Dataset Datasheet: Customer Support Chatbot Corpus

1. Motivation

This dataset was created to train a multilingual customer support chatbot capable of answering common product questions in English and German.

2. Composition

- **Data Type:** Text conversations (user queries + agent responses)
- **Languages:** English (70%), German (30%)
- **Size:** 120,000 dialogue pairs
- **Domains:** E-commerce, technical support, FAQs

3. Collection Process

The dataset was compiled from anonymised chat logs and manually cleaned by human annotators.
Sensitive information (names, emails, phone numbers) was removed automatically.

4. Recommended Uses

- Training intent recognition models
- Fine-tuning LLMs for conversational tone
- Evaluating multilingual response quality

5. Not Recommended For

- Sentiment analysis (data is neutral)
- Legal or medical chatbots (not domain-specific)

6. Ethical Considerations

Bias may exist toward polite, service-oriented tone due to training sources.
All data was anonymised following GDPR compliance standards.

7. Maintenance and Updates

- Version: 1.2
- Last Updated: November 2025
- Maintained by: Leslie Amadi
- Contact: dr.leslieamadi@lesliewrites.tech

This datasheet follows the guidelines from "Datasheets for Datasets" (Gebru et al., 2018).