

BUILDING COMMONS FOR CLINICAL LLMS CONFERENCE

Paris, October 1 & 2, 2026

CONFERENCE PROCEEDINGS



Organized by:



With the support of:



TABLE OF CONTENTS

1	Organizers' Foreword	3
	1.1. The Global Imperative for Clinical AI Commons	3
	1.2. Conference key figures	4
	1.3. Conference Scientific Committee	4
	1.4. Conference Program	5
2	Keynotes	7
	2.1. Two Perspectives on Clinical LLMs: Source Fidelity and Cultural Adaptation	7
	2.2. MeditronFO: A Fully Open Pipeline to Create Medical Expert LLMs	8
	2.3. Guideline for Responsible Development, Validation, Implementation, and Monitoring of Generative Artificial Intelligence in Healthcare: The GENAISIS Framework and Guideline	9
	2.4. Building Ecosystem to Develop Medical LLM in Native Language	10
	2.5. From Answers to Action: Building the Evaluation Stack for Healthcare AI	11
	2.6. PARTAGES : Advanced Development of Digital Commons for Generative Artificial Intelligence in Health	12
3	Poster Abstracts	13
4	Panel Discussions	17
	4.1. Panel Discussion 1 - Building Commons for Clinical LLMs: Models, Training Dataset, Cross-Lingual Challenges, and Governance	17
	4.2. Panel Discussion 2 - Deploying LLM-Based Solutions into Clinical Practice: Challenges, Governance & Perspectives	19
5	Workshop Outcomes	21
	5.2. Build Together Workshop	21
	5.3. Evaluate Together Workshop	22
	5.4. Deploy Together Workshop	22
6	Looking Ahead	23

1. Organizers' Foreword

1.1. The Global Imperative for Clinical AI Commons

While Large Language Models (LLMs) offer transformative potential for healthcare professionals, their development is subject to regulatory frameworks governing clinical data access. Beyond this hurdle, most generic models lack the domain expertise required for clinical accuracy, a limitation that is even more pronounced in environments where languages other than English are spoken. Furthermore, given the non-deterministic nature of LLMs, their use in a medical context presents inherent patient safety risks. This entails high precision requirements and requires establishing standardized clinical benchmarks to evaluate model performance across diverse medical specialties.

Currently, we are facing a fragmented landscape of local initiatives where multiple projects are developed in isolation, with limited sharing and reuse. This silo situation is primarily driven by two major challenges:

- **Heterogeneous healthcare settings, there are different clinical practices, patient populations, and data and information systems across various institutions and countries.**
- **Sensitive data and regulatory constraints, clinical LLMs trained on sensitive data may themselves carry privacy risks, limiting their reuse and sharing.**

Because of these regulations, researchers and hospitals often find themselves working in silos, unable to centralize data or easily share their results.

About a year and a half ago, during the AI Summit in Paris, it became clear that the obstacles we face in France are shared by many teams and countries around the world. We realized that we need to join forces to build commons and open resources, enabling collaboration, reuse and scale. This realization is the genesis of the "Building Commons for Clinical LLMs" conference.

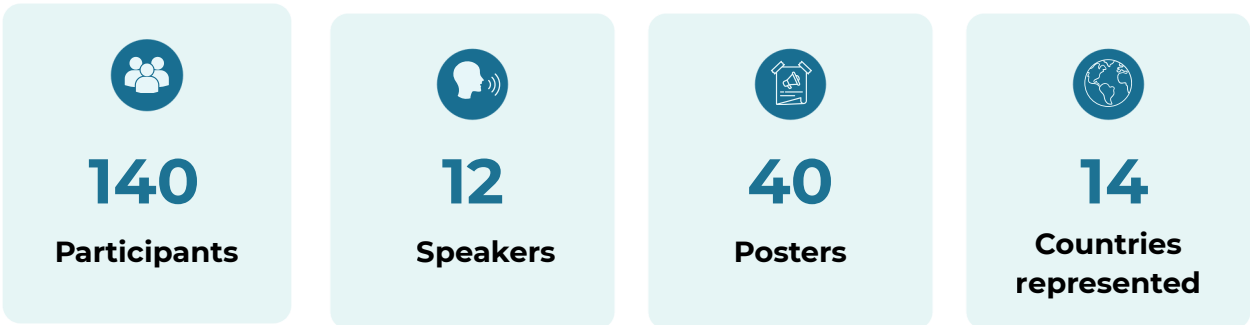
The objectives of this event were to:

- **Present recent research and findings on LLMs in healthcare, with a particular focus on open-source and collaborative approaches.**
- **Foster discussion and collaboration on the major regulatory, scientific, technical, and ethical challenges associated with LLMs in healthcare.**
- **Open a dialogue with industry stakeholders to compare research findings with real-world applications, limitations, and the conditions for deployment.**
- **Address the associated regulatory and ethical challenges, in relation to existing and emerging frameworks, and their impact.**

The French Health Data Hub is proud to have organized this event alongside LISN (Laboratoire Interdisciplinaire des Sciences du Numérique), Limics (Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé), and LIG (Laboratoire d'informatique de Grenoble), with the support of AI for Île-de-France. We hope these proceedings serve as a foundation for continued international collaboration and the development of shared clinical digital commons.

1.2. Conference key figures

We are very proud to have brought together a diverse and international community to share their projects and thoughts for the development of shared resources to support AI in healthcare.



1.3. Conference Scientific Committee



Emmanuel Bacry
Scientific Director, French Health Data Hub, CNRS, Université Paris Dauphine - PSL, PR[AI]RIE - PSAI



Judith Abécassis
Research scientist in the Soda team, INRIA, Saclay



Rémi Flictoeaux
AP-HP, Department of medical information



Chuyuan Li
CNRS researcher at LIG, Université Grenoble Alpes



Aurélie Névéol
CNRS Research Director, Head of STL department, LISN, CNRS, Université Paris-Saclay



François Portet
Profesor, LIG, CNRS, Université Grenoble Alpes



Xavier Tannier
Profesor, Limics, Sorbonne Université



Pierre Zweigenbaum
Emeritus CNRS Senior Researcher, LISN, CNRS, Université Paris Saclay



1.4. Conference Program

Day 1 - Keynotes, Panels & Posters (October 1, 2026)

 Time	 Session & Topics	 Speaker(s) / Details
09:15	Keynote: Two Perspectives on Clinical LLMs: Source Fidelity and Cultural Adaptation	Julia Ive (University College London, UK)
09:45	Keynote: MeditronFO: a Fully Open Pipeline to Create Medical Expert LLMs	Lars Klein (EPFL, Switzerland)
10:15	Panel Discussion: Building Commons for Clinical LLMs: Models, Training Dataset, Cross-Lingual Challenges, and Governance	Panelists & Moderator
10h45	 Break	
11h45	Keynote: Guideline for Responsible Development, Validation, Implementation, and Monitoring of Generative Artificial Intelligence in Healthcare: The GENAISIS Framework and Guideline	Carl Moons (UMC Utrecht, Netherlands)
12h15	 Lunch & Poster Session	
14h	Keynote: Building Ecosystem to Develop Medical LLM in Native Language	Jun Seita (RIKEN, Japan)
14h30	Keynote: From Answers to Action: Building the Evaluation Stack for Healthcare AI	Suhana Bedi (Stanford Medicine, USA)
15h	Panel Discussion: Deploying LLM-Based Solutions into Clinical Practice: Challenges, Governance & Perspectives	Panelists & Moderator
16h	 Break	
16h30	Keynote: PARTAGES: Advanced Development of Digital Commons for Generative Artificial Intelligence in Health	Aurélie Névéol (LISN, France)
17h	Cocktail & Poster Session	

Day 2 - Collaborative Workshops (October 2, 2026)

 Time	 Session & Topics	 Speaker(s) / Details
09:00	Welcome and Initial Briefing	
09:10	3 Simultaneous Collaborative Workshops	• Build Together: Developing shared, multilingual, and reusable clinical LLM resources • Evaluate Together : Designing robust and clinically relevant evaluation approaches • Deploy Together : Addressing deployment, integration, and monitoring challenges in real-world healthcare settings
11:10	 Break	
11:35	Sharing Workshop Results and Feedback	
12:25	Conference Closing	

2. Keynotes

This section brings together the keynote addresses delivered by international researchers working on clinical AI and Natural Language Processing. The talks explore key dimensions of clinical LLMs, including open model architectures, safety and evaluation frameworks, domain adaptation, and real-world deployment challenges.

2.1. Two Perspectives on Clinical LLMs: Source Fidelity and Cultural Adaptation



Julia Ive

Associate Professor in Healthcare AI at University College London, United Kingdom.



Abstract

This presentation brought together findings from recent research on two complementary aspects of clinical LLM use: preserving clinical information through direct extraction and copying, and adapting LLM-generated communication to different cultural contexts. Drawing on the SafeLLM pilot, the session demonstrated how retrieving relevant lines from approved UCL hospital protocols can offer a safer mode of using LLMs. Additionally, the keynote presented research on culturally adapted, LLM-assisted mental-health assistance, examining how AI can help account for cultural, religious, and socioeconomic contexts.



About the Speaker

Dr. Julia Ive is an Associate Professor in Healthcare AI at University College London and works closely with University College London Hospitals (UCLH) on applied clinical research. She has previously held academic and research positions at Queen Mary University of London and Imperial College London. Her research focuses on responsible, data-centric and human-centered applications of large language models (LLMs) in healthcare, with particular emphasis on clinical decision support, clinical and administrative question-answering, cultural adaptation of LLMs for psychotherapy support, and synthetic data generation for privacy preservation and bias mitigation. Her broader expertise includes multimodal machine learning and reinforcement learning for text generation.



[View the Presentation Slides](#)

2.2. MeditronFO : A Fully Open Pipeline to Create Medical Expert LLMs



Lars Klein

*Academic Director at LiGHT,
EPFL Switzerland*



Abstract

Most medical LLMs described as "open" only release weights and withhold everything that explains them: data provenance, curation rules, and synthetic generation pipelines. Independent audit is therefore impossible, a fundamental problem in high-stakes domains such as medicine.

To address this, this keynote introduced Fully Open Meditron (MeditronFO), a pipeline in which every stage is released. Its corpus combines eight public medical QA datasets with three clinician-vetted synthetic extensions, including QA grounded in 46,469 clinical practice guidelines. MeditronFO is not just a single model, but a mechanism to turn models into medical experts. Applied to five fully open base models, the recipe yields the first fully open medical specialists: Apertus-70B-MeditronFO gains 6.6 points over its base on aggregate medical benchmarks, and Gemma-3-27B-MeditronFO is preferred over MedGemma in 58.6% of pairwise clinical comparisons despite training only on releasable data.

Finally, the session discussed the requirements of open evaluation, highlighting Auto-MOOVE, an LLM judge calibrated against 204 human raters.



About the Speaker

As Academic Director of LiGHT (Laboratory for Intelligent Global Health and Humanitarian Response Technology), Lars leads the design and deployment of large-scale multimodal AI systems for clinical and humanitarian settings. His work bridges the gap between foundational machine learning and real-world healthcare, prioritizing reliability, efficiency, and human-centered design. By combining theoretical rigor with systems-level optimization, he drives research initiatives focused on:

- Foundation and multimodal models for health: Designing modular, multimodal, and multilingual AI systems; building foundation LLMs for global health applications. Structured and agentic reasoning: Developing representations, abstractions, and coordination mechanisms that enable reliable reasoning and task decomposition in complex, safety-critical settings, including clinical workflows.
- Efficient and reproducible inference: Creating algorithmic and systems-level approaches for scalable, cost-effective, and reproducible inference, including caching, modular architectures, and inference-time optimization.
- Safety, robustness, and evaluation: Analyzing model performance and failure modes of agentic systems, and developing principled evaluation frameworks with particular attention to safety and reliability requirements in healthcare.



[View the Presentation Slides](#)

2.3. Guideline for Responsible Development, Validation, Implementation, and Monitoring of Generative Artificial Intelligence in Healthcare: The GENAISIS Framework and Guideline



Carl Moons

*Professor of Clinical Epidemiology, Julius Center of Health Sciences and Primary Care, UMC Utrecht
UMC Utrecht's figurehead for AI
Netherlands*



Abstract

Generative AI (GenAI) and its widescale applications in healthcare have transformative potential to enhance care delivery and patient experience. However, safely and responsibly integrating GenAI-based applications remains a complex challenge for the diverse stakeholders involved. Within the high-stakes setting of healthcare, existing AI frameworks and guidelines do not fully capture the unique challenges across the GenAI lifecycle, from initial problem scoping to real-world application use, nor do they adequately address the methodological considerations that are specific to the GenAI paradigm. To address these gaps, this keynote presented the GENAISIS (GENerative AI Evaluation in Healthcare Systems and Implementation Standard) Framework and Guideline, a first comprehensive quality framework for the responsible development, validation, evaluation, implementation, and monitoring of GenAI applications in healthcare.



About the Speaker

Carl Moons is full professor and former director of research at the Julius Center, UMC Utrecht. He is Director Artificial Intelligence (AI) of the UMC Utrecht, developed the AI strategy of the UMC Utrecht, establishing 5 dedicated AI-Health labs (AI Living Technologies and Regenerative Medicine; AI Molecular Medicine; AI Imaging and Imaging guided interventions; AI Methods and Fundamentals; AI Healthy Living and Prevention). He also is the initiator, director and PI of the comprehensive nationwide program (112 Million €) called AI4HEALTH to safely, effectively and efficiently introduce Data&AI-driven innovations in the health and healthcare ecosystem. Carl has directed and written various international quality guidelines for the introduction of safe and effective prediction algorithms, medical technologies, and AI into Healthcare, which have been adopted in various countries. He leads large international consortia which develop guidelines for safety, trustworthiness and explainability of AI in Healthcare. Carl is co-initiator and co-director of the Health Innovation Netherlands (HI-NL) infrastructure. HINL is a unique nationwide infrastructure aiming to bring new biomedical technologies as soon as possible to the market, to ensure a sustainable, safe, effective, affordable and profitable MedTech and Biotech ecosystem. These biomedical innovations range from robotics, medical diagnostic and screening tests to biomarkers, decision aids, AI algorithms and eHealth applications.



[View the Presentation Slides](#)

2.4. Building Ecosystem to Develop Medical LLM in Native Language



Jun Seita

*Team Director, Medical Data & Deep Learning Team,
RIKEN Japan*



Abstract

LLMs used in clinical settings must be able to exchange accurate information in the patients' native language and comply with the various laws, regulations, and guidelines that differ from country to country. To develop an LLM that meets these conditions, it is necessary not only to prepare high quality training data but also to build an ecosystem that includes performance evaluation and safety assessment. In Japan, from 2024 to 2026, a consortium of more than 10 research institutions and companies worked on the development of a Japanese LLM for medical use and its supporting ecosystem. As a result, we released one model built from scratch and eight models based on continued pre-training, while also developing performance evaluation and safety assessment methods. The need for medical LLMs in the native language is the same in many countries, and it is important to develop the necessary ecosystem while taking a balance between standardization and localization.



About the Speaker

About the Speaker Dr. Jun Seita is Team Director (Principal Investigator) at RIKEN, Japan, where he leads 4 laboratories; Medical Science Deep Learning Team, Laboratory for AI-driven Experimental Medicine, Medical Data Deep Learning Special Team, and Medical Data Sharing Unit. He received his M.D. from the University of Tsukuba in 1996 and after completion of residency in cardiovascular surgery, he studied stem cell biology and received his Ph.D. in Medicine from the University of Tokyo. From 2006 to 2016, he was at Stanford University studying single cell biology, systems biology, data-driven science and machine learning. His research interests cover the areas where AI meets biology, medicine, and beyond. He is also professor at the University of Tsukuba, advisor at avatarin inc., and member of World Economic Forum Frontier AI Systems and Capabilities.



[View the Presentation Slides](#)

2.5. From Answers to Action: Building the Evaluation Stack for Healthcare AI



Suhana Bedi

PhD Candidate, Biomedical Data Science, Stanford University USA



Abstract

As AI adoption in healthcare accelerates, the question is no longer whether large language model (LLMs) can answer medical questions, but whether they can safely assist clinicians and act across real system

This talk presented an evaluation stack spanning three modes of agency :

- MedHELM, a holistic benchmark system built with clinician-validated taxonomies covering 121 tasks, evaluates whether models produce the right answers
- ChatEHR, deployed at Stanford Health Care with over 1,000 trainees, evaluates whether AI helps users working with real medical records
- HealthAdminBench, grounded in over 100 hours of workflow shadowing, evaluates whether autonomous agents can complete administrative tasks end to end

Across all three, a consistent gap emerges between benchmark performance and workflow readiness. Ultimately, the keynote argued that closing this gap requires treating evaluation itself as a shared commons, with open taxonomies, datasets, and leaderboards that the clinical LLM community builds and maintains together.



About the Speaker

Suhana Bedi is a PhD candidate in Biomedical Data Science at Stanford University. Her research builds evaluation frameworks for healthcare AI, including MedHELM, HealthAdminBench, and ChatEHR, with a focus on how clinical LLMs can be tested before they are deployed in real healthcare settings.



[View the Presentation Slides](#)

2.6. PARTAGES : Advanced Development of Digital Commons for Generative Artificial Intelligence in Health



Aurélie Névéol

CNRS Research Director, and Head of the STL Department, LISN France



Abstract

In healthcare, strict regulatory frameworks and data sensitivity create structural silos that hinder the open development, sharing, and evaluation of clinical Natural Language Processing (NLP) solutions. To overcome these obstacles, this keynote presented PARTAGES, a nationwide initiative led by the French Health Data Hub that brings together healthcare institutions, research laboratories, medical residents' associations, and industry partners to build sovereign, open-source digital commons for generative AI in health.

The presentation detailed the four core pillars of the PARTAGES project:

- Open Corpora and Models: the curation of PARCOMED, a fully open French language models
- Synthetic Medical Reports (PARAF): the creation of a human-authored open corpus comprising 7,394 synthetic clinical reports across 5,009 patient profiles and 21 specialties, designed with a strictly held-out test to prevent data contamination during model benchmark evaluation
- Annotated Clinical Use Cases : the annotation of datasets for key healthcare applications, including pseudonymization, biomarker extraction, and clinical report summarization
- Federated Evaluation Platform: the deployment of a secure, privacy-preserving evaluation framework built on CodaBench, enabling researchers to benchmark algorithms directly on private in-hospital data without exposing patient records.

Finally, the talk outlined future perspectives, highlighting international shared-task evaluations at CLEF eHealth 2027 and a collaborative platform designed to sustain the creation and sharing of open clinical text resources.



About the Speaker

Aurélie Névéol is a CNRS Research Director and Head of the STL Department at LISN. She received a PhD in Computer Science at INSA de Rouen in 2005 and has been a research fellow at the U.S. National Library of Medicine in Bethesda, MD for more than six years before joining CNRS in 2012.

Her research focuses on clinical and biomedical Natural Language Processing, with expertise in information extraction, knowledge representation, and model adaptability across languages and domains. She is also deeply involved in ethical issues surrounding the development, evaluation, and deployment of NLP systems in healthcare and has been a co-chair of the ACL ethics committee since 2025.



[View the Presentation Slides](#)

3. Poster Abstracts

This section presents the 41 accepted poster abstracts from the conference, showcasing recent research in clinical NLP, synthetic data generation, privacy-preserving techniques, and LLM evaluation in healthcare.

LLMs' Fairness Evaluation by Syn-STARTS: Synthesized START Triage Scenario

Chiharu Hagiwara, Naoki Nonaka, Yuhta Hashimoto, Ryo Uchimido, Jun Seita

Extending a French OSCE Dialogue Dataset with Controllable LLM-Generated Dialogues

Doria Bonzi, Tom Bourgeade, Irina Illina, Fabrice Lefèvre

Safe and Responsible Use of Healthcare Text: Towards the SAFETEXT Community Protocols

Goran Nenadic, Jaya Chaturvedi, Arlene Casey, Angus Roberts, Robert Stewart, Sarah Markham, Jessica Bell, Franz Gruber, Warren Del-Pinto, Beatrice Alex, Elizabeth Ford

Verifying ICD-10 Labels in Clinical Corpora

Lucas Dedieu, Niama El Khbir, Marco Naguib, Rémi Flicoteaux

Automatic and Ethical Synthesis of Clinical Documents: the CODEINE Project

Aurélie Névél, Olivier Ferret, Karën Fort, Bastien Rance

Generating Linguistically Diverse and Semantically Faithful Patient Reports

Alejandra Lorenzo, Adrien Coulet, Claire Gardent

Extraction of Personal Identifiers from French Clinical Reports with Open Source LLMs

Niama El Khbir, Dimitri Bouche, Perceval Wajsbürt

Context-Augmented Generation of Synthetic Radiotherapy Follow-up Reports

Sarah Bouaraba, Lorraine Goeuriot, Philippe Mulhem, Didier Schwab

Where Does the Signal Live? A Web-Data Recipe for Medical Encoder Pretraining

Bofeng Huang, Jacques Sun, Diane Bouchacourt, Nicolas Barascud, Fajwel Fogel

Training Toxicity Extraction Models on Synthetic Data

Daro Sim, Sarah Bouaraba, Lorraine Goeuriot, Christophe Mazzara, François Portet

Difficulties of Training Clinical LLMs in Minority Languages

Iker De La Iglesia, Ane G. Domingo-Aldama, Maitane Urruela, Maciej Rybinski, Aitziber Atutxa, Ander Barrena

Open Baselines and Controls for Conclusion Extraction on PARHAF (PARTAGES CU3)

Amirhossein Ghanipour Amirhandeh

Augmentation of a Clinical Note Corpus in French using LLMs

Nicolas Hiebel, Priyanshu Priya, Joël Mba Kouhoue, Jesús Lovón, Gaël Dias, Gaël Varoquaux, Judith Abécassis

Benchmarking Clinical LLMs: The German Medical Text Corpus (GeMTex)

Justin Hofenbitzer, Luise Modersohn, Christina Lohr, Frank Meineke, Markus Loeffler, Martin Boeker

PARHAF Corpus for Assessing Functional Independence Scores

Trung Hieu Ngo, Arthur Chevreux, Adrien Bazoge, Emmanuel Morin

LLMs and patients (mis)information on rare diseases: the case of rare peritoneal cancers

Antoine T. Richard, Cécile Odin, Laurent Villeneuve

Cross-corpus Evaluation of Encoder Models for French Biomedical Named Entity Recognition

Joël Mba Kouhoue, Nicolas Hiebel, Judith Abécassis, Gaël Varoquaux

AAbAAC: An Annotated Corpus for Autoimmunity Information Extraction

Fabien Maury, Solène Grosdidier, Maud de Dieuleveult, Adrien Coulet

Can LLMs Reliably Conduct Qualitative Evaluation of Generated Discharge Summaries, and Suggest Improvements?

Claire Dechaux, Basile Fuchs, Arthur Lamard, Francesco Monti, Stéphane Ohayon, Adrien Coulet, Rémi Flicoteaux, Éric Villemonte de la Clergerie

ECOSBot: Multicenter Validation of a Clinical LLM as Both Patient and Examiner

Raphaël Bentegeac, Nans Florens, Mehdi Maanaoui, Valentin Maisons, Antoine Lanot, Mickaël Bobot, Benoît Brilland, François Glowacki, Erwin Gérard, Marc Hazzan, Philippe Amouyel, Bastien Le Guellec, Aghiles Hamroun

Tanit-RI: An Ablation-Driven Recipe for Efficient Bilingual (French-English) Clinical Reasoning

Bechir Dardouri, Amen Debbiche, Kais Zhioua

Building a Reusable, Schema Constrained Clinical LLM-Pipeline for Multicenter Real World Evidence Generation

Titouan Lacombe, Arthur Perez, Camille Chaussy, Emmanuel Gilson, Isabelle Delarozière, Lorraine Houvet

Application of Large Language Models for Automated Structuring of Clinical Databases in Visceral Surgery

Anne-Charlotte Morel, Julien Kallout, Benjamin Popoff, Badisse Dahamma, Romain Lelong, Ndeye Aissatou Marone, Stefan J. Darmoni, Edouard Roussel, Julien Grosjean

Beyond FAIR Principles: Local Agentic Chatbot for Multicentric Clinical Trial Capability Sharing

E. Njomgue, V. Gurunathan, G. Desrues, H. Cochet, M. Sermesant

Bridging the Maternal Healthcare Gap in Rural Senegal: A Multilingual WhatsApp Voice Assistant in Wolof and Pulaar

Khady Kama, Daouda Tandiag Djiba

PhenoPrompt: Open-Source Clinical Phenotyping with Entity Extraction, Embedding Maps and LLM Grounded Retrieval

Shafiya Kausar, Anuradha Kar

Leveraging LLMs and RAG for the Analysis of Contrast Agents Allergies

Dinh-Viet-Toan Le, Samuel Ansel, Lucas Shorten, Aghiles Hamroun, Frédéric Dezoteux, Bastien Le Guellec



Grounded alerts, auditable decisions: rule-and-LLM pre-validation of day-hospital prescriptions

Sarra Ben Yahia, Sarah Zrek, Aurélie Chan Hew Wai

Ontology-Aligned Prompting for Semantic RDF Extraction: A Case Study in Rare Oro-dental Diseases

Socrates Waka Onyando, Ali Ayadi, Marzena Kawczynski, Agnès Bloch-Zupan, Cédric Wemmert

Neuro-Symbolic GraphRAG for Verifiable Clinical LLM Reasoning: Toward Safety-Critical Benchmarks for Clinical Commons

Negash Desalegn Welkdekiros, Gjergji Kasneci

A Generalizable Large Language Model Pipeline for Automated Extraction of Metastatic Progression Dates from Clinical Records Across Cancer Types

Paul Vallée, Lucie Caillot, François Bocquet, Florent Le Borgne

Adaptation Datasheets for Plain Language in Health: A Pragmatics-Centered Traceability Schema

Rémi Cardon, Lourdes Moreno, Òscar Garcia-Muñoz, Paloma Martínez

How To Get Clinical LLMs Right: Lessons from Machine Learning in Medical Research

Chiara Marcoccia

Evaluating Patient Safety Risks in Generative AI: A FMECA Framework for LLM-Generated Clinical Summaries

Jamil Zaghir, Lydie Bednarczyk, Julien Ehram, Maria Tcherepanova, Christian Skalaforis, Karim Gariani, Catherine Geslin, Claire-Bénédicte Rivara, Pascal Bonnabry, Laetitia Gosetto, Jean-Philippe Goldman, Laetitia Gosetto, Richard Dubos, Riccardo Bellazzi, Mina Bjelogrljic, Christophe Gaudet-Blavignac, Christian Lovis

Evaluation of the Medical Summary in the Pre-Anesthesia Consultation: A Randomized Controlled Trial to Guide the Integration of LLMs

Julien Kallout, Julien Grosjean, Stefan Darmoni, Romain Lelong, Badisse Dahamna, Inès Maillot, Maxence Goullé, Antoine Lefevre-Scelles, Mariette Renaux-Petel, Havana Micholet-Scandolo, Benjamin Popoff, Vincent Compère

How Hard Should the Exercises Be? Calibrating Supervision Difficulty for Small LLMs on Clinical Tasks

Antoine Messenger, Marco Naguib, Chuyuan Li

eCREAM-MedCorpus: Building an Open Italian Clinical Corpus

Pietro Ferrazzi, Guido Bertolini, Tiziano Labruna, Bernardo Magnini

Summarizing French clinical reports: Data curation and evaluation using document conclusions

Dimitri Bouche, Antoine Messenger, Niama El Khbir, Laura Zanella, Pierre Zweigenbaum, Aurélie Névél



VESTIGE: What De-identification Leaves Behind in Multilingual Healthcare Communication

Lisa Raithel, Christoph Otto, Antonin Meyer, Matthieu Le Cauchois, Adèle Lutun, Vera Czehmann, Katharina Kersting, Elisabeth Eberle, Sebastian Möller

Open Synthetic Resources for French Clinical NLP

Rian Touchent, Éric de la Clergerie

Automatic Medical Discharge Summary Evaluation

Andrea Franchini, Lorenzo Ruinelli, Amos Colombo, Samuel Corecco, Behnaz Rezaeifar, Sotirios Georgios Popeskou, Mathilde Marie Rochat, Dimitrios Christoforidis, Nicolò Saverio Centemero, Matteo Cerina, Fabio Rinaldi

4. Panel Discussions

The panel discussions held during the conference brought together experts from research, clinical practice and industry. The two sessions explored challenges surrounding clinical LLM development, training, evaluation, deployment, adoption and governance.

4.1. Panel Discussion 1 - Building Commons for Clinical LLMs: Models, Training Dataset, Cross-Lingual Challenges, and Governance

Moderator:

Chuyuan Li, CNRS researcher at LIG, Université Grenoble Alpes

Panelists:

Julia Ive, Associate Professor in Healthcare AI at University College London

Lars Klein, Academic Director at LiGHT, École Polytechnique Fédérale de Lausanne

Martin Krallinger, Head of the NLP4BIA at Barcelona Supercomputing Center

Bernardo Magnini, Head of the NLP research unit at Fondazione Bruno Kessler

1. Context

Despite the rapid evolution of open-source medical LLMs, a significant gap remains between promising academic models and safely deployable clinical tools. This panel explored how to build robust digital commons for healthcare, debating the optimal balance between general and specialized models, the role of synthetic versus real clinical data, and the limitations of translation in overcoming multilingual barriers.

2. Key Takeaways

Dedicated Medical Models vs. General-Purpose LLMs

There is no single definitive answer, model selection depends heavily on the use case, language, and deployment environment. General-purpose models might serve as a good starting point for highly under-resourced languages or settings lacking robust IT infrastructure, whereas specialized models or standardized pipelines are necessary for large clinical studies requiring comparable results across sites.

Characteristics of Shared Clinical Training Datasets

Acquiring real clinical data involves complex, multi-year ethical approvals and rigorous de-identification. This process, as seen in the eCREAM project, often requires disaggregating notes to protect privacy, which limits the types of tasks models can be trained on. Ultimately, data quality is paramount; establishing realistic, clinically relevant evaluation benchmarks is critical to determine the true value of any dataset.



Benefits and Limitations of Synthetic Data

Synthetic data is never a standalone solution. It must be combined with real data as a seed. Its primary benefit lies not in creating new facts (which risks hallucination) but in reshaping and rephrasing existing knowledge to prevent out-of-distribution errors at inference.

However, it suffers from a lack of variety and requires real-world data to address flaws or shift the model toward future edge cases.

Multilingual challenges: Limits of Translation

Relying solely on translating English datasets is not a viable strategy. Automatic translation from English inherits cultural biases, failing to capture how clinical notes would naturally be written in the target language. Furthermore, translation quality varies drastically depending on the language pair (e.g., Romance languages versus under-resourced languages like Romanian or Czech).

Adapting to New Languages

Native data should always be the first choice. When machine translation is used (as a form of data augmentation), it requires domain adaptation and rigorous quality control involving both professional medical translators and domain experts (e.g., cardiologists), which raises significant cost and scalability issues.

Open Challenges

Major unresolved hurdles include the prohibitive cost and scalability issues of human-validated medical translation for under-resourced languages. Additionally, the open-source community faces significant methodological bottlenecks in preserving the structural utility of real clinical records without compromising patient privacy, and in advancing toward natively multimodal clinical models (early fusion) to match proprietary frontier models.

Conclusion

To bridge the gap between research and real-world clinical application, the AI community must move beyond simply translating English data and prioritize the creation of culturally aware, high-quality multilingual datasets and clinically grounded benchmarks. Sustainable digital commons will ultimately rely on deep, cross-institutional collaborations to safely unlock real patient data and build tools that genuinely serve frontline healthcare professionals.

4.2. Panel Discussion 2 - Deploying LLM-Based Solutions into Clinical Practice: Challenges, Governance & Perspectives

Moderator:

Youness Khalil, Head of AI Division at French Health Data Hub

Panelists:

Nicolas Fiorini, Data Science & AI Director at Doctolib

Guillaume Martin, Deputy Chief Medical Officer at Synapse Medicine

David Morquin, Medical Head of Alliance Santé IA & Chief Medical AI Officer, CHU Montpellier

Jamel Zaghir, Researcher, Division of Medical Information Sciences, Hôpitaux Universitaires de Genève, Université de Genève

1. Context

As Clinical LLMs move from research to real-world use, their deployment raises a wide range of challenges that healthcare organizations and solution providers are starting to address in practice. This panel brought together two complementary perspectives, hospitals and industry, to explore these challenges, how they are being addressed today, and what can be learned from their respective experiences.

2. Key Takeaways

Legal and Architectural Readiness

Infrastructure (compute power, data privacy) is only the first step. The major challenges involve navigating complex legal regimes regarding data use (care vs. research) and building LLM-agnostic, modular architectures (like RAG systems). These architectures must orchestrate information retrieval reliably, despite paywalls blocking access to essential medical knowledge.

Generalizability & Risk Evaluation

Scaling an AI assistant across different countries (with varying languages and regulations) and medical specialties requires extensive product adaptation rather than a one-size-fits-all technical solution. Furthermore, evaluating a model for deployment must shift from measuring raw performance to rigorously evaluating potential risks, as high performance does not equate to clinical safety.

Adoption, Trust, and Accountability

Deployment is fundamentally a socio-technical process. Success requires co-designing the workflow with users and breaking down complex tasks into smaller, manageable steps (human-in-the-loop) to ensure the cognitive load of oversight does not overwhelm the clinician. Trust from patients (e.g., via chatbots) is built through pragmatic, balanced responses that provide actionable next steps without causing unnecessary alarm.



The Evolving Role of the Physician

Accountability remains a complex issue, especially with non-deterministic (probabilistic) systems where identical errors can recur despite safeguards. As AI tools integrate with clinical systems, they increasingly fall under stringent Medical Device Regulations. Furthermore, the rise of "shadow AI" (unauthorized use of AI by staff) complicates institutional governance and liability.

Monitoring & Safety

Deployment is not a finish line but the start of continuous algorithm vigilance. Because technology evolves faster than traditional clinical trials can evaluate, hospitals and providers must implement robust feedback loops and incident reporting systems. Every model update requires re-evaluation to detect unexpected behaviors and mitigate new risks before they cause irreversible harm.

Open Challenges

Significant challenges remain in establishing scalable frameworks for continuous evaluation in a landscape where technology outpaces traditional clinical trial methodologies. Furthermore, defining accountability in non-deterministic systems, standardizing risk evaluation, and overcoming the financial and legal barriers to accessing gated medical knowledge for RAG systems are ongoing struggles.

Conclusion

Successfully integrating LLMs into routine care requires a paradigm shift from traditional, one-time software deployment to continuous, modular, and heavily monitored socio-technical collaboration. To achieve scalable and safe adoption, the healthcare industry must prioritize building LLM-agnostic architectures, embedding rigorous risk-evaluation frameworks into the clinical workflow, and fostering deep collaboration between developers, clinicians, and regulators.

5. Workshop Outcomes

The second day of the conference was dedicated to collaborative workshops, with the objective of identifying potential solutions to shared challenges and exploring opportunities for international collaboration around commons for Clinical LLMs.

Participants joined one of three parallel workshops covering the main stages of the Clinical LLM lifecycle:

- Workshop 1 - Build Together : shared resources for developing and adapting Clinical LLMs
- Workshop 2 - Evaluate Together : shared approaches for robust and clinically relevant evaluation
- Workshop 3 - Deploy Together : shared solutions to support safe and scalable deployment in real-world healthcare settings

Each workshop followed the same approach: participants first discussed and prioritized the challenges that would benefit most from international collaboration, before working in small groups to propose potential commons : shared resources, methods, frameworks or infrastructures that could address these challenges collectively.

The following section summarizes the priorities identified by participants and the potential commons that emerged from these discussions.

5.1. Build Together Workshop - Key Takeaways

The workshop highlighted three main areas where shared resources could support the development of Clinical LLMs across institutions and countries:

- **Shared multilingual clinical terminologies and standards** to better align medical concepts across languages, cultures and healthcare systems, and improve the transferability of Clinical LLMs.
- **Shared approaches and tools for synthetic clinical data generation**, enabling institutions to create useful training data while addressing privacy, intellectual property and data-sharing constraints.
- **Federated datasets and training frameworks** to enable collaborative model development and local adaptation across institutions, specialties and languages without centralizing sensitive clinical data.

5.2. Evaluate Together Workshop - Key Takeaways

The workshop highlighted three complementary areas where international collaboration could strengthen the evaluation of Clinical LLMs:

- **Shared reliability and abstention guidelines** to assess the trustworthiness of Clinical LLM outputs, define when systems should express uncertainty or abstain, and adapt these mechanisms to different clinical tasks and levels of risk.
- **An international rolling clinical benchmark** based on regularly renewed, expert-validated clinical scenarios across countries and languages, to limit contamination and enable more robust and comparable evaluation over time.
- **Human-AI evaluation frameworks** to move beyond evaluating the model alone and assess the performance, safety and usability of the user-system pair within real clinical workflows.

5.3. Deploy Together Workshop - Key Takeaways

The workshop highlighted two main areas where shared approaches could facilitate the safe adoption and deployment of Clinical LLMs:

- **Shared frameworks and data models for AI integration** to facilitate the integration of Clinical LLMs into healthcare systems, support interoperability across languages and settings, and provide common approaches for assessing their use in practice.
- **International incident monitoring and reporting standards** to define common monitoring indicators, reporting workflows and shared incident registries, enabling healthcare organizations to learn collectively from failures and emerging risks.

The next step will be to explore synergies between the commons identified during the workshops, existing initiatives and funding opportunities, and to bring together interested partners through dedicated working groups. These discussions could ultimately lead to the development and launch of new international collaborative projects around Clinical LLM commons.

6. Looking Ahead

Expanding the Clinical AI Commons

The *Building Commons for Clinical LLMs* conference confirmed a simple but important message : many of the challenges surrounding Clinical LLMs are shared across countries, institutions and organizations, and not all of them need to be solved independently.

Across two days of discussions, experiences from research, healthcare institutions and industry showed the value of joining forces, mutualizing efforts and building open, reusable resources to support the development, evaluation and deployment of Clinical LLMs.

Building on this growing international community, the ambition is to continue connecting initiatives, sharing knowledge and advancing Clinical AI commons that can benefit the broader ecosystem.

We extend our gratitude to all speakers, scientific committee members, and participants for their contributions, and we invite the global community to continue building, evaluating, and deploying clinical AI commons together.

For any questions or further information, please contact us at: direction.scientifique@health-data-hub.fr.