Extending Epuskesmas With A Domain-Specific Large Language Model For Orthopaedic Documentation of Osteoarthritis And Osteoporosis: A Proof-Of-Concept Study Toward Indonesian Primary Health Care Strengthening

Ahmad Azmul A. Irfan[1*], Nur Ahmad Khatim[2,3], Achmad Zaki[1], Bisatyo Mardjikoen[1], Mansur M. Arief[3,4], Amril Nur Ismail[5], Ahmad Farhap[6]

[1]Orthopaedic Department, Faculty of Medicine UIN Syarif Hidayatullah Jakarta, Indonesia
[2]Division of Information Science, NAIST (Nara Institute of Science and Technology), Japan
[3]Artificial Intelligence Verification and Validation Lab, King Fahd University of Petroleum and Minerals, Saudi Arabia
[4]Industrial and Systems Engineering, King Fahd University of Petroleum and Minerals, Saudi Arabia
[5]General Practiotioner, uskesmas Banggae II, Indonesia
[6]Undergraduate Student, Faculty of Medicine UIN Syarif Hidayatullah Jakarta, Indonesia
Corresponding author:
E-mail : ahmad.azmul@uinjkt.ac.id

Abstract.
Background. Clinical documentation is a major driver of workload in primary care, and Indonesia's mandated transition to electronic medical records has increased the recording burden on community health centres (Puskesmas). A recent proof-of-concept study showed that a browser-based pipeline combining automatic speech recognition (ASR) with large language model (LLM) summarisation can convert Bahasa Indonesia doctor-patient conversations into ePuskesmas fields for general primary care. Musculoskeletal complaints, particularly knee osteoarthritis and osteoporosis, are common, disabling and largely manageable at primary level, yet they require documentation elements that a general-purpose template does not explicitly capture. Objective. To develop a domain-specific orthopaedic extension of the ePuskesmas LLM documentation framework and to evaluate the clinical adequacy of its generated documentation through independent clinician review. Methods. Six scripted roleplay consultations covering knee osteoarthritis and osteoporosis-related conditions were recorded in Bahasa Indonesia, transcribed with a Whisper model, and summarised with an LLM using an orthopaedic-specific prompt mapped to ePuskesmas fields. Transcripts and structured outputs were assessed independently by three practising clinicians (two general practitioners working in Puskesmas and one orthopaedic specialist) against a seven-domain rubric scored 1-5, yielding 126 ratings. Inter-rater reliability was quantified using the intraclass correlation coefficient (ICC). Results. Across 126 ratings the overall mean was 3.29 of 5 (range 2-5), corresponding to the rubric anchor acceptable (usable after moderate editing). Agreement between reviewers was high: ICC(2,k) = 0.96, with all three reviewers assigning an identical score on 73.8% of items and agreeing to within one scale point on every item. Domain means were highest for appropriateness of referral recommendation (4.00) and appropriateness of Puskesmas-level management (3.83), and intermediate for red-flag identification (3.44). Scenario means ranged from 4.19 (suspected fragility fracture) to 2.57 (moderate knee osteoarthritis with obesity and gastritis). Qualitative review identified one materially unsafe analgesic recommendation, systematic loss of pertinent negative findings, incomplete transfer of examination detail, over-triage of one chronic case and conflation of fracture risk with established diagnosis. Conclusion. A domain-specific ePuskesmas LLM is technically feasible and produces documentation drafts of acceptable but not clinically final quality. Performance was adequate for management and referral recommendations, whereas fidelity to the source consultation was the limiting factor.

Keywords: *large language models; electronic medical records; clinical documentation; primary health care; osteoarthritis; osteoporosis; Indonesia; ePuskesmas and clinical decision support.*

I. INTRODUCTION

Documentation has become one of the most time-consuming activities in clinical practice, and ambient artificial intelligence (AI) scribes have emerged as a leading strategy to reduce it. Early real-world deployments report meaningful reductions in perceived documentation effort and in burnout, while consistently emphasising that clinician editing remains necessary before a note is finalised [2,3,4].

In Indonesia the question is immediate rather than theoretical. Ministry of Health Regulation No. 24 of 2022 mandated electronic medical records across health facilities and required integration with the national SATUSEHAT health-information exchange [21], obliging community health centres (Puskesmas) to record structured data during consultations that are often very short. An assessment of information and commu-

nication technology (ICT) maturity across nine provinces found overall maturity between basic and good, with marked regional variation in human resources, software, hardware and infrastructure [22]. Any documentation support tool must therefore operate within real constraints of connectivity, hardware and staffing.

Khatim et al. recently reported a proof-of-concept system that records a doctor–patient conversation in the browser, transcribes it with a Whisper model, summarises it with a GPT-3.5 LLM, and populates structured ePuskesmas fields, demonstrating feasibility with physician validation [1]. That work established an end-to-end pipeline and an important technical foundation, but deliberately targeted general primary care documentation rather than any specific clinical domain.

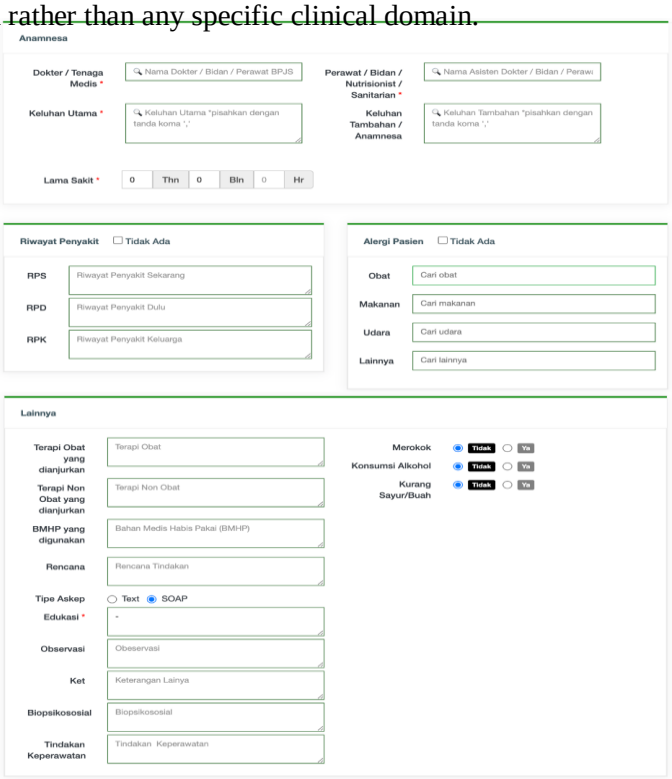

Fig. 1. The current ePuskesmas consultation interface. Free-text and structured fields that the clinician must complete manually during or after the encounter are indicated, together with the subset of fields that the proposed system is intended to pre-populate as an unverified draft.

Musculoskeletal conditions constitute a substantial clinical burden in Indonesia and are frequently encountered in primary care [24]. Osteoarthritis affects a large and rapidly growing population, with prevalence projected to rise substantially through 2050 as populations age and obesity increases [14,15]. Osteoporosis and fragility fractures impose a parallel burden in older adults, and a first fragility fracture is a strong predictor of subsequent, more disabling fractures [17,18,19]. Rehabilitation and functional care for musculoskeletal conditions are recognised as a health-systems priority for universal health coverage [20].

Both conditions are non-communicable diseases whose burden grows with population ageing, and fragility fractures generate a preventable cascade of further fractures. Puskesmas are the backbone of Indonesian primary health care and the entry point to the tiered referral system. Structured, guideline-concordant documentation at this level is therefore not only a record-keeping task but a determinant of appropriate gatekeeping, of preventive care delivery, and, in aggregate, of the quality of population health intelligence available to the health system [21,22].

Orthopaedic documentation differs from general documentation in specific ways. It should capture functional limitation, a structured red-flag screen, the elements of a musculoskeletal examination, pertinent negative findings, a conservative management plan, and explicit referral criteria and urgency. None of these is guaranteed by a general-purpose output template, which motivates domain-specific adaptation.

The present study extends the proof-of-concept of Khatim et al. [1] by adapting the original framework into a domain-specific orthopaedic module focused on osteoarthritis and osteoporosis. Where the earlier study provides the technical foundation, the present work contributes the domain adaptation: an orthopaedic ePuskesmas template, an evidence-grounded education and referral layer, and a structured clinical evalu-

ation. The study aims (i) to develop an orthopaedic-specific ePuskesmas documentation template for osteoarthritis and osteoporosis; (ii) to evaluate whether LLM-generated summaries capture clinically relevant information from simulated Bahasa Indonesia consultations, including orthopaedic history, functional limitation, red-flag symptoms and management plan; and (iii) to assess the clinical appropriateness of the generated documentation through structured, independent review by general practitioners and an orthopaedic specialist. A secondary objective examines the capacity of the system to support initial management at Puskesmas level, particularly patient education and referral decisions. As an exploratory objective, the study considers whether structured orthopaedic documentation could contribute to broader public health functions, including secondary prevention and population-level monitoring.

The system is intended as a documentation and workflow support tool, not an autonomous diagnostic system. Every generated output is a draft requiring physician verification before it enters the medical record.

Related Work

LLMs for clinical text summarisation.

Adapted LLMs can summarise clinical text to a standard comparable with, and in some tasks preferred over, that of human experts across radiology reports, progress notes and doctor–patient dialogue [5]. These results derive largely from English-language, high-resource settings and from curated text rather than from noisy spoken input in a second language.

Language models over electronic health records

Domain-adapted clinical language models such as GatorTron show that scale and clinical-domain pretraining improve downstream clinical natural-language tasks [6]. Critical appraisals nevertheless warn that foundation models for electronic health records rest on shaky foundations: they are frequently evaluated on tasks that do not reflect deployment conditions, and their usefulness to health systems remains inadequately demonstrated [7].

Documentation burden and ambient scribes

Deployments of ambient AI scribes report reduced documentation burden and improved clinician experience, alongside the consistent finding that clinicians must review and edit drafts before signing [2,3,4]. The present work belongs to this paradigm but is specialised to a low-resource, non-English, domain-specific context.

Safety, hallucination and clinical validation

LLMs can produce fluent but factually incorrect content, a well-characterised failure mode in natural-language generation [9]. Although LLMs encode substantial clinical knowledge [8], safe clinical use requires explicit safety evaluation, and frameworks now exist specifically to assess clinical safety and hallucination rates of LLMs used for medical text summarisation [10]. ASR contributes a second, independent error source: Whisper-family models can generate fluent transcription content with no basis in the audio, with harms concentrated in speech that departs from training-distribution norms [11,12]. Indonesian clinical speech is frequently code-switched, regionally variable and accented, and recognition performance degrades under exactly these conditions [25,26,27]. In such a setting the two error sources compound.

Digital health and health-system strengthening in Indonesia

A systematic scoping review of AI for health systems in low- and middle-income countries highlights limited local validation, infrastructure constraints and governance gaps [13]. In Indonesia specifically, the electronic medical record mandate and SATUSEHAT integration [21], combined with heterogeneous ICT maturity across provinces [22], define the deployment envelope for any documentation tool: it must tolerate limited connectivity, variable hardware and constrained staffing.

The original ePuskesmas framework

This study builds directly on the proof-of-concept of Khatim et al. [1], which demonstrated end-to-end feasibility of the record, transcribe, summarise, map workflow for general Bahasa Indonesia primary care documentation. To our knowledge no prior work has adapted that framework to a specific clinical domain. The present study addresses that gap for orthopaedic primary care.

III. METHODS

Study design and setting

This was a proof-of-concept study using simulated roleplay consultations designed to emulate the Indonesian Puskesmas workflow. No real patients were involved; all consultations used simulated patients. The study assessed technical feasibility and the clinical appropriateness of generated documentation, not patient outcomes.

System architecture: inherited implementation and orthopaedic extension

The module is built on the implementation described by Khatim et al. [1] and retains its four components: recording, transcription, summarisation and form population. The contribution reported here is a domain layer added at each of those stages. Inherited specification and orthopaedic addition are therefore described together below, stage by stage, so that the two can be read against each other.

Conversation recording. Audio is captured in the browser using the JavaScript MediaRecorder API, with start and stop controls and an option to upload an existing file. Capture is at a 44.1 kHz sampling rate using the browser's default codec, and audio is held locally in chunks until processing is initiated; the reference implementation uses Google Chrome and can be adapted to other browsers through their corresponding function calls. This layer includes no noise reduction, speaker diarisation or audio quality assessment. The orthopaedic addition here is procedural rather than technical: the recording protocol requires the clinician to verbalise examination procedures and findings, because a musculoskeletal record depends on examination content that an audio-only pipeline cannot otherwise obtain.

Transcription. Speech-to-text conversion uses the Whisper model accessed through the OpenAI API, configured for Indonesian, with audio automatically resampled to 16 kHz during preprocessing. Whisper was selected for its multilingual coverage and explicit support for Bahasa Indonesia, although its performance across Indonesian dialects and regional variation has not been systematically characterised and its accuracy on Indonesian medical terminology remains unevaluated. Because the model is accessed as a hosted service, transcription requires a stable internet connection. In the deployed workflow the clinician reviews and corrects the transcript for speech-recognition errors, patient-identity inconsistencies and inaccurate clinical content before summarisation [12].

Summarisation. The transcript is processed by GPT-3.5-turbo with a structured system prompt instructing the model to act as a medical transcriber familiar with Indonesian healthcare terminology and ePuskesmas documentation requirements, at a generation temperature of 0.2 to favour focused and consistent output, and emitting JSON with predefined keys. This stage carries most of the orthopaedic extension. The prompt replaces the eight output categories of the original implementation with the thirteen-field orthopaedic set and parallel SOAP note defined in Section 3.4, and directs the model to extract functional limitation, red flags, pertinent negative findings, musculoskeletal examination findings and relevant comorbidities. An Indonesian evidence layer then produces two outputs that have no counterpart in the general framework: guideline-grounded patient education, and a tiered referral recommendation specifying indication, urgency, rationale and intended destination. Both are conditioned on the scope of services available at Puskesmas level and on relevant medication-safety factors.

ePuskesmas integration. The final component populates form fields through JavaScript DOM manipulation, mapping JSON keys to form input identifiers so that no manual re-entry is required, with error handling for missing fields and formatting inconsistencies. The mapping is specific to the ePuskesmas interface version tested and does not survive interface updates. The orthopaedic extension adds the fields of the dedicated orthopaedic section to this mapping and labels every generated field as an unverified draft. The attending physician must review, edit, accept or reject each field, including its clinical evidence, medication safety, diagnostic interpretation, patient education and referral urgency, before any content enters the medical record.

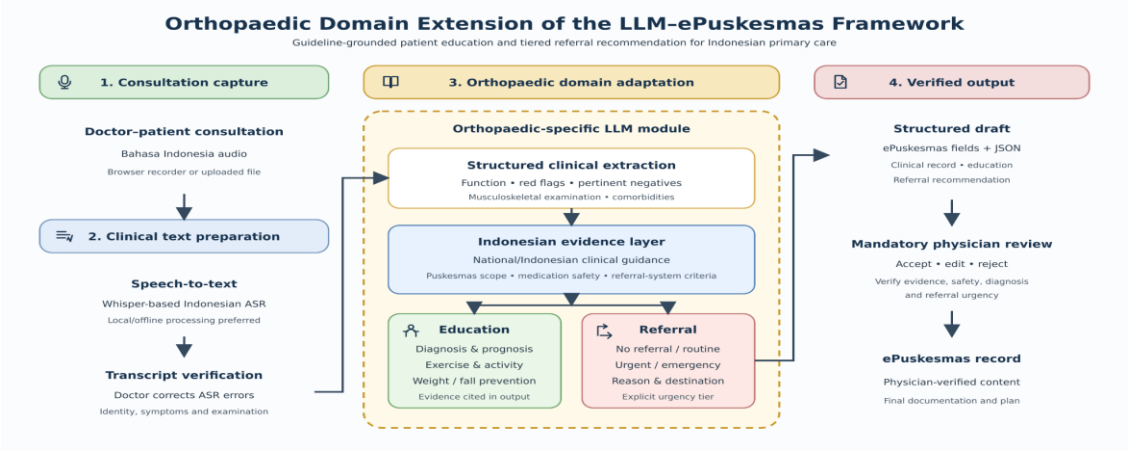

Fig. 2. Proposed orthopaedic domain extension of the LLM-based ePuskesmas framework. Recorded consultations are converted into transcripts, orthopaedic clinical information is extracted, and Indonesian evidence and guidelines are used to generate patient education and referral recommendations. All outputs require physician verification before integration into ePuskesmas.

Deployment as a browser extension

The framework was implemented as a browser extension operating alongside the existing ePuskesmas web interface, so that no modification of the underlying system is required and the clinician's established workflow is preserved. The technical pipeline was inherited from the original framework [1].

Fig. 4. Browser extension for ePuskesmas, showing the extension panel, the recording control and consent indicator, the generated draft awaiting review, and the mechanism by which the clinician accepts, edits or rejects each field before it is written into ePuskesmas.

Orthopaedic ePuskesmas template and record section

We developed an orthopaedic-specific documentation template that retains the core ePuskesmas fields defined by Khatim et al. [1] and adds an orthopaedic extension so that functional status, red flags, examination prompts and referral logic are captured explicitly. Output fields were: *Keluhan Utama; Keluhan Tambahan; Riwayat Penyakit Sekarang; Riwayat Penyakit Dahulu; Riwayat Penyakit Keluarga; Pemeriksaan Fisik yang Perlu Dicatat; Asesmen; Terapi Obat; Terapi Non-Obat; Edukasi; Rencana Tindak Lanjut; and Keterangan: (Indikasi Rujukan)*. A SOAP note (subjective; objective, including the recommended musculoskeletal examination; assessment; plan) was specified as a parallel output. Template design reflected primary-care-oriented management principles from contemporary osteoarthritis guidance, including education, exercise, weight management and staged, comorbidity-aware pharmacotherapy [15,16]. Osteoporosis content covered risk assessment, calcium and vitamin D, fall prevention, bone mineral density evaluation and fracture recognition [17,18,19,23].

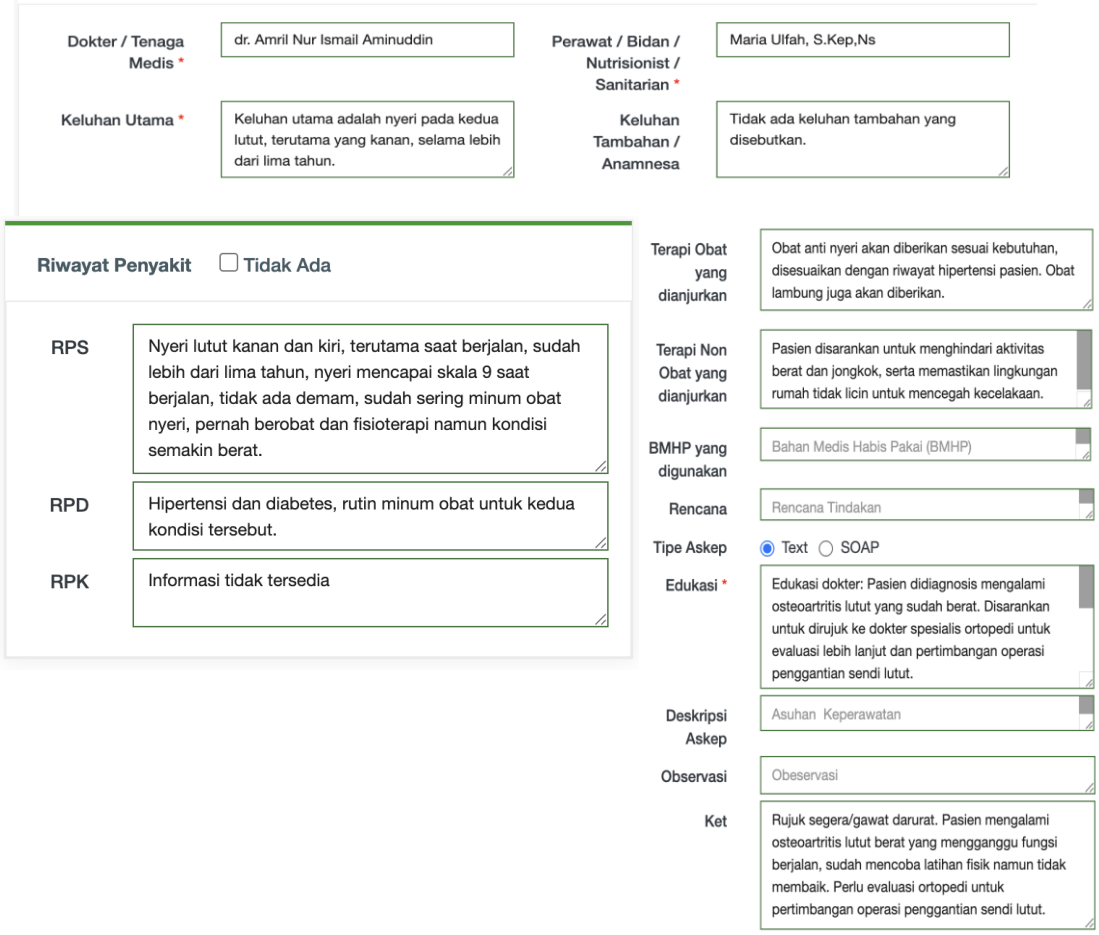

Fig. 5. Additional orthopaedic section of ePuskesmas, the novel contribution of this work, showing the orthopaedic-specific fields: functional status, red-flag checklist, conservative management plan and referral tier.

Process indicators and interoperability considerations

To explore the public health dimension, we defined process indicators derivable directly from the generated documentation: explicit red-flag screening; retention of pertinent negative findings; documentation of functional status; prevention and education counselling; weight-management counselling where body mass index was elevated; fall-prevention guidance in older adults; osteoporosis case-finding or secondary-fracture-prevention flagging; and assignment of a referral tier. As a design consideration, structured JSON output could in principle be mapped from ePuskesmas toward SATUSEHAT to support interoperable aggregated reporting [21,22]. This mapping was conceptual only and was neither implemented nor tested.

Ethical, privacy and governance considerations

Because the study used simulated patients, no patient-identifiable data were processed. The design nonetheless embeds governance requirements for real deployment: explicit patient consent to audio recording; minimisation and secure handling of audio, with a preference for local or offline processing within the Puskesmas so that sensitive audio is not transmitted externally [13]; clear labelling of every output as an unverified draft; and mandatory physician verification before any content enters the medical record.

Simulated Roleplay Experiment

1. Scenario design

Each scenario specified a simulated patient profile, learning objectives, a ready-to-record Bahasa Indonesia dialogue, the expected ePuskesmas output, the red flags to be screened, Puskesmas-level management and a referral decision. Mandatory history elements in every dialogue were pain location, duration and severity, functional limitation, trauma, fever, swelling, neurological symptoms, fall history, medication history, comorbidities and patient expectations. The six scenarios are summarised in Table 1.

Table 1. Overview of the six simulated orthopaedic scenarios.

	Simulated case	Key profile	Red-flag status	Intended disposition
S1	Mild knee OA	F, 55 y; mechanical pain ~6 mo; BMI ≈24.8	N/A	Managed at Puskesmas; review 2–4 wk
S2	Moderate knee OA with obesity	M, 62 y; BMI ≈32.8; hypertension (amlodipine), gastritis	Negative (acute)	Puskesmas care; elective referral if fails
S3	Severe knee OA	F, 68 y; cane-dependent; failed conservative care	Functional (chronic)	Planned orthopaedic referral
S4	Post-menopausal osteoporosis risk	F, 60 y; low BMI; low calcium intake; maternal hip fracture	Negative(no fracture)	Elective BMD/DXA referral; prevention
S5	Suspected fragility fracture	F, 70 y; low-energy fall; wrist swelling and deformity	**Positive**	**Urgent** referral
S6	Suspected vertebral compression fracture	F, 72 y; acute back pain after light lifting; height loss, kyphosis	**Positive**	Referral for imaging

OA, osteoarthritis; BMI, body mass index (kg/m²); BMD, bone mineral density; DXA, dual-energy X-ray absorptiometry.

Illustrative pipeline output

To illustrate pipeline behaviour we present a single representative sample from Scenario 1. The Bahasa Indonesia transcript is shown exactly as produced by the speech-to-text stage, without correction, because its recognition errors are themselves part of the evidence; an English translation is supplied for readers, although the system operates on the Bahasa Indonesia text.

Generated summary Scenario 1 (Bahasa Indonesia, as returned by the model)

{"chief_complaint": "Nyeri lutut kanan selama 6 bulan.",
"additional_complaint": "Tidak ada informasi tambahan.",
"history_of_present_illness": "Nyeri lutut kanan sisi dalam, terutama saat jalan jauh atau naik turun tangga, skala nyeri 4-5/10, kadang terdapat kaku pagi selama 10-15 menit, tidak mengganggu aktivitas sehari-hari.",
"past_medical_history": "Tidak ada riwayat penyakit lain, hanya kadang masuk angin.",
"family_history": "Informasi tidak tersedia.",
"recommended_medication_therapy": "Paracetamol untuk nyeri lutut, krim anti nyeri jika perlu.",
"recommended_non_medication_therapy": "Modifikasi aktivitas fisik, hindari naik turun tangga berlebihan, kompres dengan air hangat, jaga berat badan.",
"education": "Osteoartritis lutut ringan, modifikasi aktivitas fisik, hindari beban berat pada lutut, kompres dengan air hangat, jaga berat badan, kontrol jika gejala memburuk.",
"referral_recommendation": "Tidak perlu dirujuk",
"evidence_based_patient_education": "Osteoarthritis supported; Joint protection, weight management, low-impact aerobic activities, flexibility and muscle-strengthening exercises, avoid high-impact exercises, warning signs for worsening symptoms, follow-up in two weeks; Pedoman Nasional Pelayanan Klinis Tata Laksana Osteoartritis, 2025."
}

The corresponding generated ePuskesmas fields recorded mechanical right knee pain of approximately six months' duration, brief morning stiffness, crepitus, preserved function, a negative red-flag screen, an assessment of suspected mild knee osteoarthritis, paracetamol as required with optional topical non-steroidal anti-inflammatory drug (NSAID), quadriceps strengthening and activity modification, patient education, review at two to four weeks, and no current referral indication.

Validation Framework

1. Reviewers and procedure

Generated outputs for all six scenarios were assessed independently by three practising clinicians: two general practitioners providing care in Puskesmas settings, representing the intended primary-care users of the system, and one orthopaedic specialist, representing domain-expert judgement. Reviewers were selected

purposively for relevant clinical experience and were not involved in developing the template, the prompt or the roleplay scripts.

Each reviewer received the transcript and the structured output for every scenario, scored them against the rubric, and provided free-text comments per scenario together with an overall recommendation. Reviewers were instructed not to confer before completing all scores. Scoring used a workbook containing the rubric anchors, with entries from 1 to 5.

Reviewers assessed the transcript and the structured ePuskesmas output only. A separate SOAP note was not available at the time of review, and the SOAP component specified in Section 3.2 was therefore not evaluated in this study.

Statistical analysis

Seven domains were scored on a 1–5 Likert scale (1 = poor; 2 = fair; 3 = acceptable, usable after moderate editing; 4 = good; 5 = excellent): transcription clarity; completeness of orthopaedic history; accuracy of clinical summary; red-flag identification; appropriateness of Puskesmas-level management; appropriateness of referral recommendation; and usefulness for ePuskesmas documentation. Each reviewer produced 42 scores (7 domains × 6 scenarios), giving 126 ratings in total.

Scores are reported descriptively as means and ranges by domain, by scenario and by reviewer. Given the small number of reviewers and the purposively constructed scenario set, no hypothesis testing was performed. Inter-rater reliability across the three reviewers was quantified using a two-way random-effects, absolute-agreement intraclass correlation coefficient, reported for a single rater, ICC(2,1), and for the mean of three raters, ICC(2,k). These were supplemented by the proportion of items on which all three reviewers assigned an identical score and the proportion on which they agreed to within one scale point.

IV. RESULTS

All three reviewers completed all assigned items, giving 126 ratings with no missing data. The overall mean was 3.29 of 5, corresponding to the rubric anchor *acceptable, usable after moderate editing*. No rating of 1 (poor) was assigned. Ratings of 2 accounted for 17.5% of scores, 3 for 46.0%, 4 for 27.0% and 5 for 9.5%; 82.5% of ratings were 3 or above.

Agreement between the three reviewers was high. All three assigned an identical score on 31 of 42 items (73.8%), and on every item the three scores fell within a single scale point (42 of 42, 100%). The mean within-item range was 0.26 points. The intraclass correlation coefficient was 0.886 for a single rater, ICC(2,1), and 0.959 for the mean of three raters, ICC(2,k), indicating good to excellent reliability. The pooled domain and scenario means reported below can therefore be read as stable estimates rather than as the impression of an individual reviewer.

Scores by domain

Domain means are shown in Table 2. Performance was strongest for appropriateness of referral recommendation (4.00) and appropriateness of Puskesmas-level management (3.83), and weakest for accuracy of clinical summary (2.89), with transcription clarity, completeness of orthopaedic history and usefulness for ePuskesmas documentation all at 2.94. The orthopaedic reviewer scored at or below the general practitioners in every domain.

Table 2. Mean validation scores by domain (1-5 Likert scale; six scenarios, three reviewers; n = 126 ratings).

Scoring domain	GP 1	GP 2	GP pooled	Orthopaedic specialist	Combined mean	Range
1. Transcription clarity	3.17	2.83	3.00	2.83	2.94	2–4
2. Completeness of orthopaedic history	3.17	3.00	3.08	2.67	2.94	2–4
3. Accuracy of clinical summary	3.00	3.00	3.00	2.67	2.89	2–4
4. Red flag identification	3.50	3.50	3.50	3.33	3.44	2–5
5. Appropriateness of Puskesmas-level management	3.83	3.83	3.83	3.83	3.83	2–5
6. Appropriateness of referral recommendation	4.17	4.00	4.08	3.83	4.00	2–5

Scoring domain	GP 1	GP 2	GP pooled	Orthopaedic specialist	Combined mean	Range
7. Usefulness for ePuskesmas documentation	3.00	3.00	3.00	2.83	2.94	2–4
Overall	**3.40**	**3.31**	**3.36**	**3.14**	**3.29**	**2–5**

Scores by scenario

Per-scenario means are shown in Table 3. The suspected fragility fracture scenario (S5) scored highest (4.19), reflecting correct recognition of low-energy trauma with deformity and appropriate same-day triage. The lowest score was for moderate knee osteoarthritis with obesity and gastritis (S2, 2.57), driven by the medication-safety error described in Section 6.3. The orthopaedic reviewer scored severe knee osteoarthritis (S3) markedly lower than the general practitioners (2.57 versus 3.14), reflecting stricter expectations for documentation of deformity, range of motion, failure of conservative therapy and referral tier.

Table 3. Per-scenario mean scores across all domains and reviewers.

Scenario	S1	S2	S3	S4	S5	S6
All reviewers	3.67	2.57	2.95	3.05	4.19	3.29
GP pooled	3.71	2.64	3.14	3.07	4.21	3.36
Orthopaedic specialist	3.57	2.43	2.57	3.00	4.14	3.14

Qualitative findings

(i) A materially unsafe medication recommendation

In S2 the transcript documented that the patient had experienced gastric pain after taking strong over-the-counter analgesics, had known gastritis and was taking amlodipine for hypertension, and that the consulting doctor explicitly avoided oral NSAIDs in favour of paracetamol and topical therapy. The generated summary nevertheless presented a fenamate (mefenamic acid) as a routine recommendation. All three reviewers identified this as the single most material error in the study, and two classified the output as unsafe for use without revision. The same summary also recorded an absence of swelling although the transcript documented mild post-activity swelling.

(ii) Systematic loss of pertinent negative findings

In five of six scenarios reviewers noted that clinically important negatives present in the transcript were absent from the summary, including absence of night pain, trauma, fever, joint locking, neurological deficit, saddle anaesthesia, sphincter disturbance, night sweats and weight loss. Reviewers emphasised that in musculoskeletal documentation the recorded absence of a red flag carries evidential weight equal to its presence.

(iii) Incomplete transfer of physical examination detail

Although examination findings were verbalised during recording, gait pattern, range of motion, crepitus, joint-line tenderness, varus deformity and distal neurovascular status were inconsistently carried into the structured output.

(iv) Over-triage of a non-urgent case

In S3 the summary escalated a chronic severe osteoarthritis presentation toward emergency referral. Both the orthopaedic reviewer and the second general practitioner judged this inappropriate in the absence of acute red flags: the correct disposition was planned or urgent orthopaedic clinic referral for arthroplasty assessment with weight-bearing radiographs, not emergency department attendance.

(v) Conflation of risk with established diagnosis

In S4 the summary presented post-menopausal osteoporosis risk as an established diagnosis, and altered patient-reported details regarding regular medication use and pain character. Reviewers required explicit separation of suspected, at-risk and confirmed diagnostic states, and confirmation that bone mineral density assessment is necessary for diagnosis [19,23].

(vi) Patient identity inconsistency

In S1 and S2 reviewers observed inconsistency between the honorific used in the transcript and the demographic profile recorded in the scenario metadata, with alternation between *ibu* and *pak*. Reviewers re-

quired verification of patient identity and sex before documentation is accepted — a concrete instance of the transcription-layer risk described in the ASR literature [11].

Reviewers also identified genuine strengths. Red-flag recognition and same-day triage in S5 were rated highly by all three reviewers, with appropriate splinting, elevation, analgesia and imaging referral. In S6 the suspicion of osteoporotic vertebral compression fracture was correctly supported by acute pain after light loading, focal tenderness, kyphosis and height loss, with neurological status intact and systemic red flags explicitly excluded. Referral direction was correct in five of six scenarios, and conservative management planning was consistently appropriate to the Puskesmas level.

Public health documentation indicators

Exploratory process indicators derived from the generated documentation are shown in Table 4. Red-flag screening, referral tiering and prevention-oriented content were present across scenarios, whereas retention of pertinent negatives and of examination detail, the indicators most dependent on faithful summarisation, were substantially incomplete.

Table 4. Exploratory public health documentation indicators across six scenarios.

Indicator	Scenarios adequately captured
Red-flag screening explicitly documented	6/6
Referral tier assigned	6/6 (direction correct in 5/6)
Patient education / prevention counselling documented	6/6
Weight-management counselling where BMI elevated	2/2 (S2, S3)
Fall-prevention guidance in older-adult cases	4/4 (S3–S6)
Osteoporosis case-finding / secondary-fracture-prevention flagged	3/3 relevant (S4–S6)
Functional status documented in full	3/6
Pertinent negative findings retained	1/6
Physical examination detail transferred in full	2/6

Discussion

This study shows that a domain-specific orthopaedic extension of the ePuskesmas LLM framework is technically feasible and produces documentation of acceptable but not clinically final quality. The overall mean of 3.29 of 5 corresponds to the rubric category *usable after moderate editing*. The distribution across domains, however, is more informative than the aggregate. The system performed best at determining appropriate subsequent action from the documented findings, referral appropriateness 4.00 and appropriateness of management at Puskesmas level 3.83, and worst in the domains that reflect documentation fidelity: accuracy of the clinical summary 2.89, completeness of the orthopaedic history 2.94 and clarity of the transcribed consultation 2.94.

This dissociation is the central finding. It runs counter to the intuitive expectation that summarisation would be the reliable component and clinical judgement the fragile one. In practice the model reproduced the shape of a competent primary-care plan while losing specific content from the source consultation. Documentation fidelity, not clinical reasoning, was the limiting factor. The pattern is consistent with evidence that LLM summarisation errors concentrate in omission and unsupported assertion rather than in gross clinical implausibility [7,9,10], and it explains why these errors were easy to miss: the outputs read fluently and looked clinically sensible.

Interpreted against the baseline established by Khatim et al. [1], the domain adaptation appears to have delivered what it was designed to deliver. The orthopaedic layer supplied referral tiering, Puskesmas-scope management and guideline-grounded education, and these were the highest-scoring domains. What the adaptation did not address, and what therefore remained limiting, were properties of the inherited transcription and summarisation stages. Domain-specific prompting improved the reasoning surface of the output without improving fidelity to the source.

Three features of the implementation plausibly contribute to the observed omissions and should be treated as testable design hypotheses rather than established mechanisms. First, the output schema contained no field that obliged the model to record negatives: pertinent negatives had no dedicated destination, and content without a destination was reliably lost. Second, the generation budget of 500 output tokens forces competition between fields in a thirteen-field orthopaedic template, and the fields most likely to be sacrificed are the long, low-salience lists of absent findings and examination detail. Third, the summarisation stage car-

ried no medication-safety or contraindication check of any kind, so the recommendation of mefenamic acid in S2 reflects a population-typical prescription for knee pain rather than conditioning on the documented gastritis and prior NSAID intolerance. The corresponding design responses are direct: a dedicated pertinent-negatives array and a red-flag checklist with explicit *present*, *absent* and *not assessed* values; an output budget scaled to the schema, or field-wise generation; a rule-based contraindication layer applied to any proposed drug against documented comorbidity and current medication; an enumerated referral-urgency taxonomy to constrain triage; and an explicit diagnostic-certainty field distinguishing *suspected*, *at risk* and *confirmed*.

The loss of pertinent negatives deserves emphasis in its own right. In musculoskeletal primary care, the record that there was no night pain, no trauma, no fever and no neurological deficit is not filler: it is the documented basis for not referring and not imaging, and its absence weakens both clinical reasoning and medicolegal defensibility. The over-triage in S3 and the risk-versus-diagnosis conflation in S4 similarly show that urgency and diagnostic calibration require explicit constraint and cannot be assumed to be inherited from a general-purpose model.

The transcription-layer findings align with published evidence that Whisper models can generate confident text with no basis in the audio, with elevated risk in speech that departs from training-distribution norms [11]. Indonesian clinical speech includes code-switching, regional variation and specialised terminology, all of which degrade recognition accuracy [25,26,27]. The patient-identity inconsistency observed in S1 and S2 is a small but instructive example: an error of this kind is trivial to correct at the point of care and consequential if it reaches the record, which is the argument for mandatory clinician verification of the transcript before summarisation rather than after it.

From a public health perspective the potential value of the system remains real but conditional. Referral tiering is central to primary-care gatekeeping under Indonesia's National Health Insurance scheme, supporting appropriate access to higher-level care and cost containment [28,29]. Scenarios S5 and S6 illustrate the principle that a first fragility fracture is an opportunity for secondary prevention, and the system did flag osteoporosis case-finding and preventive follow-up [17,18,19]. Structured outputs capturing modifiable risk factors could, in aggregate, contribute to non-communicable disease monitoring if mapped to interoperable systems such as SATUSEHAT [21,22], consistent with the recognition of rehabilitation as a universal health coverage priority [20]. This value is nevertheless contingent on the fidelity problems above: indicators derived from documentation that systematically omits negatives and examination detail would misrepresent the care actually delivered.

Deployment considerations follow from the Indonesian context. Heterogeneous ICT maturity across provinces [22] and the SATUSEHAT integration mandate [21] together imply that a documentation aid must tolerate limited connectivity and variable hardware, favour local or offline processing for sensitive audio [13], and avoid widening the gap between well-resourced urban and under-resourced rural Puskesmas. The current reliance on cloud APIs for both transcription and summarisation is at odds with each of these requirements and is a priority for redesign.

Limitations

The reviewer panel comprised three clinicians (two general practitioners and one orthopaedic specialist) recruited purposively rather than through a broader sampling strategy. Although inter-rater agreement was high, ICC(2,k) = 0.96, concordance among a small group of reviewers from a similar professional and national context does not exclude shared interpretive biases or unrecognised clinical blind spots. Validation using a larger, multidisciplinary, multi-centre panel, together with blinded comparison against clinician-authored documentation, is required before these findings can support implementation.

Further limitations apply. The scenario set was small (six) and purposively constructed. Recordings did not reproduce real Puskesmas conditions, including background noise, overlapping speech, regional dialects and variable audio quality. Physical examination content was available only because it was verbalised; the system does not and cannot verify examination findings. The analysis was descriptive, with no word error rate, no automated summarisation metrics and no formal hypothesis testing, and the reliability analysis quantifies agreement between reviewers rather than accuracy against an external gold standard. The SOAP

output specified in the design was not available for review and was therefore not evaluated. Finally, the public health and surveillance value is hypothetical: the process indicators measure documentation completeness rather than verified clinical action, interoperability was conceptual only, secondary use of clinical data at population scale raises governance and consent questions, and digital tools can worsen equity where infrastructure and digital literacy are unevenly distributed [13,22].

Future Work

The immediate priority is to address the safety failures identified by reviewers and to re-validate the system with a larger, multi-centre clinician panel and blinded comparison against clinician-authored notes. Safety improvements should include schema-level retention of pertinent negatives, an automated medication-safety and contraindication check, explicit representation of diagnostic certainty, a constrained referral-urgency taxonomy, and verification of patient identity and demographic data at the transcript stage. Technical evaluation should quantify transcription error, hallucination and summary accuracy in Indonesian clinical speech, including dialect and code-switching, and should assess privacy-preserving offline models suitable for low-connectivity settings. Subsequent prospective evaluation in Puskesmas should examine documentation time, completeness, workload, edit rates, interoperability with SATUSEHAT, integration with prevention pathways, cost-effectiveness and equity.

Conclusion

This orthopaedic extension of the ePuskesmas LLM framework was technically feasible and produced documentation rated 3.29 of 5 overall, with high inter-rater reliability, ICC(2,k) = 0.96, indicating usability as a draft requiring moderate editing rather than as a final clinical record. Performance was adequate for management and referral recommendations but weaker for faithful capture of consultation content, particularly pertinent negatives and examination findings. A materially unsafe analgesic recommendation underlines that mandatory clinician verification, targeted medication-safety and triage safeguards, and prospective evaluation in Puskesmas are prerequisites for routine use.

REFERENCES

[1] Khatim NA, Irfan AAA, Arief MM. Using Large Language Models for Real-time Transcription and Summarization of Doctor-patient Interactions into ePuskesmas in Indonesia: a proof-of-concept study. arXiv preprint arXiv:2409.17054. 2024. https://arxiv.org/abs/2409.17054

[2] Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal Innov Care Deliv. 2024;5(3):CAT.23.0404. https://doi.org/10.1056/CAT.23.0404

[3] Shah SJ, Devon-Sand A, Ma SP, Jeong Y, Crowell T, Smith M, et al. Ambient Artificial Intelligence Scribes: Physician Burnout and Perspectives on Usability and Documentation Burden**.** ***J Am Med Inform Assoc.*** 2025;32(2):375–380. https://doi.org/10.1093/jamia/ocae295

[4] Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell JT, Sheinberg M, et al. Clinician Experiences with Ambient Scribe Technology to Assist with Documentation Burden and Efficiency**.** ***JAMA Netw Open***. 2025;8(2):e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637

[5] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted Large Language Models can Outperform Medical Experts in Clinical Text Summarization. Nat Med. 2024;30(4):1134–1142. https://doi.org/10.1038/s41591-024-02855-5

[6] Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A Large Language Model for Electronic Health Records. NPJ Digit Med. 2022;5(1):194. https://doi.org/10.1038/s41746-022-00742-2

[7] Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records. ***NPJ Digit Med***. 2023;6(1):135. https://doi.org/10.1038/s41746-023-00879-8

[8] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language Models Encode Clinical Knowledge. Nature. 2023;620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2

[9] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023;55(12):1–38. https://doi.org/10.1145/3571730

[10] Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Au Yeung J, et al. A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. **NPJ Digit Med**. 2025;8(1):274. https://doi.org/10.1038/s41746-025-01670-7

[11] Koenecke A, Choi ASG, Mei KX, Schellmann H, Sloane M. Careless Whisper: Speech-to-text Hallucination Harms. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). New York: ACM; 2024. p. 1672–1681. https://doi.org/10.1145/3630106.3658996

[12] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition Via Large-scale Weak Supervision. arXiv preprint arXiv:2212.04356. 2022. https://arxiv.org/abs/2212.04356

[13] Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial Intelligence for Strengthening Healthcare Systems in low- and middle-income Countries: a Systematic Scoping Review. ***NPJ Digit Med***. 2022;5(1):162. https://doi.org/10.1038/s41746-022-00700-y

[14] GBD 2021 Osteoarthritis Collaborators. Global, Regional, and National Burden of Osteoarthritis, 1990-2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Rheumatol. 2023;5(9):e508–e522. https://doi.org/10.1016/S2665-9913(23)00163-7

[15] Kloppenburg M, Namane M, Cicuttini F. Osteoarthritis. Lancet. 2025;405(10472):71–85. https://doi.org/10.1016/S0140-6736(24)02322-5

[16] National Institute for Health and Care Excellence. Osteoarthritis in Over 16s: Diagnosis and Management. NICE guideline NG226. London: NICE; 2022. https://www.nice.org.uk/guidance/ng226

[17] Gregson CL, Armstrong DJ, Bowden J, Cooper C, Edwards J, Gittoes NJL, et al. UK Clinical Guideline for the Prevention and Treatment of Osteoporosis. Arch Osteoporos. 2022;17(1):80. https://doi.org/10.1007/s11657-022-01115-8

[18] Gregson CL, Armstrong DJ, Avgerinou C, Bowden J, Cooper C, Edwards J, et al. The 2024 UK Clinical Guideline for the Prevention and Treatment of Osteoporosis. Arch Osteoporos. 2025;20(1):119. https://doi.org/10.1007/s11657-025-01588-3

[19] Morin SN, Feldman S, Funnell L, Giangregorio L, Kim S, McDonald-Blumer H, et al. Clinical Practice Guideline for Management of Osteoporosis and Facture Prevention in Canada: 2023 update**.** ***CMAJ***. 2023 ;195 (39) :E1333–E1348.https://doi.org/10.1503/cmaj.221647

[20] World Health Organization. Package of interventions for rehabilitation. Geneva: World Health Organization; 2023. https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/package-of-interventions-for-rehabilitation

[21] Kementerian Kesehatan Republik Indonesia. Peraturan Menteri Kesehatan Nomor 24 Tahun 2022 tentang Rekam Medis. Jakarta: Kemenkes RI; 2022. https://peraturan.bpk.go.id/Details/234978/permenkes-no-24-tahun-2022

[22] Aisyah DN, Setiawan AH, Lokopessy AF, Faradiba N, Setiaji S, Manikam L, Kozlakidis Z. The Information and Communication Technology Maturity Assessment at Primary Health Care Services Across 9 provinces in Indonesia: evaluation study. JMIR Med Inform. 2024;12:e55959. https://doi.org/10.2196/55959

[23] US Preventive Services Task Force. Screening for osteoporosis to prevent fractures: US Preventive Services Task Force Recommendation Statement. JAMA. 2025;333(6):498-508. https://doi.org/10.1001/jama.2024.27154

[24] Dinas Kesehatan Kabupaten Rembang. Monitoring morbiditas pasien Kabupaten Rembang berdasarkan data Platform SATUSEHAT Januari–Mei 2025. 2025 Jun 3. https://dinkes.rembangkab.go.id/monitoring-morbiditas-pasien-kabupaten-rembang-berdasarkan-data-platform-satusehat-januari-mei-2025/

[25] Tazakka RVM, Lestari D, Purwarianti A, Tanaya D, Azizah K, Sakti S. Indonesian-English code-switching Speech Recognition Using the Machine Speech Chain Based Semi-supervised Learning. In: Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages at LREC-COLING 2024. 2024. p. 143-148. https://aclanthology.org/2024.sigul-1.18/

[26] Adila A, Lestari D, Purwarianti A, Tanaya D, Azizah K, Sakti S. Enhancing Indonesian Automatic Speech Recognition: Evaluating Multilingual Models with Diverse Speech Variabilities. In: Proceedings of the 27th Conference of the Oriental COCOSDA. IEEE; 2024. https://doi.org/10.1109/O-COCOSDA64382.2024.10800336

[27] Afonja T, Olatunji T, Ogun S, Etori NA, Owodunni A, Yekini M. Performant ASR Models for Medical Entities in Accented Speech. In: Proceedings of Interspeech 2024. 2024. p. 2315–2319. https://doi.org/10.21437/Interspeech.2024-2261

[28] Mulyanto J, Wibowo Y, Kringos DS. Exploring General Practitioners' Perceptions about the Primary Care Gatekeeper Role in Indonesia. BMC Fam Pract. 2021;22:5. https://doi.org/10.1186/s12875-020-01365-w

[29] Yanthi B, Hendratini J, Sulistyo DH. Determinan Rujukan non Spesialistik dengan Kriteria TACC di FKTP Kabupaten Batang Hari tahun 2022. J Jaminan Kesehatan Nasional. 2023;3(1). https://doi.org/10.53756/jjkn.v3i1.63