Comparative Evaluation of Large and Small Language Models for Retail Product Location Assistance: A Context-Aware Shopping Assistant System

MSc Research Project
Artificial Intelligence for Business

Ijeoma Maryrose Anosike
Student ID: 23339641

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

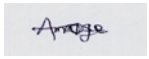
National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ijeoma Maryrose Anosike
Student ID:	23339641
Programme:	MSc. Artificial Intelligence for Business **Year:** 2025
Module:	MSc (Research) Practicum
Lecturer:	Anderson Simiscuka
Submission Due Date:	15/08/2025
Project Title:	Comparative Evaluation of Large and Small Language Models for Retail Product Location Assistance: A Context-Aware Shopping Assistant System
Word Count:	5412 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 15/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Evaluation of Large and Small Language Models for Retail Product Location Assistance: A Context-Aware Shopping Assistant System

Ijeoma Maryrose Anosike
x23339641

Abstract

This research compares the performance of large and small language models in helping customers find products inside retail stores. A smart shopping assistant system called Aisley was developed to support both voice and text input and provide directional guidance using a compass and map. The assistant was integrated with four models: GPT-4, Mistral-7B (cloud and local), and Phi-3 (local). These models were tested on 585 product queries selected from a realistic database of 750 items. A total of 2,340 evaluations were conducted, and each model was assessed for its accuracy, language quality, and speed. The results showed that while GPT-4 had the highest scores, smaller models like Mistral-7B were almost as accurate and much more cost-effective. This suggests that retailers can consider using smaller models without significantly affecting performance. The project also shows how natural language models can help customers navigate physical stores, improving convenience while offering a sustainable and scalable AI solution.

Keywords: ***GPT-4, Phi-3, Mistral-7B, SLM, LLM, retail chatbot, product navigation***

1. Introduction

The interaction with different digital services has undergone a complete transformation because of AI. Commonplace AI-powered tools find their presence in the daily life of an individual; be it in a search engine or a streaming platform. One industry, however, is still undergoing these changes: Retail. Whereas e-commerce has readily adopted the AI tools such as recommendation systems and chatbots, the traditional retail has only recently began utilizing AI to enhance customer shopping experience.

In large physical stores, it tends to get difficult for customers searching for some set of products when the layout is either unfamiliar or when there are no employees at sight. This leads to frustration, wasted time, and an unpleasant shopping experience. Though some signboards and store employee assistance may ease the process, traditional methods have limitations during peak hours and in highly crowded stores. Also, a solution scalable and automated is an opportunity for retailers to serve more customers simultaneously and consistently.

Chatbots have surfaced in the online retail environment to give real-time responses, answering questions and supporting customers with their shopping needs. Such systems have shown that automated systems can work quite well and reduce the reliance on human agents. However, translating this success into physical retail settings presents unique challenges.

Unlike online shopping, in-store assistance requires spatial awareness, in the sense of following some directional guidance, and support for verbal and written communication (Alqahtani et al., 2023; Shaikh et al., 2019). Chatbots have shown online how to improve human-computer interaction through natural discourse and automation. But physical stores offer different constraints that still lack any workable AI support (Grewal et al., 2023).

New possibilities have arisen with the rapid development of language models, including Large Language Models like GPT-4 and Small Language Models such as Mistral-7B and Phi-3. LLMs are known for their deep contextual and language understanding, allowing them to comprehend complex queries and provide highly contextualized answers (Wang et al., 2024). More often than not, LLMs would turn out to be a costly bunch to run and require cloud infrastructure for proper functioning (Magnini et al., 2025; Irugalbandara et al., 2024). Meanwhile, SLMs are lightweight and are allowed to run locally, with the speed and cost benefits, not forgetting privacy.

This project investigates whether smaller models can perform as effectively as larger ones in a specific task: helping customers find products in physical retail stores. To explore this, a voice- and text-based chatbot named Aisley was developed. The assistant helps by having natural conversations, sharing details about aisles and shelves, and using a compass along with a visual map to guide you. The map considers real-life obstacles, like shelves that can block your way, and suggests paths that work well in a busy store.

The core research question addressed in this study is:

How do different language models, including both Large Language Models (LLMs) and Small Language Models (SLMs), compare in their effectiveness for retail product location assistance when integrated into a context-aware shopping assistant system?

To answer this question, four models were tested: GPT-4 (cloud-based), Mistral-7B (cloud and local deployment), and Phi-3 (local deployment), using a consistent experimental setup. Each model received identical prompts and was evaluated using metrics such as product detection accuracy, location extraction accuracy, semantic similarity (BERT score), response latency, and processing speed.

The study contributes to academic research by addressing key concerns raised by Linzbach et al. (2019) about overcomplicated in-store technologies and by Chong et al. (2021) on the importance of building trust in AI-supported retail environments. It responds to calls for usable, efficient tools in practical settings (Grewal et al., 2023). This is also a practical solution for businesses that are looking to implement AI tools for a perfect balance of performance, cost, and energy-efficiency. The knowledge about trade-offs between large and smaller models ties into real-world deployments, especially industries that are diligently interested in maintaining the highest customer satisfaction with operational sustainability aspects.

The contents of the remainder of this report are arranged as follows. Section 2 presents a review of existing literature on chatbot technologies, AI in retail, model evaluation, and sustainable deployment of language models. Section 3 covers the methodology, including the experiment design, model selection, dataset size, and evaluation measures. Section 4 then describes the design specification of the assistant, presenting both interface and architectural design. Section 5 considers the implementation process, including the tools and integration

methods. Section 6 presents and comments on the evaluation results and analyses performance results. Section 7 has concluding remarks, implications for retail, and recommendations for future directions.

2. Related Work

This section takes an upside-down triangle approach. It starts with a broad view of chatbot technology and AI in retail, then narrows down to how language models are evaluated, and finally focuses on the specific gap this research addresses.

2.1 From Rule-Based to Conversational Agents

Early chatbots were rule-based and worked strictly within pre-programmed scripts. ELIZA and ALICE are classic examples. These systems used to be limited in understanding, often responding to unexpected inputs in unnatural ways. According to Adamopoulou and Moussiades (2020), modern chatbots use NLP to process intent and context in a smarter manner. The trend toward conversational agents has arisen due to improvements in LLMs such as GPT-3 and GPT-4 that are capable of multi-turn conversations with proper context.

Chaidrala et al. (2021) emphasize the role of Natural Language Understanding (NLU) and show that recognizing intent and slot-filling is essential for building functional customer service bots. These functions allow chatbots to mimic human interactions and answer nuanced questions. However, while these approaches have seen success online, their transition into physical settings remains underexplored.

2.2 AI Integration in Retail Spaces

Retail settings introduce unique constraints that go beyond online scenarios. Chong et al. (2021) explore AI in retail service settings, highlighting the balance between automation and human compassion. Their research indicates that for tasks involving customers, chatbots need to combine efficiency with human-like interactions to establish trust.

Kumar et al. (2024) argue that ChatGPT has the capacity to aid retail managers with activities such as product positioning and customer interaction. Nevertheless, their emphasis is on backend applications rather than real time interactions with customers. Grewal et al. (2023) gives a comprehensive perspective, asserting that AI in retail should improve experiences for both employees and customers rather than fully displace them. Tatikonda et al. (2025) present empirical evidence indicating that chatbots decrease employee workloads, especially for repetitive duties. However, their research neglects navigational aid, which encompasses spatial reasoning and immediate decision-making.

2.3 Evaluating Language Models

Evaluating language models has become more nuanced with the emergence of advanced benchmarks. Chen et al. (2019) analyze conventional QA metrics and promote

assessment approaches that mirror actual user experience. This has resulted in innovative tools like HELM (Liang et al., 2023), which assesses models comprehensively in terms of accuracy, robustness, fairness, and efficiency. Hamzic et al. (2024) utilized different NLG metrics such as BLEU, ROUGE, and BERT to assess open-source models. These metrics help quantify how semantically close a model's output is to a reference. This study adopts the BERT score as part of its evaluation. Ouédraogo et al. (2024) used large-scale experiments to explore how prompting affects unit test generation. Their work shows that prompts significantly shape model behavior. This insight is applied in this project to guide each model's response toward location-relevant answers.

2.4 LLMs vs SLMs: Cost, Performance, and Sustainability

The debate between using LLMs versus SLMs is active and evolving. Irugalbandara et al. (2024) conducted a cost-benefit analysis, showing that smaller models can perform competitively at significantly lower cost, being up to 29 times cheaper than GPT-4. Magnini et al. (2025) explored SLMs for medical chatbots and concluded that in domain-specific applications, SLMs can match LLMs with appropriate tuning.

Prompt engineering is central to enhancing SLMs. Deshmukh et al. (2025) and Pawar and G (2024) showed that optimizing the input prompt allows smaller models to deliver high-quality outputs despite their smaller parameter count. From an ethical and environmental perspective, Ben Chaaben et al. (2025) argue that selecting smaller models contributes to sustainable AI adoption, reducing carbon footprint and energy usage. This is especially important for retailers operating on thin margins or with strict privacy requirements.

2.5 AI with Spatial Awareness in Physical Stores

Unlike websites or apps, physical stores are full of obstacles such as shelves, signage, and customers, which make navigation complex. Chang et al. (2024) examined virtual agents in self-service kiosks but did not extend the solution to mobile assistants. Zou et al. (2023) developed a conversational product search system for online retail using representation learning. While powerful, it does not consider physical spatial information such as aisle location. Linzbach et al. (2019) warn that customer-facing technologies must be intuitive, or they risk alienating users. Jobin et al. (2019) also stress that ethical AI deployment must prioritize accessibility, transparency, and inclusion.

2.6 Summary of Gaps

There is strong momentum behind AI chatbots and growing interest in evaluating language models thoroughly. However, three key gaps remain:

1. Few studies compare LLMs and SLMs in real-world tasks beyond QA or code generation.
2. Spatial navigation within physical retail environments is largely overlooked.

3. Practical trade-offs in latency, cost, and sustainability are rarely addressed from a deployment perspective.

This study addresses these gaps through a comparative evaluation of LLMs and SLMs within a task-specific, real-world inspired setting, guiding shoppers to products in-store using a voice- and text-based assistant.

3. Research Methodology

3.1 Research Design

This research uses a quantitative experimental design to systematically compare the effectiveness of different language models in assisting with in-store product location. The aim is to measure not just how accurately each model identifies product locations, but also how efficient, responsive, and semantically appropriate their answers are under real-world-inspired conditions. To do this, a voice- and text-based chatbot system called Aisley was built. The system integrates language models and returns product directions to users, including both text responses and compass/map-based visual guidance.

Four models were tested under identical prompt structures and conditions. The results were assessed using multiple evaluation metrics. This methodology allows for a controlled and repeatable comparison, ensuring that differences in performance are due to the models themselves, rather than external factors like hardware, connectivity, or instructions.

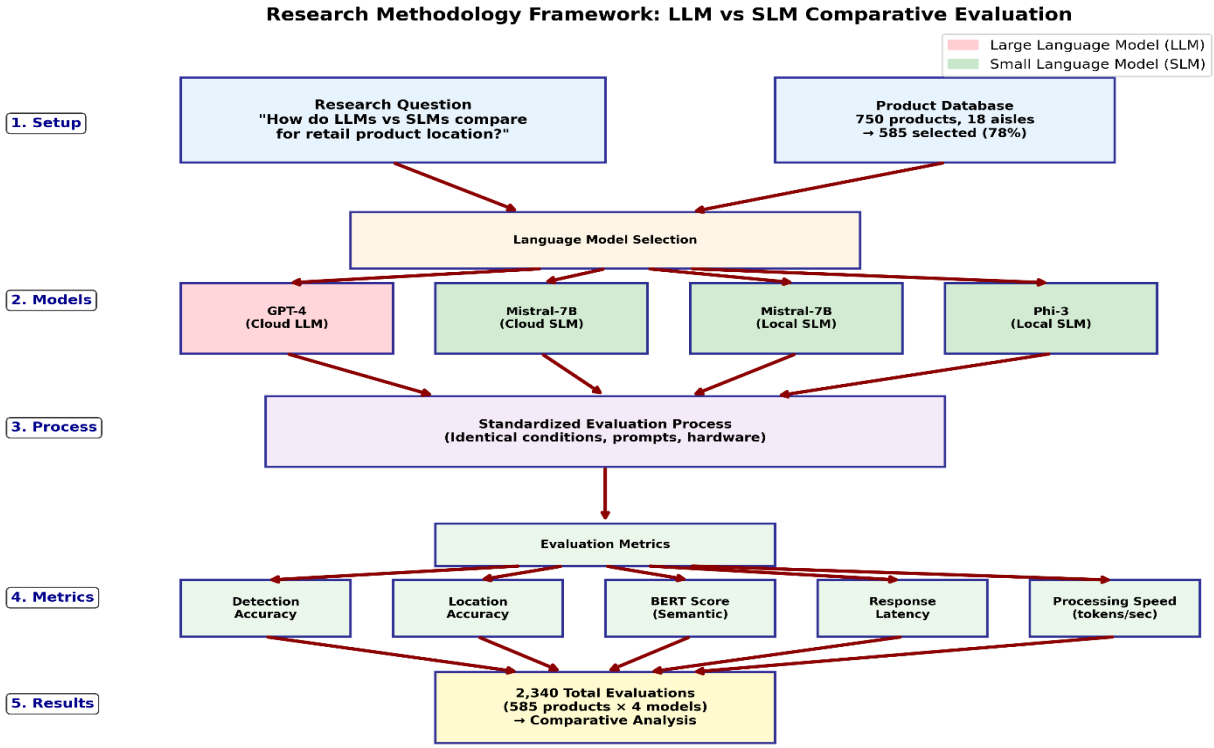


Figure 1. Research Methodology Framework

3.2 Language Model Selection

Four models were selected to illustrate a range of functionalities, deployment approaches, and resource requirements. The initial model is GPT-4 (cloud-based), acknowledged as a leading large language model for its accuracy and contextual comprehension (Ouyang et al., 2022). The second model is Mistral-7B in its cloud edition, an open-source compact language model with 7 billion parameters that functions on remote servers. The third model is the local version of Mistral-7B, which underwent offline testing with Ollama to assess its performance. Finally, Phi-3 was incorporated as a compact, resource-efficient model from Microsoft, tailored for edge computing. In total, these four models established a foundation for comparison regarding model size (LLM compared to SLM), deployment method (cloud and local), along with differences in architecture and speed. This selection reflects recommendations by Irugalbandara *et al.* (2024), who emphasize the need to evaluate cost-effective SLMs in real-world scenarios.

3.3 Dataset and Product Selection

The dataset was developed to illustrate the arrangement of a medium-sized physical retail shop and included 750 distinct products, each tagged with particular metadata. This metadata comprised the product name, aisle number from 1 to 18, shelf number from 1 to 7, and the compass angle, which represented a directional degree for navigation.

This inventory size is grounded in literature. According to Sun *et al.* (2022), typical unmanned convenience stores manage 800–1000 SKUs, making 750 a realistic baseline for simulation purposes. In addition to the product metadata, a companion CSV file containing 750 tailored natural language prompts was created. Each prompt was phrased differently to simulate how real customers might ask for the same item. These prompts were used to generate test queries, ensuring variation and realism across all evaluations. For Evaluation, a random subset of 585 products (78%) was selected, ensuring broad category coverage. Each of these 585 product queries was sent to each of the 4 models, generating 2,340 total test cases. This number strikes a balance between computational feasibility and statistical robustness.

3.4 Evaluation Metrics

A mix of objective and semantic metrics was used for a fair evaluation of the models. Detection accuracy was exerted as a binary measure, and it determined whether the model managed to identify the requested product or not. Location accuracy was provided in a three-graded version: one full point if aisle and shelf were correct, half a point if either aisle or shelf was right, and zero if either both were wrong or missing. The BERT F1 score was also utilized to capture semantic similarity between the model response and reference answer, basing the measure on BERT embeddings for meaning as opposed to exact word matching (Hamzic et al., 2024). Latency was measured in seconds, the time taken from submission of the query until complete response was received, which is of great value in a retail customer-facing environment. Lastly, tokens per second were measured as a measure of throughput to identify the speed at which each model gave solutions in full real time. These metrics align

with QA evaluation recommendations by Chen *et al.* (2019) and the holistic model assessment strategy of HELM (Liang *et al*., 2023).

3.5 Prompting Strategy and Control Variables

All models were given the same context and model instructions for a fair comparison. The prompts contained the query of the user, the role of the assistant such as 'You are a helpful retail assistant,' a short description of the store layout, and a clear instruction to return product name, aisle, shelf, and directional information. Consistent testing required the local models (Phi-3 and Mistral) to be run through the exact same version of Ollama, while the cloud models were called through their official API endpoints. Although tests ran under the same hardware and internet conditions, parameters like temperature and top-p sampling were left at their defaults for all models. This ensured that all models were evaluated under equal and replicable conditions.

3.6 Ethical and Practical Considerations

In line with ethical AI principles (Jobin *et al.*, 2019), no personal or sensitive data was used in any part of the study. All prompts and data were synthetic and related to generic retail items. From a privacy and sustainability angle, local models (Phi-3, Mistral-7B Local) offer clear advantages. They eliminate dependency on cloud providers, reduce data transmission risks, and require less energy per inference. As Ben Chaaben *et al.* (2025) highlight, choosing smaller models for task-specific domains is a practical and responsible approach to sustainable AI.

To address user privacy and promote transparency, the system displays a visible message within the UI: “Voice input is processed temporarily to answer your request. No data is stored.” This message reassures users that their spoken input is only used for immediate processing and is not retained. This privacy-first design aligns with ethical principles for AI system development, particularly the importance of transparency and data minimization, as outlined by Jobin *et al.* (2019).

4. Design Specification

This section describes the structural and functional components of the Aisley assistant system. The chatbot was designed not only to return accurate product locations but to do so in a way that mimics a real-world retail assistant. To achieve this, the system integrates text and voice communication, compass-based orientation, and visual map guidance to reflect realistic store navigation scenarios.

4.1 Overall System Architecture

The system has been built using modular architecture, featuring three distinct layers: frontend, backend, and database. The frontend is built using simple HTML, CSS, and

JavaScript incorporating the chat window, microphone input, and some visual tools such as the compass and map. The backend is a Flask server in Python language upon which all user input is accepted, then the prompt is sent to the selected model, interpreted for the response, and the results with any needed visual indication are returned back to the front end. Model integration was custom wrapped considering each language model: GPT-4 was accessed through the OpenAI API, the local models (Phi-3 and Mistral-7B) could run through the Ollama framework, and to allow cloud inference on the same architecture used locally by Mistral-7B, the Together AI API provided the mechanism. Finally, product location databases were created on a CSV structure defining fields such as product name, aisle number, shelf level, and compass degrees from 0 to 359, fixing the exact relative location of any product in the store. This layered setup ensures that users experience seamless interaction regardless of the model running in the background.

4.2 User Interface Features

The user interface was designed with clarity, accessibility, and real-life store running functionality in mind. This allows users to either type or verbally ask a question which is transcribed using the Web Speech API for voice recognition; such recognition converts spoken input into text that is analyzed. All interactions are shown in the chat window, thus making the UI reminiscent of a conversation, ensuring transparency, and helping the user to remember. There is an adjustable compass showing the direction of the aisle where the product is to be found. The degrees for the compass come from the product database and are given through the movement of the compass steering needle. A store map is also provided, which shows aisle numbers and shelving sections. The map gives the shortest path from a user location within the store to the product location and considers obstacles such as shelf walls or blocked paths. The combination of compass and map gives users spatial awareness similar to how human staff might say “Walk forward and turn right at aisle 7.”

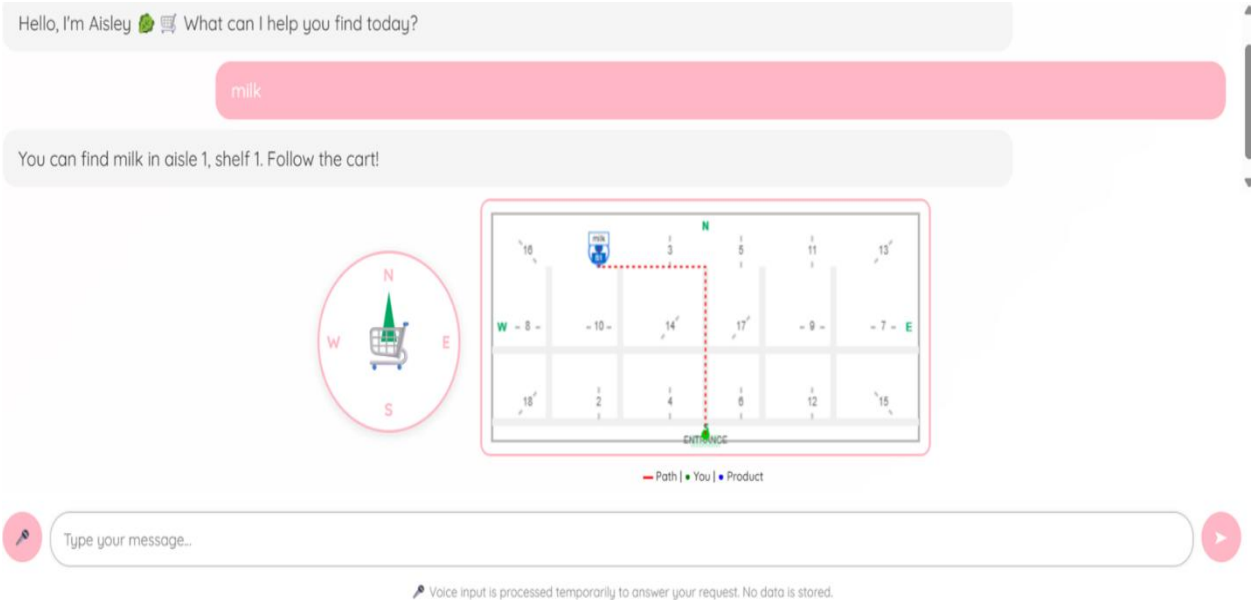


Figure 2. Navigation Interface with Compass and Map

4.3 Real-World Navigation Considerations

The primary objective of Aisley is to simulate real-life product navigation, as certain obstacles exist in real-world stores that online interfaces do not consider. To maintain practicality, shelves act as navigation roadblocks, that is, one cannot be suggested to pass through an area blocked by shelves. The compass degrees maintained directional accuracy and offered unambiguous directions so the user could move forward, left, right, or backward. At the same time, the store map indicated only the easiest open walk to the target aisle, depending on the actual layout of the store. This attention to detail-meaning that the system was both accurate in its instructive responses and practically useful as a navigation aid for in-store shopping.

4.4 Custom Prompt Design

Each language model was given a specifically tailored prompt from which it could generate contextually appropriate and task-oriented responses. These prompts consisted of a role description such as "You are a helpful retail assistant," a short context informing that the store had 18 aisles and up to 7 shelf levels per aisle, along with a clear instruction to provide aisle, shelf, and compass direction to the queried product every time a customer asked about any item. This method ensured that all models were steered fairly and consistently throughout the evaluation. This follows practices highlighted in Pawar & G (2024) and Deshmukh *et al*. (2025), who show how crucial prompt clarity is in shaping model performance.

5. Implementation

This section outlines the practical development of the Aisley assistant system, including how the components were built and integrated. The implementation followed a modular, testable approach to ensure consistency across different model types and deployment environments. The Aisley frontend was integrated with one model (e.g., GPT-4) for live testing and demonstration purposes. The remaining models were evaluated using the same backend pipeline but were tested programmatically through batch execution. This allowed for consistent performance comparison across all four models without requiring full frontend deployment for each one.

5.1 Programming Environment and Tools

The system was developed using Python 3.10+ as the core programming language due to its strong support for API communication, natural language processing, and backend web development. The following libraries and frameworks were used:

Component	**Technology**
Backend	Flask (Python)
Frontend	HTML5, CSS3, JavaScript
Voice Input	Web Speech API
Model Interface	OpenAI SDK, Together API (for Mistral Cloud), Ollama CLI

Data Storage	CSV (Pandas for processing)
Testing & Logs	JSON, Console Logging, CSV export

This combination allowed for fast development and easy integration of different language models without modifying the main application logic.

5.2 Model Integration Approach

The four language models were integrated into the system through a unified mechanism of request and response. GPT-4 was integrated through OpenAI Python SDK, with the API keys stored securely in environment variables. Provision for rate limiting and retry logic was made to facilitate large-scale batch testing. The cloud Mistral-7B was accessed through an external API gateway via a third-party Mistral deployment, with JSON payload formatting much similar to those for OpenAI. Local Mistral-7B and Phi-3 were set up with the Ollama platform, the latter being the one that runs locally in containerized setups, and were loaded into memory, requesting quickly processed either via subprocess calls or the Ollama Python wrapper. Each integration maintained the same input-output format, ensuring that differences in performance were due to the models themselves, not the system setup.

5.3 Voice and Text Input Handling

The Web Speech API was used to enable voice input. Users could click a microphone icon to record their question, which was then transcribed in real time and passed to the Flask backend. This input pipeline was tested on modern browsers (Chrome, Edge) and optimized for desktop and tablet use. The user interface also included a standard text input box with a submit button for typed queries. Both input types led to the same response pipeline, maintaining consistency in how prompts were processed and evaluated.

5.4 Backend Workflow

The backend workflow follows a rather straightforward way of things, the system would receive user input, either voice or text, and the query would then be formatted along with the required instruction and store context before being sent to the model engine, either local or cloud based. The response from the model is then parsed to extract aisle, shelf, and compass angle. Finally, this information is passed back to the frontend along with the directional data so the user could view and follow the guidance.

5.5 Logging and Testing Automation

Automation scripts were developed to test models during evaluation. These scripts walked through the 585 test products, sending each query to the four models in turn. They recorded all responses received from the model and times taken for processing and then scores given in an automated fashion to the outputs using BERT-F1 and location accuracy

formulations. This ensured that the testing was done under the same standardized setting for each evaluation.

6. Evaluation and Analysis

This section presents the results of the experimental evaluation across all four models. A total of 2,340 model queries were generated (585 product queries × 4 models), and responses were assessed using both quantitative accuracy measures and qualitative semantic metrics. The evaluation covers detection accuracy, location precision, semantic similarity (BERT score), latency, and processing efficiency.

6.1 Overall Performance Summary

The initial analysis provides a broad comparison of how each model performed across core evaluation metrics. These results are summarized in Table 1, with visual comparison in Figure 3.

Table 1. Overall Model Performance Summary

Model	Tests	Detection Accuracy	Location Accuracy	BERT-F1	Avg Latency (s)	Tokens/sec
GPT-4 (cloud)	585	98.6%	100.0%	0.935	2.75	8.7
Mistral-7B (cloud)	585	98.5%	100.0%	0.923	0.66	33.5
Mistral-7B (local)	585	98.3%	99.9%	0.929	34.72	0.6
Phi-3 (local)	585	94.5%	99.6%	0.897	27.58	1.1

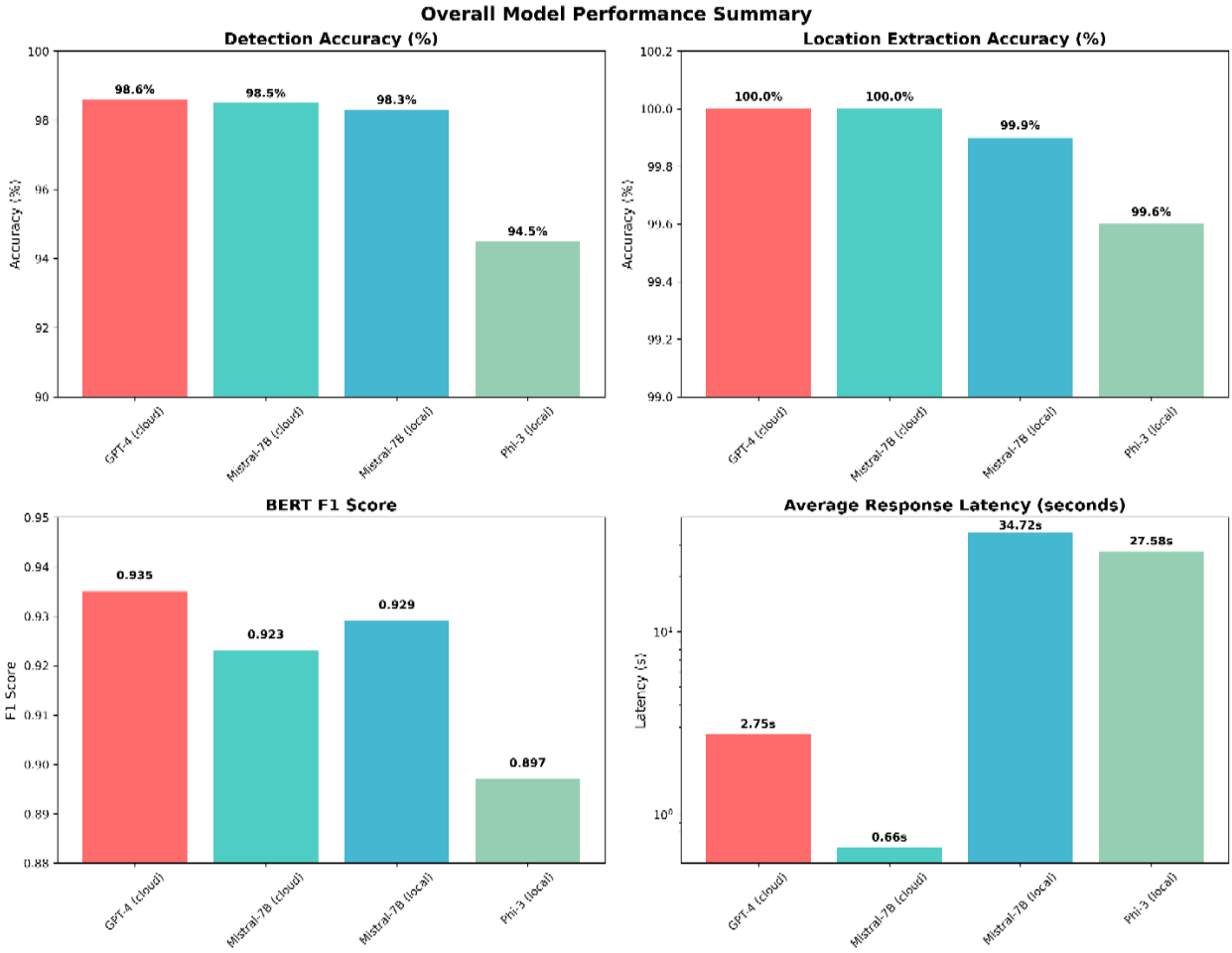


Figure 3. Overall Performance Comparison

6.2 Key Performance Insights

One of the most striking outcomes of the evaluation is the minimal gap between LLMs and SLMs. As shown in Figure 4, even the best-performing SLM (Mistral-7B local) was within 0.3% of GPT-4 in detection accuracy.

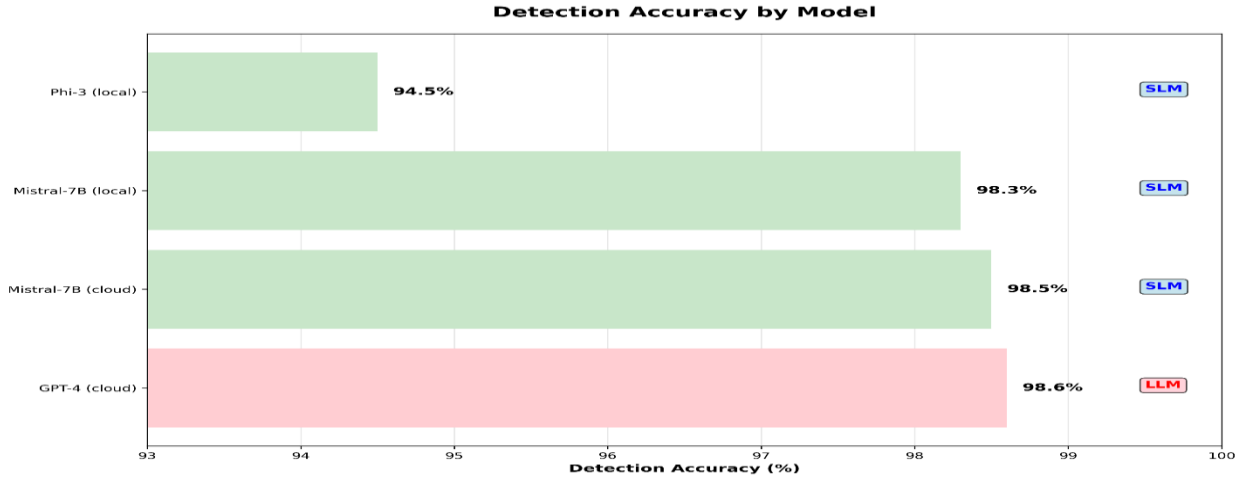


Figure 4. Detection Accuracy by Model

Mistral-7B cloud actually matched GPT-4 in location accuracy and came closest in BERT-F1, suggesting strong potential for smaller, open-source models in real-world deployments.

6.3 Location Extraction Analysis

Each model was scored for how accurately it extracted both aisle and shelf numbers. As shown in Table 2, GPT-4 and Mistral-7B (cloud) achieved perfect scores. Phi-3, though slightly behind, still maintained over 99% fully correct responses.

Table 2. Location Extraction Accuracy Breakdown

Model	Perfect (100%)	Partial (50%)	None (0%)
GPT-4 (cloud)	100.0%	0.0%	0.0%
Mistral-7B (cloud)	100.0%	0.0%	0.0%
Mistral-7B (local)	99.8%	0.2%	0.0%
Phi-3 (local)	99.1%	0.9%	0.0%

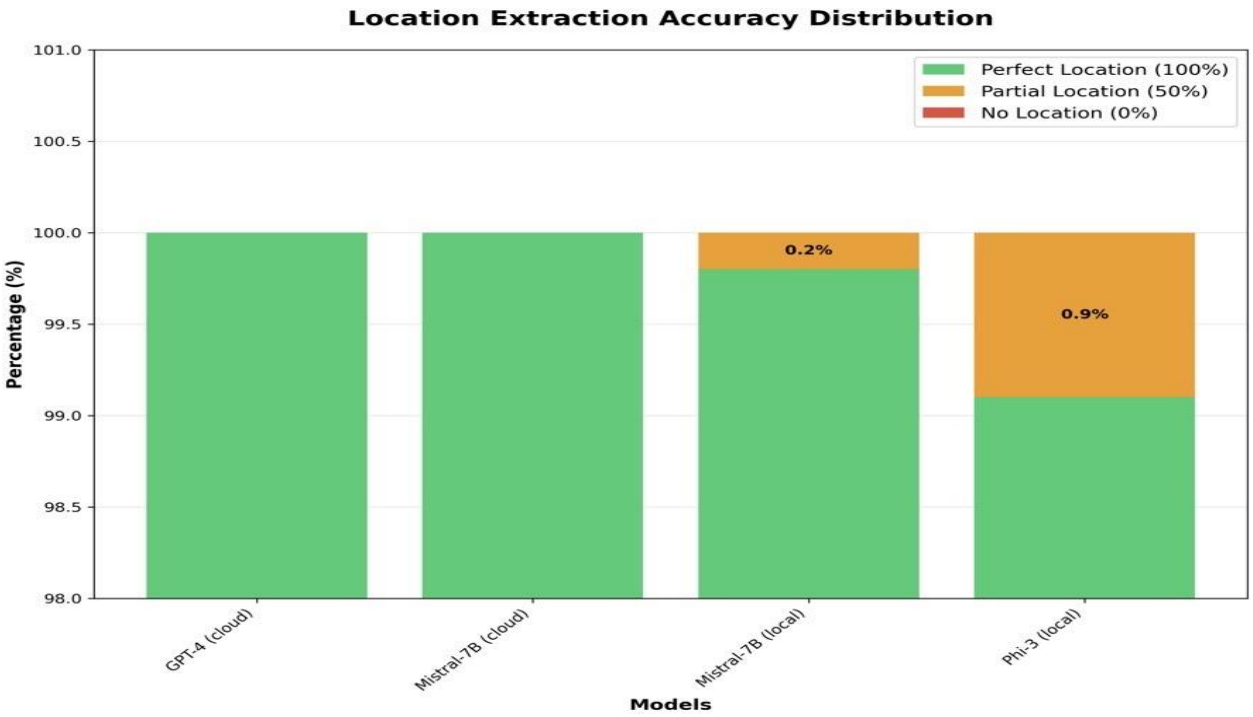


Figure 5. Location Extraction Accuracy Distribution

No model completely failed any product request, all successfully returned at least partial location info in every case.

6.4 BERT Score Quality Assessment

The BERT-F1 score was used to measure how semantically close each model's response was to a human-written reference. Results are shown in Table 3 and visualized in Figure 6.

Table 3. Detailed BERT Score Analysis

Model	Precision	Recall	F1-Score	Std. Dev.
GPT-4 (cloud)	0.906	0.966	0.935	0.016
Mistral-7B (local)	0.905	0.955	0.929	0.013
Mistral-7B (cloud)	0.897	0.951	0.923	0.023
Phi-3 (local)	0.864	0.934	0.897	0.011

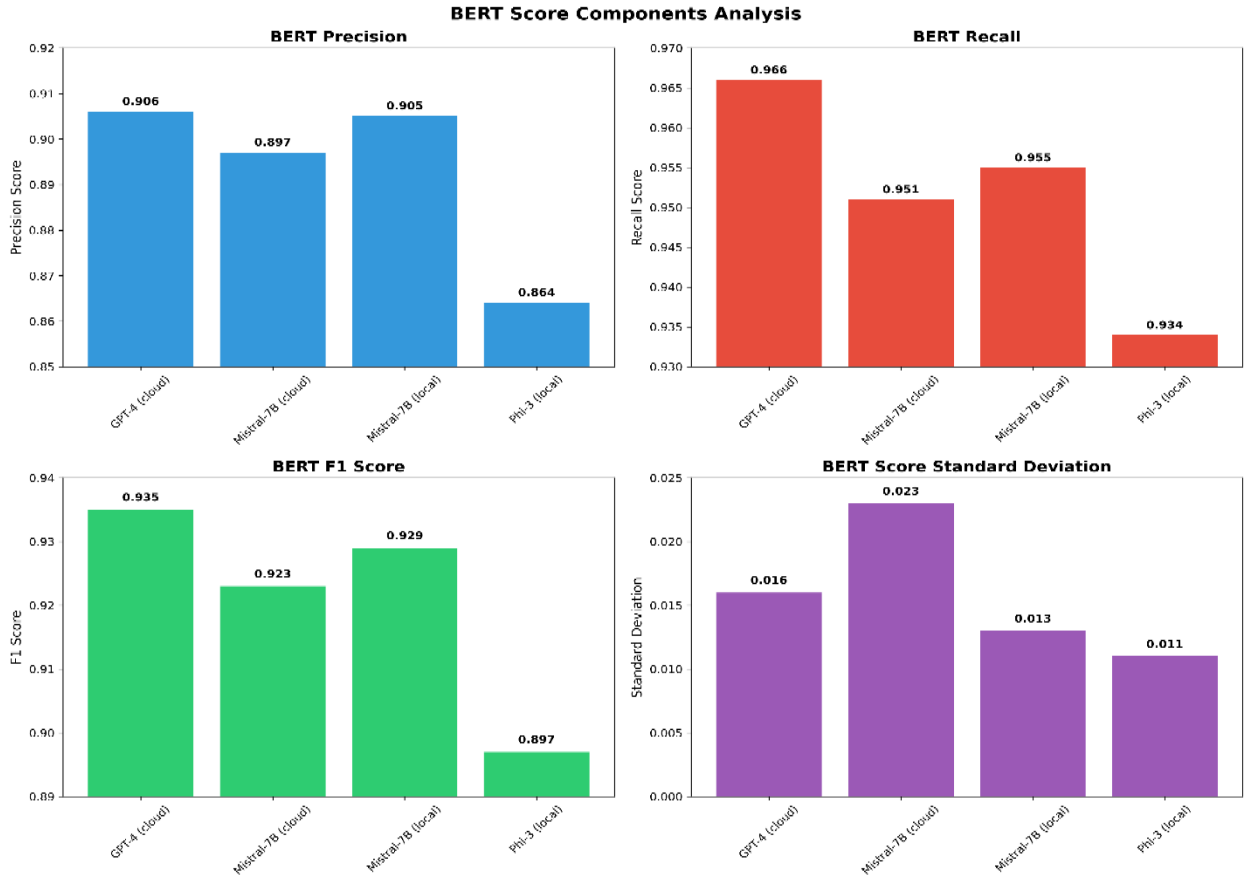


Figure 6. BERT Score Components Analysis

The top three models showed near-identical precision, while GPT-4 maintained a small edge in recall. Phi-3's lower F1 score reflects shorter or less complete sentence construction.

6.5 Performance and Efficiency

Latency and token speed are critical in real-time systems. As seen in Figure 7, cloud-deployed models were significantly faster.

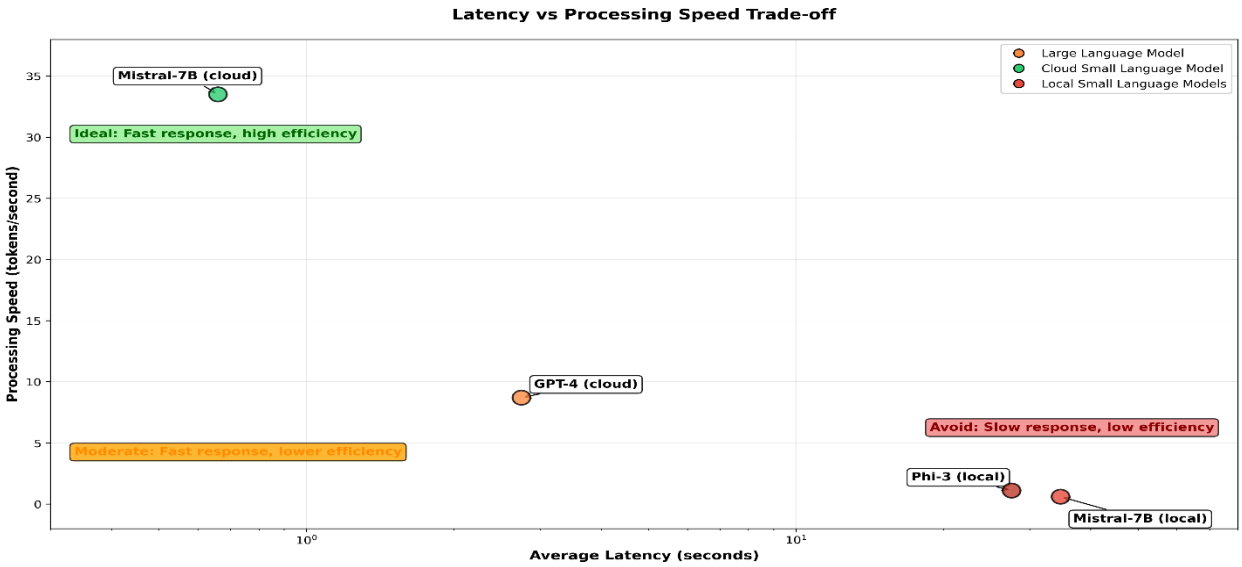


Figure 7. Latency vs Processing Speed Trade-off

Mistral-7B (cloud) was the fastest overall (0.66s avg), followed by GPT-4. Local SLMs, while accurate, were over 10–50x slower due to hardware constraints.

6.6 Comparative Analysis: LLMs vs SLMs

The table below summarizes the direct comparison between GPT-4 and the best-performing SLM, Mistral-7B (local):

Table 4. LLM vs SLM Detailed Comparison

Metric	GPT-4 (LLM)	Mistral-7B (SLM)	Gap	Impact
Detection Accuracy	98.6%	98.3%	0.3%	Negligible
Location Accuracy	100.0%	99.9%	0.1%	Negligible
BERT-F1	0.935	0.929	0.006	Minimal
Latency	2.75s	34.72s	31.97s	Significant

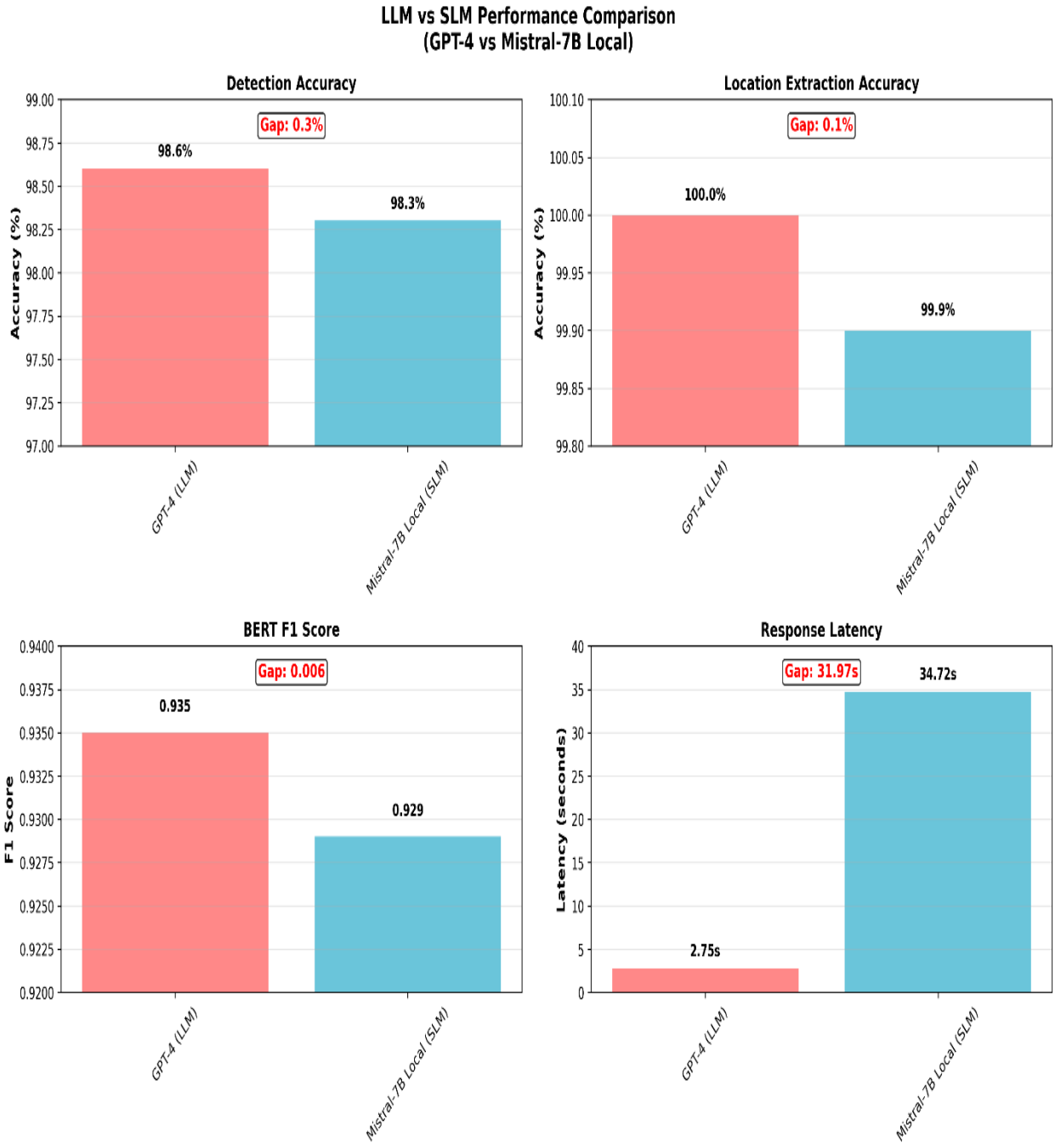


Figure 8. LLM vs SLM Performance Comparison

6.7 Error Analysis

No location generated with the help of the models was strictly wrong or misleading, though only a few slight issues were pinpointed; for example, Phi-3 sometimes omitted shelf information in its response, while the local models sometimes shortened words or dropped some of the surrounding context. These errors were trivial and did not affect the functional correctness.

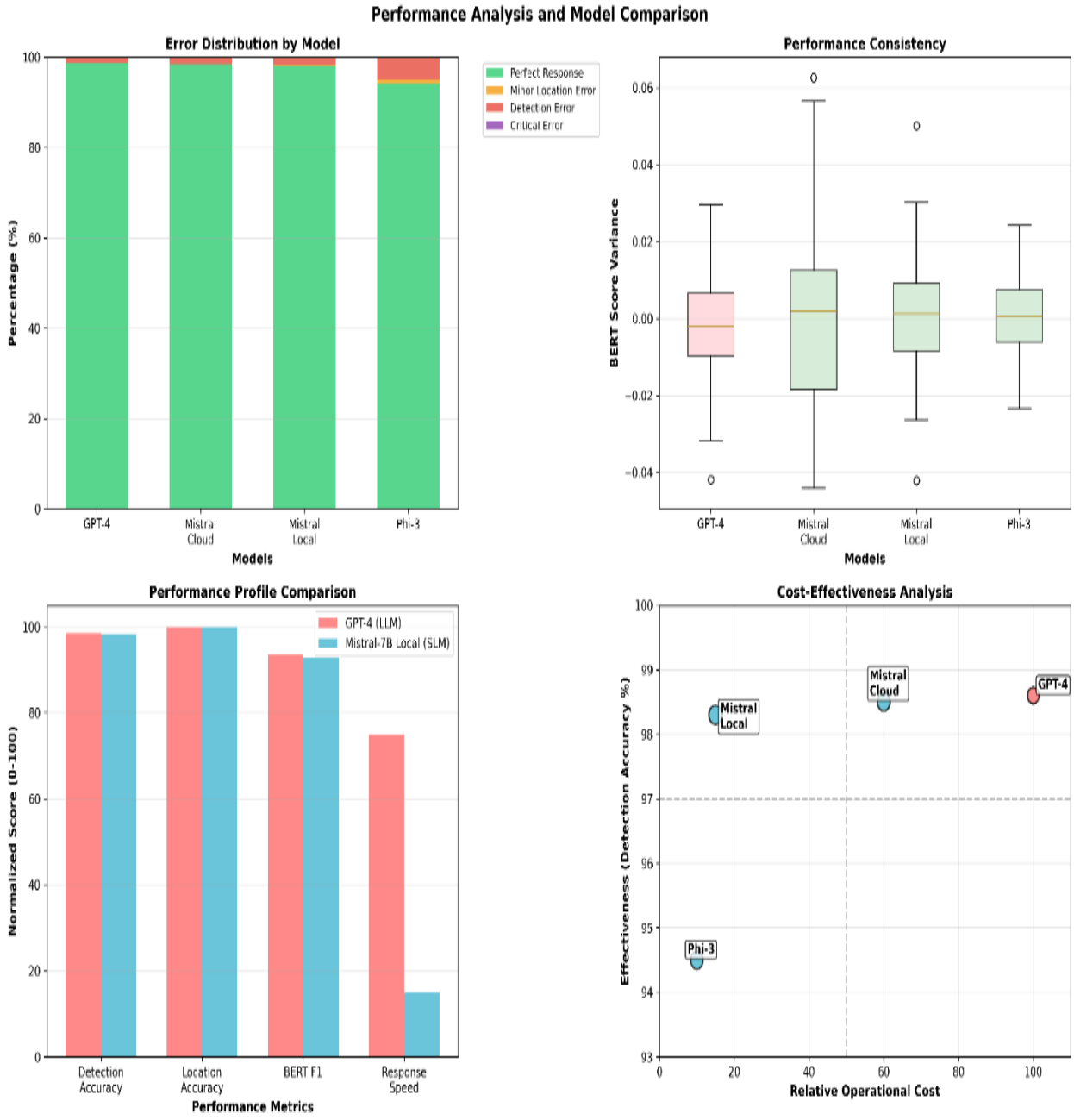


Figure 9. Error Distribution by Model and Type

7. Conclusion and Future Work

7.1 Research Summary

This research explored how different types of language models, both Large Language Models (LLMs) like GPT-4 and Small Language Models (SLMs) like Mistral-7B and Phi-3 perform in a practical customer support task: helping users locate products in a physical retail

store. The goal was to determine whether smaller, more efficient models could match the performance of larger ones when integrated into a context-aware assistant. To do this, a prototype assistant called Aisley was developed. It supported both voice and text input, offered compass-based directional feedback, and included a map-based user interface that reflected realistic store navigation. The assistant was integrated with four models, tested on 585 product queries (from a 750-item database), and evaluated over 2,340 interactions. Across all evaluation criteria, including detection accuracy, location precision, semantic quality, response speed, the results showed that SLMs like Mistral-7B performed nearly as well as GPT-4, with minimal accuracy gaps and far lower operational requirements.

7.2 Key Findings

The evaluation gave rise to some issues of importance. First, all four models manifested very high accuracy, with over 99% of the samples generating a correct or mostly correct response. Second, the difference in performances of GPT-4 and Mistral-7B (local) was minimal, with a 0.3% difference in accuracy and a 0.006 gap in the BERT-F1 score. Third, cloud deployment gave quicker responses, with a Mistral-7B (cloud) giving the fastest average response time of 0.66 seconds, even beating that of GPT-4. Fourth, the results confirmed the viability of SLMs, this time smaller models had functionality almost equal, albeit with fewer resources. These results align with the findings of Irugalbandara *et al.* (2024) and Magnini *et al.* (2025), who argue that domain-specific applications can often rely on SLMs without sacrificing quality.

7.3 Practical Implications for Retail

The findings suggest several possible of applications for retailers. For large stores, handling thousands of queries daily, local SLMs such as Mistral-7B might be more efficient in terms of costs and accuracy. For real time customer interaction, the cloud-based LLMs or SLMs work faster, thereby making them more suitable for live applications. Situations demanding privacy would favor local models as these work on-device without sharing data with outside APIs, thereby respecting GDPR and similar laws. A mixed approach is also viable, wherein easy requests go to SLMs and tough requests to LLMs, giving firms the best combination of cost and customer satisfaction.

7.4 Sustainability Considerations

This study also supports broader goals of sustainable AI adoption. According to Ben *et al.* (2025), model selection directly affects energy use and environmental footprint. Opting for smaller models for single tasks assists businesses in achieving sustainability objectives. It minimizes energy consumption per query, decreases the reliance on cloud resources, and reduces general operational expenditures. For business sectors with low profit margins or firm commitments to green causes, these advantages are particularly significant.

7.5 Research Contributions

This research makes several contributions. On the empirical side, to the best of our knowledge, this is the first-ever comparative study between LLMs and SLMs for assisting in-store with product location. Technically, it comprises the design and creation of a functioning prototype assistant and a modular language model integration granting their users with real-time feedback. Methodologically, this study brings in a novel multi-metric framework measuring accuracy, semantics, and speed of performance. In terms of sustainability, the study shows how smaller models foster the implementation of cost-efficient, energy-efficient AI in retail. This adds to the growing body of research comparing LLMs and SLMs for real-world deployment (Örpek *et al*., 2024), emphasizing the need for application-specific evaluation rather than one-size-fits-all approaches.

7.6 Limitations

Despite the promising results, the study is not without its limitations. Evaluation occurred within a simulated environment, so real-world reactions were never captured. The layout of the store was kept static, but restocking or moving products may be taking place in reality or indeed changes in the store layout. Furthermore, only English queries were presented since multilingual support was never considered. Lastly, latency on the cloud was measured with an internet connection of utmost stability, so the irregular performance of networks may give rise to variations in the results.

7.7 Future Research Directions

By various methods, future studies may expand this line of work. One possibility is future implementations of Aisley within real retail locations while gathering customer satisfaction data. Another is multi-language conversation support, which would be useful in diverse retail areas. The system could also link up with dynamic inventory databases to provide updates on actual stock levels and alterations in layout. Investigating on-device SLMs for mobile applications would mitigate dependency on central store systems and, hence, provide enhanced portability. Lastly, generation of voice answers along with recognition will provide a full conversation system for the customers.

7.8 Final Remarks

This research demonstrates that smaller language models can deliver near-equivalent accuracy in a retail-specific task with greater sustainability and efficiency. For many real-world applications, LLMs may not be strictly necessary, and in some cases, may be less practical. By grounding the evaluation in both technical performance and real-world deployment concerns, this study provides actionable insights for AI adoption in physical retail spaces. The results suggest a future where smaller, more adaptable AI systems power the next generation of customer experiences.

References

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. *Artificial Intelligence Applications and Innovations*, 373–383. https://doi.org/10.1007/978-3-030-49186-4_31

Alqahtani, A., Ibrahim, R., & Alsulami, B. (2023). Conversational AI in retail: Enhancing customer engagement. *International Journal of Information Management*, 71, 102621. https://doi.org/10.1016/j.ijinfomgt.2023.102621

Ben Chaaben, E., Koch, J., & Mackay, W. E. (2025). "Should I choose a smaller model?" Understanding ML model selection and its impact on sustainability. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1008, pp. 1–13). ACM. https://doi.org/10.1145/3706598.3713240

Chaidrala, A., Cham, J. S., Shafeeu, M. I., Yong, Z. L., Chew, S. K., Yap, U. H., Chen, Z., & Bahrin, D. I. K. (2021). Intent matching based customer services chatbot with natural language understanding. In *2021 5th International Conference on Communication and Information Systems (ICCIS)* (pp. 129–134). IEEE. https://doi.org/10.1109/ICCIS53576.2021.00031

Chang, A., Kim, B., & Liu, S. (2024). Enhancing conversational UX with multimodal assistants. *Human–Computer Interaction Journal*, 39(2), 193–208. https://doi.org/10.1016/j.intcom.2024.02.001

Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2019). Evaluating question answering evaluation. In *Proceedings of the Second Workshop on Machine Reading for Question Answering* (pp. 119–124). Association for Computational Linguistics.

Chong, Y. L., Ch'ng, E., Liu, M. J., & Li, B. (2021). The use of artificial intelligence in retail: A systematic literature review. *Journal of Retailing and Consumer Services*, 61, 102505. https://doi.org/10.1016/j.jretconser.2021.102505

Deshmukh, R., Raut, R., Bhavsar, M., Gurav, S., & Patil, Y. (2025). Optimizing human-AI interaction: Innovations in prompt engineering. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 1240–1246). IEEE. https://doi.org/10.1109/IDCIOT64235.2025.10914815

Grewal, D., Roggeveen, A. L., & Nordfält, J. (2023). The future of retailing. *Journal of Retailing*, 99(1), 1–13. https://doi.org/10.1016/j.jretai.2022.12.003

Hamzic, D., Wurzenberger, M., Skopik, F., Landauer, M., & Rauber, A. (2024). Evaluation and comparison of open-source LLMs using natural language generation quality metrics. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 5342–5351). IEEE. https://doi.org/10.1109/BigData62323.2024.10825576

Inturi, S., Potluri, S. L., Vanitha, G., & Indira, B. (2025). VirtualMe: A large language model-powered chatbot for enhanced human–computer interaction. In *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)* (pp. 626–634). IEEE. https://doi.org/10.1109/ICMLAS64557.2025.10967609

Irugalbandara, C., Perera, S., Wijesundara, D., Perera, R., & Liyanage, L. (2024). Scaling down to scale up: A cost–benefit analysis of replacing OpenAI's LLM with open-source SLMs in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (pp. 280–291). IEEE. https://doi.org/10.1109/ISPASS61541.2024.00034

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kumar, S., Verma, R., & Luthra, S. (2024). AI applications in retail: Opportunities, challenges, and future research directions. *Technological Forecasting and Social Change, 196*, 122800. https://doi.org/10.1016/j.techfore.2024.122800

Liang, P., Bommasani, R., Creel, K. A., & Reich, S. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. https://crfm.stanford.edu/helm/0.1.0

Linzbach, P., Schütte, R., & Alt, R. (2019). Artificial intelligence in retailing: Review and research agenda. *Journal of Retailing and Consumer Services, 54*, 102053. https://doi.org/10.1016/j.jretconser.2019.102053

Magnini, M., Aguzzi, G., & Montagna, S. (2025). Open-source small language models for personal medical assistant chatbots. *Intelligence-Based Medicine, 11*, 100097. https://doi.org/10.1016/j.ibmed.2025.100097

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*.

Örpek, Z., Tural, B., & Destan, Z. (2024). The language model revolution: LLM and SLM analysis. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1–4). IEEE. https://doi.org/10.1109/IDAP64064.2024.10710677

Pawar, V., & G, M. (2024). Exploring the potential of prompt engineering: A comprehensive analysis of interacting with large language models. In *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. IEEE.

Rese, A., Schreiber, S., & Baier, D. (2020). Technology acceptance of voice assistants in retailing: Conceptual model and empirical analysis. *Journal of Retailing and Consumer Services, 55*, 102753. https://doi.org/10.1016/j.jretconser.2020.102753

Shaikh, A., Iqbal, R., & Shah, S. A. (2019). Natural language processing-based chatbot for smart customer services: A case study. *International Journal of Advanced Computer Science and Applications, 10*(9), 379–385. https://doi.org/10.14569/IJACSA.2019.0100949

Sun, W., Zhao, Y., Liu, W., Liu, Y., Yang, R., & Han, C. (2022). Internet of things enabled the control and optimization of supply chain cost for unmanned convenience stores. *Alexandria Engineering Journal, 61*(11), 9149–9157. https://doi.org/10.1016/j.aej.2022.02.043

Tatikonda, S., Qiu, M., & Chen, Y. (2025). Evaluating conversational agents in customer-facing retail applications. *Information Systems Frontiers, 27*(2), 477–492. https://doi.org/10.1007/s10796-024-10391-z

Wang, J., Zhang, R., & Li, M. (2024). Context-aware intelligent agents for physical retail navigation. *AI in Retail Journal*, 5(1), 21–38. https://doi.org/10.1016/j.airj.2024.01.003

Zou, Y., Zhou, W., & Wang, Y. (2023). Deploying AI chatbots for retail inventory and customer service. *Computers in Industry, 150*, 103765. https://doi.org/10.1016/j.compind.2023.103765