scientific reports

OPEN

Large language model driven transferable key information extraction mechanism for nonstandardized tables

Rong Hu[1,6], Ye Yang[2,6]✉, Sen Liu[2,6], Zuchen Li[3], Jingyi Liu[4], Xingchen Ding[2], Hanchi Sun[2] & Lingli Ren[5]

Extracting key information from unstructured tables poses significant challenges due to layout variability, dependence on large annotated datasets, and inability of existing methods to directly output structured formats like JSON. These limitations hinder scalability and generalization to unseen document formats. We propose the Large Language Model Driven Transferable Key Information Extraction Mechanism (LLM-TKIE), which employs text detection to identify relevant regions in document images, followed by text recognition to extract content. An LLM then performs semantic reasoning, including completeness verification and key information extraction, before organizing data into structured formats. Without fine-tuning, LLM-TKIE achieves an F1-score of 80.9 and tree edit distance-based accuracy of 88.85 on CORD, and an F1-score of 83.9 with 93.3 accuracy on SROIE, demonstrating robust generalization and structural precision. Notably, our method significantly outperforms state-of-the-art multimodal large models on unlabeled customs domain datasets by 5–8% in accuracy. Additionally, our evaluation of multiple large language models of various sizes across 15 quantization strategies provides valuable insights for selecting and optimizing LLMs for key information extraction tasks, offering practical guidance for system development.

Customs documents in international trade are often non-standardized and structurally complex, encompassing key forms such as declarations, invoices, and packing lists. As shown in Fig. 1, unlike structured tables, these documents exhibit significant variability in layout, field positioning, and textual organization, often including multi-line entries, irregular alignments, and nested information blocks. Such heterogeneity poses major challenges for rule-based extraction methods. Therefore, Key Information Extraction (KIE) plays a crucial role in automatically identifying essential entities—such as product descriptions, quantities, and HS codes—thereby supporting efficient customs clearance, automated verification, and intelligent regulatory oversight.

Existing approaches for KIE from non-standardized documents can be broadly classified into rule-based systems, deep learning-based models, and end-to-end frameworks. Rule-based methods, such as Intellix[1] and SmartFIX[2], are effective for structured layouts but fail to generalize to the variability and complexity of non-standardized documents in logistics and other industries. Deep learning-based models, including LayoutLM[3], LayoutLMv2[4], and DocFormer[5], improve accuracy by leveraging multimodal information but depend heavily on large annotated datasets, making them costly and impractical for real-world applications where labeled data is scarce. End-to-end frameworks such as OmniParser[6] and DeepSolo[7] aim to streamline extraction pipelines but typically produce word- or line-level outputs, which are insufficient for downstream automation systems that require structured outputs like JSON. These limitations hinder their deployment in industrial automation, where scalability, cost-efficiency, and adaptability to unseen document types are critical.

KIE from non-standardized tables presents three critical challenges: (1) the scarcity or complete absence of labeled training data, which restricts the adaptability of existing methods in low-resource settings[8]; (2) the inability to directly produce structured outputs, such as JSON, which are essential for downstream automation pipelines[3,9]; and (3) limited generalization capability when handling diverse and complex table formats due to inadequate semantic understanding[10,11]. To address these challenges, we propose the Large Language Model

[1]Customs and Public Management College, Shanghai Customs University, Shanghai 201204, China. [2]School of Electronic Information, Shanghai DianJi University, Shanghai 201306, China. [3]School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China. [4]School of Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China. [5]Qingdao Port International Co., Ltd., Qingdao 266011, China. [6]Rong Hu, Ye Yang and Sen Liu contributed equally to this work. ✉email: yangye@st.sdju.edu.cn

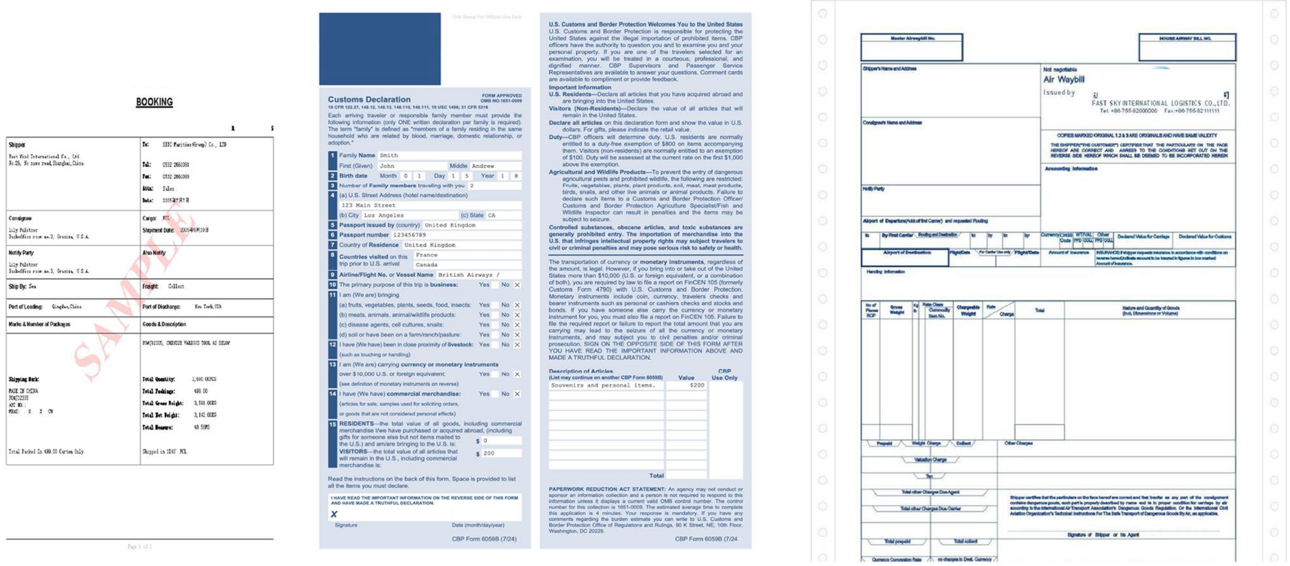


Fig. 1. Examples of non-standardized logistics documents: a shipping manifest (left), a customs declaration (center), and a bill of lading (right). These documents highlight diverse layouts and irregular field placements, posing challenges for automated extraction.

Driven Transferable Key Information Extraction Mechanism (LLM-TKIE). LLM-TKIE integrates Optical Character Recognition (OCR) with the semantic reasoning capabilities of Large Language Models (LLMs) to bridge the gap between unstructured text extraction and structured data generation. Specifically, the mechanism employs OCR for accurate text detection and recognition, followed by LLMs to interpret the extracted text, consolidate multi-line information, and transform it into structured formats, such as JSON. By leveraging pre-trained LLMs, LLM-TKIE significantly reduces dependency on annotated datasets, enhances semantic understanding, and improves adaptability to diverse document layouts[12]. This mechanism provides a robust, scalable, and accurate solution to the key challenges in KIE, demonstrating significant improvements over existing methods in real-world industrial applications.

The primary contributions of this work are summarized as follows:

- **A transferable mechanism for key information extraction from non-standardized tables based on large language models:** We propose a novel mechanism that integrates OCR with LLMs to effectively address the challenges of diverse and non-standardized document layouts. This mechanism demonstrates the ability to adapt to new table formats with as few as three annotated examples, enabling direct generation of structured outputs, such as JSON, while overcoming the generalization and scalability limitations of traditional KIE systems.
- **Competitive performance without training:** Our mechanism achieves an F1-score of 80.9 and a similarity score based on tree edit distance of 88.85 on the CORD[13] dataset, and an F1-score of 83.9 and a similarity score of 93.3 on the SROIE[14] dataset, demonstrating strong performance even without additional task-specific training.
- **Comprehensive evaluation of LLMs:** We conduct an extensive study of 85 open-source pre-trained language models, with parameter sizes ranging from 0.5 billion to 72 billion, across 15 quantization strategies. This provides valuable insights into the performance of LLMs in KIE tasks, offering practical guidelines for selecting and optimizing models in resource-constrained environments.

Related work

KIE from non-standardized tables is a crucial task in industrial automation, enabling streamlined workflows across domains such as logistics, finance, and manufacturing. Existing research in this domain can be categorized into three main areas: text detection and recognition, KIE models, and LLMs.

Text detection and recognition are critical components of document analysis, particularly for complex layouts in non-standardized tables. Traditional methods, such as sliding window and connected component analysis, have been largely supplanted by deep learning-based approaches, with architectures like EAST[15] and DBNet[16] improving boundary precision and efficiency through direct bounding box regression and differentiable binarization. Recent advancements incorporate progressive region prediction, asymmetric center positioning, and kernel expansion to enhance detection accuracy in challenging scenarios[17–20]. In text recognition, Transformer-based models such as TrOCR[21] and MAGIC[22], along with lightweight systems like PaddleOCR[23], have improved multilingual and unstructured text processing. Further enhancements include occlusion-aware recognition via graph recurrent networks[24], color-aware detection with multi-channel MSER[25], and specialized frameworks for Tibetan text using cross-sequence reasoning[26]. Additionally, query-aware Transformers have been applied for text super-resolution[27], while multi-modal knowledge transfer has been explored for few-shot text generation and object detection[28]. These developments collectively advance the robustness and adaptability of document analysis, addressing challenges in diverse and complex real-world environments.

KIE models aim to extract structured information from both structured and non-standardized tables, facilitating downstream automation workflows. Common benchmark datasets include the Consolidated Receipt Dataset (CORD)[29] and the Scanned Receipt OCR and Information Extraction dataset (SROIE)[30], which are widely used for evaluating receipt-based KIE tasks. Early rule-based approaches, while effective for predefined templates, lack adaptability to diverse layouts[6]. Traditional deep learning-based methods, including LSTM-based sequence models, multimodal pretrained models like LayoutLM[3], and end-to-end systems like OmniParser[6], face three fundamental limitations. First, these approaches heavily depend on large-scale annotated datasets for training, requiring detailed annotations specifying the position, category, and context of each word or field in a table, making the process highly labor-intensive, time-consuming, and error-prone, particularly for non-standardized tables where high-quality labeled data is difficult to obtain due to layout variability. Second, traditional KIE methods exhibit poor generalization when applied to unseen table layouts, as they rely on predefined structural patterns learned during training, which limits their ability to adapt to diverse and irregular table formats commonly found in real-world scenarios. Third, these methods are fundamentally limited by their reliance on token-level classification, predicting labels for individual tokens or words in the input text and resulting in outputs that are sequences of token-level predictions, which inherently lack the capability to produce structured outputs such as JSON or XML that are essential for real-world downstream automation tasks, requiring additional error-prone post-processing steps. Transformer-based models, such as LayoutLM[3], integrate textual and layout information, improving extraction accuracy, while end-to-end frameworks like OmniParser[7] and DeepSolo[31] unify detection, recognition, and extraction tasks, enhancing pipeline efficiency. Despite these advancements, challenges persist in cross-domain generalization and reliance on extensive labeled datasets, which hinder their applicability in dynamic and diverse industrial contexts.

LLMs have demonstrated remarkable adaptability in NLP[32,33], particularly in KIE under low-resource settings. In this study, we experiment with a wide range of LLMs, including open-source models such as Qwen, Gemma, Glm, Deepseek and Mistral, covering parameter scales from 0.5B to 72B. These models were selected for their balance between inference efficiency and generalization performance in document-level understanding tasks. These models leverage prompting techniques to enhance generalization[34] while integrating seamlessly with OCR systems[12,35]. Despite their strengths, challenges such as computational overhead and limited domain-specific adaptation persist[12,36]. Beyond NLP, LLMs have been instrumental in industrial automation, facilitating anomaly detection, task adaptability, and decision-making in unstructured environments[37–39]. In robotics, LLM-driven frameworks have been applied for hand exoskeleton control, robotic disassembly, and semantic encoding in aviation safety analysis[37,40,41]. In cybersecurity, LLM-based solutions have significantly improved phishing detection and logical anomaly identification[39,42]. The financial sector has benefited from LLM-powered information retrieval for enhanced decision-making[43,44], while voice assistant technologies have advanced human–AI interaction through LLM integration[45]. Healthcare applications include medical question answering via knowledge subgraphs, Traditional Chinese Medicine diagnostics with AcupunctureGPT, and improved taxonomy-driven entity recognition for manufacturing processes[46–48]. In education, LLMs have been utilized for interactive teacher training simulations, mitigating suspension of disbelief[49]. Additionally, they have been employed in table-to-text generation with guided planning to reduce hallucination, domain-specific dense retrieval via soft prompt tuning, and collaborative small-large model fusion for event detection in low-resource settings[50–52]. These advancements highlight the widespread applicability of LLMs in diverse fields, including robotics, cybersecurity, healthcare, education, and manufacturing.

Recent efforts have explored combining LLMs with structured knowledge representations or visual-semantic understanding for enhanced information extraction. In the medical domain, Huang et al.[53] proposed a knowledge graph (KG)-based question answering method that effectively integrates domain-specific entities to improve structured retrieval performance. Similarly, Thomas and Sangeetha[54] developed a KG-powered QA system tailored to legal document analysis, capable of modeling nested entity relations in case law. Extending this line of work, Giarelis et al.[55] presented a unified LLM-KG framework for fact-checking in public deliberation scenarios, demonstrating the synergy between pre-trained LLMs and symbolic graph reasoning.

Meanwhile, advances in visual question answering (VQA) systems offer valuable insights for document image understanding, especially in layout- and field-aware modeling. Chowdhury and Soni have introduced a series of robust VQA frameworks to address key limitations such as language prior and compositional reasoning. Their recent R-VQA model[56] integrates unified reasoning mechanisms to tackle both challenges simultaneously, while ESC-Net[57] improves visual understanding via spatial-channel attention ensembles. Further improvements are demonstrated in ENVQA[58], which enriches visual features with object-level reasoning, and QSF-VQA[59], a time-efficient and scalable inference framework. A comprehensive summary of these contributions appears in[60], highlighting strategies to mitigate reasoning failures and enhance system robustness in VQA contexts. These insights collectively inform the development of our proposed LLM-TKIE system for semi-structured document understanding.

Preliminaries

This section introduces a mechanism for extracting structured information from images of non-standardized tables. The mechanism combines OCR with LLMs to interpret the unstructured data and generate structured outputs like JSON. The following subsections describe the problem definition, the design rationale, and the system architecture.

Problem definition

The extraction of structured information from non-standardized tables represents a critical bottleneck in industrial automation workflows. Documents such as invoices, receipts, shipping manifests, and customs declarations play a vital role in domains like logistics, finance, and healthcare, yet they lack consistent layouts and well-defined structures. This variability results in key information being distributed across unpredictable and irregular formats, posing significant challenges for automated processing.

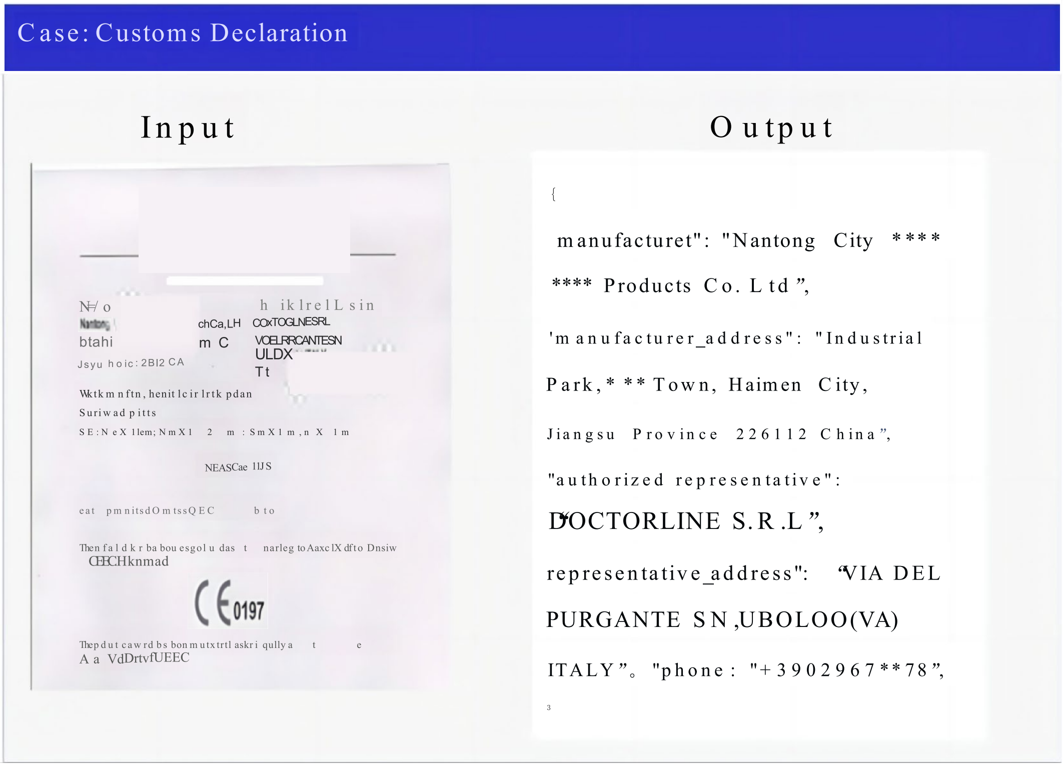
Despite advances in OCR and KIE techniques, three major challenges persist: (1) Data dependency: Existing methods heavily rely on large annotated datasets, which are expensive and time-consuming to obtain, limiting their scalability in real-world applications. (2) Generalization limitations: Current systems struggle to adapt to unseen document layouts or diverse domain-specific formats, reducing their reliability across industries. (3) Structured output generation: Most approaches fail to produce structured outputs, such as JSON or XML, which are essential for seamless integration into automation pipelines.

To address these challenges, we define the problem as follows: given a non-standardized table in a document (e.g., an invoice or receipt), the goal is to accurately extract key information (e.g., company name, date, and total amount) and output it in a structured format that is suitable for downstream applications. This requires a system that not only detects and recognizes textual content but also interprets its semantic and contextual relationships to enable robust generalization across varying layouts.

As illustrated in Fig. 2, the proposed **LLM-TKIE** addresses these challenges by integrating OCR with the reasoning capabilities of LLMs. LLM-TKIE mechanism enables accurate extraction and structured data generation with minimal reliance on annotated datasets, making it a scalable and adaptable solution for automating document processing in diverse industrial contexts.

Design rationale

The proposed **LLM-TKIE** mechanism is designed around a fundamental architectural paradigm that leverages modular decomposition and pre-trained model synergy. The core design philosophy centers on task decoupling, where visual processing and semantic understanding are deliberately separated to maximize the utilization of existing pre-trained components while minimizing domain-specific training requirements. By combining an OCR pipeline with pre-trained `DBNet++` and `SVTR_LCNet` for efficient text line detection and recognition, and leveraging the semantic reasoning and generative capabilities of LLMs, LLM-TKIE creates a robust framework that transforms heterogeneous visual layouts into structured data through intelligent prompt-driven processing. This modular architecture enables independent optimization of each component while maintaining system-level coherence through carefully designed interfaces and validation mechanisms, making LLM-TKIE a scalable solution for addressing the complexities of non-standardized tables in practical industrial scenarios.



```
{
 manufacturet": "Nantong  City  ****
**** Products  Co. Ltd ”,
'manufacturer_address": "Industrial
Park,* ** Town, Haimen  City,
Jiangsu  Province  226112  China”,
"authorized representative":
D̸OCTORLINE S.R.L”,
representative_address":   “VIA DEL
PURGANTE SN,UBOLOO(VA)
ITALY”。"phone: "+3902967**78”,
3
```

Fig. 2. Case study of receipt information extraction using LLM-TKIE. (Sensitive information has been redacted for privacy protection).

Firstly, LLM-TKIE addresses the data dependency challenge through strategic modular decomposition and knowledge transfer mechanisms. Our proposed mechanism integrates an OCR pipeline consisting of `DBNet++` for text line detection and `SVTR_LCNet` for text line recognition, fundamentally transforming the visual understanding problem into a pure text processing task. By focusing on detecting and recognizing text lines instead of analyzing entire table structures, LLM-TKIE avoids the need for detailed table annotations and leverages the robust capabilities of pre-trained OCR models. The subsequent semantic processing employs LLMs such as GPT-3[12] and LLaMA[35], which utilize their extensive pretraining on large-scale unlabeled text to interpret table semantics and extract key information through prompt-based learning. This design paradigm shifts from data-intensive supervised training to knowledge transfer through carefully crafted prompts, enabling the system to generate structured outputs such as JSON with minimal reliance on labeled datasets and significantly reducing the data dependency bottleneck for scalable and efficient key information extraction.

Secondly, LLM-TKIE enhances generalization capability through layout-agnostic processing and few-shot adaptation strategies. The system bypasses the need for table structure analysis by employing pre-trained `DBNet++` for text detection and `SVTR_LCNet` for text recognition, creating a pipeline that focuses exclusively on text line detection and recognition while decoupling the task from specific table layouts. This approach transforms diverse and irregular table formats into normalized textual representations, effectively reducing layout-specific variability and enhancing adaptability to heterogeneous document structures. The recognized text is subsequently processed by an LLM using few-shot prompting mechanisms, where 1-3 annotated examples provide sufficient contextual guidance for the model to understand task-specific requirements and field mapping relationships. This design enables LLM-TKIE to achieve strong generalization across varying table formats and domains by leveraging the LLM's pre-trained knowledge of semantic relationships and linguistic patterns, ensuring reliable performance in practical scenarios without requiring extensive layout-specific training.

Finally, LLM-TKIE enables direct structured output generation through constrained generative modeling and validation-driven refinement mechanisms. We leverage the generative capabilities of LLMs to directly produce structured outputs through prompt-based learning, where carefully designed task-specific prompts guide the LLM to generate key information in predefined structured formats such as JSON within its generative process. This approach incorporates explicit format specifications, JSON templates, and few-shot examples within the prompt design to constrain the generation process and ensure structural compliance. To further enhance robustness and consistency, we employ multi-stage regularization techniques including completeness verification, format validation through regular expressions, and iterative refinement through re-prompting mechanisms that validate and extract JSON strings from the generated output, eliminating errors and ambiguities. This validation-driven approach enables our proposed LLM-TKIE mechanism to bypass the token-level classification paradigm and produce structured outputs that can be seamlessly integrated into real-world industrial workflows, significantly improving both efficiency and reliability in key information extraction.

System overview and architecture

The proposed **LLM-TKIE** mechanism integrates OCR with LLMs to extract key information from non-standardized tables and generate structured outputs. The system architecture is composed of three primary components: text detection, text recognition, and key information extraction with structured output generation.

In the first stage, **text detection**, the input images of non-standardized tables are processed using the pre-trained `DBNet++` model. This module identifies and localizes text regions by predicting bounding boxes represented as vertex coordinates. The focus on detecting text lines allows the system to isolate textual content without requiring detailed analysis of the overall table structure.

The second stage, **text recognition**, uses the `SVTR_LCNet` model to process the bounding boxes from the text detection module. Each detected text region is independently recognized, and the recognized text sequences are aggregated to form a unified text block representing the content of the table.

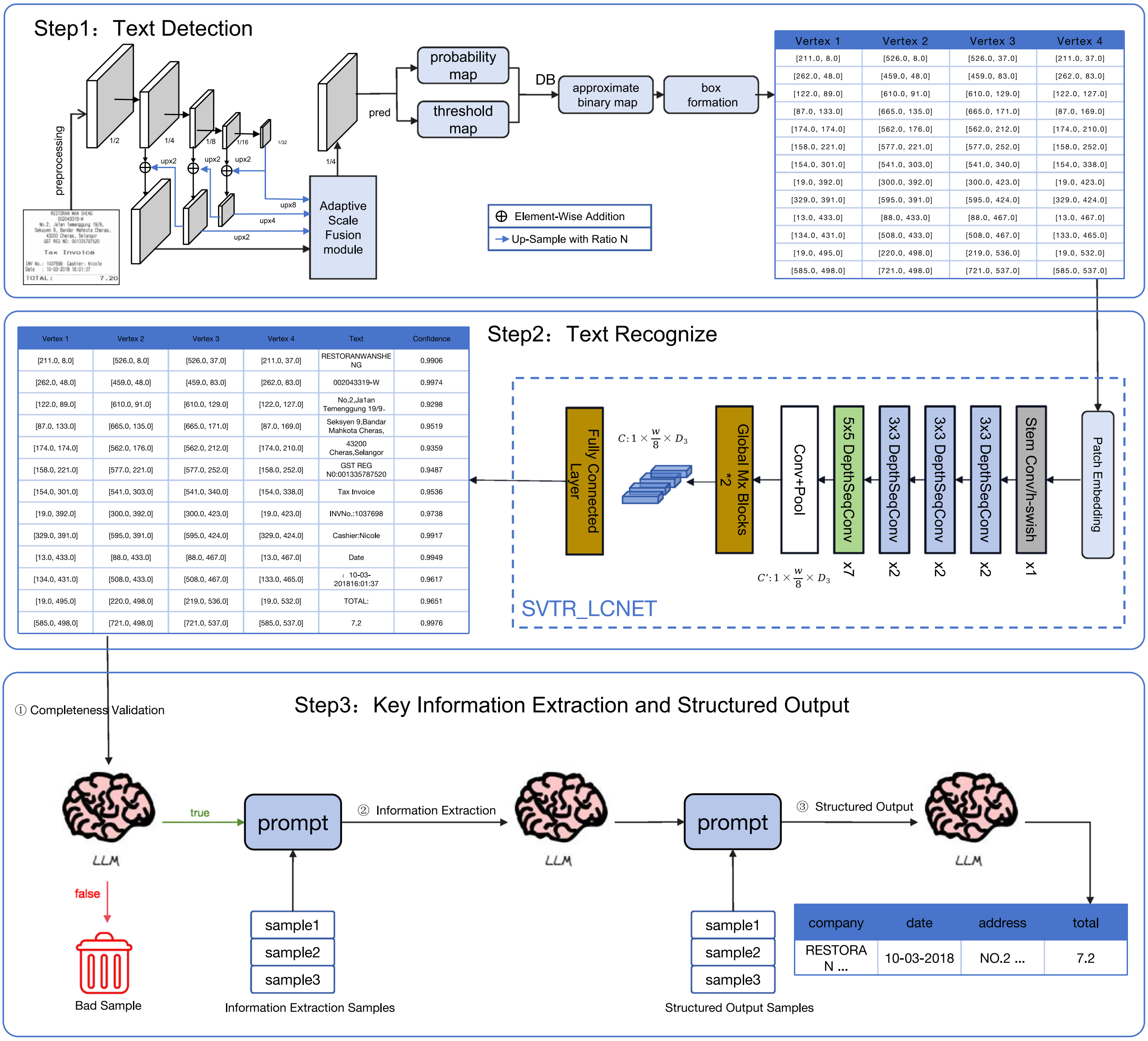
In the final stage, **key information extraction and structured output generation**, the aggregated text block is input into the LLM module for further processing. Initially, an information completeness check is conducted to determine whether the extracted text block contains sufficient data for downstream tasks. Incomplete samples are flagged for review, while complete samples proceed to the information extraction step. Using few-shot prompting techniques, the LLM extracts specified fields and generates structured outputs, such as JSON. Regularization techniques are applied to validate and refine the structured output to ensure consistency and accuracy. Figure 3 provides an overview of the system workflow and architecture, illustrating the interaction between the text detection, text recognition, and key information extraction modules.

Method

This chapter introduces the LLM-TKIE mechanism, an end-to-end pipeline designed to address the challenges of KIE from non-standardized tables. The mechanism consists of three core stages: **Text Detection**, which identifies and localizes text regions; **Text Recognition**, which extracts textual content; and **Key Information Extraction and Structured Output Generation**, which employs large language models to extract key fields and generate structured outputs such as JSON.

Text detection

The text detection stage in the LLM-TKIE mechanism identifies and localizes text regions within non-standardized table images, consisting of preprocessing, detection modeling, and post-processing.First, image preprocessing standardizes input images by removing alpha channels, retaining only RGB channels, and normalizing pixel values to mitigate lighting variations and noise:



RESTORAN WAN SHENG
002043319-W
No.2, Jalan Temenggung 19/9,
Seksyen 9, Bandar Mahkota Cheras,
43200 Cheras, Selangor
GST REG NO: 001335787520

Tax Invoice

INV No.: 1037698 Cashier: Nicole
Date : 10-03-2018 16:01:37
TOTAL: 7.20

Vertex 1	Vertex 2	Vertex 3	Vertex 4
[211.0, 8.0]	[526.0, 8.0]	[526.0, 37.0]	[211.0, 37.0]
[262.0, 48.0]	[459.0, 48.0]	[459.0, 83.0]	[262.0, 83.0]
[122.0, 89.0]	[610.0, 91.0]	[610.0, 129.0]	[122.0, 127.0]
[87.0, 133.0]	[665.0, 135.0]	[665.0, 171.0]	[87.0, 169.0]
[174.0, 174.0]	[562.0, 176.0]	[562.0, 212.0]	[174.0, 210.0]
[158.0, 221.0]	[577.0, 221.0]	[577.0, 252.0]	[158.0, 252.0]
[154.0, 301.0]	[541.0, 303.0]	[541.0, 340.0]	[154.0, 338.0]
[19.0, 392.0]	[300.0, 392.0]	[300.0, 423.0]	[19.0, 423.0]
[329.0, 391.0]	[595.0, 391.0]	[595.0, 424.0]	[329.0, 424.0]
[13.0, 433.0]	[88.0, 433.0]	[88.0, 467.0]	[13.0, 467.0]
[134.0, 431.0]	[508.0, 433.0]	[508.0, 467.0]	[133.0, 465.0]
[19.0, 495.0]	[220.0, 498.0]	[219.0, 536.0]	[19.0, 532.0]
[585.0, 498.0]	[721.0, 498.0]	[721.0, 537.0]	[585.0, 537.0]

Vertex 1	Vertex 2	Vertex 3	Vertex 4	Text	Confidence
[211.0, 8.0]	[526.0, 8.0]	[526.0, 37.0]	[211.0, 37.0]	RESTORANWANSHENG	0.9906
[262.0, 48.0]	[459.0, 48.0]	[459.0, 83.0]	[262.0, 83.0]	002043319-W	0.9974
[122.0, 89.0]	[610.0, 91.0]	[610.0, 129.0]	[122.0, 127.0]	No.2,Ja1an Temenggung 19/9，	0.9298
[87.0, 133.0]	[665.0, 135.0]	[665.0, 171.0]	[87.0, 169.0]	Seksyen 9,Bandar Mahkota Cheras,	0.9519
[174.0, 174.0]	[562.0, 176.0]	[562.0, 212.0]	[174.0, 210.0]	43200 Cheras,Selangor	0.9359
[158.0, 221.0]	[577.0, 221.0]	[577.0, 252.0]	[158.0, 252.0]	GST REG N0:001335787520	0.9487
[154.0, 301.0]	[541.0, 303.0]	[541.0, 340.0]	[154.0, 338.0]	Tax Invoice	0.9536
[19.0, 392.0]	[300.0, 392.0]	[300.0, 423.0]	[19.0, 423.0]	INVNo.:1037698	0.9738
[329.0, 391.0]	[595.0, 391.0]	[595.0, 424.0]	[329.0, 424.0]	Cashier:Nicole	0.9917
[13.0, 433.0]	[88.0, 433.0]	[88.0, 467.0]	[13.0, 467.0]	Date	0.9949
[134.0, 431.0]	[508.0, 433.0]	[508.0, 467.0]	[133.0, 465.0]	： 10-03-201816:01:37	0.9617
[19.0, 495.0]	[220.0, 498.0]	[219.0, 536.0]	[19.0, 532.0]	TOTAL:	0.9651
[585.0, 498.0]	[721.0, 498.0]	[721.0, 537.0]	[585.0, 537.0]	7.2	0.9976

company	date	address	total
RESTORA N ...	10-03-2018	NO.2 ...	7.2

Fig. 3. System workflow and architecture of LLM-TKIE for KIE from non-standardized tables.

$$\text{Normalized\_Pixel} = \frac{\text{Pixel\_Value} - \mu}{\sigma}, \tag{1}$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$. Images are resized to a standard resolution while preserving aspect ratios to avoid distortion. Additionally, for samples flagged by the Information Completeness Check as incomplete or incorrectly oriented, a rotation handling procedure is applied, systematically rotating the images by 90°, 180°, or 270° and reprocessing them. This effectively addresses orientation issues commonly encountered in real-world document images.

Next, the DBNet++ model detects text regions using a ResNet-50 backbone to extract multi-scale features, which are fused into a unified representation by the Adaptive Scale Fusion (ASF) module. The model outputs a probability map (P), indicating the likelihood of each pixel belonging to a text region, and an adaptive threshold map (T) used for binarization. A binary segmentation map is then obtained:

$$B(p) = \begin{cases} 1, & \text{if } P(p) > T(p), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

During training, a differentiable approximation enables gradient propagation to refine boundary detection accuracy.

Finally, connected component analysis groups adjacent pixels into distinct text regions. Bounding boxes around these regions are generated using the Minimum Area Bounding Rectangle Algorithm, effectively accommodating irregular and curved text layouts. Each bounding box (B_i) is assigned a confidence score based on enclosed pixel probabilities:

$$\mathrm{Conf}(B_i) = \frac{1}{|B_i|} \sum_{p \in B_i} P(p), \tag{3}$$

where $|B_i|$ represents the number of pixels within B_i. Bounding boxes with confidence scores below 0.9 are discarded to reduce false positives.The resulting set of high-confidence bounding boxes serves as input to the subsequent text recognition module, facilitating accurate key information extraction from complex table structures.

Text recognition

The text recognition stage in the LLM-TKIE mechanism extracts meaningful textual content from detected bounding boxes. Each localized text region is sequentially processed to ensure accurate recognition and robust adaptation to diverse table layouts. This stage converts detected text into a structured text block, which serves as input for subsequent key information extraction.

For recognition, we employ the `SVTR_LCNet` model, a lightweight yet effective architecture integrating convolutional and transformer-based components. This hybrid design enhances robustness against variations in font styles, orientations, and text lengths while maintaining computational efficiency. The model is applied to cropped and resized text regions, generating both character sequences and confidence scores.

To refine recognition results, we adopt a Connectionist Temporal Classification (CTC) decoding mechanism, which eliminates redundant characters and aligns input-output sequences without requiring strict positional correspondence. A confidence threshold ($\tau = 0.9$) is applied to discard low-confidence predictions, mitigating recognition errors. Finally, the recognized sequences are aggregated while preserving the spatial order of the bounding boxes, ensuring logical consistency across different table layouts.This recognition process bridges the gap between text detection and key information extraction, providing a structured textual representation for downstream processing.

Key information extraction and structured output generation

Information completeness check

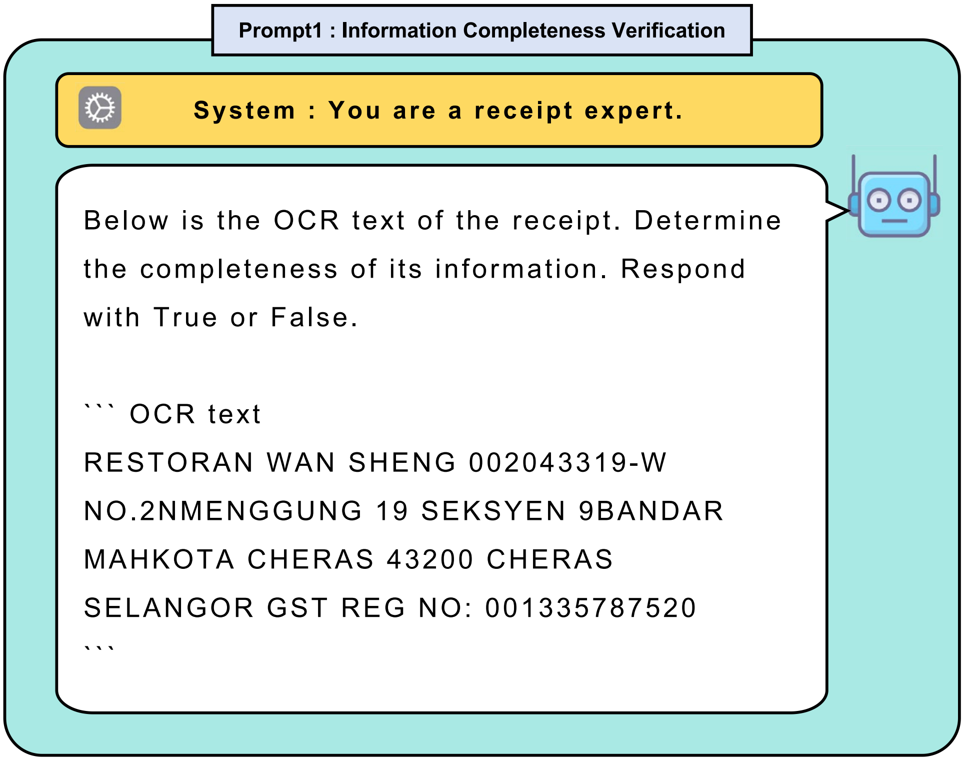
The completeness validation stage ensures that the output of the **Text Recognition** module, referred to as the *OCR Result*, includes all essential fields required for downstream KIE. Here, the *OCR Result* specifically refers to the unified text block generated by aggregating recognized text from all bounding boxes detected during the **Text Detection** and subsequently processed by the **Text Recognition** stage. This validation stage is critical for identifying incomplete or erroneous OCR results that may negatively impact the accuracy of subsequent processing.

To validate the completeness of the *OCR Result*, the system employs **Prompt1** (see Fig. 4), which is designed to verify the presence of key fields such as company name, date, and total amount. The prompt leverages a few-shot learning mechanism by providing examples of both *True* (complete) and *False* (incomplete) cases. This configuration enables the Large Language Model (LLM) to learn the task-specific requirements and generalize effectively across diverse document layouts and formats. The few-shot approach is particularly advantageous in scenarios with minimal labeled data, allowing the system to maintain high validation accuracy even in complex, non-standardized table formats.

When a sample is classified as **"False"** (incomplete) by the LLM, it is flagged as a *bad sample* and stored in a *Bad Sample Repository* for reprocessing. In real-world scenarios, incomplete results may arise due to poor image quality, missing fields, or incorrect OCR detection caused by document rotation (e.g., 90°, 180°, or 270°). To address such issues, flagged samples undergo a systematic rotation pipeline. Specifically, the samples are rotated by 90°, 180°, and 270° before being reintroduced into the OCR pipeline. Each rotated version is reprocessed through text detection and recognition, followed by another round of completeness validation. If the rotated samples still fail validation, they are permanently stored in the *Bad Sample Repository*, indicating that the incompleteness likely stems from irrecoverable issues such as low image resolution or missing content in the original document.

The validation workflow is illustrated in Fig. 4, which showcases the design of **Prompt1**. This prompt evaluates the integrity of the *OCR Result* by verifying the presence of required fields and their structural correctness. The integration of this validation mechanism ensures that only high-quality and complete *OCR Results* proceed to the key information extraction stage, thus enhancing the overall robustness and reliability of the LLM-TKIE mechanism.

By implementing a targeted validation mechanism for detecting incomplete or erroneous OCR results, the system ensures that only verified text blocks are passed to the subsequent key information extraction phase. This validation step systematically addresses issues such as document rotation, noise, and varying image quality by introducing a rotation-based reprocessing strategy and leveraging large language models to assess information completeness. While not entirely eliminating the impact of poor-quality images or missing fields, this approach effectively filters problematic samples and identifies cases requiring further intervention. Such a process minimizes the propagation of errors to downstream stages, maintaining consistency and improving the overall robustness of the LLM-TKIE pipeline.



Prompt1 : Information Completeness Verification

System : You are a receipt expert.

Below is the OCR text of the receipt. Determine the completeness of its information. Respond with True or False.

``` OCR text
RESTORAN WAN SHENG 002043319-W
NO.2NMENGGUNG 19 SEKSYEN 9BANDAR
MAHKOTA CHERAS 43200 CHERAS
SELANGOR GST REG NO: 001335787520
```

Fig. 4. Prompt1 for verifying the completeness of OCR results. This design leverages few-shot learning to enable the LLM to generalize effectively across diverse document layouts by using examples of both “True” and “False” cases.

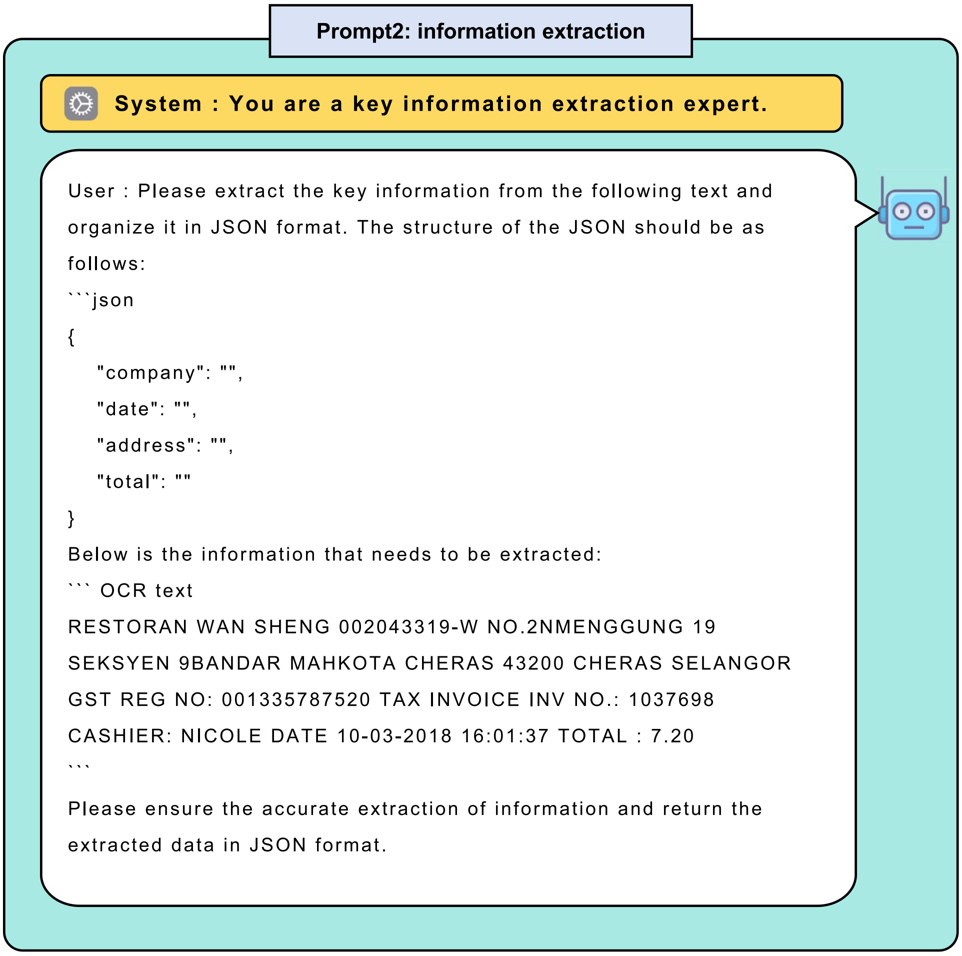
Few-shot prompting for key information extraction with large language models

The inclusion of both the JSON template and annotated examples in the prompt introduces a strong inductive bias that effectively guides the LLM to generate structured outputs directly from unstructured OCR text. This prompt is meticulously designed with four distinct components: the system prompt, the guiding instruction, the JSON template specifying the fields to be extracted, and the recognized OCR text serving as input for extraction. The combination of these elements creates a well-structured and intuitive format that directs the LLM’s attention to the essential aspects of the task, ensuring both accuracy and consistency in the generated outputs.

The system prompt, as shown in Fig. 5, explicitly frames the task for the LLM by defining its role as a key information extraction expert. This is followed by the guiding instruction, which provides clear directives for extracting the required fields and organizing them into a predefined JSON format. The JSON template further strengthens this mechanism by delineating the exact structure of the desired output, including fields such as “company,” “date,” “address,” and “total.” Finally, the OCR text input, derived from real-world documents, serves as the raw source of information to be processed.

This particular prompt design is based on examples from the SROIE dataset[14], a widely used dataset containing structured annotations of receipts. By tailoring the prompt to the format and content of SROIE data, the mechanism ensures that the LLM can generalize effectively to similar real-world scenarios. The annotated examples included in the prompt demonstrate the mappings between input text segments and JSON fields (e.g., “RESTORAN WAN SHENG” to the “company” field and “10-03-2018” to the “date” field), enabling the LLM to learn the semantic relationships necessary for extraction. By providing one to three annotated examples, the few-shot learning approach significantly enhances the LLM’s ability to generalize to unseen table layouts and document types with minimal reliance on extensive labeled datasets.

The structured combination of these components not only decouples the output schema from the positional arrangement of fields in the input but also minimizes the need for post-processing, as the LLM generates outputs directly in the required JSON format (see in Fig. 6). This design is particularly advantageous for handling diverse table layouts and variable content structures, ensuring robust performance across a wide range of applications. The prompt’s inherent transferability stems from its ability to generalize the extraction logic from a minimal set of examples. Specifically, by including just three annotated examples, the mechanism achieves remarkable performance even on table layouts and document structures it has not encountered before. This few-shot approach enables the LLM to quickly adapt to unseen formats, leveraging the semantic understanding demonstrated in the provided examples to extract the required fields accurately. Such transferability significantly reduces the dependency on extensive labeled datasets, making the mechanism highly scalable and practical for deployment across various industries where non-standardized table layouts are prevalent.



Prompt2: information extraction

System : You are a key information extraction expert.

~~~
User : Please extract the key information from the following text and
organize it in JSON format. The structure of the JSON should be as
follows:
```json
{
   "company": "",
   "date": "",
   "address": "",
   "total": ""
}
Below is the information that needs to be extracted:
``` OCR text
RESTORAN WAN SHENG 002043319-W NO.2NMENGGUNG 19
SEKSYEN 9BANDAR MAHKOTA CHERAS 43200 CHERAS SELANGOR
GST REG NO: 001335787520 TAX INVOICE INV NO.: 1037698
CASHIER: NICOLE DATE 10-03-2018 16:01:37 TOTAL : 7.20
```
Please ensure the accurate extraction of information and return the
extracted data in JSON format.
~~~

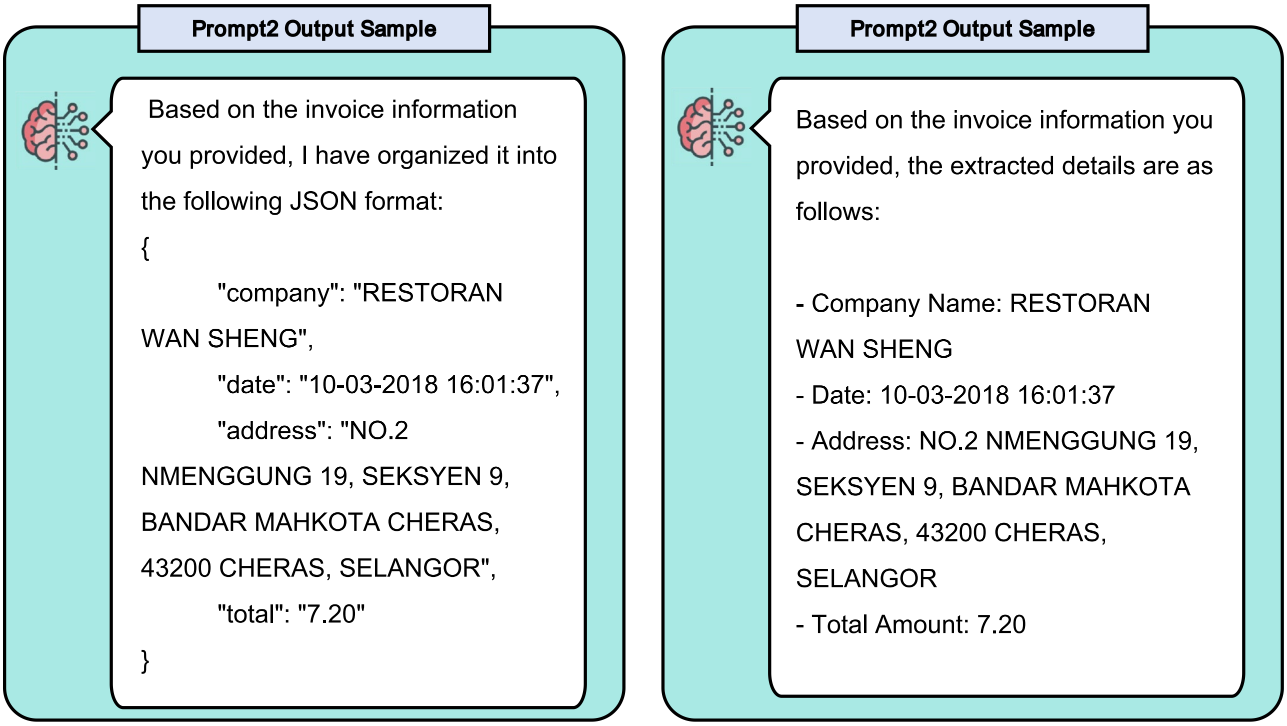
Fig. 5. Prompt2 for few-shot KIE.

Structured output generation

The structured output generation process, as shown in Algorithm 1, adopts a systematic two-step approach to ensure the reliability and correctness of the extracted data. The initial input, referred to as OCRResult, is processed using Prompt2 to produce a preliminary KIE result. This result approximates the desired JSON structure but may contain extraneous or malformed outputs. A regular expression-based method is then applied to validate the KIE result's adherence to the predefined JSON schema. If the validation succeeds, the structured JSON is directly extracted for downstream workflows. However, if the validation fails, the system invokes Prompt3 to enforce strict adherence to the JSON schema. Prompt3 takes the KIE result as input and regenerates the output, ensuring compliance with the required structure. The reprocessed output is subjected to a second round of validation, and only JSON-compliant data is retained.

The structured combination of these components not only decouples the output schema from the positional arrangement of fields in the input but also minimizes the need for post-processing, as the LLM generates outputs directly in the required JSON format (see in Fig. 6). This design is particularly advantageous for handling diverse table layouts and variable content structures, ensuring robust performance across a wide range of applications. The prompt's inherent transferability stems from its ability to generalize the extraction logic from a minimal set of examples. Specifically, by including just three annotated examples, the mechanism achieves remarkable performance even on table layouts and document structures it has not encountered before. This few-shot approach enables the LLM to quickly adapt to unseen formats, leveraging the semantic understanding demonstrated in the provided examples to extract the required fields accurately. Such transferability significantly reduces the dependency on extensive labeled datasets, making the mechanism highly scalable and practical for deployment across various industries where non-standardized table layouts are prevalent.

Algorithm 1 demonstrates the robustness of the structured output generation process, which combines regular expression-based validation with iterative refinement through re-prompting. The process begins by generating the Key Information Extraction (KIE) result from the OCRResult using Prompt2. This KIE result serves as an intermediary output, approximating the desired JSON structure. The first validation step employs a predefined regular expression to assess whether the KIE result adheres to the required JSON schema. If the KIE result is valid, it is directly passed to downstream processes.



Prompt2 Output Sample

Based on the invoice information you provided, I have organized it into the following JSON format:

```
{
        "company": "RESTORAN WAN SHENG",
        "date": "10-03-2018 16:01:37",
        "address": "NO.2 NMENGGUNG 19, SEKSYEN 9, BANDAR MAHKOTA CHERAS, 43200 CHERAS, SELANGOR",
        "total": "7.20"
}
```

Prompt2 Output Sample

Based on the invoice information you provided, the extracted details are as follows:

- Company Name: RESTORAN WAN SHENG
- Date: 10-03-2018 16:01:37
- Address: NO.2 NMENGGUNG 19, SEKSYEN 9, BANDAR MAHKOTA CHERAS, 43200 CHERAS, SELANGOR
- Total Amount: 7.20

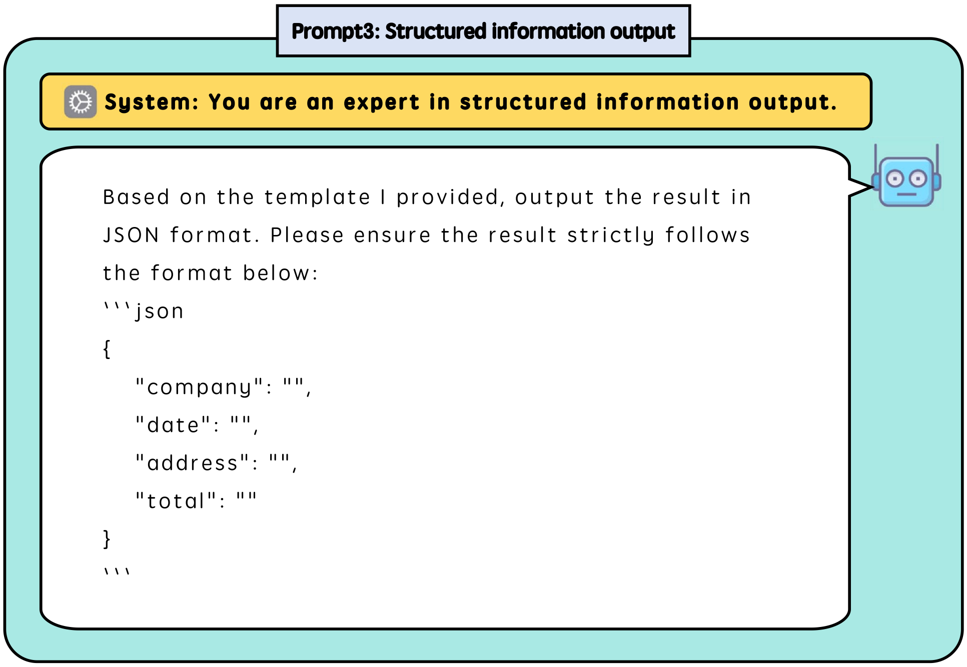
Fig. 6. JSON output generated during the key information extraction phase using LLM.

```
procedure GENERATESTRUCTUREDOUTPUT(OCRResult, Prompt2, Prompt3, JSONTemplate)
    Input: OCRResult (raw extracted text), Prompt2, Prompt3, JSONTemplate
    Output: Validated JSON Output
    Step 1: Generate the Key Information Extraction (KIE) result using Prompt2
    KIEResult ← CallLLM(Prompt2, OCRResult)
    Step 2: Validate the KIE result using a regular expression
    if RegexValidate(KIEResult, JSONTemplate) then
        return ExtractJSON(KIEResult)                          ▷ Directly extract validated JSON
    else
        Step 3: Re-prompt the LLM with Prompt3 for strict JSON compliance
        ReformattedOutput ← CallLLM(Prompt3, KIEResult)
        if RegexValidate(ReformattedOutput, JSONTemplate) then
            return ExtractJSON(ReformattedOutput)              ▷ Extract re-validated JSON
        else
            Raise Error: JSON formatting failed
        end if
    end if
end procedure
```

Algorithm 1. Structured output generation with validation and re-prompting.

In cases where the KIE result fails validation, Prompt3 (see in Fig. 7) is invoked to correct the output. Unlike Prompt2, which processes raw OCR results, Prompt3 is specifically designed to take the KIE result as input and ensure strict compliance with the JSON schema. The reprocessed output is then subjected to a second round of validation to confirm its adherence to the predefined schema. Only results that successfully pass this validation step are extracted as the final structured JSON output. If both validation steps fail, the system raises an error, signaling a failure to produce schema-compliant data.

This two-step validation and re-prompting mechanism effectively mitigates issues arising from malformed or extraneous outputs. By leveraging regular expressions for precise validation and re-prompting for iterative refinement, the proposed approach ensures the generation of high-quality, schema-compliant structured



Prompt3: Structured information output

System: You are an expert in structured information output.

Based on the template I provided, output the result in JSON format. Please ensure the result strictly follows the format below:

~~~
```json
{
   "company": "",
   "date": "",
   "address": "",
   "total": ""
}
```
~~~

Fig. 7. Prompt3 for ensuring structured output generation.

outputs. Furthermore, the integration of these mechanisms into the information extraction pipeline enhances the overall reliability and robustness of downstream workflows.

Results

The experiment evaluates the performance of LLM-TKIE in KIE from non-standardized tables without using task-specific training data. The model processes unseen document layouts using pre-trained knowledge and few-shot prompting, extracting structured information without fine-tuning. The evaluation includes various table structures with different text

Experimental setup

Experiments were conducted on a computational server equipped with four NVIDIA A100 80GB PCIe GPUs, an Intel Xeon Platinum 8375C CPU, and 256GB of RAM. The system operated on Ubuntu 22.04 LTS with Python 3.10. For models with fewer than 100 billion parameters, inference was performed on a single A100 GPU to optimize computational efficiency.

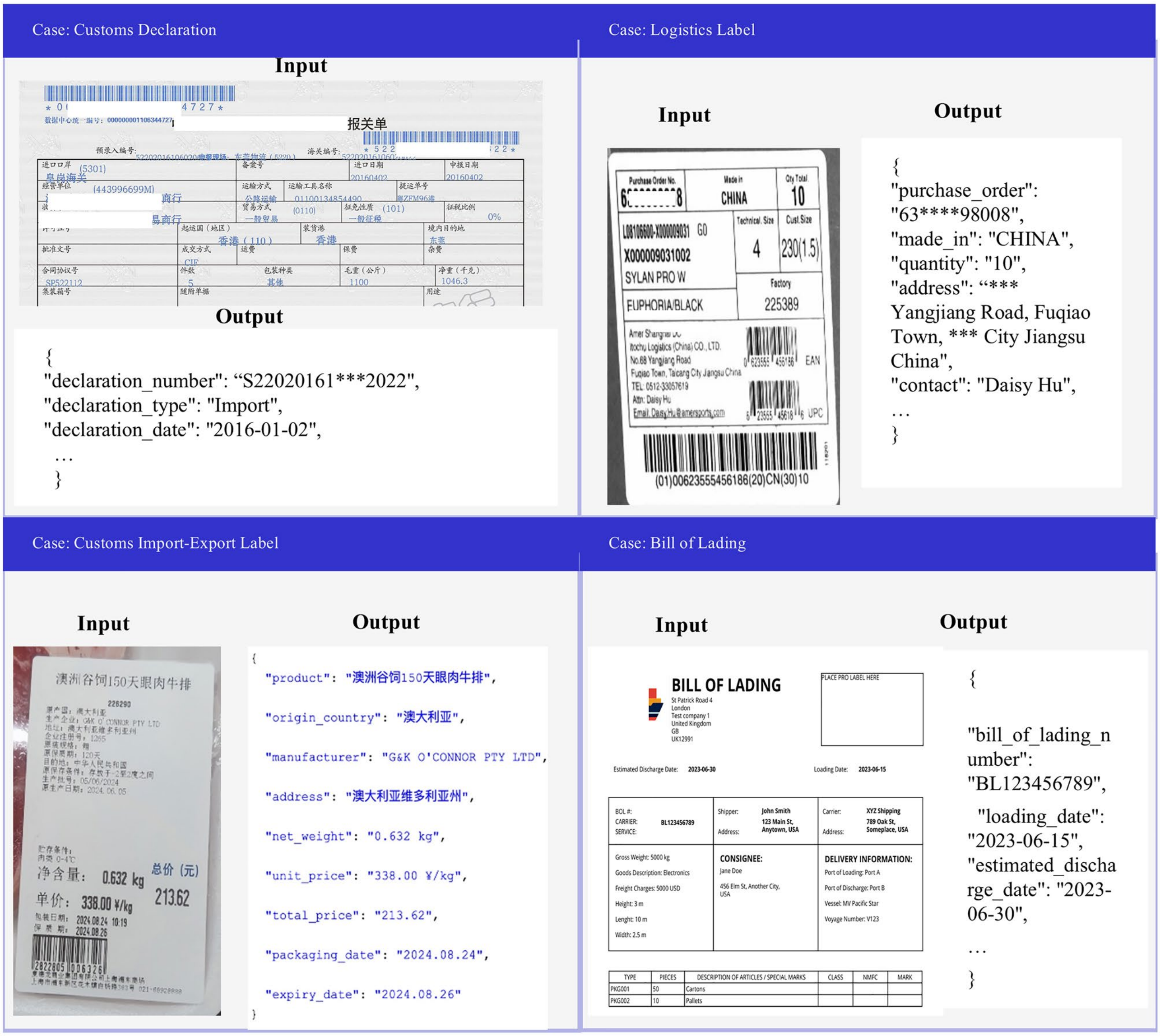
Text detection and recognition were performed using pre-trained PaddleOCR models. The text detection component employed DBNet++[61], while the text recognition utilized the SVTR_LCNet model from PPOCRv4[62]. These models were pre-trained on general-purpose datasets but were not fine-tuned on the evaluation datasets, ensuring an unbiased assessment.

The evaluation was conducted on the CORD (Consolidated Receipt Dataset)[13] and SROIE (Scanned Receipts OCR and Information Extraction)[14] datasets. SROIE focuses on extracting structured fields such as company names, dates, and invoice totals from scanned receipts, whereas CORD provides real-world receipts with hierarchical annotations. To assess generalization capability, the model was not fine-tuned on these datasets; instead, a few-shot learning approach was adopted using three sample documents, with testing performed on previously unseen document types.

The performance of the proposed key information extraction (KIE) framework was evaluated using three metrics: F1 Score, Tree-Edit Distance-Based Accuracy (TED-based Accuracy), and Time Efficiency. TED-based Accuracy (referred to as Acc in subsequent sections) quantifies the structural similarity between the predicted and ground-truth hierarchical representations. These metrics collectively assess the system's accuracy, structured output integrity, and computational efficiency, aligning with established evaluation protocols in KIE research[6].

Case study

To illustrate the effectiveness and generalizability of the proposed LLM-TKIE mechanism under few-shot conditions, we conducted case studies on four unseen, real-world document images(see in Fig. 8): a customs declaration, a logistics label, a customs import-export label and a bill of lading. Each scenario demonstrates the system's ability to accurately extract structured key information from non-standardized documents using minimal annotated examples.These examples collectively verify the LLM-TKIE mechanism's practical robustness, adaptability, and suitability for real-world automated information extraction applications in various industries.


Case: Customs Declaration
Input
报关单
Output
Case: Logistics Label
Input
Output
Case: Customs Import-Export Label
Input
Output
Case: Bill of Lading
Input
Output

```
{
"declaration_number": "S22020161***2022",
"declaration_type": "Import",
"declaration_date": "2016-01-02",
 …
}
```

Purchase Order No.	Made in	Qty Total
6[redacted]8	CHINA	10
L08106600-X000009031 G0 X000009031002 SYLAN PRO W	Technical. Size 4	Cust.Size 230(1.5)
EUPHORIA/BLACK	Factory 225389	

Amer Shanghai ...
Itochu Logistics (China) CO., LTD.
No.68 Yangjiang Road
Fuqiao Town, Taicang City Jiangsu China
TEL: 0512-33057619
Attn: Daisy Hu
Email: Daisy.Hu@amersports.com

0 623555 456186 EAN
6 23555 45618 6 UPC
(01)00623555456186(20)CN(30)10

```
{
"purchase_order": "63****98008",
"made_in": "CHINA",
"quantity": "10",
"address": "*** Yangjiang Road, Fuqiao Town, *** City Jiangsu China",
"contact": "Daisy Hu",
…
}
```

```
{
  "product": "澳洲谷饲150天眼肉牛排",
  "origin_country": "澳大利亚",
  "manufacturer": "G&K O'CONNOR PTY LTD",
  "address": "澳大利亚维多利亚州",
  "net_weight": "0.632 kg",
  "unit_price": "338.00 ¥/kg",
  "total_price": "213.62",
  "packaging_date": "2024.08.24",
  "expiry_date": "2024.08.26"
}
```

BILL OF LADING
St Patrick Road 4
London
Test company 1
United Kingdom
GB
UK12991

PLACE PRO LABEL HERE

Estimated Discharge Date: 2023-06-30   Loading Date: 2023-06-15

BOL #: CARRIER: BL123456789 SERVICE:	Shipper: John Smith Address: 123 Main St, Anytown, USA	Carrier: XYZ Shipping Address: 789 Oak St, Someplace, USA
Gross Weight: 5000 kg Goods Description: Electronics Freight Charges: 5000 USD Height: 3 m Lenght: 10 m Width: 2.5 m	CONSIGNEE: Jane Doe 456 Elm St, Another City, USA	DELIVERY INFORMATION: Port of Loading: Port A Port of Discharge: Port B Vessel: MV Pacific Star Voyage Number: V123

TYPE	PIECES	DESCRIPTION OF ARTICLES / SPECIAL MARKS	CLASS	NMFC	MARK
PKG001	50	Cartons			
PKG002	10	Pallets			

```
{
"bill_of_lading_number": "BL123456789",
 "loading_date": "2023-06-15",
"estimated_discharge_date": "2023-06-30",
…
}
```

Fig. 8. Four real-world customs scenarios. (Sensitive information has been redacted for privacy protection).

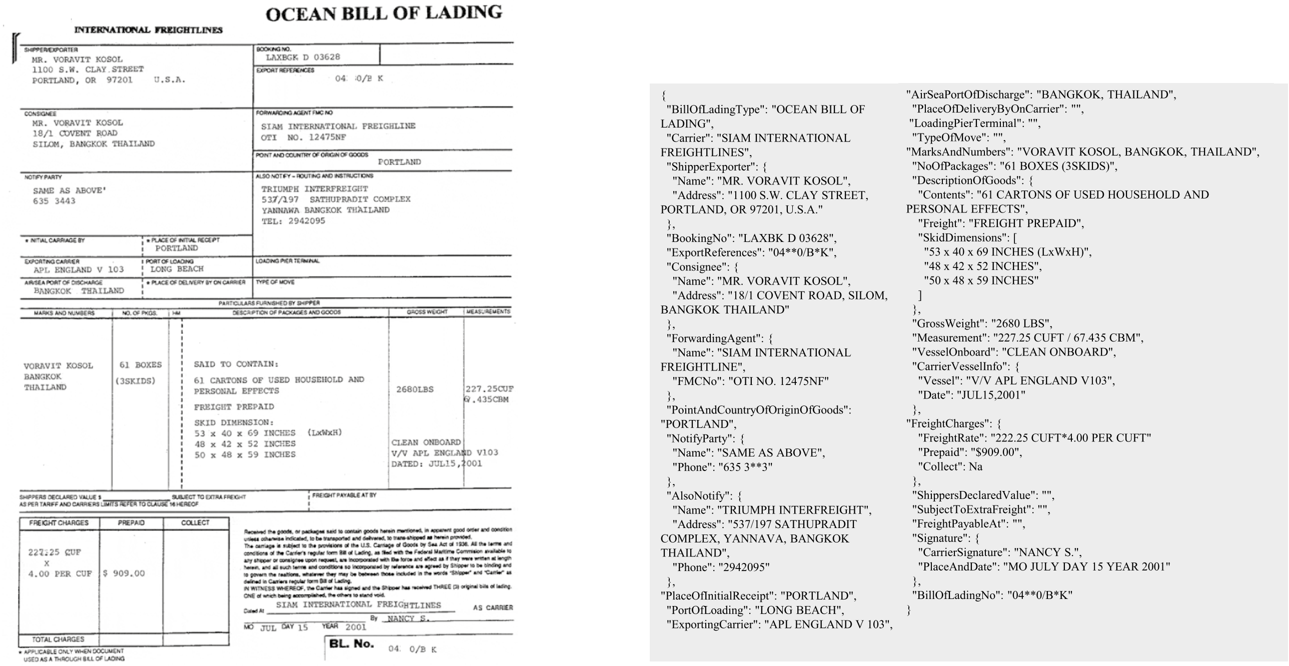
Case study on complex and incomplete tables

To further demonstrate the robustness and real-world applicability of LLM-TKIE, we present a representative example involving a complex and partially incomplete table structure (as shown in Fig. 9a). Figure 9 illustrates an actual Ocean Bill of Lading document with dense layout, field irregularities, and semi-structured content. It includes duplicate headers (e.g., *Portland* appearing multiple times), missing key-value alignments (e.g., blank *Freight Payable At* field), and embedded nested information (e.g., *Freight Charges* with units, rates, and amounts in a multi-row layout).

Despite these challenges, our system correctly parses over 20 distinct fields and preserves their hierarchical semantics, as shown in the structured JSON output (Fig. 9b). This case highlights LLM-TKIE's ability to:

- Infer semantic roles of text spans without relying on strict visual alignment.
- Handle duplicate, incomplete, or irregular fields through contextual understanding.
- Preserve field nesting (e.g., *Forwarding Agent*, *Carrier Vessel Info*) in the final structured format.

This case study affirms LLM-TKIE's suitability for real-world industrial documents, where non-standard layouts and incomplete structures are common. We believe that future enhancements can further improve robustness by incorporating visual-layout priors or adaptive self-refinement mechanisms.



INTERNATIONAL FREIGHTLINES

OCEAN BILL OF LADING

SHIPPER/EXPORTER
MR. VORAVIT KOSOL
1100 S.W. CLAY STREET
PORTLAND, OR 97201 U.S.A.

BOOKING NO.
LAXBGK D 03628

EXPORT REFERENCES
04 0/B K

CONSIGNEE
MR. VORAVIT KOSOL
18/1 COVENT ROAD
SILOM, BANGKOK THAILAND

FORWARDING AGENT FMC NO
SIAM INTERNATIONAL FREIGHLINE
OTI NO. 12475NF

POINT AND COUNTRY OF ORIGIN OF GOODS
PORTLAND

NOTIFY PARTY
SAME AS ABOVE'
635 3443

ALSO NOTIFY – ROUTING AND INSTRUCTIONS
TRIUMPH INTERFREIGHT
537/197 SATHUPRADIT COMPLEX
YANNAWA BANGKOK THAILAND
TEL: 2942095

* INITIAL CARRIAGE BY

* PLACE OF INITIAL RECEIPT
PORTLAND

EXPORTING CARRIER
APL ENGLAND V 103

PORT OF LOADING
LONG BEACH

LOADING PIER TERMINAL

AIR/SEA PORT OF DISCHARGE
BANGKOK THAILAND

* PLACE OF DELIVERY BY ON CARRIER

TYPE OF MOVE

PARTICULARS FURNISHED BY SHIPPER

MARKS AND NUMBERS	NO. OF PKGS.	HM	DESCRIPTION OF PACKAGES AND GOODS	GROSS WEIGHT	MEASUREMENTS
VORAVIT KOSOL BANGKOK THAILAND	61 BOXES (3SKIDS)		SAID TO CONTAIN: 61 CARTONS OF USED HOUSEHOLD AND PERSONAL EFFECTS FREIGHT PREPAID SKID DIMENSION: 53 x 40 x 69 INCHES (LxWxH) 48 x 42 x 52 INCHES 50 x 48 x 59 INCHES	2680LBS CLEAN ONBOARD V/V APL ENGLAND V103 DATED: JUL15,2001	227.25CUF 6.435CBM

SHIPPERS DECLARED VALUE $ ______ SUBJECT TO EXTRA FREIGHT
AS PER TARIFF AND CARRIERS LIMITS REFER TO CLAUSE 16 HEREOF

FREIGHT PAYABLE AT BY

FREIGHT CHARGES	PREPAID	COLLECT
227.25 CUF X 4.00 PER CUF	$ 909.00	
TOTAL CHARGES		

* APPLICABLE ONLY WHEN DOCUMENT USED AS A THROUGH BILL OF LADING

Received the goods, or packages said to contain goods herein mentioned, in apparent good order and condition unless otherwise indicated, to be transported and delivered, to trans-shipped as herein provided.
The carriage is subject to the provisions of the U.S. Carriage of Goods by Sea Act of 1936. All the terms and conditions of the Carrier's regular form Bill of Lading, as filed with the Federal Maritime Commision available to any shipper or consignee upon request, are incorporated with the force and effect as if they were written at length herein, and all such terms and conditions so incorporated by reference are agreed by Shipper to be binding and to govern the realtions, whatever they may be between those included in the words "Shipper" and "Carrier" as defined in Carriers regular form Bill of Lading.
IN WITNESS WHEREOF, the Carrier has signed and the Shipper has received THREE (3) original bills of lading. ONE of which being accomplished, the others to stand void.

Dated At SIAM INTERNATIONAL FREIGHTLINES AS CARRIER
By NANCY S.
MO JUL DAY 15 YEAR 2001

BL. No. 04 0/B K

(a) Complex input document (Bill of Lading)

```
{
  "BillOfLadingType": "OCEAN BILL OF
LADING",
  "Carrier": "SIAM INTERNATIONAL
FREIGHTLINES",
  "ShipperExporter": {
    "Name": "MR. VORAVIT KOSOL",
    "Address": "1100 S.W. CLAY STREET,
PORTLAND, OR 97201, U.S.A."
  },
  "BookingNo": "LAXBK D 03628",
  "ExportReferences": "04**0/B*K",
  "Consignee": {
    "Name": "MR. VORAVIT KOSOL",
    "Address": "18/1 COVENT ROAD, SILOM,
BANGKOK THAILAND"
  },
  "ForwardingAgent": {
    "Name": "SIAM INTERNATIONAL
FREIGHTLINE",
    "FMCNo": "OTI NO. 12475NF"
  },
  "PointAndCountryOfOriginOfGoods":
"PORTLAND",
  "NotifyParty": {
    "Name": "SAME AS ABOVE",
    "Phone": "635 3**3"
  },
  "AlsoNotify": {
    "Name": "TRIUMPH INTERFREIGHT",
    "Address": "537/197 SATHUPRADIT
COMPLEX, YANNAVA, BANGKOK
THAILAND",
    "Phone": "2942095"
  },
"PlaceOfInitialReceipt": "PORTLAND",
  "PortOfLoading": "LONG BEACH",
  "ExportingCarrier": "APL ENGLAND V 103",
"AirSeaPortOfDischarge": "BANGKOK, THAILAND",
  "PlaceOfDeliveryByOnCarrier": "",
 "LoadingPierTerminal": "",
  "TypeOfMove": "",
"MarksAndNumbers": "VORAVIT KOSOL, BANGKOK, THAILAND",
  "NoOfPackages": "61 BOXES (3SKIDS)",
  "DescriptionOfGoods": {
    "Contents": "61 CARTONS OF USED HOUSEHOLD AND
PERSONAL EFFECTS",
    "Freight": "FREIGHT PREPAID",
    "SkidDimensions": [
      "53 x 40 x 69 INCHES (LxWxH)",
      "48 x 42 x 52 INCHES",
      "50 x 48 x 59 INCHES"
    ]
  },
  "GrossWeight": "2680 LBS",
  "Measurement": "227.25 CUFT / 67.435 CBM",
  "VesselOnboard": "CLEAN ONBOARD",
  "CarrierVesselInfo": {
    "Vessel": "V/V APL ENGLAND V103",
    "Date": "JUL15,2001"
  },
"FreightCharges": {
    "FreightRate": "222.25 CUFT*4.00 PER CUFT"
    "Prepaid": "$909.00",
    "Collect": Na
  },
  "ShippersDeclaredValue": "",
  "SubjectToExtraFreight": "",
  "FreightPayableAt": "",
  "Signature": {
    "CarrierSignature": "NANCY S.",
    "PlaceAndDate": "MO JULY DAY 15 YEAR 2001"
  },
  "BillOfLadingNo": "04**0/B*K"
}
```

(b) Output JSON result generated by LLM-TKIE

Fig. 9. An end-to-end example illustrating the extraction process of LLM-TKIE. Given a complex semi-structured document (**a**), our method produces a highly structured output (**b**), correctly aligning textual fields, handling nested structures, and preserving semantic consistency even in cases with missing or duplicated headers.

Methods	CORD F1	CORD Acc	SROIE F1	SROIE Acc
TRIE	–	–	82.1	–
Donut	84.1	90.9	83.2	92.8
Dessurt	82.5	–	84.9	–
DocParser	84.5	–	87.3	–
SeRum	80.5	85.8	85.6	92.8
OmniParser	84.8	88.0	85.6	93.6
LLM-TKIE	80.9	88.85	83.9	93.3

Table 1. Performance comparison of key information extraction models on CORD and SROIE datasets. Acc refers to TED-based accuracy.

End-to-end key information extraction comparison

Experimental background and setup

This section compares several state-of-the-art KIE models, including TRIE, Donut, Dessurt, DocParser, SeRum, OmniParser, and the proposed LLM-TKIE framework. LLM-TKIE uses pre-trained components (DBNet++ for text detection, SVTR LCNet for text recognition) and the quantized LLaMA 3.1:70B model. Unlike other models, which were trained on the CORD and SROIE datasets, LLM-TKIE was not fine-tuned or trained on these datasets. Table 1 presents the performance results for the different KIE models on both datasets.

Despite achieving competitive accuracy across both datasets, we observe that methods such as DocParser and OmniParser slightly outperform LLM-TKIE in terms of CORD F1 scores. This is largely attributed to their supervised fine-tuning on domain-specific templates and specialized layout encoders tailored for receipts and forms. In contrast, LLM-TKIE adopts a generalist approach relying solely on in-context learning without additional domain-specific training. In future work, we plan to incorporate lightweight layout encoders and domain-adaptive instruction tuning, along with structure-constrained decoding mechanisms, to enhance field alignment and robustness in layout-sensitive scenarios.

Analysis of experimental results

The results show that LLM-TKIE, despite not being fine-tuned on the datasets, achieves competitive performance. On the CORD dataset, it reaches an F1 score of 80.9% and an ACC score of 88.85%, indicating slight field-level inaccuracies but strong overall accuracy in generating structured outputs. On the SROIE dataset, LLM-TKIE achieves an F1 score of 83.9% and an ACC score of 93.3%, demonstrating robust performance in structured output generation. Compared to models like Donut and OmniParser, LLM-TKIE remains competitive, particularly in tree-based similarity accuracy, underscoring its ability to perform well without extensive labeled data.

Performance analysis in few-shot settings

A key strength of the LLM-TKIE mechanism is its ability to perform KIE with minimal reliance on large annotated datasets, unlike models requiring extensive supervised training on datasets such as CORD and SROIE. In few-shot settings, the mechanism demonstrates strong adaptability to unseen document formats and new domains, making it well-suited for low-resource environments. While its F1 score may be slightly lower in certain field-level tasks, its tree-based similarity accuracy highlights clear advantages. This combination of generalization capability and structured output generation ensures its effectiveness for real-world KIE applications.

Performance comparison of open source models

This section presents a comprehensive evaluation of various open-source LLMs within the proposed LLM-TKIE framework, using the SROIE dataset[14]. The primary objective of these experiments is to identify the most suitable LLM for KIE tasks, optimizing for both extraction accuracy and computational efficiency in the LLM-TKIE mechanism.

Impact of model size on performance

All models utilize the Q4_0 quantization method to ensure consistency. Results in Fig. 10 illustrate a clear trade-off between accuracy and inference speed as model size increases. Smaller models (< 1B parameters), although fast in inference, show limited extraction capabilities. Conversely, medium-sized models (around 7–9B parameters) achieve significantly improved accuracy while maintaining acceptable inference times for practical scenarios.

Models larger than 16B parameters provide marginal accuracy improvements with diminishing returns, accompanied by substantial inference delays. For example, inference times of models such as Qwen2 72B surpass 10 s, limiting their real-time applicability. Overall, models in the 7–9B parameter range offer an optimal balance between extraction performance and inference efficiency, making them suitable for real-world deployment.

Impact of quantization methods on performance

We investigate how various quantization methods impact the accuracy and inference efficiency of different-sized models. Results in Fig. 11 reveal that:

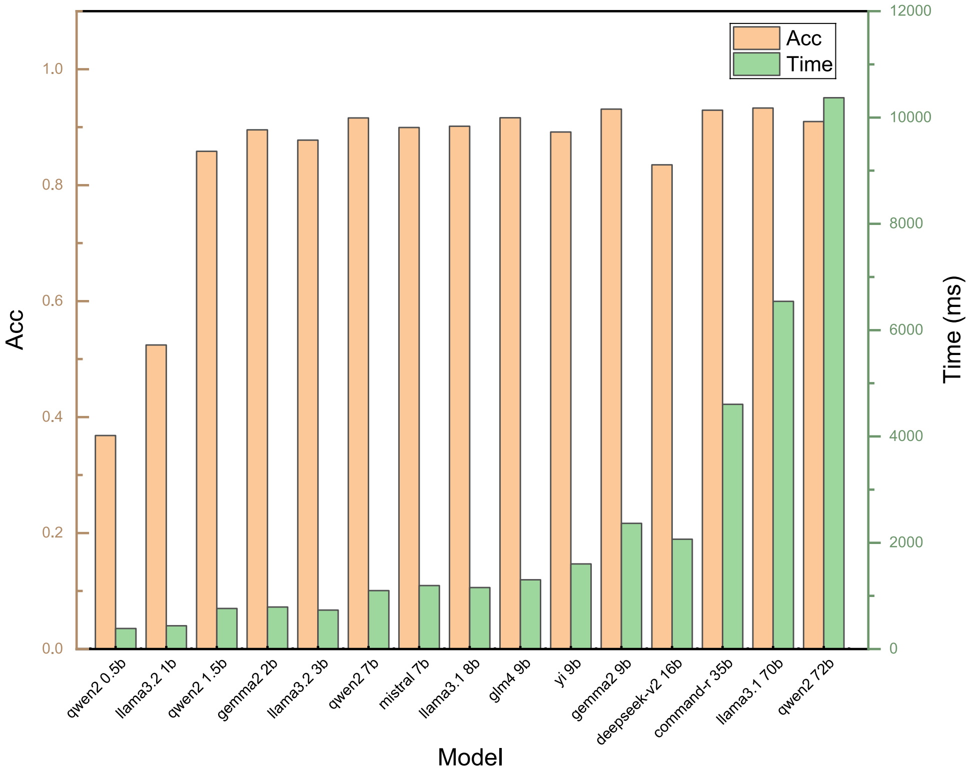
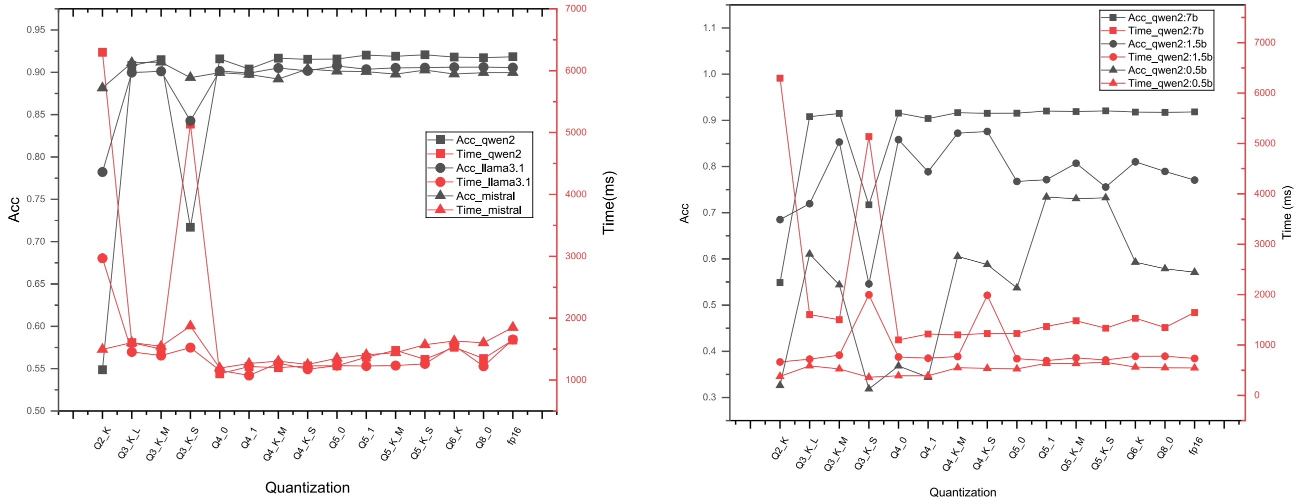


Fig. 10. Impact of model parameter size on key information extraction accuracy and inference time using Q4_0 quantization method.



(a) Quantization impact on large-scale models.

(b) Quantization impact on smaller Qwen2 models.

Fig. 11. Impact of quantization on (**a**) different large-scale models and (**b**) smaller Qwen2 models.

For **large-scale models** (Fig. 11a), such as Qwen2-7B, LLaMA3-8B, and Mistral, accuracy remains high and stable (Acc $\geq$ 0.9) when using quantization levels at or above Q4_0. At these levels, inference time is substantially reduced—from over 6000 ms (FP16) to under 1500 ms—without compromising extraction performance. In contrast, aggressive low-bit quantization (e.g., Q2_K, Q3_K) leads to steep accuracy drops, particularly for LLaMA3 and Qwen2, suggesting that bit-width reduction below 4-bit can severely affect reasoning quality in structured extraction.

For **smaller Qwen2 models** (Fig. 11b), including 2.5B, 1.5B, and 0.5B variants, quantization sensitivity becomes more pronounced. Accuracy fluctuates widely even around Q4 levels, and under Q3_K settings, performance can drop below 0.5. This instability is accompanied by latency improvements, but the trade-off is less favorable compared to larger models. Interestingly, Qwen2-2.5B performs more consistently than its smaller counterparts, indicating that model size plays a critical role in robustness to quantization.

Takeaways:

- Q4_0 emerges as the optimal quantization level for maintaining high accuracy while significantly accelerating inference, especially for large models.
- Smaller models are more vulnerable to accuracy degradation under quantization, particularly at sub-4-bit levels.
- These results emphasize the importance of model-specific quantization tuning in KIE scenarios, where structured consistency and semantic understanding are crucial.

Performance comparison between large language models and multimodal models

To evaluate the effectiveness of KIE from non-standardized customs documents, we compare two paradigms: (1) a language-only approach based on our proposed LLM-TKIE framework, and (2) a multimodal large model (MLM) approach enhanced with instruction tuning. Both systems generate structured outputs (e.g., JSON) containing the extracted key fields.

To ensure fair evaluation and avoid potential contamination from public pretraining data, we construct a proprietary, manually annotated dataset of import-export customs documents (see in Fig. 8 case: Customs Import-Export Lable), unseen during the pretraining phase of any evaluated model. The LLM-TKIE model is prompted with OCR-recognized text, while the MLM receives raw document images and corresponding task prompts.

As shown in Fig. 12, our method outperforms both similarly sized language models and all evaluated multimodal models in terms of accuracy on this private dataset. In particular, we observe that multimodal models underperform their language-only counterparts of comparable size by 5–8%, suggesting that visual modalities may not provide a clear advantage for KIE tasks in this domain. Moreover, our method exhibits superior transferability and generalization to complex layouts and unseen field structures, confirming its robustness in low-resource, domain-specific scenarios.

Failure case analysis

Despite demonstrating strong average performance across multiple benchmarks, our proposed LLM-TKIE system is still susceptible to several types of systematic failure. To illustrate the limitations, we present in Fig. 13 three representative failure cases, each highlighting a distinct type of error:

- **Hallucinated Output:** In this case, as shown in Fig. 13a the model generates structured fields that do not exist in the input document, such as fictitious entity names, fabricated addresses, or invented regulatory codes

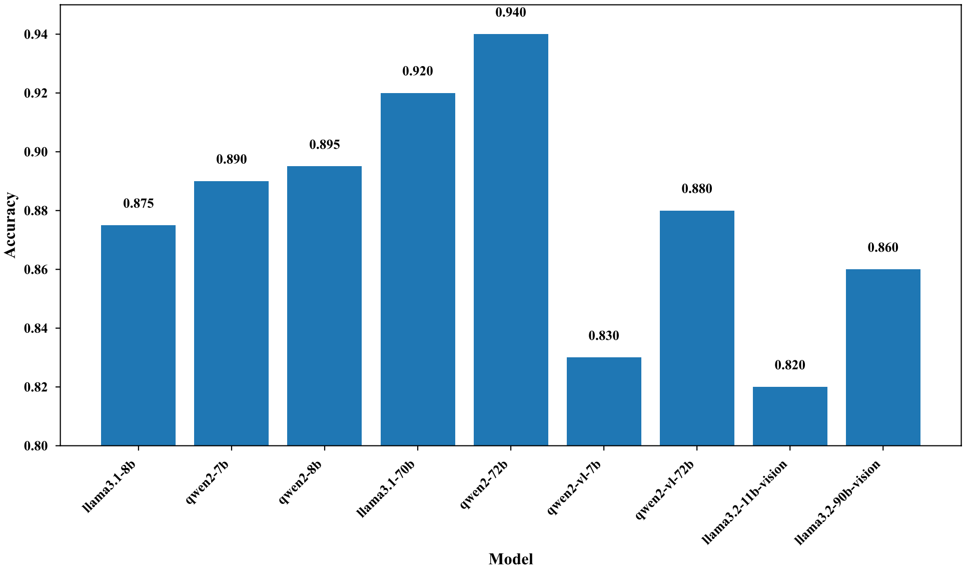
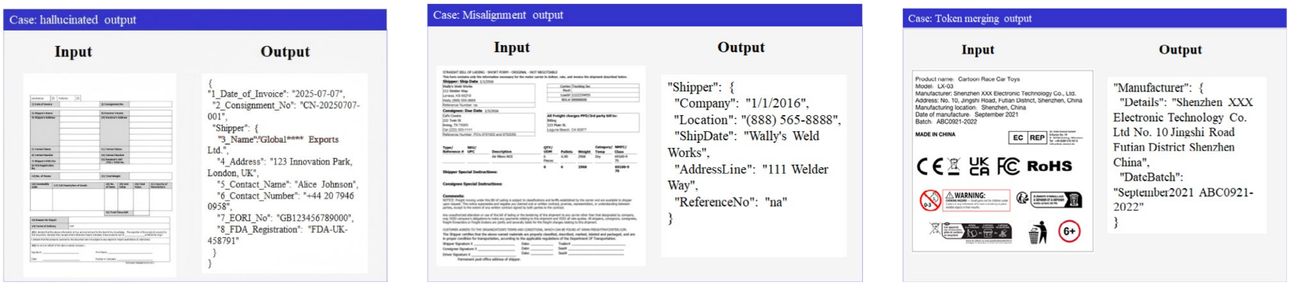


Fig. 12. Performance comparison of language and multimodal models on customs KIE task.


Case: hallucinated output
Input
Output
Case: Misalignment output
Input
Output
Case: Token merging output
Input
Output

```
{
"1_Date_of_Invoice": "2025-07-07",
 "2_Consignment_No": "CN-20250707-
001",
 "Shipper": {
  "3_Name":"Global**** Exports
Ltd.",
  "4_Address": "123 Innovation Park,
London, UK",
  "5_Contact_Name": "Alice Johnson",
  "6_Contact_Number": "+44 20 7946
0958",
  "7_EORI_No": "GB123456789000",
  "8_FDA_Registration": "FDA-UK-
458791"
 }
}
```

```
"Shipper": {
 "Company": "1/1/2016",
 "Location": "(888) 565-8888",
 "ShipDate": "Wally's Weld
Works",
 "AddressLine": "111 Welder
Way",
 "ReferenceNo": "na"
}
```

Product name: Cartoon Race Car Toys
Model: LX-03
Manufacturer: Shenzhen XXX Electronic Technology Co., Ltd.
Address: No. 10, Jingshi Road, Futian District, Shenzhen, China
Manufacturing location: Shenzhen, China
Date of manufacture: September 2021
Batch: ABC0921-2022
MADE IN CHINA

```
"Manufacturer": {
 "Details": "Shenzhen XXX
Electronic Technology Co.
Ltd No. 10 Jingshi Road
Futian District Shenzhen
China",
 "DateBatch":
"September2021 ABC0921-
2022"
}
```

(a) Hallucinated Output

(b) Field Misalignment

(c) Token Merging

Fig. 13. Representative failure cases with input document and erroneous output pairs.

(e.g., "FDA-UK-458791"). This typically occurs in low-quality scans or under domain shift, where the prompt encourages the model to complete a schema regardless of content fidelity. It reflects a classical hallucination behavior observed in large generative models. Future work may explore confidence calibration, trust scoring, or schema-aware filtering to mitigate this issue.

- **Field Misalignment:** Here, as shown in Fig. 13b text segments are incorrectly mapped to target fields. For instance, the date "1/1/2016" is assigned to the "Company" field, and the phone number is misclassified as "Location." Such errors often stem from OCR mis-segmentation or visual layout ambiguity, where adjacent fields are misinterpreted due to overlapping visual anchors or irregular alignment. One solution could involve incorporating lightweight spatial encoding or positional priors to constrain entity-to-field alignment.
- **Token Merging:** In this failure type, as shown in Fig. 13c adjacent fields such as manufacturer name, address, and production date are merged into a single flattened string. This generally happens in densely packed layouts with poor text separation, where the OCR fails to delineate field boundaries. As a result, the LLM sees the content as a run-on sentence and incorrectly collapses fields into one. Addressing this requires layout-aware segmentation or prefix-aware decoding mechanisms.

Overall, these failure modes suggest future improvements should focus on: (1) enhancing the robustness of layout-aware parsing, (2) calibrating hallucination risk via confidence-aware decoding, and (3) improving structural grounding via schema-constrained output filtering or multi-pass extraction.

Discussion

LLM-TKIE may suffer from hallucination, particularly under few-shot prompting, where LLMs fabricate structured field values that are not LLM-TKIE systems are susceptible to hallucination, particularly in schema-driven few-shot prompting scenarios, where large language models may generate structured field values that are not present in the original input, such as fabricated codes or inferred entity names. This behavior is often amplified by rigid prompt templates that compel the model to populate all fields, regardless of content availability.

Additionally, layout inconsistencies or OCR errors can result in label misalignment, field merging, or malformed output structures. To enhance system reliability, future improvements should focus on incorporating field-level confidence calibration, enforcing schema-constrained decoding, and adopting iterative verification strategies. Integrating spatially-aware vision-language models or lightweight verifier modules may further mitigate hallucination risks and improve robustness in visually noisy document contexts.

We further evaluated the framework's robustness across LLM variants and prompt styles. Experiments with Qwen2-7B, LLaMA3-8B, and Mistral-7B under consistent settings showed model-specific tendencies: Qwen2 yielded more complete outputs but introduced occasional hallucinations; LLaMA3 was more conservative but precise; Mistral displayed mid-range behavior sensitive to layout complexity. Prompt variations (e.g., "Extract structured JSON" vs. "Return key-value pairs") also caused shifts in output structure and field interpretation. These findings reveal two challenges: prompt sensitivity and model-specific biases. Future work will explore prompt ensemble/dropout techniques, model-aware confidence calibration, and fallback strategies using alternative prompts or smaller models to enhance robustness.

LLM-TKIE is currently optimized for sequential, high-fidelity extraction from individual documents, fitting practical scenarios in logistics, customs, and compliance. However, batch and concurrent document processing remain underexplored. Such settings pose challenges in prompt reuse, memory efficiency, and decoding latency. Handling heterogeneous document structures within a batch or managing interleaved turns requires advanced routing and scheduling. Future directions include designing batch-aware prompt templates, employing accelerated decoding methods (e.g., vLLM, speculative decoding), and integrating lightweight document classification modules to streamline multi-document pipelines.

Finally, as a generative system, LLM-TKIE faces ethical challenges, especially hallucination and overconfident predictions in ambiguous or out-of-distribution inputs. These risks may compromise downstream automation. We employ constrained decoding and structure-aware postprocessing to enforce consistency. In future versions, we recommend incorporating verifier modules or vision-aligned grounding to detect hallucinated entities. For sensitive domains such as legal or medical document analysis, human-in-the-loop auditing remains essential to ensure accountability and safe deployment.

Conclusion

This paper introduced the LLM-TKIE mechanism, designed to address critical limitations in existing methods for KIE. By coupling OCR with the semantic reasoning capabilities of LLMs, LLM-TKIE enables robust and efficient extraction of structured data from diverse and complex table layouts. The framework leverages few-shot prompting strategies, demonstrating strong adaptability to new document formats while reducing reliance on extensive labeled datasets.

LLM-TKIE achieves competitive performance across industry-standard datasets, including F1-scores of 80.9 and 83.9 on the CORD and SROIE datasets, respectively, with high structural accuracy validated through tree-edit distance metrics. On our private dataset, the proposed method outperforms both similarly sized language models and all evaluated multimodal models in terms of accuracy. These results highlight the scalability and transferability of LLM-TKIE to real-world applications, even in resource-constrained environments. Furthermore, a comprehensive evaluation of pre-trained models and quantization strategies provides practical insights for optimizing KIE systems in terms of computational efficiency and inference time.

Despite its advantages, LLM-TKIE still faces certain limitations. First, its reliance on generic prompting without fine-tuning can limit field-level consistency when handling documents with highly irregular layouts or ambiguous semantics. Second, while performance is competitive, several task-specific models (e.g., DocParser, OmniParser) achieve marginally higher F1-scores on certain public benchmarks due to their tight integration of layout priors and domain-specific supervision.

To close this performance gap, future improvements could include incorporating lightweight layout-aware modules, enhancing visual-textual alignment via multi-modal adapters, and exploring constraint-based decoding mechanisms to enforce structural consistency. Additionally, the integration of user feedback loops or active learning strategies may improve the system's adaptability to low-resource document types and long-tail entities.

In summary, LLM-TKIE represents a scalable, transferable, and efficient solution for industrial automation workflows in document intelligence. Future work will explore extending support to multi-page and multi-modal documents, expanding to non-English or multilingual formats, and deploying the system under edge-device constraints with dynamic model compression and streaming inference strategies.

Data availability

Data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Received: 2 May 2025; Accepted: 8 August 2025
Published online: 14 August 2025

References

1. Schuster, D. et al. Intellix—end-user trained information extraction for document archiving. In *2013 12th International Conference on Document Analysis and Recognition*, 101–105 (2013).
2. Dengel, A. R. & Klein, B. smartFIX: A requirements-driven system for document analysis and understanding. In Goos, G. et al. (eds.) *Document Analysis Systems V*, vol. 2423, 433–444 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2002). Series Title: Lecture Notes in Computer Science.

3. Xu, Y. et al. 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*, KDD '20, 1192–1200 (Association for Computing Machinery, New York, NY, USA, 2020).
4. Xu, Y. et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 2020 International Conference on Document Analysis and Recognition (ICDAR)*, 365–376 (2020).
5. Appalaraju, S. & Manmatha, R. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 990–999 (2021).
6. Li, P., Zhou, Z., Yang, M. & Shi, W. Omniparser: A unified framework for multi-task document parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3775–3784 (2020).
7. Wang, F., Yang, Y., Zhang, Z. & Bai, X. Deepsolo: End-to-end model for text detection and recognition in complex documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4597–4606 (2022).
8. Lin, X. V. et al. Few-shot learning with multilingual generative language models. In Goldberg, Y., Kozareva, Z. & Zhang, Y. (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9019–9052 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
9. Achiam, J. et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).
10. Radford, A. et al. Learning transferable visual models from natural language supervision. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 8748–8763 (PMLR, 2021).
11. Zhang, R., Liu, Y. & Bai, X. Dessurt: Deep end-to-end structured text recognition framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4712–4720 (2022).
12. Brown, T. et al. Language models are few-shot learners. in (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.)) *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020).
13. Abdallah, A. et al. Coru: Comprehensive post-ocr parsing and receipt understanding dataset (2024). arXiv:2406.04493.
14. Huang, Z. et al. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (IEEE, 2019).
15. Zhou, X. et al. East: An efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2642–2651 (2017).
16. Liao, M., Wan, Z., Yao, C., Chen, K. & Bai, X. Real-time scene text detection with differentiable binarization. *Proc. AAAI Conf. Artif. Intell.* **34**, 11474–11481 (2020).
17. Yang, C., Chen, M., Yuan, Y. & Wang, Q. Zoom text detector. *IEEE Transactions on Neural Networks and Learning Systems* 1–13 (2023).
18. Zhong, Y. et al. Prpn: Progressive region prediction network for natural scene text detection. *Knowledge-Based Syst.* **236**, 107767 (2022).
19. Zhu, B., Liu, F., Chen, X., Tang, Q. & Philip Chen, C. Acp-net: Asymmetric center positioning network for real-time text detection. *Knowledge-Based Syst.* **305**, 112603 (2024).
20. Zhu, B., Chen, X., Tang, Q., Chen, C. P. & Liu, F. Ek-net++: Real-time scene text detection with expand kernel distance and epoch adaptive weight. *Expert Syst. Appl.* **267**, 126159 (2025).
21. Li, M. et al. Trocr: Transformer-based optical character recognition with pre-trained models. arXiv preprint arXiv:2109.10282 (2021).
22. Zhang, J.-Y., Liu, X.-Q., Xue, Z.-Y., Luo, X. & Xu, X.-S. Magic: Multi-granularity domain adaptation for text recognition. *Pattern Recognit.* **161**, 111229 (2025).
23. Li, C., Jin, L., Sun, X. & Tang, R. Paddleocr: A practical ultra lightweight ocr system. arXiv preprint arXiv:2012.05707 (2021).
24. Das, A., Palaiahnakote, S., Banerjee, A., Antonacopoulos, A. & Pal, U. Soft set-based mser end-to-end system for occluded scene text detection, recognition and prediction. *Knowledge-Based Syst.* **305**, 112593 (2024).
25. Tong, G., Dong, M., Sun, X. & Song, Y. Natural scene text detection and recognition based on saturation-incorporated multi-channel mser. *Knowledge-Based Syst.* **250**, 109040 (2022).
26. Ke, W., Liu, Y., Yang, X., Wei, J. & Hou, Q. Align, enhance and read: Scene Tibetan text recognition with cross-sequence reasoning. *Appl. Soft Comput.* **169**, 112548 (2025).
27. Liu, C. et al. Qt-textsr: Enhancing scene text image super-resolution via efficient interaction with text recognition using a query-aware transformer. *Neurocomputing* **620**, 129241 (2025).
28. Du, Y. et al. Text generation and multi-modal knowledge transfer for few-shot object detection. *Pattern Recognit.* **161**, 111283 (2025).
29. Park, S. et al. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019* (2019).
30. Huang, Z. et al. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1516–1520 (IEEE, 2019).
31. Zhu, X., Tang, L. & Wang, J. Trie: End-to-end text recognition and information extraction. *Proc. AAAI Conf. Artif. Intell.* **35**, 443–451 (2021).
32. Mamede, S. & Schmidt, H. G. Making large language models into reliable physician assistants. *Nat. Med.* 1–2 (2025).
33. McDuff, D. et al. Towards accurate differential diagnosis with large language models. *Nature* 1–7 (2025).
34. Kleinig, O. et al. How to use large language models in ophthalmology: From prompt engineering to protecting confidentiality. *Eye* **38**, 649–653 (2024).
35. Touvron, H. et al. Llama: Open and efficient foundation language models (2023). arXiv:2302.13971.
36. Augenstein, I. et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Machine Intell.* **6**, 852–863 (2024).
37. Chen, W. et al. Llm-enabled incremental learning framework for hand exoskeleton control. *IEEE Trans. Automation Sci. Eng.* 1–10 (2024).
38. Li, X. et al. Toward cognitive digital twin system of human–robot collaboration manipulation. *IEEE Trans. Automation Sci. Eng.* 1–14 (2024).
39. Zhang, Y., Cao, Y., Xu, X. & Shen, W. Logicode: An llm-driven framework for logical anomaly detection. *IEEE Trans. Automation Sci. Eng.* 1–0 (2024).
40. Foo, G., Kara, S. & Pagnucco, M. Artificial learning for part identification in robotic disassembly through automatic rule generation in an ontology. *IEEE Trans. Automation Sci. Eng.* **20**, 296–309 (2023).
41. Gao, Y., Zhu, G., Duan, Y. & Mao, J. Semantic encoding algorithm for classification and retrieval of aviation safety reports. *IEEE Trans. Automation Sci. Eng.* 1–8 (2024).
42. Rashid, F., Ranaweera, N., Doyle, B. & Seneviratne, S. Llms are one-shot url classifiers and explainers. *Comput. Netw.* 111004 (2024).
43. Aarab, I. Llm-based ir-system for bank supervisors. *Knowledge-Based Syst.* 112914 (2024).
44. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).
45. Mahmood, A., Wang, J., Yao, B., Wang, D. & Huang, C.-M. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *Int. J. Human-Computer Stud.* **195**, 103406 (2025).

46. Zeng, Z. et al. Kosel: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Syst.* **309**, 112837 (2025).
47. Li, S. et al. Taming large language models to implement diagnosis and evaluating the generation of llms at the semantic similarity level in acupuncture and moxibustion. *Expert Syst. Appl.* **264**, 125920 (2025).
48. Liu, X., Erkoyuncu, J. A., Fuh, J. Y. H., Lu, W. F. & Li, B. Knowledge extraction for additive manufacturing process via named entity recognition with llms. *Robotics Computer-Integrated Manufact.* **93**, 102900 (2025).
49. Zheng, L. et al. Teaching via llm-enhanced simulations: Authenticity and barriers to suspension of disbelief. *Internet Higher Educ.* **65**, 100990 (2025).
50. Zhao, S. & Sun, X. Enabling controllable table-to-text generation via prompting large language models with guided planning. *Knowledge-Based Syst.* **304**, 112571 (2024).
51. Peng, Z., Wu, X., Wang, Q. & Fang, Y. Soft prompt tuning for augmenting dense retrieval with large language models. *Knowledge-Based Syst.* **309**, 112758 (2025).
52. Yan, Y. et al. Collaborate slm and llm with latent answers for event detection. *Knowledge-Based Syst.* **305**, 112684 (2024).
53. Huang, X., Zhang, J., Xu, Z., Ou, L. & Tong, J. A knowledge graph based question answering method for medical domain. *PeerJ Comput. Sci.* **7**, e667 (2021).
54. Thomas, A. & Sangeetha, S. Knowledge graph based question-answering system for effective case law analysis. In *Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, 291–300 (Springer, 2022).
55. Giarelis, N., Mastrokostas, C. & Karacapilidis, N. A unified llm-kg framework to assist fact-checking in public deliberation. In *Proceedings of the First Workshop on Language-Driven Deliberation Technology (DELITE)@ LREC-COLING 2024*, 13–19 (2024).
56. Chowdhury, S. & Soni, B. R-vqa: A robust visual question answering model. *Knowledge-Based Syst.* **309**, 112827. https://doi.org/10.1016/j.knosys.2024.112827 (2025).
57. Chowdhury, S. & Soni, B. Beyond words: Esc-net revolutionizes vqa by elevating visual features and defying language priors. *Computational Intell.* **40**, e70010. https://doi.org/10.1111/coin.70010 (2024)
58. Chowdhury, S. & Soni, B. Envqa: Improving visual question answering model by enriching the visual feature. *Eng. Appl. Artif. Intell.* **142**, 109948. https://doi.org/10.1016/j.engappai.2024.109948 (2025).
59. Chowdhury, S. & Soni, B. Qsfvqa: A time efficient, scalable and optimized vqa framework. *Arabian J. Sci. Eng.* **48**, 10479–10491 (2023).
60. Chowdhury, S. & Soni, B. Handling language prior and compositional reasoning issues in visual question answering system. *Neurocomputing* **635**, 129906 (2025).
61. Ch'ng, C. K. & Chan, C. S. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 935–942 (2017).
62. Li, C. et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system (2022). arXiv:2206.03001.

Acknowledgements

This work was partly supported by the General Administration of Customs of the People's Republic of China under the project titled "Research on the Customs Supervision System of the China-Europe Railway Express from the Perspective of International Supply Chain Hubs" (Project Approval Number: 2021HK248).

Author contributions

Y.Y. conceived and supervised the study. R.H., Y.Y., and S.L. contributed equally to the development of the methodology, model design, and manuscript preparation. Z.L. and J.L. provided support in system implementation and experimental validation. X.D. and H.S. were responsible for data processing, visualization, and result analysis. L.R. offered domain expertise and practical insights from the perspective of customs operations. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.