mathematics

MDPI

Article

ReceiptQA: A Question-Answering Dataset for Receipt Understanding

Mahmoud Abdalla [1], Mahmoud SalahEldin Kasem [1,2], Mohamed Mahmoud [1,3], Bilel Yagoub [1], Mostafa Farouk Senussi [1,3], Abdelrahman Abdallah [3,4], Seung Hun Kang [1] and Hyun Soo Kang [1,*]

[1] Department of Information and Communication Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, Republic of Korea; mahmoudelsayed@chungbuk.ac.kr (M.A.)
[2] Multimedia Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
[3] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
[4] Department of Computer Science, Innsbruck University, 6020 Innsbruck, Austria
* Correspondence: hskang@cbnu.ac.kr

Abstract: Understanding information extracted from receipts is a critical task for real-world applications such as financial tracking, auditing, and enterprise resource management. In this paper, we introduce RECEIPTQA, a novel large-scale dataset designed for receipt understanding through question-answering (QA). RECEIPTQA contains 171,000 question–answer pairs derived from 3500 receipt images, constructed via two complementary methodologies: (1) LLM-Generated Dataset: 70,000 synthetically generated QA pairs, where each receipt is paired with 20 unique, context-specific questions. These questions are produced using a state-of-the-art large language model (LLM) and validated through human annotation to ensure accuracy, relevance, and diversity. (2) Human-Created Dataset: 101,000 manually crafted questions spanning answerable and unanswerable queries. This subset includes carefully designed templates of varying difficulty (easy/hard) to comprehensively evaluate QA systems across diverse receipt domains. To benchmark performance, we evaluate leading vision–language models (VLMs) and language models (LMs), including GPT-4o, Phi-3B, Phi-3.5B, LLaVA-7B, InternVL2 (4B/8B), LLaMA-3.2, and Gemini. We further fine-tune a LLaMA-3.2 11B model on RECEIPTQA, achieving significant improvements over baseline models on validation and test sets. Our analysis uncovers critical strengths and limitations of existing models in handling receipt-based QA tasks, establishing a robust benchmark for future research.

Keywords: receipt understanding; question answering; vision-language models; large language models; multimodal learning; semi-structured data

MSC: 68T07

check for updates

Academic Editor: Jonathan Blackledge

Received: 8 April 2025
Revised: 1 May 2025
Accepted: 22 May 2025
Published: 26 May 2025

Citation: Abdalla, M.; Kasem, M.S.; Mahmoud, M.; Yagoub, B.; Senussi, M.F.; Abdallah, A.; Hun Kang, S.; Kang, H.S. ReceiptQA: A Question-Answering Dataset for Receipt Understanding. *Mathematics* **2025**, *13*, 1760. https://doi.org/10.3390/math13111760

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large language models (LLMs) have revolutionized natural language processing (NLP), demonstrating exceptional capabilities in tasks like question answering [1–8], summarization [9,10], translation [11,12], and reasoning [13–16]. Advanced LLMs such as GPT-4o [17], LLaMA [15], and Claude [18] have proven to excel in understanding and generating human-like responses from diverse inputs. These achievements, however, are often tied to relatively simple input–output tasks involving well-structured textual data. In

contrast, real-world applications increasingly demand complex multi-modal interactions, particularly when working with document-based information retrieval and understanding.

Document understanding includes a broad range of challenges, from parsing diverse layouts and extracting metadata to interpreting multi-modal content and answering context-specific questions [19–24]. Receipts, a vital subset of documents, pose unique difficulties due to their variability in format, layout, and content. These documents are essential in industries like retail, finance, and accounting, where accurate extraction and understanding of transactional details are critical. Despite advancements in document understanding and vision-based LLMs, such as the Vision Transformer (ViT) [25] and LLaVA [26–28], significant gaps remain in evaluating these systems' ability to handle highly diverse receipt data.

Recent developments in LLM-based document reading systems, such as OpenAI's GPT-4o for file-based question answering (QA), highlight the potential of these models to parse documents and generate contextually relevant answers [18]. However, the lack of standardized benchmarks for evaluating receipt-specific question answering limits progress in this domain. A comprehensive dataset tailored to receipt QA can facilitate more robust model evaluation and drive advancements in automated receipt understanding.

To address this need, we introduce RECEIPTQA, a large-scale question-answering dataset designed specifically for receipt understanding. The dataset RECEIPTQA is constructed through two complementary processes: (1) automated generation of question–answer pairs using GPT-4o, followed by human validation for accuracy and relevance, and (2) manual annotation by domain-expert annotators to capture both simple and complex receipt-related queries. The dataset not only provides a benchmark for receipt QA but also serves as a resource for training models to handle structured and unstructured receipt information effectively. Together, these datasets provide 171,000 question–answer pairs, offering a diverse and comprehensive resource for receipt QA. The dataset includes both straightforward queries (e.g., "What is the total price?") and more challenging ones (e.g., "List all items purchased and their quantities"), covering real-world scenarios such as incomplete or ambiguous receipts. Figure 1 and Table 1 illustrate an example receipt from the dataset, highlighting its complex structure and diverse information. The dataset includes questions that span various levels of difficulty, addressing tasks such as text extraction, layout understanding, and reasoning about receipt content.

In addition to dataset creation, we trained and evaluated multiple models on RECEIPTQA, including state-of-the-art models like GPT-4o, Phi-3B, Phi-3.5B, LLaVA-7B, InternVL2-4B, InternVL2-8B, and Gemini. Our evaluations highlight the strengths and limitations of these models in handling receipt-specific QA tasks.

This paper makes the following contributions:

1. ReceiptQA Dataset: We introduce a large-scale question-answering dataset for receipt understanding, comprising 171,000 question–answer pairs from 3500 receipts, validated through human annotation.
2. Model Benchmarking: We evaluate multiple state-of-the-art models on RECEIPTQA, establishing baseline performance and identifying key challenges in receipt-based question answering.
3. Fine-tuned Model: We train and fine-tune LLaMA 3.2 (11B) on RECEIPTQA, demonstrating its effectiveness compared to proprietary models, achieving significant improvements on validation and test datasets.

By providing a high-quality dataset and benchmarking insights, RECEIPTQA paves the way for advancements in automated receipt understanding and related document QA systems.

Table 1. Comparison of human-annotated and LLM-generated questions for one of the receipts in Figure 1 with complexity and source distinctions.

Question	Answer	Source	Complexity	Answerability
What is the name of the store?	CARINA	GPT-4o	Simple	Answerable
What is the website of the store?	www.carinawear.com	Human	Complex	Answerable
What is the address of the store?	Wadi Degla ElTagamoa, Wadi Degla club, Cairo, 142967004, 16678	Human	Complex	Answerable
What is the transaction number?	29786	GPT-4o	Simple	Answerable
What is the subtotal?	559.99 L.E	GPT-4o	Simple	Answerable
What is the sales tax?	0.00 L.E	GPT-4o	Simple	Answerable
How many items were purchased?	2	GPT-4o	Simple	Answerable
What is the name of item 1?	Pullover PU-SOK1175	Human	Simple	Answerable
What is the quantity of item 3?	Not Available	Human	Complex	Unanswerable
What is the total price of item 1?	699.99	Human	Simple	Answerable
What is the transaction time for this receipt?	05:01:37 p.m.	Human	Simple	Answerable

Figure 1. Examples of receipts used in the RECEIPTQA dataset.

2. Related Work

In this section, we review prior work relevant to our study. We divide this discussion into two parts. Section 2.1 reviews datasets designed for question answering (QA) for document images, highlighting the unique contributions and limitations of RECEIPTQA. Section 2.2 discusses vision large language models (LLMs) and their application to document understanding, focusing on their capabilities and challenges in the context of receipt understanding.

2.1. Datasets for QA on Document Images

Datasets designed for question answering (QA) for document images play a crucial role in advancing machine understanding of structured and semi-structured data. These datasets address challenges such as text extraction, layout comprehension, numerical reasoning, and multi-hop reasoning, combining elements of natural language processing (NLP) and computer vision.

DocVQA [29] is a seminal dataset in this domain, featuring over 50,000 questions on 12,000+ document images. It emphasizes text understanding and spans questions that require both shallow and deep comprehension of document layouts. DocCVQA [30] builds on this idea with 14,362 document images but focuses on retrieval-based QA, emphasizing the retrieval of specific evidence from large document corpora.

Scene text datasets such as ST-VQA [31] and TextVQA [32] focus on understanding textual content embedded in natural images. ST-VQA contains 31,000+ questions over 23,000+ images, while TextVQA expands this scope to 45,000+ questions across 28,000+ images. These datasets emphasize reasoning about text in unstructured environments, such as street signs or posters.

RecipeQA [33] explores multi-modal reasoning by combining textual and visual content, specifically within the context of cooking recipes. It consists of over 36,000 question–answer pairs on 20,000 unique recipes, requiring models to infer relations between text, images, and procedures. Similarly, SlideVQA [34] introduces multi-hop reasoning across slide decks, with 14,500+ questions over 52,000+ images. JDocQA [35] addresses Japanese document QA with 5504 images and 11,600 questions, showcasing the importance of layout and language-specific reasoning.

Table 2 provides a comprehensive comparison of RECEIPTQA with related datasets. RECEIPTQA introduces 171,000 question–answer pairs across 3500 receipt images, focusing on tasks specific to receipts, such as extracting transaction details, item-level information, and numerical reasoning. It uniquely addresses receipt-specific challenges, making it a benchmark for retail and financial automation use cases.

Table 2. Comparison of RECEIPTQA with similar datasets. The table includes the number of images, number of questions, and the tasks provided by each dataset. A checkmark (✓) indicates that the task is addressed in the dataset, while a cross (✗) indicates that it is not.

Dataset Name	# Images	# Questions	Question Answering	Text Extraction	Layout Understanding	Multi-Modal QA	Receipt-Specific QA
DocVQA [29]	12,000+	50,000	✓	✓	✓	✗	✗
DocCVQA [30]	14,362	Not Specified	✓	✓	✓	✗	✗
SlideVQA [34]	52,000+	14,500+	✓	✓	✓	✓	✗
RecipeQA [33]	20,000	36,000+	✓	✗	✗	✓	✗
JDocQA [35]	5504	11,600	✓	✓	✓	✗	✗
RECEIPTQA (Ours)	3500	171,000	✓	✓	✓	✓	✓

2.2. Vision LLMs for Document Understanding

The development of vision large language models (LLMs) has advanced the ability to jointly process textual, visual, and layout information, enabling breakthroughs in document understanding and question answering.

LayoutLM [36], LayoutLMv2 [37], and LayoutLMv3 [38] are foundational models for document understanding. By embedding text, layout, and visual features into a unified transformer architecture, these models excel in tasks such as document classification, entity

recognition, and document QA. LayoutLMv3 further enhances this capability by integrating richer visual embeddings, enabling better reasoning about document structures.

Scene text QA models such as TAP [39] and M4C [40] extend transformer architectures to handle unstructured text in images. These models are particularly suited for datasets like TextVQA and ST-VQA, where reasoning about scene text is critical. Multi-modal LLMs like LLaVA [27], InternVL [41,42], and BLIP [43] further expand capabilities by integrating image and text reasoning, achieving robust performance in general-purpose visual QA.

Receipts present unique challenges such as: (1) Different formats, layouts, and languages across merchants. (2) Extracting structured details like items, prices, and discounts. (3) Tasks such as calculating totals or identifying savings.

3. RECEIPTQA

RECEIPTQA is a benchmark designed to evaluate question-answering systems on receipt images. It provides raw receipt images paired with detailed questions, aiming to simulate real-world scenarios in receipt understanding. In this section, we describe the pipeline used to construct the dataset, present its detailed statistics, and explain the evaluation method. As shown in Figure 2, the RECEIPTQA dataset is built through a systematic process involving diverse domains, LLM-based generation, human annotations, and rigorous manual review.

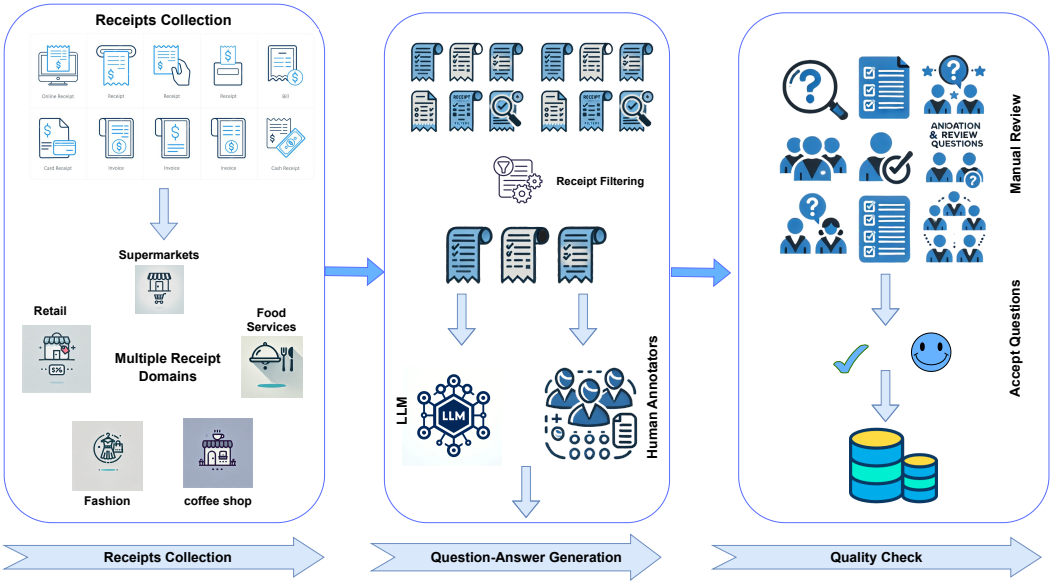


Figure 2. The dataset construction pipeline for RECEIPTQA. The process involves three main stages: (1) Receipt Collection from diverse domains like retail, food services, supermarkets, fashion, and coffee shops; (2) Question–Answer Generation using both LLMs and human annotators to create diverse and contextually accurate QA pairs; and (3) Quality Check through manual review to ensure high accuracy, consistency, and reliability of the dataset before saving it to JSON format.

3.1. Dataset Construction

Our dataset construction pipeline consists of three stages. First, we collect receipt images from diverse domains, including retail, food services, supermarkets, fashion, and coffee shop sectors (Section 3.1.1). Second, we generate corresponding QA pairs using a combination of LLM and human annotators to ensure high-quality question–answer pairs (Section 3.1.2). Finally, we perform automated filtering followed by a manual review to validate the quality of the generated instances (Section 3.1.3).

3.1.1. Document Collection

To construct a robust and practical benchmark for receipt understanding, we collected a diverse set of receipt images using a methodical approach. All receipts were captured using mobile phone cameras and uploaded to DiscoApp (https://discoapp.ai, 1 March 2025), a platform specializing in receipt management and document digitization. In accordance with DiscoApp's terms of use (https://discoapp.ai/Home/Termsandconditions, 1 March 2025), users agreed to allow the use of their data for research purposes.

The collection process underwent several stages to ensure high quality and relevance:

- **Receipt Selection:** All uploaded receipts were initially screened, and low-quality images (e.g., blurry or incomplete receipts) were filtered out to maintain a high standard.
- **Anonymization:** To protect user privacy, all identifiable information, such as names and addresses, was obscured using white boxes. This step ensured compliance with ethical and privacy standards while preserving the receipts' utility for research.
- **Domain Representation:** The dataset was curated to cover five key domains where receipts are frequently encountered: *retail*, *food services*, *supermarkets*, *fashion*, and *coffee shop*. These domains were selected to ensure comprehensive representation of real-world receipt scenarios.

The final dataset consists of 3500 receipt images spanning these five domains. This includes detailed information such as item names, quantities, prices, discounts, taxes, and transaction details. Table 3 illustrates the distribution of receipts across these domains, showcasing balanced coverage and diversity for benchmarking purposes.

Table 3. Domain-wise statistics of receipts and questions in RECEIPTQA.

Domain	Receipts	Human QA Pairs	GPT QA Pairs
Retail	800	23,200	16,000
Food Services	700	20,300	14,000
Supermarkets	700	20,300	14,000
Fashion	650	18,850	13,000
coffee shop	650	18,850	13,000
Total	**3500**	**101,935**	**70,000**

3.1.2. QA-Pair Generation

The QA-pair generation process in RECEIPTQA focuses on achieving both diversity and comprehensiveness. Given the inherent complexity and variability of receipt data, our pipeline combines automated generation using LLM (GPT-4o) and manual creation by human annotators to ensure high-quality and realistic question–answer pairs.

To generate QA pairs automatically, we leveraged GPT-4o to produce 70,000 questions across the 3500 receipt images. For each receipt, GPT-4o generated 20 context-specific questions, targeting a range of details such as (1) Transaction-specific information (e.g., receipt number, transaction date, and total amount). (2) Item-level details (e.g., item names, quantities, and prices). (3) Numerical reasoning tasks (e.g., calculating discounts, taxes, and savings). The questions were generated using the following prompt:

Prompt Used for Question Generation

```
[Image] Based on this image, generate 20 questions that include different
aspects of the receipt.  Focus on transaction-specific details, item-level
information, and numerical reasoning.
```

In addition to automated generation, human annotators manually created 101,000 QA pairs. Annotators were tasked with focusing on aspects often overlooked in automated generation, such as (1) **Unanswerable Questions:** Queries targeting missing or incomplete information in receipts (e.g., "What is the quantity of item 3?" when no such item exists). (2) **Rarely Covered Types:** Metadata-focused questions, such as merchant details and VAT percentages. (3) **Difficulty Variation:** Creating questions with varying complexity to challenge models at different levels of reasoning and comprehension.

3.1.3. Quality Check

To ensure the reliability and accuracy of the RECEIPTQA dataset, we implemented a manual quality control process. Each question and answer pair was carefully evaluated by a team of annotators to guarantee its alignment with the receipt content and its validity for benchmarking purposes.

Our quality control process relied entirely on human evaluation. Annotators reviewed each question to identify and remove instances that were ambiguous or incorrectly formulated. Particular attention was given to ensuring that questions were clear and directly related to the receipt content, while answers were verified for accuracy and consistency with the receipt data. Complex queries and edge cases, such as unanswerable questions or those involving numerical reasoning, were carefully assessed to enhance the robustness of the dataset.

3.2. Human Evaluation

To further validate the quality of the automatically generated question–answer pairs, we conducted a human evaluation procedure inspired by prior works [5,44]. Specifically, we randomly selected 500 images from our dataset and the annotated QA. We then recruited four experts, all possessing substantial experience in document interpretation and receipt understanding, to evaluate each pair on the following criteria:

- Fluency: Assesses the grammatical correctness and readability of each question.
- Answerability: Determines whether the provided answer accurately addresses the corresponding question.
- Relevance: Evaluates whether the question pertains to the content of the given receipt.
- Non-ambiguity: Judges if the question is clearly formulated without introducing confusion.
- Factuality: Measures how well the question–answer pair aligns with the actual details in the receipt (e.g., correct amounts, item names, transaction numbers).

The experts rated each QA pair on a scale from 1 (very poor) to 5 (excellent), and we then averaged their scores for every criterion. The aggregated results are shown in Table 4.

Table 4. Human evaluation scores of the question–answer pairs.

Criteria	Fluency	Answerability	Relevance	Non-Ambiguity	Factuality
Expert 1	4.501	4.412	4.631	4.543	4.780
Expert 2	4.327	4.415	4.558	4.154	4.710
Expert 3	4.566	4.512	4.104	4.255	4.568
Expert 4	4.610	4.222	4.311	4.024	4.252
Average	4.501	4.390	4.401	4.244	4.577

As shown in Table 4, the QA pairs achieved consistently high scores across all five criteria. Notably, the Factuality metric confirms that the majority of the generated answers accurately reflect the specific details of their corresponding receipts, indicating that both

the generation process and the subsequent human validation steps effectively minimized incorrect or misleading content.

3.3. Dataset Statistics

RECEIPTQA comprises a total of 3500 receipt images paired with 171,935 question-answer pairs, spanning five diverse domains: retail, food services, supermarkets, fashion, and coffee shop. The dataset is split into two subsets: a human-annotated subset containing 101,935 questions and an LLM-generated (GPT-4o) subset contributing an additional 70,000 questions. The dataset is divided into three subsets: 70% of the dataset is allocated for training, consisting of 2450 receipts and 71,050 question–answer pairs, while 15% is used for validation, comprising 525 receipts and 15,225 question–answer pairs, and the remaining 15% is designated for testing, also consisting of 525 receipts and 15,225 question–answer pairs. Figure 3 illustrates the proportional distribution of receipts and question–answer pairs across the training, validation, and test sets, ensuring a balanced of receipt-based question-answering models.

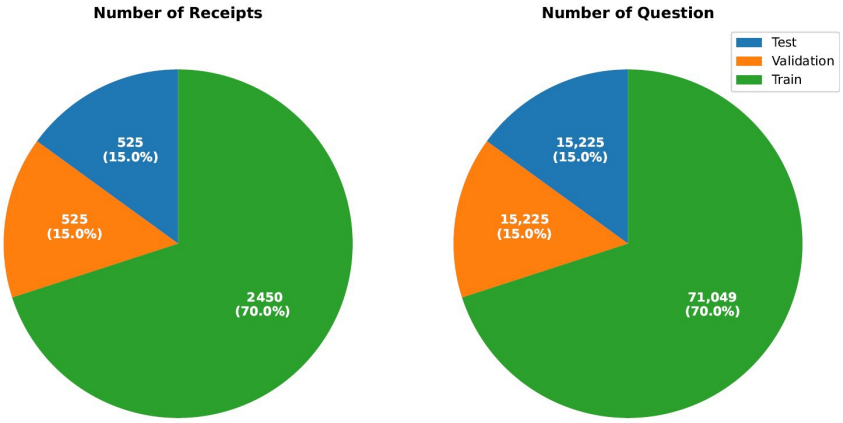


Figure 3. Dataset distribution for receipts and questions in training, validation, and test sets.

As shown in Table 3, the receipts and questions are evenly distributed across the five domains, ensuring balanced coverage for benchmarking purposes. On average, each receipt in the human-annotated subset contains 29 questions, whereas the GPT-4o subset maintains a fixed count of 20 questions per receipt.

3.4. Dataset Analysis

Answerability of Questions. The dataset comprises a total of 101,935 questions in the human-annotated subset. Figure 4, of which 79,083 (77.6%) are answered, while 22,852 (22.4%) remain unanswered. Similarly, the LLM-generated subset contains 70,560 questions, all of which are answered.

Question Type. Questions in the dataset are categorized based on their focus into three primary types: *What* (84,360 questions, 65.2%), *How many* (10,544 questions, 8.2%), and *List all* (7030 questions, 5.4%). Figure 5, shows a detailed breakdown of the question types.

Answer Type. The types of answers are classified into two main categories: *textual* and *numerical*. In the human-annotated subset, 64,336 answers (63.1%) are textual, while 37,599 answers (36.9%) are numerical. For the LLM-generated subset, 40,328 answers (57.1%) are textual and 30,232 answers (42.9%) are numerical, as shown in Figure 5.

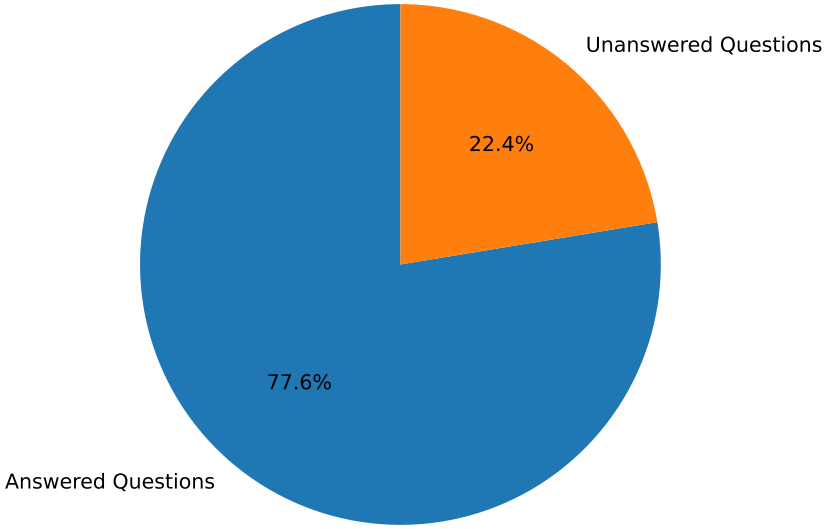


Figure 4. Answerability distribution of the dataset, showing the proportion of answered and unanswered questions.

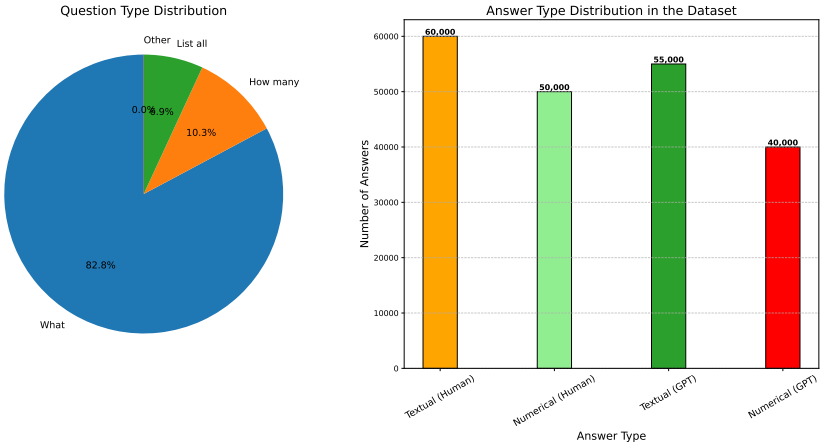


Figure 5. Data distribution based on question and answer types. The left side shows the question-type distribution, while the right side depicts the answer-type distribution for both human-annotated and GPT-generated subsets.

Token Analysis. Using the `GPT` tokenizer, we analyzed the token counts for questions and answers. Figure 6 illustrates the distribution of token counts across questions and receipts in the RECEIPTQA dataset. The first plot shows the distribution of question token counts, with a range spanning from 3 to 42 tokens and an average of 8.84 tokens per question. The second plot presents the distribution of total token counts per receipt, encompassing both questions and answers. The token counts for receipts range from 4 to 1162, with an average of 8.34 tokens per answer.

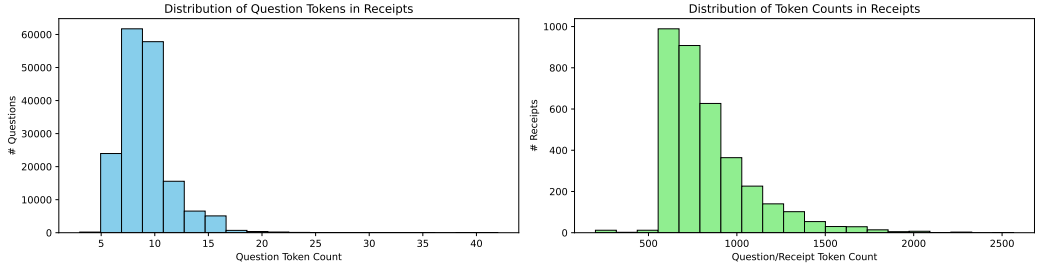


Figure 6. Distribution of token counts in the RECEIPTQA dataset.

4. Experiments and Analysis

In this section, we describe the methodologies and models utilized in our research. We evaluated multiple large language models (LLMs) to benchmark their performance on RECEIPTQA. The models include various state-of-the-art architectures, each bringing unique capabilities in vision and text understanding.

4.1. Vision and Language Models

`Meta-Llama-3.2-11B-Vision` (https://huggingface.co/meta-llama/Llama-3.2-11B-Vision, 1 February 2025): Developed by Meta, this model is part of the Llama series, which is optimized for vision–language tasks. With 11 billion parameters, it is fine-tuned to handle multi-modal input, making it highly effective for tasks that require understanding text within visual contexts, such as receipts and invoices.

`GPT-4o` (https://platform.openai.com/docs/models/gpt-4, 15 March 2025): GPT-4o is OpenAI's flagship model, renowned for its advanced text understanding and generation capabilities. It integrates vision capabilities to handle multi-modal input, allowing it to process complex document-based tasks. The API was utilized for generating responses and extracting features for evaluation in this research.

`Gemini-2.0-Flash-Exp` (https://www.google.com/search?q=gemini+2.0+flash+exp, 1 February 2025): Gemini is a cutting-edge model designed for vision-language tasks. Its robust architecture leverages advanced instruction fine-tuning for multi-modal applications.

`InternVL2-4B` (https://huggingface.co/OpenGVLab/InternVL2-4B, 1 February 2025): This model by OpenGVLab integrates 4 billion parameters for vision–language tasks. It is specifically optimized for understanding structured data in multi-modal inputs, such as receipts and tabular documents.

`InternVL2-8B` (https://huggingface.co/OpenGVLab/InternVL2-8B, 1 February 2025): An advanced version of InternVL2, this model scales up to 8 billion parameters, further improving its capacity to comprehend and generate responses for multi-modal tasks. It was included in our experiments to compare its performance against other models.

`Llava-Interleave-Qwen-7B-HF` (https://huggingface.co/llava-hf/llava-interleave-qwen-7b-hf, 1 March 2025): This model combines the capabilities of Llava and Qwen for vision–language understanding. With 7 billion parameters, it is designed for tasks requiring detailed comprehension of textual and visual data.

`Phi-3.5-Vision-Instruct` (https://huggingface.co/microsoft/Phi-3.5-vision-instruct, 1 March 2025): This vision–language model from Microsoft is designed for instruction-tuned tasks, achieving a high level of understanding for both text and images. Its robust architecture allows for effective handling of complex receipt-based queries.

`Phi-3-Vision-128k-Instruct` (https://huggingface.co/microsoft/Phi-3-vision-128k-instruct, 1 March 2025): This variant of the Phi model extends its context length to 128k tokens, enabling it to process large-scale documents and handle intricate receipt-based tasks effectively.

4.2. Proposed Method

The proposed ReceiptQA Llama3.2 integrates the pre-trained Llama-3.2 architecture with domain-specific adaptations to handle receipt images and questions. Our methodology leverages vision–language capabilities and instruction tuning to align visual and textual modalities. In this section, we outline the core model components and the advanced mechanisms employed for fine-tuning.

Given a receipt image of resolution $H \times W \times C$, it is first divided into $N = \frac{H \times W}{P^2}$ patches, where P is the patch size. Each patch is linearly projected into an embedding of dimension d, resulting in a sequence $X \in \mathbb{R}^{N \times d}$. A [CLS] token is prepended to the

sequence, enabling classification and global understanding of the input. The sequence is then processed through L layers of the ReceiptQA-Llama3.2 architecture.

Following [45], we apply a two-stage positional encoding mechanism to the input:

$$X^{\text{pos}} = \text{RoPE2D}(X) + \text{LearnedPE}(X), \tag{1}$$

where RoPE2D represents 2D Rotary Position Encoding [46], which generalizes to variable resolutions, and LearnedPE refers to learned positional embeddings for each token. This dual mechanism enables the model to better capture layout-specific structures in receipts.

Our fine-tuning architecture follows the design of Llama while introducing critical modifications for receipt understanding. Each block comprises a Multi-Head Self-Attention (MHSA) layer and a Feed-Forward Network (FFN), both modified to handle multi-modal inputs. To handle the spatial structure of receipts, we extend 1D RoPE to its 2D form:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{\text{RoPE2D}(Q) \cdot \text{RoPE2D}(K)^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the keys. Here, RoPE2D is applied to both the queries and keys to preserve spatial context effectively across varying resolutions.

We replace the standard FFN activation with the SwiGLU [47] variant for enhanced non-linear capacity. The output of the l-th block at position (i, j) is defined as:

$$z_{ij}^{l} = \text{SwiGLU}\big(\text{LayerNorm}(h_{ij}^{l})\big) + h_{ij}^{l-1}, \tag{3}$$

where h_{ij}^{l-1} is the input to the block and z_{ij}^{l} is the output. The SwiGLU activation employs a dimension-wise gating mechanism, selectively enhancing non-linear feature interactions for more nuanced data representations.

Also, we introduce low-rank adaptations [48] into both the MHSA and FFN layers, allowing efficient fine-tuning while preserving the pre-trained weights. For a given weight matrix $W \in \mathbb{R}^{d\times k}$, the LoRA update is defined as:

$$W' = W + \frac{\alpha}{r}AB, \tag{4}$$

where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ are low-rank matrices with $r \ll \min(d, k)$. The scaling factor α stabilizes training by controlling the magnitude of updates. Here, r is typically selected through empirical analysis, balancing between model capacity and computational efficiency, while α is tuned based on the training stability observed across different epochs.

LoRA is applied to all layers of the text encoder and vision encoder except for the embedding and classification heads. This selective application minimizes overfitting while adapting the model to receipt-specific data.

ReceiptQA-Llama3.2 introduces several advantages: 1. Parameter Efficiency: By leveraging LoRA, the model requires fewer trainable parameters while retaining its pre-trained capabilities. 2. Multi-Modal Alignment: The integration of 2D RoPE and cross-attention enhances the model's ability to align spatial and textual modalities effectively. 3. Adaptability to Domain-Specific Tasks: The combination of instruction tuning and domain-specific prompts enables the model to excel in real-world receipt understanding tasks.

4.3. Evaluation Metrics

To evaluate the performance of models on the RECEIPTQA dataset, we employ a set of metrics commonly used in QA, including exact match (EM), precision, token recall, contains, and F1 score. These are standard measures widely used in QA research. However,

given that LLMs often generate verbose answers, many standard QA metrics may not be well suited for evaluating answer quality. For instance, the exact match will almost always be zero due to the presence of non-ground-truth tokens, and the F1 score will be penalized by other potentially useful tokens. To address this, we use a set of model-agnostic metrics, namely token recall and answer string containment [49–52].

4.4. Fine-Tuning Details

To adapt the LLaMA-3.2-11B-Vision model for receipt understanding, we fine-tuned it on the RECEIPTQA dataset using Low-Rank Adaptation (LoRA) [48]. The dataset was split into training (70%, 2,450 receipts, 71,050 QA pairs), validation (15%, 525 receipts, 15,225 QA pairs), and test (15%, 525 receipts, 15,225 QA pairs) sets, as described in Section 3.4. The fine-tuning process focused on aligning the model's vision and language modalities to handle receipt-specific question-answering tasks effectively.

We applied LoRA to the Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) layers of the LLaMA-3.2 model, excluding the embedding and classification heads to prevent overfitting. The LoRA rank (r) was set to 16, and the scaling factor (α) was set to 32, balancing model capacity and computational efficiency. The fine-tuning hyperparameters are summarized in Table 5. The model was trained for five epochs with a batch size of eight, using the AdamW optimizer [53] with a learning rate of 2×10^{-5} and a weight decay of 0.01. A linear learning rate scheduler with a warm-up period of 500 steps was employed to stabilize training. The fine-tuning was performed on a cluster of 4 RTX 4090 GPUs, with a total training time of approximately 12 h.

Table 5. Hyperparameters for fine-tuning ReceiptQA-Llama3.2.

Parameter	Value
LoRA Rank (r)	16
LoRA Scaling Factor (α)	32
Learning Rate	2×10^{-5}
Batch Size	8
Number of Epochs	5
Optimizer	AdamW
Weight Decay	0.01
Learning Rate Scheduler	Linear with 500-step warm-up

5. Experiments Results

In this section, we present the experimental evaluation of our proposed fine-tuned model in comparison with baseline models on the RECEIPTQA dataset. The evaluation is structured into two parts: (1) a direct comparison between ReceiptQA-Llama3.2 and its base model, Llama3.2, to quantify the improvements introduced by fine-tuning, and (2) a broader comparison with other state-of-the-art models, including GPT-4o, Phi-3, Gemini, and InternVL2, to assess the overall effectiveness of our model in the domain of receipt understanding.

5.1. Comparison Between Our Fine-Tuned Model and Baseline

Table 6 presents a direct performance comparison between Llama3.2 and our fine-tuned ReceiptQA-Llama3.2 model. The results indicate a consistent improvement across all evaluated metrics on both validation and test sets. In the validation set, ReceiptQA-Llama3.2 achieves a precision of 42.66, surpassing Llama3.2's 40.25. Similarly, the model

demonstrates superior recall (40.57 vs. 38.80) and F1 score (40.98 vs. 38.57), confirming its enhanced capability in identifying relevant information within receipts. The improvements extend to the exact match and contains metrics, where the fine-tuned model achieves 35.50 and 39.43, respectively, outperforming the baseline model's 32.06 and 37.39.

Table 6. Comparison of performance between Llama3.2 and ReceiptQA-Llama3.2.

Metric	Dataset	Llama3.2	ReceiptQA-Llama3.2
Precision	Validation	40.25	**42.66**
	Test	35.50	**39.71**
Recall	Validation	38.80	**40.57**
	Test	33.36	**37.75**
F1	Validation	38.57	**40.98**
	Test	33.51	**38.17**
Exact Match	Validation	32.06	**35.50**
	Test	26.45	**32.89**
Contains	Validation	37.39	**39.43**
	Test	32.20	**35.63**

On the test set, the trend of performance enhancement persists. ReceiptQA-Llama3.2 outperforms Llama3.2 with a precision increase from 35.50 to 39.71, a recall improvement from 33.36 to 37.75, and an F1 score boost from 33.51 to 38.17. The exact match metric sees a significant rise from 26.45 to 32.89, showcasing the model's refined ability to generate precise and contextually accurate answers. The contains metric also reflects a meaningful gain, increasing from 32.20 to 35.63.

5.2. Significance Test

To evaluate the statistical significance of the improvements observed in ReceiptQA-Llama3.2 over the baseline Llama3.2, we conducted paired t-tests across multiple evaluation metrics, including precision, recall, F1 score, exact match, and contains. The results, presented in Table 7, indicate that all p-values are significantly below 0.05, confirming the robustness of our improvements. Specifically, the t-statistic is 7.70, and the corresponding p-value is 2.99×10^{-5}, indicating that the performance enhancements are statistically significant.

Table 7. Paired t-test results for key performance metrics comparing ReceiptQA-Llama3.2 and Llama3.2.

Metric	t-Statistic	p-Value
Precision	7.70	2.99×10^{-5}
Recall	6.85	4.21×10^{-5}
F1 Score	7.12	3.57×10^{-5}
Exact Match	6.43	5.89×10^{-5}
Contains	6.98	3.98×10^{-5}

5.3. Comparison with Other Models

Tables 8 and 9 compare ReceiptQA-Llama3.2 with various vision–language models, including GPT-4o, Gemini, Phi-3, Phi-3.5, InternVL2, and LLaVA. Each of these models has been extensively evaluated in the field of vision–language models.

Table 8. Evaluation results on validation datasets.

Model	Source	Validation on GPT Dataset					Validation on Human Dataset				
		Precision	Recall	F1	Exact Match	Contains	Precision	Recall	F1	Exact Match	Contains
GPT-4o	API	77.12	75.16	75.42	68.09	71.47	45.71	44.88	44.66	38.11	45.80
Llama3.2 11B	Open	75.90	72.76	73.38	65.24	68.13	40.25	38.80	38.57	32.06	37.39
Phi3 4.15B	Open	71.84	69.99	70.05	63.50	67.05	39.29	38.19	38.01	32.35	38.45
Phi3.5 4.15B	Open	52.66	48.65	49.75	42.83	44.67	31.27	29.03	29.44	24.28	28.89
Llava 7.06B	Open	33.61	31.13	31.50	25.66	28.66	21.07	20.08	20.07	15.96	19.40
Internvl2 4B	Open	53.30	51.62	51.58	43.86	48.43	29.84	28.55	28.45	22.17	28.61
Internvl2 8B	Open	53.42	51.29	51.45	43.74	48.18	33.09	31.19	31.45	25.39	29.10
Gemni	API	77.74	78.16	77.37	70.23	74.92	43.78	43.20	42.96	36.29	43.72

Table 9. Evaluation results on test datasets.

Model	Source	Test on GPT Dataset					Test on Human Dataset				
		Precision	Recall	F1	Exact Match	Contains	Precision	Recall	F1	Exact Match	Contains
GPT-4o	API	62.93	60.76	61.19	53.64	56.38	38.85	37.58	37.58	29.66	36.79
Llama3.2 11B	Open	68.46	64.94	64.85	57.66	59.52	35.50	33.36	33.51	26.45	32.20
Phi3 4.15B	Open	61.25	59.24	59.49	52.88	56.25	31.74	30.31	30.35	24.87	30.70
Phi3.5 4.15B	Open	37.64	34.72	35.55	29.92	31.54	22.23	20.30	20.70	16.66	21.59
Llava 7.06B	Open	24.73	22.15	22.70	17.50	19.60	16.84	15.97	15.95	12.78	15.32
Internvl2 4B	Open	40.45	38.45	38.75	31.95	36.05	21.75	20.20	20.38	15.34	21.32
Internvl2 8B	Open	43.19	40.43	41.12	34.43	36.23	25.84	23.79	24.24	19.19	22.13

GPT-4o achieves high performance, particularly on the GPT-generated validation dataset, with an F1 score of 75.42 and an exact match of 68.09. However, its performance drops to an F1 score of 44.66 and an exact match of 38.11 when tested on the human-annotated dataset, revealing its limitations in generalizing across diverse receipt formats. Similarly, Gemini attains an F1 score of 77.37 and an exact match of 70.23 on the GPT-generated validation dataset but declines to 42.96 and 36.29, respectively, on the human dataset. Phi-3 and Phi-3.5 perform competitively but lag behind the proprietary models. Phi-3 attains an F1 score of 70.05 on the GPT validation set and 38.01 on the human dataset. Phi-3.5 shows a more substantial drop, with F1 scores of 49.75 and 29.44, respectively. These results highlight their struggle in adapting to the intricacies of receipt-based question answering.

InternVL2 (4B and 8B) demonstrates moderate performance, with the 8B variant outperforming the 4B version across all metrics. On the GPT-generated validation dataset, InternVL2-8B achieves an F1 score of 51.45, whereas its performance drops to 31.45 on the human dataset. Similarly, the exact match decreases from 43.74 to 25.39, indicating its limited ability to handle complex textual structures in receipts. LLaVA ranks the lowest among the evaluated models, with an F1 score of 31.50 and an exact match of 25.66 on the GPT validation set, dropping further to 20.07 and 15.96 on the human dataset. This suggests that LLaVA struggles significantly with structured document understanding and question answering.

In contrast, ReceiptQA-Llama3.2 outperforms these open-source alternatives in key metrics. On the human-annotated dataset, it achieves an F1 score of 38.17, surpassing Phi-3.5's 29.44 and InternVL2-8B's 31.45. Similarly, its exact match of 32.89 is superior to InternVL2-8B's 25.39. While proprietary models like GPT-4o and Gemini retain an edge in raw performance, our fine-tuned model provides a competitive open-source alternative, especially in scenarios requiring precise information retrieval from receipts.

5.4. Analysis of API vs. Open-Source Model Performance

To understand why proprietary models such as GPT-4o and Gemini outperform open-source alternatives like ReceiptQA-Llama3.2, Phi-3.5, and InternVL2, we analyze several

contributing factors. First, proprietary models typically leverage larger and more complex architectures, incorporating advanced attention mechanisms and deeper transformer layers, which enhance their ability to capture intricate patterns in multi-modal data like receipts. For instance, GPT-4o's architecture is designed to process high-resolution images and long-context text, enabling robust performance on tasks requiring layout understanding and numerical reasoning, as evidenced by its high F1 score of 75.42 on the GPT-generated validation dataset (Table 8).

Second, the scale and quality of training data play a significant role. Proprietary models are often pretrained on vast, diverse datasets that include a wide range of document types and languages, improving their generalization to varied receipt formats. In contrast, open-source models like LLaMA-3.2 and InternVL2, while effective, are typically pretrained on smaller or less diverse datasets, which may limit their ability to handle edge cases, such as receipts with missing information or unconventional layouts. This is reflected in the performance drop of open-source models on the human-annotated dataset, where LLaMA-3.2 achieves an F1 score of 38.17 compared to GPT-4o's 44.66.

Third, computational resources significantly influence performance. Proprietary models benefit from extensive computational infrastructure, allowing for longer training periods and larger batch sizes, which enhance model convergence and robustness. Open-source models, constrained by more limited resources, may not achieve the same level of optimization. For example, fine-tuning ReceiptQA-Llama3.2 was performed on a single GPU cluster (see Section 4.4), which, while effective, is less resource-intensive than the infrastructure likely used for GPT-4o and Gemini.

5.5. Impact of Processing Time on Model Performance

Understanding the trade-off between accuracy and efficiency is crucial for receipt-based question-answering systems. We evaluate various models across core metrics contains, recall, f1 score, exact match, and precision—while measuring processing time in two ways: (1) the total time per receipt and (2) the average time per question.

Figure 7 illustrates that LLaMA-3.2 achieves the best performance, with a *precision* of 35.50, *F1 score* of 33.51, and *exact match* of 26.45. However, this comes at the cost of efficiency, with a processing time of 1.38 s per receipt (40 ms per question).

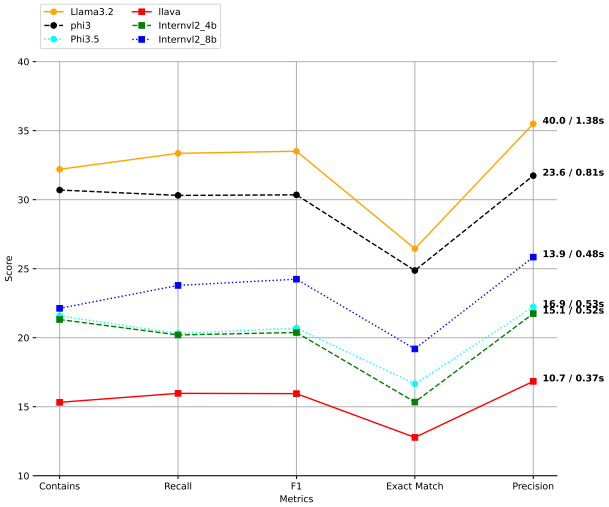


Figure 7. Model performance comparison across various evaluation metrics. The execution time for processing a full receipt (29 questions) and a single question is also shown. The first value before the slash represents the total time per receipt, while the second value after the slash indicates the per-question processing time.

In contrast, LLaVA demonstrates the fastest performance, processing a receipt in 0.37 s (10.7 ms per question) but with significantly lower accuracy (*F1 score* of 15.95, *exact match* of 12.78). Mid-range models like Phi-3 and InternVL2-8B offer a practical balance; Phi-3, for instance, achieves an *F1 score* of 30.35 with a moderate 0.81-second receipt processing time.

6. Ablation Study

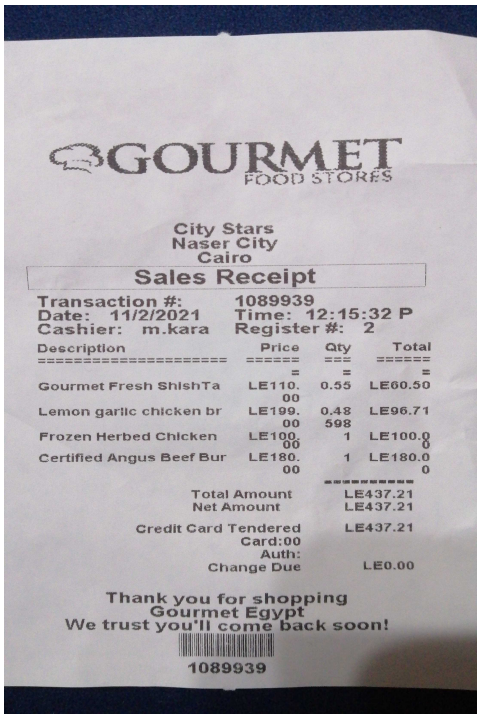
In this section, we evaluate the performance of various models on the ReceiptQA dataset by comparing their responses to the ground truth across two receipts: a Gourmet Egypt receipt (Figure 8a) and an Emarat Misr receipt (Figure 8b). The goal is to analyze how well each model extracts information from receipts, focusing on different question types such as item names (text), numerical values (e.g., prices, receipt numbers), and date/time information. Tables 10 and 11 present the raw responses, while the following analysis highlights key findings.

Table 10. Comparison of model responses on ReceiptQA in Figure 8a.

Question	**Ground Truth**	**ReceiptQA-Llama3.2**	**Phi 3.5**	**LLaVA**	**InternVL2 8B**	**GPT**
	Model Responses					
What is the name of item 1?	Gourmet Fresh ShishTa	Gourmet Fresh ShishTa	Lemon garlic chicken br	Gourmet Fresh Shishita	Certified Angus Beef Bur	Gourmet Fresh ShishTa
How many items are in this receipt?	4	4	3	4	5	4
What is the merchant's name on this receipt?	Gourmet Egypt	Gourmet Food Stores.	Gourmet	Gourmet food stores	Gourmet Food Stores	Gourmet Egypt
What is the receipt number for this receipt?	1089939	1089939	1089939	10849939	1089939	1089939
What is the transaction time for this receipt?	12:15:32 P	12:15:32 P	12:15:32	12:15	12:15:32 P	12:15:32 P
What is the TransactionDate for this receipt?	11/2/2021	11/2/2021	11/2/2021	11/15/02	11202121	11/2/2021
What is the total card amount paid in this receipt?	LE437.21	437.21	647.21	10843721	437.21	LE437.21

Table 11. Comparison of model responses on ReceiptQA in Figure 8b.

Question	**Ground Truth**	**ReceiptQA-Llama3.2**	**Phi 3.5**	**LLaVA**	**InternVL2 8B**	**GPT**
	Model Responses					
What is the name of item 1?	JUHAYNA MIX CHOCOLAT pcs	Juhayna Mix Chocolat	JUHANNA M/W XA 4428	JUHANNA MIKA	JUJINA YEMIX CHOCOLOAT M	JUHAYNA MIX CHOCOLAT pcs
What is the total price of item 1?	10.0	10.00	10.00	1000.00	6.14	10.00
What is the name of item 2?	AQUAFINA WATER 600 M pcs	AQUAFINA WATER 600 M PCS.	JUHANNA MIX WATER	JUHANNA MIKA	chocolat pos	AQUAFINA WATER 600 M pcs
What is the total price of item 2?	7.0	7.00	10.00	610.00	35.00	7.00
What is the name of item 3?	HEETS RUSSET'20 pcs	HEETS RUSSET 20 pcs	500.1	JUHANNA MIKA	MIX CHOCOLAT	HEETS RUSSET'20 pcs
What is the transaction time for this receipt?	08:07	08:07	15:05:07	0815	1500688036	08:07
What is the TransactionDate for this receipt?	15/06/23	15/06/23	15.08.07	08/15	15/06/08	15/06/23



(**a**) Test receipt 1

(**b**) Test receipt 2

Figure 8. Samples of test dataset.

Model Responses

In Tables 10 and 11 are the detailed responses of each model for the following two receipts.

We evaluate the models across three question types: text based (e.g., item names, merchant names), numerical (e.g., prices, receipt numbers), and date/time (e.g., transaction date and time). Accuracy is calculated as the percentage of correct responses compared to the ground truth.

For the Gourmet Egypt receipt (Table 10), LLaMA 3.2 and GPT show strong performance on text-based questions, both achieving 66.7% (two out of three). LLaVA correctly answers only one out of three, while Phi 3.5 and InternVL2 8B fail to get any correct. On numerical questions, all models except LLaVA perform similarly, each with 66.7% accuracy. LLaVA struggles with number parsing, misinterpreting the total amount significantly. For date/time extraction, LLaMA 3.2 and GPT again stand out, with both models correctly extracting the date and time, while the others fail to do so.

On the Emarat Misr receipt (Table 11), LLaMA 3.2 achieves perfect accuracy (3/3) on text-based item names, followed by GPT with 66.7% accuracy. Other models, including Phi 3.5, LLaVA, and InternVL2 8B, fail to retrieve any correct item names. In numerical price extraction, LLaMA 3.2, Phi 3.5, and GPT all achieve 100% accuracy, while LLaVA and InternVL2 misidentify the amounts. For date/time questions, LLaMA 3.2 and GPT again correctly identify both values, while other models provide incorrect or invalid outputs.

7. Dataset Use

LLM Evaluation and Training RECEIPTQA can be used to evaluate and fine-tune large language models (LLMs) for receipt-based question answering. It supports research on hallucination detection, prompt engineering, and domain-specific model optimization.

Information Extraction and Financial Applications RECEIPTQA enables the development of models for automated information extraction, expense tracking, and financial analytics by simulating real-world receipt scenarios with missing or ambiguous data.

Multimodal RAG Systems The dataset supports training and evaluating retrieval-augmented generation (RAG) systems that combine visual and textual cues for improved document understanding.

Transfer Learning and Cross-Domain Applications RECEIPTQA can facilitate transfer learning for other semi-structured documents, such as invoices and tickets, helping models generalize beyond receipts to diverse document types.

8. Conclusions

In this paper, we introduced RECEIPTQA, a large-scale dataset designed to advance receipt understanding through question answering. The dataset consists of 171,000 question–answer pairs derived from 3500 receipt images, created using a combination of LLM-generated and human-annotated questions. Our experiments with state-of-the-art vision–language models revealed significant performance variations, with high-precision models like LLaMA-3.2 (11B) achieving superior accuracy at the cost of increased processing time, while lightweight models like LLaVA prioritized efficiency but exhibited lower performance. This highlights the fundamental trade-off between accuracy and speed in document-based QA tasks. RECEIPTQA provides a robust benchmark for future research, offering insights into structured data extraction, numerical reasoning, and layout understanding. Future work can explore retrieval-augmented generation techniques, model optimization strategies, and multilingual dataset extensions to further enhance the capabilities of receipt-based question-answering systems.

Author Contributions: Conceptualization, M.A., M.S.K. and M.M.; Methodology, M.A., M.S.K., M.F.S. and A.A.; Software, M.A., M.M., B.Y., M.F.S. and S.H.K.; Validation, M.A. and H.S.K.; Formal analysis, M.A. and M.F.S.; Investigation, H.S.K.; Resources, H.S.K.; Data curation, M.A.; Writing—original draft, M.A.; Writing—review & editing, M.A. and H.S.K.; Visualization, H.S.K.; Supervision, H.S.K.; Project administration, H.S.K.; Funding acquisition, H.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (RS-2023-NR076833, 50%) and partly by the Innovative Human Resource Development for Local Intellectualization program through an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (Ministry of Science and ICT (MSIT)) (IITP-2025-RS-2020-II201462, 50%).

Data Availability Statement: The datasets used in this paper are public datasets, available in https://github.com/MahmoudElsayedMahmoud/ReceiptQA (1 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Majumdar, A.; Ajay, A.; Zhang, X.; Putta, P.; Yenamandra, S.; Henaff, M.; Silwal, S.; Mcvay, P.; Maksymets, O.; Arnaud, S.; et al. Openeqa: Embodied question answering in the era of foundation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16488–16498.
2. Abdallah, A.; Piryani, B.; Wallat, J.; Anand, A.; Jatowt, A. TempRetriever: Fusion-based Temporal Dense Passage Retrieval for Time-Sensitive Questions. *arXiv* **2025**, arXiv:2502.21024.
3. Piryani, B.; Abdallah, A.; Mozafari, J.; Jatowt, A. Detecting Temporal Ambiguity in Questions. *arXiv* **2024**, arXiv:2409.17046.
4. Abdallah, A.; Mozafari, J.; Piryani, B.; Ali, M.; Jatowt, A. From Retrieval to Generation: Comparing Different Approaches. *arXiv* **2025**, arXiv:2502.20245.
5. Abdallah, A.; Kasem, M.; Abdalla, M.; Mahmoud, M.; Elkasaby, M.; Elbendary, Y.; Jatowt, A. Arabicaqa: A comprehensive dataset for arabic question answering. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024; pp. 2049–2059.
6. Abdallah, A.; Piryani, B.; Mozafari, J.; Ali, M.; Jatowt, A. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. *arXiv* **2025**, arXiv:2502.02464.

7. Abdallah, A.; Mozafari, J.; Piryani, B.; Abdelgwad, M.M.; Jatowt, A. DynRank: Improving Passage Retrieval with Dynamic Zero-Shot Prompting Based on Question Classification. *arXiv* **2024**, arXiv:2412.00600.
8. Abdallah, A.; Mozafari, J.; Piryani, B.; Jatowt, A. ASRank: Zero-Shot Re-Ranking with Answer Scent for Document Retrieval. *arXiv* **2025**, arXiv:2501.15245.
9. Zhang, Y.; Jin, H.; Meng, D.; Wang, J.; Tan, J. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv* **2024**, arXiv:2403.02901.
10. Zhang, H.; Yu, P.S.; Zhang, J. A systematic survey of text summarization: From statistical methods to large language models. *arXiv* **2024**, arXiv:2406.11289. [CrossRef]
11. Koshkin, R.; Sudoh, K.; Nakamura, S. Transllama: Llm-based simultaneous translation system. *arXiv* **2024**, arXiv:2402.04636.
12. Huang, H.; Wu, S.; Liang, X.; Wang, B.; Shi, Y.; Wu, P.; Yang, M.; Zhao, T. Towards making the most of llm for translation quality estimation. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 375–386.
13. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
14. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **2024**, *18*, 186345. [CrossRef]
15. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
16. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
17. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
18. Anthropic. Claude 3 Haiku: Our Fastest Model Yet. 2024. Available online: https://www.anthropic.com/news/claude-3-haiku (accessed on 1 March 2025).
19. Liu, Z.; Huang, D.; Huang, K.; Li, Z.; Zhao, J. Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 4513–4519.
20. Toiganbayeva, N.; Kasem, M.; Abdimanap, G.; Bostanbekov, K.; Abdallah, A.; Alimova, A.; Nurseitov, D. Kohtd: Kazakh Offline Handwritten Text Dataset. *Signal Process. Image Commun.* **2022**, *108*, 116827. [CrossRef]
21. Abdallah, A.; Eberharter, D.; Pfister, Z.; Jatowt, A. A survey of recent approaches to form understanding in scanned documents. *Artif. Intell. Rev.* **2024**, *57*, 342. [CrossRef]
22. Abdallah, A.; Abdalla, M.; Kasem, M.S.; Mahmoud, M.; Abdelhalim, I.; Elkasaby, M.; ElBendary, Y.; Jatowt, A. Coru: Comprehensive post-ocr parsing and receipt understanding dataset. *arXiv* **2024**, arXiv:2406.04493.
23. Chen, Z.Z.; Ma, J.; Zhang, X.; Hao, N.; Yan, A.; Nourbakhsh, A.; Yang, X.; McAuley, J.; Petzold, L.; Wang, W.Y. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. *arXiv* **2024**, arXiv:2405.01769.
24. Abdallah, A.; Berendeyev, A.; Nuradin, I.; Nurseitov, D. Tncr: Table net detection and classification dataset. *Neurocomputing* **2022**, *473*, 79–97. [CrossRef]
25. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
26. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. *arXiv* **2023**, arXiv:2310.03744.
27. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *arXiv* **2023**, arXiv:2304.08485.
28. Pang, R.Y.; Parrish, A.; Joshi, N.; Nangia, N.; Phang, J.; Chen, A.; Padmakumar, V.; Ma, J.; Thompson, J.; He, H.; et al. QuALITY: Question Answering with Long Input Texts, Yes! In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 5336–5358.
29. Mathew, M.; Karatzas, D.; Jawahar, C. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 2200–2209.
30. Tito, R.; Karatzas, D.; Valveny, E. Document collection visual question answering. In Proceedings of the Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, 5–10 September 2021; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2021; pp. 778–792.
31. Biten, A.F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Valveny, E.; Jawahar, C.; Karatzas, D. Scene Text Visual Question Answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
32. Singh, A.; Natarjan, V.; Shah, M.; Jiang, Y.; Chen, X.; Parikh, D.; Rohrbach, M. Towards VQA Models That Can Read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.

33. Yagcioglu, S.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv* **2018**, arXiv:1809.00812.
34. Tanaka, R.; Nishida, K.; Nishida, K.; Hasegawa, T.; Saito, I.; Saito, K. Slidevqa: A dataset for document visual question answering on multiple images. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 13636–13645.
35. Onami, E.; Kurita, S.; Miyanishi, T.; Watanabe, T. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. *arXiv* **2024**, arXiv:2403.19454.
36. Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M. Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 1192–1200.
37. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florêncio, D.A.F.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual, 1–6 August 2021; pp. 2579–2591.
38. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv* **2022**, arXiv:2204.08387.
39. Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; Luo, J. TAP: Text-Aware Pre-training for Text-VQA and Text-Caption. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
40. Hu, R.; Singh, A.; Darrell, T.; Rohrbach, M. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9992–10002.
41. Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv* **2024**, arXiv:2412.05271.
42. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 24185–24198.
43. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
44. Wang, J.; Jatowt, A.; Yoshikawa, M. ArchivalQA: A Large-scale Benchmark Dataset for Open-Domain Question Answering over Historical News Collections. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 3025–3035.
45. Chu, X.; Su, J.; Zhang, B.; Shen, C. Visionllama: A unified llama backbone for vision tasks. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–18.
46. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
47. Shazeer, N. Glu variants improve transformer. *arXiv* **2020**, arXiv:2002.05202.
48. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
49. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv* **2022**, arXiv:2212.10511.
50. Adlakha, V.; BehnamGhader, P.; Lu, X.H.; Meade, N.; Reddy, S. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv* **2023**, arXiv:2307.16877. [CrossRef]
51. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
52. Mozafari, J.; Abdallah, A.; Piryani, B.; Jatowt, A. Exploring Hint Generation Approaches in Open-Domain Question Answering. *arXiv* **2024**, arXiv:2409.16096.
53. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.