Comparison of GPT-4o and MedGemma on IDRiD

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Model** | **Variant** | **Accuracy** | **F1 Score** | **Sensitivity** | **Specificity** |
| **IDRiD** | GPT-4O | Zero-Shot | 0.71 | 0.12 | 0.06 | 1.0 |
| **IDRiD** | GPT-4O | + CNN (80%) | 0.79 | 0.48 | 0.31 | 1.0 |
| **IDRiD** | GPT-4O | + CNN (95%) | 0.82 | 0.58 | 0.41 | 1.0 |
| **IDRiD** | GPT-4O | + CNN (99%) | 0.8 | 0.51 | 0.34 | 1.0 |
| **IDRiD** | GPT-4O | + Clinical Context | 0.73 | 0.22 | 0.12 | 1.0 |
| **IDRiD** | GPT-4O | Continuous (Threshold 20) | 0.67 | 0.51 | 0.56 | 0.72 |
| **IDRiD** | MedGemma | Zero-Shot | 0.78 | 0.68 | 0.75 | 0.79 |
| **IDRiD** | MedGemma | + CNN (80%) | 0.81 | 0.74 | 0.88 | 0.77 |
| IDRiD | MedGemma | + CNN (95%) | 0.82 | 0.75 | 0.88 | 0.79 |
| **IDRiD** | MedGemma | + CNN (99%) | 0.79 | 0.72 | 0.88 | 0.75 |
| **IDRiD** | MedGemma | + Clinical Context | 0.82 | 0.72 | 0.75 | 0.85 |
| **IDRiD** | MedGemma | Continuous (Threshold 85) | 0.76 | 0.68 | 0.84 | 0.72 |

Sensitivity: The proportion of actual referral cases (true positives) that the model correctly identifies. **High sensitivity** means fewer missed referrals.

**Specificity:** The proportion of non-referral cases (true negatives) that the model correctly identifies. **High specificity** helps reduce unnecessary follow-ups or resource use.

#### ****F1 Score:**** The harmonic mean of precision and sensitivity. It balances false positives and false negatives.

\*\*\*\*\*\*\*

* **MedGemma consistently outperformed GPT-4O** across most metrics, especially F1 Score and Sensitivity.
* **GPT-4O achieved perfect Specificity (1.0)** in most configurations, meaning it rarely over-predicted referrals — but at the cost of **very low Sensitivity**, meaning it often missed true referrals.
* **MedGemma had a better balance**, maintaining high Sensitivity (up to 0.88) while preserving reasonable Specificity (0.75–0.85).

**🧠 Model Insights**

**GPT-4O**

* **Zero-shot:** Accuracy 0.71 but extremely low Sensitivity (0.06) and F1 (0.12); very conservative referral behavior.
* **Adding CNN (80–99%)** improved F1 and Sensitivity slightly, with best results at 95% threshold (F1: 0.58, Sens: 0.41).
* **Continuous Output** boosted Sensitivity to 0.56 and F1 to 0.51, but dropped Specificity to 0.72 — a better tradeoff.
* **Adding Clinical Context** did not meaningfully improve performance (F1: 0.22, Sens: 0.12).

**MedGemma**

* **Zero-shot** started strong (F1: 0.68, Sens: 0.75), outperforming GPT-4O baseline.
* **CNN thresholds (80–95%)** further boosted performance:
  + Best config: **+CNN (95%)**, with F1: 0.75 and Sensitivity: 0.88.
* **Clinical context** also helped here (Sens: 0.75, Specificity: 0.85).
* **Continuous Output** maintained solid balance (F1: 0.68, Sens: 0.84).

**✅ Key Takeaways**

* **MedGemma is more balanced and reliable** for referral classification, particularly with CNN integration.
* **GPT-4O needs calibration or contextual tuning** (e.g., continuous scoring) to achieve usable recall.
* **CNN confidence thresholds help both models**, but more significantly for GPT-4O.
* **Clinical context** benefits MedGemma, but not GPT-4O in this setup.

GPT-4o and MedGemma on Messidor Dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Model** | **Variant** | **Accuracy** | **F1 Score** | **Sensitivity** | **Specificity** |
| **Messidor** | GPT-4O | Zero-Shot | 0.61 | 0.06 | 0.03 | 1.0 |
| **Messidor** | GPT-4O | + Clinical Context | 0.6 | 0.16 | 0.09 | 0.98 |
| **Messidor** | GPT-4O | Continuous (Threshold 20) | 0.85 | 0.79 | 0.74 | 0.91 |
| **Messidor** | MedGemma | Zero-Shot | 0.91 | 0.88 | 0.87 | 0.93 |
| **Messidor** | MedGemma | + Clinical Context | 0.9 | 0.87 | 0.84 | 0.94 |
| **Messidor** | MedGemma | Continuous (Threshold 85) | 0.93 | 0.9 | 0.86 | 0.97 |