

# Enhancing Embryo Image Interpretability through Language Models

Alberto León<sup>1</sup>, Isaac Aguilar<sup>2</sup>, Omar Paredes<sup>1,2</sup>

<sup>1</sup> Biodigital Innovation Lab,  
Translational Bioengineering Department, CUCEI,  
Mexico

<sup>2</sup> R&D Department, IVF 2.0 LTD,  
United States of America

jose.aguirre5645@alumnos.udg.mx, isaac@ivf20.com,  
omar.paredes@academicos.udg.mx

**Abstract.** Large Language Models (LLMs) have rapidly transformed numerous disciplines following the release of ChatGPT. The emergence of open-source models like LLaMA, Mistral, Phi and R1 has accelerated the Language Models adoption across diverse computational domains. Building on this foundation, Vision-Language Models (VLMs) and Multimodal Models (MMs) have extended these capabilities to process and interpret visual data alongside text. This study explores the application of Small Language Models (SMLs) in embryo evaluation, a critical area in assisted reproductive procedures. We demonstrate that SMLs can effectively interpret outputs from specialized deep learning systems, translating complex embryo features into accessible natural language descriptions for clinicians. Through comparative analysis of four models (DeepSeek, Llama 3.2 Vision, Llava:7b and Qwen2), we identify DeepSeek as most effective in generating descriptions that balance detail and conciseness. Our approach addresses two key challenges in medical AI: bridging the “black box” gap between complex deep learning and human-readable explanations and providing computationally efficient solutions suitable for clinical settings where data privacy is paramount. By enabling interpretability of specialized AI systems through lightweight language models, our methodology offers a promising direction for enhancing embryologist decision-making in *In Vitro Fertilization* (IVF) procedures while maintaining practical deployment capabilities in resource-constrained environments.

**Keywords:** Language models, small language models, interpretability, embryo.

## 1 Introduction

Language Models (LMs) have revolutionized numerous disciplines in recent years, particularly following the widespread adoption of ChatGPT as a breakthrough technology. Similar applications using language models to generate feedback through chat, text, and even images have quickly followed. Subsequently, open-source models including Llama, Mistral, Phi, and R1 have emerged, leading to the development of

language models as solutions for virtually any computational task which offer similar capabilities with lower computational requirements. These models have practical advantages in clinical settings, where efficiency and privacy are critical being the last one something important in clinical field to keep privacy of patient record and not sending this sensitive information to any other vendor, service provider or company.

The natural evolution of these models has been the emergence of Vision and Language Models (VLMs), which are LMs capable of "seeing" images, extracting their content, and generating responses to queries related to visual content. Recent work has explored their application in ophthalmology, radiology, microscopy, and embryology [1,2,5,6]. However, literature lacks focused studies on their use in reproductive medicine. Our study builds on these advances, presenting SLMs as interpretable layers for deep learning systems that process embryo images. Unlike LLMs, SLMs are deployable in constrained environments, enabling privacy-preserving workflows in sensitive medical domains.

Our study builds on this trend by proposing the use of these models to enhance interpretability of AI systems used in specialized medical image evaluation, enabling multimodality capability previously limited to private models such as ChatGPT [OpenAI] and Claude [Anthropic]. This limitation has presented obstacles for medical applications where patient data privacy is paramount.

Deep learning systems have long been characterized as "black boxes" [3], making it difficult or impossible to understand how they process information and reach conclusions. The interpretability challenge varies depending on the neural network architecture and outputs generated. As Afnan et al. assert [4], "the sacrifice of interpretability is something that cannot be justified". Language models, with their ability to abstract information and analyze it within broader contexts (dependent on their training data), may clarify complex values and provide better interpretability for sophisticated deep learning models, particularly for non-specialists. Recent work demonstrates this potential, using vision models and ChatGPT as tools for generating medical image diagnoses [4,5].

In this paper, we present a proof-of-concept protocol using SLMs as interpreters of results from other deep learning models whose outputs are abstract to human understanding. SLMs are particularly suitable for this application as they are computationally efficient and have narrower context windows, potentially reducing the risk of output deviation. We apply this approach to assisted reproduction, specifically focusing on the critical task of embryo evaluation for transfer and feature description. This application demonstrates how SLMs can bridge the gap between complex AI outputs and human-understandable information in sensitive medical contexts.

## 2 Methodology

### 2.1 Small Language Model Selection

In this study, we evaluated several state-of-the-art Small Language Models (SLMs) with multimodal capabilities, selected for their balance of performance and efficiency. These models represent different approaches to combining language and vision capabilities while maintaining reasonable computational requirements as show in Table

**Table 1.** SMLs used, their parameters length and publisher.

Model	Parameters	Publisher
Llama 3.2 Vision	11B	Meta
Llava	7B	Haotuan Liu
MiniCPM-V	8B	OpenBMB
DeepSeek	14B	DeepSeek

1. Each model offers distinct advantages in processing medical imagery and interpreting complex features.

**Llama 3.2** Is a foundational language model developed by Meta, distinguished by its dense Transformer architecture. It demonstrates multilingual capabilities and strong logical reasoning, showing performance comparable to GPT-4 in various evaluation tasks. The model was trained on 15.6 billion multilingual tokens, with particular emphasis on data quality and diversity [6]. For this work, we utilized the Vision variant with 11 billion parameters.

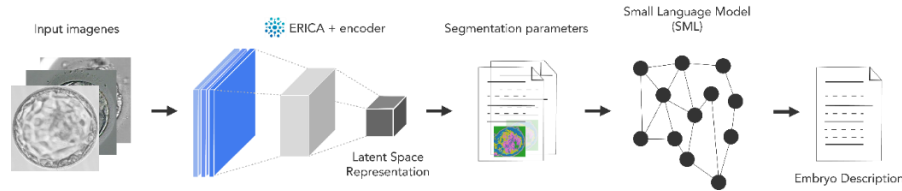
**Llava** Large Language and Vision Assistant (Llava) is a multimodal model developed using the "Visual Instruction Tuning" strategy. It combines a visual encoder based on CLIP with a Llama language model and was trained on data automatically generated by GPT-4. Llava excels at following detailed visual instructions and executing complex reasoning tasks at a level comparable to GPT-4 [7]. For this research, we employed the 7 billion parameter variant.

**MiniCPM-V** belongs to a family of lightweight multimodal models developed by the OpenBMB team, specifically designed for efficient operation in constrained systems such as mobile or embedded devices. The most recent model (version 2.6) integrates adaptive vision techniques with the Qwen2 language model. It distinguishes itself through its ability to process high-resolution images (up to 1.8 million pixels) and, when combined with quantization processes, offers exceptional performance and efficiency [8]. For our work, we utilized the 8 billion parameter version.

**R1** is an advanced language model developed by DeepSeek-AI, specifically designed to enhance reasoning abilities through reinforcement learning. The final version combines initial cold-start data, supervised fine-tuning, and reasoning-oriented reinforcement learning, enabling it to achieve results comparable to leading models such as OpenAI-o1-1217 in cognitive tasks and complex questions. The model offers distilled variants based on architecture like Qwen and Llama, making it more lightweight [9]. For this research, we employed the 14 billion parameter version.

## 2.2 Experimental Design

**Research Collaboration and Data Acquisition** This research was conducted in collaboration with IVF 2.0 LTD, which provided anonymized embryo images and access to their proprietary ERICA API [10] for image processing and feature extraction. A total of 30 anonymized embryo images were supplied for testing across the different SMLs.



**Fig. 1.** Pipeline to extract features and parameters for the SML to generate a description of an embryo image.

**Image Preprocessing and Feature Extraction** The raw images provided by IVF 2.0 were processed by our team to ensure consistency. We cropped each embryo image to standardize dimensions across the dataset, preparing them for subsequent feature extraction.

The standardized images were processed through the ERICA API [10], which extracted morphological zone features in tensor format and generated segmentation masks. These features represented key embryological characteristics used by specialists in embryo evaluation, such as Pellucid Zone, Trophectoderm and Inner Cell Mass segmentation briefly described on Figure 1.

We designed two primary testing protocols:

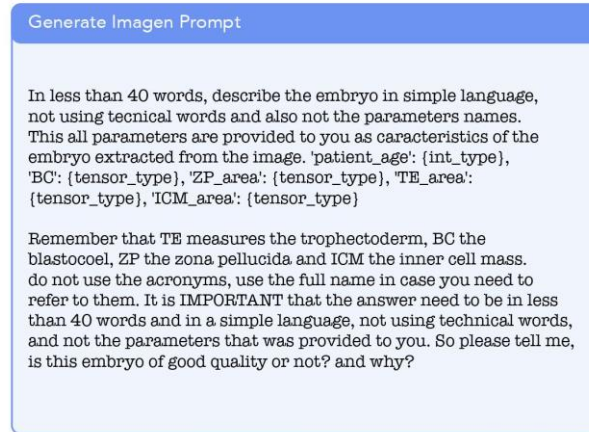
**Feature Interpretation Test:** SMLs were prompted with comprehensive embryo features and segmentation masks to interpret the image.

**Quality Assessment Test:** SMLs evaluated embryo quality based on the parameters extracted from images and characteristics provided by ERICA.

## 2.3 Implementation Framework

**Technical Integration and Experimental Process** We accessed the language models through the Ollama platform [11], which provides open-source language models via a straightforward Python API. Model versions from the Ollama repository.

Using the Ollama API, we generated a fresh chat instance for each model-image combination to prevent cross-contamination between descriptions. For each image, we created a prompt containing both the image and its extracted characteristics as shown in Figure 2. The resulting descriptions were saved in a data frame for subsequent analysis, as outlined in the next section.



**Fig. 2.** Prompt used to generate the descriptions of each image for all SMLs.

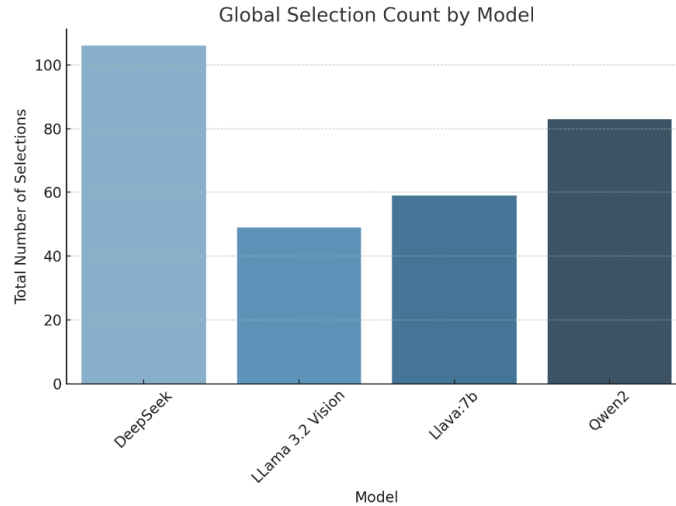
**Technical Requirements and Reproducibility Guidelines** Our testing environment consisted of a computer equipped with a Nvidia RTX 4080 (16GB VRAM) graphics card, Intel 14700K processor, and 64GB DDR5 RAM. The most computationally intensive component was the Ollama Chat API interacting with the language models, which primarily utilized the GPU.

To replicate similar experiments, the following are necessary: the Ollama package, Python for scripting, and appropriate models selected from ollama.com. Researchers may also employ vision model encoders or implement alternative feature encoders such as OpenAI's CLIP or convolutional network models. Hardware specifications may vary as the primary computational demand is on the GPU for model inference.

### 3 Results

We analyzed participants' preferences regarding descriptions generated by four different language models: DeepSeek, Llama 3.2 Vision, Llava:7b, and Qwen2. Each model provided a description for 27 embryo images, resulting in a total of 108 descriptions. Participants selected the description they considered most appropriate or appealing for each image.

Our analysis revealed that DeepSeek was the most preferred model, receiving 106 total selections across all participants. Qwen2 followed with 83 selections, while Llava:7b and Llama 3.2 Vision received 59 and 49 selections respectively see in Figure 3. This distribution suggests a clear preference hierarchy among the models when interpreting morphological features.



**Fig. 3.** Global selection counts of descriptions generated by each model.

**Table 2.** Selection for each model and average number of words on the model descriptions.

Model	Selection count	Avg # words
DeepSeek R1	106	45.35
MiniCPM-V	83	56.25
Llava	59	75.32
Llama 3.2 Vision	49	34.46


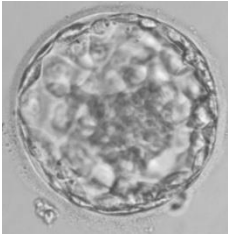
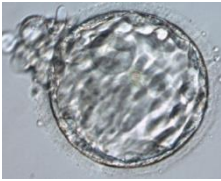
As clarification for the number of images from 30 to 27, these 3 remaining images were corrupted and unable to process so they were discarded from the experiment.

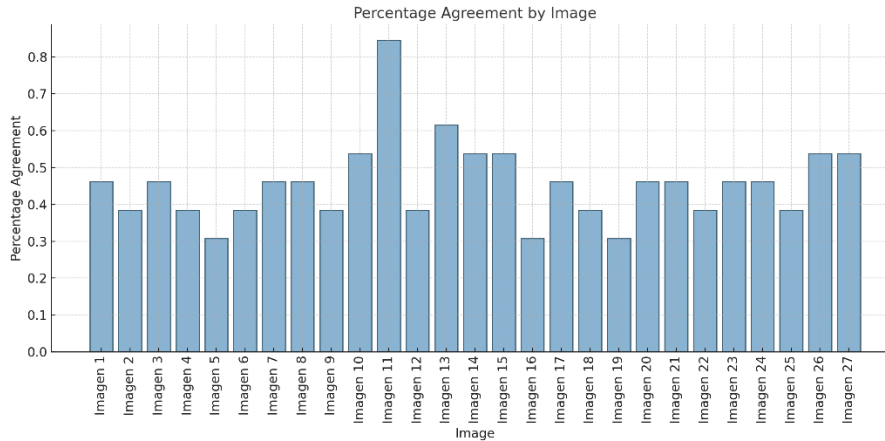
To investigate whether description length influenced participants' preferences, we analyzed the average word count of descriptions generated by each model see in Table 2. Llava:7b produced the longest descriptions with an average of approximately 75 words per description, followed by Qwen2 with about 56 words, DeepSeek with about 45 words, and Llama 3.2 Vision with the shortest descriptions averaging only 34 words.

This analysis reveals that longer descriptions did not necessarily correlate with higher preference rates. Despite generating substantially longer descriptions, Llava:7b ranked third in overall participant preferences. In contrast, DeepSeek, with its more concise descriptions of moderate length, was significantly favored by participants. This suggests that moderate description length may optimize clarity and general acceptance among users, potentially striking an ideal balance between providing sufficient detail while remaining accessible.

Interestingly, Llama 3.2 Vision, which produced the shortest descriptions, received the fewest selections. This indicates that excessively brief descriptions may fail to provide enough detail to satisfy participants' information needs when interpreting complex embryological features. The relationship between description length and preference appears to follow an inverted U-shaped curve, with optimal length falling in the moderate range represented by DeepSeek's descriptions.

**Table 3.** Descriptions generated by each model of embryo images 1, 5 & 11.

Images	Llama3.2 vision	Llava 7B	MiniCPM-v (Qwen2)	DeepSeek R1
	The embryo appears to have <b>**poor quality**</b> . While it is euploid (normal chromosome count), the measurements show unusual proportions between key structures like the inner cell mass and trophectoderm, and some areas are too small or misshapen, which can affect its developmental potential.	Based on the image's characteristics, it is difficult to determine if the embryo is of good quality as there are several factors affecting its quality. The size of the embryo, which includes the trophectoderm (TE), inner cell mass (ICM), and zona pellucida (ZP), may not be enough to assess its quality. Furthermore, other factors such as the number of cells, the presence of fragmentation or abnormalities in the embryo's shape, and the overall appearance cannot be assessed without further information.	The embryo appears to have various irregularities such as decreased circularity, increased eccentricity, and low area-to-perimeter ratio. These factors suggest it may be of lower quality compared to a well-formed embryo with high circularity and proper proportions between size metrics. The presence of an atypical ICM also indicates potential issues in the embryo's development.	The embryo appears to have <b>**poor quality**</b> . While it is euploid (normal chromosome count), the measurements show unusual proportions between key structures like the inner cell mass and trophectoderm, and some areas are too small or misshapen, which can affect its developmental potential.
	This embryo appears to be of good quality. The zona pellucida is thick enough, the inner cell mass is well-formed, and the trophoblast layer is even. However, it's always best to consult a doctor for an accurate assessment.	This embryo has a large and well-defined inner cell mass (ICM) and trophectoderm (TE), suggesting it may be of high quality. The large size of the blastocoele (BC) is typical for an early-stage embryo, while the zona pellucida (ZP) appears to be thinning out. Overall, these characteristics suggest that this embryo is of good quality and likely has a good chance of implanting successfully.	This embryo appears euploid with healthy cell division. The blastocoele has no significant defects, zona pellucida looks intact but slightly elongated in shape, the inner cell mass seems well developed without any signs of abnormality, and the trophectoderm is present too. Based on these features, it can be considered as a good quality embryo.	This embryo appears to be of good quality because it is euploid (the correct number of chromosomes), has normal-sized areas for its structures, and shows proper development of the inner cell mass, which will form the future baby.
	The embryo appears to be of good quality. It has a normal number of chromosomes (euploid) and its cells are well-organized. The measurements suggest a healthy balance between the outer layers and the inner cell mass.	Based on the image, the embryo appears to be of poor quality. The reason for this assessment is the low values for BC_circularity (0.257) and ICM_circularity (0.634), which indicate that the embryo has an irregular shape. Additionally, the TE_ICM_area_ratio (0.844) suggests a lack of proper development in the inner cell mass. It is important to note that these assessments are based solely on the image provided and do not consider other important factors such as chromosomal status or the overall health of the embryo.	The embryo appears euploid with a well-defined structure. However, it has low circularity values for BC, ZP, ICM, and TE, indicating potential issues in cell shape and organization. Overall, the quality seems to be average but could improve.	This embryo appears to have irregular shapes for its blastocoele, trophectoderm, and inner cell mass, with very low circularity in all areas. The zona pellucida also has a highly irregular shape. These features suggest the embryo may not be of optimal quality due to fragmentation or poor organization.



**Fig. 4.** Percentage of Agreement by image.

The preference patterns observed suggest that when interpreting specialized medical imagery, particularly in the context of embryo evaluation, users value descriptions that balance comprehensiveness with conciseness. This finding has important implications for the design of AI-assisted medical interpretation systems, suggesting that optimizing for moderate description length may enhance user satisfaction and potentially improve the clinical utility of such systems.

Participants evaluators were from a diverse but short number of professionals, such as embryologists, biologists and technicians in the field of medicine (medical engineers) in total 11 participants the method to get the selections and feedback was through a form (feedback was no considered in this paper because tend to be subjective and limited to the words of the description y future experiments we will try get a description to compare model vs professional), so they can see image + description by each model as see in Table 3.

## 4 Discussion

### 4.1 Variability in Model Agreement

The analysis of percentage agreement for descriptions generated by language models for embryo images revealed significant variability across the dataset Figure 4. Image 11 demonstrated the highest consensus among language models, exceeding an 80% agreement level. This indicates a strong convergence towards a particular standardized description, which was selected as the gold standard due to its precision and clarity.

The gold-standard description for image 11 Table 3, provides a clear, detailed characterization of structural irregularities in the blastocoel, trophoctoderm, inner cell mass, and zona pellucida, emphasizing critical quality indicators such as fragmentation and low circularity. This precision likely facilitated high agreement among language models, as clear pathological indicators were easily identifiable, leading to a unified interpretation.



Conversely, image 5 Table 3 exhibited the lowest percentage agreement and greatest dispersion among generated descriptions. The descriptions generated for this image varied significantly, highlighting different morphological features and resulting in contrasting interpretations of embryo quality. For example, one description emphasized a thinning zona pellucida and a large blastocoel as indicators of good quality and high implantation potential, while another pointed out an elongated zona pellucida with less definitive implications for embryo viability. Such discrepancies reflect inherent ambiguity in the embryo's morphological features, resulting in varied diagnostic interpretations by the language models.

#### **4.2 Interpretability Challenges in Embryo Assessment**

This variability underscores the inherent complexity of accurately evaluating embryos with intermediate or subtle morphological features. Embryos with clearly identifiable abnormal characteristics such as those seen in third image from Table. 3, facilitate higher levels of agreement due to the unambiguous pathological features. Conversely, embryos like second imagen Table. 3, with subtle or ambiguous morphological indicators, pose challenges to automated diagnostic tools, resulting in decreased consensus and diagnostic confidence.

These findings highlight important limitations and opportunities for improvement in the use of language models for embryo assessment. Specifically, training language models with standardized, detailed, and precise descriptions as benchmarks appears essential to enhance their diagnostic accuracy, particularly for embryos presenting intermediate or unclear morphological characteristics. Future research could focus on incorporating highly standardized references during training phases to improve model convergence and accuracy, thus facilitating more consistent and reliable embryo assessments.

#### **4.3 Model Performance and User Feedback**

The performance of the models in terms of efficiency and hallucination met our initial expectations. The survey conducted among biology professionals yielded predominantly positive feedback regarding the models' ability to accurately describe embryo characteristics. Participants specifically noted the models' effectiveness in evaluating cell quality, suggesting that the descriptions generated by SMLs were clinically relevant and potentially useful in practice.

DeepSeek emerged as the preferred model among professionals, likely due to its optimal balance of description length and detail. This preference pattern suggests that concise, yet comprehensive descriptions may be most effective for clinical interpretation of embryological features. The moderate description length (averaging 45 words) provided by DeepSeek appears to strike an ideal balance between providing sufficient detail while remaining accessible to users.

#### **4.4 Implications for AI Interpretability**

Our findings demonstrate that Small Language Models (SMLs) can effectively generate descriptions and evaluations based on analysis from another AI model or Deep

Learning system with relatively little context. This capability represents a promising approach to AI interpretability, particularly in specialized medical domains like embryology. By translating complex feature representations from deep learning models into natural language descriptions, SMLs can potentially bridge the gap between sophisticated AI systems and human practitioners.

This interpretability layer addresses a critical limitation of deep learning systems in healthcare—their "black box" nature. By providing human-understandable explanations of AI-derived assessments, SMLs may enhance trust, facilitate error detection, and improve the clinical utility of AI systems in assisted reproduction technologies.

In summary, this analysis emphasizes the importance of clearly defined, standardized descriptors in training language models for embryological applications, particularly in cases where the morphological quality of embryos is not clearly delineated. The study demonstrates the potential of SMLs as interpretability tools for more complex AI systems, suggesting a promising direction for developing more robust assistants for interpretability tasks in healthcare and beyond.

## **5 Conclusion and Future Work**

### **5.1 Summary of Findings**

This research demonstrates that SMLs can effectively generate accurate descriptions of embryonic images using a Zero-Shot approach. Their key advantage is providing interpretability for complex Deep Learning models through accessible natural language descriptions. Among the tested models, DeepSeek performed best by balancing detail and conciseness, suggesting moderate description length optimizes clarity for specialist users.

### **5.2 Limitations**

Our study has several limitations: the accuracy of SML-generated descriptions requires systematic validation against expert assessments; our evaluation relies on preference judgments rather than objective measures; and the sample size (27 images) limits generalizability across the full spectrum of embryo morphologies encountered clinically.

### **5.3 Future Work**

Future research directions include:

- Collecting descriptions from senior embryologists to measure semantic distance with SML outputs.
- Evaluating textual quality using standard metrics like BLEU scores.
- Training models with specialized datasets containing detailed embryo descriptions and clinical outcomes.
- Investigating integration into clinical embryology workflows.

- Exploring additional data modalities like time-lapse imaging or genetic testing results.

## 6 Conclusion

SMLs show promise as interpretability tools for complex AI systems in reproductive medicine. By translating abstract features into accessible descriptions, they address the "black box" problem in healthcare AI. With domain-specific training, these models could develop into robust tools for embryo evaluation that enhance accuracy and clinical utility in assisted reproduction.

**Acknowledgments.** This study was possible thanks to IVF 2.0 LTD for providing data and computational resources. Thanks to Gustavo Guzmán for the figures.

## References

1. Hong Lee, K., Lee, R., Eun Kwon Y.: Validation of a Deep Learning Chat X-ray Interpretation Model: Integration Large-Scale AI and Large Language Models for Comparative Analysis with ChatGPT. *Diagnostics* (2023).
2. Lingxuan, Z., Weiming M., Ancheng L.: Step into The Era Of Large Multimodal Models: A Pilot Study On Chatgpt-4(Visions) Ability To Interpret Radiological Images. *International Journal of Surgery* (2024).
3. Lee, T., Natalwala J., Chapple, V.: A brief history of artificial intelligence embryo selection: from black-box to glass-box (2024). doi: <https://doi.org/10.1093/humrep/dead254>.
4. Afnan, M.A.M., Liu, Y., Conitzer V.: Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction* (2021): hoab040.
5. Lim, G., Elangovan, K., Jin, L.: Vision Language Models in Ophthalmology (2024).
6. Bosbach., W., Jan F., Senge, B.: Ability Of Chatgpt To Generate Competent Radiology Reports For Distal Radius Fracture By Use Of RSNA Template Items And Integrated AO Classifier. *Current problems in diagnostic radiology* (2023).
7. Meta AI.: Llama 3: A Family of Large Language Models. (2024) url: <https://arxiv.org/abs/2402.06687>.
8. Liu, H., Li, C., Wu, Q.: Visual Instruction Tuning. (2023) url: <https://arxiv.org/abs/2304.08485>.
9. OpenBMB. (2024). MiniCPM-V: Lightweight Multimodal Models for Efficient Deployment. <https://arxiv.org/abs/2402.15733>
10. DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. <https://arxiv.org/abs/2501.12948>
11. Chavez, A., Flores-Saiffe, A., Mendizabal, G., Drakeley, A., Cohen, J., Embryo Ranking Intelligent Classification (ERICA): Artificial Intelligence clinical assistant prediction embryo ploidy and implantation (2020). <https://doi.org/10.1016/j.rbmo.2020.07.003>.
12. Ollama Homepage, <https://ollama.com/> last accessed 2025/03/07.
13. Zhang, J., Huang, J., Jin, S., & Lu, S. (2023). Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Patter Analysis and Machine Intelligence*, 46, 5625-5644. <https://www.semanticscholar.org/reader/f38bf22a5ceef785d6a15263fad3f22f623a3e6b>