Role-Adapted Clinical Report Generation for Ultrasound Measurements in Low-Resource Settings

Ayoub Nainia[1,2], **Tanya Akumu**[2], **Noussair Lazrak**[2], **Karim Lekadir**[2]
[1] Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, Paris, France
[2] Artificial Intelligence in Medicine Lab (BCN-AIM), Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
ayoub.nainia@sorbonne-universite.fr, tanya.akumu@ub.edu,
noussair.lazrak@ub.edu, karim.lekadir@ub.edu

Abstract

Obstetric ultrasound is critical for monitoring fetal growth, yet in many low-resource settings, healthcare workers who perform or receive ultrasound measurements lack the training to interpret them clinically. We present a system that automatically generates role-adapted clinical reports from fetal biometry measurements, targeting six healthcare worker roles across three expertise levels. The system combines Retrieval-Augmented Generation (RAG) from a knowledge base extracted from the World Health Organization (WHO) Manual of Diagnostic Ultrasound with deterministic fetal growth percentile computation based on INTERGROWTH-21st international standards. The knowledge base is designed for multilingual extensibility: since the source material is from an official WHO document, entries can be translated into any target language by domain experts or machine translation services. A key design principle is that clinical decision support (red, yellow, and green alerts) is derived deterministically from percentile thresholds, not from the language model, ensuring safety regardless of LLM output quality. Evaluation demonstrates sub-millimeter accuracy in percentile computation, 100% correctness in decision support classification, measurable readability differentiation across roles (Flesch-Kincaid grade 8.8 for community health workers vs. 11-13 for clinical roles), and 98% factual consistency across 42 generated reports spanning seven clinical scenarios. The system is designed for local deployment without internet connectivity.

Keywords: clinical NLP, low-resource settings, fetal ultrasound, report generation, decision support, RAG

1. Introduction

Fetal growth monitoring through ultrasound biometry is a cornerstone of antenatal care. Measurements such as head circumference (HC), biparietal diameter (BPD), abdominal circumference (AC), femur length (FL), and estimated fetal weight (EFW) are compared against population-based reference standards to identify growth abnormalities that may require clinical intervention (Papageorghiou et al., 2014).

However, in many low and middle-income countries (LMICs), the healthcare workers who obtain or receive these measurements, such as community health workers, nurses, and midwives, often lack the specialized training to interpret percentile values, identify abnormal growth patterns, or formulate appropriate clinical recommendations (World Health Organization, 2013). This interpretation gap means that even when ultrasound equipment is available, the clinical value of the measurements may be lost.

Large language models (LLMs) have shown promise in generating clinical text (Singhal et al., 2023), but deploying them in medical settings raises safety concerns: an LLM might generate plausible but incorrect clinical recommendations. This risk is amplified in low-resource settings where there may be no specialist available to catch errors.

We address these challenges with a system that:

1. Generates role-adapted clinical reports from ultrasound measurements, adjusting language complexity and clinical detail for six healthcare worker roles across three expertise levels (basic, intermediate, advanced).

2. Computes fetal growth percentiles deterministically using INTERGROWTH-21st equations (Papageorghiou et al., 2014; Stirnemann et al., 2020), ensuring that the quantitative assessment is independent of the LLM.

3. Provides deterministic decision support (red/yellow/green alerts) based on percentile thresholds, not LLM output, which is a critical safety design for contexts where the user cannot independently verify clinical recommendations.

4. Supports multilingual extensibility through a modular knowledge base extracted from the WHO Manual of Diagnostic Ultrasound, Volume 2. Since entries are sourced from an official WHO document, they can be translated into any language via professional translators or machine translation, enabling deployment across diverse linguistic settings.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 262–269
16 May 2026.

2. Related Work

Large language models have advanced rapidly in medical applications, from clinical question answering (Singhal et al., 2023) to clinical text summarization, where adapted LLMs can match or outperform medical experts (Van Veen et al., 2024). Thirunavukarasu et al. (2023) review LLM deployment across clinical domains, identifying promising applications alongside persistent concerns about factual reliability. However, most work targets high-resource settings with specialist oversight, leaving a gap for environments where such oversight is unavailable.

For medical report generation specifically, earlier approaches used encoder-decoder architectures to produce text from medical images (Li et al., 2018; Chen et al., 2020). Retrieval-augmented generation (RAG) offers an alternative by grounding generation in authoritative sources (Lewis et al., 2020). Xiong et al. (2024) benchmarked medical RAG across five QA datasets, showing that retrieval quality critically determines whether RAG helps or harms performance. Our system uses RAG to ground narratives in WHO reference material, while relying on deterministic computation rather than the LLM for quantitative assessments.

A key challenge is that LLM-generated clinical text is prone to hallucination. Pal et al. (2023) introduced Med-HALT, the first benchmark for medical hallucinations, revealing significant variation across models. Jeblick et al. (2024) found that ChatGPT-simplified radiology reports occasionally contain incorrect or potentially harmful statements, even when the overall quality is high. Wiest et al. (2024) showed that locally-deployed open-source LLMs can achieve high accuracy for clinical information extraction while preserving data privacy, which is an approach our system also adopts. We go further by ensuring that safety-critical decision support is never derived from LLM output.

Despite these advances, most clinical NLP research remains focused on English-language EHRs (Electronic Health Records) in high-resource hospitals (Johnson et al., 2016). Ciecierski-Holmes et al. (2022) review AI deployments in LMICs and identify persistent barriers including the lack of representative datasets in under-resourced languages. For obstetric care specifically, Della Ripa et al. (2025) surveyed providers in African LMICs and found that clinical decision support alongside AI-enabled point-of-care ultrasound is deemed essential for improving antenatal outcomes, which is a need our system directly addresses.

Adapting medical text for different readers is well-established in health literacy research (Nutbeam, 2008). Recent work has applied LLMs to this task: Spina et al. (2024) demonstrated that GPT models can tailor information complexity to specific education levels, and Jeblick et al. (2024) showed ChatGPT can simplify radiology reports for patients. However, these studies focus on binary adaptation (expert vs. patient), whereas our system implements graduated adaptation across six healthcare worker roles at three expertise levels.

Our percentile computations build on the INTERGROWTH-21st international fetal growth standards (Papageorghiou et al., 2014) and the Hadlock formula for estimated fetal weight (EFW) (Hadlock et al., 1985), with EFW reference curves from Stirnemann et al. (2020).

3. System Architecture

Our system consists of four components: (1) a multilingually extensible knowledge base with retrieval pipeline, (2) an INTERGROWTH-21st percentile computation module, (3) a role-adapted prompt construction engine, and (4) a deterministic decision support layer. Figure 1 illustrates the overall architecture.

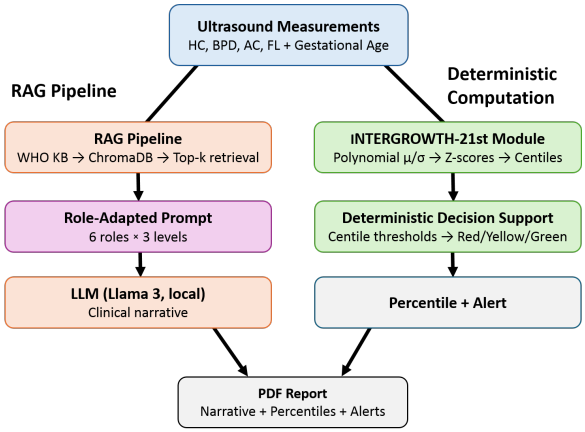

Figure 1: System architecture. Measurements flow through two independent paths: the RAG pipeline generates a clinical narrative via a locally-deployed LLM, while the INTERGROWTH-21st module computes percentiles that feed a deterministic decision support layer. The decision support alerts are never derived from LLM output.

3.1. Knowledge Base and RAG Pipeline

The knowledge base contains 171 entries extracted from the WHO Manual of Diagnostic Ultrasound, Volume 2 (World Health Organization, 2013). Each entry comprises a text passage with metadata (chapter, header, section, page number). Since the source material is from an authoritative WHO document, entries can be translated into any target language by professional translators for clinical-grade accuracy or via machine translation APIs for

rapid prototyping. As a proof of concept, we include machine-translated Swahili entries to demonstrate multilingual extensibility for East African deployment scenarios.

For retrieval, entries are encoded using the `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer (Reimers and Gurevych, 2019) and stored in a ChromaDB vector database. Given a set of ultrasound measurements, the system generates search queries from measurement types and values, retrieves the top-k most relevant knowledge base entries, and includes them as context in the LLM prompt.

The architecture is LLM-agnostic: any model served through Ollama can be used as the generation backend. In our experiments, we use Llama 3 8B (Grattafiori et al., 2024) as it offers a good balance between capability and resource requirements for local deployment. For non-English output, a multilingual model variant can be used.

3.2. INTERGROWTH-21st Percentile Computation

Rather than relying on the LLM to interpret whether measurements are normal, we compute percentiles deterministically using the published INTERGROWTH-21st equations. For fetal biometry (HC, BPD, AC, FL, OFD), the z-score is computed as:

$$z = \frac{x - \mu(\mathrm{GA})}{\sigma(\mathrm{GA})} \tag{1}$$

where x is the measured value and μ and σ are gestational-age-specific polynomial functions derived from the INTERGROWTH-21st study. The centile is then obtained via the standard normal CDF: $C = \Phi(z)$.

For estimated fetal weight, we first compute EFW using the Hadlock formula from HC, AC, and FL measurements (Hadlock et al., 1985), then compute the z-score using the Box-Cox Power Exponential (BCPE) distribution parameters published by Stirnemann et al. (2020):

$$z_{\mathrm{EFW}} = \frac{(\mathrm{EFW}/\mu)^{\lambda} - 1}{\lambda \cdot \sigma} \tag{2}$$

where λ, μ, and σ are gestational-age-specific parameters. The valid gestational age range is 14^{+0} to 40^{+6} weeks for biometry and 18^{+0} to 40^{+6} weeks for EFW.

3.3. Role-Based Adaptation

The system defines six healthcare worker roles organized into three expertise levels, as shown in Table 1.

Level	Roles	Language Style
Basic	CHW	Plain language, no jargon, clear action items
Intermediate	Nurse, Midwife	Medical terms with explanations
Advanced	Clin. Officer, Sonographer, GP	Full terminology, differentials, mgmt. plan

Table 1: Healthcare worker roles and expertise levels. CHW = Community Health Worker, GP = General Practitioner.

Role adaptation is achieved through role-specific instructions injected into a structured prompt template. The prompt takes the form: *"You are a medical assistant helping a {role} analyze ultrasound measurements. Use ONLY the reference knowledge below."* It then provides (1) role-specific instructions, (2) the raw measurements with INTERGROWTH-21st percentile assessments, (3) retrieved knowledge base entries as reference context, and (4) general instructions requiring the LLM to use the percentile assessment as the primary reference, cite sources, and not invent information beyond the provided context.

Role-specific instructions control language complexity. For example, the Community Health Worker prompt instructs: *"Explain in simple, non-technical language [...] Use plain terms instead of medical jargon (e.g., 'baby head size' instead of 'head circumference'). End with a clear ACTION: either 'Continue routine care' or 'REFER to hospital'."* In contrast, the General Practitioner prompt requests *"full medical terminology, percentile interpretation, differential diagnosis considerations, and recommended management plan."* Intermediate roles (Nurse, Midwife) receive instructions that use standard medical terminology but include explanations of complex findings.

3.4. Deterministic Decision Support

A critical design decision is that clinical urgency classification is *never* derived from LLM output. Instead, a deterministic algorithm evaluates the INTERGROWTH-21st percentiles:

- **Red (URGENT):** Any measurement $<$ 3rd or $>$ 97th percentile.
- **Yellow (FOLLOW UP):** Any measurement in the 3rd-10th or 90th-97th percentile range (and none red).
- **Green (NORMAL):** All measurements between the 10th and 90th percentile.

The worst flag across all measurements determines the overall severity. Each role level receives a severity-appropriate message: basic-level workers receive simple referral instructions ("Refer to hospital immediately"), while advanced-level workers receive clinical guidance ("Consider detailed anatomical survey, Doppler assessment, and specialist referral").

This separation ensures that even if the LLM generates an inaccurate narrative, the action recommendation remains correct.

4. Evaluation

We evaluate six aspects of the system: (1) percentile computation accuracy, (2) decision support correctness, (3) report readability across roles, (4) factual consistency of generated reports, (5) qualitative role differentiation, and (6) comparison against a template baseline.

Evaluations 1-2 use the project's deterministic test suite: 94 unit tests that verify percentile computation against the published INTERGROWTH-21st reference tables and decision support classification against hand-labeled cases covering normal, borderline, and abnormal scenarios, including boundary conditions (e.g., a centile of exactly 3.0 classified as yellow vs. 2.9 as red). Of these, 17 specifically target the decision support classifier (Section 3.4).

Evaluations 3-6 use 42 reports generated by Llama 3 (8B, Q4_K_M quantization). We designed seven clinical scenarios that span the full severity spectrum: three normal cases at different gestational ages (16, 30, and 38 weeks) to test age-dependent interpretation, one borderline-low case (measurements near the 5th percentile), one abnormal-low case (measurements below the 3rd percentile), one case with multiple abnormal measurements, and one abnormal-high/macrosomia case (measurements above the 97th percentile). Each scenario was generated for all six roles, yielding $7 \times 6 = 42$ reports. While the dataset is modest in size, the scenarios were selected to cover every decision support category and clinical direction of abnormality (low and high), providing systematic coverage of the system's behavior space. Factual consistency was evaluated using automated negation-aware checks that verify agreement between each report's narrative content and the computed percentile data (details in Section 4.4).

4.1. Percentile Accuracy

We compared the equation-computed mean (μ) values against the official INTERGROWTH-21st percentile tables for gestational ages 14-40 weeks. Table 2 shows the results.

Measure	GA Range	Max Err. (mm)	Mean Err. (mm)
HC	14-40 wk	0.05	0.02
BPD	14-40 wk	0.05	0.02
FL	14-40 wk	0.05	0.03

Table 2: Equation accuracy vs. INTERGROWTH-21st reference tables (50th percentile) across $N = 27$ gestational age weeks per measurement.

Scenario	Cases	Correct	Acc.
Normal (10th-90th)	4	4	100%
Borderline (3-10/90-97)	5	5	100%
Abnormal (<3/>97)	4	4	100%
Red-overrides-yellow	1	1	100%
Boundary + edge cases	3	3	100%
Total	**17**	**17**	**100%**

Table 3: Decision support classification accuracy. All cases classified correctly, including boundaries (e.g., centile of exactly 3.0 $\rightarrow$ yellow; 2.9 $\rightarrow$ red).

HC, BPD, and FL show sub-millimeter accuracy across all 27 gestational age weeks, confirming faithful implementation of the published equations. The EFW centile computation was validated by round-trip testing: feeding reference table values (3rd, 50th, 97th percentiles) through the centile function yields maximum errors of 0.003 centile units for p50 across 19 GA weeks (22-40).

4.2. Decision Support Safety

We tested the decision support classification with 17 test cases covering normal, borderline, and abnormal scenarios, including edge cases at classification boundaries (Table 3).

All 17 cases were classified correctly, including the red-overrides-yellow propagation test.

4.3. Readability Across Roles

We computed standard readability metrics for all 42 generated reports. Table 4 shows the results per role, averaged across seven scenarios.

Reports for the basic-level role (CHW) are written at the lowest reading level (FK grade 8.8, corresponding to 8th-9th grade), while clinical roles (Nurse through GP) produce reports at FK grade 10-14 (college-level prose). The Flesch reading ease confirms this: CHW reports score 49.0, while most clinical roles score 27-37.

The Sonographer role is a notable exception among advanced roles, producing reports with a lower FK grade (10.2) and higher reading ease (42.6) than expected. Inspection reveals that Sonographer reports are structured as con-

Role	Level	FK Grade	FK Ease	Fog
CHW	basic	8.8	49.0	13.0
Nurse	interm.	11.2	36.5	15.4
Midwife	interm.	13.6	27.1	18.1
Clin. Officer	adv.	12.0	32.3	16.4
Sonographer	adv.	10.2	42.6	14.6
GP	adv.	11.9	31.3	16.7

Table 4: Readability metrics per role, averaged across 7 scenarios ($N = 42$ reports). FK = Flesch-Kincaid; FK grade and Gunning Fog correspond to US school grade levels; FK reading ease ranges from 0 (hardest) to 100 (easiest).

Scenario	Reports	Consistent	Acc.
Normal (16w/30w/38w)	18	18	100%
Borderline low	6	6	100%
Abnormal low	6	6	100%
Multiple abnormal	6	6	100%
Abnormal high (macrosomia)	6	5	83%
Total	**42**	**41**	**98%**

Table 5: Factual consistency of generated reports. Negation-aware matching distinguishes "no signs of growth restriction" from "growth restriction is suspected."

cise per-measurement technical assessments with short, declarative sentences (e.g., "The measured value of 279.0 mm falls within the normal range"), whereas GP and Clinical Officer reports use longer discursive paragraphs with subordinate clauses. This reflects a genuine difference in clinical communication style: sonographers typically produce structured technical findings, while clinicians produce narrative interpretations. The readability metrics thus capture both prompt-driven adaptation and role-appropriate discourse patterns.

4.4. Factual Consistency

We evaluated whether generated reports contradict the known percentile data using automated negation-aware checks across all 42 reports (Table 5).

Consistency checks verify that: (1) reports for normal measurements do not contain non-negated alarm terms (e.g., "urgent," "growth restriction"), (2) reports for abnormal measurements express appropriate concern, and (3) mentioned percentile values do not deviate from computed values by more than 30 percentage points.

Of 42 reports, 41 passed (98%). The single failure occurred in the macrosomia scenario (measurements $>$99th percentile): the Sonographer report mentioned "intrauterine growth restriction" despite all measurements being *above* the expected range. This error traces to the RAG pipeline, which retrieves growth restriction content for any abnormal measurement query without distinguishing the direction of abnormality. While the deterministic decision support correctly flagged this case as urgent, the LLM narrative contained clinically contradictory information, illustrating precisely why the system's safety-critical decisions are not derived from LLM output.

Community Health Worker (basic level):

"[...] The baby's head size is *smaller than normal* for this age. [...] I recommend **REFERRAL** to the hospital for further evaluation. This referral is needed because the baby's head size is significantly below expected."

General Practitioner (advanced level):

"[...] Head Circumference: 240.0 mm, significantly below the expected mean of 278.4 mm (0.0th percentile, z-score -4.15). [...] This may indicate a *small-for-gestational-age* (SGA) fetus. [...] Proper clinical management of intrauterine growth restriction requires maternal hospitalization and strict fetal surveillance, including a non-stress test and serial Doppler velocity waveform measurements."

Figure 2: Excerpted reports for the same abnormal scenario. The CHW report uses plain language and a clear REFER action; the GP report includes z-scores, differential diagnosis (SGA), and a detailed management plan.

4.5. Qualitative Role Differentiation

Figure 2 shows excerpted outputs for the same abnormal scenario (HC at 0th percentile, 30 weeks) generated for a Community Health Worker and a General Practitioner.

The CHW report uses accessible language ("baby's head size is smaller than normal") and ends with a clear action (REFER), while the GP report includes z-scores, references to SGA classification, and specific clinical management recommendations (Doppler assessment, non-stress test). This qualitative difference, combined with the quantitative readability results in Table 4, confirms that role adaptation is functioning as designed.

4.6. Template Baseline Comparison

To quantify the value added by the LLM over simple rule-based generation, we compared the RAG-generated reports against a template baseline that fills measurement values and decision support recommendations into fixed sentence patterns (Table 6).

LLM-generated reports are 2-4$\times$ longer than templates and contain nearly twice as many medical

Level	Source	FK Grade	Words	Med. Terms
Basic	LLM	8.8	196	5.3
	Tmpl	9.8	65	1.4
Interm.	LLM	12.4	251	9.2
	Tmpl	9.9	90	5.1
Adv.	LLM	11.3	333	9.0
	Tmpl	9.9	90	5.1

Table 6: LLM-generated reports vs. template baseline. Words = average word count; Med. Terms = count of 20 clinical terms (e.g., "differential," "Doppler," "SGA") found per report.

terms (5-9 vs. 1-5). Critically, the LLM produces meaningful readability differentiation between roles (FK grade 8.8 for basic vs. 12.4 for intermediate), while templates produce nearly uniform readability (FK 9.8-9.9) regardless of the target role. The templates also lack clinical context: they report measurements and a recommendation but cannot provide differential diagnosis considerations, management plans, or references to clinical guidelines that the RAG-grounded LLM incorporates.

5. Discussion

5.1. Design Principles for Clinical Safety

Our central design principle, that decision support must be deterministic, not LLM-generated, addresses a fundamental concern with deploying language models in clinical settings. The LLM provides valuable clinical narrative that contextualizes the measurements, but it does not determine whether a patient needs referral. This separation allows the system to be useful even when LLM output quality varies, which is particularly important when using smaller, locally-deployed models (8B parameters, 4-bit quantization) in low-resource settings.

The factual consistency evaluation (Section 4.4) provides evidence that this design works in practice: across 42 generated reports spanning seven scenarios, 41 were factually consistent (98%). The single failure, a Sonographer report that mentioned "intrauterine growth restriction" for a macrosomia case, illustrates a known RAG limitation: retrieval queries based on abnormal measurements do not distinguish the direction of abnormality. Importantly, this narrative error did not affect the deterministic decision support, which correctly classified the case as urgent. The template baseline comparison (Section 4.6) further validates the LLM's contribution: RAG-grounded reports provide 2-4$\times$ more content, meaningful readability differentiation, and clinical context (differential diagnoses, management plans) that static templates cannot offer.

5.2. Low-Resource Deployment

The system runs entirely locally using Ollama, requiring no internet connectivity or cloud API access. This is critical for deployment in settings with limited or unreliable internet infrastructure. The locally-deployed LLM also addresses data privacy concerns, as patient measurements never leave the local device.

The knowledge base architecture is designed for multilingual extensibility. Since entries are sourced from an official WHO document, they can be translated into any target language by local healthcare organizations or professional translators to produce clinically validated versions. The modular design allows new languages to be integrated without modifying the retrieval or generation pipeline.

5.3. Limitations

We have not yet conducted a user study with healthcare workers; readability metrics provide an objective proxy for language complexity but do not capture whether reports are understandable, actionable, or trustworthy for the intended audience. The factual consistency evaluation relies on negation-aware keyword matching, which covers the most critical error class (misidentifying normal vs. abnormal) but does not replace expert annotation. Although the knowledge base architecture supports multilingual deployment and we include machine-translated Swahili entries as a proof of concept, no direct evaluation of multilingual output quality was performed; validating retrieval and generation in non-English settings remains necessary before deployment. The knowledge base scope is also limited to 171 entries from a single WHO manual, which suffices for the fetal biometry use case but may lack coverage for more complex or ambiguous clinical scenarios. Finally, our evaluation uses a single model (Llama 3, 8B parameters); performance may vary across models. User studies, expert-annotated evaluation, multilingual evaluation, and multi-model comparison are planned as future work.

6. Conclusion

We presented a system for generating role-adapted clinical reports from obstetric ultrasound measurements, designed for healthcare workers in low-resource settings. The system combines RAG from a multilingually extensible WHO knowledge base with deterministic INTERGROWTH-21st percentile computation and a safety-focused decision support layer. Evaluation across 42 reports spanning seven clinical scenarios demonstrates: (1) sub-millimeter accuracy in percentile computation, (2) perfect decision support classification across 17 test cases,

(3) measurable readability differentiation across roles (FK grade 8.8 for community health workers vs. 11-14 for clinical roles), (4) 98% factual consistency with a single instructive failure that validates the deterministic safety layer, and (5) substantial improvement over template baselines in content richness and role differentiation. The system is designed for local deployment without internet connectivity.

Ethics Statement

This system is designed as a clinical decision support tool and is not intended to replace professional medical judgment. The deterministic decision support layer is designed to prioritize caution, recommending referral for any measurement outside normal ranges. The system has not been validated in a clinical setting and should not be used for clinical decision-making without appropriate validation and regulatory approval.

Acknowledgments

This research was funded by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (AIMIX project – Grant Agreement No. 101044779). This work is co-funded by the European Union's Horizon Europe research and innovation programme through Cofund SOUND.AI under the Marie Skłodowska-Curie Grant Agreement No. 101081674.

7. Bibliographical References

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449. Association for Computational Linguistics.

Tomasz Ciecierski-Holmes, Raj Singh, Matthew Axt, Stefan Axt, and Samia Brenner. 2022. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *npj Digital Medicine*, 5:162.

Sabrina Della Ripa, Christopher Oduor, Anthony Wanyoro, Peter Nguhiu, Elizabeth Rakers, and Cynthia K. Stanton. 2025. AI-enabled obstetric point-of-care ultrasound as an emerging technology in low- and middle-income countries: provider and health system perspectives. *BMC Pregnancy and Childbirth*, 25:729.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. 2024. The Llama 3 herd of models.

Frank P. Hadlock, R.B. Harrist, Ralph S. Sharman, Russell L. Deter, and Seung K. Park. 1985. Estimation of fetal weight with the use of head, body, and femur measurements—a prospective study. *American Journal of Obstetrics and Gynecology*, 151(3):333–337.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, and Michael Ingrisch. 2024. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European Radiology*, 34:2817–2825.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems*, volume 31.

Don Nutbeam. 2008. The evolving concept of health literacy. *Social Science & Medicine*, 67(12):2072–2078.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334. Association for Computational Linguistics.

Aris T Papageorghiou, Eric O Ohuma, Douglas G Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A Jaffer, Enrico

Bertino, Michael G Gravett, Manorama Purwar, et al. 2014. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st project. *The Lancet*, 384(9946):869–879.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Aidin Spina, Saman Andalib, Daniel Flores, Rishi Vermani, Faris F Halaseh, and Ariana M Nelson. 2024. Evaluation of generative language models in personalizing medical information: Instrument validation study. *JMIR AI*, 3:e54371.

Julien Stirnemann, Laurent J Salomon, and Aris T Papageorghiou. 2020. INTERGROWTH-21st standards for Hadlock's estimation of fetal weight. *Ultrasound in Obstetrics & Gynecology*, 56(6):946–948.

Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30:1134–1142.

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine*, 7:257.

World Health Organization. 2013. *Manual of Diagnostic Ultrasound*, 2nd edition, volume 2. World Health Organization, Geneva.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251. Association for Computational Linguistics.