Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

VozConsultAI: A Provider-Agnostic Architecture and a Clinically Grounded Evaluation of Open-Weight Large Language Models for SOAP Note Extraction in Brazilian Portuguese Telehealth

Nilton Gomes Furtado

nilton.furtado@ua.pt

Universidade de Aveiro https://orcid.org/0000-0002-6372-4693

Julia Vasconcelos Furtado

Universidade de Aveiro https://orcid.org/0000-0002-4335-6272

Case Report

Keywords: clinical documentation, SOAP notes, automatic speech recognition, large language models, telehealth, Brazilian Portuguese, hallucination, evaluation metrics, microservice architecture, digital health

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10255415/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Clinical documentation consumes a substantial share of health professionals' time. It is a well-documented driver of burnout, a problem amplified by the rapid expansion of telehealth in large public health systems such as the Brazilian Unified Health System (SUS). This paper presents VozConsultAI, an end-to-end solution that transcribes teleconsultation audio and automatically structures the resulting text into SOAP (Subjective, Objective, Assessment, Plan) clinical notes in Brazilian Portuguese. We make three contributions. First, we describe an open, provider-agnostic microservice architecture for asynchronous clinical audio processing that abstracts heterogeneous automatic speech recognition providers behind a single interface, offering per-requester FIFO ordering, at-least-once delivery with idempotent processing, and automatic recovery of orphaned jobs. Second, we propose a clinically grounded evaluation protocol for automated SOAP extraction built on three orthogonal dimensions - semantic and structural correctness, information coverage, and clinical grounding - consolidated into a safety-weighted Clinical Documentation Quality Index that explicitly penalizes hallucinated content, a failure mode that traditional classification metrics do not capture. Third, we report what is, to our knowledge, the first comparative benchmark of open-weight large language models for SOAP extraction in Brazilian Portuguese, covering five models across thirty synthetic teleconsultation transcripts from six medical specialties. DeepSeek V3.2 achieved the best consolidated quality (0.87) and the highest clinical grounding (0.96). At the same time, the results show that model scale alone does not explain performance and that the assessment and plan sections account for most errors. We discuss deployment implications for the SUS, including data-sovereignty constraints under Brazilian data-protection law that affect support for on-premises open-weight models.

Categories: C.2.4, D.2.11, H.3.3, I.2.1, I.2.7, J.3

1 Introduction

Approximately 80% of the information stored in electronic health records (EHRs) remains unstructured, mostly narrative text generated during or after medical consultations, thereby limiting its reusability, interoperability, and value for clinical decision support (Li et al., 2022). The problem is particularly acute in telehealth, where the clinician-patient interaction is predominantly spoken, and clinicians must manually convert dialogue into structured records. Evidence consistently shows that physicians may spend more time on documentation than on direct care, and that documentation burden is directly associated with professional burnout (National Academies of Sciences, 2019; Moy et al., 2021).

In Brazil, this challenge acquires a distinctive scale. The Unified Health System (SUS) serves more than 200 million people, and the national digital health strategy has accelerated the adoption of teleconsultation in primary care (Ministério da Saúde, 2020). However, the tooling available for automated clinical documentation remains overwhelmingly Anglophone: commercial ambient scribes target English-speaking markets, and the few domain-adapted language resources for Portuguese clinical text are scarce and fragmented (Schneider et al., 2020). Public institutions face an additional

constraint that commercial cloud scribes do not address well: under the Brazilian General Data Protection Law (LGPD), clinical audio and text are sensitive personal data, and processing them on third-party infrastructure outside institutional control raises legal and sovereignty concerns (Lei No. 13.709, 2018).

This paper presents the design and initial experimental results of VozConsultAI, a project carried out under a technical cooperation agreement (TED No. 31/2024) with the Brazilian Ministry of Health. The system captures consultation audio in the browser, transcribes it through an abstraction layer that can route requests to cloud or self-hosted automatic speech recognition (ASR) engines, and structures the transcript into SOAP (Subjective, Objective, Assessment, Plan) notes (Lisenby et al., 2018) intended for integration with Brazilian public digital health platforms, notably the e-SUS APS ecosystem.

Our contributions are threefold:

1. C1 - an open, provider-agnostic microservice architecture for reliable asynchronous processing of clinical audio, with per-requester FIFO ordering, at-least-once delivery neutralized by idempotent claims, lease-based failure detection, and automatic recovery of orphaned jobs (Section 3);
2. C2 - a clinically grounded evaluation protocol for automated SOAP extraction based on three orthogonal dimensions - Semantic and Structural Correctness (SSC), Information Coverage and Completeness (ICC), and a Clinical Grounding Index (CGI) - consolidated into a safety-weighted Clinical Documentation Quality Index (CDQI) that explicitly penalizes hallucination (Section 5);
3. C3 - the first comparative benchmark, to our knowledge, of open-weight large language models (LLMs) for SOAP extraction in Brazilian Portuguese, covering five models over thirty synthetic teleconsultation transcripts from six specialties, with an analysis of the trade-offs among quality, hardware requirements, licensing, and data sovereignty (Section 6).

The remainder of the paper is organized as follows. Section 2 reviews related work on clinical ASR, digital scribes, SOAP extraction, and evaluation of clinical text generation. Section 3 describes the VozConsultAI architecture. Section 4 presents the three SOAP extraction strategies investigated by the project. Section 5 formalizes the evaluation protocol. Section 6 reports the experimental comparison of open-weight models. Section 7 discusses implications, limitations, and threats to validity, and Section 8 concludes.

2 Background and Related Work

2.1 Documentation burden and digital scribes

The clinical documentation burden and its association with clinician burnout are extensively documented (NASEM, 2019; Moy et al., 2021). Digital scribes - systems that capture the clinician-patient conversation and draft clinical notes automatically - have been proposed as a mitigation strategy for over a decade (Finley et al., 2018; van Buchem et al., 2021). Commercial ambient documentation

products (e.g., Nuance DAX, Abridge, AWS HealthScribe) demonstrate the pipeline's feasibility at scale. Still, they are proprietary, priced per encounter, focused on English, and require routing sensitive audio to vendor clouds - four characteristics that limit their applicability to large public health systems in the Global South (Falcetta et al., 2023).

2.2 ASR for clinical speech

Modern ASR models trained on general-purpose corpora, such as Whisper (Radford et al., 2023), exhibit strong multilingual performance but degrade on clinical speech due to specialized vocabulary, spontaneous conversational patterns, disfluencies, and acoustic variability of real consultation settings. Fine-tuning on domain data improves robustness but depends on the availability of representative corpora; for Brazilian Portuguese, open resources such as CORAA (Candido Junior et al., 2023) cover spontaneous speech but not clinical discourse, which motivated the construction of a dedicated clinical speech corpus within the VozConsultAI project, organized by a taxonomy of fifteen semantic macro-domains and eighteen transversal linguistic classes designed jointly for ASR and TTS purposes.

2.3 SOAP note extraction

Automatic structuring of consultation transcripts into SOAP notes has been addressed with modular summarization (Krishna et al., 2021), sentence-level classification with pre-trained encoders (Devlin et al., 2019), and, more recently, instruction-following LLMs (Ouyang et al., 2022). For Portuguese, de Oliveira et al. (2024) trained domain-specific SOAP classifiers over free-text clinical notes and showed that domain-adapted pre-training outperforms generic models; BioBERTpt provides clinical/biomedical pre-training for Portuguese encoders (Schneider et al., 2020). Benchmark datasets for the English doctor-patient dialogue-to-note task, such as ACI-BENCH (Yim et al., 2023), have no Brazilian Portuguese counterpart; the evaluation resources and results reported here are a step toward filling that gap.

2.4 Evaluating clinical text generation

Traditional evaluation of note generation relies on n-gram overlap (ROUGE/BLEU), classification metrics (F1), or generic semantic similarity, none of which adequately penalizes clinically dangerous failure modes, in particular hallucination - the generation of content unsupported by the source (Ji et al., 2023; Maynez et al., 2020). In clinical applications, plausible-but-wrong content is more harmful than omission, and inference-heavy sections (Assessment, Plan) are the most exposed (Nori et al., 2023). LLM-as-a-judge protocols (Zheng et al., 2023) enable scalable semantic comparison but require careful design and human validation. Our protocol (Section 5) combines atomic clinical-unit decomposition, an auxiliary LLM judge with conservative equivalence criteria and human validation, and an explicit grounding dimension.

2.5 Positioning

In summary, prior work either (i) targets English with proprietary cloud pipelines, (ii) addresses isolated components (ASR or classification) rather than a deployable end-to-end architecture, or (iii) evaluates with metrics blind to hallucination. VozConsultAI differs in that it combines an open architecture

designed for public-sector deployment constraints (LGPD, heterogeneous providers, horizontal scale) with a hallucination-aware evaluation protocol and a benchmark of open-weight models for Brazilian Portuguese.

3 The VozConsultAI Architecture

The VozConsultAI pipeline covers audio capture in the browser during the teleconsultation, asynchronous transcription, SOAP structuring, and delivery of the draft note for professional review. This section describes the transcription layer - the Transcription Broker - which is the architectural core of the system (Fig. 1). The design premise is to decouple the requester (the telehealth application) from the executor (an ASR engine), so that slow, costly processing happens outside the synchronous request path, with reliable result delivery via signed webhooks.

3.1 Design principles

Five principles govern the design: (1) decoupling - the external boundary (broker) is isolated from orchestration (gateway) and execution (worker); (2) horizontal scalability - multiple workers consume partitioned queues, so adding capacity means adding consumers; (3) per-requester ordering - messages from the same requester are processed FIFO through a consistent-hash exchange; (4) at-least-once delivery combined with idempotent operations, so duplicated messages are harmless; and (5) automatic recovery - jobs claimed by a worker that crashes are detected through lease expiry and republished without manual intervention.

3.2 Components

The broker is the only externally exposed service and acts as a backend-for-frontend: it authenticates callers (API keys with argon2-hashed secrets for applications; JWT for human administrators), uploads the audio artifact to S3-compatible object storage (MinIO), resolves the task type to a destination gateway, and keeps a local copy of job state so that status queries do not traverse the core. The gateway is the internal orchestrator and single source of truth for job lifecycle: it persists jobs, publishes them to a RabbitMQ consistent-hash exchange keyed by requester identifier, consumes results, and dispatches HMAC-SHA256-signed webhooks. The worker encapsulates the ASR engine behind a generic process(payload, artifactUri) contract - the reference implementation runs Whisper (Radford et al., 2023), but the contract is engine-agnostic. Workers claim jobs through an idempotent conditional UPDATE that acquires a time-limited lease, renews the lease via heartbeat during processing, and acknowledges the message only after the result is persisted. The dispatcher is a scheduled job that scans for orphaned jobs - those whose lease expired while still in processing state - and either republishes them (until a per-task maximum number of attempts) or closes them as failed, still triggering the webhook so the client is always notified.

3.3 Provider abstraction

A central function of the broker is to hide provider heterogeneity. Each supported transcription service - Google Cloud Speech-to-Text, Amazon Transcribe, Azure AI Speech, OpenAI Whisper API, AssemblyAI, or the self-hosted Whisper worker - is integrated through an adapter that translates the canonical request into the provider-specific API and normalizes the response into the project's canonical format. Provider selection is a broker-side policy configurable by cost, latency, accuracy, language support (including Brazilian Portuguese and speaker diarization), and resilience (fallback on unavailability). The application never learns which provider served a request, thereby eliminating vendor lock-in and enabling A/B comparisons of engines in production. The self-hosted path also keeps audio within the institutional infrastructure, which is decisive for LGPD compliance (Section 3.5).

3.4 Reliability guarantees

Correctness properties emerge from the combination of durable messaging, database leases, and idempotency. Per-requester FIFO follows from consistent hashing on the requester identifier: all messages of one requester land in the same partitioned queue. At-least-once delivery follows from durable queues and persistent messages; duplication - expected both from broker republication and from RabbitMQ redelivery - is neutralized by the idempotent claim, whose second execution matches zero rows and acknowledges silently. Webhook processing on the broker side is likewise idempotent: a job already terminated returns a no-op. Failure scenarios and mitigations were analyzed systematically: worker crash (lease expiry plus dispatcher), message-bus restart (durability), broker unavailability during webhook (negative acknowledgment and retry, with lazy state reconciliation as fallback), and permanent worker loss (bounded attempts, terminal failure with notification).

3.5 Security and data sovereignty

All inter-service trust boundaries are explicit: API-key authentication at the external edge, HMAC-signed webhooks on the return path, and per-service database roles with least privilege. Because clinical audio is sensitive personal data under the LGPD (Lei No. 13.709, 2018), the architecture treats data residency as a first-class routing criterion: the same pipeline can serve a fully on-premises deployment (self-hosted Whisper plus local LLM extraction) or a hybrid deployment where non-sensitive workloads use cloud providers. This flexibility motivated the evaluation focus on open-weight models in Section 6, since only open-weight models can seed a local inference engine.

3.6 Processing flows and data model

Four flows cover the job lifecycle. Submit: at the end of a teleconsultation, the application posts the recorded audio to the broker, which validates the credential, stores the artifact in object storage, delegates the task to the gateway, and immediately returns a job identifier - the consultation is never blocked waiting for processing. Processing: a worker claims the job via an idempotent conditional update, downloads the audio, runs the ASR engine, and persists the result before acknowledging the message, so a mid-processing crash never loses the job. Webhook: the gateway consumes the result and notifies the broker with an HMAC-signed payload; from then on, client status queries are answered from the broker's local state without a round trip to the core. Recovery: on each scheduled execution, the

dispatcher selects expired-lease jobs with row-level locking that skips contended rows, republishing or terminally failing them within a transaction that publishes before committing.

The data model deliberately uses two independent logical databases with no foreign keys between them - one owned by the broker (organizations, users, API keys, task-type catalog, and a local job mirror) and one shared by gateway, worker, and dispatcher (task types with default lease and attempt parameters, and the job table with webhook URL, lock ownership, lease expiry, and attempt counters). The only coupling is the job UUID, which lets each side evolve independently. The worker intentionally has no database of its own: it writes job state directly to the gateway's database, maintaining a single source of truth without an extra round of communication.

4 SOAP Extraction Strategies

The SOAP extractor receives the consultation transcript and produces the four-component structure: Subjective (patient-reported complaints and history), Objective (measurable clinical data), Assessment (clinical interpretation and diagnosis), and Plan (interventions and follow-up) (Lisenby et al., 2018). The project investigates three complementary strategies, chosen to span the trade-off space between computational cost, data requirements, and quality.

4.1 Prompt-engineered extraction

The first strategy adapts a pre-existing instruction-following LLM through prompt engineering (Wang et al., 2025; Patil et al., 2024). The prompt (i) defines the task explicitly over the four SOAP categories; (ii) fixes a structured JSON output schema to enable downstream persistence; (iii) contextualizes the model for medical language, abbreviations, and Brazilian clinical terminology; (iv) instructs the model to extract only information present in the transcript, with explicit mechanisms for empty or unclassifiable content; and (v) applies uniform classification criteria across transcripts to guarantee comparability. A hard requirement of the project is local executability, which excludes proprietary cloud-only models from the candidate set; a frontier commercial model is retained only as an external reference baseline.

4.2 Hybrid classification plus generation

The second strategy segments the transcript into sentences, classifies each sentence into a SOAP category with a fine-tuned BERT-family encoder (Devlin et al., 2019; de Oliveira et al., 2024), and then passes the per-category groups to a generative model that consolidates them into a coherent, readable note. The discriminative stage offers precise, auditable categorization; the generative stage normalizes language and removes redundancy. The open questions this strategy must answer empirically are the cost of building the sentence-level training set for Portuguese clinical dialogue and whether the two-stage pipeline justifies its additional computational cost over the single-stage approaches.

4.3 Fine-tuned generation

The third strategy fine-tunes a local generative model on transcript-to-SOAP pairs, initially synthetic (avoiding privacy issues during development) and progressively enriched with real consultation data collected under the project's ethics framework. Parameter-efficient methods - LoRA and QLoRA (Hu et al., 2022; Dettmers et al., 2023) - keep the computational cost of adaptation compatible with the hardware envelope of public institutions. This strategy addresses a structural weakness of locally deployed models: publicly released weights often lack the full alignment stack of their hosted counterparts, and quantization required for modest hardware degrades fluency (Ouyang et al., 2022), both of which domain fine-tuning can partially compensate.

5 A Clinically Grounded Evaluation Protocol

Failures of a SOAP extractor are not homogeneous in clinical risk: misclassifying a correctly captured finding is less dangerous than omitting it, and both are less dangerous than inventing one. The protocol therefore evaluates three orthogonal questions: Is what was generated correct and well-structured? Was everything important covered? Was anything invented? - and consolidates them into a single safety-weighted index. All definitions operate over clinical units: semantically atomic clinical items (a symptom, a finding, a diagnosis, an action) obtained by decomposing both the reference annotation and the model output.

5.1 Metrics

Semantic and Structural Correctness (SSC). For each generated component, each unit is matched against the reference components, producing a correspondence matrix; the exact-support rate of a component is the fraction of its units supported by the corresponding reference component (the diagonal of the matrix), averaged over transcripts, and the overall SSC averages the four components. SSC captures whether generated content is semantically correct and assigned to the appropriate SOAP section, including distinctions between semantically close categories such as diagnostic assessment and therapeutic plan.

Information Coverage and Completeness (ICC). The dual matrix is computed in the opposite direction: each reference unit is matched against the generated components. A reference unit not found anywhere in the generated note is an omission; the ICC of a component is the complement of the mean omission rate, and the overall ICC averages the components. A model can achieve perfect SSC while capturing half of the relevant information; ICC exposes exactly this failure.

Clinical Grounding Index (CGI). For each generated unit, the protocol searches the original transcript for an explicit textual span that semantically supports it; absence of explicit support cannot be resolved by inference and marks the unit as ungrounded - a potential hallucination (Ji et al., 2023; Maynez et al., 2020). The CGI of a component is the fraction of grounded units, averaged over transcripts and components. CGI shifts evaluation from structural correctness to fidelity, addressing the safety, ethics, and regulatory dimension that overlap- or F1-based metrics ignore.

Clinical Documentation Quality Index (CDQI). For model ranking, the three dimensions are consolidated in two forms: a safety-oriented weighted mean assigning weight 4 to CGI and 3 each to SSC and ICC, reflecting that hallucination is the most critical failure mode in this domain, and a conservative form defined as the minimum of the three dimensions, which strongly penalizes any single severe weakness.

Formally, let each reference annotation and each generated annotation for a transcript t consist of four component sets of clinical units. For SSC, each unit of a generated component c-hat is compared against every reference component, yielding a correspondence matrix per transcript whose diagonal entries count exactly supported units; the exact-support rate of c-hat in t is the ratio of its diagonal count to its size. The SSC of a component is the mean of that ratio over the n transcripts. Overall, SSC is the mean over the four components. For ICC, the dual-matrix counts for each reference component c are the units whose information does not appear anywhere in the generated annotation (i.e., a general absence of coverage). ICC of c is one minus the mean absence ratio, and overall ICC is the mean over components. For CGI, each generated unit is checked for explicit support in the transcript; the per-component index is the fraction of supported units, averaged over transcripts and then over components. A component set may legitimately be empty (for example, no objective findings in a brief triage call), and the protocol treats empty sets without penalty on either side.

5.2 Operationalization

The protocol is operationalized in three phases, combining human annotation with an auxiliary LLM judge (Zheng et al., 2023). Phase 1 (segmentation): an LLM proposes decomposing reference and generated notes into clinical units, which humans then validate. Phase 2 (semantic alignment): an auxiliary judge - a different model from the one under evaluation, run at temperature zero - receives the reference unit, the generated unit, and an explicit conservative definition of semantic equivalence, returning a binary decision. Phase 3 (grounding verification): the judge searches the transcript for explicit supporting evidence for each generated unit. The full protocol is itself the reference implementation for the project's cloud evaluation service, and every decision is auditable.

AI-tool disclosure. In line with COPE guidance and the journal's AI policy, we disclose that large language models were used as components of the evaluated system and as auxiliary judges inside the evaluation protocol, as described above; specific model names and versions are listed in Section 6.

6 Experimental Evaluation

6.1 Setup

We evaluated five open-weight models spanning distinct families, scales, licenses, and hardware envelopes (Table 1): gpt-oss-120b (OpenAI, 2025), Llama 3.3 70B Instruct (Meta, 2024), Mistral Small and Mistral 7B Instruct v0.3 (Mistral AI, 2024), and DeepSeek V3.2 (DeepSeek-AI, 2025). Open-weight availability was a hard inclusion criterion: only models whose weights can be downloaded may seed the project's on-premise inference engine and satisfy the sovereignty requirements discussed in Section 3.5.

Proprietary frontier models accessible only through cloud APIs were surveyed but not prioritized, as they cannot serve the local engine. All models received the same prompt-engineered extraction task (Section 4.1). The test set comprises 30 synthetic teleconsultation transcripts in Brazilian Portuguese covering six specialties: general practice, pediatrics, gynecology and obstetrics, psychiatry, cardiology, and infectious diseases. Reference SOAP annotations follow the protocol of Section 5, with human-validated segmentation and a temperature-zero auxiliary judge for alignment and grounding.

Table 1
Open-weight models included in the benchmark.

Model	Developer (country)	License	Scale	Hardware (indicative)	PT-BR support
gpt-oss-120b	OpenAI (USA)	Apache 2.0	117B MoE (~ 5.1B active)	1x 80 GB GPU	limited (EN-centric)
Llama 3.3 70B Instruct	Meta (USA)	Llama Community	70B dense	~ 2x 80 GB GPU	official
Mistral Small	Mistral AI (EU)	Apache 2.0	intermediate	multi-GPU (medium)	yes
Mistral 7B Instruct v0.3	Mistral AI (EU)	Apache 2.0	7B dense	1x consumer GPU	partial
DeepSeek V3.2	DeepSeek (China)	MIT	671B MoE (~ 37B active)	multi-GPU (large)	partial (validated)

6.2 Results

Tables 2–4 report the per-component and overall scores for the three dimensions, while Table 5 and Fig. 2 consolidate them into the CDQI. Table 2 presents the correctness results: on SSC, DeepSeek V3.2 and Llama 3.3 tied for the best overall score (0.75), with small differences across models, indicating that structural classification is largely within reach of all candidates; the Assessment component was uniformly the weakest (0.40–0.58), as it demands clinical inference.

Table 2
Semantic and Structural Correctness (SSC) by SOAP component. Coverage, as reported in Table 3, discriminates the models far more sharply than correctness does. On ICC, DeepSeek V3.2 achieved 0.88, and gpt-oss-120b achieved 0.84, while Mistral 7B dropped to 0.56, exposing the limitations of smaller models in capturing the totality of clinically relevant information.

Component	gpt-oss-120b	Mistral-7B	Llama 3.3	Mistral Small	DeepSeek V3.2
Subjective	0.62	0.79	0.81	0.65	0.72
Objective	0.93	0.90	0.90	0.89	0.93
Assessment	0.58	0.40	0.47	0.41	0.54
Plan	0.79	0.83	0.80	0.80	0.80
Overall SSC	0.73	0.73	0.75	0.69	0.75

Table 3
Information Coverage and Completeness (ICC) by SOAP component.

Component	gpt-oss-120b	Mistral-7B	Llama 3.3	Mistral Small	DeepSeek V3.2
Subjective	0.77	0.28	0.60	0.70	0.86
Objective	0.97	0.62	0.89	0.94	0.96
Assessment	0.81	0.76	0.85	0.79	0.80
Plan	0.81	0.56	0.66	0.72	0.91
Overall ICC	0.84	0.56	0.75	0.79	0.88

Table 4
turns to grounding, the safety-critical dimension. On CGI, DeepSeek V3.2 led with 0.96, followed by gpt-oss-120b with 0.92; grounding of the Objective component was near-perfect for all models, whereas the Assessment component again concentrated the ungrounded content.

Component	gpt-oss-120b	Mistral-7B	Llama 3.3	Mistral Small	DeepSeek V3.2
Subjective	0.91	0.87	0.91	0.91	0.97
Objective	0.99	0.97	0.98	1.00	1.00
Assessment	0.86	0.65	0.70	0.62	0.87
Plan	0.91	0.97	0.93	0.95	0.99
Overall CGI	0.92	0.87	0.88	0.87	0.96

Table 4: *Clinical Grounding Index (CGI) by SOAP component.*

Table 5
consolidates the three dimensions into the CDQI under both aggregation forms, confirming the same ordering of models under the safety-weighted mean and, under the conservative form, revealing that Llama 3.3 matches DeepSeek V3.2 once each model is judged by its weakest dimension.

Model	CDQI (safety-weighted)	CDQI (conservative)
DeepSeek V3.2	0.87	0.75
gpt-oss-120b	0.84	0.73
Llama 3.3 70B	0.80	0.75
Mistral Small	0.79	0.69
Mistral 7B v0.3	0.74	0.56

Table 5: *Consolidated Clinical Documentation Quality Index (weights: CGI 4; SSC 3; ICC 3; conservative = minimum dimension).*

Figure 2 offers a visual synthesis of the same results, making immediately apparent that the models differ far more in coverage (ICC) and grounding (CGI) than in classification (SSC), and that the consolidated index tracks the grounding dimension closely.

6.3 Analysis

Three findings stand out. First, scale is not destiny: DeepSeek V3.2, the largest model, led the consolidated ranking, but Llama 3.3 matched its SSC and its conservative CDQI at a fraction of the parameter count, indicating that architecture, training quality, and alignment matter as much as size. Second, the discriminating dimensions are coverage and grounding, not classification: inter-model variance concentrated in ICC and CGI, confirming that assigning a sentence to the right SOAP section is easier than being complete and honest about it - and justifying an evaluation protocol that separates these dimensions instead of collapsing them into a single F1. Third, inference-heavy components concentrate risk: Assessment and Plan exhibit both the lowest correctness and the lowest grounding, exactly where clinical extrapolation beyond available evidence is most tempting (Nori et al., 2023); drafts for these sections require the closest professional review.

Beyond quality, deployment criteria differentiate the candidates. All five models are installable on-premises, satisfying LGPD-driven sovereignty requirements; they differ in hardware envelope (from a consumer GPU for Mistral 7B to a robust multi-GPU setup for DeepSeek V3.2) and licensing (Apache 2.0 and MIT are fully permissive; the Llama license imposes usage conditions). For the project's local engine, the pragmatic recommendation is a model-agnostic serving layer - already provided by the broker - keeping a top-quality candidate (DeepSeek V3.2) and a best-balance candidate (gpt-oss-120b or Llama 3.3 70B) simultaneously supported.

A qualitative reading of the error patterns reinforces the quantitative picture. Ungrounded units in the Assessment section typically consist of diagnostic refinements that are clinically plausible given the symptoms but never stated by the professional - precisely the failure mode that CGI is designed to expose and that an overlap metric would often reward, since the invented diagnosis shares vocabulary with the dialogue. Omissions are concentrated in the Subjective section of the smaller models, which tend to compress multiple reported complaints into a single summary, thereby omitting clinically relevant negatives (denied symptoms). Misclassifications between Assessment and Plan - a therapeutic hypothesis recorded as a decided intervention - are the dominant SSC error across all models and carry practical risk, as they may convert a possibility into an apparent prescription in the record.

7 Discussion

7.1 Implications for public digital health

For the SUS, characterized by continental scale, regional diversity, and high assistential demand, automated SOAP drafting has direct potential to reduce administrative burden, standardize records, and

improve care continuity; higher-quality structured records additionally strengthen epidemiological surveillance, clinical audit, and secondary use of data for research and management (Ministério da Saúde, 2020). The architecture's provider abstraction has a policy dimension: it lets a public institution switch between self-hosted engines and commercial clouds by configuration, converting a legal constraint (data sovereignty) into a routing criterion rather than a system redesign.

7.2 Relation to commercial ambient scribes

Commercial ambient documentation services demonstrate that the audio-to-note pipeline is commercially viable. Still, their unit economics and data flows are poorly matched to a universal public system: per-encounter pricing multiplied by the SUS consultation volume is prohibitive, English-centric models underperform on Brazilian Portuguese clinical discourse, and routing sensitive audio through foreign clouds conflicts with the sovereignty posture required of public health data controllers under the LGPD (Falcetta et al., 2023; Lei No. 13.709, 2018). The combination reported here - open-weight models at CDQI 0.87 on institutional hardware, behind a broker that can still reach commercial engines when policy allows - suggests that a public system need not choose between capability and control. The open evaluation protocol also gives procurement a lever that vendor benchmarks do not: hallucination-aware acceptance criteria that can be written directly into contracts.

7.3 Limitations and threats to validity

Internal validity. Reference annotations and judge decisions involve an auxiliary LLM. However, the judge differs from the evaluated models, runs at temperature zero under conservative equivalence criteria, and has human-validated segmentation; residual judge bias cannot be excluded (Zheng et al., 2023). Construct validity. CDQI weights encode a specific safety prioritization; we mitigate arbitrariness by also reporting the conservative (minimum) form and all raw dimensions. External validity. The benchmark uses 30 synthetic transcripts; synthetic dialogue underrepresents the disfluency, code-switching, and acoustic noise of real teleconsultations, and n = 30 limits statistical power - we deliberately refrain from significance claims and present results as a first reproducible baseline. Validation on real teleconsultation data collected under the project's ethics framework, enlargement of the test set, and inclusion of Portuguese-native models (e.g., Sabia-family) are the immediate next steps. Finally, the end-to-end effect of ASR errors on downstream SOAP quality was not yet measured; the present evaluation isolates the extraction stage.

8 Conclusions and Future Work

This paper presented VozConsultAI, an open, provider-agnostic architecture for automated clinical documentation in Brazilian Portuguese telehealth, together with a clinically grounded evaluation protocol and the first comparative benchmark of open-weight LLMs for SOAP extraction in this language. The results show that open-weight models already reach useful quality levels (CDQI up to 0.87, grounding up to 0.96) under strict sovereignty constraints, that coverage and hallucination - not sentence classification

- are the real technical frontier. That assessment and plan drafting must remain under close professional review.

Future work proceeds on four fronts: (i) fine-tuning a local model on the project's growing clinical corpus of Brazilian Portuguese, including specialization in medical jargon; (ii) validating the protocol and the selected models on real teleconsultation recordings with health professionals, including structured feedback collection; (iii) measuring the compound effect of the full pipeline (ASR plus extraction) rather than the extraction stage in isolation; and (iv) integrating the system with the e-SUS APS platform and the V4H telehealth infrastructure, closing the loop from consultation audio to reviewed clinical record.

Declarations

Ethical Approval

The experimental evaluation reported in this paper used exclusively synthetic transcripts and involved no human participants or identifiable personal data. The project's subsequent phases involving real consultation recordings are conducted with the approval of the competent research ethics committee.

Conflicts of Interest

The authors declare that no competing interests exist.

Funding

This work was carried out under the Decentralized Execution Agreement TED No. 31/2024 between the Brazilian Ministry of Health and the Federal Institute of Education, Science and Technology of Paraiba (IFPB).

Acknowledgements

The authors thank the VozConsultAI project team for the technical deliverables that ground Sections 3–6 of this paper.

Data Availability

The system components are available as open repositories of the VozConsultAI organization (broker-api, vozai-gateway, vozai-worker, vozai-dispatcher, text-to-soap) at github.com/VozConsultAI.

References

1. Candido Junior A, Casanova E, Soares A et al (2023) CORAA ASR: A large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. Lang Resour Evaluation 57:1139–1171. https://doi.org/10.1007/s10579-022-09621-4
2. DeepSeek-AI (2025) DeepSeek-V3.2 model card. Hugging Face. https://huggingface.co/deepseek-ai
3. de Oliveira JM, Antunes RS, da Costa CA (2024) SOAP classifier for free-text clinical notes with domain-specific pre-trained language models. Expert Syst Appl 245:123046. https://doi.org/10.1016/j.eswa.2023.123046
4. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2023) QLoRA: Efficient finetuning of quantized LLMs. Adv Neural Inf Process Syst, 36
5. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186). Association for Computational Linguistics
6. Falcetta FS, de Almeida FK, Lemos JCS, Goldim JR, da Costa CA (2023) Automatic documentation of professional health interactions: A systematic review. Artif Intell Med 137:102487. https://doi.org/10.1016/j.artmed.2023.102487
7. Finley G, Edwards E, Robinson A et al (2018) An automated medical scribe for documenting clinical encounters. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 11–15). Association for Computational Linguistics
8. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: Low-rank adaptation of large language models. In the International Conference on Learning Representations (ICLR 2022)
9. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM-CSUR 55(12):1–38. https://doi.org/10.1145/3571730
10. Krishna K, Khosla S, Bigham J, Lipton ZC (2021) Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)* (pp. 4958–4972). Association for Computational Linguistics
11. Lei nº 13 (2018) Diário Oficial da União. Brasília, Brasil, 709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais)
12. Li I, Pan J, Goldwasser J, Verma N, Wong S, Nuzumlalı MY, Rosand B, Li Y, Zhang M, Chang D, Taylor RA, Krumholz HM, Radev D (2022) Neural natural language processing for unstructured data in electronic health records: A review. Comput Sci Rev 46:100511. https://doi.org/10.1016/j.cosrev.2022.100511
13. Lisenby KM, Andrus MR, Jackson CW, Stevenson TL, Fan S, Gaillard P, Carroll DG (2018) Ambulatory care preceptors' perceptions on SOAP note writing in advanced pharmacy practice experiences

(APPEs). Currents Pharm Teach Learn 10(12):1574–1578

14. Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906–1919). Association for Computational Linguistics
15. Meta (2024) Llama 3.3 70B Instruct model card. Hugging Face. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
16. Ministério da Saúde (2020) *Estratégia de Saúde Digital para o Brasil 2020–2028*. Ministério da Saúde
17. Mistral AI (2024) *Model documentation and weights*. Hugging Face. https://huggingface.co/mistralai
18. Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, Rossetti SC (2021) Measurement of clinical documentation burden among physicians and nurses using electronic health records: A scoping review. J Am Med Inform Assoc 28(5):998–1008. https://doi.org/10.1093/jamia/ocaa325
19. National Academies of Sciences, Engineering, and Medicine (2019) Taking action against clinician burnout: A systems approach to professional well-being. National Academies. https://doi.org/10.17226/25521
20. Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of GPT-4 on medical challenge problems. arXiv. https://doi.org/10.48550/arXiv.2303.13375
21. OpenAI (2025) GPT-OSS-120 B and GPT-OSS-20 B model cards. Hugging Face. https://huggingface.co/openai/gpt-oss-120b
22. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 35:27730–27744
23. Patil R, Heston TF, Bhuse V (2024) Prompt engineering in healthcare. Electronics 13(15):2961. https://doi.org/10.3390/electronics13152961
24. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2023) Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023) (pp. 28492–28518). PMLR
25. Schneider ETR, de Souza JVA, Knafou J et al (2020) BioBERTpt: A Portuguese neural language model for clinical named entity recognition. In Proceedings of the 3rd Clinical Natural Language Processing Workshop (pp. 65–72). Association for Computational Linguistics
26. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW (2021) The digital scribe in clinical practice: A scoping review and research agenda. npj Digit Med 4:57. https://doi.org/10.1038/s41746-021-00432-5
27. Wang J, Shi E, Yu S et al (2025) Prompt engineering for healthcare: Methodologies and applications. Meta-Radiology 3:100190
28. Yim W, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M (2023) ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Sci Data 10:586.

https://doi.org/10.1038/s41597-023-02487-3

29. Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing EP, Zhang H, Gonzalez JE, Stoica I (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, p 36

Figures

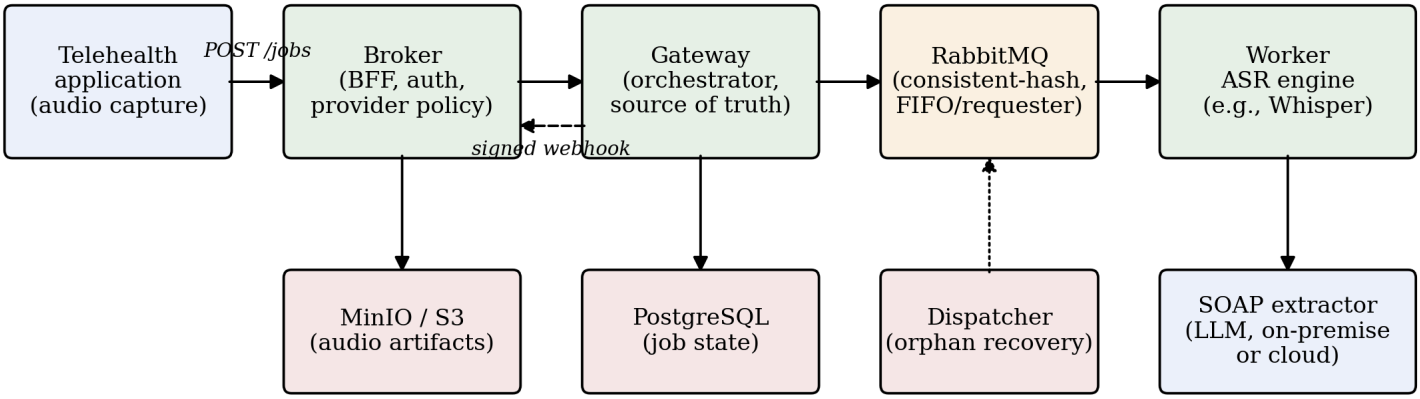


Figure 1

High-level architecture of the VozConsultAI transcription layer. Solid arrows represent the request path; the dashed arrow represents the signed webhook return; the dotted arrow represents republication of orphaned jobs

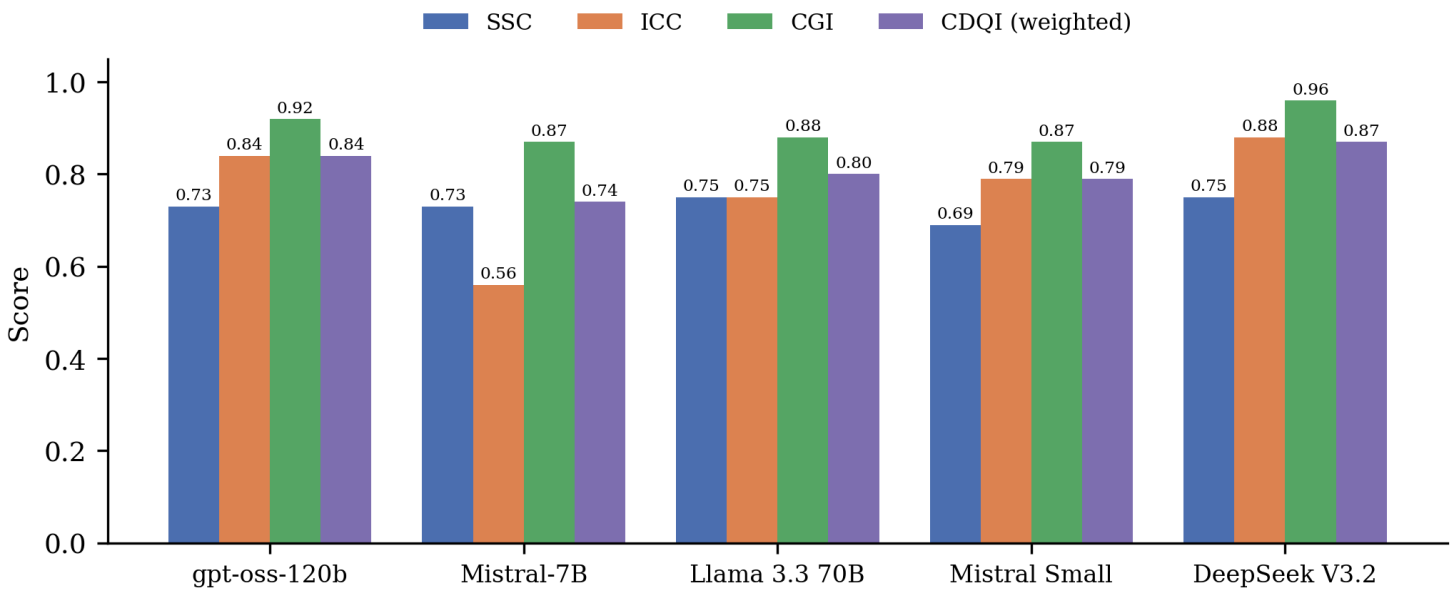


Figure 2

Overall scores per model on the three dimensions and the safety-weighted CDQI

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- VozConsultAITables.xlsx