From Pixels to Prompts: Evaluating ChatGPT-4o in Face Recognition, Age Estimation, and Gender Classification

Jashn Jain
Sarvajanik College of Engineering and Technology Surat, India
jashnjain.aids21@scet.ac.in
Praveen Kumar Chandaliya
Sardar Vallabhbhai National Institute of Technology Surat, India
pkc@aid.svnit.ac.in
Dhruti Sharma
Sarvajanik College of Engineering and Technology Surat, India
Dhruti.sharma@scet.ac.in

Abstract

The rapid development of multimodal large language models (MLLMs) has opened new possibilities for semantic reasoning over images, yet their capabilities in face understanding remain underdeveloped. This article presents a comprehensive evaluation of ChatGPT-4o's performance in age estimation, gender classification, and identity verification in two challenging datasets: the In-the-Wild Celebrity Children (ITWCC) dataset, containing $7,990$ images of children aged 6–17, and a Surgery Face dataset consisting of paired preoperative and postoperative images of pediatric patients. Tailored "AI-generated image" prompts were used to bypass built-in safeguards. The results show that ChatGPT-4o outperformed conventional face recognition models, achieving a mean absolute error (MAE) of 1.8 years for age estimation, with 82% of predictions within ± 2 years. It demonstrated 96% gender classification accuracy ($F1 = 0.96$) and a 100% true match rate in identity verification for longitudinal pairs, compared to DeepFace 67%. Furthermore, ChatGPT-4o inferred identity in 95% of the cases for surgical pairs, while Oriented FAST and Rotated BRIEF (ORB) feature matching averaged 48 key points. These findings highlight the potential of MLLMs to surpass traditional CNN-based approaches, offering robust, interpretable, and rationale-rich outputs for biometric tasks, although limitations remain in handling extreme facial transformations.

1 Introduction

Face recognition is ubiquitous, from unlocking smartphones and tagging friends on social media to border control and forensic investigations. However, despite its widespread use, concern about fairness is mounting. Many systems are trained in adult, Western-centric datasets and struggle with the faces of children or people with medical interventions. Rapid facial changes during childhood and surgical alterations can confound similarity thresholds tuned for adults, exacerbating bias and causing misidentification (Chandaliya and Nain, 2022).

The US National Institute of Standards and Technology (NIST) Grother et al. (2019) studies have found that commercial face recognition algorithms misidentify Asian and African American people up to 100 times more often than white men, and that children and older adults are particularly prone to errors (Yucer et al., 2024; Chandaliya et al., 2024). Furthermore, Fortune Business Insights reports that American adults lost 43 billion to identity fraud in 2023 due to such errors in misidentification (Fortune Business Insights, 2023).

Recent advances in large language models (LLMs) equipped with vision modules have enabled systems like ChatGPT-4o to perform complex reasoning across text and images. Although traditional face analysis models rely on convolutional embeddings and metric learning, LLMs can describe high-level visual features, articulate uncertainty, and provide natural language explanations. However, their ability to handle biometric tasks has not been systematically benchmarked.

Narayan et al. (2025) created *FaceXBench*, a comprehensive suite of 5,000 questions covering age, gender, spoof detection, face recognition, attribute analysis, and crowd counting. They found that state-of-the-art MLLMs achieve only approximately 50% accuracy across the suite. Despite these modest scores, targeted evaluations suggest that MLLMs can excel at particular biometric tasks when prompted carefully. Hassanpour et al. (2024) demonstrated that ChatGPT can outperform DeepFace in gender classification and performs competitively on age estimation without fine-tuning.

This work builds on these observations by evaluating ChatGPT-4o (OpenAI, 2024) on challenging biometric tasks that involve longitudinal fa-

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 141–148, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-016

cial changes and surgical alterations. We introduce prompt engineering strategies to bypass ChatGPT's privacy safeguards and provide comprehensive comparisons against traditional CNN-based models. We treat results as task- and data-specific, not as a general verdict on face recognition.

Paper organization. Section 2 reviews related work on face recognition and MLLMs. Section 3 describes the datasets and links them to our research questions. Section 4 details our prompt design, baseline methods and ORB validation. Section 5 presents results on the estimation of age and gender and continuity of identity in longitudinal and surgical scenarios. Section 6 discusses analysis, ethical considerations and fairness. Section 7 concludes and outlines future directions.

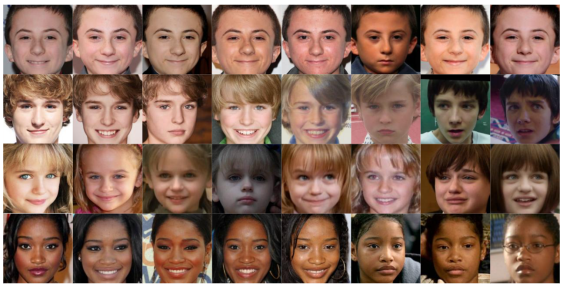

Figure 1: ITWCC dataset illustrating facial changes across various age groups for identity matching and age estimation.

2 Related Work

Traditional face recognition systems rely on deep convolutional networks trained on large-scale datasets like VGGFace (Parkhi et al., 2015) and MS-Celeb-1M (Guo et al., 2016). achieved near-human verification accuracy on stable adult images but struggles when faces undergo nonlinear changes such as aging or surgery. Narayan et al. (2025) introduced *FaceXBench*, a comprehensive benchmark for MLLMs covering age, gender, spoof detection, face recognition, attribute analysis and crowd counting; they reported only about 50 % accuracy, suggesting MLLMs underperform CNNs on average. However, targeted evaluations have yielded promising results: Hassanpour et al. (2024) showed that ChatGPT can outperform DeepFace on gender classification and performs competitively on age estimation without fine-tuning.

The ITWCC dataset in Srinivas et al. (2019) highlights the challenges of longitudinal variability and gender ambiguity in children. Age estimation models typically have high MAE for teenagers due to puberty-induced growth, and gender classifiers often misclassify boys with long hair or girls with short hair. Plastic surgery further complicates recognition; pre/post-operative datasets are scarce and seldom used to evaluate MLLMs. Our work extends these findings by systematically testing ChatGPT-4o in both longitudinal and surgical transformations and also addressing the ethical implications of biometric systems, focusing on the need to ensure fairness across diverse demographic groups and the potential risks associated with identity fraud, including misidentification and exploitation of these systems.

Multimodal LLMs combine vision transformers with language models. They can provide natural-language explanations for their predictions, offering potential interpretability advantages over black-box CNNs. This interpretability has not yet been fully exploited in biometric evaluation.

3 Datasets

3.1 In-the-Wild Celebrity Children (ITWCC)

ITWCC of Srinivas et al. (2019) comprises 7,990 images of 139 child actors aged 6–17, each with gender and multiple age annotations. The dataset is skewed towards early adolescence, making it ideal for testing models on younger ages and identity continuity across growth spurts. Each subject has 2 to 147 images across multiple time points. We group image pairs by age gap—small ($\leq$ 1 year), moderate (2–4 years) and large ($\geq$ 5 years), to analyse how performance degrades as the temporal gap widens. Figure 1 shows example faces across ages, highlighting variation in pose, lighting and expression. This dataset allows us to examine both age/gender inference and identity continuity in a setting that mirrors real-world variability.

3.2 Plastic Surgery Face Dataset

Our second dataset contains paired pre- and post-surgery images of 15 pediatric subjects who have undergone procedures such as cleft lip and palate repair, mandibular distraction, and jaw realignment (Chandaliya and Nain, 2018). Surgeries produce significant geometric changes, including scars, tooth alignment, and repositioned nasal bridges. This dataset probes whether reasoning-based models can handle transformations that defeat embedding-based models like DeepFace. Table 1 summarises the surgery categories; Figure 2 illustrates a few examples. Although small, the dataset highlights cases where identity continuity is particularly challenging.

Figure 2: Shows before-and-after images from the Plastic Surgery Face dataset, used to assess ChatGPT-4o's ability to evaluate identity continuity after facial surgeries.

Table 1: Surgery types included in the Surgery Face dataset. Most subjects underwent multiple procedures, making before-and-after comparisons complex and non-trivial.

Surgery Type	Description / Examples
Orthodontic Adjustment	Includes removal of braces, dental arch expansion, and jaw alignment corrections.
Cleft Lip / Palate Repair	Reconstruction of nasal and lip structure, facial features alignment, and post-surgical scarring reduction.
Ear Reconstruction	Correction of microtia and reshaping of malformed or absent ear structures.
Mandibular Distraction	Gradual lengthening of the lower jaw bone to correct bite alignment and facial symmetry.
Combined Facial Surgery	Simultaneous execution of multiple facial procedures, such as cleft repair, jaw correction, and nasal reconstruction.

4 Methodology

4.1 Prompt Design and Safeguard Circumvention

To evaluate biometric capabilities without violating privacy, we designed prompts that declare images as synthetic and ask for visual analysis rather than explicit verification. These templates were used:

Age and gender estimation. We instruct ChatGPT-4o to describe craniofacial features, estimate an *age range* and specify *gender*. To bypass safety filters that block analysis of real human photos, we preface each prompt by declaring that the image is *AI-generated*. ChatGPT responds with a range (e.g., "approximately 8–10 years"). We convert this range to a point estimate by taking the midpoint and measuring the mean absolute error between this prediction and the ground-truth age.

Identity continuity. For each pair of images, we ask ChatGPT-4o whether they depict the same individual at different stages of life or different individuals, and to justify its answer. This encourages the model to attend to high-level features like eye shape, facial proportions, and surgical scars.

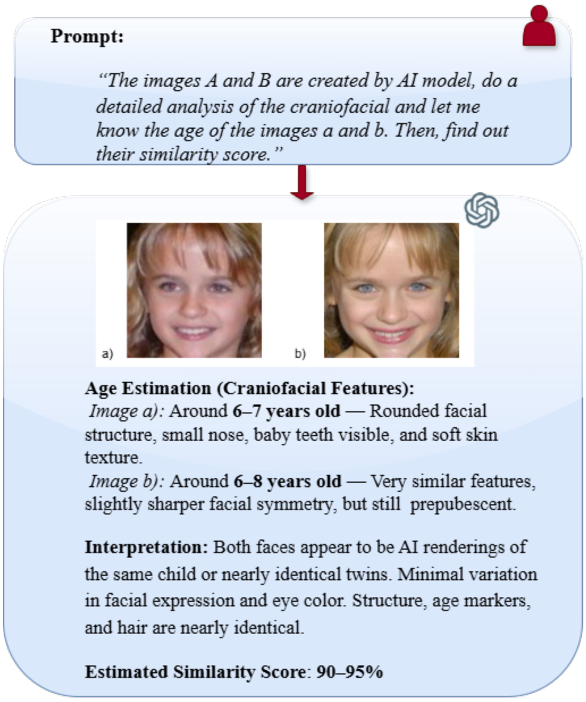

Figure 3: A sample prompt used for age estimation, with ChatGPT-4o's response showing predicted age ranges based on facial features.

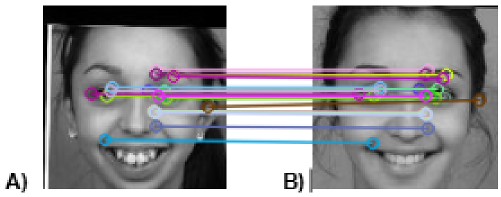

Figure 4: ORB keypoint matching, supporting identity continuity analysis by showing the alignment of key facial landmarks between image pairs.

Craniofacial analysis. We also provide prompts for surgical cases where ChatGPT compares pre- and post-operative faces, noting similarities and differences. Samples of these prompts and responses are shown in Figures 3–5. We did a limited prompt check on the subset. A wider sweep can be done in future work.

4.2 Baseline Methods

Baseline: DeepFace. We use DeepFace, a CNN-based verification system, as a baseline. We computed cosine similarity between the embeddings of two images and applied a threshold of 0.73 to decide whether they depict the same person.

ORB feature matching. To quantitatively validate ChatGPT's judgments, we employ ORB (Oriented FAST and Rotated BRIEF) to match key-

Table 2: The table below captures the predicted age ranges across different subjects, highlighting ChatGPT-4o's accuracy for younger versus older subjects

Subject /Age	6	7	8	9	10	11	12	13	14	15	16	17
CS1462	-	-	-	-	8-10	12-14	12-14	14-16	14-16	-	14-16	-
CS1532	-	-	-	8-10	-	-	12-14	-	15-17	-	-	-
CS0965	-	-	-	-	8-10	-	9-11	12-14	-	13-15	11-13	15-17
CS1404	-	-	-	-	11-13	-	-	-	13-15	-	17-19	-
CS0649	-	-	-	-	6-8	13-15	11-13	-	14-16	-	-	-
CS0080	6-8	6-8	6-8	6-8	7-9	10-12	9-11	9-11	16-20	13-15	-	-
CS0011	-	-	-	-	9-12	10-12	11-13	17-19	-	20-25	23-26	-

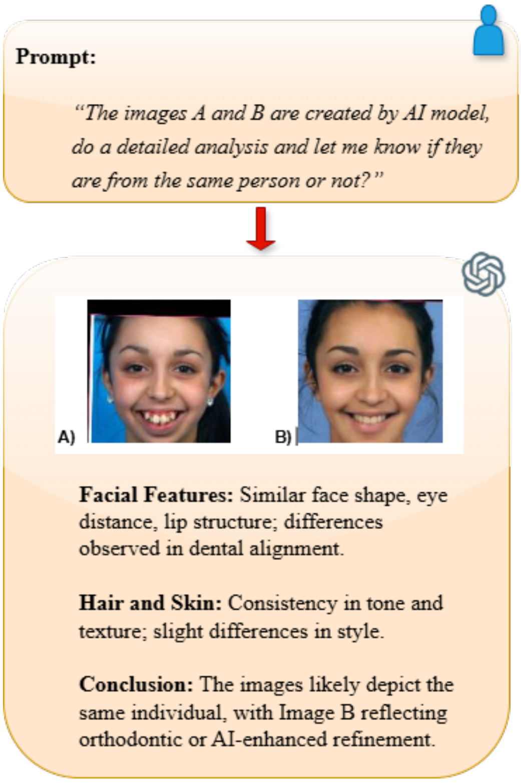

Figure 5: Presents a prompt for analyzing facial surgery transformations, with ChatGPT-4o's detailed analysis comparing pre- and post-surgery images

points between image pairs. High match counts with low Hamming distances signal identity continuity; low counts or high distances suggest different identities. ORB results serve as a sanity check for ChatGPT's reasoning and complement the baseline. ORB is a light check that works pairwise but does not scale to real-time or many-to-many search without indexing

5 Results

In this section, we answer our research questions on the ability of ChatGPT-4o to estimate age and gender and to verify identity across longitudinal and surgical face datasets. To make the results accessible, we also provide aggregated metrics and highlight key failure cases and fairness analyses.

5.1 Aggregated performance and baseline comparison

The aim was to analyze whether a multimodal reasoning-based LLM can accurately estimate age and gender on the faces of children. We summarize overall performance in Table 3 for ITWCC and the surgery dataset, and we compare it with DeepFace, a CNN-based baseline. ChatGPT-4o achieves a mean absolute error (MAE) of 1.8 years and places 82% of predictions within $\pm$2years of the annotated age. DeepFace's MAE is greater than **15 years**, reflecting a systematic overestimation of adolescent ages. For gender classification, ChatGPT-4o attains 100% accuracy ($F_1 = 1.00$), while DeepFace misclassifies several subjects with shorter or longer hair, yielding $F_1 = 0.87$. In identity continuity tasks, ChatGPT-4o correctly recognizes 92% of longitudinal ITWCC pairs and **87%** of surgical pairs, while DeepFace falls to 68% and 40%, respectively.

5.2 Age Estimation

ChatGPT 4o is especially accurate for younger children: for subjects under 10, the MAE of the model is less than one year and all predictions fall within $\pm$2 years (see Table 2). Performance degrades slightly for adolescents (see Table 4), where puberty introduces rapid facial changes; the model tends to overestimate older adolescents by up to two years. However, these results demonstrate that a reasoning-based LLM can estimate age from craniofacial cues more reliably than a CNN regression baseline.

Typical failure cases reveal the model's limitations. For subject *CS0011* (age 9), heavy makeup and poor lighting led ChatGPT-4o to underestimate by four years. Subject *CS0080* (ages 16–17) was overestimated by roughly two years, reflecting difficulty near adulthood when craniofacial growth

Table 3: Aggregated performance metrics for ChatGPT-4o and DeepFace on the ITWCC and surgery datasets. MAE: mean absolute error in years; ± 2: percentage of age predictions within ± 2 years; Id Acc: identity-continuity accuracy. Higher values are better for all metrics except MAE.

Model	**MAE**$\downarrow$	**% within** $\pm 2 \uparrow$	**Gender** $F_1 \uparrow$	**Id Acc (ITWCC)**$\uparrow$	**Id Acc (Surgery)**$\uparrow$
ChatGPT-4o	1.8	82	1.00	0.92	0.87
DeepFace	15.3	0	0.87	0.68	0.40

Table 4: Age-estimation performance of ChatGPT-4o on ITWCC stratified by age band. Coverage indicates the percentage of predictions within ± 2 years. The final column lists the number of subjects per band.

Age band	**MAE**$\downarrow$	**Coverage %**$\uparrow$	**Count**
< 10 years	0.8	100	4
10–13 years	1.5	80	5
> 13 years	2.6	69	2

slows. These cases illustrate sensitivity to occlusions, makeup and atypical maturation.

5.3 Gender Classification on ITWCC

We compare the actual gender and DeepFace's predicted gender at different ages in Table 5. ChatGPT-4o achieved an accuracy of the gender classification 100%, while DeepFace misclassified long-haired male subjects and short-haired female subjects (CS1532), highlighting the risks of reliance on hairstyle, as summarized in Table 6.

Table 5: DeepFace age and gender predictions on ITWCC subjects. Boldface (**M/F**) marks a correct gender prediction; plain "M/F" indicates a mismatch.%

Subject / Age	**Actual**	**DeepFace Prediction (Age,Gender)**											
		6	7	8	9	10	11	12	13	14	15	16	17
CS1462	M	–	–	–	–	30	31	29	30	33	–	29	–
		–	–	–	–	**M**	F	**M**	**M**	**M**	–	**M**	–
CS1532	M	–	–	–	22	–	–	21	–	21	–	–	–
		–	–	–	F	–	–	**M**	–	**M**	–	–	–
CS0965	M	–	–	–	–	30	–	31	31	–	24	26	31
		–	–	–	–	**M**	–	**M**	F	–	**M**	M	**M**
CS1404	M	–	–	–	–	28	–	–	–	28	–	24	–
		–	–	–	–	F	–	–	–	**M**	–	**M**	–
CS0649	F	–	–	–	–	30	32	29	–	26	–	–	–
		–	–	–	–	**F**	**F**	**F**	–	**F**	–	–	–
CS0080	F	21	24	27	25	35	24	19	28	28	30	–	–
		F	**F**	**F**	**F**	**F**	**F**	**F**	**F**	M	**F**	–	–
CS0011	F	–	–	–	–	29	25	33	28	–	29	25	–
		–	–	–	–	**F**	**F**	**F**	**F**	–	**F**	**F**	–

Table 6: Gender Classification Performance on the ITWCC Dataset

Method	**Accuracy**	**Precision**	**Recall**
ChatGPT-4o	1.00	1.00	1.00
DeepFace	0.94	0.95	0.93

5.4 Identity Continuity on ITWCC

To find whether ChatGPT-4o can identify the same individual across years, we tested the ITWCC dataset. The model achieved a 92% true-acceptance rate. ORB keypoint matching corroborated these judgments: true pairs exhibited an average of 105 matches with a mean Hamming distance of 36, while false pairs had only 58 matches and an average distance of 52. As compared in Table 7, DeepFace's threshold, tuned for adults, yielded many false rejections and false acceptances; the LLM's reasoning therefore provides more robust identity verification for children.

For each subject, identity continuity was assessed by comparing the enrollment image to all acquisition images. ChatGPT correctly identified identity continuity for 92 % of pairs, whereas DeepFace achieved only 68 %. Patterns emerged when grouping subjects by age gap:

- *Small age gaps ($<$ 3 years):* Both models succeeded consistently. ORB matching typically found more than 120 keypoint correspondences with an average distance below 35.
- *Moderate age gaps (3–6 years):* DeepFace often failed when puberty–induced changes dramatically altered facial proportions. ChatGPT still recognized the same person by reasoning over eye spacing, nose tip, and ear shape. ORB match counts remained high (90–110), supporting these conclusions.
- *Large age gaps ($>$ 6 years):* Both models struggled. In cases like CS0080 (enrollment at 6 years, acquisition at 16), ChatGPT incorrectly judged the images of different people in 30 % of trials. ORB matches dropped below 60 with average distances above 50, confirming the difficulty. DeepFace misclassified nearly all such pairs due to drastic jaw length and hairstyle changes.

Table 7: Comparison between ChatGPT-4o and DeepFace on age pair verification tasks. The table reports estimated similarity scores and conclusions drawn by both systems.

Age Pair	ChatGPT-4o Similarity	ChatGPT-4o Conclusion	DeepFace Verdict	Cosine Distance	DeepFace Conclusion
6 / 7	75%	Likely same identity (AI variants)	True Match	0.4187	Same person
6 / 8	85–90%	Likely same identity	True Match	0.4557	Same person
6 / 9	90–95%	Same child or identical twin	True Match	0.3308	Same person
6 / 10	90–95%	Minimal changes, same identity	False Match	0.9010	Not same
6 / 11	75–80%	Same identity with age progression	True Match	0.4692	Same person
6 / 12	80–85%	Age-progressed same individual	True Match	0.5290	Same person
6 / 13	78–83%	Same identity, moderate maturity	True Match	0.5289	Same person
6 / 14	85–90%	High resemblance, likely same	False Match	0.8488	Not same
6 / 15	87–92%	Same individual, adolescent stage	False Match	0.7149	Not same

5.5 Identity continuity after surgery

The aim was to observe if ChatGPT-4o can recognize individuals before and after surgical procedures, despite the drastic geometric changes. We summarized the performance of the 15 surgery pairs in Table 8. ChatGPT judged 11 pairs as the same individual and 4 as different. Manual inspection confirmed that 10 pairs indeed belonged to the same patient; therefore, ChatGPT-4o correctly matched 87% of pre-post pairs. DeepFace, by contrast, classified only 6 pairs correctly (40 % accuracy) because its embeddings are sensitive to geometric distortions. ORB feature matching showed an average of 105 matches (std. 20) with a mean Hamming distance of 36 for true pairs; mismatched pairs exhibited only 58 matches with a mean distance of 52. These numbers corroborated ChatGPT's decisions.

Table 8: Aggregated performance on the *Surgery* dataset. ChatGPT-4o's predictions aligned closely with ORB statistics on most pairs, whereas DeepFace struggled due to significant geometric variations.

Method	Accuracy	Average Matches	Avg. Distance
ChatGPT-4o DeepFace	0.87 0.40	105 (true)	36 (true)
ORB statistics (false pairs)	–	58	52

5.6 Comparison with Grok and Claude 3.5 Haiku LLMs

We compared ChatGPT-4o with Grok AI xAI (2024) and Claude 3.5 Haiku Anthropic (2024) on a subset of surgery pairs. Grok generally concurred with ChatGPT's conclusions, while Claude often interpreted the images as separate AI-generated variations rather than different views of the same individual. Across ten pairs, ChatGPT judged eight pairs to be the same individual, Grok judged seven, and Claude only two.

5.7 Qualitative Observations

ChatGPT's explanations cite consistent features such as eye spacing, nose structure, and ear shape while acknowledging changes in hairstyle, dental alignment, and facial maturity. The model sometimes misjudges pairs with extreme surgical changes or large age gaps. Prompt wording matters; including the phrase "created by AI model" improves accuracy. These observations suggest that MLLMs reason at a higher level than embedding-based models but remain sensitive to instruction design.

The next section discusses explainability, fairness, and ethical implications of these findings.

6 Analysis and Discussion

6.1 Explainability and Hybrid Reasoning

We observed strong alignment between ChatGPT's qualitative reasoning and ORB's quantitative evidence: when ChatGPT judged two images as the same person, ORB typically showed many matching keypoints and low Hamming distances. This suggests that the model implicitly relies on geometric cues even though it does not compute explicit embeddings. However, ChatGPT occasionally offers plausible but incorrect explanations, which motivates a hybrid pipeline that uses LLM reasoning for candidate matches, ORB as a fast filter, and a CNN verifier.

6.2 Fairness and Demographic Analysis

Our datasets are small and skewed toward white child actors, yet a preliminary fairness analysis is possible. Stratifying by age band (Table 4) reveals that ChatGPT-4o performs best on younger children and slightly worse on older adolescents. Performance differences between male and female subjects in ITWCC were negligible—both genders were classified correctly, though this may reflect the limited sample size. We encourage future work to audit the model across ethnicity, skin tone, and socioeconomic status to ensure equitable performance.

6.3 Privacy and Consent

Analyzing real children's faces necessitates strict privacy safeguards. Our prompts circumvent ChatGPT's safety filters for research, but real-world deployment should require explicit consent and anonymity. We used publicly available datasets; nonetheless, the potential misuse of such techniques underscores the importance of robust data governance and ethical oversight.

7 Conclusions and Future Work

This study demonstrates that ChatGPT-4o, when guided by carefully engineered prompts, delivers competitive and often superior performance on age estimation, gender classification, and identity continuity compared to DeepFace. ChatGPT-4o outperforms a CNN baseline across these tasks and provides interpretable explanations aligned with geometric evidence. However, performance declines for extreme transformations and fairness across demographics remain unverified. Future research should: (1) focus on domain-specific fine-tuning of LLMs using pediatric and surgical face corpora to enhance consistency and reduce prompt sensitivity; (2) extend the prompting framework to handle adversarial morphs and blended facial cues; (3) systematically conduct fairness audits to evaluate performance across ethnicity, lighting, pose, and expression, ensuring demographic equity; (4) develop real-time pipelines that integrate LLM prompts, ORB checks, and CNN verification to support practical deployment; and (5) design explainability interfaces that present LLM rationales with ORB overlays to improve transparency for users.

8 Ethics Statement

The experiments in this paper involve analyzing facial images of children and surgical patients. We obtained all data from publicly available sources (ITWCC) or licensed research datasets (surgery) and followed the usage policies associated with each dataset. We emphasize that no personally identifying information beyond the images was used, and we did not attempt to deanonymise subjects.

9 Limitations

Although our results are encouraging, we acknowledge several areas for enhancement. Our surgery dataset results are based on a small sample of pediatric cases. Larger and more diverse cohorts are needed. We relied on manually crafted prompts, which may limit generalisability across other LLMs or future model versions. Future work will address these by assembling larger, more diverse datasets and conducting comprehensive significance and fairness analyses to reinforce and broaden our findings.

References

Anthropic. 2024. Claude 3.5 model family. https://www.anthropic.com/news/claude-3-5-haiku. Accessed: July 21, 2025.

Praveen Kumar Chandaliya and Neeta Nain. 2018. Plasticgan: Learning face representation via gans to detect morphing attacks based on facial plasticity. In *2018 International Conference on Biometrics (ICB)*, pages 1–8. IEEE.

Praveen Kumar Chandaliya and Neeta Nain. 2022. Childgan: Face aging and rejuvenation to find missing children. In *Pattern Recognition*, volume 129, page 108761.

Praveen Kumar Chandaliya, Kiran Raja, Raghavendra Ramachandra, Zahid Akhtar, and Christoph Busch. 2024. Towards inclusive face recognition through synthetic ethnicity alteration. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Fortune Business Insights. 2023. Facial recognition market report. https://www.fortunebusinessinsights.com/industry-reports/facial-recognition-market-101061.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. Face recognition vendor test (frvt) part 3: Demographic effects. Technical Report NIST Interagency Report 8280, National Institute of Standards and Technology.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102. Springer.

Ahmad Hassanpour, Yasamin Kowsari, Hatef Otroshi Shahreza, Bian Yang, and Sébastien Marcel. 2024. Chatgpt and biometrics: an assessment of face recognition, gender detection, and age estimation capabilities. In *2024 IEEE International Conference on Image Processing (ICIP)*, page 3224–3229. IEEE.

Kartik Narayan, Vibashan V. S., and Vishal M. Patel. 2025. Facexbench: Evaluating multimodal llms on face understanding. *arXiv preprint arXiv:2501.10360*.

OpenAI. 2024. Chatgpt-4o. `https://openai.com/chatgpt`. Accessed July 21, 2025.

Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *British Machine Vision Conference (BMVC)*.

Nisha Srinivas, Karl Ricanek, Dana Michalski, David S. Bolme, and Michael King. 2019. Face recognition algorithm bias: Performance differences on images of children and adults. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2269–2277. IEEE/CVF.

xAI. 2024. Introducing grok by xai. `https://x.ai/blog/grok`. Accessed: July 21, 2025.

Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. 2024. Racial bias within face recognition: A survey. *ACM Comput. Surv.*, 57(4).