

Doctolib

Machine Learning **Internships**

All our internships

Doctolib for Doctors

- Leverage reinforcement learning to improve Alfred, an intent-detection agent performance
- Build an automated eval-driven loop that tunes intent-detection agent
- Evaluate & optimize the latency in healthcare
- Fine-tune a Multi-task Clinical Document Creator
- Improve our consultation assistant templates

Voice recognition

- Improve evaluation of our phone assistant conversations
- Voice-to-voice conversational AI with task-specific sub-agent orchestration for Phone Assistant
- Post-training small language models for phone assistant agents
- Streaming ASR - evaluation and domain fine-tuning for healthcare voice applications

Doctolib for Patients

- Improve LLM Post-Training
- Detect Medical safety issues
- Build an evaluation model leveraging feedback from patients & health pros
- Learn to deliver health suggestions that support patients
- Clinical Perception from live audio & video streams
- Build full-duplex spoken health assessments
- Synthetic patient trajectories for training & evaluating AI

Doctolib Evaluation

- Secure our Loop Engineer that improves the quality of our agentic systems



DOCTOLIB FOR DOCTORS

Leverage reinforcement learning to improve Alfred, an intent-detection agent performance

Context:

Doctolib is building core agentic systems to generate clinical outputs such as patient summaries, referral letters or handle medical coding. As these systems move from deterministic pipelines to agents that retrieve patient data and call tools on their own, they become harder to evaluate. They also cost more and respond more slowly. The goal is to get the latency and cost down without degrading clinical quality.

Mission:

The intern will design and run a framework that measures agent performance, finds latency and cost bottlenecks, and tests optimizations against clear quality, speed, and cost targets.

- **Evaluate agents:** build evaluation methods that go beyond the final output to cover full trajectories, tool calls, and retrieval steps, on top of the existing evaluation framework.
- **Profile performance:** instrument flows to break down latency and cost per step
- **Test optimizations:** try levers such as prompt and context compression, model routing (smaller models for simpler steps), caching, parallel tool calls, early stopping, fewer retrieval rounds and/or post training of open weight models.
- **Measure trade-offs:** use A/B tests and LLM-judge scores to quantify each optimization's effect on quality, latency, and cost.



DOCTOLIB FOR DOCTORS

Build an automated eval-driven loop that tunes intent-detection agent

Context:

Alfred is a Customer Support agent at Doctolib. Alfred's internal agents are tuned by hand today. We edit prompts, configurations or parameters, then run the evals, using calibrated offline and online scorers. This internship builds an eval-driven optimization loop for one Alfred internal agent, e.g. intent detection or RAG answer generation. The loop proposes candidate changes, scores each one against the scorers, and promotes a change when the gain is statistically sound.

Mission:

- **Build the closed loop:** candidate generation (prompts, few-shot examples, instructions, model choice), scoring through the existing eval system, and promotion of a candidate.
- **Optimize several objectives together** (answer quality, tone-of-voice adherence, latency, cost). Guard against overfitting the eval set and gaming the scorers.
- **Benchmark the auto-pilot against the current manual tuning**, and package it so other Alfred agents and future subagents can reuse it.

Potential outcome:

A working auto-pilot that measurably improves one Alfred agent without manual iteration



DOCTOLIB FOR DOCTORS

Evaluate & optimize the latency in healthcare

Context:

Doctolib is building core agentic systems to generate clinical outputs such as patient summaries, referral letters or handle medical coding. As these systems move from deterministic pipelines to agents that retrieve patient data and call tools on their own, they become harder to evaluate. They also cost more and respond more slowly. The goal is to get the latency and cost down without degrading clinical quality.

Mission:

The intern will design and run a framework that measures agent performance, finds latency and cost bottlenecks, and tests optimizations against clear quality, speed, and cost targets.

- **Evaluate agents:** build evaluation methods that go beyond the final output to cover full trajectories, tool calls, and retrieval steps, on top of the existing evaluation framework.
- **Profile performance:** instrument flows to break down latency and cost per step.
- **Test optimizations:** try levers such as prompt and context compression, model routing (smaller models for simpler steps), caching, parallel tool calls, early stopping, fewer retrieval rounds and/or post training of open weight models.
- **Measure trade-offs:** use A/B tests and LLM-judge scores to quantify each optimization's effect on quality, latency, and cost.



DOCTOLIB FOR DOCTORS

Fine-tune a Multi-task Clinical Document Creator

Context:

Health Practitioners write a lot of documentation every day: from consultation notes, referral and report letters to colleagues, prescriptions, messages to their patients... This time-consuming process is also error-prone: they might miss important information from documents, letters etc..

Our Consultation Assistant and Letter Generation services aim to assist them in these tasks. Such use cases will only grow over time. Current frontier models (GPT, Gemini, Claude variants) however are not adapted to generating clinical documents in a concise, accurate way using correct medical terminology.

Mission:

The internship will aim at fine-tuning a model to tackle this issue. This will include refining datasets, defining training procedures aimed at these particular aspects, fine-tune and benchmark candidate models and validate the approach with A/B tests with real users. Patient safety is paramount and thus the intern will also need to integrate guardrails in the fine-tuning procedure.

Potential outcome: A fine-tuning recipe to build models apt to produce high quality clinical documentation and a fine-tuned model validated on at least one use case, ideally on several.



DOCTOLIB FOR DOCTORS

Improve our consultation assistant templates

Context:

In Doctolib's Consultation Assistant, practitioners define customizable templates that set the content and structure of their clinical notes. At the end of each consultation, the template and the transcript are used together to generate the note. Practitioners often edit the note before saving it. Today these edits are only used for analytics and never feed back into the product. The internship will turn them into a live feedback loop.

Mission:

- **Analyse edits:** compare generated and edited notes and sort the edits into types (added section, deleted section, added instruction-like content, reformulation).
- **Trace edits to the template:** separate the edits that come from template gaps from generation errors, and find the ones that recur across consultations.
- **Propose a new template version:** turn recurring edits into concrete changes (add or remove a section, add an instruction) and suggest them to the practitioner at the end of the consultation.
- **Explore practitioner memory (bonus):** capture edits that recur across all of a practitioner's templates, such as style, phrasing or abbreviations, as a memory that sits on top of the template.

Potential Outcome: A feedback step at the end of the consultation that suggests template updates based on the practitioner's edits and creates a new template version once they accept. Success would be a measurable drop in post-generation edits.



DOCTOLIB FOR PATIENTS

Improve LLM Post-Training

Context:

The Health Companion answers patients' health questions and guides them to the right care. It runs on an agentic system that mostly relies on prompted frontier models. Prompting has limits: some medical errors keep coming back, fixing one behaviour often breaks another, and the model has no reliable sense of when not to answer. A first open-weight model, fine-tuned with SFT on one health topic (sleep), already runs in production.

Mission:

Find which post-training methods best improve medical accuracy, answer reliability and clinical reasoning. Medical answers are open-ended, so their correctness can't be checked automatically. The intern will:

- **Build training data** from our evaluation vignettes, case-specific rubrics and doctor feedback.
- **Compare methods:** SFT, preference optimization (e.g. DPO) and RL (e.g. GRPO). Use verifiable rewards where a ground truth exists (urgency level, care pathway), and rubric-based rewards from LLM judges where it doesn't.
- **Make answers reliable:** reward abstention and escalation rather than confident guessing, and guard against reward hacking.
- **Extend to clinical reasoning** in multi-turn health assessment, training against our patient simulator.

This topic pairs with How to build an Evaluation Model for Doctolib Health Companion (reward model) and Detecting medical safety issues and turning them into training signals (training data).

Potential Outcome:

A reusable post-training pipeline for Health Companion agents, and at least one post-trained model that beats its prompted baseline on medical accuracy and safety and is tested in production.

DOCTOLIB FOR PATIENTS

Detect Medical safety issues

Context:

Doctolib's Health Companion is a medical conversational assistant used by patients to answer medical questions and guide them on the care pathway. There are however a few cases where the assistant is not able to answer the patient's ask safely. We currently have a mechanism to detect safety issues with the Health Companion, which are then reviewed by our internal doctors. Based on their feedback on safety issues, we then manually make updates to our agentic system (usually the prompts) to improve our assistant's safety.

Mission:

- Improve the detection mechanism of safety issues.
- It is currently a mix of rules and a simple classification system. This will require collaborating with Doctolib's internal medical expertise team, as well as machine learning engineering skills
- Turn the detected issues into training signals.
- After issues are properly detected and sent for doctor review, we collect their feedback and usually act on it manually. The goal is to instead create proper training samples from this feedback, which we could use for post-training our model

Potential Outcome:

- A new system to detect safety issues for the health companion, deployed on our full user base
- A data pipeline turning medical feedback into training samples, used for the different post-training efforts across the Health Companion team



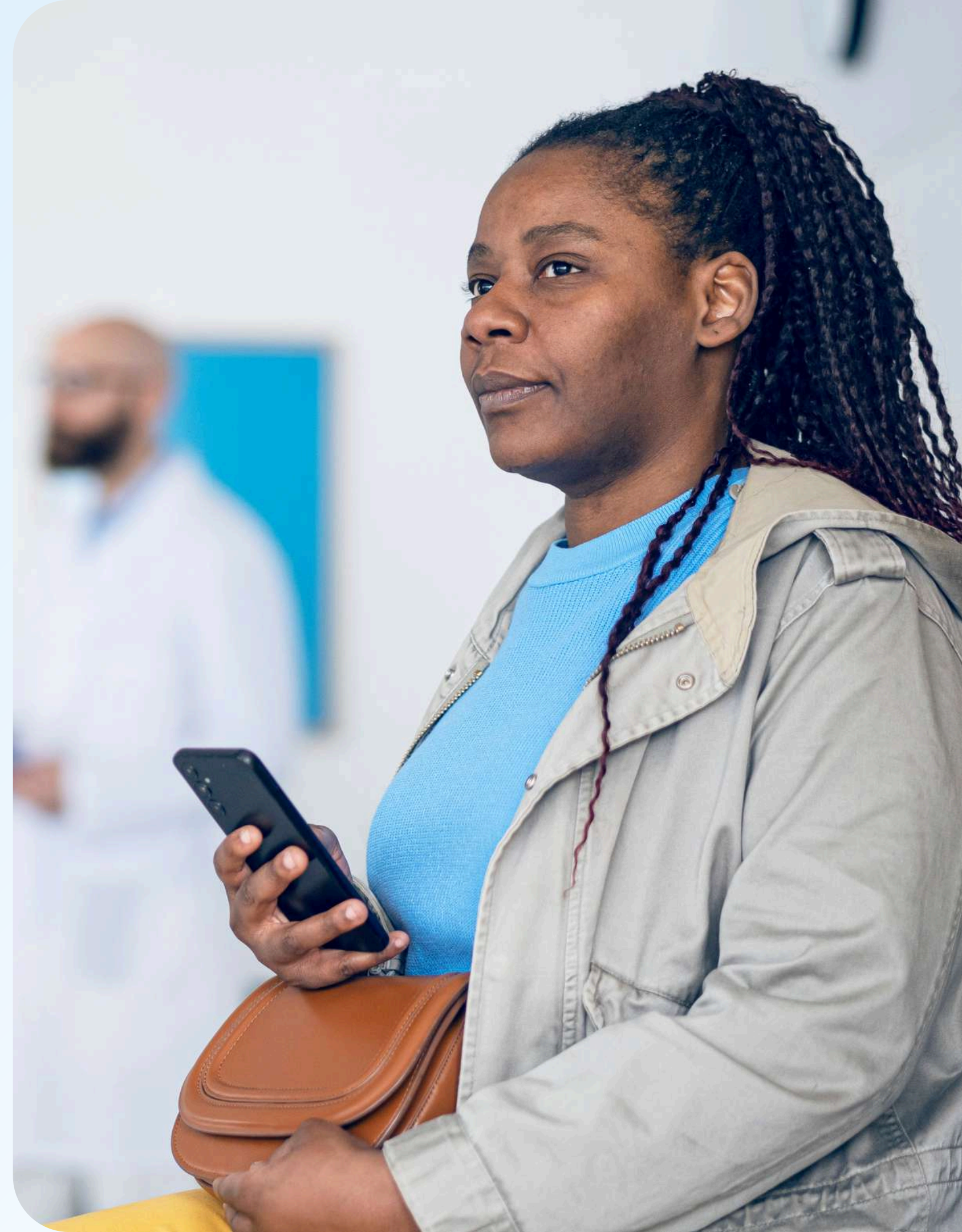
DOCTOLIB FOR PATIENTS

Build an evaluation model leveraging feedback from patients & health pros

Context:

Doctolib has recently released a new medical conversational assistant to guide patients into the appropriate care pathway and answer their health questions. Doctors are currently reviewing sampled conversations to monitor the medical safety of the assistant, while patients can also submit feedback through the app. We would like to leverage feedback from doctors and patients to build an evaluation model that can automatically assess new conversations across several dimensions, such as medical safety and usefulness.

Potential Outcome: An evaluation model aligned with doctors and patients feedback will enable us to scale the online monitoring of the assistant as well as unlock offline learning strategies to improve the assistant.





DOCTOLIB FOR PATIENTS

Learn to deliver health suggestions that support patients

Context:

The Health Companion suggests health actions through push notifications, content and in-chat suggestions, but delivery is tuned for clicks, not for whether patients act or get healthier. Doctolib can observe real outcomes after a suggestion (bookings, attendance, follow-ups), which generic AI chats cannot. Testing every policy on live patients is not an option: randomization needs medical sign-off, and a policy optimized for bookings can cause over-referral or stop reaching patients who respond less.

Mission:

Build and validate a method to estimate from historical logs how much a suggestion changes patient behavior, and how a different delivery policy would have performed. The pilot case is triage → booking.

The intern will:

- Estimate causal impact from logs: define outcomes (including booking quality) and measure the incremental effect of a nudge with observational causal inference.
- Model uplift: identify patients who act only when nudged, and compare uplift approaches (meta-learners, causal forests).
- Evaluate policies off-policy: estimate how new timing or format policies would have performed (IPS, doubly robust), and study when these estimates can be trusted.
- Learn adaptive policies: train contextual bandit or offline RL policies on outcomes instead of clicks, with guardrails on care appropriateness and equity.
- Possible extensions: other health actions (vaccination, screening), a patient-behavior simulator calibrated on observed outcomes, or cross-channel coordination.

Potential Outcome:

A validated offline evaluation and optimization pipeline for health-suggestion delivery, with first evidence on triage → booking and a publishable contribution on outcome-driven policy learning.

DOCTOLIB FOR PATIENTS

Clinical Perception from live audio & video streams

Context:

The Health Companion assesses patients through text chat, with optional photos, so visual and auditory cues that a doctor would notice are lost. Inspired by [AMIE Video \(Palepu et al., 2026\)](#), this internship builds a perception agent. The agent watches the patient's camera and microphone during a consultation and passes grounded findings to the health assessment agents.

Mission:

- define a taxonomy of triage-relevant cues (e.g., skin lesions, swelling, laterality, breathing effort, cough, slurred speech, gait, visible distress) and build a benchmark from synthetic, public and internally recorded clips;
- benchmark multimodal models, fine-tuning where needed;
- design guided examination steps, such as asking the patient to move the camera or compare both sides, and check that each step was performed;
- move from single frames to robust streaming under real-world conditions.

Safety is central throughout. The work will measure and reduce hallucinated observations and laterality errors, and add calibrated confidence and abstention.

This topic pairs with Full-duplex spoken health assessment. The two systems are combined into one audio-video prototype, which the team will compare against the current text and image chat in a study with patient actors and medic review. No real patient data is in scope.

Potential Outcome: A perception agent producing grounded, calibrated findings from live audio and video, with a reusable perception benchmark and a characterisation of failure modes.





DOCTOLIB FOR PATIENTS

Build full-duplex spoken health assessments

Context:

The Health Companion assesses patients through text chat. This is slow for patients and penalises those with limited digital or health literacy. Inspired by [AMIE Video \(Palepu et al., 2026\)](#), this internship turns health assessment into a natural real-time spoken conversation without losing clinical rigour or safety guardrails.

Mission:

- compare end-to-end duplex speech models (e.g., [Moshi](#), Gemini Live, OpenAI Realtime) with a cascaded pipeline built on our in-house ASR and TTS;
- design a talker/planner architecture in which the existing health assessment agents reason in the background while a low-latency agent handles the conversation;
- handle turn-taking, interruptions and silence while still asking every question that matters for triage.

Evaluation runs at scale with a voice-enabled patient simulator built on our vignettes, varying voices, accents and noise. It compares voice with text on triage accuracy, under-triage and red-flag elicitation, as well as latency and conversational quality.

This topic pairs with *Clinical perception from live audio and video streams*. Until the perception agent is ready, the simulator injects visual cues as text, and the two systems are later combined into one audio-video prototype for the team's human evaluation study. The duplex stack should be shared with the Voice team's *Voice-to-Voice conversational AI with task-specific sub-agent orchestration for Phone Assistance* topic. No real patient data is in scope.

Potential Outcome:

A full-duplex spoken health assessment prototype integrated with the existing agents, plus an audio simulation and evaluation pipeline quantifying the effect of voice on medical safety and conversational quality.

DOCTOLIB FOR PATIENTS

Synthetic patient trajectories for training & evaluating AI

Context:

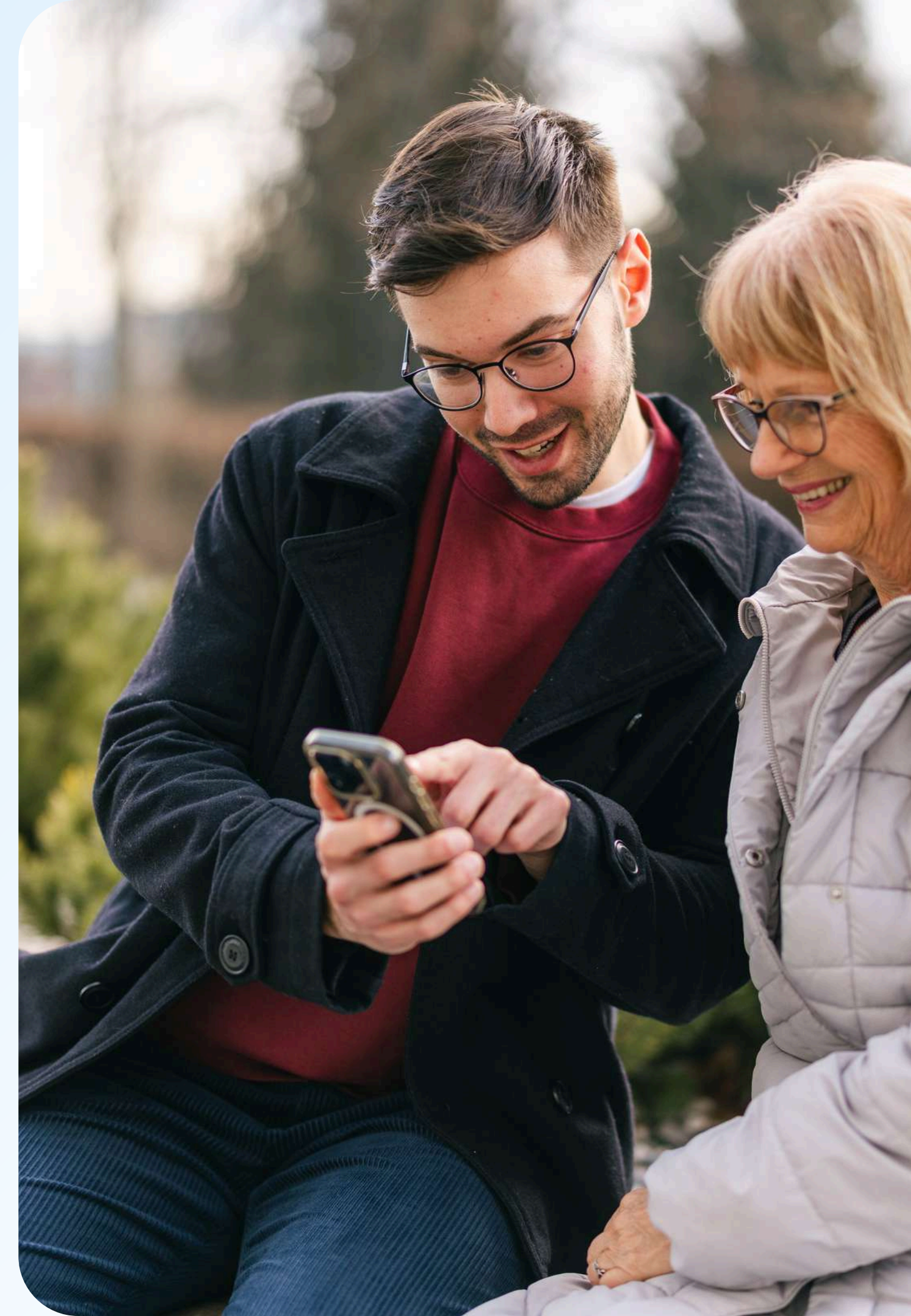
Training and evaluating clinical AI systems requires realistic patient cases. A useful case is rarely a single message: it is a trajectory of symptoms, consultations, investigations, prescriptions and follow-up unfolding over time. Hand-written vignettes are too few and too homogeneous to capture this variability, while real records cannot be circulated to models or teams without privacy risk. This internship studies how to generate synthetic patient trajectories that preserve the statistical and clinical structure of the Doctolib population without being traceable to any individual, learning from de-identified data.

Mission:

- Formalize the vignette: with the medical team, define a trajectory schema (presenting complaint, history, care journey, latent ground truth such as urgency and diagnosis), and build a small annotated reference set
- Model trajectories: learn from heterogeneous de-identified sources (Care events, coded conditions, clinical notes, AI assistant interactions) and compare generation strategies, from rule-based and LLM-based approaches to learned sequence models
- Assess privacy: evaluate each synthetic set against re-identification and membership-inference attacks
- Assess fidelity and coverage: compare synthetic and real distributions (demographics, conditions, languages, rare and high-risk cases)
- Assess utility: use the vignettes to drive a patient simulator, and quantify their contribution to post-training and evaluation relative to hand-written vignettes

Potential Outcome:

A trajectory generation pipeline with a documented privacy evaluation, a reusable set of synthetic patients available to any AI team for training and evaluation, and empirical evidence of its effect on post-training and simulator-based evaluation. The work is a natural basis for publication on privacy preserving synthetic patient generation.





DOCTOLIB EVALUATION

Secure our Loop Engineer that improves the quality of our agentic systems

Context:

At Doctolib, we are building an evaluation product for agentic systems. It continuously tests AI agents against large sets of scenarios, scores their behavior, and flags regressions. The first system it covers is the Health Companion, our AI that talks to patients. The product is designed to serve any agentic system we build. On top of it sits the Loop Engineer, an autopilot. When a test fails, it investigates, explains what went wrong, and proposes a fix as a draft code change for a human to review.

Mission:

It works, but how good is it really? A wrong diagnosis wastes reviewers' time. A fix that only hides the problem, for example by loosening a test, is worse because it makes the evaluated system look healthier than it is. Nobody has yet measured rigorously when such a loop deserves our trust, or how to tune it so that it generalizes from one agentic system to the next. This internship makes the Loop Engineer measurable and tunable.

The intern will:

- Build a benchmark from months of real incidents: the failure, the correct root cause, what the loop proposed, and whether humans accepted it.
- Define and measure "a good recommendation": correct diagnosis, fix that really works, no metric gaming.
- Teach the loop to know what it doesn't know: when to propose a fix, when to ask a human, when to stay silent.
- Compare loop designs (one agent vs. several, with or without memory of past incidents, self-critique) and tune them without overfitting to the past.
- Test generalization: check that a loop calibrated on the Health Companion still gives sensible recommendations on other agentic systems, and identify what must be configured per system and what can be shared.

Everything is tested offline by replaying past incidents and injecting synthetic bugs with known causes, then in shadow mode next to the real loop. No real patient data is involved.

Potential Outcome:

calibrated Loop Engineer integrated in the evaluation product, a reusable incident benchmark, and a measured answer to "how much autonomy does an improvement loop deserve, and does the answer transfer across systems?"

VOICE TECHNOLOGY

Improve the evaluation of our phone assistant conversations

Context:

The Phone Assistant is evaluated offline by running a simulated caller against the real agents and scoring each conversation with LLM judges. Both ends of this loop are weak:

- The simulated caller mostly plays happy paths, because we have little structured knowledge of how real conversations go off script.
- The judges offer us an automated source of truth, but they are often using prompts and decision matrix close to the agent ones. This creates a non null risk of error propagation.

Mission:

The internship would start from production traces and follow two directions.

- Replaying real non-happy paths offline. Extract and characterise the ways production conversations deviate from the happy path, then build the component that replays them in offline simulations. Scenarios could be granular, specific to conversations parts and even assembled to replay a variety of use cases.
- Making judges independent from the agents they evaluate. Explore and measure judge designs rather correlated to business rules than production agents prompts. This could draw on current multi-agent approaches, such as voting across several judges or adversarial debate between them.

Potential Outcome:

A controllable and reliable agentic evaluation suite representative of the variety production observed behaviours and approaching human level annotation trustfulness.



VOICE TECHNOLOGY

Voice-to-voice conversational AI with task-specific sub-agent orchestration for Phone Assistant

Context:

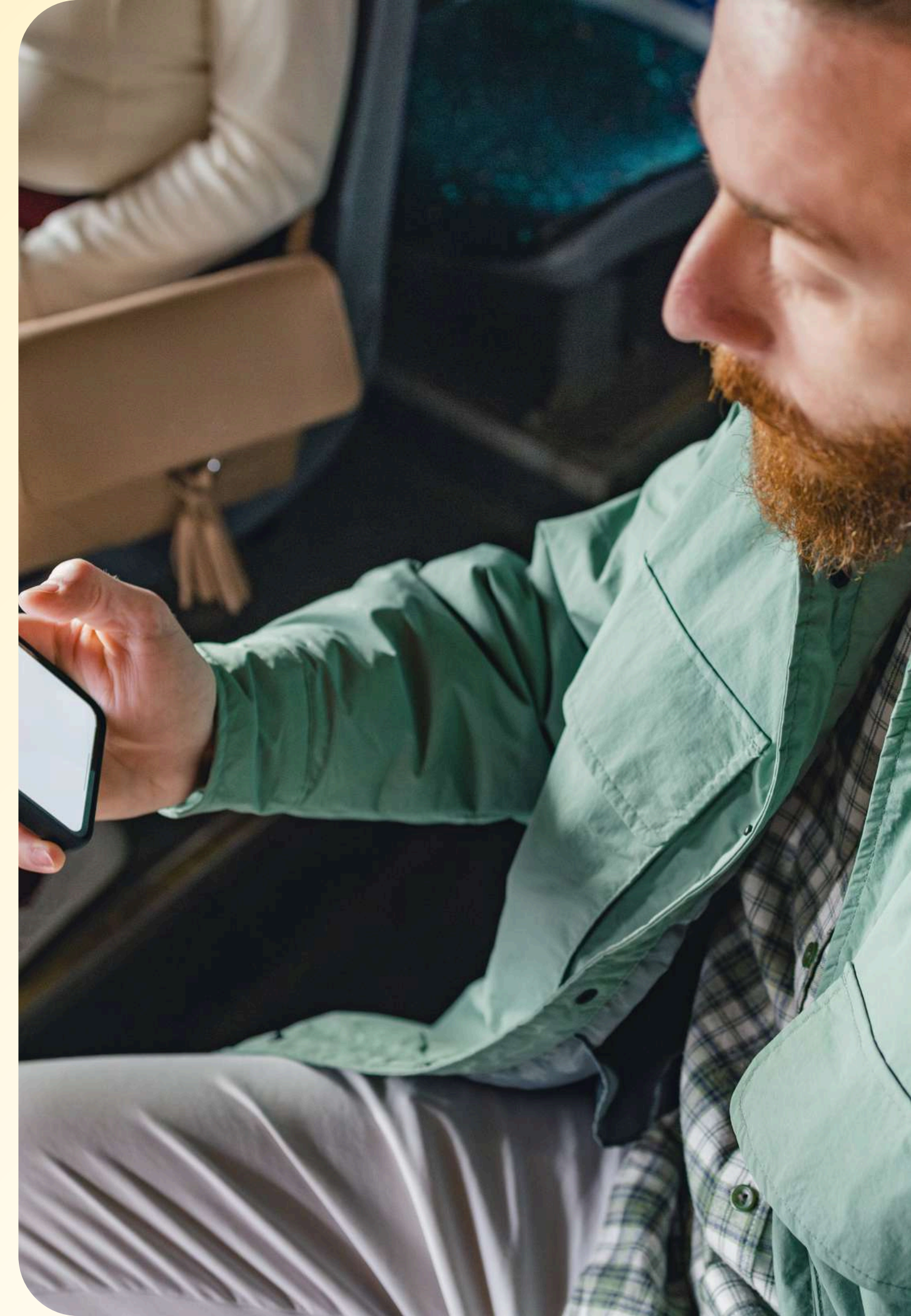
Doctolib's Phone Assistant answers patient calls for healthcare practices and handles tasks such as identifying the caller, booking or cancelling an appointment, and routing the patient to the practice. It runs on a cascaded pipeline: speech-to-text, then an LLM agent, then text-to-speech. That architecture is controllable, but it discards everything the transcript does not carry such as tone, hesitation, urgency, and adds latency at each hop.

Speech-native (voice-to-voice) models process and generate audio directly. They promise lower latency, better turn-taking and more natural handling of interruptions, but their ability to drive reliable agentic behaviour is much less established, especially on telephony-grade audio and in a healthcare setting.

Mission:

Survey the state of the art, prototype a voice-to-voice agent on one Phone Assistant use case, and benchmark it against the current implementation. Integration into the Phone Assistant product is out of scope.

- Literature and model review of speech-native models for conversational and agentic use: architectures, tool-calling support, latency, multi-language support, deployment constraints.
- Characterise the gap: identify and quantify what voice-to-voice captures that traditional voice agent pipelines lose (prosody, emotion, interruptions), and assess which of these can actually change the outcome of a call.
- Design the orchestration: specify how the speech model delegates to task-specific sub-agents (booking, patient identification, routing) while keeping voice continuity. Compare against a hybrid design where a text agent handles the task.
- Prototype a self-contained agent for one mocked use case
- Evaluate and assess trade-offs between voice-to-voice and the traditional pipeline: task success, latency, robustness to telephony-grade audio, instruction adherence, hallucination rate, cost per call, controllability and traceability.



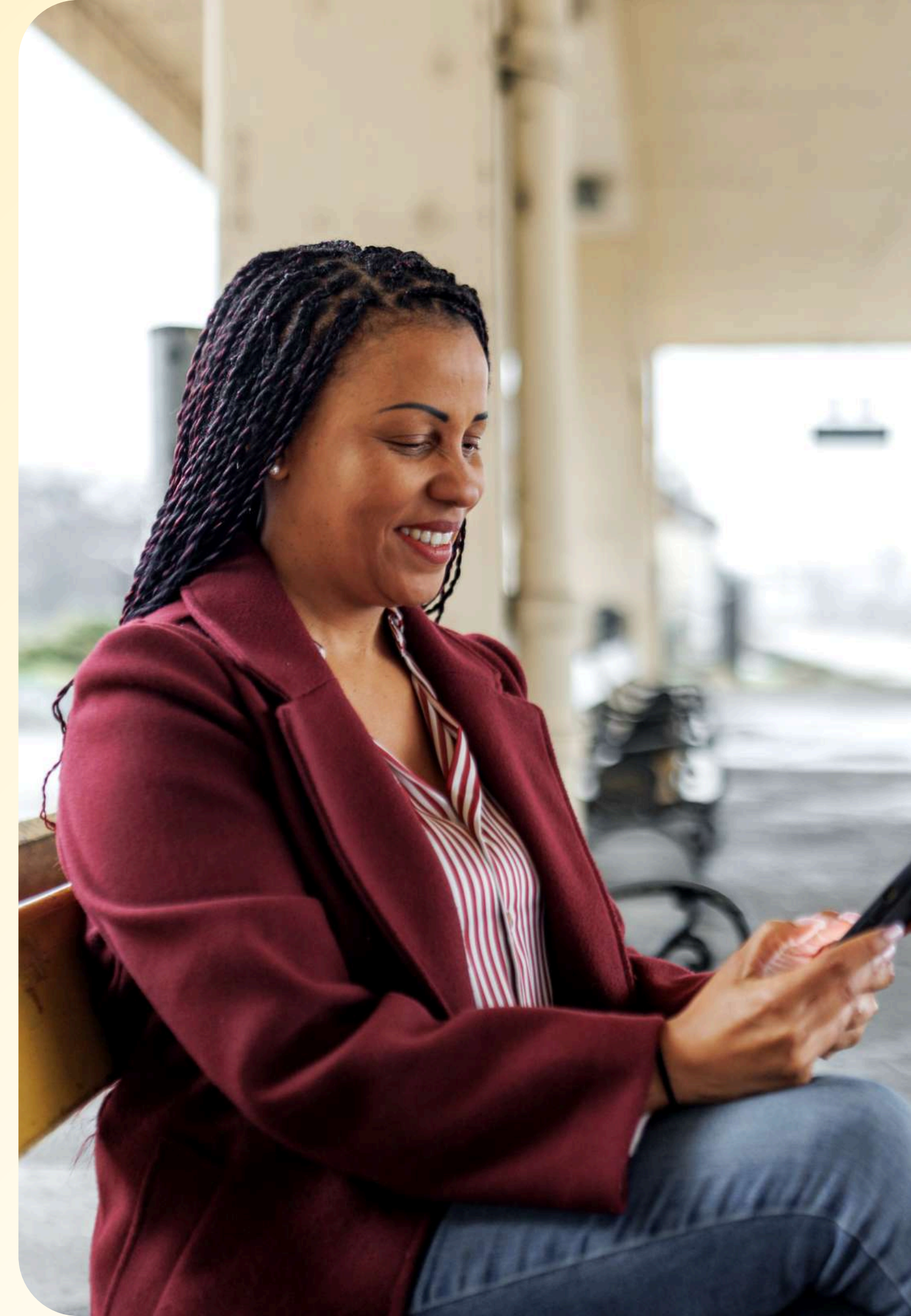
VOICE TECHNOLOGY

Post-training small language models for phone assistant agents

Context:

Doctolib's Phone Assistant relies on many small, specialized agents running under strict latency constraints, currently powered by prompted general-purpose LLMs. The internship will explore whether smaller post-trained models can match or outperform them on agentic tasks such as intent classification, structured data extraction, tool-call optimization for booking, and embedded guardrails. Work includes building datasets from de-identified and synthetic call data, benchmarking candidate models (e.g., Qwen, Gemma, DoctoBERT) against prompted baselines, and comparing post-training strategies (SFT, DPO, RL). The intern will also study how to host post-trained models on our infrastructure within latency and cost budgets.

Potential Outcome: A reusable post-training and evaluation pipeline for PA agents, at least one post-trained agent tested in production, and guidance on model choice, training method, and hosting per agent type.



VOICE TECHNOLOGY

Streaming ASR - evaluation and domain fine-tuning for healthcare voice applications

Context:

Doctolib's Speech team runs ASR in production for three products: Consultation Assistant (CA), Phone Assistant (PA) and Medical Dictation.

Today, only Medical Dictation runs in streaming mode, and it does so with an offline model. CA and PA do not stream, but both could. PA stands to gain the most, since streaming would cut latency in its interactions with the user. Yet most of our evaluation and fine-tuning work focuses on offline, full-utterance decoding.

This internship will benchmark streaming ASR models, adapt them to our data, and evaluate them. It will build a proper evaluation framework for streaming ASR and use it to determine whether streaming brings real benefits to our products. The study will look at how fine-tuned streaming models can lower latency while keeping transcription quality high.

Mission:

- Extend our existing ASR evaluation pipeline to simulate real-time audio feeding and to capture partial and final hypotheses with timestamps. The framework should allow a fair comparison between streaming and offline models (through appropriate audio chunking, for instance).
- Benchmark candidate streaming models against our already-tested offline models.
- Adapt streaming models to our data, in particular medical speech in French and German, through fine-tuning.
- Evaluate the adapted models and compare model configurations with the framework, measuring both quality and latency.
- Assess whether streaming delivers measurable benefits for CA, PA and Dictation.

Potential outcome:

- An evaluation framework for streaming ASR models, integrated into our earlier work on harmonizing the evaluation pipelines for speech-related models.
- A benchmark of streaming models, before and after adaptation to our data, against our offline models, with a clear answer on where streaming pays off in our products.
- If results are conclusive, deployment of a candidate model to production for a shadow run or A/B test.



Doctolib

**We look forward to
meeting you!**

