Doctolib

Machine Learning Engineer Interns

Join the Artificial Intelligence team in Digital Health

About Doctolib

An innovative digital health company of

2,900

Doctolibers in France, Germany, Italy, and the Netherlands

900+

Tech & product, AI experts

80 million

patients use Doctolib

400,000

health professionals across Europe

More than **30 locations**

B-corp certification

Certified B Corporation

We leverage AI ethically across our products to empower patients and health professionals.

Discover our AI vision here and learn about our first AI hackathon here!

Interns for April 2026

What tasks or projects could I contribute to as a Machine Learning Engineer Intern?

Summary

12 topics

01
Comprehensive Evaluation of a Clinical AI System
(MLE)

02
High Quality and Robust ASR through Pretraining
(MLE)

03
Health Agentic Evaluation Framework
(MLE)

04
Scaling DoctoBERT: Multilingual Encoder Model for Medical Data, Continuum Pretraining and Finetuning on Specific Use Cases
(MLE)

05
Feedback-Driven Adaptive Classification for AI agents
(MLE)

06
Finetuning a RAG Model for Clinical Information Retrieval
(MLE)

07
Finetuning a Model for Precise Fact Matching in Clinical Text
(MLE)

08
Sentiment Analysis in Patient Calls
(MLE)

09
Adaptation to Accents and Environments
(MLE)

10
Text-to-Speech (TTS) Evaluation
(MLE)

11
Synthetic Data Generation: Develop methods to generate realistic synthetic data for training and evaluating machine learning models.
(MLE)

12
Scalable Real-Time Speech Infrastructure
(ML OPS)

Topic 1

Comprehensive Evaluation of a Clinical AI System

(MLE)

Description

This project delves into the intricate challenge of evaluating complex Large Language Model (LLM) systems.

As LLMs become more sophisticated and integrated into various applications, the need for a comprehensive and robust evaluation framework is paramount.

- Interns will research and implement novel methodologies to assess LLMs not just on accuracy, but also on aspects like relevance, reasoning, safety, and real-world applicability.
- This involves exploring metrics beyond traditional benchmarks, considering human-in-the-loop evaluation, and developing automated tools to streamline the process.
- A key focus will be on creating an adaptive framework that can evolve with the rapid advancements in LLM technology.

Such a framework entails several challenges: variance and biases of scores produced by LLM as judges, reliance on time-consuming annotations.

This work will be crucial in ensuring that our LLM-powered products are reliable and trustworthy.

Interns will gain deep insights into the internal workings of LLMs and the critical importance of rigorous validation. (see Croxford et al., 2025, Arora et al., 2025, Peyrard, 2020, Malinar et al., 2025)

Potential Outcome

A comprehensive and expandable framework for evaluating complex LLM systems at scale, complete with documentation and implementation.

Topic 2

High Quality and Robust ASR through Pretraining

(MLE)

Description

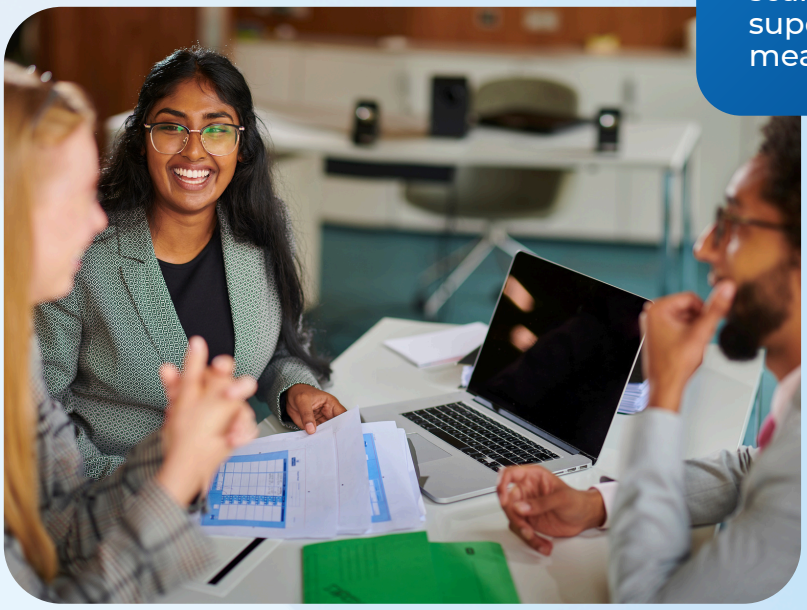
Doctolib develops state-of-the-art speech-to-text models for healthcare, built on Conformer-style architectures and deployed in real clinical workflows.

So far, our research has focused on fine-tuning with annotated and synthetic data derived from medical texts, production usage, and user feedback, as well as domain adaptation through language models.

This internship will explore how self-supervised speech representation learning (e.g., wav2vec 2.0, HuBERT, data2vec, or newer approaches) can be used to enhance robustness and generalization across languages, accents, background noise, and medical specialties.

Potential Outcome

A self-supervised medical ASR model trained on large-scale unlabelled audio, showing improved robustness and scalability compared to supervised baselines and measurable gains in real use.



Topic 3

Health Agentic EvaluationFramework

(MLE)

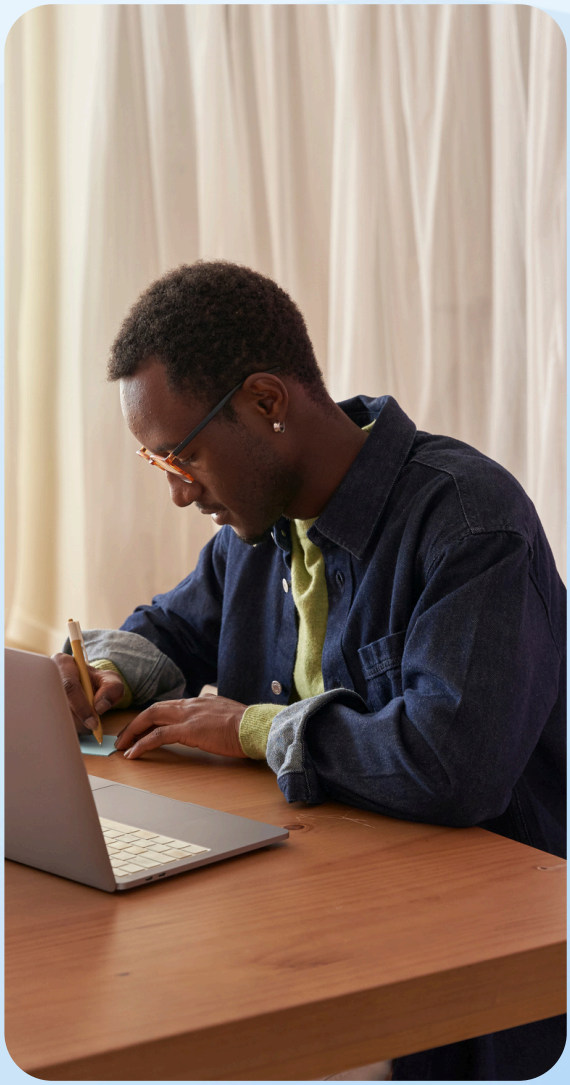
Description

This project focuses on the challenge of evaluating compound AI systems on healthcare-specific safety and usefulness metrics, using offline evaluation on historical clinical workflows and counterfactual replay to assess agent policies without any risk to patients.

Your work will extend our framework, originally developed for a clinical interview assistant, to more specialized medical agents (triage, diagnosis, knowledge search, coding).

- The role will require running agentic system simulations, handling evaluation datasets, and developing evaluation modules.
- You will ground methods in recent advances on agent evaluation, healthcare AI safety, and structured LLM tool-use

(e.g., Mukherjee, Subhabrata, et al. "Polaris: A safety-focused LLM constellation architecture for healthcare." arXiv:2403.13313 (2024); Tu, Tao, et al. "Towards conversational diagnostic artificial intelligence." Nature (2025): 1–9).



Potential Outcome

A modular, offline evaluation framework for healthcare agentic systems, tested across multiple specialized medical agents (triage, diagnosis, knowledge search, coding).

Topic 4

Scaling DoctoBERT:

Multilingual Encoder Model for Medical Data, Continuum Pretraining and Finetuning on Specific Use Cases

(MLE)

Description

This project focuses on scaling and extending DoctoBERT, a medical language model built on ModernBERT, into a multilingual setting.

- Interns will work on developing a multilingual encoder designed specifically for medical data, a domain with unique linguistic challenges and a critical demand for accuracy.
- The work will involve large-scale training on distributed systems with advanced optimization methods, as well as extensive data curation: collecting, cleaning, and using LLMs to synthesize a wide corpus of medical text in multiple European languages.
- Interns will fine-tune the multilingual DoctoBERT on targeted medical applications, such as ICD coding, named entity recognition, and experimenting with approaches like GliNER for more general-purpose usage.
- The role will require handling diverse datasets, managing computational resources, and evaluating model performance across languages and tasks.
- The ultimate goal is to build a powerful and versatile language model that can support medical professionals and researchers worldwide.

This is a unique opportunity to take part in cutting-edge medical NLP research.

Potential Outcome

A scaled, multilingual DoctoBERT model, fine-tuned for specific medical use cases, along with performance benchmarks and insights into multilingual medical NLP.

Topic 5

Feedback-Driven Adaptive Classification for AI agents

(MLE)

Description

This project focuses on developing adaptive classification in AI agents that learn directly from user interactions and annotations.

The current zero-shot classification layer in the customer support agent is effective but limited in handling new or evolving cases.

Feedback-driven methods will refine classifications over time by leveraging positive and negative examples obtained from user responses and annotations.

Rather than relying on a static pipeline, a multi-agent workflow with roles such as proposer, reviewer, and finaliser will be designed to ensure an accurate and scalable classification process.

In parallel, robust evaluation strategies for adaptive multi-classification will measure how effectively the model adjusts to evolving user behavior.

Interns will gain hands-on experience in system design, feedback-driven adaptation, and evaluation within AI-driven applications.

Potential Outcome

A production-grade classification AI agent that surpasses the current zero-shot LLM baseline. The outcome includes improved classification accuracy over time and robust evaluation methods for adaptive classification, providing a reliable way to measure and enhance AI agent performance in production environments.

Topic 6

Finetuning a RAG Model for Clinical Information Retrieval

(MLE)

Description

This internship focuses on fine-tuning a Retrieval Augmented Generation (RAG) model for the highly specialized domain of medical information retrieval.

RAG systems combine the strengths of retrieval-based and generative models, allowing them to access external knowledge bases for more accurate and grounded responses.

- Interns will work on processing medical guidelines to serve as the knowledge source for the RAG system.
- The fine-tuning process will involve adapting the retrieval part to understand a user query, identify relevant clinical entities, and generate coherent and factually accurate responses to clinical questions.

This project is vital for improving the efficiency and accuracy of our medical agents.

Interns will gain hands-on experience with state-of-the-art RAG systems and contribute to critical applications in healthcare.

Potential Outcome

A fine-tuned RAG system optimized for clinical information retrieval, demonstrating improved accuracy and relevance of generated responses.

Topic 7

Finetuning a Model for Precise Fact Matching in Clinical Text

(MLE)

Description

This project aims at comparing medical texts generated by medical chatbots against expert answers. Usually, this is done with off-the-shelf LLMs, however we want to get higher performance at a lower cost by finetuning a BERT (or similar) model.

- This project will involve building a dataset, finetuning a model, comparing it against a baseline and making it usable for a wide range of use cases.

Interns will gain deep hands-on experience of evaluation of LLM-based systems in healthcare.

Potential Outcome

A model will be finetuned on this specific task , outperform our current baseline based on LLMs and prompt engineering and become our standard way to estimate medical LLM Agent performance.

Topic 8

Sentiment Analysis in Patient Calls

(MLE)

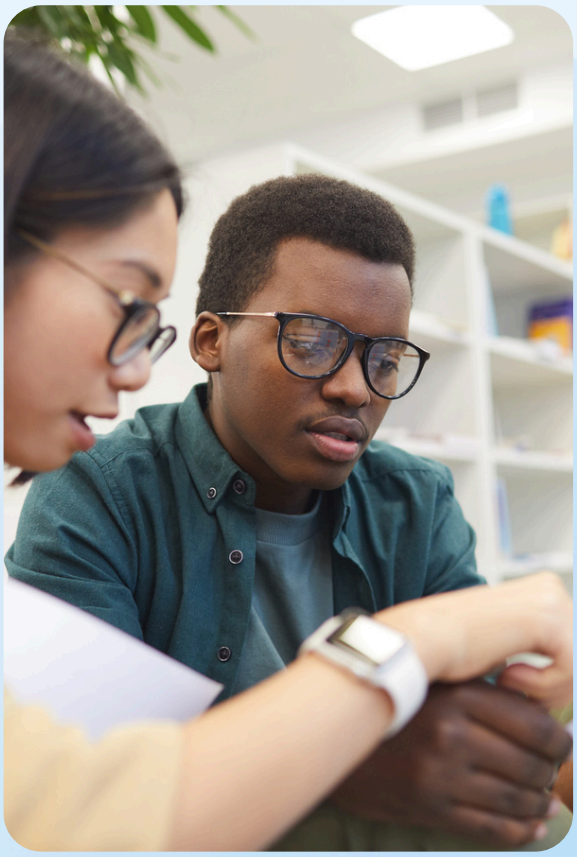
Description

This project involves developing a cutting-edge sentiment analysis system specifically designed for patient calls and communications exchanged through Doctolib's phone assistant.

The primary objective is to automatically identify and flag messages that indicate potentially urgent situations or express negative sentiments.

Interns will be responsible for the entire pipeline, from data annotation to model deployment.

- This includes curating and labeling a dataset of patient communications for sentiment and urgency, experimenting with transformer-based models such as BERT for text and HuBERT for audio, exploring a multimodal approach, and optimizing the system for both accuracy and real-time performance.
- A crucial aspect will be to differentiate subtle nuances in patient language that might signal distress or a critical need for intervention.



This work has direct implications for improving patient care by enabling faster response times and more targeted support.

Potential Outcome

A deployed sentiment analysis system capable of detecting urgent situations and negative sentiments in patient communications, with documented performance metrics.

Topic 9

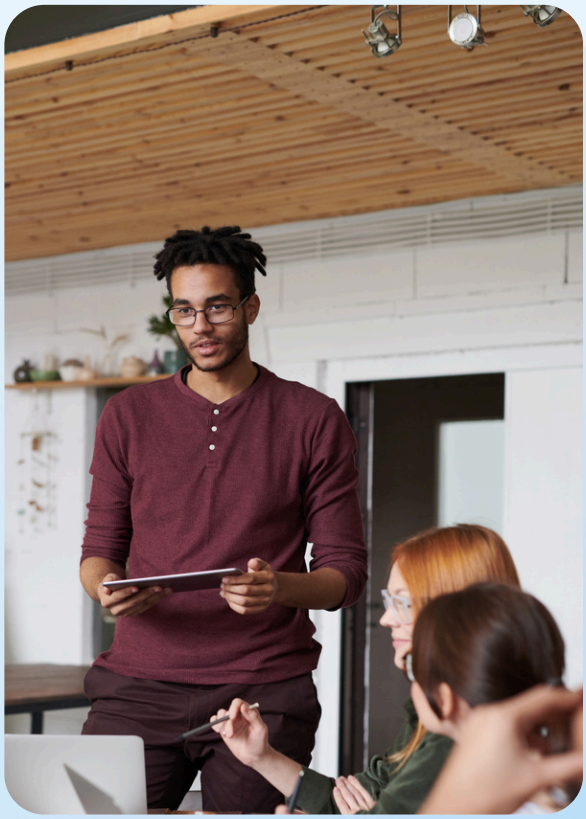
Adaptation to Accents and Environments

(MLE)

Description

This critical project addresses the challenge of fairness in Automatic Speech Recognition (ASR) systems by focusing on reducing the performance gap for different accents, regions (specifically French and German), and background noise environments.

- Interns will explore state-of-the-art fine-tuning techniques such as Parameter-Efficient Fine Tuning and Domain-Adaptive Pretraining to efficiently adapt existing ASR models to diverse vocal characteristics without extensive retraining.
- Data augmentation strategies will be key to creating robust datasets that represent a wide range of real-world acoustic conditions.
- A significant part of this project will involve defining and implementing fairness metrics to quantitatively assess and track improvements in ASR performance across different demographic groups and environmental conditions.



This work is essential for building inclusive and equitable speech technologies.

Potential Outcome

An ASR model with reduced WER gaps across different French accents and noise environments, accompanied by a detailed analysis of fairness metrics and implemented adaptation techniques.

Topic 10

Text-to-Speech (TTS) Evaluation

(MLE)

Description

This internship focuses on the critical task of evaluating the quality of Text-to-Speech (TTS) systems, particularly for applications involving patient communications in French and German.

- Interns will be responsible for developing and implementing a comprehensive evaluation framework, which can potentially be extended into a broader TTS benchmarking suite.
- This will involve researching and applying various speech and audio analysis techniques to objectively measure aspects like naturalness, intelligibility, prosody, and emotional congruence.

A key aspect will be the use of established evaluation metrics, such as Mean Opinion Score (MOS), which often requires human perception studies, alongside automated metrics.

Interns will interact with different TTS APIs and compare their performance against the defined quality criteria.

The goal is to ensure that our TTS systems deliver high-quality, empathetic, and clear audio for sensitive patient interactions.

Potential Outcome

A comprehensive evaluation framework for TTS systems, including objective and subjective metrics, and a comparative analysis of different TTS systems for patient communications in French.

Topic 11

Synthetic Data Generation:

Develop methods to generate realistic synthetic data for training and evaluating machine learning models.

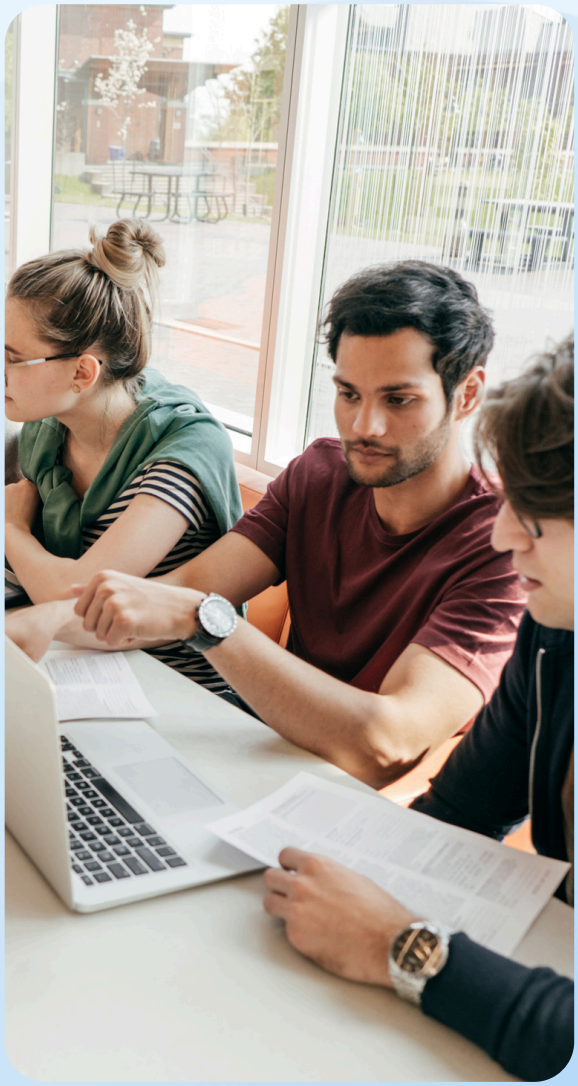
(MLE)

Description

In the healthcare sector getting access to data and especially annotated data is costly.

This innovative project is dedicated to developing methods for generating realistic synthetic data, a crucial capability for training and evaluating machine learning models.

- This will involve understanding the structure and nuances of clinical dialogues, developing intelligent data pipelines to transform real data into synthetic counterparts, and ensuring the generated data maintains statistical properties and representativeness.
- The ability to create high-quality synthetic data significantly accelerates model development and enables testing in scenarios where real data is scarce or inaccessible.



Potential Outcome

Implemented methods and a prototype system for generating realistic synthetic data, including mock clinical conversations, along with a thorough evaluation of data quality and utility for ML training.

Topic 12

Scalable Real-Time Speech Infrastructure

(ML OPS)

Description

Doctolib operates state-of-the-art medical speech recognition models that power products such as our Consultation Assistant, Medical Dictation, and Telephone Assistant.

These models run in both batch and real-time streaming modes. While the batch infrastructure is already robust and widely deployed, the streaming stack is less mature and needs to evolve into a scalable, production-grade system.

The internship project will focus on designing and implementing the next generation of our real-time speech infrastructure.

- This involves analyzing the limitations of the current approach, exploring distributed and fault-tolerant architectures (e.g., task orchestration frameworks and message queues), and building a secure, high-performance system that can support large-scale usage.

Since most speech-related product requirements converge toward streaming, this system will become a core pillar of Doctolib's speech AI capabilities.

Potential Outcome

A scalable and production-ready real-time speech infrastructure, with benchmarks demonstrating improved throughput, reliability, and fault tolerance.

Doctolib

careers.doctolib.com