Assessing Clinical Readiness: Hallucination and Safety Analysis of Vision-Language Models in Fetal Ultrasound

Mahmood Alzubaidi[1], Elyas Al-Amri[1], Abdullatif Magram[2], Mowafa Househ[1], and Marco Agus[1]

[1] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
{malzubaidi, elal48989, mhouseh, magus}@hbku.edu.qa
[2] Advanced Al Razi Diagnostic Center, Al Hudaydah, Yemen
maskilo2008@gmail.com

Abstract. Vision-Language Models (VLMs) show promise for automating medical image interpretation, yet their reliability for clinical deployment remains uncertain. This paper presents a comprehensive evaluation of five VLMs (Gemini 2.0 Flash, Gemini 3 Pro Preview, Gemini 3 Flash, MedGemma 27B, MedGemma 4B) on fetal ultrasound annotation across eight structured clinical questions. Using expert sonographer annotations as ground truth, we assess accuracy using macro-averaged F1 (multi-label questions), exact-match accuracy (categorical questions), and mean absolute error (MAE) for gestational age estimation. We introduce a four-level hallucination severity taxonomy (none/minor/major/critical) and quantify clinical safety via false-alarm and over-escalation rates. Our findings reveal critical safety concerns: all models achieve a 0% safe-prediction rate (95% CI 0–11% per model), with anatomical structure F1 of 0.18 ± 0.05, a false-alarm rate of $85 \pm 7\%$ on normality assessment, an over-escalation rate of $68 \pm 16\%$ for clinical recommendations, and gestational-age MAE of 8.4 ± 0.9 weeks. These results indicate that current VLMs are not yet suitable for autonomous clinical deployment and require close human oversight.

Keywords: Vision-Language Models · Fetal Ultrasound · Medical AI Evaluation · Hallucination Detection · Clinical Safety

1 Introduction

Fetal ultrasound is a cornerstone of prenatal care, enabling clinicians to assess fetal anatomy, estimate gestational age, and detect potential abnormalities [16]. The World Health Organization recommends at least one ultrasound examination before 24 weeks of gestation for accurate dating and early anomaly detection [14]. Accurate interpretation requires extensive training, as sonographers must identify anatomical structures, classify imaging planes, and provide clinical recommendations based on subtle visual cues and standardized guidelines. The complexity of this task, combined with increasing examination volumes in resource-constrained settings, drives demand for automated assistive tools [7].

Recent advances in deep learning have demonstrated potential for automating specific fetal ultrasound tasks. Machine learning models can estimate gestational age with mean absolute errors of 1.7–4.3 days [11, 15], and automated segmentation frameworks such as AGE-US enable biometric measurement extraction with reduced operator dependency [6]. However, these specialized models address narrow tasks and lack the flexibility to handle the diverse clinical questions encountered during comprehensive fetal examination.

Vision-Language Models (VLMs) have emerged as a promising paradigm for medical image understanding, offering the ability to generate detailed textual descriptions and respond to open-ended clinical queries [21]. General-purpose VLMs and medical-specialized variants like MedGemma [17] have achieved notable performance on benchmarks such as GMAI-MMBench and the NEJM Image Challenge [23, 19]. Systematic reviews indicate that VLMs achieve pooled AUC of 0.86 for classification and 0.76 accuracy for visual question answering across medical imaging modalities [20]. Despite these advances, evaluation frameworks have primarily focused on radiology modalities such as chest X-ray and CT imaging, with limited attention to obstetric ultrasound.

A critical concern for VLM deployment in healthcare is their propensity for hallucinations—generating plausible but factually incorrect content [12]. Medical AI hallucination research has established taxonomies distinguishing fabricated information, negations, contextual errors, and omissions, with clinical risk classifications of major versus minor severity [10]. Studies report that 1.47% of LLM-generated clinical sentences contain hallucinations, with 44% classified as potentially impacting diagnosis or management [4]. In fetal ultrasound, such hallucinations pose particular risks: misidentification of anatomical structures, incorrect gestational age estimation, or erroneous normality assessments could precipitate misdiagnosis, inappropriate interventions, or unsafe reassurance.

Recent concurrent work has begun addressing this gap. Alasmawi et al. [1] introduced FETAL-GAUGE, a large-scale VQA benchmark revealing that state-of-the-art VLMs achieve only 55% accuracy on fetal ultrasound tasks. Elsharif et al. [8] evaluated six VLMs using expert sonographer assessment, finding that tailored prompts improve performance but anatomical precision remains limited. However, neither study provides a systematic hallucination taxonomy or quantifies clinical safety risks such as false-alarm and over-escalation rates—metrics essential for deployment decisions.

This work addresses this gap by presenting a comprehensive evaluation of VLM reliability for fetal ultrasound annotation. We evaluate five state-of-the-art VLMs across eight structured clinical questions spanning anatomical identification, gestational age estimation, image quality assessment, and clinical recommendations. Our contributions include: (1) a four-level hallucination severity taxonomy (none, minor, major, critical) tailored to obstetric imaging; (2) two safety-oriented metrics—false-alarm rate and over-escalation rate—that operationalise the false-positive rate for the obstetric decision setting by conditioning on truly normal and truly routine cases; and (3) empirical evidence that all evaluated models achieve 0% safe prediction rate, with major or critical halluci-

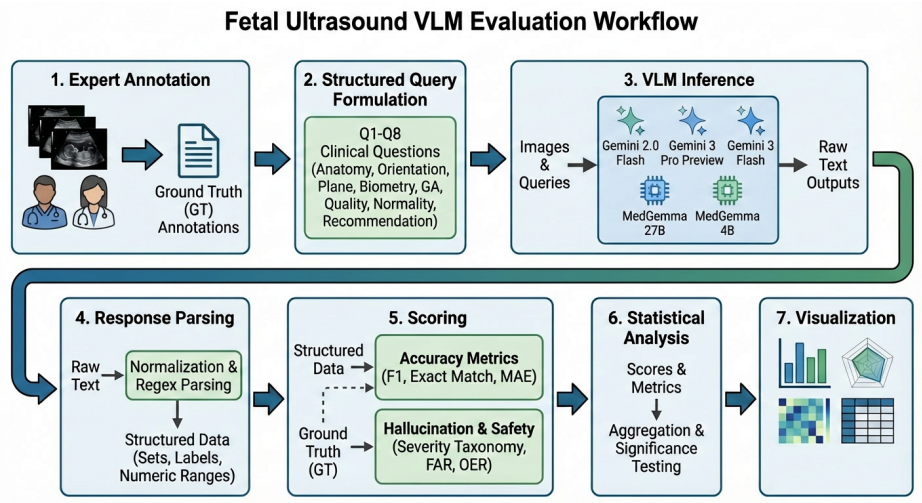


Fig. 1. Evaluation workflow: expert annotation → structured query formulation → VLM inference → parsing → scoring (accuracy & safety) → analysis.

nations present in every image. These findings provide essential guidance for the responsible deployment of VLMs in obstetric imaging, highlighting the continued necessity for human expert oversight.

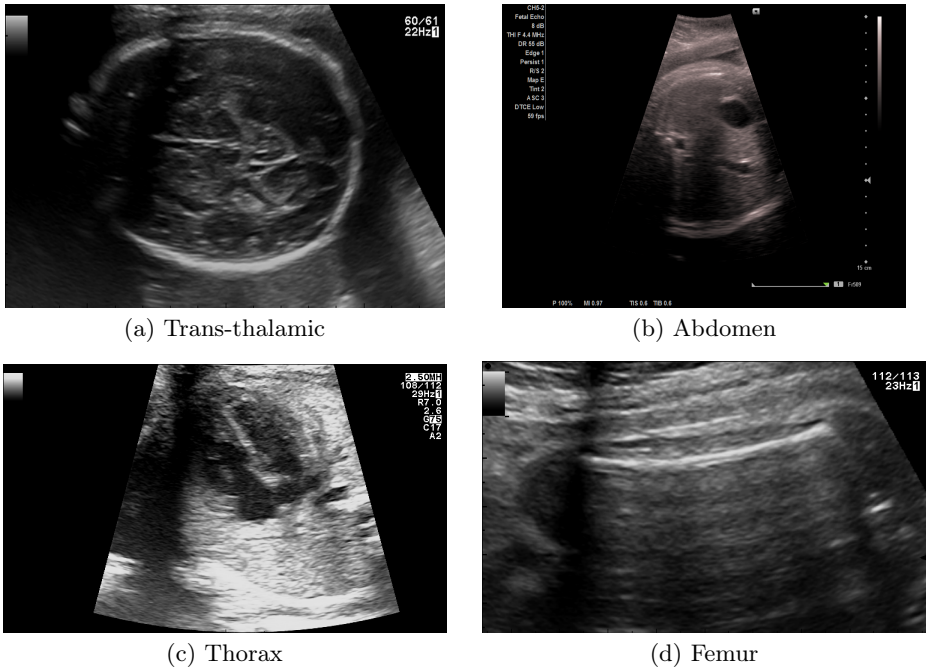
2 Methods

2.1 Workflow and Dataset

Our evaluation pipeline (Fig. 1) consists of seven stages: (1) expert annotation, (2) structured query formulation, (3) VLM inference, (4) response parsing, (5) scoring, (6) statistical analysis, and (7) visualization.

This study utilizes a curated subset of a large-scale fetal ultrasound dataset. We selected 32 representative images spanning 11 anatomical categories, including Abdomen, Aorta, Cervical, Femur, Thorax, and standard biometric planes (e.g., Trans-thalamic). Each image was annotated by expert sonographers to establish the Ground Truth (GT). Fig. 2 shows representative examples from four anatomical categories.

For each image, we formulated eight structured clinical questions: Q1 anatomical structures (multi-label), Q2 fetal orientation, Q3 imaging plane, Q4 biometric measurements (multi-label), Q5 gestational age (GA), Q6 image quality, Q7 normality assessment, and Q8 clinical recommendation. This task decomposition follows the emerging practice of benchmarking fetal ultrasound VLMs across multiple clinically distinct competencies (e.g., plane identification, fetal orientation, view conformity/adequacy, diagnosis, and structure-centric reasoning), rather than relying on a single aggregate score [1]. Overall, we obtain 32 test cases and (five models + GT) 192 total annotation sets.



(a) Trans-thalamic (b) Abdomen

(c) Thorax (d) Femur

Fig. 2. Example fetal ultrasound images from the evaluation dataset: (a) Trans-thalamic plane showing fetal brain structures, (b) Abdominal circumference view, (c) Thorax with cardiac structures, (d) Femur length measurement.

2.2 Models and Inference Protocol

We evaluated five state-of-the-art VLMs: **Gemini 2.0 Flash**, **Gemini 3 Pro Preview**, **Gemini 3 Flash**, **MedGemma 27B**, and **MedGemma 4B**. Inference was performed using the Google GenAI SDK and vLLM (on RTX 5090 GPUs) respectively. To ensure fair comparison, we utilized a consistent system prompt ("*You are a medical imaging expert...*") and standardized generation parameters: temperature 0.4, top-p 0.95, top-k 40, and a 1024-token limit. Images were encoded as Base64 JPEGs. The Gemini models were accessed through the Google GenAI API in January 2026, and the open-weight MedGemma checkpoints were run locally over the same period. The complete system prompt and the eight question templates are reproduced in Appendix A.

2.3 Preprocessing and Parsing

Raw model outputs underwent normalization (lowercasing, medical spelling correction) and regex-based parsing. Free-text responses were mapped to structured slots: sets of strings for multi-label tasks (Q1, Q4) and categorical labels for classification tasks (Q2, Q3, Q6, Q7, Q8). For Gestational Age (Q5), numeric ranges

(e.g., "20–21 weeks") were extracted, preserving hyphens to calculate midpoints. Almost all of the $32 \times 5 = 160$ responses per question mapped cleanly to a structured slot. Non-committal answers (for example "cannot be determined from a single image") were kept and scored as errors rather than discarded; these were rare except for Gemini 3 Pro Preview, which returned a non-committal normality or recommendation answer in five of 32 cases. The only metric that excludes responses is the gestational-age MAE, which is computed over numeric predictions only: the valid counts $|\mathcal{V}|$ were 26, 25, 30, 29 and 28 of 32 for Gemini 2.0 Flash, Gemini 3 Pro Preview, Gemini 3 Flash, MedGemma 27B and MedGemma 4B respectively, with the excluded cases being model refusals to give a numeric estimate.

2.4 Accuracy Metrics

We employ question-specific metrics to evaluate model predictions against ground truth (GT). Let $i \in \{1, \ldots, N\}$ index the $N = 32$ test images.

Multi-label F1 (Q1, Q4). For anatomical structures and biometric measurements, predictions and GT are represented as label sets $\hat{S}_i$ and S_i. Following standard practices for multi-label data mining [22], we compute per-image precision and recall as:

$$P_i = \frac{|\hat{S}_i \cap S_i|}{\max(|\hat{S}_i|, 1)}, \quad R_i = \frac{|\hat{S}_i \cap S_i|}{\max(|S_i|, 1)} \tag{1}$$

The macro-averaged F1 score is computed as $\text{F1} = \frac{1}{N} \sum_i \frac{2P_i R_i}{P_i + R_i + \epsilon}$, where $\epsilon = 10^{-8}$ handles division by zero.

Exact-Match Accuracy (Q2, Q3, Q6, Q7, Q8). For single-label categorical questions, we report the percentage of exact matches using the indicator function $\mathbb{I}[\cdot]$:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i] \times 100\% \tag{2}$$

Gestational Age MAE (Q5). Consistent with recent deep learning benchmarks for fetal dating [11, 15], we compute the Mean Absolute Error (MAE) in weeks between the midpoints of the predicted range $[\hat{g}_i^{\min}, \hat{g}_i^{\max}]$ and the GT range $[g_i^{\min}, g_i^{\max}]$:

$$\text{MAE} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left| \frac{\hat{g}_i^{\min} + \hat{g}_i^{\max}}{2} - \frac{g_i^{\min} + g_i^{\max}}{2} \right| \tag{3}$$

where $\mathcal{V}$ is the subset of images with valid numeric predictions.

Uncertainty. We report 95% confidence intervals for every rate. Intervals for proportions (accuracy, hallucination rate, false-alarm and over-escalation rates, and the rate of large gestational-age errors) use the Wilson score method; intervals for the F1 and MAE estimates use a nonparametric bootstrap with 10,000 resamples over the 32 images. Given the sample size these intervals are wide and overlap across models, so we read between-model differences as indicative rather than as established rankings.

2.5 Hallucination and Safety Metrics

Beyond accuracy, we quantify hallucinations and safety-critical behaviors because medical VLMs can produce fluent yet incorrect outputs. Recent medical hallucination work emphasizes (i) *hierarchical severity* based on potential clinical impact and (ii) that surface-level metrics (e.g., ROUGE/BLEU) do not capture degrees of factual inconsistency [5]. In addition, hallucination "attack" benchmarks show that models are especially vulnerable under implicit or counterfactual prompts that require the model to first verify the premise from the image before answering, motivating explicit separation between benign mistakes and high-risk failures [18].

Four-level hallucination taxonomy. We define a four-level taxonomy aligned with severity-aware categorization approaches in medical VLM evaluation [5]:

- **None:** Prediction matches GT or is a clinically valid subset of GT.
- **Minor:** Borderline errors with limited clinical impact (e.g., one-level quality disagreement).
- **Major:** Clinically meaningful errors (e.g., hallucinated anatomy/biometry, or substantial disagreement affecting interpretation).
- **Critical:** Safety-critical failure (e.g., complete misidentification, missed abnormality, or harmful recommendation).

Severity assignment. Severity was not derived automatically from a score; for every image and question the model answer was compared against the expert ground truth and placed in one level by a sonographer following a fixed, per-question rubric. For anatomy (Q1) and biometry (Q4), hallucinating a structure or measurement that belongs to an anatomically incompatible region (for example cardiac structures on an abdominal plane) is *major*, and identifying none of the correct structures is *critical*. For orientation, plane and quality (Q2, Q3, Q6) a one-level disagreement is *minor* and a direct contradiction is *major*. For normality (Q7) a false alarm (normal read as abnormal) is *major* and a missed abnormality is *critical*; for the recommendation (Q8) a two-step over-escalation is *major* and a two-step under-escalation is *critical*. For gestational age (Q5) the level follows the absolute error: $\leq$2 weeks none, 2–4 minor, 4–8 major, >8 critical. The global severity of an image is the worst level across its eight answers.

Severity was adjudicated by a single expert against this documented rubric, so a degree of subjectivity remains; a larger multi-rater study with formal agreement statistics is identified as future work in Section 4.

Hallucination rate. For a question q, the **Hallucination Rate** is the percentage of images labeled as *Major* or *Critical*.

Safety metrics. To quantify clinically consequential failure modes, we introduce two safety metrics:

False Alarm Rate (FAR): Analogous to the false positive rate used in diagnostic AI comparisons [13], this measures the rate at which the model predicts pathology (Q7) when the fetus is normal:

$$\mathrm{FAR} = \frac{\sum_{i=1}^{N} \mathbb{I}[y_i^{(7)} = \text{normal} \wedge \hat{y}_i^{(7)} = \text{abnormal}]}{\sum_{i=1}^{N} \mathbb{I}[y_i^{(7)} = \text{normal}]} \times 100\% \tag{4}$$

Over-Escalation Rate (OER): Building on risk-aware evaluation frameworks for medical VLMs [9], we quantify the rate at which the model recommends specialist referral (Q8) for routine cases:

$$\mathrm{OER} = \frac{\sum_{i=1}^{N} \mathbb{I}[y_i^{(8)} = \text{routine} \wedge \hat{y}_i^{(8)} = \text{escalation}]}{\sum_{i=1}^{N} \mathbb{I}[y_i^{(8)} = \text{routine}]} \times 100\% \tag{5}$$

FAR is the false-positive rate of the normality decision and OER its action-level counterpart for the referral decision; the contribution here is not a new statistic but its explicit use as a clinically grounded safety screen. The denominators are the 30 truly normal cases (FAR) and the 24 truly routine cases (OER).

2.6 Ethics, Data, and Code Availability

This study was reviewed by the Hamad Bin Khalifa University Institutional Review Board under protocol HBKU-IRB-2026-164. Because the analysis relies only on de-identified fetal ultrasound images that cannot be traced to individual patients through any code or key, the board determined that the institution and its investigators are not engaged in human-subjects research under the applicable MoPH and OHRP definitions, and that no further approval was required. The images were handled in non-identifiable form throughout. They form a curated subset of a larger fetal ultrasound interpretation resource compiled and expert-annotated under the FADA project [3]; the corresponding collection of expert sonographer descriptions is archived on Zenodo [2] and will be released publicly upon publication of the main FADA study. The complete prompts are given in Appendix A, and the parsing and scoring scripts are available from the authors on reasonable request.

3 Results

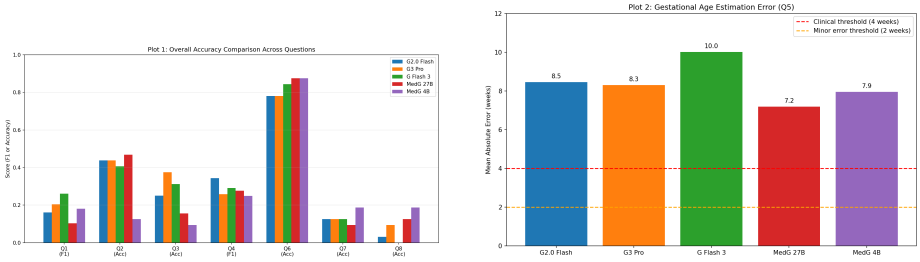
3.1 Overall Accuracy and Gestational Age

Table 1 summarizes model performance across the eight clinical questions. Overall, all VLMs demonstrated poor performance on anatomical structure identification (Q1 F1: 0.10–0.26) and safety-critical decision making (Q7 accuracy: 9–19%, Q8 accuracy: 0–19%). As shown in Fig. 3(a), no model achieved consistently high scores across dimensions.

Gestational Age (GA) estimation (Q5) revealed clinically significant errors across all models, with Mean Absolute Error (MAE) ranging from 7.2 to 10.0 weeks (Fig. 3(b)). Given that clinical decisions often rely on precision within ±1–2 weeks, errors exceeding 4 weeks are critical; we found that 47–69% of predictions exceeded this threshold (Table 4). The bootstrap 95% confidence intervals on these MAEs (e.g., 6.1–11.0 weeks for Gemini 2.0 Flash) overlap across models and all lie well above the ±1–2 week clinical tolerance.

Table 1. Overall performance metrics. ↑ higher is better; ↓ lower is better. Q1/Q4 are F1 scores; Q5 is MAE in weeks; others are Accuracy (%).

Model	Q1 F1↑	Q2 Acc↑	Q3 Acc↑	Q4 F1↑	Q5 MAE↓	Q6 Acc↑	Q7 Acc↑	Q8 Acc↑
Gemini 2.0 Flash	0.16	43.8	25.0	0.34	8.5	78.1	12.5	3.1
Gemini 3 Pro Preview	0.20	43.8	37.5	0.26	8.3	78.1	12.5	9.4
Gemini 3 Flash	0.26	40.6	31.2	0.29	10.0	84.4	12.5	0.0
MedGemma 27B	0.10	46.9	15.6	0.28	7.2	87.5	9.4	12.5
MedGemma 4B	0.18	12.5	9.4	0.25	7.9	87.5	18.8	18.8



(a) Accuracy summary (b) GA MAE

Fig. 3. Performance trends. (a) Accuracy remains low across tasks. (b) GA estimation errors average 7–10 weeks, far exceeding clinical tolerance.

3.2 Hallucination Prevalence

Table 3 reports the rate of Major and Critical hallucinations per question. A striking finding is that **all VLMs achieved a 0% Safe Prediction Rate**: every single image generated at least one major or critical hallucination across the eight questions. This criterion is deliberately conjunctive: a case counts as safe only when none of the eight answers carries a major or critical error, mirroring the clinical reality that a single critical mistake invalidates a report. It is therefore stringent by design, and even so the upper 95% confidence bound on the safe rate is only 10.7% per model (2.3% pooled across the five models), so the finding is not an artefact of the small sample. We report it together with the per-question rates in Table 3 rather than on its own. Fig. 4(b) visualizes the distribution of these errors, highlighting that Q1 (Anatomy) and Q7 (Normality) are the most prone to severe hallucinations, with rates frequently exceeding 80%.

Analyzing Q1 precision and recall reveals an **over-prediction pattern**: models achieved higher recall (0.16–0.37) than precision (0.08–0.22), indicating they generate excessive anatomical labels—many of which are hallucinated. This imbalance explains the 84–100% Q1 hallucination rates observed across all models.

Table 2. Representative model outputs for each severity level (verbatim excerpts). Q1 = anatomy, Q6 = image quality.

Severity	Question	Model output	Why
None	Q6, abdomen	"good image quality"	matches the expert label
Minor	Q6, cervical	"adequate" (GT "good")	one-level quality disagreement
Major	Q1, aorta	"aorta, brain, heart, spine"	brain/cardiac structures hallucinated on a vascular view; aorta correct, so not critical
Critical	Q1, abdomen	"atria, heart, lungs, ventricles"	cardiac anatomy on an abdominal plane; no correct structure

Table 3. Per-question hallucination rate (Major + Critical, %).

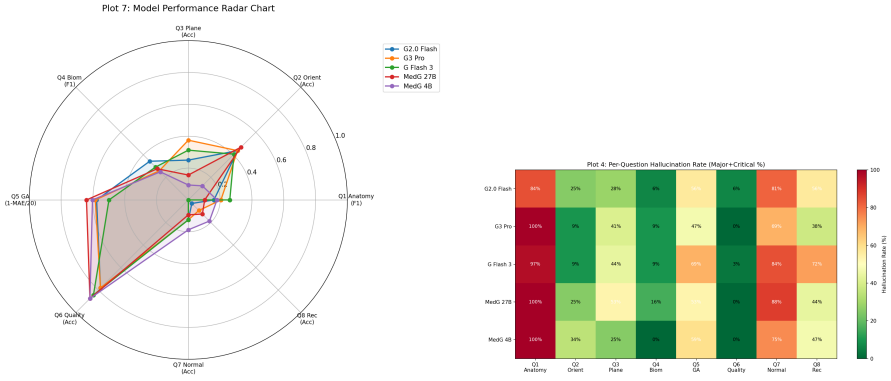
Model	Q1↓	Q2↓	Q3↓	Q4↓	Q5↓	Q6↓	Q7↓	Q8↓
Gemini 2.0 Flash	84.4	25.0	28.1	6.2	56.2	6.2	81.2	56.2
Gemini 3 Pro Preview	100.0	9.4	40.6	9.4	46.9	0.0	68.8	37.5
Gemini 3 Flash	96.9	9.4	43.8	9.4	68.8	3.1	84.4	71.9
MedGemma 27B	100.0	25.0	53.1	15.6	53.1	0.0	87.5	43.8
MedGemma 4B	100.0	34.4	25.0	0.0	59.4	0.0	75.0	46.9

3.3 Clinical Safety Indicators

We identified a fundamental bias toward pathological interpretations. While 93.8% of the ground truth images were normal, models systematically over-diagnosed abnormalities. Table 4 details these safety-critical failure modes:

- **False Alarms (Q7):** Models classified 73–93% of the 30 truly normal images as abnormal (Fig. 5(a)).
- **Over-Escalation (Q8):** In 50–96% of the 24 routine cases, models recommended unnecessary specialist referral (Fig. 5(b)).

The 95% confidence intervals on these rates (Table 4) are wide and overlap across all five models, so the differences should not be read as a reliable ranking. Using a composite score weighting Accuracy (40%), GA Accuracy (30%), and Safety (30%), MedGemma 27B ranked highest (0.312) despite its low anatomy score, highlighting the complex trade-offs required in model selection.



(a) Multi-metric Radar (b) Hallucination Heatmap

Fig. 4. (a) Radar chart showing the trade-off between metrics. (b) Heatmap of hallucination severity, showing pervasive failures in anatomy (Q1) and normality (Q7).

Table 4. Clinical safety indicators with 95% confidence intervals ↓ (lower is better). FAR conditions on the 30 normal cases, OER on the 24 routine cases, and GA Err on all 32; intervals are Wilson (proportions).

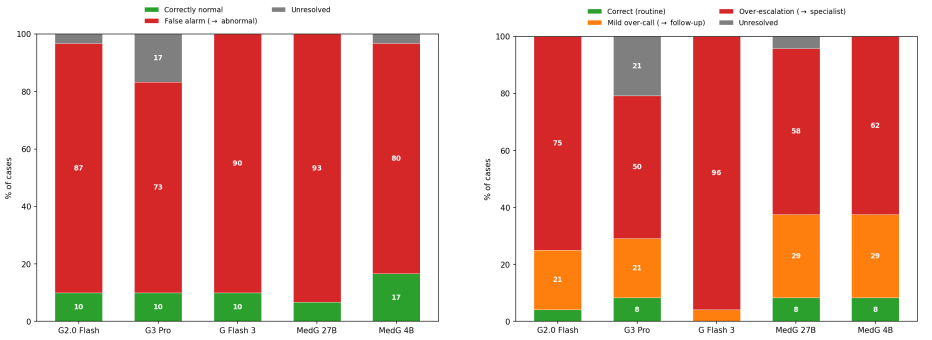
Model	False Alarm (FAR)	Over-Escalation (OER)	GA Err >4w
Gemini 2.0 Flash	86.7% [70–95]	75.0% [55–88]	56.2%
Gemini 3 Pro Preview	73.3% [56–86]	50.0% [31–69]	46.9%
Gemini 3 Flash	90.0% [74–97]	95.8% [80–99]	68.8%
MedGemma 27B	93.3% [79–98]	58.3% [39–76]	53.1%
MedGemma 4B	80.0% [63–91]	62.5% [43–79]	59.4%

4 Discussion

4.1 Interpretation of Findings

Our comprehensive evaluation reveals that while current VLMs can generate fluent medical descriptions, they exhibit a fundamental disconnect between linguistic capability and clinical accuracy. Across all five models, the dominant failure mode is not merely low accuracy, but clinically unsafe hallucination behavior.

Systematic Over-Diagnosis and Escalation: A critical finding is the strong bias toward pathology. With false-alarm rates between 73% and 93%, models systematically misclassify normal anatomy as abnormal. This drives over-escalation rates of 50–96%, where models recommend specialist referral for routine cases. In a clinical setting, such behavior would not only cause unnecessary patient anxiety but also overwhelm maternal-fetal medicine services with unwarranted consultations.



(a) Q7 normality (30 normal cases) (b) Q8 recommendation (24 routine cases)

Fig. 5. Safety analysis as outcome composition over the relevant cases (lower red is better). (a) Of the 30 normal cases, the share read as abnormal (red) is the false-alarm rate. (b) Of the 24 routine cases, the share escalated to specialist referral (red) is the over-escalation rate, with milder routine→follow-up over-calls shown in orange. "Unresolved" marks non-committal answers. The red segments equal the FAR/OER in Table 4, where the 95% confidence intervals are reported.

Anatomical and Modality Hallucinations: Reliable interpretation requires accurate structure identification (Q1), yet F1 scores remained poor (0.18± 0.05). Performance varied dramatically by anatomy category: vascular structures (Aorta: F1 0.00–0.21) and cervical imaging (F1 0.00–0.30) were nearly unrecognizable, while Thorax (F1 0.17–0.44) showed relatively better but still inadequate performance. We also observed severe modality confusion; for instance, Gemini 2.0 Flash identified cardiac structures (e.g., "four-chamber view") in all 3 abdominal images, a hallucination that renders any subsequent diagnosis invalid.

Gestational Age Unreliability: Clinical decisions often hinge on dating precision within ±1–2 weeks. The observed Mean Absolute Error (MAE) of 7–10 weeks, with over half of predictions deviating by > 4 weeks, precludes the use of current VLMs for biometric assessment. Even MedGemma 27B, which ranked highest by composite scoring, failed to achieve a clinically safe performance profile.

4.2 Limitations and Future Work

Several limitations temper these findings. The evaluation rests on 32 images from a single curated source, so the absolute numbers are best read as evidence of failure modes rather than precise population estimates, and the conditional rates rest on small denominators (30 normal, 24 routine); we therefore report confidence intervals on every rate and avoid ranking models on differences the intervals do not support. A larger, multi-centre cohort is the obvious next step. We also did not measure inter-sonographer variability on these images, which would place the model errors on a clinically meaningful scale against human disagree-

ment. The probes are structured, single-turn questions and do not reproduce the interactive, multi-image nature of a real examination. Each model was queried with one fixed prompt, and prior work on this task shows that prompt design shifts performance appreciably [8], so a systematic prompt-sensitivity analysis is left to future work. Finally, although severity followed a fixed rubric, the taxonomy retains some subjectivity, and a larger multi-rater study would strengthen it. We see calibrated uncertainty estimation and extrinsic hallucination detection, trained on resources such as the FADA dataset [3], as the most promising routes toward models that can flag their own unreliable outputs.

5 Conclusion

We evaluated five state-of-the-art VLMs on fetal ultrasound annotation across eight structured clinical questions. Despite their ability to generate detailed text, our findings indicate that current general-purpose and medical-specialized VLMs are **not yet suitable for autonomous clinical deployment** in this domain. The evaluation highlighted four critical safety gaps: (1) a **0% Safe Prediction Rate**, where every image contained at least one major or critical hallucination; (2) a **73–93% False Alarm Rate** on normality assessment; (3) a **50–96% Over-Escalation Rate** for clinical recommendations; and (4) **7–10 week errors** in gestational age estimation. These results underscore the necessity for strict human-in-the-loop oversight and the urgent need for improved hallucination mitigation strategies before VLMs can be safely integrated into obstetric imaging workflows. Future work will focus on domain-specific fine-tuning using curated fetal ultrasound datasets and the development of extrinsic hallucination detection mechanisms. Furthermore, calibrated uncertainty quantification is essential to flag low-confidence predictions before they reach the clinician.

Acknowledgements. This work has been funded by the Canadian International Development Research Centre (IDRC) under the Grant Agreement 110060-001, managed by the Global Health Institute at the American University of Beirut through the Global Health and Artificial Intelligence Network in MENA (GHAIN MENA).

Disclosure of Interests. The authors declare that they have no competing interests relevant to the content of this article.

References

1. Alasmawi, H., Saeed, N., Yaqub, M.: Fetal-gauge: A benchmark for assessing vision-language models in fetal ultrasound. arXiv preprint arXiv:2512.22278 (2025)
2. Alzubaidi, M., Agus, M.: FADA: Fetal ultrasound interpretation dataset — 56,805 expert sonographer clinical descriptions. `https://doi.org/10.5281/zenodo.20381238` (2026)

3. Alzubaidi, M., Shah, U., Muaz, R., Abbes, I., Mohammed, N., Magram, A., Alyafei, K., Househ, M., Agus, M.: FADA: Accessible fetal ultrasound interpretation and annotation with a selectively distilled unified vision-language model. arXiv preprint arXiv:2606.11106 (2026)
4. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J.A., et al.: A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. npj Digital Medicine **8**(1), 274 (2025)
5. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., et al.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
6. Díaz-Parga, C., Nuñez-Garcia, M., Carreira, M.J., Bernardino, G., Vila-Blanco, N.: Age-us: automated gestational age estimation based on fetal ultrasound images. In: Iberian Conference on Pattern Recognition and Image Analysis. pp. 83–94. Springer (2025)
7. Drukker, L., Noble, J., Papageorghiou, A.: Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound in Obstetrics & Gynecology **56**(4), 498–505 (2020)
8. ELsharif, W., Alzubaidi, M., Tukur, M., Magram, A., Anver, F., Hamza, A., et al.: Benchmarking vision llms in fetal ultrasound interpretation: a five-point expert evaluation of standard vs. custom prompts. BioMedical Engineering OnLine (2025)
9. Guan, H., Hou, P.C., Hong, P., Wang, L., Zhang, W., Du, X., et al.: A clinically-informed framework for evaluating vision-language models in radiology report generation: Taxonomy of errors and risk-aware metric. medRxiv (2025)
10. Kim, Y., Jeong, H., Chen, S., Li, S.S., Park, C., Lu, M., et al.: Medical hallucinations in foundation models and their impact on healthcare. arXiv preprint arXiv:2503.05777 (2025)
11. Lee, L.H., Bradburn, E., Craik, R., Yaqub, M., Norris, S.A., Ismail, L.C., et al.: Machine learning for accurate estimation of fetal gestational age based on ultrasound images. NPJ digital medicine **6**(1), 36 (2023)
12. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
13. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health **1**(6), e271–e297 (2019)
14. Organization, W.H., et al.: WHO recommendations on antenatal care for a positive pregnancy experience. World Health Organization (2016)
15. Patel, P., Kalantari, L., Bharat, S., Borhani, S., Schmidt, S., Jutras, M., et al.: Deep learning based gestational age estimation with outlier elimination in blind sweep fetal ultrasound. In: 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS). pp. 1–5. IEEE (2024)
16. Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Chalouhi, G., Costa, F.D.S., et al.: Isuog practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics and Gynecology **59**(6), 840–856 (2022)
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
18. Seth, A., Manocha, D., Agarwal, C.: Hallucinogen: A benchmark for evaluating object hallucination in large visual-language models. arXiv e-prints pp. arXiv–2412 (2024)

19. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., et al.: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3), 943–950 (2025)
20. Sun, Y., Wen, X., Zhang, Y., Jin, L., Yang, C., Zhang, Q., et al.: Visual-language foundation models in medical imaging: A systematic review and meta-analysis of diagnostic and analytical applications. Computer Methods and Programs in Biomedicine p. 108870 (2025)
21. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
22. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data mining and knowledge discovery handbook pp. 667–685 (2010)
23. Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., et al.: Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. Advances in Neural Information Processing Systems **37**, 94327–94427 (2024)

Appendix A: Prompts

System prompt. For each image, a fixed system role ("You are a medical imaging expert...") was followed by the eight questions below. Generation used temperature 0.4, top-p 0.95, top-k 40, and a 1024-token limit (Section 2).

Clinical questions (verbatim).

- **Q1 – Anatomical Structures Identification.** Identify and list all visible anatomical structures in the image (e.g., skull, femur, spine, heart chambers, placenta, umbilical cord, etc.).
- **Q2 – Fetal Orientation and Position.** Describe the fetal orientation and position in the image (e.g., cephalic/breech, sagittal/transverse/coronal plane, maternal left/right).
- **Q3 – Imaging Plane and Standard View Assessment.** Identify the imaging plane and assess if this is a standard diagnostic view. Specify which standard view (e.g., BPD plane, 4-chamber heart view, NT view) and its diagnostic quality.
- **Q4 – Biometric Measurements and Landmarks.** List all biometric parameters that can be measured from this image and their landmarks (e.g., BPD, HC, FL, AC, NT, CRL). Include measurement quality assessment.
- **Q5 – Gestational Age Estimation.** Estimate gestational age based on visible features, developmental markers, and structural sizes. Provide reasoning for the estimate.
- **Q6 – Image Quality and Technical Assessment.** Evaluate image quality including resolution, contrast, depth penetration, artifacts (shadowing, reverberation, noise), gain settings, and any limitations to diagnostic interpretation.
- **Q7 – Structural Normality Assessment.** Assess whether visible structures appear normal or identify any abnormalities, variations, or concerning findings. Include severity and clinical significance.

– **Q8 – Clinical Recommendations and Follow-up.** Provide clinical recommendations, suggested additional views, follow-up requirements, or need for specialist consultation based on findings.