Model Evaluation Report: Multilingual Chatbot Model v2.0

1. Objective

Evaluate the performance of the chatbot model (v2.0) against baseline model (v1.3) using multilingual benchmark datasets.

2. Evaluation Datasets

Dataset	Language	Size	Purpose
Customer Support Eval Set	English	5,000 pairs	Customer intent recognition
Kundenhilfe Eval Set	German	2,500 pairs	Multilingual accuracy
General QA Eval	Mixed	3,000 pairs	Conversational fluency

3. Evaluation Metrics

Metric	Definition	Target	Achieved
Accuracy	Correct responses / total responses	90%	91.5%
Precision	True positives / (true + false positives)	85%	88%
Recall	True positives / (true + false negatives)	80%	84%
F1 Score	Harmonic mean of precision and recall	82%	86%

4. Results Summary

The multilingual chatbot model v2.0 shows consistent improvement in both English and German datasets, with a +4% increase in accuracy and +6% in F1 Score compared to v1.3.

5. Interpretation

- Improved **intent recognition** for technical queries
- Slightly lower performance in **casual conversation tone**
- Model generalises well across languages, but minor German grammar errors remain

6. Visual Summary

Accuracy trend over model versions:

v1.0 → 82% v1.3 → 87% v2.0 → 91.5%

7. Next Steps

- Fine-tune for German slang and colloquial expressions

- Add additional low-resource language data (Hungarian, Ukrainian)

8. Last Updated

Date: 7 November 2025
Author: Leslie Amadi
Version: 2.0