

## **Supplemental Tables**

| <b>Task</b>    | <b>Gender</b>     | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b>  | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|-------------------|---------------------|---------------------|---------------------|----------------------------------|----------------------------------|
| Abdominal pain | Male<br>(n=104)   | 0.93<br>[0.89-0.97] | 0.96<br>[0.93-0.99] | 0.92<br>[0.88-0.96] | 0.85<br>[0.79-0.91]              | 0.98<br>[0.96-1]                 |
|                | Female<br>(n=146) | 0.93<br>[0.88-0.98] | 0.96<br>[0.92-1]    | 0.92<br>[0.87-0.97] | 0.82<br>[0.75-0.89]              | 0.98<br>[0.96-1]                 |
| Diarrhea       | Male<br>(n=104)   | 0.88<br>[0.83-0.93] | 0.91<br>[0.86-0.96] | 0.86<br>[0.80-0.92] | 0.84<br>[0.78-0.90]              | 0.92<br>[0.88-0.96]              |
|                | Female<br>(n=146) | 0.88<br>[0.82-0.94] | 0.87<br>[0.81-0.93] | 0.90<br>[0.84-0.96] | 0.87<br>[0.81-0.93]              | 0.90<br>[0.84-0.96]              |
| Fecal blood    | Male<br>(n=104)   | 0.93<br>[0.89-0.97] | 0.88<br>[0.83-0.93] | 0.95<br>[0.91-0.99] | 0.88<br>[0.83-0.93]              | 0.95<br>[0.91-0.99]              |
|                | Female<br>(n=146) | 0.97<br>[0.94-1]    | 1.00                | 0.96<br>[0.92-1]    | 0.90<br>[0.84-0.96]              | 1.00                             |

*Supplemental Table 1: Bias assessment of PaLM-2 based on gender, conducted on the test set of 250 notes at*

*Stanford University. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Model</b>   | <b>Disease</b> | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b>  | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|----------------|---------------------|---------------------|---------------------|----------------------------------|----------------------------------|
| Abdominal pain | CD (n=138)     | 0.93<br>[0.89-0.97] | 0.96<br>[0.93-0.99] | 0.91<br>[0.86-0.96] | 0.85<br>[0.79-0.91]              | 0.98<br>[0.96-1]                 |
|                | UC (n=112)     | 0.94<br>[0.90-0.98] | 0.97<br>[0.94-1]    | 0.93<br>[0.88-0.98] | 0.82<br>[0.75-0.89]              | 0.99<br>[0.97-1]                 |
| Diarrhea       | CD (n=138)     | 0.91<br>[0.86-0.96] | 0.95<br>[0.91-0.99] | 0.88<br>[0.83-0.93] | 0.87<br>[0.81-0.93]              | 0.96<br>[0.93-0.99]              |
|                | UC (n=112)     | 0.85<br>[0.78-0.92] | 0.81<br>[0.74-0.88] | 0.88<br>[0.82-0.94] | 0.83<br>[0.76-0.90]              | 0.86<br>[0.80-0.92]              |
| Fecal blood    | CD (n=138)     | 0.94<br>[0.90-0.98] | 0.88<br>[0.83-0.93] | 0.96<br>[0.93-0.99] | 0.88<br>[0.83-0.93]              | 0.96<br>[0.93-0.99]              |
|                | UC (n=112)     | 0.96<br>[0.92-1]    | 0.97<br>[0.94-1]    | 0.95<br>[0.91-0.99] | 0.89<br>[0.83-0.95]              | 0.99<br>[0.97-1]                 |

*Supplemental Table 2: Bias assessment of PaLM-2 based on diagnosis—Crohn’s Disease (CD) versus Ulcerative colitis (UC). This was conducted on the test set of 250 notes at Stanford University. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Task</b>    | <b>Race</b>         | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b>  | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|---------------------|---------------------|---------------------|---------------------|----------------------------------|----------------------------------|
| Abdominal pain | White<br>(n=150)    | 0.92<br>[0.88-0.96] | 0.95<br>[0.92-0.98] | 0.91<br>[0.86-0.96] | 0.78<br>[0.71-0.85]              | 0.98<br>[0.96-1]                 |
|                | Non-white<br>(n=98) | 0.95<br>[0.91-0.99] | 0.97<br>[0.94-1]    | 0.93<br>[0.88-0.98] | 0.90<br>[0.84-0.96]              | 0.98<br>[0.95-1]                 |
| Diarrhea       | White<br>(n=150)    | 0.91<br>[0.86-0.96] | 0.92<br>[0.88-0.96] | 0.90<br>[0.85-0.95] | 0.86<br>[0.80-0.92]              | 0.94<br>[0.90-0.98]              |
|                | Non-white<br>(n=98) | 0.85<br>[0.78-0.92] | 0.85<br>[0.78-0.92] | 0.84<br>[0.77-0.91] | 0.84<br>[0.77-0.91]              | 0.86<br>[0.79-0.93]              |
| Fecal blood    | White<br>(n=150)    | 0.96<br>[0.93-0.99] | 0.94<br>[0.90-0.98] | 0.97<br>[0.94-1]    | 0.89<br>[0.84-0.94]              | 0.98<br>[0.96-1]                 |
|                | Non-white<br>(n=98) | 0.94<br>[0.89-0.99] | 0.91<br>[0.85-0.97] | 0.95<br>[0.91-0.99] | 0.91<br>[0.85-0.97]              | 0.95<br>[0.91-0.99]              |

*Supplemental Table 3: Bias assessment of PaLM-2 based on race conducted on the test set of notes at Stanford University. In our cohort, the category of Non-White race included Black (n=6), Pacific Islander (n=2), Native American (n=2), Asian (n=22), and Other (n=65). Two clinical encounters had an unknown race and were excluded. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Task</b>    | <b>Age Group</b>     | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b>  | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|----------------------|---------------------|---------------------|---------------------|----------------------------------|----------------------------------|
| Abdominal pain | Below median (n=125) | 0.92<br>[0.87-0.97] | 0.94<br>[0.90-0.98] | 0.91<br>[0.86-0.96] | 0.81<br>[0.74-0.88]              | 0.98<br>[0.96-1.00]              |
|                | Above median (n=125) | 0.93<br>[0.88-0.98] | 0.96<br>[0.92-1]    | 0.92<br>[0.87-0.97] | 0.82<br>[0.75-0.89]              | 0.98<br>[0.96-1.00]              |
| Diarrhea       | Below median (n=125) | 0.88<br>[0.83-0.93] | 0.91<br>[0.86-0.96] | 0.86<br>[0.80-0.92] | 0.84<br>[0.78-0.90]              | 0.92<br>[0.88-0.96]              |
|                | Above median (n=125) | 0.88<br>[0.82-0.94] | 0.87<br>[0.81-0.93] | 0.90<br>[0.84-0.96] | 0.87<br>[0.81-0.93]              | 0.90<br>[0.84-0.96]              |
| Fecal blood    | Below median (n=125) | 0.93<br>[0.89-0.97] | 0.88<br>[0.83-0.93] | 0.95<br>[0.91-0.99] | 0.88<br>[0.83-0.93]              | 0.95<br>[0.91-0.99]              |
|                | Above median (n=125) | 0.97<br>[0.94-1]    | 1.00                | 0.96<br>[0.92-1]    | 0.90<br>[0.84-0.96]              | 1.00                             |

*Supplemental Table 4: Bias assessment of PaLM-2 based on age, conducted on the test set of 250 notes at Stanford University. We divided our cohort into above and below the median age of 35 years old. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Task</b>    | <b>Gender</b>     | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b>  | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|-------------------|---------------------|---------------------|---------------------|----------------------------------|----------------------------------|
| Abdominal pain | Male<br>(n=146)   | 0.96<br>[0.93-0.99] | 0.90<br>[0.85-0.95] | 0.99<br>[0.97-1]    | 0.98<br>[0.96-1]                 | 0.95<br>[0.91-0.99]              |
|                | Female<br>(n=104) | 0.93<br>[0.88-0.98] | 0.86<br>[0.79-0.93] | 0.96<br>[0.92-1]    | 0.89<br>[0.83-0.95]              | 0.95<br>[0.91-0.99]              |
| Diarrhea       | Male<br>(n=146)   | 0.89<br>[0.84-0.94] | 0.80<br>[0.74-0.86] | 0.96<br>[0.93-0.99] | 0.95<br>[0.91-0.99]              | 0.86<br>[0.80-0.92]              |
|                | Female<br>(n=104) | 0.91<br>[0.85-0.97] | 0.80<br>[0.71-0.88] | 1                   | 1                                | 0.87<br>[0.81-0.93]              |
| Fecal blood    | Male<br>(n=146)   | 0.95<br>[0.91-0.99] | 0.88<br>[0.83-0.93] | 0.98<br>[0.96-1]    | 0.95 [0.91-0.99]                 | 0.95<br>[0.91-0.99]              |
|                | Female<br>(n=104) | 0.99<br>[0.97-1]    | 0.96<br>[0.92-1]    | 1                   | 1                                | 0.99<br>[0.97-1]                 |

*Supplemental Table 5: Bias assessment of GPT-4 based on gender, conducted on the test set of 250 notes at Stanford University. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Model</b>   | <b>Disease</b> | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b> | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|----------------|---------------------|---------------------|--------------------|----------------------------------|----------------------------------|
| Abdominal pain | CD<br>(n=138)  | 0.96<br>[0.93-0.99] | 0.89<br>[0.84-0.94] | 0.99<br>[0.97-1]   | 0.98<br>[0.96-1]                 | 0.95<br>[0.91-0.99]              |
|                | UC<br>(n=112)  | 0.94<br>[0.90-0.98] | 0.86<br>[0.80-0.92] | 0.96<br>[0.92-1]   | 0.89<br>[0.83-0.95]              | 0.95<br>[0.91-0.99]              |
| Diarrhea       | CD<br>(n=138)  | 0.88<br>[0.83-0.93] | 0.77<br>[0.70-0.84] | 0.97<br>[0.94-1]   | 0.96<br>[0.93-0.99]              | 0.84<br>[0.78-0.90]              |
|                | UC<br>(n=112)  | 0.92<br>[0.87-0.97] | 0.83<br>[0.76-0.90] | 0.98<br>[0.95-1]   | 0.98<br>[0.95-1]                 | 0.89<br>[0.83-0.95]              |
| Fecal blood    | CD<br>(n=138)  | 0.98<br>[0.96-1]    | 0.91<br>[0.86-0.96] | 1                  | 1                                | 0.97<br>[0.94-1]                 |
|                | UC<br>(n=112)  | 0.96<br>[0.92-1]    | 0.91<br>[0.86-0.96] | 0.98<br>[0.95-1]   | 0.94<br>[0.90-0.98]              | 0.96<br>[0.92-1]                 |

*Supplemental Table 6: Bias assessment of GPT-4 based on diagnosis—Crohn's Disease (CD) versus Ulcerative colitis (UC). This was conducted on the test set of 250 notes at Stanford University. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Task</b>    | <b>Race</b>         | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b> | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|---------------------|---------------------|---------------------|--------------------|----------------------------------|----------------------------------|
| Abdominal pain | White<br>(n=150)    | 0.95<br>[0.92-0.98] | 0.87<br>[0.82-0.92] | 0.97<br>[0.94-1]   | 0.92<br>[0.88-0.96]              | 0.96<br>[0.93-0.99]              |
|                | Non-white<br>(n=98) | 0.96<br>[0.92-1]    | 0.92<br>[0.87-0.97] | 0.98<br>[0.95-1]   | 0.97<br>[0.94-1]                 | 0.95<br>[0.91-0.99]              |
| Diarrhea       | White<br>(n=150)    | 0.91<br>[0.86-0.96] | 0.78<br>[0.71-0.85] | 0.98<br>[0.96-1]   | 0.98<br>[0.96-1]                 | 0.87<br>[0.82-0.92]              |
|                | Non-white<br>(n=98) | 0.89<br>[0.83-0.95] | 0.81<br>[0.73-0.89] | 0.96<br>[0.92-1]   | 0.95<br>[0.91-0.99]              | 0.84<br>[0.77-0.91]              |
| Fecal blood    | White<br>(n=150)    | 0.97<br>[0.94-1]    | 0.91<br>[0.86-0.96] | 0.99<br>[0.97-1]   | 0.97<br>[0.94-1]                 | 0.97<br>[0.94-1]                 |
|                | Non-white<br>(n=98) | 0.97<br>[0.94-1]    | 0.90<br>[0.84-0.96] | 1                  | 1                                | 0.96<br>[0.92-1]                 |

*Supplemental Table 7: Bias assessment of GPT-4 based on race conducted on the test set of notes at Stanford University. In our cohort, the category of Non-White race included Black (n=6), Pacific Islander (n=2), Native American (n=2), Asian (n=22), and Other (n=65). Two clinical encounters had an unknown race and were excluded. Metrics are reported as percentages with their associated 95% confidence intervals.*

| <b>Model</b>   | <b>Age Group</b>     | <b>Accuracy</b>     | <b>Sensitivity</b>  | <b>Specificity</b>  | <b>Positive Predictive Value</b> | <b>Negative Predictive Value</b> |
|----------------|----------------------|---------------------|---------------------|---------------------|----------------------------------|----------------------------------|
| Abdominal pain | Below median (n=125) | 0.94<br>[0.90-0.98] | 0.90<br>[0.85-0.95] | 0.96<br>[0.93-0.99] | 0.92<br>[0.88-0.97]              | 0.95<br>[0.92-0.99]              |
|                | Above median (n=125) | 0.94<br>[0.90-0.98] | 0.86<br>[0.80-0.92] | 0.98<br>[0.96-1.00] | 0.97<br>[0.94-1.00]              | 0.93<br>[0.88-0.97]              |
| Diarrhea       | Below median (n=125) | 0.90<br>[0.85-0.95] | 0.83<br>[0.76-0.90] | 0.97<br>[0.94-1.00] | 0.96<br>[0.93-0.99]              | 0.87<br>[0.81-0.93]              |
|                | Above median (n=125) | 0.90<br>[0.85-0.95] | 0.77<br>[0.70-0.84] | 0.98<br>[0.96-1.00] | 0.98<br>[0.96-1.00]              | 0.86<br>[0.80-0.92]              |
| Fecal blood    | Below median (n=125) | 0.97<br>[0.94-1.00] | 0.89<br>[0.84-0.94] | 1.00                | 1.00                             | 0.96<br>[0.93-0.99]              |
|                | Above median (n=125) | 0.97<br>[0.94-1.00] | 0.93<br>[0.88-0.97] | 0.98<br>[0.96-1.00] | 0.93<br>[0.88-0.97]              | 0.98<br>[0.96-1.00]              |

*Supplemental Table 8: Bias assessment of GPT-4 based on age, conducted on the test set of 250 notes at Stanford University. We divided our cohort into above and below the median age of 35 years old. Metrics are reported as percentages with their associated 95% confidence intervals.*