Appendix A: Complete Results for A/B Test

Additional information about intent distribution is shown in Table 5. Further, Table 4 shows the complete results about the real-world A/B test that we conducted in collaboration with Jacaranda Health. Only the intents which have been predicted with a precision score greater than 0.75 appear in this table. The column (*Answer "Yes"*) indicates the percentage of SMS queries for which the sender of that SMS query (i.e., a pregnant mother) was satisfied by the information provided through the AI model. Similarly, the column (with *Answer "No"*) presents the percentage of SMS queries for which the sender of that SMS query was not satisfied with the information provided through the AI model. The column called "*Total Questions*" corresponds to the percentage of SMS queries (which belong to each intent) that were assigned to either TRIM-AI or Vertex AI model.

Finally, the row (named *Grand Total*) demonstrates that 51.8% of SMS queries have been assigned to TRIM-AI model while the remaining 48.2% queries have been assigned to Vertex AI model. For all queries sent to the TRIM-AI model, users think 79.51% of them have been answered in a satisfactory manner. For all queries sent to the Vertex-AI model, users think 78.89% of them have been answered in an unsatisfactory manner.

Appendix B: Further Statistical Analysis for A/B Test

Further, we report on the statistical significance of the results obtained in our A/B test. Our overall dataset contains 5323 messages in total. As mentioned in the paper, this occurs when one intent among the predicted intent distribution crosses the 75% confidence threshold.

Across all intents, the performance improvement of TRIM-AI (over Vertex AI) is not statistically significant (we conducted a two-proportion Z-test and got a p-value greater than 0.05). However, of all the intents evaluated in the A/B test, TRIM-AI achieves statistically significant benefits over the Vertex AI model in two intents (i.e., *baby_milestone_general* and *baby_jaundice*). Importantly, *baby_jaundice* is a highly critical intent. The lack of statistical significance across all intents is because of the small scale of our pilot study (which was only run over a two week period).

Intent Name	**TRIM-AI**			**Vertex AI**		
	Answer ”No”	Answer ”Yes”	Total Question	Answer ”No”	Answer ”Yes”	Total Question
baby_constipation	19.51%	80.49%	46.59%	19.15%	80.85%	53.41%
baby_general	30.49%	69.51%	59.85%	27.27%	72.73%	40.15%
baby_hiccups	0.00%	100.00%	73.68%	0.00%	100.00%	26.32%
baby_jaundice	0.00%	100.00%	44.44%	40.00%	60.00%	55.56%
baby_milestone_general	10.26%	89.74%	52.70%	28.57%	71.43%	47.30%
baby_milestone_teething	100.00%	0.00%	21.05%	20.00%	80.00%	78.95%
breastfeeding	29.00%	71.00%	60.61%	24.62%	75.38%	39.39%
edd	25.37%	74.63%	46.53%	36.36%	63.64%	53.47%
family_planning	21.10%	78.90%	50.46%	28.97%	71.03%	49.54%
fatigue	60.00%	40.00%	26.32%	28.57%	71.43%	73.68%
fetal_movement	20.93%	79.07%	58.90%	11.67%	88.33%	41.10%
linda_mama	21.05%	78.95%	40.43%	17.86%	82.14%	59.57%
medication_general	22.58%	77.42%	53.45%	7.41%	92.59%	46.55%
ok_thanks	15.35%	84.65%	62.93%	15.97%	84.03%	37.07%
pain_stomach	22.73%	77.27%	53.99%	25.33%	74.67%	46.01%
pregnancy_general	36.06%	63.94%	55.37%	24.06%	75.94%	44.63%
survey_response	17.97%	82.03%	49.27%	20.04%	79.96%	50.73%
ultrasound	15.38%	84.62%	29.21%	23.81%	76.19%	70.79%
urination_uti	17.39%	82.61%	68.66%	23.81%	76.19%	31.34%
Grand total	20.49%	79.51%	51.80%	21.11%	78.89%	48.20%

Table 4: A/B tests results based on intents whose precision scores are greater than 0.75 after AI prediction.

Intent type	**Counts**	**Percentage**
survey_response	2823	53.03%
pregnancy_general	596	11.20%
ok_thanks	321	6.03%
family_planning	216	4.06%
breastfeeding	165	3.10%
pain_stomach	163	3.06%
fetal_movement	146	2.74%
edd	144	2.71%
baby_general	137	2.57%
urination_uti	134	2.52%
linda_mama	94	1.77%
ultrasound	89	1.67%
baby_constipation	88	1.65%
baby_milestone_general	74	1.39%
medication_general	58	1.09%
baby_hiccups	19	0.36%
fatigue	19	0.36%
baby_milestone_teething	19	0.36%
baby_jaundice	18	0.33%
Total	5323	100.00%

Table 5: Intent distribution over all test samples in A/B test.