

Multimodal Information Extraction of Supermarket Leaflets

Ilyesse Hettenbach, Gabriel Schurr

HKA, Karlsruhe, Germany

{heil1012, scga1011}@h-ka.de

<https://github.com/ilyii/leaflets>

Abstract

TODO

1 Introduction

1.1 Motivation

If one thinks of a paper-heavy country, Germany is likely to come to mind. The country’s affinity for printed materials is deeply ingrained in its culture, with activities such as browsing through supermarket leaflets serving as a familiar ritual for many individuals. These leaflets have long been an essential medium for consumers to plan their purchases, discover special offers, and compare prices. The tactile nature of printed leaflets, the excitement of spotting discounts, and the satisfaction of making a well-informed shopping decision contribute to their enduring appeal. However, the rapid advancement of digital technology is transforming consumer behavior. While some individuals still find browsing through physical leaflets a nostalgic and even relaxing experience, younger generations are increasingly turning to digital alternatives for their shopping needs. The modern shopper expects convenience, efficiency, and instant access to information, yet the digital transformation of supermarket leaflets lags significantly behind other aspects of e-commerce and retail technology.

Despite the existence of applications that compile digital versions of these leaflets, their usability and overall quality remain suboptimal. Many digital leaflets are plain documents, failing to leverage the full potential of interactivity, searchability, and intelligent recommendation systems. There is a pressing need for sophisticated digital solutions that not only present supermarket deals but also enable consumers to interact with them in a meaningful way.

Possibly, one can imagine a seamless digital platform where users can search for specific items across multiple supermarket chains, apply dietary

filters to instantly highlight suitable products, or receive personalized recommendations based on past preferences. A tool that allows consumers to create a dynamic shopping list, track price trends over time, and receive real-time updates on the best deals would revolutionize the way people approach grocery shopping. Yet, these features remain largely absent from existing digital leaflet solutions. Furthermore, while price comparison websites and applications have become commonplace for electronics, fashion, and travel, the grocery sector remains an overlooked frontier. Consumers are often left navigating multiple supermarket websites or manually cross-referencing prices to find the best deals—a tedious and inefficient process. A comprehensive digital ecosystem that integrates grocery price comparisons, personalized discounts, and AI-driven shopping assistants could bridge this gap and act as a workaround, offering an enhanced, data-driven shopping experience tailored to individual needs.

The future of supermarket deal discovery should not be a static PDF or an unorganized collection of scanned pages—it should be a dynamic, interactive, and intelligent experience that aligns with the evolving digital landscape.

1.2 Problem Statement

The primary objective of this research is to develop an intelligent system for extracting, structuring, and utilizing information from supermarket leaflets in a digital format. Given a set of supermarket leaflets $\mathcal{L} = \{L_1, L_2, \dots, L_n\}$, each containing a collection of product deals with specific attributes such as product name, brand, price, discount, product image and other product specifications, the goal is to extract and normalize these attributes to enable efficient querying and comparison across multiple supermarket chains.

Each leaflet L_i consists of a set of m product entries $P_i = \{p_1, p_2, \dots, p_m\}$, where each product

p_j is associated with a set of varying attributes:

$$p_j = \{a_1, a_2, \dots, a_k\}, \quad (1)$$

where each attribute a_k represents a specific entity. Common attributes include:

- Product Name
- Brand
- Original Price
- Deal Price
- Unit (e.g., weight, volume)
- Product Image

among others. Each attribute may or may not be present in a given product entry, and the format and layout of these attributes can vary significantly across different leaflets.

To enable advanced applications such as the abovementioned, one has to tackle different challenges, such as:

1. Detecting and separating deals p from the leaflet L .
2. Extracting structured information from the detected deals.
3. Normalizing the extracted information to ensure consistency and comparability.
4. Validating the extracted information to ensure accuracy.

Therefore, a function

$$f : \mathcal{L} \rightarrow \mathcal{D}$$

has to be designed, where \mathcal{D} represents the structured information extracted from the leaflets. The function f should minimize the extraction error E .

1.3 Contributions

This work presents a comprehensive approach to digitalizing and structuring supermarket leaflet information by leveraging advanced computer vision, OCR, and language models. The key contributions of this research are as follows:

Development of a Detection and Extraction Pipeline for Supermarket Leaflets. We propose a modular and efficient pipeline for detecting, extracting, and structuring information from supermarket leaflets. This pipeline integrates OCR-based

text extraction, image processing, and layout understanding to accurately associate textual and visual elements such as product names, prices, descriptions, and images. Additionally, we incorporate normalization techniques to standardize extracted data across different retailers and formats.

Design of an Interactive Application for Deal Browsing and Price Comparison. To facilitate practical use, we wrap the extracted data into a user-friendly application that enables consumers to browse, filter, and compare deals across multiple supermarket chains. This application includes basic functionalities to search, track and compare deals across different supermarket chains.

Comprehensive Evaluation of Information Extraction Models. We systematically evaluate multiple state-of-the-art models for supermarket leaflet information extraction, including hybrid OCR+LLM approaches, end-to-end Vision-Language Models (VLMs), and Large Vision-Language Models (LVLMs). Our evaluation covers accuracy metrics such as per-entity correctness, Levenshtein distance, and robustness across different supermarket layouts and textual formats. This analysis provides insights into the trade-offs between computational efficiency and extraction performance.

Training of Custom Deal Detection and Information Extraction Models. To further improve extraction performance, we develop and fine-tune specialized models for deal detection and information extraction. This includes training custom object detection models to identify product regions within leaflets, as well as domain-specific sequence-to-sequence models for structured text extraction. These models are designed to enhance recognition accuracy, particularly for complex layouts and noisy scan conditions.

Creation of a Benchmark Dataset: Leaflet-IE. We introduce **Leaflet-IE**, a novel dataset curated for supermarket leaflet information extraction. The dataset consists of annotated images spanning multiple supermarket chains, with labeled entities including product names, prices, discounts, brands, and categories. This dataset serves as a benchmark for future research in structured information extraction from promotional materials and contributes to advancing OCR and multimodal learning for real-world retail applications.

2 Related Works

Deal Detection. Optical Character Recognition. OCR-Model-Driven methods use OCR tools to acquire text and bounding box information. Subsequently, they rely on the models to integrate text, layout, and visual data.

Information Extraction.

3 Nature of Supermarket Leaflet Data

Supermarket leaflets constitute a complex and heterogeneous data source, characterized by multimodal content that combines structured and unstructured information. This data presents unique challenges for computational processing due to its inherent ambiguities, multi-instance object representations, and spatial dependencies. The reader is encouraged to simultaneously inspect the visual representation of these challenges in [Figure 1](#) while proceeding with the discussion.

One of the primary challenges lies in the ambiguity of price recognition, where a single product may be associated with multiple price points, such as discounted prices, bulk purchase offers, or tiered pricing structures. This ambiguity is further exacerbated by errors in punctuation extraction, particularly in recognizing decimal places, which often arise from font-specific artifacts or noisy backgrounds. Resolving such issues requires advanced plausibility checks, contextual reasoning, and robust numerical interpretation mechanisms to ensure accurate data extraction. Additionally, the non-strict one-to-one mapping between textual and visual elements introduces further complexity. For instance, a single product image may represent multiple items, or textual descriptions may not align precisely with their corresponding visual representations. This misalignment necessitates sophisticated multimodal fusion techniques to bridge the gap between visual and textual data, ensuring accurate association and interpretation of product information.

The inconsistent layouts of supermarket leaflets further complicate data extraction. Promotional deals and product information are often placed in varying positions across different leaflets, making it difficult to design a universal extraction method that generalizes across diverse formats. Text elements frequently overlap with product images or other textual information, leading to misidentification or incorrect parsing of data. These challenges are compounded by the variability in fonts, sizes,

and weights used within a single leaflet, which complicates the differentiation between product names, prices, and promotional tags. The presence of multilingual content and special characters adds another layer of complexity, as leaflets may contain text in multiple languages or stylized symbols that require optical character recognition (OCR) and natural language processing (NLP) models capable of handling diverse linguistic structures and alphabet systems.

Promotion labels and discount representations introduce additional challenges due to their non-standard or highly stylized text formats. Identifying the exact nature of promotions, such as "Buy 1, get 1 free" or "20% off," requires precise extraction and interpretation, particularly when the text is embedded in noisy backgrounds or overlaps with other elements. Furthermore, the temporal context of deals and promotions adds a critical dimension to the extraction process. Some offers may only be valid for specific periods, and extracting and associating the correct date ranges with these offers can be challenging, especially when the information is not explicitly stated or is presented in ambiguous formats.

Therefore, the nature of supermarket leaflet data necessitates an advanced toolkit of visual and textual processing, analysis, extraction and generation techniques to effectively interpret and utilize the information contained within these documents. The subsequent sections will delve into the methodologies and models used to address these challenges and extract structured information from supermarket leaflets.

4 Deal Detection

4.1 Datasets

4.2 Model

4.3 Experiments

5 Information Extraction of Supermarket Deals

Subsequently to the detection and separation of supermarket deals, the overarching goal of using these in applications requires the extraction of the various useful information from these deals. The task of extracting meaningful and structured information from supermarket deal images is a complex and multi-faceted problem, requiring the integration of various modalities, overcoming challenges posed by noisy data, and leveraging deep learning

techniques. This section provides a detailed exploration of the underlying challenges, methodologies, and the theoretical and practical framework needed to perform such extraction.

5.1 Challenges in Information Extraction for Supermarket Deals

Supermarket deal extraction involves a series of challenges that require careful attention and sophisticated methods to resolve. These challenges include the diversity of layout and format in the images, the multimodal nature of the data, the frequent occurrence of Optical Character Recognition (OCR) errors, and the specific linguistic and cultural issues posed by the German language.

High Variety in Layouts and Visual Elements. Supermarket deals exhibit considerable variability in layout, color schemes, font choices, and text sizes, which hinders automatic detection and extraction. Deals are often printed with superscripted decimals or without clear separation between integer and fractional components of the price (e.g., "2.29" might appear as "229"). This phenomenon exacerbates OCR difficulties. Additionally, original prices may be struck through, while deal prices are sometimes printed in drastically different fonts than those the OCR model was trained on. Furthermore, overlapping elements—such as product images or additional textual information—often interfere with text extraction and localization. These diverse visual elements require robust and flexible models capable of accommodating such variability.

Multimodal Information. Supermarket deal images consist of both visual (e.g., product images) and textual (e.g., brand names, prices, and product descriptions) modalities, each contributing different types of information. While text provides rich information about the product’s identity, pricing, and units, the image serves as a supplementary modality that aids in product identification, brand recognition, and other visual cues. Multimodal fusion, where information from both text and image domains is integrated, is a key challenge, and existing models must effectively leverage both sources to provide high-quality extraction.

Conversion Errors and Linguistic Challenges. OCR systems frequently introduce errors, particularly when dealing with non-standard fonts, noisy backgrounds, or small text. This is particularly problematic when extracting numerical data such as prices. Common errors include misinterpretation of decimals and digits (e.g., "229" instead of

"2.29") and missing characters. To address this, OCR post-processing steps, such as error correction using context-based reasoning or dedicated error models, are required. Additionally, the presence of proper nouns—particularly brand names—adds complexity, as these entities may not appear in typical language models, making their extraction difficult. Furthermore, the German language presents its own set of challenges, including compound words, specific punctuation conventions, and diverse word forms, all of which must be accounted for in any robust extraction model.

Ambiguities. The presence of overlapping elements—whether textual or graphical—can lead to a degradation in the extraction accuracy. This challenge is compounded when the overlap involves essential information, such as when the original price is partially obscured by a product image or strike-through marks. Furthermore, ambiguity in labeling and the context in which different pieces of information appear in the deal image can create difficulties in associating text with the correct object (e.g., associating a price with a product rather than the surrounding descriptive text).

5.2 Formal Approach

Formally, the task of information extraction from a deal image can be defined as the identification of specific entities, including the product name, brand, original price, deal price, and unit. We define the following set of output labels:

$$\mathcal{Y} = \{y_i\}_{i=1}^n, \quad (2)$$

where y_i represents the i -th entity in the deal image. In the following, the entities are defined as:

- $y_{\text{product_name}}$: The name of the product.
- y_{brand} : The brand of the product.
- $y_{\text{original_price}}$: The original price of the product.
- $y_{\text{deal_price}}$: The deal price of the product.
- y_{unit} : The unit of the product (e.g., weight, volume).

To extract these entities, one aims to learn a function $f(\cdot)$ that maps an input image I to the desired outputs:

$$f : I \rightarrow \mathcal{Y}, \quad (3)$$

where I is the deal image, and \mathcal{Y} represents the extracted entities.

The choice of the function $f(\cdot)$ is an architectural decision that depends on the specific requirements of the task, the nature of the data, the available resources, and the desired performance metrics.

5.3 Architectural Approaches to Information Extraction

Among the various existing architectural approaches to information extraction, several methodologies have been developed to address the inherent complexities in supermarket deal images. These methods can be categorized into classical approaches, multi-stage traditional models, hybrid OCR-LLM systems, and end-to-end deep learning frameworks. Each paradigm offers distinct advantages, and their applicability depends on computational resources, dataset characteristics, and performance requirements.

5.4 Traditional Multi-Stage Methods

Traditional architectures decompose the problem into sequential sub-tasks and solve these independently through sub-task-specific systems. While these methods have been studied extensively and with high variability in methodology, they are often divided into three main stages: text detection, text recognition, and key-value pair extraction.

Text Detection. Text detection is the process of identifying regions in the image that contain textual information. Commonly used methods include rule-based systems, as well as mostly small object detection neural networks to detect regions of textural interest by predicting M bounding boxes $\{b_i\}_{i=1}^M$.

Text Recognition. Each bounding box b_i is used to extract the corresponding region in the image, which is passed to a text recognition model to convert the visual representation into a textual one. Since this task is often more complex than it might appear due to decoding errors, there exists a high variety of models that can be used for this task, such as CRNN (?) or transformer-based recognizers. Given a set of M bounding boxes $\{b_i\}_{i=1}^M$, the goal is to predict the corresponding textual entities $\{t_i\}_{i=1}^M$.

$$\{t_i\}_{i=1}^M = \arg \max_{\{t_i\}_{i=1}^M} P(\{t_i\}_{i=1}^M | \{b_i\}_{i=1}^M). \quad (4)$$

Text detection and recognition, together with

pre-and post-processing steps, is often referred to and standardized as Optical Character Recognition (OCR).

Key-Value Pair Extraction: Finally, extracted text is structured into meaningful entities. Named Entity Recognition (NER) models or Graph Neural Networks (GNNs) can be used to learn associations between extracted elements:

$$\mathcal{A}^* = \arg \max_{\mathcal{A} \subseteq E} \sum_{(i,j) \in \mathcal{A}} \Psi(v_i, v_j), \quad (5)$$

where E represents textual adjacency relationships.

While these types of approaches are widely used, they inherit a high complexity due to the need for multiple models with each being trained, evaluated, tested and managed independently, which also leads to a higher computational overhead due to input-output transformations between the stages. Furthermore, the lack of end-to-end training can lead to suboptimal performance and error propagation, which is often seen when applying these types of models to edge cases or unseen data.

5.5 Hybrid OCR-LLM Models

The rise of large language models (LLMs) has led to the development of hybrid OCR-LLM models that combine the strengths of OCR tools with the linguistic capabilities of LLMs. While these approaches still rely on a more traditional process to convert visually appearing text into a machine-readable form, the LLM is used to enhance the robustness of the OCR output by providing contextual reasoning capabilities.

$$\hat{y} = F_\theta(T), \quad (6)$$

where T is the OCR output and F_θ is a pretrained, general-purpose LLM that is being prompted with a specific task.

5.6 End-to-End Models

The

End-to-End (E2E) approaches eliminate pipeline fragmentation by learning a unified mapping $F : I \rightarrow \mathcal{D}$. Two major architectures dominate this space:

Vision Encoder-Decoder Models: Models such as Donut (?) encode input images via a Vision Transformer (ViT) and generate structured outputs via an autoregressive decoder:

$$P(Y|I) = \prod_{t=1}^T P(y_t | I, y_{<t}). \quad (7)$$

These methods improve robustness by directly training on image-text pairs without explicit OCR supervision.

Large Vision-Language Models (LVLMs): Multimodal transformers such as BLIP (?) and LLaVA (?) leverage joint vision-text embeddings to infer structured outputs. Given an image-text embedding $z = f_{\theta}(I, T)$, task-specific decoding is performed via:

$$\hat{y} = \text{Decoder}(z), \quad (8)$$

where Decoder is a domain-adapted transformer head.

E2E models significantly reduce error propagation but require large-scale annotated datasets for effective training. Their adoption is growing with advancements in multimodal pretraining strategies and self-supervised learning.

5.7 Dataset Creation

Since the availability of German supermarket leaflet data is literally not findable, a custom dataset was created by manually annotating a collection of supermarket deal images. The dataset includes images from the most common supermarket chains, each with unique layout, fonts, and colors to, in general, ensure the highest diversity and robustness possible. Each image is annotated with a corresponding label file that includes the desired entities as well as the image identifier. The dataset is split into a training (80%) and a validation set (20%) for each model to be trained and evaluated on. Table Table 1 provides an overview of the dataset with the sample size for each entity and Figure 2 shows some sample images with their annotations.

Table 1: Leaflet-IE Dataset

Entity	Sample Size
image_id	372
brand	357
product_name	370
original_price	286
deal_price	372
weight	369

Even though the Leaflet-IE dataset is relatively small, the reader will be able to see that the dataset is sufficient to evaluate the performance of different IE approaches as well as to train an competitive

end-to-end model. However, as the reader may notice, the dataset is solely focused on single product deals at this point and explicitly does not include

6 Experiments

The primary objective of these experiments is to identify the most effective approach for extracting structured information from supermarket leaflets. The evaluation considers general performance trends and use cases to determine candidate methodologies, including Hybrid OCR + LLM, end-to-end Vision-Encoder-Decoder models, and LVLMs. The order of the experiments is as follows: - Plain Deal OCR Performance - OCR + LLM Performance - Donut Fine-Tuning Performance - MAIN: Comparison of Hybrid OCR+LLM, Donut, and LVLM

6.1 Implementation Details

The experiments were conducted on two distinct computational workstations - station 1 equipped with an NVIDIA RTX 4080 GPU and station 2 with an NVIDIA GTX 1080 GPU. Therefore, the experiments could be conducted in parallel, ensuring efficient model training and evaluation.

The software stack includegraphics Python with PyTorch as the primary deep learning framework, HuggingFace Transformers for leveraging pre-trained models, and OpenCV for image preprocessing and data augmentation. PyTorch, HuggingFace Transformers, and various other libraries for data pre-and post-processing, visualization, data management and monitoring ¹.

6.2 Plain Deal OCR Performance

The aim of this experiment is to evaluate the performance of various OCR models in extracting text from supermarket leaflets. The evaluation focuses on the accuracies of the extracted entities from Table 1 and the impact of different normalization levels on the recognition performance. The results provide insights into the robustness and accuracy of different OCR models in handling the complexities of supermarket leaflet data.

The most commonly used OCR models include:

- Tesseract,
- EasyOCR,
- PaddleOCR,

¹The codebase is available at <https://github.com/ilyii/leaflets>

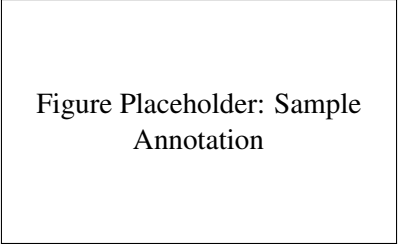


Figure Placeholder: Sample
Annotation

Figure 1: Sample images with annotations from the custom dataset.

- DocTR.

These models have been trained on a variety of datasets and are known for their robustness and accuracy in text extraction. Since it is likely that none of the OCR models have seen this data before, all 372 samples of the Leaflet-IE dataset were used for evaluation. While OCR alone is not capable of generating key-value mappings, the evaluation used a different set of metrics:

- **Accuracy:** Measures whether the expected value is present in the OCR output.
- **N-gram Accuracy:** Determines if an n-gram substring of the expected value exists in the OCR output, computed for $n \in \left[\frac{|v|}{2}, |v| - 1\right]$, where $|v|$ represents the string length of the ground truth entity value.

Given that language-specific OCR errors are common, the evaluation was conducted at different levels of normalization to improve comparability. The text normalization included the following layers:

- **Level 1:** Text stripping and string conversion.
- **Level 2:** Lowercasing and basic replacements.
- **Level 3:** Use-case specific replacements.
- **Level 4:** Punctuation and whitespace removal.

Each level includes the normalization steps of the previous levels. Therefore, the normalization procedure can be thought of as a gradual alignment of the OCR output to the ground truth to investigate the OCR model's robustness to different normalization levels.

Results. Figure 3 presents a comparative analysis of various OCR models and normalization levels based on per-entity accuracy. The results highlight several key trends regarding the recognition performance of different entities.

In general, the OCR models exhibit superior performance in extracting words compared to numerical values. This discrepancy is particularly evident in entity-wise accuracy variations. At the baseline level (normalization level 1), the deal price entity consistently shows the lowest accuracy, whereas the brand entity achieves the highest. The accuracy of other entities varies significantly, with PaddleOCR, EasyOCR, and docTR displaying comparable performance, while Tesseract lags behind.

The overall accuracy improves with increasing normalization levels, with normalization level 4 yielding the highest accuracy gains. At this level, Tesseract continues to struggle with deal price recognition, failing in over 90% of cases. In contrast, the other OCR models achieve accuracies exceeding 70% across nearly all entity types.

Among the different entities, brand names are consistently recognized with the highest accuracy. This can be attributed to their simpler linguistic structure and distinct visual presentation, as brand names are often bolded or highlighted in promotional material. In contrast, product names tend to have more complex linguistic structures, making exact string matching more challenging.

The recognition of unit information lags behind brand names but remains more accurate than deal prices. The reduced accuracy can be explained by the smaller font size of unit labels, their less prominent placement, and the high variability in unit measurements (e.g., "kg", "g", "Stück"), which complicates recognition.

A significant gap is observed between deal price and original price accuracy. This can be attributed to the typographic differences between the two: deal prices are often emphasized using distinctive fonts and superscripted decimals, making them harder for OCR models to interpret correctly. In contrast, original prices are typically displayed in a standard inline format, which is easier to recognize.

When evaluating n-gram accuracy per-entity in Figure 4, the results reveal a similar trend to per-

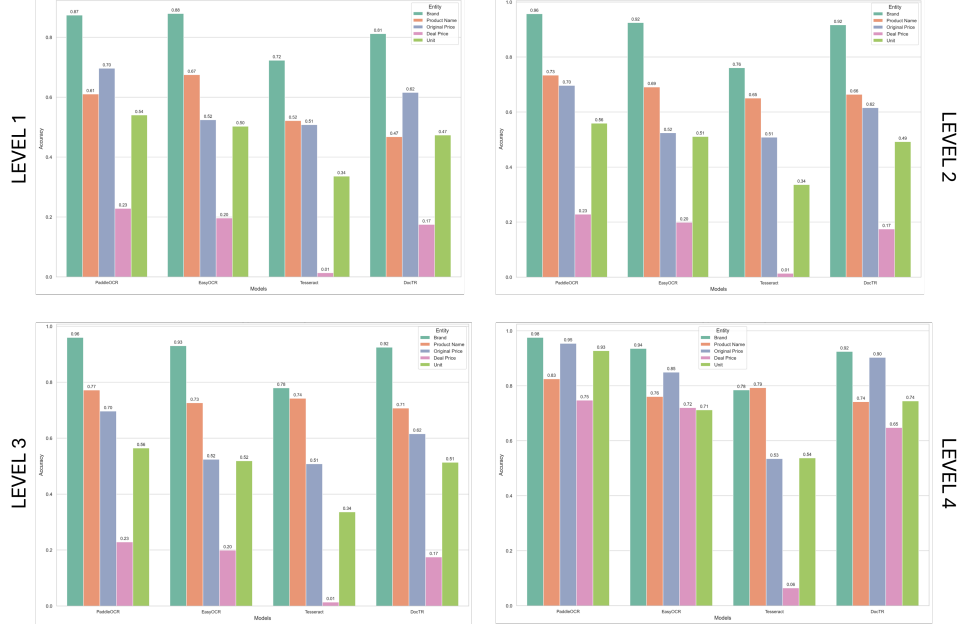


Figure 2: Accuracy per Entity of OCR Models at Different Normalization Levels.

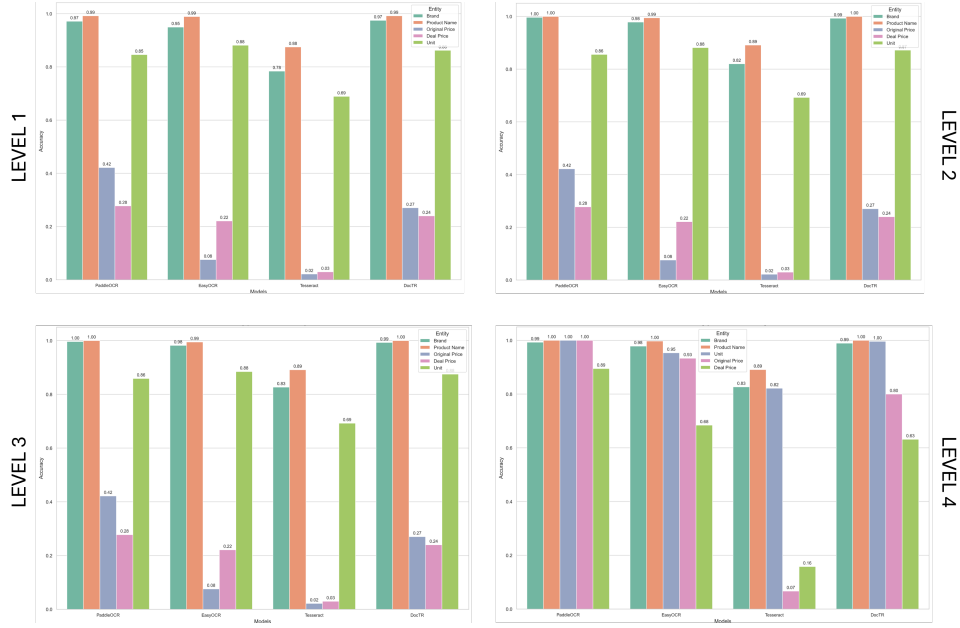


Figure 3: N-gram Accuracy per Entity of OCR Models at Different Normalization Levels.

entity accuracy. When comparing the OCR models, PaddleOCR consistently outperforms the other models especially in the price extractions while Tesseract lacks behind in all entities.

The most significant observation is that the deal price and original price exhibit in lower normalization levels significantly worse accuracies. Since the length of n-gram sequences increase superlinearly, longer entity values, precisely the brand and product name, more likely achieve higher accuracies, which is also reflected in the results.

At the highest normalization level, most models achieve 100% or close to 100% n-gram accuracy in the majority of entites. Additionally, here the discrepancy of Tesseract is most visible, where it cannot nearly achieve comparable scores. Also, one can observe that the unit entity drops significantly at level 4 across all models. This can be attributed to the high variability in unit measurements, which makes it challenging to match the exact n-grams.

6.2.1 Hybrid-OCR+LLM Performance

In addition to evaluating the OCR capabilities, the performance of OCR models was assessed in conjunction with large language models (LLMs). The primary questions addressed were:

- Which LLM is best suited to work with these OCR outputs?
- Which OCR and LLM combination works best for information extraction (IE)?

All 372 samples of the Leaflet-IE dataset were used for evaluation. It is assumed that neither the OCR models nor the LLMs have previously encountered this data, similar to the OCR evaluation. Since the task is to extract structured information, the evaluation metrics were adapted to reflect the performance of the OCR-LLM combination in generating key-value mappings. The evaluation metrics include:

- **Accuracy (Per Entity):** Measures the accuracy (of each extracted entity).
- **Levenshtein Distance:** Calculates the distance between the prediction and the ground truth.

The LLMs were prompted to extract entities from the OCR output. The specific prompt used can be found in the appendix (see Appendix ??).

Models Evaluated:

- **Llama 3.1 [8b, Q4]:** llama3.1_8b,
- **Qwen 2.5 [1.5b, Q8]:** qwen2.5_1.5b-instruct-q8_0,
- **Llama 3.2 [3b, Q8]:** llama3.2_3b-instruct-q8_0,
- **Qwen 2.5 [7b, Q4]:** qwen2.5_7b,

where the LLM name and version is additionally specified through the model size (in billions of parameters) and the quantization level.

To avoid confusion with the normalization used in the OCR experiment, the OCR results were kept raw, as normalization may remove valuable information. The LLMs are expected to handle such inconsistencies and errors in the input. Instead, here the entity values were evaluated in both raw form and with a normalization applied for the comparability between the LLM prediction and the ground

truth value. The normalization procedure includes usual character replacements and filtering based on linguistic artifacts commonly encountered in OCR errors. The specific normalization procedure can be found in the appendix

Results: The entity-averaged accuracies and Levenshtein distances for the raw OCR output are depicted in Figures 5 and 6, respectively.

The averaged accuracies reveal that a higher number of parameters is more beneficial for the LLMs in this task than a higher quantization level. Specifically, the Llama 3.1:8b and Qwen 2.5:7b models achieve significantly higher accuracy scores than their counterparts with a lower number of parameters. Qwen 2.5:1.5b in Q8 performs worse than Llama 3.2:3b in Q8, but Qwen 2.5:7b in Q4 is overall the best. These differences can be attributed to multiple factors, such as the pre-training data of the models, the model architecture, or the prompt used to extract the entities. DocTR and PaddleOCR achieve equal or better results than Tesseract and EasyOCR. Specifically, Tesseract consistently performs the worst among the evaluated models.

When regarding the levenshtein distances, the results show similar trends. DocTR delivers by far the best OCR input across all OCR models. Surprisingly, Tesseract achieves lower levenshtein distances than PaddleOCR and EasyOCR, with EasyOCR being the worst across all models. This could be due to the fact that Tesseract delivers more often slightly wrong but still similar results, while EasyOCR delivers less often wrong but more different results. The LLMs show similar trends as in the accuracy evaluation, but Llama 3.1:8b now slightly outperforms Qwen 2.5:7b. The differences between the LLMs are less pronounced than in the accuracy evaluation, which can be attributed to the unnormalized levenshtein distance calculation used.

6.3 Donut Fine-Tuning Performance

Experiment Question: - Is it possible to train a Vision Encoder Decoder model to extract end2end information from supermarket leaflet deal images?

Setup:

Results:

6.4 Comparison of Hybrid OCR+LLM, Donut, and LVLM

This section presents a comparative evaluation of a hybrid OCR+LLM pipeline against end-to-

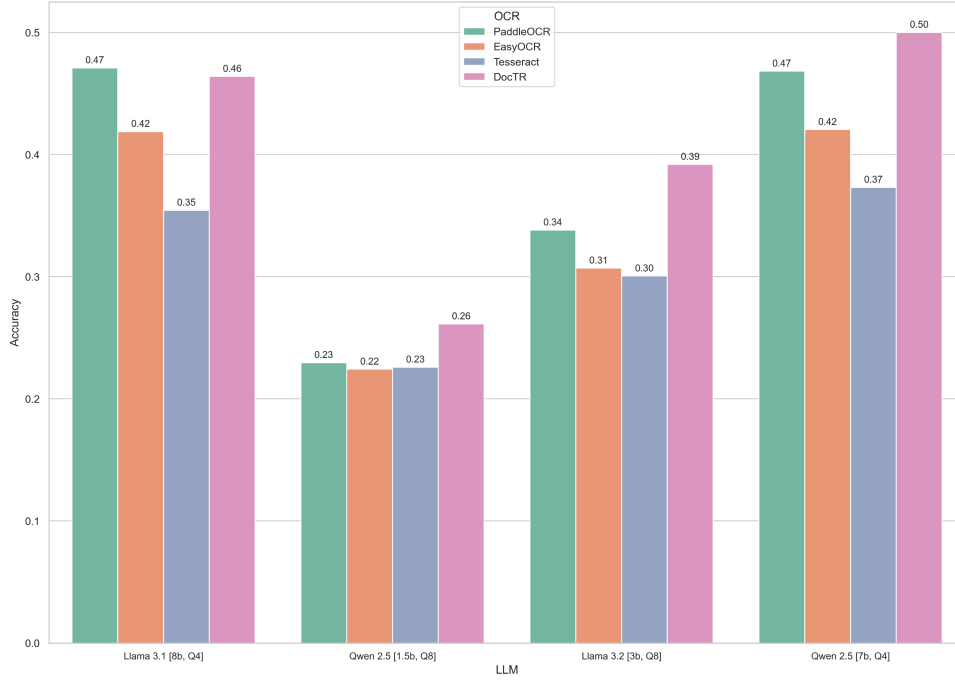


Figure 4: Entity-Averaged Accuracy of LLM Models per Raw OCR Output.

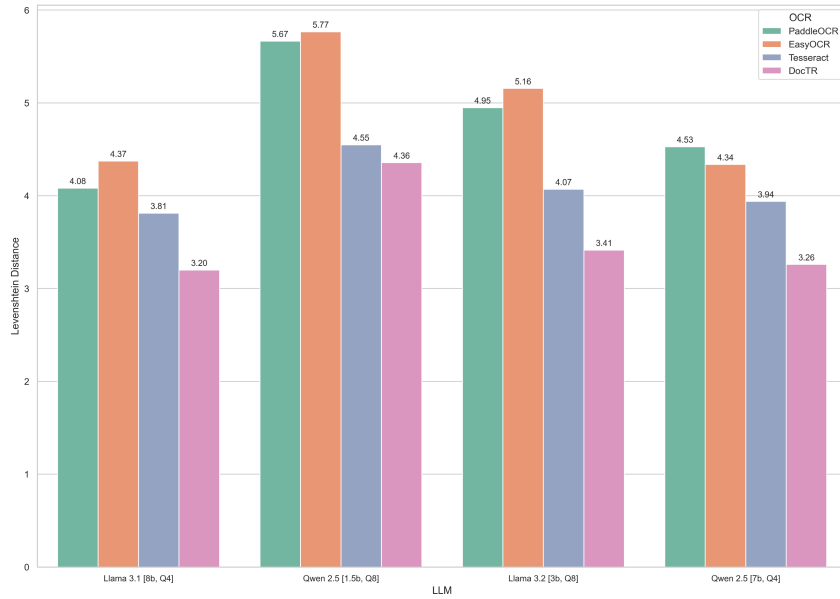


Figure 5: Entity-Averaged Levenshtein Distance of OCR-LLM Models (Raw OCR Output).

end models, specifically Donut and Large Vision-Language Models (LVLMs), for information extraction (IE) from supermarket leaflets. The assessment focuses on model performance using the best OCR+LLM configuration and two end-to-end approaches.

6.4.1 Experimental Setup

Models. The following models were selected for evaluation:

- **OCR+LLM:** DocTR was utilized as the OCR engine, with Qwen 2.5-7B and LLaMA 3.1-8B serving as LLMs due to their comparable performance.
- **Donut:** An end-to-end Vision Encoder-Decoder model optimized for document understanding.
- **LVLMs:** MiniCPM-V and LLaMA 3.2-Vision were chosen as large vision-language

models for comparison.

Dataset. The Leaflet-IE validation subset was used for evaluation since Donut had been trained on the train subset.

Evaluation Metrics. The models were evaluated using:

- **Accuracy:** The percentage of correctly extracted entity values.
- **Levenshtein Distance:** The edit distance between predicted and ground truth values at the entity level.

Normalization. To ensure consistency, the same normalization approach from the OCR+LLM evaluation was applied to both model predictions and ground truth values.

6.4.2 Results and Discussion

Per-Entity Accuracy Without Normalization (Figure 7)

- The OCR+LLM pipeline performed significantly worse than both Donut and LVLMs. Specifically, its accuracy for original and deal price entities was close to 0%.
- LVLMs outperformed all other models, particularly in extracting price-related information. Their superior performance is attributed to their ability to leverage spatial information and contextual cues.
- Donut achieved competitive results, closely trailing LVLMs, but showed a notable drop in accuracy for brand recognition.

Per-Entity Levenshtein Distance Without Normalization (Figure 8)

- As expected from the accuracy results, LVLMs and Donut exhibited low Levenshtein distances, indicating high precision.
- The OCR+LLM approach had significantly higher edit distances, particularly for brand recognition, suggesting a consistent difficulty in extracting brand names accurately.
- Among all entities, brand names showed the highest Levenshtein distance across models, which is uncommon in IE tasks.

Per-Entity Accuracy With Normalization (Figure 9)

- Applying normalization significantly improved accuracy, especially for OCR+LLM. Interestingly, its price detection performance exceeded that of Donut in some cases.
- Despite the improvements, OCR+LLM remained notably weaker than LVLMs and Donut.
- Donut displayed stable and balanced accuracy across all entities, whereas LVLMs exhibited higher variance.
- LVLMs retained their leading performance, particularly in price extraction.

Per-Entity Levenshtein Distance With Normalization (Figure 10)

- Levenshtein distances decreased across all models post-normalization, especially for OCR+LLM.
- LVLMs achieved the lowest Levenshtein distances but maintained high variance.
- Donut demonstrated remarkable consistency across all entities.
- The brand entity continued to exhibit the highest Levenshtein distances across models.

6.4.3 Conclusions

- **LVLMs are the most effective for IE tasks**, particularly when spatial information is crucial. However, their computational requirements are significantly higher, making them less practical in resource-constrained environments.
- **Donut provides a strong trade-off between accuracy and efficiency.** Given the limited training resources, Donut has the potential to outperform LVLMs in the long run for this specific task. However, its generalization ability is limited compared to LVLMs, which can be applied to a wider range of tasks.
- **OCR+LLM is considerably weaker** than the other approaches. Its performance limitations are primarily attributed to:
 - Lack of spatial and visual information.

- Error propagation from OCR to LLM.
- Insufficient training data for robust IE.

Nevertheless, OCR+LLM remains a viable option when computational resources are constrained, and the task complexity is moderate.

7 Application

7.1 Database

8 Application

9 Database

10 Front End / Webapp

11 Conclusion

11.1 Summary

11.2 Future Work

ww

A Appendices

A.1 OCR Evaluation:Normalization Procedure

A.2 LLM Prompt

TODO

A.3 OCR + LLM Evaluation: Normalization Procedure

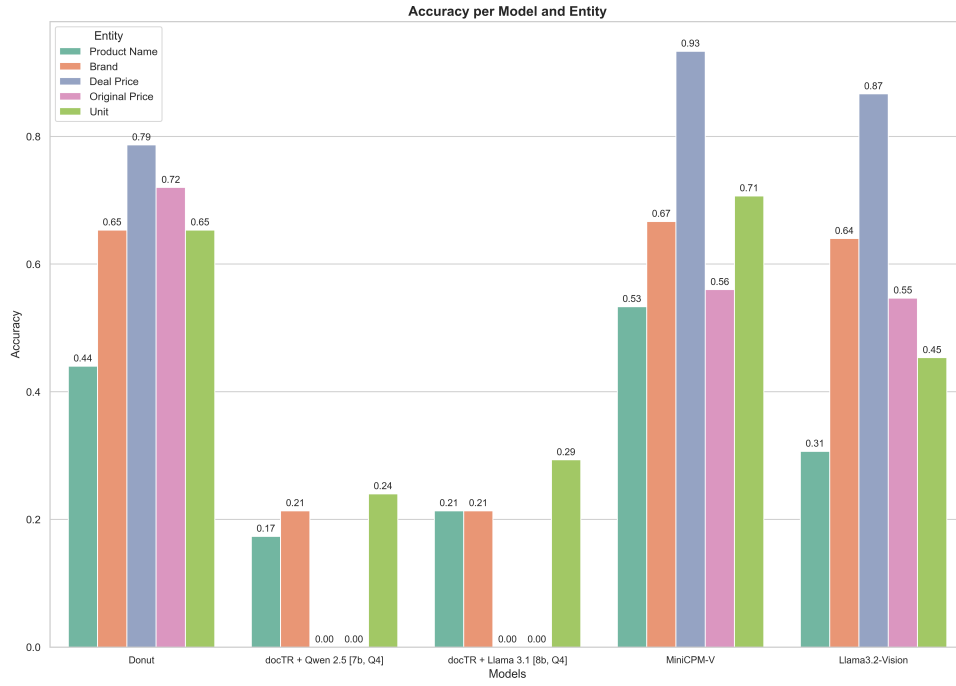


Figure 6: Per-Entity Accuracy without normalization.

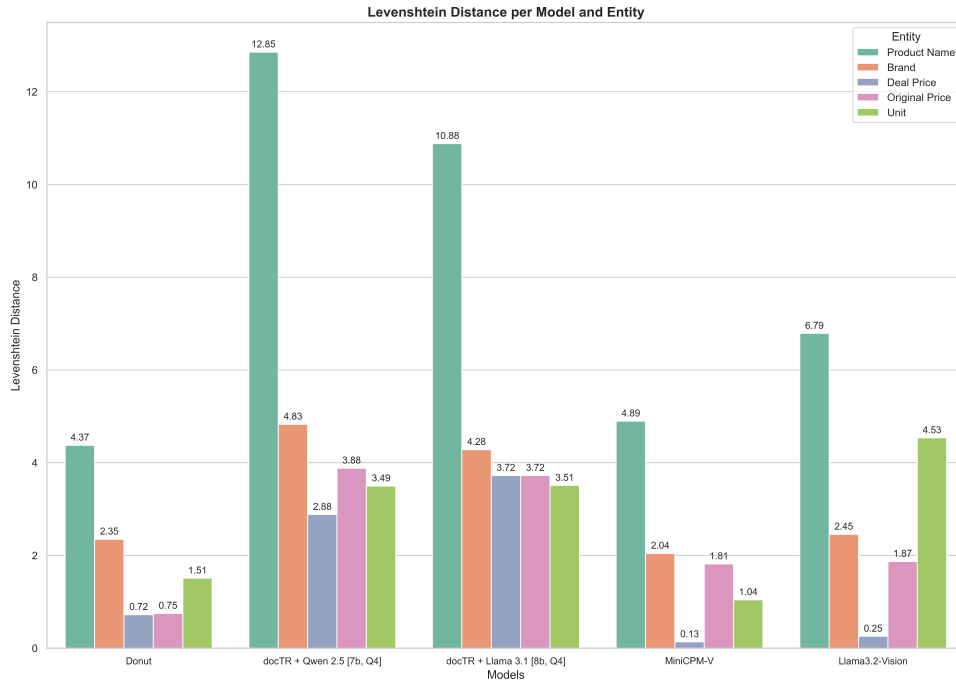


Figure 7: Per-Entity Levenshtein Distance without normalization.

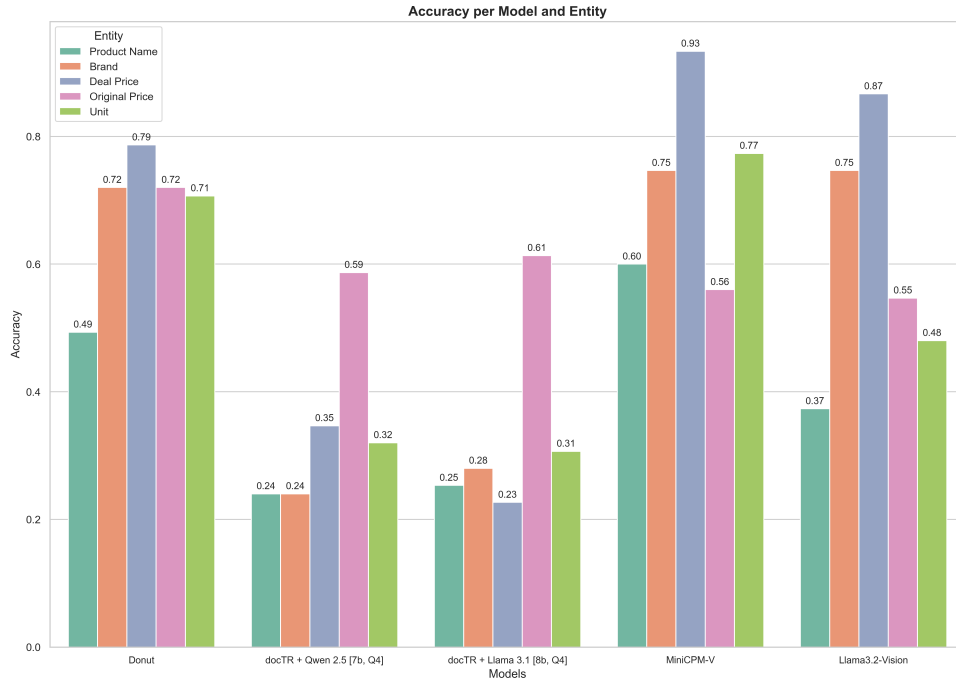


Figure 8: Per-Entity Accuracy with normalization.

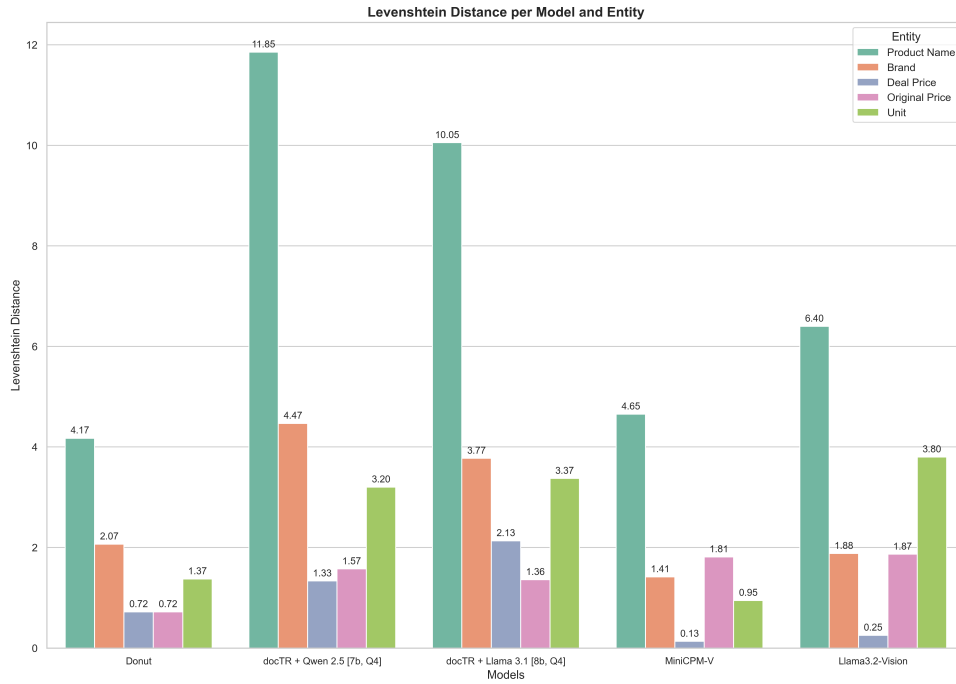


Figure 9: Per-Entity Levenshtein Distance with normalization.