International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

# Assistify: Customer Service Chatbot

**Utkarsh Singh**
R.I.T

**Abstract — Despite the widespread adoption of chatbots for customer service, small businesses continue to lack the resources required to develop and customize such solutions effectively. In this work, we propose an approach that enables small businesses to construct tailored customer service chatbots by fine-tuning an open-weight large language model (LLM) using QLoRA [1] on a domain-specific customer service dataset. The proposed method leverages parameter-efficient transfer learning to adapt a pre-trained LLM to the customer service domain, preserving its general natural language understanding and generation capabilities while specializing its behaviour. To extend the chatbot's functional scope beyond the fine-tuning corpus, we integrate a Retrieval-Augmented Generation (RAG) [4] module that retrieves relevant product information, warranty details, and installation guides from PDF documentation at inference time. This allows the chatbot to ground its responses in accurate, business-specific information rather than relying solely on parametric knowledge. In addition, we log chatbot conversations and apply sentiment analysis to the resulting interaction data, forming a feedback loop that surfaces unsatisfactory exchanges for human follow-up and informs iterative improvement of the system. Taken together, these components constitute a cost-effective pipeline through which small businesses can deploy personalized customer service chatbots without incurring the cost of large-scale data collection or bespoke system engineering.**

**Keywords— Help, Start, Menu, Support, Contact, Order Status, Track Order**

## I. INTRODUCTION

Chatbots have become a standard feature of customer service infrastructure, promising streamlined interactions and improved operational efficiency. For many small businesses, however, this promise remains largely unrealized: despite recognizing the potential benefits of deploying such systems, resource constraints and technical complexity continue to place capable chatbot technology out of reach. This leaves small businesses at a structural disadvantage relative to larger competitors that can afford custom-built or heavily licensed solutions. In response, we present a pragmatic, resource-conscious approach to building tailored customer service chatbots, combining parameter-efficient fine-tuning, retrieval augmentation, and a lightweight feedback mechanism into a single pipeline suited to small-business constraints.

Central to our approach is the fine-tuning of an open-weight LLM using QLoRA [1] on a specialized customer service dataset. This fine-tuning step adapts a general-purpose pre-trained model to the linguistic patterns and task structure of customer service dialogue, without requiring the computational budget of full-parameter fine-tuning or the data volume needed to train a model from scratch.

Effective customer service, however, requires more than fluent text generation: it requires access to accurate, business-specific facts that change over time and cannot be reasonably baked into model weights. To address this, we integrate a Retrieval-Augmented Generation (RAG) [4] module into the system. This module retrieves relevant passages from a small corpus of product documentation, warranty information, and installation guides supplied by the business, and conditions the LLM's generation on this retrieved context. This allows the chatbot to produce responses that are both fluent and factually grounded in the business's own documentation, rather than solely dependent on what the base model has memorized.

Finally, to support continuous improvement without additional manual annotation effort, we log chatbot conversations and apply sentiment analysis to flag interactions that likely leŁ the customer dissatisfied, routing these cases to a human representative. Taken together, fine-tuning, retrieval augmentation, and sentiment-based feedback form a compact, cost-effective pipeline that allows small businesses to deploy a customer service chatbot tailored to their own products and customers, without the overhead typically associated with such systems.

## II. LITERATURE SURVEY

Large language models have demonstrated strong general-purpose capabilities, but adapting and deploying them efficiently remains a central challenge, particularly in resource-constrained settings. A substantial body of work has focused on parameter-efficient fine-tuning (PEFT) as a means of

specializing large pre-trained models without the cost of full fine-tuning. Hu et al. [3] introduced LoRA, which injects trainable low-rank decomposition matrices into a frozen backbone, drastically reducing the number of trainable parameters required to adapt a model such as GPT to a down-stream task. Dettmers et al. [1] extended this idea with QLoRA, combining 4-bit quantization of the base model with LoRA adapters, which allows fine-tuning of large LLMs on a single consumer-grade GPU with minimal loss in output quality. Xu et al. [8] provide a broader critical review of PEFT methods for pretrained language models, situating LoRA- and adapter-based approaches within the wider landscape of efficient adaptation techniques. In practice, these methods are most commonly applied through unified training toolkits; Zheng et al. [9] introduce LLaMA-Factory, a widely adopted open-source framework that provides a single interface for parameter-efficient fine-tuning across a large number of open-weight LLMs, and we adopt a toolkit in this same spirit as the fine-tuning backbone of our system. Industry-oriented surveys have also tracked this shiŁ toward efficient fine-tuning in practice, with SuperAnnotate AI Inc. [7] documenting the growing adoption of PEFT methods for LLM customization across production settings.

In parallel, researchers have explored methods for controlling and augmenting LLM outputs with external knowledge, rather than relying purely on parametric memory. Bsharat et al. [6] studied principled instruction design as a means of steering instruction-tuned models toward more truthful and controllable generation without any additional fine-tuning. Lewis et al. [4] introduced Retrieval-Augmented Generation (RAG), which conditions a language model's generation on passages retrieved from an external corpus, substantially improving factual grounding on knowl-edge-intensive tasks. Khattab and Zaharia [5] proposed ColBERT, an efficient late-interaction retrieval architecture that supports fast, high-quality passage retrieval at scale and has become a common retrieval backbone for RAG-style systems.

Limitations of Current Solutions. Despite this progress, existing large language model solu-tions still present significant obstacles for resource-constrained adopters such as small businesses:

- Cost and Accessibility: The largest frontier models remain expensive to query at scale or effectively out of reach for self-hosting, making them a poor fit for small businesses operating under tight budgets.
- Latency: Reliance on large hosted models can introduce response latency inconsistent with the near-real-time expectations of customer service chat.
- Lack of Domain Specificity: General-purpose LLMs, out of the box, typically lack the fine-grained product and policy knowledge required for a specific business, and closing this gap naively would require costly, business-specific data collection.
- Incomplete Solutions: Strong language modeling capability alone does not constitute a deployable customer service system; a complete solution additionally requires retrieval over business documentation, a delivery interface, and a mechanism for measuring and improving quality over time, none of which is provided by a language model in isolation.
- To address these limitations, our approach combines efficient adaptation, retrieval augmentation, and lightweight feedback analysis into a single system:
- We use QLoRA [1], applied through an open-source fine-tuning toolkit [9], to efficiently adapt an open-weight LLM to customer service dialogue on commodity hardware.
- We integrate a RAG module [4] to retrieve relevant product information, warranty details, and installation guides from business documentation, allowing the chatbot to ground its responses in accurate, up-to-date content.
- We log chatbot conversations and apply sentiment analysis to the resulting transcripts, creating a low-overhead feedback loop that flags unsatisfactory interactions for human review.

This combination is intended to give small businesses a practical path to a capable, personalized customer service chatbot, without the data collection and engineering overhead typically associated with building such systems from scratch.
Datasets and Data Preparation

Our fine-tuning corpus is the Bitext customer-support LLM chatbot training dataset, a publicly available collection of conversational exchanges curated by Bitext Innovations for training and evaluating customer service dialogue systems. The dataset consists of paired customer queries and corresponding customer service responses spanning a broad range of intents typical of real-world support interactions, and is provided in a structured format designed to facilitate direct use in supervised fine-tuning pipelines.

Each conversational exchange in the raw dataset consists of a customer query followed by a corresponding support response. As part of data preparation, we reformat every exchange into an instruction-style template matching the chat format of the base model selected for fine-tuning, wrapping the customer query and target response in the model's designated turn-boundary tokens. This preserves the semantic content and pairing of query and response while ensuring compatibility

with the fine-tuning pipeline and the base model's expected prompt structure.

```
text = "<s>[INST] Customer query: How can I track my order? [/INST]"
"Customer service response: You can track your order by logging into your account and
navigating to the 'Order History' section.</s>"

text = "<s>[INST] Customer query: What is your return policy? [/INST]"
"Customer service response: Our return policy allows for returns within 30 days of
purchase, provided the item is in its original condition.</s>"

text = "<s>[INST] Customer query: I'm having trouble setting up my device. [/INST]"
"Customer service response: We apologize for the inconvenience. Please refer to
the installation guide provided with your device, or contact our support team for
assistance.</s>"
```

Figure 1: Instruction-formatted fine-tuning example for the base language model

This structured, format-aligned preparation of the Bitext dataset is intended to give the fine-tuned model a consistent grounding in customer service dialogue patterns, prior to the addition of retrieval-based grounding in business-specific documentation described in the next section.

## III. METHODOLOGY

Our methodology combines three components: parameter-efficient fine-tuning of an open-weight LLM, a retrieval-augmented generation module for business-specific grounding, and a sentiment-based feedback loop. The pipeline begins with fine-tuning Llama 3.1 8B Instruct [2], a compact, widely supported open-weight model with a large fine-tuning ecosystem, using an open-source parameter-efficient fine-tuning toolkit built around the LLaMA-Factory framework [9]. This step adapts the pre-trained model to the customer service domain, so that it can understand and generate contextually appropriate responses to customer queries drawn from the same distribution as the Bitext training data.

Fine-tuning is performed on a 4-bit quantized version of the base model using QLoRA [1]. QLoRA combines quantization of the frozen base weights with trainable low-rank adapters, allowing the model to be adapted to a new domain on a single consumer-grade or low-cost cloud GPU, while retaining accuracy close to that of full-parameter fine-tuning. This keeps the computational cost of adaptation low enough to be practical for a small business, without requiring specialized infrastructure. To supplement the fine-tuned model's built-in knowledge with accurate, business-specific facts, we integrate a Retrieval-Augmented Generation (RAG) module into the system. This module uses the ColBERT [5] late-interaction retriever to index and search a small corpus of PDF documents containing product information, warranty details, and installation guides. At inference time, relevant passages retrieved from this corpus are supplied to the LLM as additional context, allowing the chatbot to tailor its responses to the specific product or policy referenced in a given customer inquiry, rather than relying solely on what was seen during fine-tuning.

The fine-tuned model and RAG module are orchestrated using LlamaIndex, which manages indexing, retrieval, and prompt construction for interaction with the LLM. This is exposed through a lightweight Gradio interface, which is in turn wrapped by a Flask backend and a React frontend to provide a chat interface deployable on a business's website. For inference-time hosting, the fine-tuned model is served on an on-demand GPU cloud instance, allowing it to handle customer traffic without requiring the business to maintain dedicated hardware.

To further improve the chatbot's behaviour in situations not directly covered by the fine-tuning data, a small number of example conversations are included as few-shot context within the retrieval corpus itself. These examples act as an in-context guide for handling ambiguous or unfamiliar queries, helping the chatbot produce coherent, appropriately scoped responses even outside the immediate coverage of its fine-tuning data.

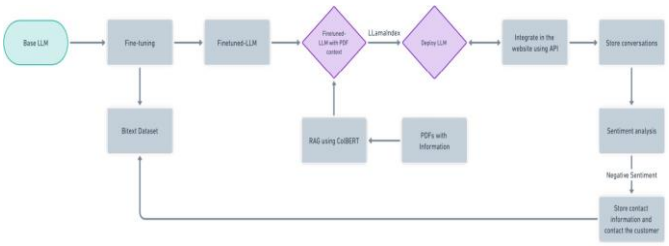


Figure 2: Flowchart of a data processing pipeline, starting with a base language model and ending with sentiment analysis and conversation storage.

Finally, to support ongoing monitoring and improvement, a sentiment analysis module is integrated into the deployed system. This module applies a Hugging Face sentiment analysis pipeline to logged conversations between the chatbot and customers, identifying interactions in which the chatbot's responses are likely to have produced negative customer sentiment. When such an interaction is flagged, the associated customer contact information is recorded in a dedicated follow-up queue, so that a human representative can intervene. This feedback loop provides a low-overhead mechanism for identifying weaknesses in the chatbot's responses and prioritizing them for review, supporting incremental improvement of the system over time without requiring continuous manual monitoring of every conversation.

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

Taken as a whole, this methodology is intended to give small businesses a practical, cost-effective route to a customer service chatbot that is both linguistically capable and factually grounded in their own documentation, while providing a mechanism for surfacing and addressing weak points in its behaviour through ongoing use.

## IV. RESULT

Our implementation shows that the proposed pipeline is effective at producing a tailored, low-cost customer service chatbot for small businesses. The fine-tuned Llama 3.1 8B Instruct model, augmented with the RAG module, handled a range of representative customer queries with contextually appropriate and factually grounded responses, as illustrated in Figure 3. Integration of the ColBERT retriever allowed the system to reliably surface and incorporate relevant passages from the indexed PDF documentation into its responses.

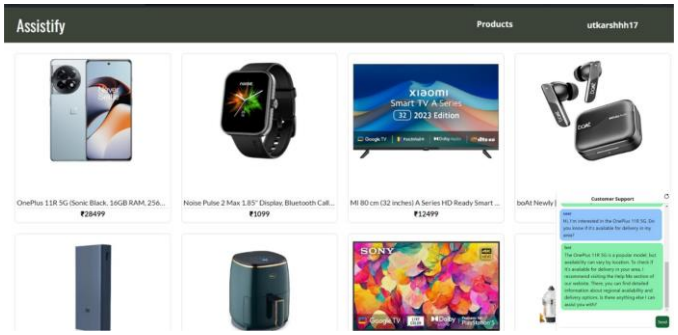


Figure 3: Chatbot interface integrated in the website

A central design goal of this work was resource efficiency, so that the resulting system remains accessible to small businesses without dedicated ML infrastructure. Fine-tuning and inference with the QLoRA-adapted model fit comfortably within the memory budget of a single low-cost cloud GPU instance, reflecting the efficiency gains that quantized, low-rank adaptation provides over full-parameter fine-tuning. This allows the system to be hosted affordably on-demand rather than requiring dedicated, always-on hardware, which we view as an important factor for small-business adoption.

To assess response quality, we compared our chatbot's outputs against a frontier proprietary model on a set of 30 representative customer service queries, using BERT-based semantic similarity as an automatic proxy for response quality. Our chatbot achieved an average semantic similarity score of 0.81 relative to the frontier model's responses across these 30 queries, suggesting that the fine-tuned, retrieval-augmented pipeline produces responses broadly comparable in relevance and coherence to a substantially larger, general-purpose system.

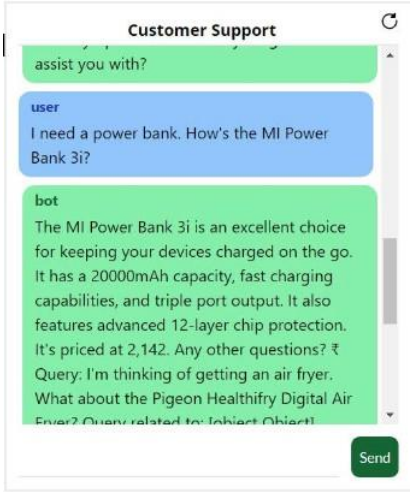


Figure 4: Sample conversation with the customer

Sentiment analysis of logged conversations indicated a predominantly positive distribution of customer sentiment, with a small proportion of interactions correctly flagged as negative and routed to the human follow-up queue. The inclusion of example conversations in the retrieval corpus appeared to help the chatbot produce more coherent responses in less-familiar scenarios. The deployed Gradio-Flask-React interface, hosted on an on-demand GPU instance, provided smooth end-to-end interaction throughout testing. Together, these observations suggest that the proposed methodology provides a viable and resource-efficient customer service solution for small businesses.

## V. CONCLUSION AND FUTURE WORK

This work presents a resource-efficient methodology for building tailored customer service chatbots for small businesses. By fine-tuning Llama 3.1 8B Instruct with QLoRA via an open-source PEFT toolkit, we adapt a general-purpose open-weight model to the customer service domain at low computational cost. Integrating a RAG module built on ColBERT further grounds the chatbot's responses in business-specific documentation, and a lightweight sentiment-analysis feedback loop supports ongoing quality monitoring without continuous manual review. Together, these compo-nents form a practical, end-to-end pipeline for small-business customer service automation.

Several directions remain open for future work. Expanding the fine-tuning corpus with more diverse, real-world customer

interactions could further improve the chatbot's generalization to edge cases. Incorporating multilingual base models would allow the system to serve customers across a broader range of linguistic markets. Personalization features that condition responses on individual customer history could improve relevance over repeated interactions. Finally, exploring smaller, edge-deployable models could reduce latency and hosting cost further, particularly for businesses operating in regions with limited or unreliable internet infrastructure. We view these directions as natural extensions of the current pipeline rather than departures from it.

## REFERENCES

1. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," Advances in Neural Information Processing Systems (NeurIPS), 2023.
2. AI-Meta, "The Llama 3 Herd of Models," arXiv:2407.21783, 2024.
3. E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations (ICLR), 2022.
4. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), 2020.
5. O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020.
6. S. M. Bsharat, A. Myrzakhan, and Z. Shen, "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4," arXiv:2312.16171, 2023.
7. SuperAnnotate AI Inc., "Fine-tuning Large Language Models (LLMs) in 2023," SuperAnnotate Blog, n.d.
8. L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment," arXiv:2312.12148, 2023.
9. Y. Zheng, R. Zhang, J. Zhang, Y. Ye, and Z. Luo, "LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models," Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2024.
10. Khaleel Khan Mohammed. "The Future is Cloud: Modernizing Big Data for the Cloud Era