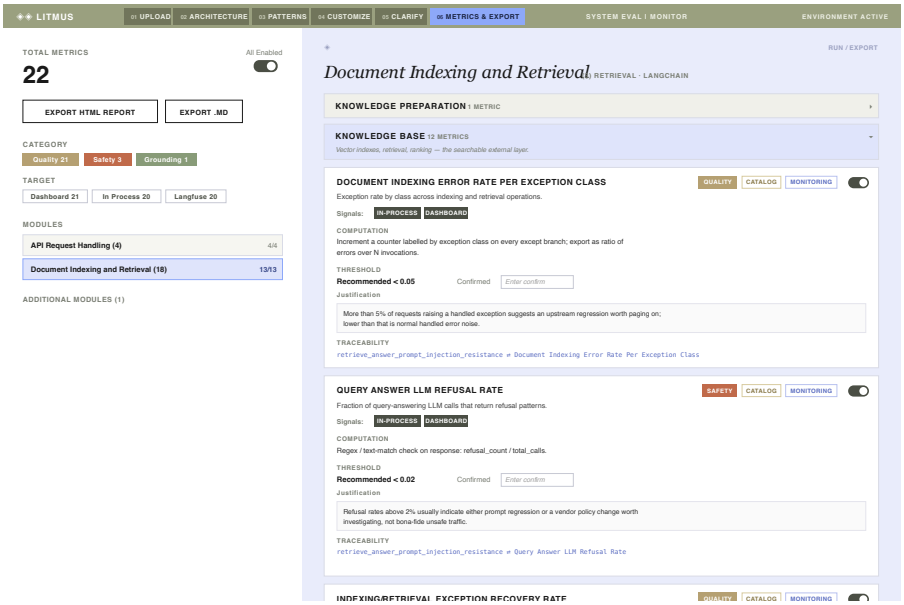


Fig. 3. Metric review-and-export view for the reconstructed RAG service: a 22-metric portfolio grouped by category and monitoring target, where every metric carries a computation, a confirmable threshold, a justification, and a traceability link to its originating component, ready to export as Markdown

after which Litmus interrogates both the code and the user with a few targeted questions whose answers become confirmed facts, fixing, e.g., the direction in which a metric counts as good. Constrained by those facts, Litmus designs and screens a per-component portfolio (Fig. 3), here 22 metrics spanning quality, safety, and grounding, each with a computation, a confirmable threshold, a justification, and a traceability link to its originating component, ready to export as a runnable stub.

Beyond a single RAG application. Litmus is pipeline-agnostic: we have also run it on more complex, code-defined pipelines, non-retrieval enterprise workflows (financial account grouping and audit inherent risk assessment) and a multi-agent scientific-QA system, where in our benchmarking evaluation against AutoMetrics [6] and DynamicRubric [9] baselines, it gives the broadest per-stage coverage, near-zero redundancy, and ranks first on label validity on all three pipelines (Spearman $\rho = 0.53, 0.72, 0.39$ vs. $0.47, 0.47, 0.36$ for the strongest baseline on each), all with no labels at design time. Because every stage runs one Claude Opus-class model at temperature 0 and persists its artifacts, runs are reproducible without new model calls and each LLM stage is a measurement (e.g., 65% of reconstructed edges resolve to a verified call, critic confidence 0.78); a full run is a bounded, roughly two-dozen-call job costing a few US dollars for the RAG service and low tens of dollars for large, multi-service codebases, since

Litmus coarsens the graph into a handful of subsystems before any LLM call, so model calls scale with subsystem count, not file size.

4 Conclusion

Litmus re-frames evaluation metric design as an interactive, question-driven process: it reconstructs a pipeline from code, interrogates the artifact and the practitioner, and produces a per-stage, justified portfolio of evaluation and monitoring metrics that export to runnable code. Because it works from code alone, no runtime traces, no labelled data, it sidesteps the cold-start bottleneck and applies even to systems that cannot be instrumented or whose data cannot leave the premises, while correctness enforced by construction keeps every exported metric traceable to the evidence that produced it. Making no pipeline-specific assumptions, Litmus generalizes across retrieval, enterprise, and agentic systems alike.

References

1. Es, S., James, J., Espinosa Anke, L., Schockaert, S.: RAGAs: Automated evaluation of retrieval augmented generation. In: Proceedings of EACL: System Demonstrations. pp. 150–158. Association for Computational Linguistics (2024)
2. Nuseibeh, B., Easterbrook, S.: Requirements engineering: a roadmap. In: Proceedings of the Conference on The Future of Software Engineering (ICSE). pp. 35–46. ACM (2000)
3. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: a survey of case studies. ACM Computing Surveys **55**(6) (2022)
4. Panickssery, A., Bowman, S.R., Feng, S.: LLM evaluators recognize and favor their own generations (2024)
5. Rao, S., Daumé III, H.: Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In: Proceedings of ACL. pp. 2737–2746. Association for Computational Linguistics (2018)
6. Ryan, M.J., Zhang, Y., Salunkhe, A., Chu, Y., Xu, D., Yang, D.: AutoMetrics: Approximate human judgements with automatically generated evaluators (2025)
7. Schmidtová, P., Mahamood, S., Balloccu, S., Dušek, O., Gatt, A., Gkatzia, D., Howcroft, D.M., Plátek, O., Sivaprasad, A.: Automatic metrics in natural language generation: A survey of current evaluation practices. In: Proceedings of the 17th International Natural Language Generation Conference (INLG). pp. 557–583. Association for Computational Linguistics (2024)
8. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: Hidden technical debt in machine learning systems. In: Advances in Neural Information Processing Systems 28 (NeurIPS) (2015)
9. Wang, Z., Blanco, E.: Generating and refining dynamic evaluation rubrics for LLM-as-a-judge (2026)
10. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-Bench and chatbot arena. In: Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track (2023)

YOTO - You Only Train Once: Template-Specific Table Extraction from a Single Example

Gaspar Deloin[1][0009-0007-2449-5385], Aurélie Joseph[1][0000-0002-5499-6355], and Vincent Poulain d'Andecy[1][0009-0008-5515-3561]

Yooz, Immeuble le sequoia, Parc d'Andron, 30470 Aimargues, France
{firstname.lastname}@getyooz.com
https://www.getyooz.com/

Abstract. Extracting product tables from business documents such as invoices is challenging: while a given organisation consistently uses the same invoice layout, layouts differ widely across organisations, and many templates are too rare to appear in any training dataset. We address this in a one-shot setting. Given a single annotated document for a known template, the system must reliably extract product tables from all invoices sharing that layout. Our approach extends the star graph framework of d'Andecy et al. to tabular structures, inferring header positions, column boundaries, and row sequences from a single example without any large-scale training data. We evaluate on a filtered subset of Inv3D (112 invoices, 37 templates) and a private dataset (3 templates, 10 documents). On Inv3D, our best run achieves 97.82% recall and 95.37% precision, within 2 points of a proprietary method that first performs TD and then TSR, trained on private data, while outperforming NuExtract3 (74.48% / 66.33%) in both accuracy and speed. On the private dataset, our method reaches 100% precision and recall. A page is processed in 0.3 seconds, making it 100 times faster than NuExtract3.

Keywords: Table extraction · One-shot learning · Invoice understanding · Document analysis · Template-based extraction

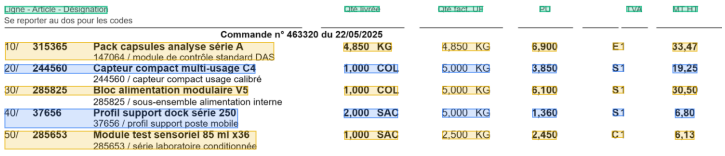
1 Introduction

Extracting structured data from business documents such as invoices is a core task in industrial document processing. A key challenge arises from the diversity of document layouts: each organisation produces its own invoice format, which we refer to as a template. Within a template, all invoices share a consistent fixed layout, but templates vary significantly across organisations. This high variability makes it impractical to collect large annotated datasets for fine-tuning, as new templates may appear at any time with only a few annotated examples available. Moreover, in real-world deployment, document streams are continuously evolving, and cases not represented in the training distribution may frequently occur, limiting the reliability of generic document understanding systems.

We focus specifically on product tables, which are structured as rows and columns where elements in fixed positions share consistent semantics across rows. We restrict our scope to tabular data with a uniform structure, assuming no variation in cell height.

Our goal is to extract tables such as the one shown in Figure 1 in a one-shot setting: given a single annotated document for a given template, the system should reliably extract the table from any other invoice sharing that template. Beyond accuracy, inference speed is a practical requirement for industrial deployment.

We propose a one-shot method that learns to extract product tables from a single annotated example, relying on a simple and efficient structural model. Our method achieves performance competitive with a proprietary TD+TSR model trained on private data, and reaches 100% precision and recall on low-variability templates, while running significantly faster than LLM-based approaches.



Ligne	Article	Désignation	Qté livrée	Qté facturée	PU		TVA	Mt HT
Se reporter au dos pour les codes								
		Commande n° 463320 du 22/05/2025						
10/	315365	Pack capsules analyse série A 147064 / module de contrôle standard DAS	4,850 KG	4,850 KG	6,900		E1	33,47
20/	244560	Capteur compact multi-usage C4 244560 / capteur compact usage calibré	1,000 COL	5,000 KG	3,850		S1	19,25
30/	285825	Bloc alimentation modulaire V5 285825 / sous-ensemble alimentation interne	1,000 COL	5,000 KG	5,100		S1	30,50
40/	37656	Profil support dock série 250 37656 / profil support poste mobile	2,000 SAC	5,000 KG	1,360		S1	6,80
50/	285653	Module test sensoriel 85 ml x36 285653 / série laboratoire conditionnée	1,000 SAC	2,500 KG	2,450		C1	6,13

(a) Extraction by a generic method

Ligne	Article	Désignation	Qté livrée	Qté facturée	PU		TVA	Mt HT
Se reporter au dos pour les codes								
		Commande n° 463320 du 22/05/2025						
10/	315365	Pack capsules analyse série A 147064 / module de contrôle standard DAS	4,850 KG	4,850 KG	6,900		E1	33,47
20/	244560	Capteur compact multi-usage C4 244560 / capteur compact usage calibré	1,000 COL	5,000 KG	3,850		S1	19,25
30/	285825	Bloc alimentation modulaire V5 285825 / sous-ensemble alimentation interne	1,000 COL	5,000 KG	5,100		S1	30,50
40/	37656	Profil support dock série 250 37656 / profil support poste mobile	2,000 SAC	5,000 KG	1,360		S1	6,80

(b) Expected output

Fig. 1: Invoice product table extracted by a generic TD+TSR method (a) compared with expected output (b). Columns *Ligne*, *Article*, and *Désignation* are wrongly merged due to a black line spanning all three headers, and *TVA* is also incorrectly merged with a neighboring column. Row extraction is unstable, with text below rows mistakenly included, highlighting limitations of generic table extraction on invoice layouts.

2 Related Works

2.1 Table Extraction

In recent years, table extraction has significantly improved with the rise of deep learning, particularly transformer-based object detectors such as DETR (End-to-End Object Detection with Transformers) and its table-specific extension Table Transformer (TATR)[10], supported by large-scale datasets PubTables-1M[10] and FinTabNet[11].

There are two main approaches. The first is a two-step pipeline that begins with Table Detection (TD) to locate the table region, followed by Table Structure Recognition (TSR) to analyze its internal layout. The second is a direct end-to-end TSR approach that predicts the table structure without an explicit prior detection stage. TSR methods can be broadly categorized into three main families: detection-based approaches[10]; which model rows, columns, and cells as object detection tasks, sequence-based approaches [7,6,4], which represent structure as token sequences (e.g., HTML, compact encodings, or pointer-based outputs); and regression-based approaches [5], which directly estimate cell coordinates, often enhanced with self-supervised pre-training.

Although these supervised methods are effective, they frequently have limited generalization, particularly when applied to noisy real-world business documents with complex and irregular layouts.

More recently, vision–language models and specialized document LLMs have emerged as an alternative paradigm. They can overcome limitations of purely visual approaches and better handle complex or ambiguous table structures. NuExtract3 [8] is one such model, designed for structured extraction from documents and capable of handling tables in a zero-shot setting.

2.2 Template-based approach for specific fields

A complementary line of work focuses on extracting specific fields from documents with known layouts, rather than parsing arbitrary table structures. Dhakal et al.[2] proposed a one-shot learning method that identifies field coordinates and searches nearby regions for similar patterns, assuming fields appear in consistent locations.

In contrast to the local neighborhood-based approach of Dhakal et al.[2], d'Andecy et al.[1] report over 92% successful extraction per field in a one-shot learning setup, using an incremental, structure-driven model based on a star graph.

While template-based methods effectively extract individual fields and general table extraction approaches perform well on standard benchmarks, no existing work specifically addresses extracting tables with a known template-defined structure.

3 Proposed methods

Our method uses only OCR text and word coordinates for spatial reasoning, bypassing all visual features and pixel-level processing.

3.1 Learn

The learning phase is structured around three steps: (1) selecting and memorizing the header zone, (2) delimiting the columns, and (3) recording one or more rows and formalizing their structure. An additional step automatically derives structural constraints from the annotated examples.

Step 1: Header zone selection This step is directly based on the framework of d'Andecy et al.[1]. In their approach, a star graph is built for each field to extract: the central node represents the target field, while surrounding words with stable relative positions act as satellite nodes. During learning, the model memorizes the spatial relationships between the target field and its satellites. At inference time, these satellites jointly vote for the most likely location of the field.

We apply this mechanism to each table header. We assume that headers within a table are close enough for their surrounding words to serve as reliable satellite nodes, allowing each header position to be memorized and later retrieved.

Step 2: Column delimitation During annotation, the user defines each column according to its estimated width. Column learning is then based on the relative spatial relationships between headers and columns. Since each header is assumed to correspond to a column, the model learns the distances from each header to the boundaries of its associated column.

For columns without headers, their positions are learnt from neighboring columns, using the same principle of relative distance with respect to adjacent columns that do have headers. By default, columns are assumed to extend downward until the end of the document.

Step 3: Row recording and formalization For rows, we learn the most likely sequence of cells using a directed graph. Each cell C_i is connected to every cell C_j, with $i \neq j$, located to its right, and for each pair (C_i, C_j), displacement vectors δ_{ij} are computed. In addition, each cell is horizontally expanded to cover the full width of its corresponding column.

When multiple rows are annotated, the model gains additional information for identifying optional cells. It also learns the expected content type of each cell, for example whether a given position is likely to contain only numerical values (amounts, tax rates, etc.). During inference, this information helps reject implausible row sequences.

Automatic constraint learning From the annotated examples, the model automatically derives structural constraints, including the minimum number of headers required to trigger an extraction and the minimum number of non-optional cells required to validate a row.

3.2 Extraction

Extraction is performed in two stages: first locating the table through header detection, then searching for row sequences within the inferred columns.

Using the star graph mechanism from d'Andecy et al.[1], the model locates each header by having its satellite nodes vote for the most likely position. If

fewer than α headers are detected, the extraction is aborted. Otherwise, column boundaries are inferred from the learnt header-to-boundary distances, and headerless columns are positioned relative to the other detected columns.

With columns established, we select all words W inside each column that satisfy the learnt content type constraints. Each selected word is horizontally expanded to span the full width of its column, and when several expanded words overlap significantly, only one representative is kept. The retained words then cast votes for possible next cells, progressively building row sequences from left to right, as illustrated in Figure 2.

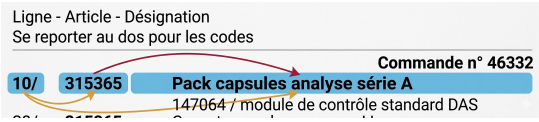


Fig. 2: Example of row sequence voting.

For instance, Figure 2 yields two candidate sequences:

- 10 → 315365 → Pack capsules analyse série A
- 10 → Pack capsules analyse série A

Since the objective is to find the longest valid sequence, the first is retained. A row is only returned if it contains at least θ non-optional cells.

4 Experiments

4.1 Datasets

We used a subset of the Inv3D dataset [3], filtering by font size, font family, template, and margin (rounded to 6 pixels), which resulted in 37 distinct templates. Since the original annotations did not include table headers, all templates were manually re-annotated. We then retained only those templates with at least three images after filtering to ensure a minimum number of examples per layout, yielding a final dataset of 112 invoices across the 37 templates.

We also evaluate on a private test dataset comprising 10 documents across 3 templates, which is incrementally extended as new invoices are collected over time. The dataset includes one template illustrated in Figure 1.

4.2 Metrics

For table structure recognition, we use GRITS[9], and specifically GRITS-CON which overcomes the limitations of previous metrics by comparing cell content between the ground truth and predicted tables. This allows us to evaluate the correctness of the extracted table content directly, independently of structural encoding choices.

4.3 Baselines

We compare our method against two baselines. The first is a proprietary model trained on private invoice data that follows a classical TD+TSR pipeline, representing a strong domain-specific reference. The second is NuExtract3 [8], a large language model, developed by NuMind, used in a zero-shot setting to assess how well a modern vision–language model handles template-specific table extraction without any adaptation.

4.4 Results

Table 1 reports GRITS-CON scores for all methods. The private model serves as a reference for comparison. Our memorization module is evaluated in a one-shot setting, with results shown for the best (most informative annotation) and worst (least informative annotation) runs.

We set $\alpha = 0.5$ (minimum header fraction for extraction) and $\theta = 0.75$ (minimum non-optional cell fraction per row), chosen empirically on a validation set and found to be robust across configurations.

Table 1: GRITS-CON results (%).

	Inv3D		Private dataset	
Method	Recall	Precision	Recall	Precision
Private model (TD+TSR)	99.03	97.36	70.19	84.24
Ours – best run	97.82	95.37	100.00	100.00
Ours – worst run	70.79	90.32	100.00	100.00
NuExtract3	74.48	66.33	47.05	50.13

Our method achieves competitive results with the private model on the best run, with a precision gap of less than 2 points and a recall gap of about 1 point. It is worth noting that these results are particularly meaningful given that our evaluation set was specifically designed to include cases where the standard TD+TSR approach struggles either due to table detection failures or TSR errors. Best runs occur when the annotated document sits at a midpoint among the other invoices of the same template, both in terms of text size and word positions. In such cases, the learnt spatial relationships generalize well to the remaining invoices.

The gap widens on the worst run, mostly on recall. Despite the constraints imposed by the template structure, Inv3D exhibits variations in OCR output and text positioning across invoices of the same template, which the model fails to handle when the annotated document is not representative enough. This contrasts with the private dataset, where each template corresponds to a single supplier: the resulting invoices are much more homogeneous, and our method achieves 100% precision and recall on all 3 private templates.

Figure 3 illustrates this with the worst-performing template from Inv3D, where performance is low regardless of which invoice is chosen for annotation. The two invoices show significant differences in column widths and overall table width, in addition to OCR-level variations in text size, making it impossible for any single annotated document to fully represent the others.

NuExtract3 performs worst on both datasets, showing that zero-shot LLM-based extraction remains unreliable for structured table extraction in business documents. Performance drops further on the private dataset (47.05% recall, 50.13% precision), likely due to its more visually ambiguous layouts.

DESCRIPTION	QTY	UNIT PRICE	TOTAL
Set Of 12 Fork Candles	9	2.89	26.01
Cream Slice Flannel Pink Spot	2	5.79	11.58
Children'S Circus Parade Mug	7	3.29	23.03

(a) Table A

DESCRIPTION	QTY	UNIT PRICE	TOTAL
Victorian glass hanging t-light	9	2.46	22.14
Clear drawer knob acrylic edwardian	3	1.25	3.75
Christmas retrospot star wood	4	1.66	6.64
Doormat welcome puppies	10	7.08	70.80

(b) Table B

Fig. 3: Two tables from the worst-performing Inv3D template, showing variability in column widths and table width.

In terms of inference speed, the private TD+TSR method processes a page in about 1 second on an Intel i7-1370P CPU. On the same CPU, our method is the fastest, processing a page in approximately 0.3 seconds. By comparison, NuExtract3 requires around 30 seconds per page when run on an NVIDIA Quadro RTX 8000 GPU.

5 Conclusion

We presented a one-shot method for template-specific table extraction in business documents. By extending the star graph framework of d'Andecy et al.[1] to tabular structures, our approach learns column boundaries and row sequences from a single annotated document, without requiring a large training dataset. Experiments on Inv3D and a private dataset show that our method reaches competitive accuracy with a proprietary TD+TSR pipeline, achieves 100% precision and recall on low-variability templates, and runs at 0.3 seconds per page, significantly faster than LLM-based approaches.

The main limitation lies in sensitivity to annotation quality: when the annotated document is not representative of the intra-template variability, recall

drops. This is more pronounced on Inv3D, where the same template can exhibit significant differences in text size, word positions, and table width across invoices. Future work will focus on extending the method to tables with variable cell height and growing the private evaluation dataset as new templates are collected. Beyond tables, a natural direction is to handle mixed documents, which combine product tables with key-value fields, requiring the two extraction modes to work jointly. Further extensions include support for spanning cells within tables, and more generally, documents that fall entirely outside the tabular scope considered here.

References

1. D'Andecy, V.P., Hartmann, E., Rusiñol, M.: Field extraction by hybrid incremental and a-priori structural templates. In: 13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018. pp. 251–256. IEEE Computer Society (2018). https://doi.org/10.1109/DAS.2018.29, https://doi.org/10.1109/DAS.2018.29
2. Dhakal, P., Munikar, M., Dahal, B.: One-shot template matching for automatic document data capture. In: Proceedings of Artificial Intelligence for Transforming Business and Society (AITB). vol. 1, pp. 1–6. IEEE (2019)
3. Hertlein, F., Naumann, A., Philipp, P.: Inv3d: a high-resolution 3d invoice dataset for template-guided single-image document unwarping. Int. J. Document Anal. Recognit. **26**(3), 175–186 (2023). https://doi.org/10.1007/S10032-023-00434-X, https://doi.org/10.1007/s10032-023-00434-x
4. Khang, M., Hong, T.: TFLOP: Table structure recognition framework with layout pointer mechanism. In: International Joint Conference on Artificial Intelligence (IJCAI) (2024)
5. Long, R., Xing, H., Yang, Z., Zheng, Q., Yu, Z., Huang, F., Yao, C.: LORE++: Logical location regression network for table structure recognition with pre-training. Pattern Recognition **157**, 110816 (2025)
6. Lysak, M., Nassar, A., Livathinos, N., Auer, C., Staar, P.W.J.: Optimized table tokenization for table structure recognition. In: Document Analysis and Recognition (ICDAR) (2023)
7. Nassar, A., Livathinos, N., Lysak, M., Staar, P.: TableFormer: Table structure understanding with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4614–4623 (2022)
8. NuMind: Nuextract3. Hugging Face model card (2026), https://huggingface.co/numind/NuExtract3, accessed: 2026-06-01
9. Smock, B., Pesala, R., Abraham, R.: Grits: Grid table similarity metric for table structure recognition. ArXiv **abs/2203.12555** (2022), https://api.semanticscholar.org/CorpusID:247618872
10. Smock, B., Pesala, R., Abraham, R.: PubTables-1M: Towards comprehensive table extraction from unstructured documents. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4634–4642 (2022)
11. Smock, B., Pesala, R., Abraham, R.: Aligning benchmark datasets for table structure recognition. In: International Conference on Document Analysis and Recognition (ICDAR). pp. 371–386 (2023)

Benchmarking Vision-Language Models for French PDF-to-Markdown Conversion

Bruno Rigal[1], Victor Dupriez[1], Alexis Mignon[1], Ronan Le Hy[1], and Nicolas Mery[2]

[1] Probayes, La Poste
bruno.rigal@probayes.com
victor.dupriez@probayes.com
alexis.mignon@probayes.com
ronan.lehy@probayes.com
[2] OpenValue, La Poste
nicolas.mery@openvalue.fr

Abstract. We introduce a French-focused benchmark of difficult pages selected via model-disagreement sampling from a corpus of 60,000 documents, covering handwritten forms, complex layouts, dense tables, and graphics-rich pages. Evaluation is performed with unit-test-style checks that target concrete failure modes (text presence, reading order, and local table constraints) combined with category-specific normalization designed to discount presentation-only variance. Across 25 models, we observe substantially higher robustness for the strongest proprietary models on handwriting and forms, while several open-weights systems remain competitive on standard printed layouts.

Keywords: OCR · PDF-to-Markdown · French documents · benchmarking · vision-language models

1 Introduction

The conversion of unstructured PDF documents into structured text formats such as Markdown is a critical step for modern document pipelines, especially for Retrieval-Augmented Generation (RAG). Errors introduced at this stage propagate directly to downstream retrieval, summarization, and grounding.

Existing benchmarks [2,5,3,6] cover document VQA, OCR robustness, and end-to-end parsing (Table 1), but share two key limitations. First, they are mostly focused to English and Chinese, leaving French and other languages under-evaluated. Second, they rely on global edit-distance metrics (character error rate, TEDS) that become unreliable once transcription is mostly correct: residual differences are dominated by linearization and formatting choices rather than recognition errors. This is acute for PDF-to-Markdown, where semantically equivalent outputs can still produce large string distances, and where reading order is intrinsically ambiguous for multi-column and figure-heavy layouts.

The OlmOCR benchmark [7,8] avoids global distances by checking small, machine-checkable facts per page. This unit-test framing is effective for diagnosing structural errors; however, it focuses on English PDFs and remains sensitive to benign formatting differences. Our manual audit shows that for strong models, over 50% of detected failures stem from annotation noise rather than genuine transcription errors [9]. Furthermore, many existing datasets under-represent failure modes critical for real-world pipelines: tiny printed text, long multi-column articles, dense tables, mixed handwritten–printed forms, and graphics-rich layouts. Practitioners lack a benchmark that simultaneously (i) targets French documents, (ii) emphasizes hard real-world layouts and handwriting, and (iii) evaluates outputs aligned with RAG usage where semantic completeness matters more than exact string fidelity.

Table 1: Comparison of existing OCR/document-parsing benchmarks.

Benchmark	Size	Languages	Document Types	Metrics
OCRBench v2	850 pages	EN, ZH	Slides, Papers, Web	Edit Distance
CC-OCR	300 pages	10 langs (incl. FR)	Magazines, Brochures	Edit Distance / BLEU
OmniDocBench 1.5	1355 pages	EN, ZH	Academic, Manuals, HW	Text, TEDS, CDM
OlmOCR Bench	1403 pages	EN, ZH	Legal, Academic, Manuals	Unit Tests

Contributions. We make three contributions: (i) a French-focused benchmark available on HuggingFace of difficult PDF pages selected via model-disagreement sampling, spanning handwriting, forms, complex layouts, and dense tables; (ii) a unit-test-based evaluation suite with category-specific normalization profiles designed to discount presentation-only variance; and (iii) a comparative study of 25 proprietary and open-weights models under a unified conversion pipeline.

2 Methodology

2.1 Dataset construction

The task is to convert PDF page images into Markdown, requiring OCR, layout analysis (reading order, headers, lists), table recognition, and interpretation of non-textual elements such as handwritten fields and scientific graphics. The source corpus consists of approximately 60,000 French documents from CCPDF and Gallica. To ensure the benchmark targets processing limitations rather than trivial cases, we employed an adversarial selection strategy. Documents were transcribed by two different VLMs (dots.ocr and MinerU2.5), and we used the edit distance between their outputs as a proxy for model disagreement. Pages with the largest disagreements were selected as candidates for "difficulty," which intentionally biases the benchmark toward pages where current conversion systems are unstable rather than implying an absolute notion of difficulty.

The resulting dataset is categorized into specific challenge types:

- Long tiny text: Dense pages requiring sustained generation window attention with low-resolution or small font sizes.
- Multi-column: Complex flows requiring correct reading order capabilities.
- Long tables: Data-dense structures spanning large vertical space.
- Manuscript/handwritten: Historical documents and contemporary handwritten notes.
- Forms: Semi-structured documents with handwritten entries. These forms were written manually by a diverse set of participants to ensure variability of handwriting styles.
- Graphics: Scientific figures and charts requiring textual description. While many VLMs were not trained explicitly for this task, we include it to reflect real-world use cases where graphics are often the most information-dense part of a document, and omitting them is operationally equivalent to dropping content. From our experience developing products based on such OCR systems, the lack of high quality graphics parsing is a major bottleneck not addressed by existing OCR benchmarks.

Mathematical equations were explicitly excluded, as they are extensively covered in other benchmarks [6,7].

2.2 Test generation and verification

Tests were generated semi-automatically using category-specific prompts and verified via a human-in-the-loop process. For each selected page, gemini-3-pro-preview was prompted with tailored instructions (e.g., "identify the tiniest text spans", "list handwritten fields", "extract table cells around a given header") and its textual candidates were converted into machine-checkable unit tests.

Per category: tiny text, multi-column, and graphics pages yielded TextPresenceTest items asserting that a short verbatim span appears in the output; handwritten and forms pages used TextPresenceTest with stricter normalization tolerating minor spacing and Unicode differences; long tables produced TableTest instances encoding local structural constraints from manual inspection.

2.3 Test types and normalization

All unit tests inherit from a common class, which is responsible for normalizing both the reference text (from the annotation) and the candidate text (from the model output) before comparison. The normalization is designed to reduce spurious failures caused by harmless formatting differences while still being sensitive to genuine content errors, in the spirit of unit-test-based OCR evaluation [8].

Concretely, normalization includes: (i) Markdown/HTML cleanup (flatten breaks, remove emphasis markers, normalize tables, collapse whitespace); (ii) Unicode harmonization with a curated symbol mapping (quotes, dashes, French guillemets); (iii) optional ASCII projection of accented characters; and (iv) optional alphanumeric filtering, layout-insensitive spacing, and per-test fine-grained masking.

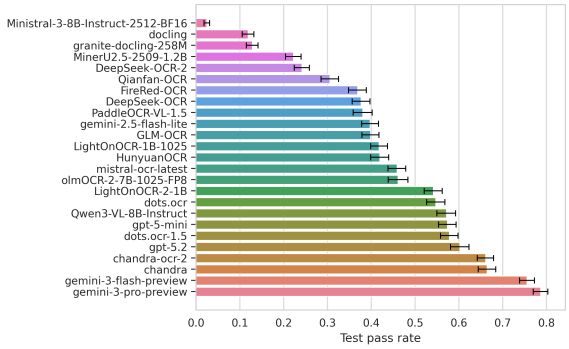


Fig. 1: Overall benchmark results by model (aggregate unit-test pass rate).

By tuning these options per test type and category, we obtain a spectrum of strictness that reduces false negatives from superficial formatting variation while keeping failures interpretable.

3 Evaluation protocol

Experiments were conducted using the `vlmparse` library (`https://github.com/ld-lab-pulsia/vlmparse`), a unified wrapper for heterogeneous VLMs and OCR backends that supports asynchronous processing and Docker-managed local servers. Inference was parallelized across 32 threads on an NVIDIA A100 (80 GB VRAM) with a timeout of 500s; throughput comparisons are reported for local models only.

We evaluated 25 models spanning proprietary APIs and open-weights systems including OlmOCR [7,8], LightOnOCR [10], dots.ocr [4], DeepSeek-OCR [11], and PaddleOCR-VL [1]. The primary metric is the unit-test pass rate (aggregating TextPresenceTest, TextOrderTest, and TableTest), which enables targeted diagnosis of failure modes; benchmark code is available at github.com/ld-lab-pulsia/benchpdf2md.

4 Results

4.1 Overall performance

Gemini 3 Pro Preview achieved the highest average score (0.76), followed by Gemini 3 Flash Preview (0.74); the best open-weights model was Chandra (0.66). Figure 1 and Table 2 summarize results by model and category.

4.2 Category-specific analysis

Performance varied drastically across document categories. Handwritten and forms proved most discriminatory. Gemini 3 Pro retained relative competence

Table 2: Category results by model (mean scores).

	baseline	forms	graphics	handwritten	long_table	multicolumn	tiny_text	Time per page [s]	All categories
gemini-3-pro-preview	0.965	**0.725**	0.773	**0.600**	0.813	**0.867**	**0.831**		**0.786**
gemini-3-flash-preview	0.964	0.684	0.734	0.582	**0.828**	**0.867**	0.804		0.755
datalab-to/chandra	0.996	0.375	0.748	0.212	0.722	0.794	0.765	4.290	0.664
datalab-to/chandra-ocr-2	0.996	0.433	0.768	0.242	0.709	0.758	0.731	2.029	0.661
gpt-5.2	0.998	0.481	0.802	0.206	0.732	0.739	0.535		0.602
kristaller486/dots.ocr-1.5	0.994	0.467	0.278	0.176	0.566	0.836	0.799	2.629	0.578
gpt-5-mini	**1.000**	0.416	**0.816**	0.182	0.660	0.770	0.506		0.574
Qwen3-VL-8B-Instruct	0.913	0.450	0.377	0.218	0.572	0.770	0.754	4.146	0.571
rednote-hilab/dots.ocr	0.988	0.351	0.269	0.079	0.628	0.782	0.765	2.432	0.547
lightonai/LightOnOCR-2-1B	0.990	0.357	0.326	0.127	0.640	0.806	0.671	1.207	0.542
allenai/olmOCR-2-7B-1025-FP8	0.999	0.392	0.357	0.127	0.614	0.764	0.438	1.107	0.461
mistral-ocr-latest	0.993	0.388	0.286	0.170	0.444	0.733	0.600		0.459
tencent/HunyuanOCR	0.978	0.251	0.278	0.036	0.372	0.727	0.671	4.467	0.419
lightonai/LightOnOCR-1B-1025	0.996	0.216	0.297	0.012	0.406	0.673	0.602	1.085	0.418
GLM-OCR	0.951	0.251	0.198	0.036	0.348	0.600	0.743	**0.600**	0.398
gemini-2.5-flash-lite	0.970	0.388	0.422	0.127	0.205	0.588	0.589		0.397
PaddleOCR-VL-1.5	0.961	0.206	0.278	0.012	0.125	0.673	0.714	4.056	0.381
deepseek-ai/DeepSeek-OCR	**1.000**	0.124	0.368	0.012	0.382	0.630	0.506	0.893	0.377
FireRedTeam/FireRed-OCR	0.964	0.213	0.283	0.024	0.133	0.667	0.664	3.241	0.369
baidu/Qianfan-OCR	0.990	0.251	0.283	0.012	0.352	0.558	0.334	5.261	0.306
deepseek-ai/DeepSeek-OCR-2	0.991	0.096	0.278	0.000	0.281	0.382	0.284	1.470	0.242
opendatalab/MinerU2.5-2509-1.2B	0.795	0.100	0.246	0.000	0.093	0.236	0.405	0.898	0.222
ibm-granite/granite-docling-258M	0.877	0.031	0.187	0.006	0.067	0.333	0.180		0.128
docling	0.999	0.031	0.119	0.000	0.195	0.055	0.138		0.119
Ministral-3-8B-Instruct-2512-BF16	0.349	0.031	0.096	0.000	0.017	0.024	0.001		0.025

(0.60 on handwritten, 0.72 on forms), whereas many smaller models failed completely. For example, granite-docling and MinerU2.5 scored near zero on unstructured handwritten text. As expected graphics were not extracted by non-generalist VLMs with the exception of chandra. Multi-column and tables remained challenging but were better solved than handwriting: high-performing models achieved scores above 0.80 on multi-column text, indicating robust reading order detection. Figure 2 provides a category-by-model breakdown.

4.3 Throughput and resolution effects

Inference speed varies inversely with model size and architectural complexity. A clear Pareto tradeoff is observed with Chandra and Chandra-2 models being the strongest, but also slowest on the Pareto front, while LightOnOCR-2 is a good intermediate, reasonably fast and competitive. At the other end, an interesting model is GLM-OCR, which achieves an excellent speed–accuracy trade-off. Its small model size keeps inference cost low, while speculative decoding accelerates generation; crucially in our integration, it processes only non-native text regions rather than the full page image, avoiding redundant computation on digitally born content. Figure 3a summarizes the accuracy–throughput trade-off across models.

5 Discussion

On this benchmark, performance differences concentrate in a small set of operationally relevant failure modes. Handwriting and forms remain the most challenging categories, where many models fail on hard-to-decode historical documents.

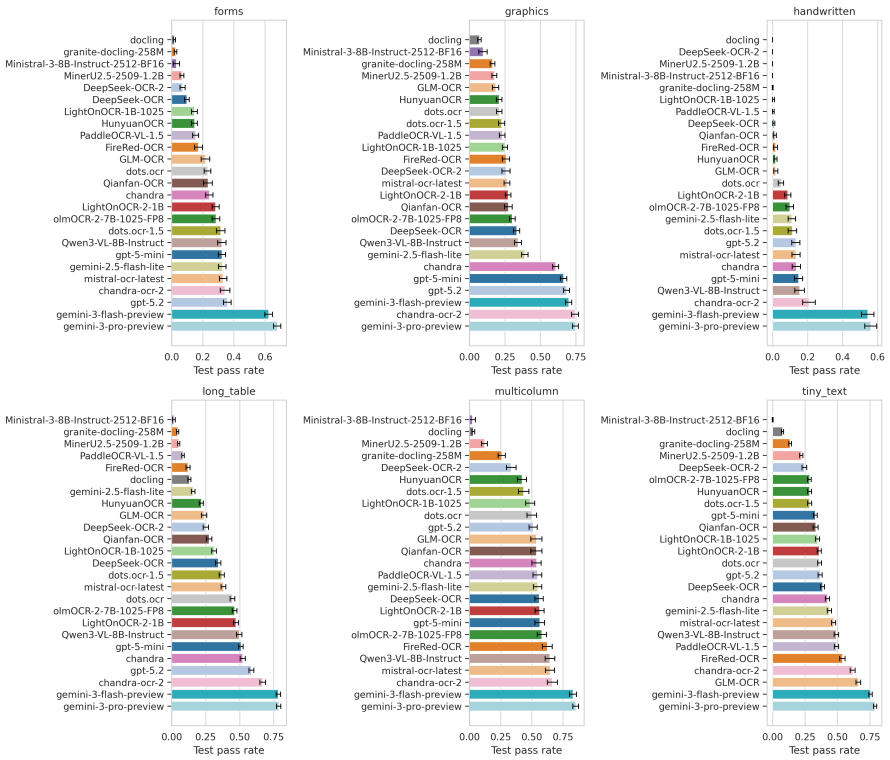


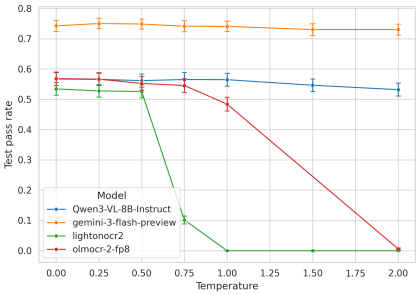
Fig. 2: Per-category performance by model (mean scores).

For long and dense pages, several smaller VLMs exhibit decoding instability—repetition/looping behaviors and incomplete page coverage—reducing unit-test pass rates despite otherwise plausible local OCR. Decoding temperature modulates this trade-off (Figure 3b): smaller open models amplify local errors into repeated spans or structurally inconsistent Markdown at higher randomness, while overly low temperatures risk "stuck" continuations on long pages.

In addition, we observed a systematic failure mode for dots.ocr : incomplete page transcription. A plausible explanation is that the model has been trained with a strong bias to ignore pictures/illustrations; however, in real-world PDFs, "pictures" can be text-dense (e.g., scanned callouts, legends embedded in figures, or rasterized colored blocks). On colorful layouts, this bias can cause dots.ocr to skip entire regions that actually contain relevant printed text, leading to partial Markdown outputs even when the locally transcribed regions look correct. This is an additional reason to avoid transcription without graphics.

Layout-heavy categories (multi-column pages and long tables) primarily expose reading-order and structure-preservation errors. Even when text recognition is strong, small deviations in region traversal or table linearization can break local

(a) Accuracy–throughput trade-off across models.



(b) Effect of decoding temperature on unit-test pass rate.

structural constraints, motivating evaluation with targeted tests that discount presentation-only variance while remaining sensitive to missing or corrupted content.

6 Limitations

This benchmark has several limitations. First, the exclusive focus on French documents limits generalizability to other languages, especially non-Latin scripts. Second, throughput comparisons are hardware-specific and sensitive to quantization choices. Third, PDFs from CCPDF (a Common Crawl subset) may overlap with model pre-training data; contamination cannot be ruled out. The selection procedure emphasizes pages that remain difficult even for the strongest model in our pool (Gemini 3 Pro), but overlap cannot be excluded. However, the forms category (manually written by diverse participants) was not exposed to crawl-based overlap. The alignment of conclusions/model classment between this category and others across the benchmark suggests limited data leakage impact. Finally, model performance may depend on inference implementation details; we welcome corrections and will update the leaderboard accordingly.

7 Conclusion

We presented a French-focused PDF-to-Markdown benchmark combining adversarial page selection, a diverse document taxonomy, and unit-test evaluation with category-specific normalization. Across 25 models, proprietary frontier systems dominate on handwriting and forms while open-weights models remain competitive on standard layouts; graphics parsing is broadly unaddressed. We release the benchmark and evaluation code to support reproducible comparison and community extension.

Acknowledgments. This benchmark used documents provided by Gallica under a restricted use. We also acknowledge the AllenAI OlmOCR-bench, whose public codebase we adapted extensively [7,8].

References

1. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., Zhang, Y., Zhang, Y., Zheng, H., Zhang, J., Zhang, J., Liu, Y., Yu, D., Ma, Y.: Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model (Nov 2025), http://arxiv.org/abs/2510.14528
2. Ding, Y., Luo, S., Dai, Y., Jiang, Y., Li, Z., Martin, G., Peng, Y.: A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends (Jul 2025), http://arxiv.org/abs/2507.09861
3. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., Tang, J., Chen, W., Jin, L., Liu, Y., Bai, X.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning (Jun 2025), http://arxiv.org/abs/2501.00321
4. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots.ocr: Multilingual document layout parsing in a single vision-language model (Dec 2025), http://arxiv.org/abs/2512.02498
5. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: On the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (Dec 2024). https://doi.org/10.1007/s11432-024-4235-6, https://link.springer.com/10.1007/s11432-024-4235-6
6. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24838–24848 (2025), http://openaccess.thecvf.com/content/CVPR2025/html/Ouyang_OmniDocBench_Benchmarking_Diverse_PDF_Document_Parsing_with_Comprehensive_Annotations_CVPR_2025_paper.html
7. Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Rangapur, A., Wilhelm, C., Lo, K., Soldaini, L.: olmocr: Unlocking trillions of tokens in pdfs with vision language models (Jul 2025), http://arxiv.org/abs/2502.18443
8. Poznanski, J., Soldaini, L., Lo, K.: olmocr 2: Unit test rewards for document ocr (Oct 2025), http://arxiv.org/abs/2510.19817
9. Rigal, B., Dupriez, V., Mignon, A., Le Hy, R., Mery, N.: Benchmarking vision-language models for french pdf-to-markdown conversion (Feb 2026), http://arxiv.org/abs/2602.11960
10. Taghadouini, S., Cavaillès, A., Aubertin, B.: Lightonocr: A 1b end-to-end multilingual vision-language model for state-of-the-art ocr (Jan 2026), http://arxiv.org/abs/2601.14251
11. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression (Oct 2025), http://arxiv.org/abs/2510.18234

From Pen Strokes to Sleep States: Detecting Low-Recovery Days Using Sigma-Lognormal Handwriting Features

Chisa Tanaka[1][0009-0004-2647-3284], Andrew Vargo[1][0000-0001-6605-0113], Anna Scius-Bertrand[2][0000-0002-8464-3059], Andreas Fischer[3][0000-0003-0069-3436], and Koichi Kise[1][0000-0001-5779-6968]

[1] Osaka Metropolitan University, Osaka, Japan
sp25241r@st.omu.ac.jp, awv@omu.ac.jp, kise@omu.ac.jp
[2] University of Applied Sciences and Arts Western Switzerland, Fribourg, Switzerland
anna.scius-bertrand@unifr.ch
[3] University of Fribourg, Fribourg, Switzerland
andreas.fischer@unifr.ch

Abstract. While handwriting has traditionally been studied for character recognition and disease classification, its potential to reflect day-to-day physiological fluctuations in healthy individuals remains unexplored. This study examines whether daily variations in sleep-related recovery states can be inferred from online handwriting dynamics. We propose a personalized binary classification framework that detects low-recovery days using features derived from the Sigma-Lognormal model, which captures the neuromotor generation process of pen strokes. In a 28-day in-the-wild study involving 13 university students, handwriting was recorded three times daily, and nocturnal cardiac indicators were measured using a wearable ring. For each participant, the lowest (or highest) quartile of four sleep-related metrics—heart rate variability (HRV), lowest heart rate, average heart rate, and total sleep duration—defined the positive class. Leave-One-Day-Out cross-validation showed that Precision-Recall AUC (PR-AUC) significantly exceeded the chance rate for all four variables after correction for multiple comparisons, with the strongest performance observed for cardiac-related variables. Furthermore, classification performance did not differ significantly across task types or recording timings, demonstrating that subtle within-person autonomic recovery fluctuations can be detected from everyday handwriting.

Keywords: Handwriting analysis · Sleep quality · Sigma-Lognormal model · Binary classification · Heart rate variability

1 Introduction

Online handwriting data acquired via digital pens and tablet devices contain rich information related to the writer's motor control, including not only char-

acter shape information but also writing speed, pen pressure, and inter-stroke time intervals. Recent research has leveraged such kinematic features for ADHD classification [1], early diagnosis of neurodegenerative diseases from handwritten signatures [6], and analysis of motor control changes associated with aging [7], showing that handwriting is a rich source of information reflecting the writer's neurological and cognitive characteristics.

However, these studies primarily target pathological conditions or developmental stages, and estimating day-to-day physical condition among healthy individuals remains unexplored. In this study, we focus on sleep duration and sleep-related cardiac indicators as targets of estimation. Heart rate variability (HRV) and heart rate during sleep (resting heart rate) are objective indicators of autonomic nervous system recovery, and decreases in HRV or increases in resting heart rate are associated with accumulated stress [3]. Furthermore, poor sleep quality has been reported to impair fine motor control [9], suggesting that changes in sleep quality may be reflected in daytime handwriting patterns. Such cardiac indicators are currently measured using wearable devices, but high cost and the burden of continuous wear and maintenance limit broad adoption. If sleep quality could be estimated from everyday handwriting activities, non-invasive sleep monitoring could be provided to a wider audience, such as children who already use tablets at school.

This study proposes a binary classification framework that detects days with poor sleep quality (low-recovery days) from handwriting features based on the Sigma-Lognormal model. The main contributions of this study are as follows:

- We introduce a previously unexplored application domain: estimating sleep quality from handwriting features.
- Using 28-day in-the-wild data from 13 participants evaluated with person-dependent leave-one-day-out cross-validation, we confirmed that classification performance significantly exceeded the chance rate for all four sleep variables (after false discovery rate (FDR) correction).
- No significant differences in classification performance were observed across five handwriting task types or three daily session timings (after FDR correction), showing that the proposed approach is largely robust to task and timing variations within the evaluated conditions.

2 Related Work

Handwriting features extracted from the Sigma-Lognormal model have been used for ADHD classification in children [1], early diagnosis of neurodegenerative diseases from handwritten signatures [6], and analysis of motor control changes associated with aging [7], demonstrating that the model can capture neuromotor changes attributable to diseases or developmental stages. However, these targets are all pathological conditions or long-term developmental changes, and no studies have estimated day-to-day variations in physical condition among healthy individuals from handwriting. Regarding the relationship between sleep and handwriting, Venevtseva et al. [9] showed in a study of university students

Table 1: Five basic features extracted from the Sigma-Lognormal model. The mean and standard deviation of each feature are computed, yielding ten features in total.

Feature	Description	Type
D	Amplitude of motor command	Model parameter
t_0	Onset time of motor command	Model parameter
nblog	Number of lognormal distributions	Reconstruction index
SNR	Signal-to-noise ratio	Reconstruction index
SNR/nblog	Reconstruction accuracy per distribution	Reconstruction index

that poor sleep quality impairs fine motor control. As a directly relevant study, Jasper et al. [2] analyzed basic handwriting kinematics (writing speed, fluency, etc.) under a 40-hour sleep deprivation protocol in a laboratory setting, confirming a circadian rhythm in writing speed while fluency remained unchanged. However, the study did not employ Sigma-Lognormal model parameters, and the relationship with natural variations in sleep quality under in-the-wild conditions has not been examined. This study attempts to classify sleep-related indicators under in-the-wild conditions using finer-grained features based on the Sigma-Lognormal model for the first time.

3 Proposed Method

3.1 Sigma-Lognormal Model

Handwriting is a form of human movement and can be described mathematically by motor theory [8]. The Sigma-Lognormal model [5] describes the velocity profile of a single stroke as a superposition of lognormal distributions, expressed by:

$$\boldsymbol{v}(t) = \sum_{i=1}^{N} D_i \begin{bmatrix} \cos(\theta_i(t)) \\ \sin(\theta_i(t)) \end{bmatrix} \Lambda_i(t; t_{0i}, \mu_i, \sigma_i^2), \quad N \geq 2 \tag{1}$$

where D_i is the amplitude of the motor command at time t_{0i}, and Λ_i is the motor response represented by a lognormal distribution. We extract parameters using the implementation by Lai et al. [4][4] (mean SNR 24.1 ± 4.4 dB), obtaining the five basic features shown in Table 1. Since each handwriting task contains multiple strokes, we compute the mean and standard deviation across the entire task, yielding ten features in total.

3.2 Sleep Indicators

From data recorded by the Oura Ring during sleep, we define four target variables representing sleep quality: total sleep duration (Total Sleep; sleep quantity),

[4] https://github.com/LaiSongxuan/SynSig2Vec

average heart rate variability (Avg HRV; reflecting autonomic recovery), lowest heart rate (Lowest HR; the deepest resting state), and average heart rate (Avg HR; elevated values are associated with accumulated stress [3]).

3.3 Binary Classification Framework

We address a binary classification problem of detecting "days with poor sleep quality." For each target variable, a per-user quartile threshold is set in the direction of poorer sleep quality: for sleep duration and HRV, values below Q25 are defined as the "problematic" (positive) class; for heart rate, values above Q75 are positive. This results in approximately 25% of samples being positive, creating an imbalanced classification problem. A separate binary classifier is trained and evaluated independently for each of the four target variables. As the classifier, we adopt random forest (RF), which captures nonlinear interactions among features and provides stable performance even on small samples.

4 Dataset

The dataset used in this study consists of handwriting data and sleep data, newly collected for the purpose of this research.

Handwriting data were collected using IoT Paper, a tablet device provided for research purposes by Wacom Co., Ltd., with a sampling frequency of 480 Hz, recording online handwriting data including x- and y-coordinates, pen pressure, and tilt in InkML format. Sleep data were measured using Oura Ring (Gen 3 and Gen 4), recording heart rate and HRV during sleep. The validity of Oura Ring has been confirmed through comparison with polysomnography [10]. Measurement data were obtained through the Oura API v2[5].

Seventeen university students (4 female, 13 male) participated in the experiment, all first-time users of the IoT Paper. Handwriting data were collected three times per day (after waking, after lunch, and before bedtime), consisting of shape drawing tasks (circle, triangle, and square, each drawn three times per session) and phrase writing tasks in which participants wrote a positive phrase ("アイスを食べた" (I ate ice cream)) and a negative phrase ("コーヒーこぼした" (I spilled coffee)). For participants who ate only two meals per day, the after-lunch task was performed after their first meal (provided that at least two hours had elapsed since waking). For new Oura Ring users (13 participants), a break-in wearing period of at least one week was established before the experiment; continuous wearing from bedtime to waking was mandatory while daytime wearing was optional.

For each timing independently, the experiment was continued until 28 days of paired data were collected. Of the 17 participants, 4 were excluded from the analysis because they could not collect 28 days of complete data due to frequent missed tasks. Data from the remaining 13 participants (28 days × 3 timings × 5

[5] https://cloud.ouraring.com/v2/docs

tasks = 420 samples per participant) were used (the sleep label is shared within each day).

The target sleep indicators are: total sleep duration (Total Sleep; mean 6.6 h, SD 1.5), average heart rate variability (Avg HRV; mean 56.3 ms, SD 18.7), lowest heart rate (Lowest HR; mean 51.8 bpm, SD 7.2), and average heart rate (Avg HR; mean 57.4 bpm, SD 7.9). Participants are labeled P1–P13 in ascending order of their median Total Sleep. Inter-participant variability is large (e.g., the median Total Sleep ranges from 4.8 to 7.3 hours), and intra-participant variability also differs across individuals.

5 Experimental Results and Discussion

5.1 Evaluation Method

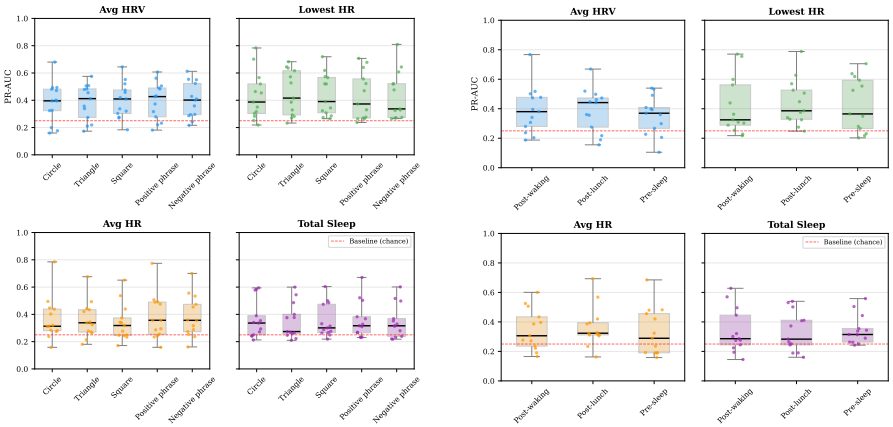
Given the large inter-individual variability observed in sleep indicators, we adopt a person-dependent modeling strategy, training a classifier for each participant. We use Leave-One-Day-Out cross-validation (LODOCV), in which one day's data (three sessions $\times$ five tasks = 15 samples) is held out as the test set and the remaining days' data are used for training. As evaluation metrics, we use PR-AUC (area under the Precision-Recall curve) and Recall@25%, which are suitable for class imbalance. For both metrics, the chance rate (the expected performance of a random classifier) equals the positive class ratio (approximately 0.25); values above 0.25 indicate performance better than chance.

5.2 Experimental Results

The handwriting tasks consist of five types: circle, triangle, square, a positive phrase, and a negative phrase. The sessions were conducted three times per day: after waking, after lunch, and before bedtime.

Table 2 shows the PR-AUC and Recall@25% (mean across all tasks, user mean $\pm$ standard deviation) and the results of the Wilcoxon signed-rank test for each target variable. The Wilcoxon test evaluated whether each user's mean value significantly exceeded the chance rate using a one-sided test. FDR correction (Benjamini-Hochberg method, $q = 0.05$) was applied to the 4 target variables $\times$ 2 metrics = 8 comparisons, and all 8 tests remained significant. This indicates that, under both metrics, sleep-related cardiac indicators and total sleep duration can be identified from handwriting features with accuracy above random chance. Lowest HR showed the highest classification performance (PR-AUC 0.438).

Figures 1 (a) and (b) show the distribution of PR-AUC by task and by timing. Each point represents one user's value, and box plots show the distribution across users; the red dashed line represents the chance rate. For each target variable, we assessed differences across tasks and across timings using Friedman tests (Table 3). After FDR correction (Benjamini-Hochberg method) over the 4 target variables $\times$ 2 metrics $\times$ 2 effects = 16 comparisons, no significant differences were observed for any combination. This suggests that equivalent classification is achievable from any handwriting task and timing.



(a) By task (b) By timing

Fig. 1: Distribution of PR-AUC. Each point represents one user's result. The red dashed line indicates the chance rate.

Table 2: PR-AUC and Recall@25% for each target variable (user mean ± SD) with Wilcoxon test results. p: raw p-value; q: FDR-corrected p-value (Benjamini-Hochberg, 8 comparisons). The chance rate (random classifier) is 0.25 for both metrics. Significance levels: * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$.

	PR-AUC				Recall@25%			
Target variable	mean ± SD	p	q	sig.	mean ± SD	p	q	sig.
Avg HRV	0.391 ± 0.131	0.0023	0.0046	**	0.416 ± 0.139	0.0023	0.0046	**
Lowest HR	0.438 ± 0.164	0.0001	0.0008	***	0.350 ± 0.174	0.0327	0.0374	*
Avg HR	0.365 ± 0.146	0.0031	0.0050	**	0.344 ± 0.185	0.0471	0.0471	*
Total Sleep	0.352 ± 0.125	0.0023	0.0046	**	0.322 ± 0.108	0.0199	0.0265	*

Additionally, large individual differences in PR-AUC across users can be observed in Fig. 1. For example, for Lowest HR, one user achieved a notably high PR-AUC of 0.729, while several users remained near the chance rate. The pattern of individual differences varies by target variable; a user who shows high performance for Lowest HR does not necessarily show high performance for Total Sleep.

5.3 Discussion

The finding that handwriting features achieved significant classification performance for all four sleep variables is consistent with the hypothesis that changes in sleep quality affect daytime fine motor control via the autonomic nervous system. The parameters of the Sigma-Lognormal model reflect neuromuscular

Table 3: Results of Friedman tests. Tests were conducted for task effect (5 levels) and timing effect (3 levels) on PR-AUC and Recall@25% respectively (16 tests in total). q: FDR-corrected p-value (Benjamini-Hochberg, 16 comparisons). None remained significant after correction.

	Task effect						Timing effect					
	PR-AUC			R@25%			PR-AUC			R@25%		
Target	χ^2	p	q	χ^2	p	q	χ^2	p	q	χ^2	p	q
Avg HRV	2.40	0.663	0.847	2.26	0.688	0.847	0.46	0.794	0.907	6.84	0.033	0.344
Lowest HR	6.09	0.192	0.680	3.31	0.507	0.847	4.31	0.116	0.619	1.10	0.576	0.847
Avg HR	9.85	0.043	0.344	0.40	0.982	0.982	2.92	0.232	0.680	0.91	0.633	0.847
Total Sleep	2.65	0.619	0.847	0.82	0.935	0.982	1.08	0.584	0.847	2.73	0.255	0.680

control characteristics [8], and changes in recovery state due to sleep may be reflected in these parameters. Unlike prior work that relied on basic kinematics such as writing speed and fluency [2], the Sigma-Lognormal model may capture subtler motor variations that reflect daily recovery fluctuations. Cardiac-related variables (Lowest HR: PR-AUC 0.438, Avg HRV: 0.391, Avg HR: 0.365) showed higher classification performance than Total Sleep (0.352), likely because heart rate and HRV directly reflect the autonomic recovery state [9], whereas total sleep duration is an indicator of sleep "quantity" with a more indirect correspondence to autonomic recovery.

The absence of significant task and timing effects after FDR correction suggests that the influence of sleep is not limited to specific motor patterns but permeates handwriting motor control in general. Since the Sigma-Lognormal model captures dynamic properties of movement rather than character shape, prediction from everyday handwriting activities is feasible without requiring users to perform specific tasks.

6 Conclusion

Using Sigma-Lognormal model-based handwriting features, we evaluated a binary classification framework for detecting low-recovery days with 28-day in-the-wild data from 13 participants. LODOCV evaluation confirmed that PR-AUC significantly exceeded the chance rate for all four sleep variables (highest: Lowest HR 0.438, all tests remaining significant after FDR correction), and no significant differences were observed across task types or session timings, demonstrating the robustness of the proposed approach. These results constitute the first empirical evidence that handwriting features can be used to estimate daily health fluctuations in healthy individuals, suggesting the feasibility of non-invasive sleep monitoring using tablet devices. Future work includes investigating user characteristics associated with the large individual differences in classification performance, expanding the number and diversity of participants, and validating

applicability to natural handwriting activities (e.g., note-taking) in educational settings.

Acknowledgments. The authors would like to express their sincere gratitude to Mr. Toshihiko Horie and Mr. Takahiro Yamamoto of Wacom Co., Ltd. for lending the Wacom IoT Paper used in this study. This research was supported by the JST research program CRONOS (Grant No. JPMJCS24K4).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Faci, N., Nguyen, H., Laniel, P., Gauthier, B., Beauchamp, M., Nakagawa, M., Plamondon, R.: Classifying the Kinematics of Fast Pen Strokes in Children with ADHD using Different Machine Learning Models, chap. 5, pp. 117–142. World Scientific (2021). https://doi.org/10.1142/9789811226830_0005
2. Jasper, I., Häußler, A., Marquardt, C., Hermsdörfer, J.: Circadian rhythm in handwriting. Journal of Sleep Research **18**(2), 264–271 (2009). https://doi.org/10.1111/j.1365-2869.2008.00706.x
3. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y., Koo, B.H.: Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investigation **15**(3), 235–245 (2018). https://doi.org/10.30773/pi.2017.08.17
4. Lai, S., Jin, L., Zhu, Y., Li, Z., Lin, L.: SynSig2Vec: Forgery-free learning of dynamic signature representations by sigma lognormal-based synthesis and 1D CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6472–6485 (2022). https://doi.org/10.1109/TPAMI.2021.3087619
5. O'Reilly, C., Plamondon, R.: Development of a sigma–lognormal representation for on-line signatures. Pattern Recognition **42**(12), 3324–3337 (2009). https://doi.org/10.1016/j.patcog.2008.10.017
6. Pirlo, G., Diaz, M., Ferrer, M., Impedovo, D., Occhionero, F., Zurlo, U.: Early diagnosis of neurodegenerative diseases by handwritten signature analysis. In: New Trends in Image Analysis and Processing – ICIAP 2015 Workshops, Lecture Notes in Computer Science, vol. 9281, pp. 290–297. Springer (2015). https://doi.org/10.1007/978-3-319-23222-5_36
7. Plamondon, R., O'Reilly, C., Rémi, C., Duval, T.: The lognormal handwriter: learning, performing, and declining. Frontiers in Psychology **4**, 945 (2013). https://doi.org/10.3389/fpsyg.2013.00945
8. Plamondon, R.: The Lognormality Principle: A Personalized Survey, chap. 1, pp. 1–39. World Scientific (2021). https://doi.org/10.1142/9789811226830_0001
9. Venevtseva, Y., Melnikov, A., Nesterova, S.: Sleep and fine motor skills: the influence of sex and the level of physical activity. Neuroscience and Behavioral Physiology **54**, 961–966 (2024). https://doi.org/10.1007/s11055-024-01697-5
10. de Zambotti, M., Rosas, L., Colrain, I., Baker, F.: The sleep of the ring: Comparison of the Ōura sleep tracker against polysomnography. Behavioral Sleep Medicine **17**(2), 124–136 (2019). https://doi.org/10.1080/15402002.2017.1300587

Reconstructing Historical Manuscripts through MSI: The Potential of Contrast in Assessing Image Quality and Legibility

Anna Breger[1] [0000−0001−8878−5743]

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
{ab2864}@cam.ac.uk

Abstract. Digital restoration of historical manuscript images aims to improve readability while preserving the authenticity of cultural heritage documents. However, evaluating quality of restored manuscripts remains challenging, where readability is often subjective and expert annotations are scarce. This study investigates the suitability of contrast-based image quality measures to assess quality and legibility of reconstructed manuscript images from multi-spectral imaging. Two experiments were conducted with publicly-available data sets, facilitating manual quality scores by experts and full-reference image quality measures as reference evaluations. The results show that potential contrast achieves the highest correlation with expert ratings, while contrast-to-noise ratio demonstrates the strongest agreement with full-reference quality measures. Overall, contrast-based measures consistently outperform general image quality measures, demonstrating their potential as objective indicators of manuscript legibility and reconstruction quality.

Keywords: Historical Manuscript Data · IQA · Potential Contrast

1 Introduction

Digital restoration of historical manuscripts remains a central topic in document image analysis due to severe degradations that affect many cultural heritage collections. Common degradations such as ink fading, bleed-through and paper aging significantly reduce legibility and hinder both human interpretation and downstream processing tasks such as transcription and recognition. In recent years, substantial progress has been made with data-driven and deep learning based approaches for document analysis and enhancement [14] and have also improved readability of hand-written texts through task-aware objectives, see e.g. [13]. Moreover, multi-spectral imaging (MSI) has emerged as a powerful non-invasive tool for the analysis and restoration of historical manuscripts by acquiring images across ultraviolet, visible, and near-infrared wavelengths. Together with advanced post-processing, MSI has demonstrated its power in revealing information that has otherwise not been visible, see e.g. [8],[12],[9].

Automated evaluation of text legibility remains a challenging and largely open problem [6], although highly needed for task-driven restoration frameworks

and larger amounts of data. Common image quality assessment (IQA) measures are primarily designed for natural images and often fail to capture legibility in degraded documents. In order to understand suitability of IQA measures dedicated evaluation protocols and comprehensive data sets would be needed for existing as well as newly developed approaches. However, available datasets that enable the assessment of IQA methods for historical manuscript legibility remain very limited. Some data sets have been developed, but never published, e.g. [19], others had been made available in the past but disappeared over time, e.g. [20]. The few available data sets include the subjective assessment framework SALAMI, that provides expert-based legibility annotations for manuscript images [5], as well as the Parchment data set, that includes artificially degraded parchment patches and their untreated references, in its original form [11] and modified [4]. Both data sets are based on MSI of the degraded manuscripts.

Potential Contrast (PC) has been proposed in [21] as a task-dependent IQA measure that estimates the maximum achievable contrast between pixel classes under arbitrary intensity transformations. Unlike traditional contrast measures, it explicitly accounts for the possibility that relevant information may become visible only after suitable grayscale transformations, making it particularly relevant for cultural heritage applications incorporating the possibility that the user might adjust the contrast in the viewing process. Its formulation allows user-guidance through annotated pixel sample sets and has been shown to be effective in identifying informative structures in degraded document images, e.g. [10]. Recently, it has been modified to be independent from the underlying data type, cf. [17], and throughout our experiments we will employ this Normalized Potential Contrast (NPC).

Despite its promising properties, the suitability of (N)PC as an IQA measure for historical manuscript legibility has not yet been systematically investigated. In particular, its relationship to text legibility and its behavior compared to other IQA measures when judging manuscript enhancement quality remains unclear. Given the increasing reliance on objective measures for evaluating restoration pipelines, we aim to provide first insights here. We design 2 experiments with the described data sets, analyzing the rank correlation of NPC values and expert ratings as well as reference-based quality evaluation across common degradations such as ink fading and bleed-through. For comparison we also report the contrast measure contrast-to-noise ratio (CNR), simple image statistics such as the root mean square contrast (RMSC) and entropy, as well as common no-reference (NR) IQA measures BRISQUE [15], NIQE [16] and PIQE [22]. CNR can be understood as a direct complementing contrast measures by using the same sample masks in the computation as NPC. BRISQUE, NIQE and PIQE have been developed for natural images and do not require any reference information.

2 Data and Methods

2.1 Experiment 1: SALAMI Data [3][5]

The SALAMI dataset (Subjective Assessments of Legibility in Ancient Manuscript Images), published in 2020, contains grayscale manuscript images (900×900

pixels) derived from MSI data of 48 historical manuscripts, where 5 output reconstructions of 50 distinct image regions are shared publicly. The resulting 250 images had been annotated region-wise by 20 experts with philology and paleography background regarding legibility, yielding score maps for each image. The data set serves as a benchmark for developing and evaluating quantitative measures of legibility and digital restoration quality in historical manuscript imaging. A first IQA analysis with the data set was published by the data collectors in [6], reporting the Spearman Rank Correlation Coefficients (SRCC) of several quality measures and the mean score maps, including simple image statistics, text detection and NR IQA measures, over several window sizes and dedicated text areas. (N)PC had not been included in the study as it requires defined foreground and background pixels, which also holds true for CNR.

For our complementing experiments we choose from the 50 distinct images the ones that contain writing over the whole image and where at least one of the provided 5 versions of each image allows automated text and background extraction via direct thresholding yielding objective binary annotation masks. This selection process results in 55 test images (11 images with 5 versions each), in which we compute 10000 random patches of the size 200×400 to account for the region-wise score maps, respectively. In Figure 1 we show two examples of random patches and corresponding mean score maps.

2.2 Experiment 2: Parchment Data [4] and Random Reconstructions

The Parchment data set, published in 2019, and available at [2], is a modified subset of the original data set [11] published in 2017. The image data is based on multispectral acquisitions of 18th-century parchment manuscript fragments written with iron-gall ink, imaged both before and after controlled artificial degradation treatments. It comprises 22 parchment patches with different treatment, including untreated control samples, and provides registered multispectral image data with 21 spectral bands from 400 to 950 nm. The degradation procedures simulate common forms of manuscript deterioration, enabling direct evaluation of the degraded images with their intact counterparts serving as a ground truth.

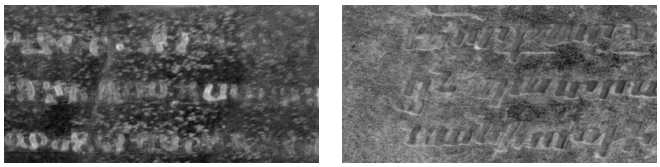
Here, we chose 4 patches (208R, 305R, 309R, 602V) of the provided data set, restricting the experiment to patches with severe degradations/illegible parts after treatment, as well as requiring to have some signal information of the degraded bits in at least one of the 21 MSI bands, i.e. enforcing possible reconstruction of the illegible notation, see examples in Figure 2. Furthermore, we cropped the images to the degraded parts, e.g. if the lower half became illegible after the treatment, but the upper half remained perfectly intact, then only the lower half was evaluated to ensure feasible results representing the degraded regions. Binary masks for text and background were created manually on small regions therein, employing for the annotation process the untreated ground truth representation corresponding to the degraded image parts. For each image 15 minutes were set to create the masks in the software GIMP, mimicking a real research situation with constraint timelines.

Next, we use the MSI data of the treated patches to create with random orthogonal projections a set of 100000 reconstructed grayscale outputs per image to be evaluated for quality beyond the 25 provided outputs per image.

Dimension reduction with orthogonal projections. *Let $x \in \mathbb{R}^d$ be a high-dimensional image. A random lower-dimensional representation is given by $px \in V$, where $V \subset \mathbb{R}^d$ denotes a k-dimensional linear subspace with $k < d$ and $p \in \mathbb{R}^{d \times d}$ an orthogonal projection distributed according to the unique orthogonally invariant probability measure on the Grassmannian. The dimension d is reduced to k in practice via elements $q \in \mathbb{R}^{k \times d}$ of the Stiefel manifold with $q^T q = p$.*

In our case, we do have $d = 21$ MSI bands and reduce the images directly to a grayscale output with $k = 1$ and $m = 720 \times 720$ pixels. Random projections can be computed efficiently via the QR composition, cf. [7], and it has been shown that they give a good representation of the whole space whilst preserving important properties, cf. [1].

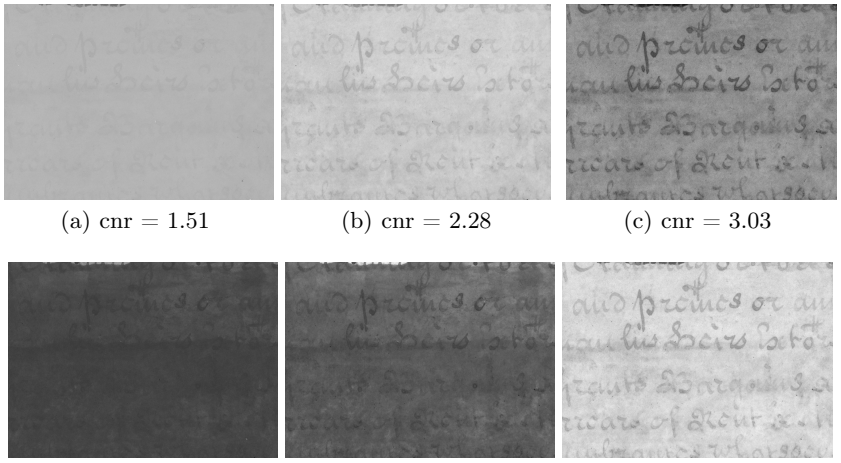
With ground truth images available, we employ 3 full-reference (FR) IQA measures that have shown stable behavior to assess structural information or, in particular, text legibility, as references of quality by comparing the random reconstruction to the untreated ground truth data. Namely, we employ HaarPSI [18] based on Haar wavelets, Pearson correlation [4] and the multi-scale SSIM [23]. As suggested in the original paper [4], we implement the evaluation invariant to polarity. Then, the SRCC is computed between the tested quality measures and these 3 reference measures. As an additional experiment, we compare which MSI band of the 21 available bands is rated highest by the tested quality measures for each of the 4 images. Please find the results in the Appendix.



(a) npc = 0.56 (b) npc = 0.13

(c) mean score map (d) mean score map

Fig. 1. Visual examples of two random patches (a)(b) in the SALAMI data set with corresponding NPC values and mean score maps (c)(d) stored as uint8 images, where white (255) denotes the best and black (0) the worst score. The values of NPC are in [0,1], with 1 denoting the best judgement. The left example shows good matching between the NPC and the expert ratings with an average score over the map of 126.54, the right example shows a pitfall of NPC, judging the quality poorly although it is rated with an average score of 188.95 by the experts.



(a) cnr = 1.51 (b) cnr = 2.28 (c) cnr = 3.03

(d) npc = 0.40 (e) npc = 0.59 (f) npc = 0.77

Fig. 2. Visual examples of random reconstructions of the MSI data of fol.305R. The images are ordered in each row in regards to the quality assessment by CNR (top), NPC (bottom), with increasing quality values from the left to the right, where the right gives the highest quality value, respectively, within the 100000 random reconstructions.

3 Results

In the quantitative results we will state the absolute Spearman Rank Correlation Coefficient (SRCC) of the measures' quality evaluation and the references, indicating how well the tested measures align with the reference judgement. In the qualitative analysis we exemplify results of the highest achieving IQA measures.

3.1 Experiment 1

In Table 1 the correlation results of the IQA measures and mean score maps are shown. The highest result was achieved by NPC with a mean correlation of 0.6784 over the 10000 random patches and 0.6561 for the full image comparison. In Figure 1, we visualize 2 random patches with their NPC score and the corresponding mean score maps. The first example shows a patch where the NPC value and score map align well, whereas in the second example NPC fails.

IQA method	BRISQUE	CNR	ENTROPY	NPC	NIQE	PIQE	RMSC
SRCC $mean_{10k}$	0.0562	0.5332	0.4805	**0.6806**	0.2031	0.1630	0.5896
$variance_{10k}$	0.0021	0.0039	0.0009	0.0005	0.0005	0.0002	0.0012
SRCC *full*	0.1709	0.4627	0.5136	**0.6561**	0.2341	0.1558	0.6150

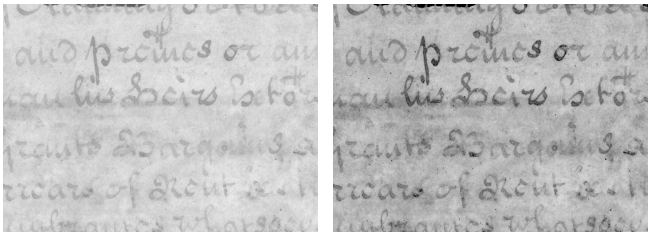
Table 1. Absolute SRCC values for the tested IQA measures and the mean annotation scores over all 55 images with the 10000 random patches of size 200×400 (mean and variance) as well as the full images. The highest correlation value is highlighted in bold.

3.2 Experiment 2

We report in Table 2 the absolute SRCC over 100000 random reconstructions for each of the 4 images between the tested IQA measures and the chosen set of FR-IQA measures, i.e. Haar-PSI, Pearson Correlation and MS-SSIM, serving as a reference judgement set. Moreover, in Figure 2, we show 3 random reconstructions ordered by IQA values from CNR and NPC, the two measures that obtained the highest correlation results. It is important to note that NPC is invariant to linear brightness/contrast changes and we display in Figure 3 a visual result when adjusting the image with intensity transformations.

IQA	BRISQUE	CNR	ENTROPY	NPC	NIQE	PIQE	RMSC
Image 1	.08,.10,.12	**.84,.83,.91**	.64,.55,.47	.76,.75,.86	.16,.27,.24	.01,.23,.24	.01,.03,.04
Mean	0.10	**0.86**	0.55	0.79	0.22	0.16	0.03
Image 2	.13,.18,.18	.84,**.95**,**.94**	.90,.55,.75	**.85**,.94,**.94**	.58,.32,.52	.17,.01,.01	.77,.34,.59
Mean	0.16	**0.91**	0.73	**0.91**	0.47	0.06	0.57
Image 3	.06,.06,.09	**.86**,**.96**,**.88**	.66,.48,.64	.79,.76,.79	.68,.50,.67	.42, .32, .43	.55,.34,.53
Mean	0.07	**0.90**	0.59	.78	0.61	0.39	0.48
Image 4	.11,.09,.00	**.53,.94,.58**	.23,.53,.07	.50,.82,.50	.50,.47,.49	.29,.40,.33	.22,.49,.06
Mean	0.07	**.68**	0.27	0.61	0.49	0.34	0.26

Table 2. Absolute SRCC values of the tested IQA measures and the reference measures HaarPSI, Pearson Correlation and MS-SSIM, respectively as well as the mean, over all 100000 random reconstructions. The highest correlation values are highlighted in bold.



(a) npc = 0.77 (b) npc = 0.77

Fig. 3. NPC is invariant to linear brightness/contrast changes in the image. This property does not hold for the reference measures it has been compared to in Experiment 2, influencing the correlation result.

4 Discussion and Limitations

For Experiment 1 we can observe in Table 1 that NPC clearly yields the highest correlation to the mean score maps for the random patches as well as the full images, with the patch-wise computation yielding a higher result. This behavior is expected, since the mean score maps were created region-wise and the NPC computation over the whole image might not reflect local quality changes. The

low variance confirms stability across the random patches. It is notable that the 3 contrast measures (NPC, RMSC and CNR) yield the 3 best results, indicating that indeed contrast notions can be helpful to identify image restoration quality as well as legibility. Nevertheless, NPC by no means acts perfectly, which is reflected in the overall correlation value of 0.68 and possible mismatches as displayed in Figure 1.

For Experiment 2, which has a more complex design since manual annotations are not available in the employed data set, we can observe in Table 2 that CNR outperforms all tested measures, followed by NPC. This different behavior in comparison to Experiment 1 is not surprising because NPC is invariant to intensity transformations, which aligns well with quality ratings where humans might adjust the intensity when viewing, but does not align necessarily well with FR-IQA measures that are computed in the intensity scaling that is given. In Figure 3 we show the visual difference that can be obtained while NPC remains the same. This property can be useful in practice when it is possible to adjust the brightness/contrast of an image output, e.g. with an image viewer. Figure 2 demonstrates that both, CNR and NPC, produce meaningful orders of image quality.

Limitations of this study include that expert-annotated quality scores were not available for the Parchment dataset; therefore, 3 FR-IQA measures were used as references instead. While these measures capture meaningful aspects of image quality, they do not necessarily represent human expert judgment. In addition, NPC and CNR depend on manually created masks, introducing subjectivity into the evaluation. Finally, larger and more diverse datasets are needed to improve robustness and generalizability and to capture a wider range of manuscript degradations. Further research shall study the behavior of recent deep-learning based and other document-specific image quality measures and include different image transformations to better assess their suitability for legibility evaluation in historical manuscripts. Note that Experiment 2 adds to the study [6], where results with complementing image quality measures may be found.

5 Summary

In conclusion, the results indicate that contrast-based measures are valuable tools for assessing image restoration quality and legibility in historical manuscripts, tested here on images derived from MSI data. NPC showed the strongest performance when compared to human-derived quality maps, while CNR achieved the best results against full-reference image quality measures. Although both measures demonstrate great potential for legibility evaluation, limitations in the study are given. Future research shall explore additional quality measures and larger datasets to improve the reliability and generalizability of the study.

Acknowledgments. A.B. acknowledges funding by the Cambridge Centre for Data-Driven Discovery and Accelerate Programme (grant LEAH/011) through a donation by Schmidt Sciences.
Disclosure of Interests. The author has no competing interests to declare.

References

1. Breger, A., et al.: On orthogonal projections for dimension reduction. Journal of Mathematical Imaging and Vision **62**(3) (2020)
2. Brenner, S.: On the use of artificially degraded manuscripts for quality assessment of readability enhancement methods – dataset & code (2019). `https://doi.org/10.5281/zenodo.2650152`
3. Brenner, S.: SALAMI - Subjective Assessments of Legibility in Ancient Manuscript Images (2020). `https://doi.org/10.5281/zenodo.4270352`
4. Brenner, S., Sablatnig, R.: On the use of artificially degraded manuscripts for quality assessment of readability enhancement methods. In: Proceedings of the ARW & OAGM Workshop 2019. Verlag der Technischen Universität Graz (2019)
5. Brenner, S., Sablatnig, R.: Subjective assessments of legibility in ancient manuscript images - the salami dataset. In: Pattern Recognition. ICPR International Workshops and Challenges (2021)
6. Brenner, S., et al.: Estimating human legibility in historic manuscript images - a baseline. In: Document Analysis and Recognition - Proceedings of ICDAR (2021)
7. Chikuse, Y.: Statistics on Special Manifolds, Lecture Notes in Statistics, vol. 174. Springer-Verlag, New York, NY (2003)
8. Duffy, C.: Multi-spectral imaging at the british library. In: 3rd Digital Heritage International Congress. pp. 1–4 (2018)
9. Easton, R.L., et al.: Standardized system for multispectral imaging of palimpsests. vol. 7531. International Society for Optics and Photonics, SPIE (2010)
10. Faigenbaum-Golovin, S., et al.: Multispectral imaging reveals biblical-period inscription unnoticed for half a century. PLOS ONE **12**(6), 1–10 (2017)
11. Giacometti, A., et al.: Evaluating multispectral image processing methods for the analysis of primary historical texts. Digital Scholarship in the Humanities **32** (2017)
12. Janke, A., et al.: A second look at multispectral data of late medieval music manuscripts. Manuscript Studies **9**(1), 90–117 (2024)
13. Khamekhem Jemni, S., et al.: Enhance to read better: A multi-task adversarial network for handwritten document image enhancement. Pattern Recognition (2022)
14. Lombardi, F., Marinai, S.: Deep learning for historical document analysis and recognition-a survey. J Imaging **6**(10) (2020)
15. Mittal, A., et al.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
16. Mittal, A., et al.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013)
17. Peaslee, W., et al.: Potential contrast: Properties, equivalences, and generalization to multiple classes. In: Proceedings of 33rd EUSIPCO Conference (2025)
18. Reisenhofer, R., et al.: A Haar wavelet-based perceptual similarity index for image quality assessment. Signal Processing: Image Communication **61** (2018)
19. Shahkolaei, A., et al.: Mhdid: A multi-distortion historical document image database. In: IEEE Workshop ASAR (2018)
20. Shahkolaei, A., et al.: Subjective and objective quality assessment of degraded document images. Journal of Cultural Heritage **30**, 199–209 (2018)
21. Shaus, A., et al.: Potential contrast - a new image quality measure. In: Proc. IS&T Intl. Symp. on Electronic Imaging. pp. 52–58 (2017)
22. Venkatanath, N., et al.: Blind image quality evaluation using perception based features. In: International Conference on SPC. IEEE (2015)
23. Wang, Z., et al.: Multi-scale structural similarity for image quality assessment. In: The 37th Asilomar Conference on Signals, Systems & Computers. IEEE (2003)

DAS 2026

WORKSHOP PROCEEDINGS
SHORT AND DEMO PAPERS

SEVENTEENTH INTERNATIONAL WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS

WORKSHOP VENUE

Vienna, Austria

3 – 4 September 2026

Satellite workshop of ICDAR 2026, the International Conference on Document Analysis and Recognition

LET'S GET IN TOUCH

EMAIL
das-2026@googlegroups.com

WEB
das2026.seecs.edu.pk

LINKEDIN
linkedin.com/company/102434587

X
@IAPR_TC11 · **@IAPR_TC12**

IAPR-sponsored workshop
Satellite workshop of ICDAR 2026

SEE YOU IN VIENNA

das2026.seecs.edu.pk