UniKIE-Bench: Benchmarking Large Multimodal Models for Key Information Extraction in Visual Documents

Yifan Ji[1*], Zhipeng Xu[1*], Zhenghao Liu[1†], Zulong Chen[3], Qian Zhang[3], Zhibo Yang[3], Junyang Lin[3], Yu Gu[1], Ge Yu[1] and Maosong Sun[2]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Alibaba Group, Hangzhou, China

Abstract

Key Information Extraction (KIE) from real-world documents remains challenging due to substantial variations in layout structures, visual quality, and task-specific information requirements. Recent Large Multimodal Models (LMMs) have shown promising potential for performing end-to-end KIE directly from document images. To enable a comprehensive and systematic evaluation across realistic and diverse application scenarios, we introduce UniKIE-Bench, a unified benchmark designed to rigorously evaluate the KIE capabilities of LMMs. UniKIE-Bench consists of two complementary tracks: a constrained-category KIE track with scenario-predefined schemas that reflect practical application needs, and an open-category KIE track that extracts any key information that is explicitly present in the document. Experiments on 15 state-of-the-art LMMs reveal substantial performance degradation under diverse schema definitions, long-tail key fields, and complex layouts, along with pronounced performance disparities across different document types and scenarios. These findings underscore persistent challenges in grounding accuracy and layout-aware reasoning for LMM-based KIE. All codes and datasets are available at `https://github.com/NEUIR/UNIKIE-BENCH`.

1 Introduction

Key Information Extraction (KIE) aims to identify and extract critical fields from visually rich and semantically diverse documents (Tang et al., 1994; Cesarini et al., 2002; Mao et al., 2003). It serves as a fundamental component for downstream applications such as automated accounting, financial auditing, and enterprise knowledge management (Nasar et al., 2018; Oral et al., 2020; Skalický et al., 2022). Despite extensive progress, KIE in real-world settings remains highly challenging due to substantial variations in document layouts, capture conditions, linguistic patterns, and field definitions (Gbada et al., 2025). Addressing these challenges requires fine-grained multimodal understanding capabilities that tightly couple visual structure with textual semantics (Zhang et al., 2023; Ding et al., 2024; Rombach and Fettke, 2025).

Recent advances in Large Multimodal Models (LMMs) have renewed interest in KIE tasks (Bai et al., 2025b; Comanici et al., 2025; Vafaie et al., 2025). Unlike traditional approaches that rely on explicit and modular pipelines for text, layout, and visual processing, LMMs extract information from document images in an end-to-end manner. This design mitigates error propagation across modules and enables strong generalization across a wide range of visual tasks (Liu et al., 2024c; Bai et al., 2025c; Yang et al., 2025). Despite these advances, existing benchmarks remain insufficient for comprehensively evaluating LMMs' KIE capabilities under realistic document settings (Ebrahimi et al., 2022; He et al., 2023; Wei et al., 2025). In particular, most KIE benchmarks are designed for specific application scenarios and rely on heterogeneous annotation schemas, task formulations, or evaluation metrics, making it difficult to support consistent, end-to-end evaluation across diverse scenarios (Laatiri et al., 2023; Kuang et al., 2023). Furthermore, the predefined fields in these datasets are often tightly coupled to scenario-specific requirements, limiting their ability to systematically cover the diverse types of key information encountered in real-world documents (Laatiri et al., 2023; Kuang et al., 2023; Shi et al., 2023).

To address these limitations, we introduce UniKIE-Bench, a comprehensive benchmark for the systematic evaluation of KIE capabilities in LMMs. As shown in Figure 1, UniKIE-Bench is built upon a unified, schema-guided structured prediction formulation, which enables

[*] indicates equal contribution.
[†] indicates corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6331–6352
July 2-7, 2026 ©2026 Association for Computational Linguistics

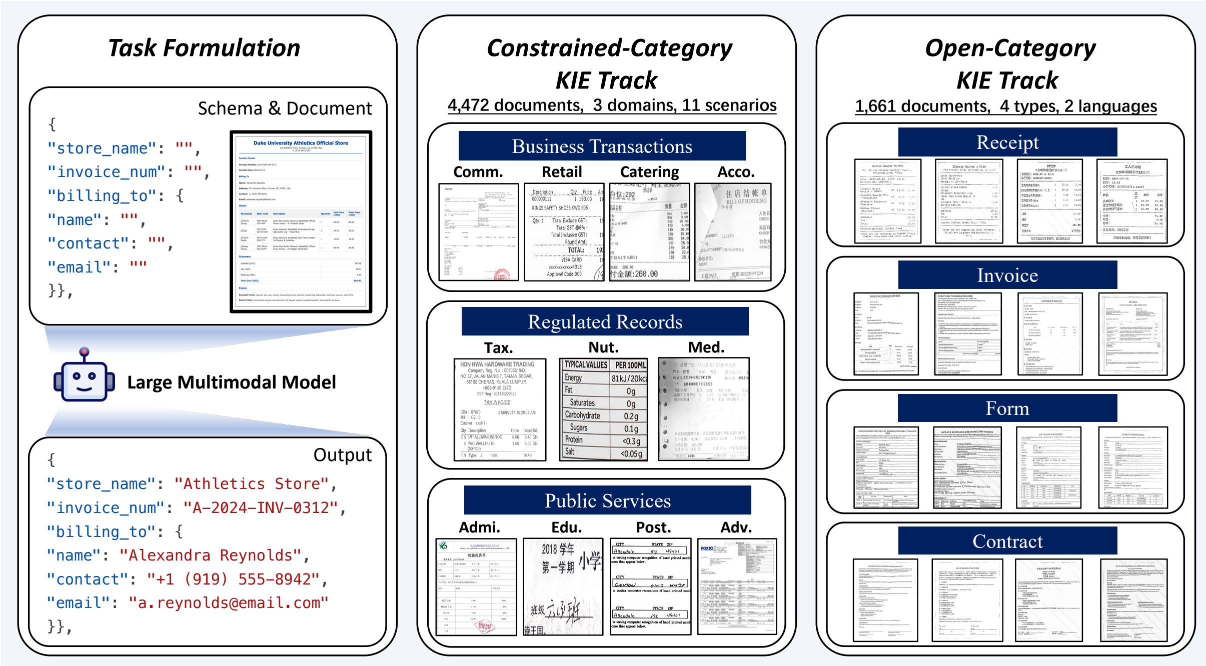


Figure 1: Overview of our UNIKIE-BENCH. UNIKIE-BENCH is built upon a schema-guided KIE formulation to enable end-to-end evaluation of LMMs. It comprises two complementary evaluation tracks: the constrained-category KIE track and the open-category KIE track.

a consistent evaluation paradigm across diverse document types and annotation standards. Specifically, UNIKIE-BENCH comprises two tracks: a constrained-category KIE track targeting specific application scenarios, and an open-category KIE track designed for general-purpose key information extraction. The constrained-category KIE track incorporates task-specific information requirements derived from a wide range of real-world applications. These requirements are organized by their application scenario, enabling fine-grained, scenario-aware evaluation under well-defined schemas. In contrast, the open-category KIE track requires models to extract all key information explicitly present in each document, evaluating the general-purpose extraction capability of LMMs. Together, these two tracks enable UNIKIE-BENCH to capture both targeted, application-driven performance and holistic extraction ability across heterogeneous documents.

Our experiments on 15 LMMs demonstrate that state-of-the-art LMMs still suffer from substantial limitations in KIE. In particular, performance degrades more severely when confronted with diverse schemas, long-tail fields, and complex document layouts, and this degradation is further exacerbated by realistic visual artifacts. We further observe pronounced performance disparities across document types, application scenarios, and field categories, revealing persistent challenges in schema grounding, layout-aware reasoning, and robust information extraction under different visual conditions. These findings underscore the necessity of a unified and comprehensive benchmark that captures the structural and visual complexities of real-world documents. We hope that UNIKIE-BENCH will facilitate more rigorous evaluation of LMM-based KIE systems and provide a solid foundation for assessing the generalizability and reliability of LMMs in document understanding.

2 Related Work

Key Information Extraction (KIE) is a fundamental task in document understanding, aiming to extract structured fields from visually rich documents (Aiello et al., 2002; Gao et al., 2011; Kim et al., 2022). Earlier research typically depends on Optical Character Recognition (OCR) to recognize textual content in the document images and employs graph neural networks or sequence labeling models to detect semantic relationships within the text (Palm et al., 2017; Yu et al., 2021; Wang et al., 2021; Hui et al., 2025). While effective, these methods rely heavily on accurate OCR results, leading recognition errors to cascade into field extraction (Kim et al., 2022; Cheng et al., 2022; Dhouib et al., 2023). Moreover, OCR text fails to adequately preserve the original layout features, which may weaken or omit layout cues beneficial for extraction (Lee et al., 2022; Liu et al., 2024b). While some works explore encoding layout features into language models to enable layout-aware language modeling (Xu et al., 2020; Wang et al., 2024; Huang et al., 2022), their performance remains tightly dependent on OCR accuracy (Zhang et al., 2025b; Barboule et al., 2025).

More recent efforts have explored using Large Multimodal Models (LMMs) to capture key information from document images (Ke et al., 2025; Wang et al., 2025a). Previous work adapts LMMs to text-rich documents by tailoring their visual encoders to effectively model dense text and complex layouts (Liu et al., 2024a; Yu et al., 2024; Zhang et al., 2025a). These methods typically employ filtering modules to remove content-irrelevant visual tokens (Liu et al., 2024a; Zhang et al., 2025a; Liu et al., 2024d) or aggregate multi-scale visual features (Park et al., 2024; Hu et al., 2025) to enhance the visual text perception of LMMs (Ding et al., 2024). Recent research focuses on optimizing LMMs with document-oriented training objectives to enhance their document understanding capability (Ye et al., 2023; Nacson et al., 2025). MiniCPM-V adopts an OCR-oriented pretraining strategy that corrupts textual regions in document images and trains the model to reconstruct the original text (Yu et al., 2025). Kosmos-2.5 and Qwen2.5-VL further convert document pages into Markdown or HTML codes, offering a unified alignment target for text recognition and layout perception (Lv et al., 2023; Bai et al., 2025b).

Despite recent progress in LMMs, evaluating their capabilities in KIE tasks remains challenging (Shi et al., 2023). Early KIE benchmarks focus on a specific document type, such as receipts, invoices, and orders (Huang et al., 2019; Šimsa et al., 2023; Abdallah et al., 2024). To enable broader evaluation, several recent document understanding benchmarks incorporate multiple KIE datasets into their evaluation suites (Ding et al., 2024; Ke et al., 2025). OCRBench introduces a dedicated KIE track that evaluates receipt and bill understanding by requiring LMMs to answer queries for predefined fields, thereby formulating KIE as a constrained question answering task over document images (Liu et al., 2024c; Fu et al., 2025). CC-OCR further expands the evaluation to business documents and requires LMMs to generate result dict conditioned on predefined field schemas, rather than answering independent queries field by field (Yang et al., 2025). Despite their contributions, these benchmarks primarily focus on homogeneous document types and narrow scenarios, limiting their ability to evaluate KIE performance on real-world documents comprehensively.

3 UniKIE-Bench: A Comprehensive Layout-Variant KIE Benchmark

In this section, we formalize the Key Information Extraction (KIE) task and introduce the evaluation tracks in Sec. 3.1. We then present detailed statistics of the proposed benchmark in Sec. 3.2. The data curation process is described in Sec. 3.3. Finally, we compare our benchmark with existing KIE benchmarks in Sec. 3.4.

3.1 Task Formulation of KIE

To address the KIE task, existing KIE benchmarks (Liu et al., 2024c; Fu et al., 2025) typically adopt a question-answering style formulation. Specifically, given a document image x and a set of target fields $\mathcal{F}$, the inquiry for each field $f \in \mathcal{F}$ is verbalized as a field-specific query $q(f)$, and the model $\mathcal{M}$ generates a response y_f to predict the value of each field independently:

$$y_f = \mathcal{M}(x, q(f)), \; f \in \mathcal{F}. \tag{1}$$

The final prediction $\mathbf{y}^{\mathrm{QA}}$ is obtained by aggregating all field-level outputs:

$$\mathbf{y}^{\mathrm{QA}} = \{f : y_f\}_{f \in \mathcal{F}}. \tag{2}$$

While simple and flexible, this formulation requires multiple independent inferences for KIE and fails to capture the structural relationships among fields.

In contrast, UniKIE-Bench formulates KIE as a schema-guided structured prediction task that performs extraction in one inference step. Each instance is represented by a schema $s = (\mathcal{F}, \mathcal{R})$, where $\mathcal{F}$ denotes the set of target fields, and $\mathcal{R}$ specifies the relations among them. The model is required to generate a structured output based on the document image x and the schema s:

$$\mathbf{y}^{\mathrm{SG}} = \mathcal{M}(x, s), \tag{3}$$

where the output $\mathbf{y}^{\mathrm{SG}}$ is a schema-aligned structured prediction that assigns a value to each field.

To comprehensively evaluate KIE capabilities, UniKIE-Bench introduces two complementary

Domain	Scenario	#Samples	#Fields
Business Transactions	Commercial	620	8.56
	Retail	347	4.00
	Catering Services	212	5.73
	Accommodation	40	7.00
Public Services	Administrative	385	5.83
	Education	320	3.01
	Postal Label	500	4.00
	Advertisement	71	3.65
Regulated Records	Tax-Compliant	987	8.00
	Medical Services	240	4.00
	Nutrition Label	750	8.43

Table 1: Dataset Statistics of the Constrained-Category KIE Track. **#Field** indicates the averaged field number.

Language	Type	#Samples	#Fields
Chinese	Receipt	202	9.88
	Form	208	17.75
	Invoice	207	12.11
	Contract	208	17.38
English	Receipt	207	13.61
	Form	210	15.90
	Invoice	212	17.36
	Contract	207	17.93

Table 2: Dataset statistics of the Open-Category KIE Track. **#Field** indicates the averaged field number.

evaluation tracks under a schema-guided task formulation. The *constrained-category KIE track* targets specific application scenarios, where documents are organized by scenario, and each scenario is associated with a small number of predefined extraction schemas. These schemas specify the target fields and output structures, reflecting practical extraction requirements in the real world.

In contrast, the *open-category KIE track* removes scenario-level assumptions by defining extraction schemas at the document level. Rather than relying on shared predefined schemas across documents, each document is paired with a unique schema derived from its own content. This design results in heterogeneous information demands and evaluates whether LMMs can accurately perform KIE over diverse and previously unseen field types.

3.2 Data Statistics of Different Evaluation Tracks of UniKIE-Bench

The data statistics of constrained-category and open-category KIE tracks in UniKIE-Bench are shown in Table 1 and Table 2, respectively.

The constrained-category track is organized under a hierarchical taxonomy spanning 3 domains and 11 real-world application scenarios, covering business, public service, and regulated document settings. The dataset exhibits substantial variation in both sample scale and schema complexity, with the average number of fields per document ranging from 3 to 9 across scenarios. Business transaction documents show moderate information requirements, while public service documents generally involve more compact schemas. In contrast, regulated records contain more fields on average. This distribution reflects realistic differences across application scenarios and enables fine-grained evaluation of model robustness under varying scenarios.

The open-category track organizes documents by language and document type. While both languages encompass the same set of document types, English documents consistently exhibit more complex schemas. This difference is particularly pronounced for Form and Invoice documents, in which the average number of fields exceeds 15. These statistics indicate systematic variations in information density across languages and document types, further highlighting the heterogeneous extraction requirements inherent in open-category KIE.

3.3 Data Curation of UniKIE-Bench

In this section, we describe the data curation process of UniKIE-Bench, which is designed to support a systematic evaluation under the schema-guided KIE formulation.

Constrained-Category KIE. We curate data from existing benchmarks and organize the constrained-category KIE track according to application scenarios.

Specifically, we collect document images from publicly available document understanding datasets, which are manually assigned to scenarios based on their functional roles and content characteristics. There are three domains (Business Transactions, Public Services, and Regulated Records) that are defined, which comprise 11 application scenarios that span a broad spectrum of real-world document types. Detailed data sources are provided in Appendix A.2. For each scenario, we identify common information requirements and consolidate key fields into scenario-predefined schemas. When annotations from source datasets are compatible with the schema-guided KIE formulation, they are mapped to the corresponding schema fields; otherwise, the documents are re-annotated accordingly. The detailed curation procedures for this track are described in Appendix A.3. To ensure high annota-

Benchmark	LMM-Ready	Track		Taxonomy		Size
		Constrained	Open	#Domains	#Scenarios	
OCRBench (Liu et al., 2024c)	✓	✓	✗	–	3	200
DocILE (Šimsa et al., 2023)	✗	✓	✗	1	–	600
OCRBenchV2 (Fu et al., 2025)	✓	✓	✗	–	6	800
RealKIE (Townsend et al., 2024)	✗	✓	✗	1	5	1,867
CC-OCR (Yang et al., 2025)	✓	✓	✓	–	6	2,008
UNIKIE-BENCH	✓	✓	✓	3	11	6,133

Table 3: Comparison between our UNIKIE-BENCH and Previous KIE Benchmarks. We conduct a detailed comparison of existing KIE benchmarks along four dimensions: support for end-to-end evaluation, task tracks, taxonomy, and dataset size. The symbol "–" indicates that the corresponding benchmark does not define or report.

tion quality for the constrained-category KIE track, we employ a multi-stage quality control process. For converted annotations, we manually verify field-level semantic consistency with the scenario schema. For re-annotated samples, we conduct model-assisted verification by leveraging advanced LMMs and resolving any discrepancies through human review.

Open-Category KIE. We categorize documents in the open-category KIE track along two dimensions, language and document type, resulting in a collection that spans 2 languages (English and Chinese) and 4 representative document types (receipt, form, invoice, and contract).

We collect documents via a document reconstruction pipeline grounded in real-world document examples. For each document type, we first curate a small set of representative documents from practical scenarios and include them as in-context examples in the prompt. The LMM is prompted conditioned on these examples to generate a document description that characterizes the content elements that typically appear in such documents. Based on this description, we further prompt an LLM to generate executable HTML code that instantiates the described content and layout, explicitly defining the document structure and the spatial arrangement of textual elements. Rendering the HTML code yields the final document images used in this track. To further enhance visual realism, we subsequently introduce lightweight noise that reflects real-world conditions. More curation details are provided in Appendix A.4, while additional authenticity analyses are presented in Appendix A.5. To get the ground truth key-value pairs, we follow Yang et al. (2025) to annotate documents in the open-category KIE track. Each document is processed with optical character recognition to obtain textual elements. Then we correct the OCR results, identify key information values, and organize them into a hierarchical extraction schema based on semantic relationships, such as grouping and containment.

3.4 Benchmark Comparison

We compare UNIKIE-BENCH with representative KIE benchmarks in Table 3 along four key dimensions: (i) support for end-to-end LMM evaluation, (ii) coverage of extraction tracks, (iii) document taxonomy design, and (iv) dataset scale.

As shown in Table 3, most existing KIE benchmarks exhibit notable limitations when applied to LMM evaluation. Specifically, DocILE and RealKIE depend on OCR annotations, rendering them unsuitable for end-to-end LMM evaluation. While OCRBench and OCRBenchV2 enable end-to-end evaluation, they mainly target constrained-category KIE via a question-answering formulation and offer either coarse-grained or no explicit document taxonomy. In contrast, UNIKIE-BENCH introduces a hierarchical document taxonomy encompassing 3 domains and 11 application scenarios, facilitating fine-grained analyses on different domains and scenarios. Moreover, UNIKIE-BENCH unifies constrained-category and open-category KIE tracks under a schema-guided formulation and scales the dataset to 6,133 documents, resulting in a more comprehensive and realistic benchmark for LMM-based KIE evaluation.

4 Evaluation Protocol

In this section, we describe the evaluation protocol of UNIKIE-BENCH, including the evaluation metrics, baseline models, and implementation details.

Baselines. We evaluate a diverse set of representative LMMs as baselines, covering both proprietary and open-source models with varying

Method	Business Transactions				Public Services				Regulated Records			Avg.
	Ret.	Cat.	Com.	Acco.	Post.	Admi.	Edu.	Adv.	Tax.	Med.	Nut.	
Close-source LMMs												
Claude-Sonnet-4.5	69.24	73.65	45.92	69.29	79.45	69.20	39.46	85.06	67.79	63.96	67.24	66.39
GPT-4o	75.79	79.33	49.67	67.86	82.95	64.26	47.77	73.56	75.33	67.57	76.51	69.15
GPT-5	71.90	71.33	45.94	69.29	80.95	62.79	50.26	75.48	72.05	67.71	75.83	67.59
Qwen-VL-Max	86.67	84.68	49.33	76.07	84.85	79.96	82.46	82.76	77.94	75.62	75.14	77.77
Qwen3-VL-Plus	89.12	86.45	50.57	77.86	83.55	78.91	87.44	83.14	78.93	76.46	89.72	80.20
Gemini-3-Pro	89.27	83.54	53.87	81.43	90.74	84.32	80.71	90.54	81.54	77.60	92.47	82.37
Open-source LMMs												
SmolVLM2-2.2B	17.80	57.22	28.52	18.92	34.98	8.63	3.87	27.16	34.97	18.00	10.51	23.69
Gemma-3-12B	54.32	68.73	30.68	53.93	72.30	37.14	17.56	36.78	56.81	48.54	52.20	48.09
InternVL3.5-8B	77.23	72.23	41.64	67.14	73.90	72.10	63.08	73.08	72.81	65.94	65.53	67.70
Ministral-3-8B	6.47	57.01	26.24	43.39	75.80	32.37	29.54	19.53	65.85	51.46	6.17	37.62
MiniCPM-V4.5-8B	68.37	74.91	45.91	67.76	83.60	71.89	63.21	79.31	69.57	70.00	44.18	67.15
GLM-4.1V-9B	62.32	77.33	46.67	76.43	76.46	78.75	65.83	81.99	70.68	72.08	60.47	69.91
Kimi-VL-A3B	86.49	63.12	42.87	61.79	83.40	68.72	80.37	72.41	77.80	64.93	25.72	66.14
MiMo-VL-7B-RL	77.45	77.66	49.01	67.50	76.70	76.81	76.65	66.28	74.86	68.70	46.10	68.88
Qwen3-VL-8B	88.26	86.23	51.14	71.07	83.80	78.90	87.39	81.23	78.74	77.19	86.36	79.12

Table 4: Overall Performance for Representative LMMs on the Constrained-Category KIE Track of UNIKIE-BENCH. Detailed model information is provided in Appendix A.7.

model scales, architectures, and training strategies. Specifically, the baselines include Gemini-3-Pro, Qwen-VL family (Bai et al., 2025b), GPT family, Claude-Sonnet-4.5, SmolVLM2-2.2B (Marafioti et al., 2025), Ministral-3-8B (Mistral AI, 2025), InternVL3.5-8B (Wang et al., 2025b), GLM-4.1V-9B (Team et al., 2025c), MiMo-VL-7B-RL (Team et al., 2025a), MiniCPM-V-4.5-8B (Yu et al., 2025), and Kimi-VL-A3B (Team et al., 2025b). More details are provided in Appendix A.7.

Evaluation Metrics. We evaluate KIE performance using a field-level F1 score, following prior work (Yang et al., 2025; Kim et al., 2022). Each schema-defined field is evaluated independently. A prediction is considered correct only if the extracted value exactly matches the corresponding ground-truth annotation; any character-level mismatch is treated as an error. We release the implementation of all normalization rules and metrics used during model evaluation as a part of our proposed UNIKIE-BENCH.

Implementation Details. For all evaluated LMMs, we adopt a unified prompt template across evaluation tracks to ensure fair comparison; the full prompt is provided in Appendix A.6. All document images are resized such that the total number of pixels does not exceed 1,605,632. Proprietary models are accessed via their official SDKs, while all open-source models are deployed using vLLM with Flash-Attention. We set the temperature to 0 for all models to eliminate sampling variability.

5 Results and Analysis

This section presents a comprehensive evaluation of representative LMMs on UNIKIE-BENCH, with in-depth analyses across diverse application scenarios, document types, and languages.

5.1 Overall Performance on Different KIE Tracks in UNIKIE-BENCH

This subsection presents the performance of various models on both the Constrained Category and Open Category KIE tracks in UNIKIE-BENCH.

KIE for Constrained Category. Table 4 displays the performance of representative LMMs on the constrained-category KIE track of UNIKIE-BENCH, with results across different scenarios.

A clear performance gap is observed between closed-source and open-source LMMs, underscoring the significance of strong multimodal alignment in KIE. Among the proprietary LMMs, Gemini-3-Pro achieves the best overall performance, followed by Qwen3-VL-Plus and Qwen-VL-Max, all of which exhibit stable and consistently strong results across business, public service, and regulated document scenarios. In contrast, most open-source models underperform, particularly in the public service and regulated document categories, where long text spans, domain-specific terminology, and implicit field dependencies impose higher demands on visual understanding. Despite similar model scales, open-source LMMs show substantial performance variation, suggesting that scale alone does

Method	Chinese				English				Avg.
	Receipt	Form	Invoice	Contract	Receipt	Form	Invoice	Contract	
Closed-source LMMs									
Claude-Sonnet-4.5	37.59	31.20	29.12	20.88	41.64	49.22	44.18	63.87	39.71
GPT-4o	39.66	30.82	35.31	16.67	78.51	48.09	57.08	67.35	46.69
GPT-5	39.61	31.97	37.64	17.47	80.64	50.09	57.50	64.24	47.39
Qwen-VL-Max	69.94	69.75	66.62	69.68	70.76	65.52	70.80	76.57	69.95
Qwen3-VL-Plus	71.14	69.87	69.59	70.95	68.27	66.82	73.45	74.81	70.61
Gemini-3-Pro	78.50	79.17	78.63	75.96	93.57	80.41	87.06	79.90	81.65
Open-source LMMs									
SmolVLM2-2.2B	5.74	1.51	3.05	2.15	36.49	11.64	19.73	22.36	12.83
Gemma-3-12B	25.71	14.17	17.36	9.42	76.24	38.32	30.23	35.77	30.90
InternVL3.5-8B	54.26	37.83	40.97	39.43	79.46	43.64	49.41	50.06	49.38
Ministral-3-8B	25.21	35.80	24.42	26.18	58.54	49.01	55.66	57.82	41.58
MiniCPM-V4.5-8B	52.33	42.69	44.54	39.46	72.41	47.46	56.44	65.49	52.60
GLM-4.1V-9B	55.19	56.63	55.52	54.68	63.83	54.66	59.31	68.70	58.56
Kimi-VL-A3B	62.98	52.85	50.72	54.94	73.91	55.96	52.67	64.64	58.58
MiMo-VL-7B-RL	57.63	56.60	51.93	65.90	71.34	52.83	57.64	70.55	60.55
Qwen3-VL-8B	68.93	64.70	63.38	65.59	71.40	61.46	67.34	76.08	67.36

Table 5: Overall Performance of Representative LMMs on the Open-Category KIE Track of UNIKIE-BENCH. Detailed model information is provided in Appendix A.7.

not account for the observed differences. Notably, Qwen3-VL-8B emerges as a strong open-source baseline, approaching the performance of some proprietary LMMs on average, but still shows degradation in the most challenging categories.

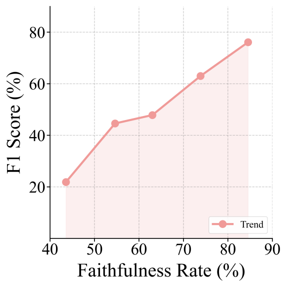
KIE for Open Category. Table 5 presents the overall performance of representative LMMs on the Open-Category KIE track.

In general, closed-source models outperform their open-source counterparts, with Gemini-3-Pro achieving the highest average F1 score and demonstrating robust performance across both languages and document types. We observe a consistent and substantial performance drop from English to Chinese across most models, highlighting a persistent cross-lingual robustness challenge in open-category KIE. This disparity is likely influenced by the dominance of English document distributions in vision-language pretraining. Additionally, the compact glyph structure, high character density, and absence of explicit word boundaries in Chinese texts further complicate visual recognition and boundary alignment, which negatively impacts KIE performance. The performance varies significantly across different document types. Specifically, receipt documents consistently present the least difficulty for all models, yielding the highest scores due to their relatively regular layouts and well-defined key–value structures. In contrast, form documents show the greatest challenge, especially for smaller and open-source models. This suggests that the fragmented layouts and flexible field boundaries in forms make field localization and extraction more prone to errors. Invoice and contract documents show intermediate performance but exhibit different behaviors across languages: while contracts perform reasonably well in English, Chinese contracts remain particularly challenging, especially for smaller-scale LMMs.

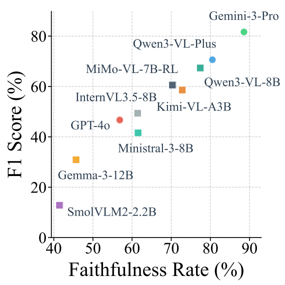
5.2 Faithfulness Analysis of LMMs in KIE

We further investigate the faithfulness of LMMs in KIE by examining whether the predicted field values are grounded in the input document image, rather than arising from visual perception errors or hallucinated content. Specifically, we conduct this analysis on the open-category KIE track, which benefits from validated OCR results and thus enables a reliable estimation of hallucination behaviors in LMMs. An extracted value is considered faithful if it can be located in the OCR results of the corresponding document.

As shown in Figure 2a, when averaged across all evaluated models, higher F1 scores are generally associated with higher faithfulness rates. This positive correlation indicates that stronger KIE performance is typically accompanied by more reliable grounding of predicted field values in the document content. Figure 2b further provides a fine-grained analysis across different LMMs and reveals notable variations among models. Importantly, even among LMMs exhibiting comparable



(a) Correlation between the Faithfulness Rate and F1 Score.



(b) F1 Score and Faithfulness Rate across Different LMMs.

Figure 2: Faithfulness Analysis of LMMs in KIE. We analyzed the relationship between the faithfulness and the extraction performance of LMMs in KIE.

faithfulness rates, we observe substantial disparities in F1 scores. This observation suggests that while faithfulness constitutes a necessary prerequisite for conducting accurate KIE results, it is not sufficient on its own to guarantee strong extraction performance. The primary reason is that, even when models exhibit similar grounding behaviors, they can still differ substantially in their ability to capture field semantics, precisely delineate field boundaries, and correctly select the target instance among multiple candidates.

5.3 Typical Error Analysis

To better understand the remaining performance gaps, we conduct a typical error analysis. As illustrated in Figure 3, we identify several recurring error patterns, which shed light on the systematic limitations of current LMMs in key information extraction. A more detailed behavioral analysis of LMMs in KIE is provided in Appendix A.8.

Visual Perception Failure. In this error category, the model fails at the level of visual text recognition. For the tax field, the ground-truth value shown in the document is 2.32, while the model predicts 4.32. Such errors mainly arise from imperfect visual perception (e.g., digit confusion or misrecognition), rather than from semantic misunderstanding of the field.

Layout Perception Failure. These errors occur when the model correctly recognizes textual content but fails to associate fields with their corresponding spatial regions. For example, when extracting an address field, the model selects an incorrect nearby line. This error pattern indicates deficiencies in layout understanding, where the model struggles to align a target field with the correct text span when layout boundaries are ambiguous.

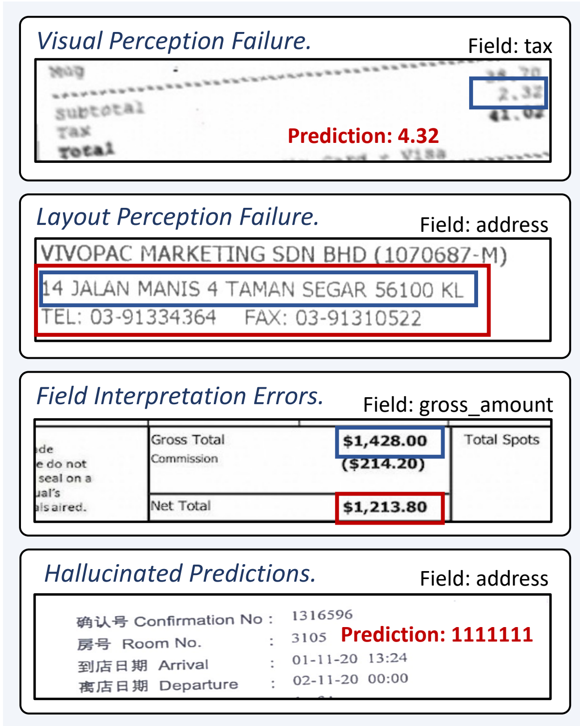


Figure 3: Typical Error Cases of LMMs in KIE. Blue boxes indicate the ground truth, while red boxes denote the model predictions.

Field Interpretation Errors. This error type refers to cases where the model successfully localizes relevant values but assigns them to incorrect fields. In our examples, the model confuses the semantic roles of fields, mapping the net total to the gross amount. Such errors reveal limitations in fine-grained field understanding, especially in scenarios requiring domain-specific knowledge or nuanced distinctions between closely related fields.

Hallucinated Predictions. Hallucinated errors correspond to predictions that are not grounded in the document content. In the illustrated case, the model outputs an address value that does not appear anywhere in the input image. Unlike perception or interpretation errors, these predictions are fabricated without visual or textual evidence. This error pattern highlights the challenge of maintaining faithfulness in KIE, particularly when target fields are missing and incomplete.

6 Conclusion

We present UNIKIE-BENCH, a unified benchmark for evaluating KIE in realistic and heterogeneous document settings. Experiments on representative LMMs show that existing models still face challenges when confronted with diverse extraction requirements, long-tail fields, and complex document layouts. We hope that UNIKIE-BENCH can serve as a reliable and practical evaluation to facilitate the development of more robust and faithful document understanding models.

Limitations

While UNIKIE-BENCH offers a systematic and comprehensive evaluation of the KIE capabilities of current LMMs, it is restricted to single-page or short-document scenarios and does not address long-document KIE. This limitation arises because long-document KIE typically depends on auxiliary components (*e.g.*, page retrieval and cross-page aggregation), whose performance often dominates overall results under the current context-length constraints of LMMs. As a consequence, isolating and evaluating the intrinsic KIE ability of LMMs in long-document settings becomes challenging. To facilitate a more controlled, interpretable, and focused assessment of the core KIE capabilities of LMMs, UNIKIE-BENCH deliberately confines its evaluation scope to short documents.

Ethical Considerations

All documents included in UNIKIE-BENCH are sourced either from publicly available datasets or generated through carefully designed privacy-preserving pipelines. Throughout the construction of the benchmark, strict measures are taken to ensure that no personal, sensitive, or identifiable information is collected, processed, or disclosed. A portion of the dataset is annotated by the collaborators of this work. No external annotators, crowd-workers, or paid participants are involved; all annotations are conducted voluntarily by contributors with relevant expertise. Before the annotation process, all annotators are fully informed of the research objectives, the intended use of the annotated data, and the evaluation protocols of the benchmark. To ensure consistency and reliability, we provide comprehensive annotation guidelines that clearly specify task definitions, annotation formats, and operational constraints. Importantly, the annotation process for key information extraction from documents does not constitute human-subjects research. It does not involve interaction with, or data collection from, external individuals, and therefore does not raise concerns related to human participant protection or personal data usage.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (No. 62576082 and No. 62461146205) and Alibaba Innovative Research Program.

References

Abdelrahman Abdallah, Daniel Eberharter, Zoe Pfister, and Adam Jatowt. 2024. A survey of recent approaches to form understanding in scanned documents. *Artificial Intelligence Review*, page 342.

Marco Aiello, C Monz, L Todoran, and M Worring. 2002. Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition*, pages 1–16.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, and 1 others. 2025a. Qwen3-vl technical report. *ArXiv preprint*, abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-vl technical report. *ArXiv preprint*, abs/2502.13923.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025c. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of ACL*, pages 3639–3664.

Camille Barboule, Benjamin Piwowarski, and Yoan Chabot. 2025. Survey on question answering over visually rich documents: Methods, challenges, and trends. *ArXiv preprint*, abs/2501.02235.

Francesca Cesarini, Marco Gori, Simone Marinai, and Giovanni Soda. 2002. Informys: A flexible invoice-like form-reader system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 730–745.

Zhanzhan Cheng, Peng Zhang, Can Li, Qiao Liang, Yunlu Xu, Pengfei Li, Shiliang Pu, Yi Niu, and Fei Wu. 2022. Trie++: towards end-to-end information extraction from visually rich documents. *ArXiv preprint*, abs/2207.06744.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261.

Mohamed Dhouib, Ghassen Bettaieb, and Aymen Shabou. 2023. Docparser: End-to-end ocr-free information extraction from visually rich documents. In *Proceedings of ICDAR*, pages 155–172.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. 2024. Deep learning based visually rich document content understanding: A survey. *ArXiv preprint*, abs/2408.01287.

Sayna Ebrahimi, Sercan O Arik, and Tomas Pfister. 2022. Test-time adaptation for visual document understanding. *ArXiv preprint*, abs/2206.07240.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, and 1 others. 2025. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *ArXiv preprint*, abs/2501.00321.

Liangcai Gao, Zhi Tang, Xiaofan Lin, Ying Liu, Ruiheng Qiu, and Yongtao Wang. 2011. Structure extraction from pdf-based book documents. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 11–20.

Hamza Gbada, Karim Kalti, and Mohamed Ali Mahjoub. 2025. Deep learning approaches for information extraction from visually rich documents: Datasets, challenges and methods. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 121–142.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *ArXiv preprint*, abs/2406.20094.

Jiabang He, Yi Hu, Lei Wang, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023. Do-good: towards distribution shift evaluation for pre-trained visual document understanding models. In *Proceedings of SIGIR*, pages 569–579.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of ACL*, pages 5817–5834.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of MM*, pages 4083–4091.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *Proceedings of ICDAR*, pages 1516–1520.

Yangyang Hui, Jiahao Liu, Qi Zhang, Tianyi Zhou, and Yonghong Song. 2025. Seg-doc: A simple yet efficient graph neural network framework for document key information extraction. *Neurocomputing*, page 130493.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6.

Wenjun Ke, Yifan Zheng, Yining Li, Hengyuan Xu, Dong Nie, Peng Wang, and Yao He. 2025. Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends. *ACM Transactions on Information Systems*, pages 1–64.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Proceedings of ECCV*, pages 498–517.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: practical dataset and end-to-end solution. In *Proceedings of ICDAR*, pages 36–53.

Seif Laatiri, Pirashanth Ratnamogan, Joël Tang, Laurent Lam, William Vanhuffel, and Fabien Caspani. 2023. Information redundancy and biases in public document information extraction benchmarks. In *Proceedings of ICDAR*, pages 280–294.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. *ArXiv preprint*, abs/2203.08411.

Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. 2024a. Hrvda: High-resolution visual document assistant. In *Proceedings of CVPR*, pages 15534–15545.

Shuhang Liu, Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Qing Wang, Jianshu Zhang, and Chenyu Liu. 2024b. See then tell: Enhancing key information extraction with vision grounding. *ArXiv preprint*, abs/2409.19573.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, page 220102.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024d. Textmonkey: An ocr-free large multimodal model for understanding document. *ArXiv preprint*, abs/2403.04473.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, and 1 others. 2023. Kosmos-2.5: A multimodal literate model. *ArXiv preprint*, abs/2309.11419.

Song Mao, Azriel Rosenfeld, and Tapas Kanungo. 2003. Document structure analysis algorithms: a literature survey. *Document recognition and retrieval X*, pages 197–207.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *ArXiv preprint*, abs/2504.05299.

Mistral AI. 2025. Introducing mistral 3.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. 2025. Docvlm: Make your vlm an efficient reader. In *Proceedings of CVPR*, pages 29005–29015.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, pages 1931–1990.

Berke Oral, Erdem Emekligil, Seçil Arslan, and Gülşen Eryiğit. 2020. Information extraction from text intensive and visually rich banking documents. *Information Processing & Management*, page 102361.

Rasmus Berg Palm, Ole Winther, and Florian Laws. 2017. Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. In *Proceedings of ICDAR*, pages 406–413.

Jaeyoo Park, Jin Y Choi, Jeonghyung Park, and Bohyung Han. 2024. Hierarchical visual feature aggregation for ocr-free document understanding. In *Proceedings of NeurIPS*, pages 105972–105996.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Alexander Michael Rombach and Peter Fettke. 2025. Deep learning based key information extraction from business documents: Systematic literature review. *ACM Computing Surveys*, pages 1–37.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *ArXiv preprint*, abs/2310.16809.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and 1 others. 2023. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer.

Matyáš Skalickỳ, Štěpán Šimsa, Michal Uřičář, and Milan Šulc. 2022. Business document information extraction: Towards practical benchmarks. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 105–117. Springer.

Yuan Yan Tang, Chang De Yan, and Ching Y. Suen. 1994. Document processing for automatic knowledge acquisition. *IEEE transactions on Knowledge and Data Engineering*, pages 3–21.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, and 1 others. 2025a. Mimo-vl technical report. *ArXiv preprint*, abs/2506.03569.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025b. Kimi-vl technical report. *ArXiv preprint*, abs/2504.07491.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, and 1 others. 2025c. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *ArXiv preprint*, abs/2507.01006.

Benjamin Townsend, Madison May, Katherine Mackowiak, and Christopher Wells. 2024. Realkie: Five novel datasets for enterprise key information extraction. *ArXiv preprint*, abs/2403.20101.

Mahsa Vafaie, Sven Hertling, Inger Banse-Strobel, Kevin Dubout, and Harald Sack. 2025. End-to-end information extraction from archival records with multimodal large language models. In *Proceedings of CIKM*, pages 6075–6083.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of ACL*, pages 8529–8548.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards robust visual information extraction in real world: New dataset and novel solution. In *Proceedings of AAAI*.

Weishi Wang, Hengchang Hu, Zhijie Zhang, Zhaochen Li, Hongxin Shao, and Daniel Dahlmeier. 2025a. Document intelligence in the era of large language models: A survey. *ArXiv preprint*, abs/2510.13366.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv preprint*, abs/2508.18265.

Kaiwen Wei, Jie Yao, Jiang Zhong, Yangyang Kang, Jingyuan Zhang, Changlong Sun, Xin Zhang, Fengmao Lv, and Li Jin. 2025. P^{2}net: Parallel pointer-based network for key information extraction with complex layouts. In *Proceedings of ACL Findings*, pages 10611–10626.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of SIGKDD*, pages 1192–1200.

Zhibo Yang, Rujiao Long, Pengfei Wang, Sibo Song, Humen Zhong, Wenqing Cheng, Xiang Bai, and Cong Yao. 2023. Modeling entities as semantic points for visual information extraction in the wild. In *Proceedings of ICCV*.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, and 1 others. 2025. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In *Proceedings of ICCV*, pages 21744–21754.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, and 1 others. 2023. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Proceedings of EMNLP Findings*, pages 2841–2858.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *ArXiv preprint*, abs/2509.18154.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2021. Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In *Proceedings of ICPR*, pages 4363–4370.

Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, Yuning Du, Shikun Feng, Xiaoguang Hu, Pengyuan Lyu, Kun Yao, Yuechen Yu, Yuliang Liu, Wanxiang Che, Errui Ding, and 8 others. 2023. Icdar 2023 competition on structured text extraction from visually-rich document images. *ArXiv preprint*, abs/2306.03287.

Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *ArXiv preprint*, abs/2404.09204.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *Proceedings of EMNLP*, pages 13716–13730.

Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2025a. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. In *Proceedings of AAAI*, pages 9923–9932.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025b. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. In *Proceedings of ICCV*, pages 17443–17453.

A Appendix

A.1 License

We strictly comply with the original licenses released with each dataset used to build UNIKIE-BENCH and do not redistribute any third-party raw document images. Among the datasets used in this work, DocILE and SROIE are released under the MIT License; CORD and SIBR are distributed under the CC BY-SA 4.0 License; and Nanonets-KIE is provided under the Apache License 2.0. In addition, FUNSD, EPHOIE, HW-FORMS, and CELL are restricted to non-commercial academic use and are included in this study in accordance with their original licensing terms. DeepForm is subject to the ProPublica Terms of Use[1]. For datasets without an explicitly stated standard license, we follow the original release terms, use them solely for research evaluation, and do not redistribute the raw data.

A.2 Data Sources for Documents in the Constrained-Category KIE Track

The document images in the constrained-category KIE track are collected from multiple existing public document understanding datasets. We detail the data sources in Table 6. Specifically, we curate documents across 3 domains: Business Transactions, Public Services, and Regulated Records, and further organize them into 11 representative application scenarios.

For Business Transactions, we include documents from commercial, retail, catering, and accommodation scenarios. These documents are mainly drawn from SIBR (Yang et al., 2023), DocILE (Šimsa et al., 2023), SROIE (Huang et al., 2019), CORD (Park et al., 2019), and CELL (Yu et al., 2023), covering typical transactional records such as receipts, invoices, and service-related documents. The Public Services domain focuses on administrative and societal documents, including administrative forms, educational documents, postal labels, and advertisements. Data in this domain are collected from CELL (Yu et al., 2023), FUNSD (Jaume et al., 2019), EPHOIE (Wang et al., 2021), HW-FORMS[2], and DeepForm[3]. Compared to business documents, these samples exhibit greater layout diversity and more heterogeneous information requirements. The Regulated Records domain consists of documents subject to stricter formatting or compliance requirements, including tax-compliant documents, medical service records, and nutrition labels. We filter these documents from Nanonets-KIE[4], SIBR (Yang et al., 2023), and POIE (Kuang et al., 2023). This domain typically involves complex schemas, fine-grained field definitions, and higher sensitivity to extraction errors, making it particularly challenging for end-to-end KIE systems.

A.3 Annotation Details for the Constrained-Category KIE Track

To curate annotations for the constrained-category KIE track, we adopt a systematic annotation curation pipeline. Specifically, we distinguish between datasets with existing KIE annotations and those that provide only OCR annotations, and apply specific processing strategies accordingly.

For datasets with existing KIE annotations, we directly use the raw field annotations or map them to our schema. However, the quality of these annotations varies substantially across datasets, with common issues including annotated field values that do not appear in the document, inconsistent capitalization, overlaps with irrelevant text, and incomplete or partially missing values. To address these issues, we correct misannotated field values, remove overlaps with irrelevant text regions, normalize formatting inconsistencies, and supplement missing values when they can be unambiguously inferred from the document. Annotations that remain ambiguous or unsupported are excluded to ensure high annotation reliability. For datasets that only provide OCR annotations that can not be converted to the end-to-end KIE formulation, we manually re-annotate them based on the scenario-predefined schema we identified before.

A.4 Data Curation Details for the Open-Category KIE Track

To construct the open-category KIE track in our benchmark, we utilize a document reconstruction pipeline to collect documents, ensuring the diversity and realism of document content and layouts while avoiding privacy and copyright concerns inherent in real-world documents.

For each language and document type, we curate representative real-world documents from practical

[1] https://projects.propublica.org/datastore/terms

[2] https://huggingface.co/datasets/ift/handwriting_forms

[3] https://github.com/jstray/deepform

[4] https://huggingface.co/datasets/nanonets/key_information_extraction

Domain	Scenario	Abbreviation	Data Source	#Test Samples	Ratio
Business Transactions	Commercial	Com.	SIBR, DocILE	620	13.86%
	Retail	Ret.	SROIE	347	7.76%
	Catering Services	Cat.	CORD, CELL	212	4.74%
	Accommodation	Acco.	SIBR	40	0.89%
Public Services	Administrative	Admi.	CELL, FUNSD	385	8.61%
	Education	Edu.	EPHOIE, CELL	320	7.16%
	Postal Label	Post.	HW-FORMS	500	11.18%
	Advertisement	Adv.	DeepForm	71	1.59%
Regulated Records	Tax-Compliant	Tax.	Nanonets-KIE	987	22.07%
	Medical Services	Med.	SIBR	240	5.37%
	Nutrition Label	Nut.	POIE	750	16.77%

Table 6: Dataset Sources for the Constrained-Category KIE Track of UNIKIE-BENCH.

scenarios and use them as in-context demonstrations, which are selected to reflect the typical content elements of the target document type. These documents are not reused in the benchmark and are solely intended to anchor the generation in realistic document patterns. Then, we prompt GPT-4o to generate a description for the target document type based on these demonstrations. These descriptions characterize the types of content elements and their structural organization (*e.g.*, headers, keys, tables, and free-form text blocks), without specifying concrete field values or copying real-world content, preventing the leakage of sensitive or proprietary information. Based on the description, we further prompt the model to instantiate a concrete document by generating executable HTML that explicitly specifies both the textual elements and their layout. To enhance the diversity of the documents, we follow Ge et al. (2024) and assign a distinct persona to each sample, conditioning generation on persona-specific attributes so that the resulting documents exhibit richer variation in content and presentation. We then render the HTML code into document images and remove samples with obvious rendering failures, obtaining the document images used in the open-category KIE track.

We further introduce lightweight noise into the rendered document images to better approximate the imperfections commonly observed in real-world scanned or photographed documents. Specifically, we first model the document images as 3D objects in Blender[5] to simulate diverse lighting conditions and crease effects encountered in real-world scenarios. This 3D-based rendering process closely mirrors practical document acquisition settings, enabling synthesized images to capture realistic illumination variations, shading patterns, and surface deformations that are difficult to reproduce with purely 2D augmentations. This yields visual characteristics that are nearly indistinguishable from those observed in real scanned or photographed documents. We then capture the document images using a vertically positioned camera and apply a diverse set of perturbations that simulate common acquisition and printing artifacts, including motion blur and mild Gaussian blur to mimic camera shake or defocus, elastic distortions to emulate paper deformation during scanning, and slight perspective transformations to reflect imperfect capture angles. Additionally, we introduce printer-related artifacts, including thin ink streaks, faded or broken ink regions, subtle variations in ink color, and horizontal noise patterns that resemble dirty drum effects. These perturbations are applied with low intensity and carefully constrained so as not to alter textual content or layout. Samples in which noise perturbations cause excessive degradation to textual content are manually identified and removed, ensuring that all retained documents remain legible and semantically reliable.

A.5 Detailed Analysis of the Documents in the Open-Category KIE Track

The open-category KIE track is constructed from automatically generated document images, which naturally raises questions about whether these documents adequately capture the realism and diversity of real-world inputs encountered in practical KIE scenarios. To address this concern, we conduct a comprehensive analysis from multiple complementary perspectives, examining document realism, content and field diversity, as well as layout variability.

Authenticity Analysis. Figure 4a reports the authenticity judgments on the 200 synthesized doc-

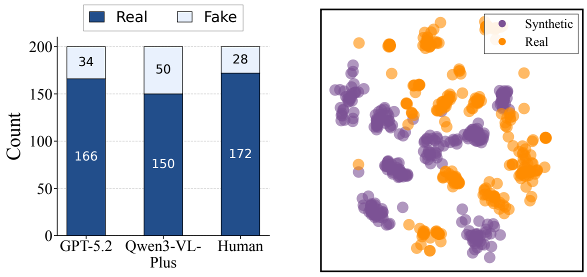
[5] `https://www.blender.org`

ument images in the open-category KIE track. Each evaluator is asked to determine whether a document is real or generated, without access to any auxiliary information. We find that only a limited number of synthesized documents are correctly identified as generated. Specifically, GPT-5.2 correctly classifies 34 out of the 200 synthesized documents as generated, while Qwen3-VL-Plus correctly identifies 50 synthesized documents. Human evaluators recognize 28 synthesized documents as generated. The majority of synthesized documents are instead judged as real, indicating that they are perceptually similar to real-world documents.

In addition to subjective authenticity judgments, we further investigate the distributional similarity between synthesized documents and real documents from the representation perspective, as shown in Figure 4b. Specifically, we employ a multimodal embedding model, gme-Qwen2-VL-2B-Instruct[6], to extract document embeddings and project both real and synthesized documents into a shared low-dimensional embedding space for visualization and comparison. The real documents are sampled from the constrained-category KIE track. The results show that the two types of documents are largely intermingled, rather than forming clearly separable clusters. This observation indicates that the synthesized documents closely resemble real documents in terms of their global visual–semantic representations. These findings provide additional evidence that the generated documents achieve a strong distributional alignment with real-world documents.

Semantic Analysis. We further investigate whether the open-category KIE track exhibits sufficient semantic diversity, rather than degenerating into a narrow set of templated field schemas or repetitive content patterns. To this end, we analyze document semantics from two complementary perspectives—field-level distributions and content-level distributions—as illustrated in Figure 5. In both analyses, we employ the Qwen3-Embedding model[7] to encode field values and full document contents into a shared semantic embedding space.

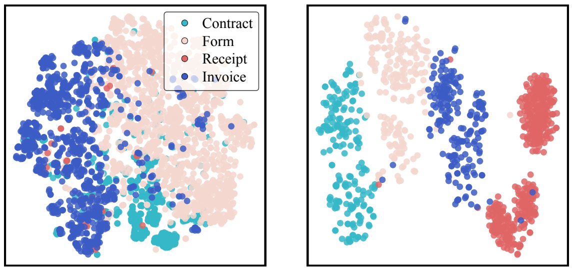
Figure 5a visualizes the distribution of document fields in a shared semantic embedding space. We observe that fields associated with different doc-



(a) Document Authenticity Analysis by Human and Advanced LMMs.

(b) Visualization of Document-Level Embeddings for Synthesized and Real Documents.

Figure 4: Document Authenticity Analysis in the Open-Category KIE Track.



(a) Visualization of Document Field Distribution in the Open-Category KIE Track.

(b) Visualization of Document Content Distribution in the Open-Category KIE Track.

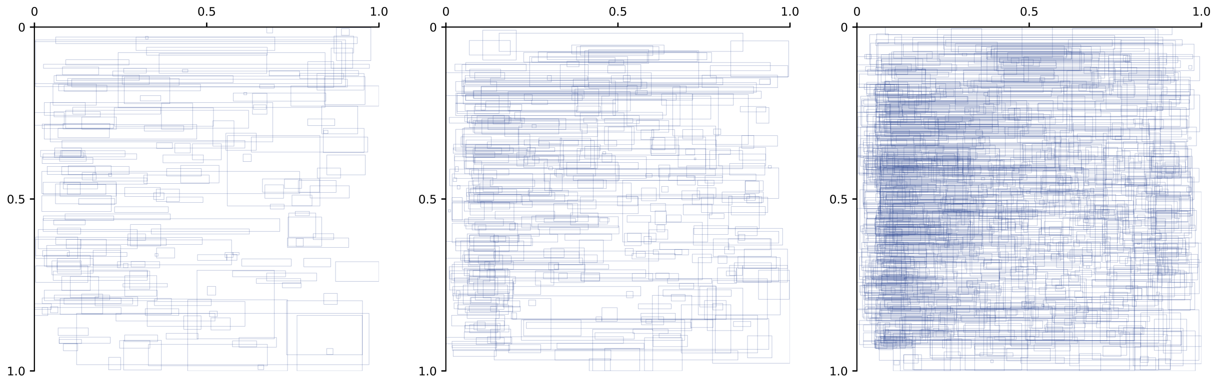
Figure 5: Diversity Analysis of Document Images in the Open-Category KIE Track.

ument types are largely intermingled rather than forming isolated, document-type-specific clusters. This pattern suggests that the open-category KIE track does not rely on rigid field templates tied to individual document categories. Instead, similar field semantics naturally occur across heterogeneous document types, while documents of the same type can exhibit diverse field compositions. Such behavior aligns with real-world KIE scenarios, where field definitions are flexible and often context-dependent.

Figure 5b presents the distribution of document content embeddings. In contrast to the field-level visualization, content representations exhibit clearer clustering across document types, reflecting semantic differences in document content. Each cluster shows substantial intra-class dispersion, indicating that the generated documents avoid trivial templating and instead cover a broad range of lexical and semantic variations within each category. This combination of inter-class separability and intra-class diversity suggests that the synthesized

[6]https://huggingface.co/Alibaba-NLP/gme-Qwen2-VL-2B-Instruct

[7]https://huggingface.co/Qwen/Qwen3-Embedding-8B



(a) N=50 (b) N=100 (c) N=300

Figure 6: Visualization of Document Layout Diversity in the Open-Category KIE Track.

documents capture meaningful content diversity.

Layout Analysis. We analyze layout diversity by visualizing where text regions appear on the document page. Figure 6 overlays normalized bounding boxes of text regions sampled from the synthesized documents into a shared coordinate space. Each subfigure corresponds to a different number of sampled bounding boxes N, allowing us to progressively reveal the underlying layout distribution as more text regions are included.

When N is small, the visualization already exhibits multiple horizontal text lines and block-level regions distributed across different positions. As N increases, these patterns become more pronounced: dense horizontal bands emerge, reflecting line-based text layouts commonly observed in real documents, while the wide spread in bounding box locations and sizes indicates substantial variation in layout structure. Importantly, even with a large number of bounding boxes overlaid, the layout does not collapse into a single dominant template.

A.6 The Prompt Template Used in UNIKIE-BENCH

To ensure fairness and reproducibility, we use the same prompt template for all evaluated tracks and models as shown in Figure 7. Specifically, we modify the prompt used by Yang et al. (2025) and remove explicit descriptions of the field, thereby requiring the model to infer the semantics of each field solely from its name and context.

A.7 Detailed LMM Configurations

In this section, we provide detailed information for the evaluated LMMs and report additional results on more models. Specifically, we show the model vendors, identifiers, and release dates in Table 7, ensuring transparency and reproducibility of our evaluation. For closed-source models, the identifiers correspond to the official names used in their SDKs, while for open-source models, they refer to the corresponding repositories on HuggingFace[8]. All open-source models are deployed on two NVIDIA A100 GPUs.

A.8 Behavioral Analysis for LMMs in KIE

In this subsection, we conduct an in-depth investigation of LMM behaviors when performing the KIE task. Specifically, we take Qwen3-VL-8B-Instruct (Bai et al., 2025a) as a representative model and analyze its attention patterns layer by layer, aiming to elucidate the behavioral mechanisms underlying the extraction process.

In the first scenario, as illustrated in Figure 8, where the field name is explicitly visible in the document, the model demonstrates a clear layer-wise transition in its attention patterns. In the shallow layers, attention is predominantly concentrated on the field label itself, suggesting that the model initially searches for and anchors to explicit field cues present in the document. As processing progresses into deeper layers, the model's attention gradually shifts away from the field label toward the corresponding field value. This transition indicates that the model first establishes a semantic anchor via the field cue and then leverages this anchor to locate and extract the associated value, thereby completing the extraction process in a relatively direct and localized manner.

In contrast, the second scenario corresponds to cases where the field name does not explicitly appear, and the target value is implicitly embedded within a broader textual context, as illustrated in

[8]https://huggingface.co/

```
<image>
Suppose you are an information extraction expert.
Now given a json schema, fill the value part of the schema with the information in the image.
Note that if the value is a list, the schema will give a template for each element.
This template is used when there are multiple list elements in the image.
Finally, only legal json is required as the output.
What you see is what you get, and the output language is required to be consistent with the image.
No explanation is required. Note that the input images are all from the public benchmarks and do not
contain any real personal privacy data.
Please output the results as required. The input json schema content is as follows:
<schema>
```

Figure 7: The Prompt Template for LMMs Used in UNIKIE-BENCH.

Model	**Vendor**	**Identifier**	**Release Date**
Closed-source LMMs			
GPT-5	OpenAI	gpt-5-2025-08-07	2025-08
GPT-4o	OpenAI	gpt-4o-2024-08-16	2024-05
Gemini-3-Pro	Google	gemini-3-pro-preview	2025-11
Claude-Sonnet-4.5	Anthropic	claude-sonnet-4-5-20250929	2025-09
Qwen3-VL-Plus	Alibaba	qwen3-vl-plus-2025-09-23	2025-09
Qwen-VL-Max	Alibaba	qwen-vl-max-2025-08-13	2025-08
Open-source LMMs			
MiniCPM-V-4.5-8B	OpenBMB	OpenBMB/MiniCPM-V-4_5	2025-08
Gemma-3-12B	Google	google/gemma-3-12b-it	2025-03
InternVL3.5-8B	OpenGVLab	OpenGVLab/InternVL3_5-8B-Instruct	2025-08
GLM-4.1V-9B	Zhipu AI	zai-org/GLM-4.1V-9B-Thinking	2025-07
Kimi-VL-A3B	Kimi	moonshotai/Kimi-VL-A3B-Instruct	2025-04
MiMo-VL-7B-RL	Xiaomi	XiaomiMiMo/MiMo-VL-7B-RL	2025-06
Qwen3-VL-8B	Alibaba	Qwen/Qwen3-VL-8B-Instruct	2025-10
SmolVLM2-2.2B	HuggingFace	HuggingFaceTB/SmolVLM2-2.2B-Instruct	2025-02
Ministral-3-8B	Mistral	mistralai/Ministral-3-8B-Instruct-2512	2025-12

Table 7: Model Information of the Evaluated LMMs.

the Figure 9. In this scenario, shallow-layer attention is predominantly allocated to approximate or semantically related field-like phrases, suggesting that the model initiates the extraction process by hypothesizing and probing potential semantic anchors in the absence of explicit field cues. At deeper layers, the model's attention no longer converges on a single location; instead, it is distributed across multiple candidate text spans that potentially contain the target value. This behavior suggests that the model engages in implicit candidate aggregation and comparison, inferring the correct extraction primarily through contextual reasoning rather than relying on explicit field cues.

A.9 Examples from UNIKIE-BENCH

To offer a more concrete perspective on the challenges posed by UNIKIE-BENCH, we present several representative document examples in this subsection. These examples span multiple evaluation tracks, document types, and application scenarios, collectively demonstrating the extensive visual, structural, and semantic diversity captured by the benchmark.

Figure 10 shows a representative receipt example from the open-category KIE track. The document exhibits a typical real-world receipt layout, with dense text, multiple itemized entries, and both explicit structural cues and implicit information embedded in free-form text. Given the input schema, the model is required to extract receipt-level metadata as well as variable-length item records. As illustrated by the prediction, even strong LMMs may produce subtle field interpretation errors, such as confusing quantities with unit prices, highlighting the difficulty of robust structured extraction under realistic layouts.

Figure 11 presents an example from the Form document type in the open-category KIE track. In contrast to receipts, this document exhibits a highly structured, multi-section layout, with fields distributed across well-defined regions, including free-text inputs, tabular expense records, and checkbox-based selections. This document type presents a distinct challenge for LMMs, as successful extraction requires not only accurate textual recognition but also precise layout-aware reasoning over heterogeneous field types. As illustrated by the example prediction, the model correctly recovers most fields, yet exhibits layout perception errors on selection-based fields, such as misidentifying or over-selecting checkbox options. These failure cases indicate that LMMs struggle to reliably associate visual selection cues with fields, underscoring the critical role of fine-grained layout understanding in KIE.

Figure 12 illustrates an example from the constrained-category KIE track under the advertisement scenario. The document is visually dense and information-rich, integrating broadcast metadata, advertiser and agency information, detailed schedule tables, and summary totals within a single invoice layout. While the layout appears regular, key fields are distributed across disparate regions of the page and expressed in heterogeneous formats. LMMs exhibit inconsistent extraction behaviors: some correctly recover the advertiser and flight dates but misestimate the gross amount, while others confuse semantically related entities. These errors suggest that robust KIE remains challenging due to subtle layout dependencies, long-range field–value associations, and the need to disambiguate closely related entities in complex business documents.

Figure 13 shows a medical services example from the constrained-category KIE track. In practical applications, such documents are frequently shared, stored, or processed with sensitive regions (*e.g.*, personal identifiers) manually redacted for privacy protection. In this example, the redaction removes critical textual evidence required by the input schema. This commonly occurring masking practice can trigger severe hallucination behaviors in LMMs. When the ground-truth value is partially or fully occluded, models may still produce confident yet unsupported predictions, effectively inventing missing content rather than abstaining. As reflected in the predictions, some LMMs hallucinate plausible personal names that are not present in the visible document, while others substitute unrelated numeric fragments as the receipt number, total amount, or billing time. This example highlights a practical and underexplored failure mode in real-world KIE deployments: even under a constrained schema, the widespread use of privacy-driven redaction can break visual grounding and lead to fluent but incorrect extractions.

A.10 Instructions for Annotators

We employ two distinct annotation workflows for the constrained-category and open-category KIE tracks. The annotation process was conducted by three collaborators of this work, who voluntarily participated as annotators. All annotators held master's degrees, had 1–3 years of experience in real-world enterprise KIE systems, and were proficient in both Chinese and English. All annotations are completed on the itag[9] platform.

The instruction for annotators in the constrained-category KIE track is shown in Figure 14. Annotators follow a predefined schema and extract the fields defined therein, grounding each field value in visible document text and marking fields as missing when the corresponding information is absent.

For the open-category KIE track, no predefined extraction schema is provided. As shown in Figure 15, annotators are required to review and correct the OCR output so that it exactly matches the visible document text, and then identify all key information units present in the document. Based on these units, annotators construct a document-specific hierarchical extraction schema that organizes fields and their semantic relations, such as grouping and containment. All annotations must remain strictly grounded in observable content, without inference, completion, or hallucination.

[9] https://www.aliyun.com/product/bigdata/learn/itag

Figure 8: Layer-wise Attention Transition with Explicit Field Cues.

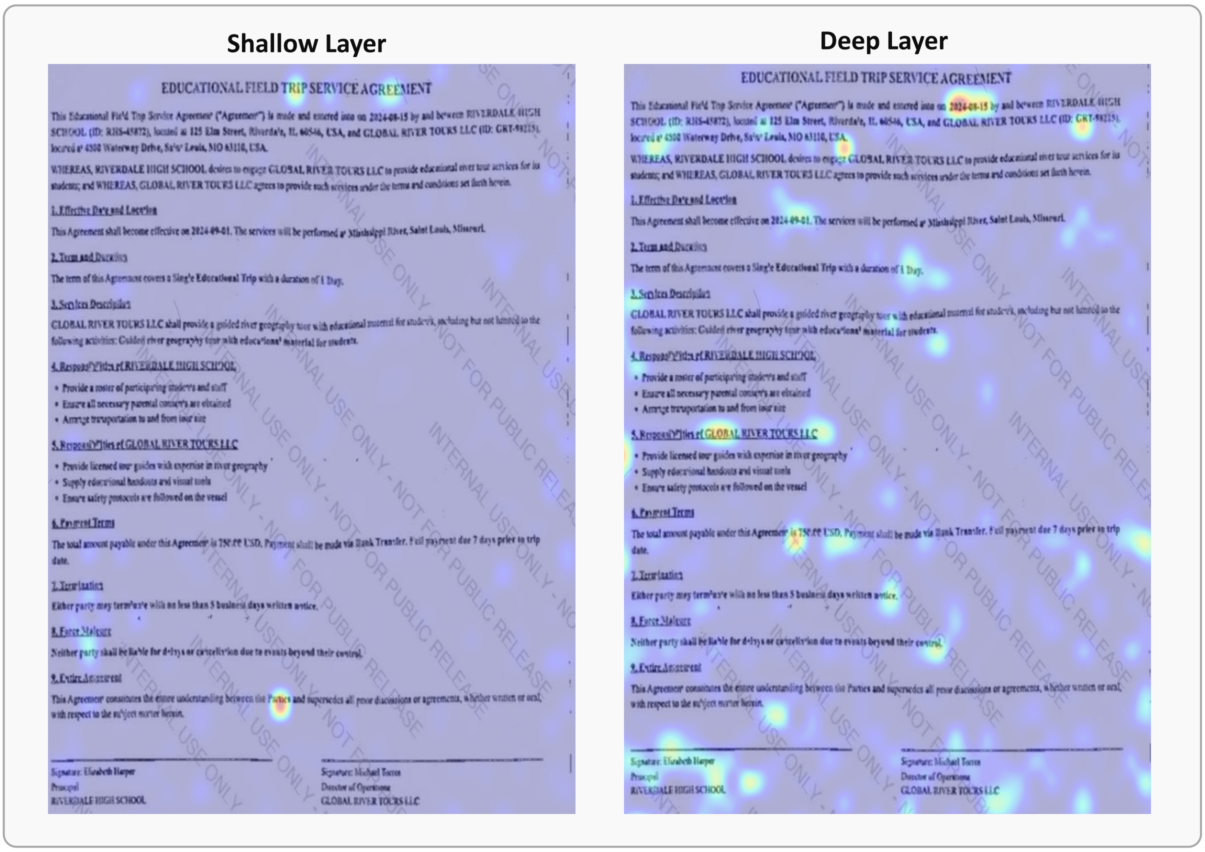
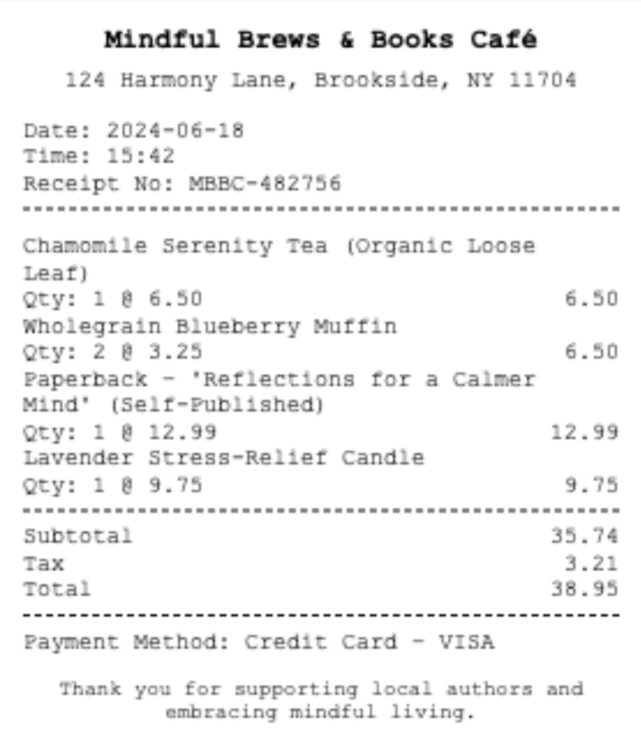


Figure 9: Layer-wise Attention Transition under Implicit Field Cues.



Mindful Brews & Books Café

124 Harmony Lane, Brookside, NY 11704

Date: 2024-06-18
Time: 15:42
Receipt No: MBBC-482756

Chamomile Serenity Tea (Organic Loose Leaf)	
Qty: 1 @ 6.50	6.50
Wholegrain Blueberry Muffin	
Qty: 2 @ 3.25	6.50
Paperback - 'Reflections for a Calmer Mind' (Self-Published)	
Qty: 1 @ 12.99	12.99
Lavender Stress-Relief Candle	
Qty: 1 @ 9.75	9.75
Subtotal	35.74
Tax	3.21
Total	38.95

Payment Method: Credit Card - VISA

Thank you for supporting local authors and embracing mindful living.

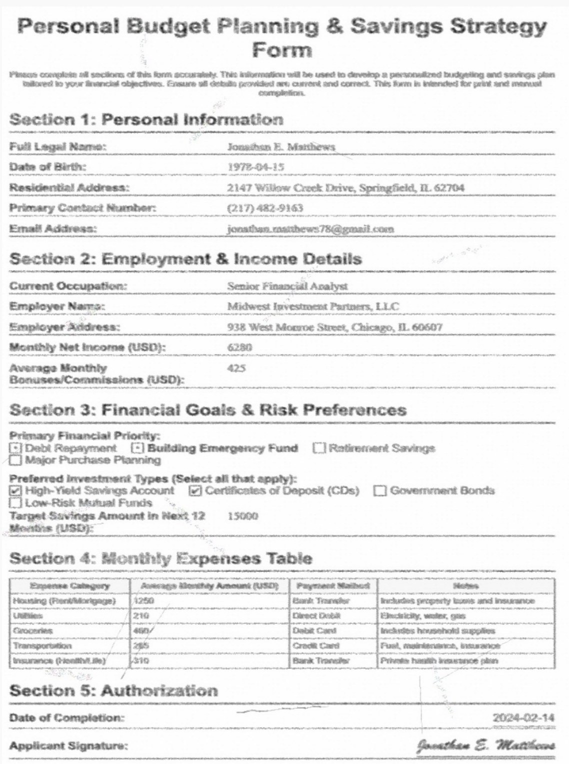
Input Schema:


```
{ "receipt_number": "",
    "items": [{
    "subtotal": "",
    "unit_price": "",
    "item_name": "",
    "quantity": ""
}...]
"total": "",
"transaction_time": "",
"store_address": "",
"store_name": ""
}
```


Qwen3-VL-Plus:


```
{ "receipt_number": "MBBC-482756",
   "items": [{
   "subtotal": "6.50",
   "unit_price": "6.50",
   "item_name": "Chamomile Serenity
Tea(Organic LooseLeaf)",
   "quantity": "1@6.50"
}...]
"total": "38.95",
"transaction_time": "15:42",
"store_address": "124 Harmony Lane,
Brookside, NY 11704",
"store_name": "Mindful Brews & Books Café"
}
```


Figure 10: Example from the Receipt Document Type in the Open-Category KIE Track.



Personal Budget Planning & Savings Strategy Form

Please complete all sections of this form accurately. This information will be used to develop a personalized budgeting and savings plan tailored to your financial objectives. Ensure all details provided are current and correct. This form is intended for print and manual completion.

Section 1: Personal Information

Full Legal Name:	Jonathan E. Matthews
Date of Birth:	1978-04-15
Residential Address:	2147 Willow Creek Drive, Springfield, IL 62704
Primary Contact Number:	(217) 482-9163
Email Address:	jonathan.matthews78@gmail.com

Section 2: Employment & Income Details

Current Occupation:	Senior Financial Analyst
Employer Name:	Midwest Investment Partners, LLC
Employer Address:	938 West Monroe Street, Chicago, IL 60607
Monthly Net Income (USD):	6280
Average Monthly Bonuses/Commissions (USD):	425

Section 3: Financial Goals & Risk Preferences

Primary Financial Priority:
☐ Debt Repayment ☐ **Building Emergency Fund** ☐ Retirement Savings
☐ Major Purchase Planning

Preferred Investment Types (Select all that apply):
☑ High-Yield Savings Account ☑ Certificates of Deposit (CDs) ☐ Government Bonds
☐ Low-Risk Mutual Funds

Target Savings Amount in Next 12 Months (USD): 15000

Section 4: Monthly Expenses Table

Expense Category	Average Monthly Amount (USD)	Payment Method	Notes
Housing (Rent/Mortgage)	1250	Bank Transfer	Includes property taxes and insurance
Utilities	210	Direct Debit	Electricity, water, gas
Groceries	460	Debit Card	Includes household supplies
Transportation	285	Credit Card	Fuel, maintenance, insurance
Insurance (Health/Life)	310	Bank Transfer	Private health insurance plan

Section 5: Authorization

Date of Completion:	2024-02-14
Applicant Signature:	Jonathan E. Matthews

Input Schema:


```
{ ...
  "Employment & Income Details": {
  "Monthly Net Income (USD)": "",
  "Employer Address": ""...},
  "Financial Goals & Risk Preferences": {
  "Primary Financial Priority": [],
  "Preferred Investment Types (Select all that
apply)": [],
}...}
```


Qwen-VL-Max:


```
{ ...
  "Employment & Income Details": {
  "Monthly Net Income (USD)": "6280",
  "Employer Address": "938 West Monroe Street,
Chicago, IL 60607"...},
  "Financial Goals & Risk Preferences": {
  "Primary Financial Priority": "Debt Repayment",
  "Preferred Investment Types (Select all that
apply)": [
     "High-Yield Savings Account",
     "Certificates of Deposit (CDs)",
     "Low-Risk Mutual Funds"]
}...}
```


Figure 11: Example from the Form Document Type in the Open-Category KIE Track.

REMIT TO Sinclair Broadcast c/o WXLV PO Box 206270 Dallas, TX 75320-6270	
Advertiser	**Friends Of Jim ONeil-R (134816)**
Agency	**PinPoint Media LLC (15192)**
Buyer	Danet,Segroves,
Salesperson	Millennium/DC, Washington DC (1108) ph: (202) 955-5342 , fx: (202) 955-5348x
Special Handling	None
Product	POLITICAL CANDIDATE (ns) (1186)
Brand	candidate (1011042)
Acct Types	**National/Political Candidate Agency BRD**
Est/External #	1277/ECR26793652
Demo	
Revision	
Comments	WXLV candidate

Invoice	7248774
Inv Date	2/23/2020
Terms	CIA
Contract	4228683
Bill Type	Standard
Period	1/27/2020 - 2/23/2020
CO-OP/Order Type	No/Normal
Package	
Gen. Date	2/25/2020 8:42:03AM

PinPoint Media LLC
4501 Ford Ave
Ste 1109
Alexandria, VA 22302
AgM

Greensboro (WXLV)

OFFICIAL BILLING INVOICE

Line	Type	Scheduled	Schedule Days to Run	Air Time	Length	Program	Copy/ISCI	Amount	Remarks
1.0	**Contract Line Remarks:**								
1.0	SPOT	20716-ABC-Good Moı	Day,Tu-1	02/04/20 7:25AM (Tu)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
2.0	SPOT	20716-ABC-Good Moı	Day,W-1	02/05/20 7:41AM (We)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
3.0	SPOT	20716-ABC-Good Moı	Day,Th-1	02/06/20 7:59AM (Th)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
4.0	SPOT	20716-ABC-Good Moı	Day,F-1	02/07/20 7:54AM (Fr)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
5.0	SPOT	20716-ABC-Good Moı	Day,M-1	02/10/20 8:13AM (Mo)	00:30	ABC-Good Morning America	DMSCJO0420H	$24.00	
6.0	SPOT	20716-ABC-Good Moı	Day,Tu-1,W-1,Th-1,F-1	02/04/20 8:59AM (Tu)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
6.0	SPOT	20716-ABC-Good Moı	Day,Tu-1,W-1,Th-1,F-1	02/05/20 8:55AM (We)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
6.0	SPOT	20716-ABC-Good Moı	Day,Tu-1,W-1,Th-1,F-1	02/06/20 8:38AM (Th)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
6.0	SPOT	20716-ABC-Good Moı	Day,Tu-1,W-1,Th-1,F-1	02/07/20 8:53AM (Fr)	00:30	ABC-Good Morning America	DMSCJO0220H	$24.00	
7.0	SPOT	20716-ABC-Good Moı	Day,M-1	02/10/20 8:59AM (Mo)	00:30	ABC-Good Morning America	DMSCJO0420H	$24.00	
8.0	News	20156-ABC-World Ne	Day,Tu-1,W-1,Th-1,F-1	02/04/20 6:29PM (Tu)	00:30	ABC-World News Monday-Friday	DMSCJO0220H	$192.00	
8.0	News	20156-ABC-World Ne	Day,Tu-1,W-1,Th-1,F-1	02/05/20 6:29PM (We)	00:30	ABC-World News Monday-Friday	DMSCJO0220H	$192.00	
8.0	News	20156-ABC-World Ne	Day,Tu-1,W-1,Th-1,F-1	02/07/20 6:29PM (Fr)	00:30	ABC-World News Monday-Friday	DMSCJO0220H	$192.00	
8.0	News			02/06/20			1 preempt @ $192.00	$0.00	Sold Out
9.0	News	20156-ABC-World Ne	Day,M-1	02/10/20 6:29PM (Mo)	00:30	ABC-World News Monday-Friday	DMSCJO0420H	$192.00	
10.0	SPOT	143476-ABC-Good M	Day,Su-1	02/09/20 8:53AM (Su)	00:30	ABC-Good Morning America Sunday	DMSCJO0420H	$16.00	
11.0	SPOT	By Orbit-[P]ABC-This	Day,Su-1	02/09/20 10:28AM (Su)	00:30	News-Full Measure with Sharyl Attkisson	DMSCJO0420H	$20.00	
12.0	News	636723-ABC-World N	Day,Sa-1	02/08/20 6:29PM (Sa)	00:30	ABC-World News Tonight (Saturday)	DMSCJO0420H	$180.00	
13.0	SPOT	20374-ABC-Good Moı	Day,Sa-3	02/08/20 8:41AM (Sa)	00:30	ABC-Good Morning American Saturday	DMSCJO0420H	$15.00	
13.0	SPOT	20374-ABC-Good Moı	Day,Sa-3	02/08/20 9:27AM (Sa)	00:30	ABC-Good Morning American Saturday	DMSCJO0420H	$15.00	
13.0	SPOT	20374-ABC-Good Moı	Day,Sa-3	02/08/20 9:53AM (Sa)	00:30	ABC-Good Morning American Saturday	DMSCJO0420H	$15.00	
14.0	SPOT	9:00:00AM-10:30:0(	Day,Su-1	02/09/20 9:40AM (Su)	00:30	ABC-This Week	DMSCJO0420H	$25.00	
15.0	SPOT	20180-ABC-Nightline	Day,M-1	02/10/20 12:39AM (Mo)	00:30	ABC-Nightline	DMSCJO0420H	$15.00	
16.0	SPOT	20716-ABC-Good Moı	Day,M-1	02/10/20 7:24AM (Mo)	00:30	ABC-Good Morning America	DMSCJO0420H	$45.00	
17.0	SPOT	20716-ABC-Good Moı	Day,M-1	02/10/20 8:41AM (Mo)	00:30	ABC-Good Morning America	DMSCJO0420H	$45.00	
18.0	SPOT	1198851-News-Full M	Day,Su-1	02/09/20 1:30AM (Su)	00:30	News-Full Measure With Sharyl Attkisson 2	DMSCJO0420H	$9.00	
19.0	SPOT	3220906-News-Ameri	Day,Sa-2	02/08/20 12:02AM (Sa)	00:30	News-America This Week	DMSCJO0420H	$10.00	
19.0	SPOT	3220906-News-Ameri	Day,Sa-2	02/08/20 12:18AM (Sa)	00:30	News-America This Week	DMSCJO0420H	$10.00	

Net 30-Payment is due within 30 days from invoice date.
Station owned or provided services by the Sinclair Broadcast Group.

Warranty – We warrant the above broadcasts were made according to the official station log. Please note that we do not notarize invoices since the purpose of placing a notary seal on a document is to guaranty the authenticity of an individual's signature and not to certify the accuracy of commercials aired.

Gross Total	**$1,428.00**
Commission	**($214.20)**
Net Total	**$1,213.80**

Total Spots 27

Officially printed on 02/27/20 03:18 PM

Page: 1

Input Schema:
{ "gross_amount": "",
"advertiser": "",
"flight_from": "",
"flight_to": ""
}

Gemini3-pro:
{ "gross_amount": "$1,428.00",
"advertiser": "Friends Of Jim ONeil-R (134816)",
"flight_from": "1/27/2020",
"flight_to": "2/23/2020"
}

MiniCPM-V-4.5-8B:
{ "gross_amount": "$1,213.80",
"advertiser": "PinPoint Media LLC(15192)",
"flight_from": "1/27/2020",
"flight_to": "2/23/2020"
}

Kimi-VL-A3B-Instruct:
{ "gross_amount": "$1,426.00",
"advertiser": "Friends Of Jim O‘ Neil-R(134816)",
"flight_from": "1/27/20",
"flight_to": "02/23/20"
}

Figure 12: Example from the Advertisement Scenario in the Constrained-Category KIE Track.

重庆市
门诊医药费专用收据
姓 [redacted]
日期 2018-07-018 No: 873710597

项 目	金额(元)
治疗费	8.00
诊察费	15.00
*卫材费	1.32
*检查费	66.00
*检验费	208.00
非甲乙药品	1552.50
乙类	343.50

账户自负 149.93 账户支付:5.00 统筹支付:10.00
现金支付:0.00
合计小写 2194.32
合计大写 贰仟壹佰玖拾肆元叁角贰分整
备注:
其他补助:0
收款单位(盖章): 收款人:
15:12:51
第一联 收据

Input Schema:
{
"姓名": "",
"收据号": "",
"合计金额": "",
"收费时间": ""
}

Qwen-VL-Max:
{
"姓名": "XXX",
"收据号": "873710597",
"合计金额": "2194.32",
"收费时间": "2018-07-XX"
}

InternVL3_5-8B-Instruct:
{
"姓名": "张某某",
"收据号": "873710597",
"合计金额": "2194.32",
"收费时间": "2018-07-18 15:12:51"
}

SmolVLM2-2_2B-Instruct:
{
"姓名": "黄晓明",
"收据号": "20180701",
"合计金额": "149.93",
"收费时间": "16:12:51"
}

Figure 13: Example from the Medical Services Scenario in the Constrained-Category KIE Track.

Objective
Given a document image and a predefined, scenario-specific extraction schema, annotate the corresponding key information fields accurately.

Annotation Inputs
- A document image
- The associated application scenario
- A predefined schema with field names

Annotation Guidelines
- Annotators are required to locate and extract field values according to the provided schema.
- Each field has a clearly defined semantic scope, and the task focuses on identifying whether the corresponding information is present and extracting its exact textual content.

Constraints
- Annotate only the fields explicitly defined in the schema; no additional fields may be introduced.
- Field values must be strictly grounded in explicitly visible text in the document image.
- If a field is not present, it should be marked as *missing / not present*.
- Semantically related fields must be distinguished strictly according to their schema definitions.
- No inference or completion of missing information is allowed.

Figure 14: Instructions for Annotators in the Constrained-Category KIE Track.

Objective
Given a document image without a predefined extraction schema, identify all semantically important information and organize it into a document-specific hierarchical extraction schema.

Annotation Inputs
- A document image
- Automatically generated OCR results

Annotation Guidelines
- Annotators are required to first review and correct the OCR output so that it exactly matches the explicitly visible text in the document image, including fixing character-level errors, missing text, and incorrect line segmentation or ordering.
- Based on the corrected OCR text, annotators should identify semantically important information units, such as amounts, dates, entity names, identifiers, clauses, or record entries.
- Annotators should then construct a document-specific extraction schema and organize the identified information accordingly.
- The schema may vary across documents and should explicitly model semantic relations among fields, such as grouping (e.g., multiple fields forming a record) and containment (e.g., clauses belonging to a section).
- Field names should be semantically clear, interpretable, and consistent within each document.
- Field values should correspond to the minimal self-contained semantic text spans present in the document.

Constraints
- All annotations must be strictly grounded in explicitly visible document content.
- Annotators must not infer, complete, or hallucinate missing or implicit information.
- OCR correction is limited to correcting recognition errors; no completion of occluded or non-existent content is allowed.
- Decorative or layout-only text that does not carry semantic meaning may be excluded.

Figure 15: Instructions for Annotators in the Open-Category KIE Track.