International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-026-00572-y

ORIGINAL PAPER

Beyond Accuracy: Understanding Model Confidence in Key Information Extraction with Conformal Prediction

Alexander Rombach[1,2] · **Nijat Mehdiyev**[1,2]

Received: 28 May 2025 / Revised: 22 December 2025 / Accepted: 19 February 2026

Abstract

Key Information Extraction (KIE) systems based on Deep Learning achieve strong token-level performance but offer no formal guarantees on prediction reliability, limiting their adoption in business-critical document workflows. In this work, we introduce a post hoc Uncertainty Quantification framework for KIE using Split Conformal Prediction (CP). After fine-tuning multimodal transformer models on a challenging receipt dataset, we reserve a held-out calibration set to derive nonconformity scores and construct entity-level prediction sets that satisfy a user-specified error rate. On unseen receipts, CP achieves tight marginal coverage (98.3% for $\alpha = 0.02$), with 70% of predictions being high-confidence singletons. A detailed analysis shows that highly structured fields such as dates and prices yield small, singleton sets with near–perfect reliability, whereas rare or semantically ambiguous fields such as tips or generic keywords produce larger sets and lower coverage. By exposing positional biases and common label confusions that standard F1-scores and document-accuracy metrics overlook, CP reveals critical risk areas for downstream automation. Finally, we demonstrate how calibrated prediction-set sizes can drive risk-aware workflows by automatically processing high-confidence extractions and flagging uncertain cases for human review, thereby enhancing the efficiency, trustworthiness and operational feasibility of real-world document-processing systems.

Keywords Key Information Extraction · Conformal Prediction · Uncertainty Quantification · Explainable Artificial Intelligence · Deep Learning

1 Introduction

Key Information Extraction (KIE) is central to business document processing, where large volumes and substantial manual effort are involved in daily operations [12]. Recent advances in Deep Learning (DL) have significantly improved KIE performance across benchmarks [18]. However, these models cannot guarantee full accuracy in real-world settings, and common evaluation procedures often overestimate performance, providing limited insight into practical reliability [14]. This gap is critical, as extraction errors can have severe downstream consequences [10]. For example, in unattended invoice processing, incorrect extraction of financial values may directly result in costly overpayments.

To mitigate such risks, human validation remains necessary in automated KIE workflows [7]. Yet, DL-based KIE models are typically black-box systems, offering little transparency into why predictions succeed or fail. Consequently, practitioners struggle to determine which outputs can be trusted. Although DL models are deployed in other high-stakes domains such as medicine, reliability there is achieved through extensive supervision, conservative decision rules and continuous monitoring rather than reliance on raw model outputs alone. In high-throughput document processing, exhaustive manual review is infeasible; instead, validation must be selectively integrated through scalable, risk-aware workflows.

Without reliable uncertainty estimates, organizations face an undesirable trade-off: either manually inspect all predictions or rely on uncalibrated model confidences that lack statistical guarantees. In practice, many systems adopt static confidence thresholds or heuristic rules (e.g., flagging predictions below a fixed confidence level). While such approaches

✉ Alexander Rombach
alexander_michael.rombach@uni-saarland.de

Nijat Mehdiyev
nijat.mehdiyev@dfki.de

[1] Saarland University, Campus D3 2, Saarbrücken, Germany

[2] German Research Center for Artificial Intelligence, Campus D3 2, Saarbrücken, Germany

Published online: 10 March 2026

may improve average performance, they provide no formal error guarantees, cannot distinguish inherently ambiguous cases from systematic model failures, and may allow high-risk errors to propagate downstream.

One promising solution is Conformal Prediction (CP), a distribution-free, post hoc framework that augments any black-box model with prediction sets offering rigorous coverage guarantees under mild exchangeability assumptions. Instead of a single point prediction, CP outputs a set of plausible labels with a guaranteed probability of containing the true value (e.g., $\geq$ 98%), while set size naturally reflects uncertainty. In this work, we apply Split CP to state-of-the-art multimodal transformer models for KIE, positioning CP not merely as a confidence estimation layer but as a unified framework for uncertainty-aware document analysis. We show that calibrated prediction sets reveal systematic, layout-dependent failure modes that conventional metrics fail to capture and directly support risk-aware automation by separating confident predictions from cases requiring human validation. Beyond post hoc verification, our approach leverages CP as both a diagnostic tool for understanding KIE model behavior and a practical mechanism for reliable large-scale deployment.

We address the following research questions: *How can uncertainty in KIE approaches be effectively quantified?* (RQ1) *How can this uncertainty be analyzed to generate meaningful insights?* (RQ2) *In which aspects do state-of-the-art KIE models exhibit uncertainty?* (RQ3) *What are the implications of these insights for real-world document processing applications?* (RQ4)

The remainder of this paper is organized as follows. Section 2 introduces the multimodal transformer–based KIE architecture and the split CP framework. Section 3 describes the use case, dataset and evaluation setup. Section 4 presents point-prediction and CP-based results with detailed uncertainty analysis. Section 5 discusses practical implications, limitations and future work. We review related literature in Sect. 6 and conclude in Sect. 7.

2 Methodology

Our methodological framework comprises two primary components: Named Entity Recognition (NER), which is responsible for KIE from business documents, and UQ, which is aimed at providing reliable statistical guarantees for these predictions. The process begins directly after the Optical Character Recognition (OCR) pre-processing step, which provides textual elements of the documents and their spatial positions. A high-level overview of our methodology and its key steps is provided in Fig. 1.

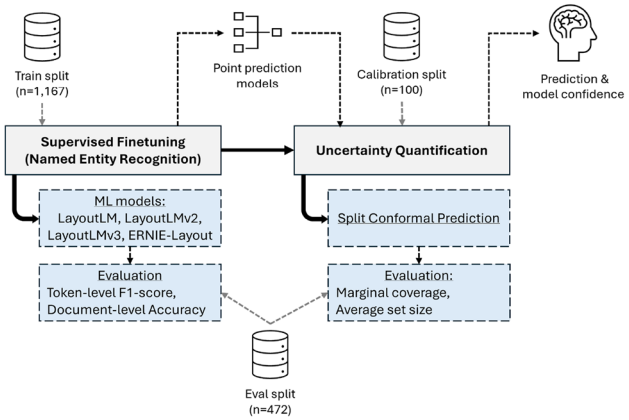

Fig. 1 Conceptual overview of the methodology

2.1 Named entity recognition

Automated KIE requires a model that can recognize every semantically relevant entity on the document page while remaining sensitive to its layout and visual appearance. OCR transcribes the document image into an ordered list of textual elements $(w_1, \ldots, w_T)$ together with bounding boxes $b_t = (x_0, y_0, x_1, y_1)$[1]. Each word w_t is further decomposed into subtokens $(x_{t,1}, \ldots, x_{t,m_t})$, and re-indexing produces the flattened sequence $(x_1, \ldots, x_{T'})$ with matching boxes $(b_1, \ldots, b_{T'})$, where $T' = \sum_{t=1}^{T} m_t$. For multimodal KIE models, an RGB crop $i_t \in \mathbb{R}^{c\times h\times h}$ centered on b_t is extracted once per word and copied to all its subtokens (see Fig. 2).

The NER problem is generally framed as a sequence-labeling task. Given the semantic fields to extract $\mathcal{C}$, the tag alphabet $\mathcal{Y}$ is defined as:

$$\mathcal{Y} = \{O\} \cup \{I-c \mid c \in \mathcal{C}\}, \text{ where } |\mathcal{Y}| = |\mathcal{C}| + 1. \tag{1}$$

A gold span of class c covering tokens $[s, e]$ is mapped to the tag sequence $I-c, \ldots, I-c$ and may include non-contiguous tokens belonging to the same semantic entity. For example, the monetary value "10,20 e" spans multiple tokens but is annotated as a single entity such as Total_v. All subtokens of a word inherit the $I-c$ tag if the parent word is entity c.

Each token x_t is embedded as:

$$h_t = E_{\text{text}}(x_t) \oplus E_{\text{pos}}(b_t) \oplus E_{\text{vis}}(i_t), \tag{2}$$

where E_{text} is a token embedding, E_{pos} a learnable two-dimensional positional map of the bounding box, and E_{vis} the projection of the image crop through a vision backbone. The embeddings are processed as parallel streams inside a

[1] The coordinates are normalized to the range [0, 1000] based on the page width and height.

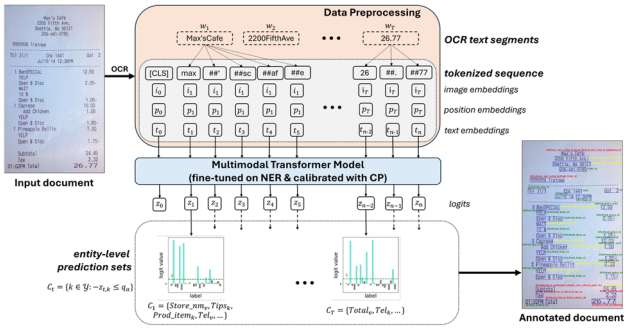

Fig. 2 Technical workflow

multimodal transformer which fuses semantic, positional and visual modalities.

For each token, the transformer produces the contextual vector h_t. A linear projection yields logits:

$$z_t = Wh_t + c \in \mathbb{R}^{|\mathcal{Y}|}, \tag{3}$$

(see Fig. 2) which are normalized via softmax to produce token-level probabilities that can be used for KIE:

$$P(y_t = k \mid x_t) = \frac{\exp(z_{t,k})}{\sum_{j \in \mathcal{Y}} \exp(z_{t,j})}. \tag{4}$$

2.2 Uncertainty quantification

Reliable KIE alone is insufficient for business workflows, particularly in financial scenarios such as invoice processing, where a single mislabeled token can have significant implications. Thus, every prediction must include a calibrated measure of uncertainty. This is achieved *post hoc* by augmenting the trained transformer model with a Split CP layer that converts a point prediction into a *set* of plausible labels, providing rigorous statistical coverage guarantees on unseen data assuming data exchangeability.

For CP, the data is partitioned into three mutually exclusive subsets. The *training* subset is used to fit the NER model parameters and is discarded afterwards. The *calibration* subset is reserved for calibrating the conformal predictor. The *test* subset is held out solely for evaluation purposes.

During calibration, for each semantic entity e, the model provides logits $\mathbf{z}_e$ corresponding to the predicted labels for each subtoken. Considering the pivotal role of the first token in determining semantic correctness, we define the non-conformity score based exclusively on the logits of the first token within each semantic entity. This choice reflects that misclassification of the entity's initial token substantially impacts downstream correctness. We follow [6] and define the non-conformity score for every entity e as the negative logit assigned to the ground-truth label y_e:

$$s_e = -z_{e,y_e} \tag{5}$$

where higher values indicate lower confidence. The multiset $\mathcal{S} = \{s_e \mid e \in \mathcal{C}\}$ constitutes an empirical sample of model uncertainty. Given a specified entity-level error budget $\alpha \in (0, 1)$, we select the $(1-\alpha)$-quantile of this empirical distribution as:

$$q_\alpha = \inf\{q \in \mathbb{R} : |\{s \in \mathcal{S} : s \le q\}| \ge (1-\alpha)|\mathcal{S}|\}, \tag{6}$$

ensuring that at most an α proportion of calibration scores exceed q_α.

During inference, the same threshold q_α applies to each new entity's first token logits z_t, forming the entity-level prediction set (see Fig. 2):

$$C_t = \{k \in \mathcal{Y} : -z_{t,k} \le q_\alpha\}. \tag{7}$$

Standard conformal theory guarantees the finite-sample coverage property at the entity level:

$$\Pr\{y_t \in C_t\} \ge 1 - \alpha, \tag{8}$$

where the probability is marginal over entities drawn exchangeably from the production stream. Importantly, this guarantee applies to the overall entity population without necessarily ensuring coverage for individual subsets or entity types.

The size of the prediction set acts as a calibrated uncertainty measure. Singleton sets reflect sufficiently high confidence, enabling automatic processing without further review. Entities assigned multiple-label sets clearly indicate ambiguity, prompting selective human intervention. Thus, the methodology efficiently targets manual verification efforts toward the most sensitive areas.

3 Experiment

3.1 Use case description

The aim of automated KIE systems is to streamline document processing workflows by identifying, extracting and structuring relevant data fields - in this case, those from receipts. The process begins with incoming documents, which may arrive via various channels including email attachments, document management systems, or direct uploads to a processing platform. These documents are first preprocessed, which typically involves OCR to make the content machine-readable. Following pre-processing, the KIE system identifies and extracts predefined key fields. Confidence scores are also

assigned to each extracted field to reflect the system's (un-)certainty. After the automated extraction phase, a human validation step is incorporated to ensure accuracy and reliability. The extracted data is passed to a validation interface, where a human reviewer is presented with both the original document and the extracted fields. During this phase, the operator reviews and verifies the extracted information, verifies its correctness and corrects any misidentified or incomplete fields. Once validation is complete, the final extracted information is transferred into structured formats (e.g., in ERP system) for downstream tasks such as automated accounting, compliance audits, customer relationship management or business analytics. In our scenario, the receipts contain an average of 40 words. This would therefore require a considerable effort to manually process every document. Using a KIE solution, the operator's workload is reduced to the validation of the provided extracted values [19].

3.2 Dataset

For our experiments, we use WildReceipt proposed by [21], which consists of 1,739 receipt images from restaurants or shops, covers 25 fields to extract and is often used in KIE related research. The labels are divided into two groups, namely keywords and their corresponding values (e.g., *Prod_price_value* and *Prod_price_key*). We removed the labels *Store_nm_key* and *Store_addr_key* because they only appear five times across all documents and never in the test set. Affected segments were relabeled with the *Others*[2] class accordingly. The dataset is annotated on segment-level meaning that e.g., the store name is considered and labeled as one element even if it covers multiple words (see Fig. 2). We chose this dataset because of its challenging nature with high layout diversity (1,192 different templates) and because the images are photographs taken on mobile devices which are therefore no high-quality scans. Another reason is the large amount of fields to extract, which allows for a deeper analysis compared to related datasets such as SROIE [8], which only covers four high-level fields [18].

We use the provided train/test split and additionally randomly selected 100 receipts from the training set for CP calibration and model validation. We end up with 1,167 images for training, 100 for validation/calibration and 472 for testing. Note that although we use 100 receipts for calibration, the actual number of calibration points is the number of tokens these 100 receipts contain - as this is the prediction task performed by the NER model. More specifically, our calibration set includes 4,074 tokens/calibration points in total, which ensures proper coverage [1]. From a dataset size standpoint, WildReceipt is sufficient for our purposes because the approach builds on large-scale pre-trained document backbones, significantly reducing data requirements for the adaption to custom data (transfer learning).

[2] *Others* represent words of a document that do not belong to any of the fields of interest

3.3 Evaluation metrics

The performance of the proposed pipeline is assessed from two complementary viewpoints: (i) the accuracy of predicted label for each sub-token (*point metrics*) and (ii) the reliability of the calibrated prediction sets generated by Split CP (*set metrics*).

Point Prediction Measures. For every field $c \in \mathcal{C}$ we accumulate token-level counts of True Positives, False Positives and False Negatives, denoted TP_c, FP_c and FN_c, respectively. The class-specific F1-score is defined by

$$F_{1,c} = \frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{9}$$

The macro average weights each label equally,

$$F_1^{\text{macro}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F_{1,c}. \tag{10}$$

To quantify end-to-end correctness on entire pages we also report the Document Accuracy Rate (DAR):

$$\mathrm{DAR} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{d \in \mathcal{D}_{\text{test}}} \mathbf{1}\big[\mathrm{FP}_d = 0 \ \wedge \ \mathrm{FN}_d = 0\big], \tag{11}$$

where FP_d and FN_d denote the numbers of mis-tagged tokens on document d and $\mathbf{1}[\cdot]$ is the indicator function.

Set Metrics. Let $C_t \subseteq \mathcal{Y}$ be the prediction set attached to token t by Split CP with risk level α, and let y_t be its groundtruth tag. Following the formulation of [1], two statistics jointly describe calibration and efficiency. Equation 12 estimates the marginal coverage achieved on the evaluation corpus, whereas Equation 13 measures the average set size and therefore the sharpness of the uncertainty estimate.

$$\widehat{\mathrm{Cov}} = \frac{1}{T'} \sum_{t=1}^{T'} \mathbf{1}\big[y_t \in C_t\big], \tag{12}$$

$$\overline{|C|} = \frac{1}{T'} \sum_{t=1}^{T'} |C_t|. \tag{13}$$

3.4 Implementation details

We fine-tuned several pre-trained transformer-based models on the NER task, specifically LayoutLM [22], LayoutLMv2 [23], LayoutLMv3 [23] and ERNIE-Layout [15], using the

HuggingFace Transformers library as our implementation framework. For each model, we selected the *large* model variants whenever available, as these typically offer improved performance through increased model size. All models have 24 transformer layers, a hidden size of 1024 and 16 attention heads; mostly the treatment of visual input and the pre-training objectives differ. The models were trained for 75 epochs with a batch size of 8. Hyperparameters such as learning rate (5e-5) and weight decay (0) were kept to default values provided by the HuggingFace Trainer class[3]. We also use the default optimizer AdamW with a linear learning rate scheduler and 0 warmup steps.

During training, the goal is to minimize the cross-entropy loss between the token-level class probabilities and the groundtruth labels, encouraging the model to leverage contextual and visual cues for predictions. Model training is guided by F_1^{macro}, with validation steps performed after each epoch. Upon completion of training, the model checkpoint corresponding to the best validation performance was selected for subsequent evaluation. Consequently, we did not employ early stopping.

Following model training, we performed the calibration step for CP according to the methodology explained in Sect. 2.2. We set $\alpha = 0.02$ to obtain very confident guarantees. This means that during inference, we can guarantee with 98% certainty that the correct label is part of the prediction set. Subsequently, the quantile threshold q_α was obtained based on the calculated nonconformity scores.

Finally, predictions were generated on the test set. For each first subtoken of the entities, a prediction set was constructed according to the threshold. This ensures that each word in the original document is associated with a distinct prediction set.

4 Results

4.1 Point prediction evaluation

The NER evaluation results on the test set are presented in Table 1. As can be seen, all models achieve relatively similar results. Nevertheless, LayoutLMv2 achieves the best overall performance with an F_1^{macro} of 0.9215. The improvements of LayoutLMv3 compared to the previous iterations therefore did not lead to improved results on this particular dataset. It must be said, however, that this is only a snapshot of a single training run. The original LayoutLM is the worst performing model, indicating that the enhancements of LayoutLMv2 can indeed lead to improved extraction results. ERNIE-Layout is not able to achieve the best results in any of the fields, but shows promising generalization performance across all fields as it performs better on average than the original LayoutLM with F_1^{macro} of 0.9019.

Table 1 Point prediction results on test set

Label[1]	LM	LMv2	LMv3	ERNIE
Date_k	0.9375	**0.9529**	0.9468	0.9278
Date_v	0.9829	**0.9852**	**0.9852**	0.9800
Others	0.9140	0.9314	**0.9380**	0.9184
Prod_item_k	0.8857	0.9151	**0.9154**	0.8909
Prod_item_v	0.9470	0.9637	**0.9719**	0.9540
Prod_price_k	0.8558	0.8785	**0.9126**	0.8621
Prod_price_v	0.9598	0.9718	**0.9804**	0.9716
Prod_qty_k	**0.9486**	0.9371	0.9474	0.9474
Prod_qty_v	0.9723	0.9685	**0.9785**	0.9658
Store_adr_v	0.9102	**0.9305**	0.9267	0.9202
Store_nm_v	0.8668	0.8828	**0.8925**	0.8734
Subtotal_k	0.8768	**0.9102**	0.8784	0.8745
Subtotal_v	0.7956	**0.9058**	0.8820	0.8396
Tax_k	0.9012	0.9261	**0.9459**	0.9065
Tax_v	0.8737	0.9058	**0.9262**	0.8848
Tel_k	0.9577	0.9691	**0.9792**	0.9744
Tel_v	0.9260	0.9494	**0.9554**	0.9541
Time_k	0.8667	**0.9091**	0.8908	0.8850
Time_v	0.9638	**0.9705**	0.9597	0.9501
Tips_k	0.8182	**0.8750**	0.8125	0.8571
Tips_v	0.5116	**0.7778**	0.7500	0.6923
Total_k	0.8715	**0.9064**	0.8851	0.8742
Total_v	0.8333	**0.8710**	0.8682	0.8393
F_1^{macro}	0.8859	**0.9215**	0.9186	0.9019
DAR	0.1543	0.2368	**0.2516**	0.1670

"*_k" denote keyword labels, "*_v" denote value labels

In general, the presented F1-scores would indicate a relatively high confidence of the respective models. Without further consideration, it could be assumed that, given a prediction is correct in around 90 percent of the time, the models are well suited for being integrated into automated real-world workflows. However, considering the DAR reveals that there are many incorrect extractions. In the best case, only around 25 percent of the test documents were classified without any errors. This would mean a large manual overhead for validating extraction results. This discrepancy shows firstly that token-level metrics are not a good indicator of real-world performance, and secondly that the KIE approaches inherently suffer from uncertainty. In the following, we further quantify and analyze these observations.

4.2 Conformal prediction evaluation

For the sake of brevity, we take a closer look at the LayoutLMv2 model, as it produced the best point prediction

[3] https://huggingface.co/docs/transformers/main_classes/trainer.

results as discussed previously. Since all approaches achieved relatively similar results, it can be assumed that the findings are also reflected in the other models. All experiments in the following are based on $\alpha = 0.02$.

4.2.1 Prediction Set Analysis

Figure 3 visualizes the distribution of prediction set sizes on the test set, while Table 2 reports the corresponding relative frequencies and marginal coverage. The histogram is strongly skewed toward small prediction sets, with a clearly dominant first bar at set size 1, followed by a rapidly decaying long tail. This indicates that the vast majority of fields can be processed automatically with a single predicted label, while larger sets occur only infrequently. Concretely, 70% of all instances have a prediction set of size 1, meaning the model makes an unambiguous prediction. An additional 24% of instances have a set size of 2 which means that in total, in 94% of the cases at most two candidates need to be considered. In contrast, prediction sets of size five or six are extremely rare and correspond to cases where the model is highly uncertain. The average prediction set width is 1.38, which is low and reflects the overall confidence of the KIE model. Importantly, the marginal coverage remains high for the most frequent set sizes, reaching 98.3% for set sizes 1 and 2. While intermediate set sizes (three and four) exhibit slightly lower marginal coverage, the largest sets achieve perfect coverage (100%), indicating that the conformal predictor behaves conservatively by expanding the prediction set when uncertainty is high. From an operational perspective, this distribution highlights that most instances can be handled automatically with minimal ambiguity, while the small fraction of uncertain cases is explicitly flagged through larger prediction sets. The rapid decay visible in the histogram demonstrates that the "cost" of uncertainty is limited and manageable, providing a practical mechanism for selective review without sacrificing coverage guarantees.

We further analyzed the prediction sets containing multiple elements to identify which labels frequently co-occur and whether recognizable patterns emerge. One observation is that the model confuses different monetary values. For example, in the prediction sets of words belonging to *Prod_price_v*, the labels *Subtotal_v* and *Total_v* also frequently appear. This suggests that the model is not always confident in differentiating between the individual monetary values. Interestingly, the labels for the keywords and the actual values of a corresponding field also often appear in the same prediction set. This is more prevalent for some fields than for others. For example, the two labels *Time_v* and *Time_k* often appear in the same prediction set. Since keywords and actual values usually represent very different words in terms of their semantics, one explanation for these confusions is their close proximity on the receipts.

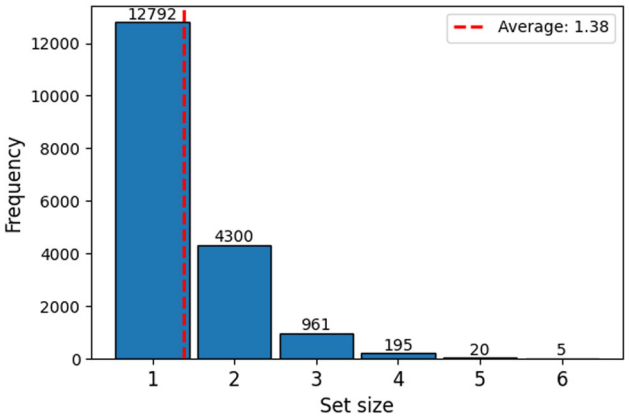

Fig. 3 Histogram of set sizes

Table 2 Statistics of set sizes

Set size	Relative frequency	Marginal coverage
1	70.00%	98.30%
2	23.53%	98.30%
3	5.26%	97.29%
4	1.07%	97.44%
5	0.11%	100%
6	0.03%	100%

Therefore, the model focuses more on the position of the words than on their corresponding text. At the same time, this further explains the confusion regarding different monetary values, as they often appear close to each other and are therefore often falsely included in the prediction set of another monetary value. Prediction sets of words belonging to the *Others* class show the highest heterogeneity, i.e., contain the most amount of different co-occurring labels. Corresponding prediction sets frequently contain *Prod_item_v, Total_v* and *Prod_price_v*.

4.2.2 Label-based Analysis

Table 3 shows findings regarding the labels in terms of average set size and marginal coverage. In general, the average set sizes range from 1.13 to 2.24. The model is most confident for the field *Prod_price_v* and least confident for *Tips_v*. To assess whether prediction sets reliably indicate high confidence, we analyze their marginal coverage. Among all tokens with singleton prediction sets, the true label is present in 98.30% of cases. This confirms that when the model is highly confident - as expressed by assigning only a single label - it is also typically correct. This strong alignment between set size and correctness suggests that the prediction set size is a meaningful proxy for uncertainty and could be used as a confidence signal in downstream decision-making systems. Besides, in most of the cases, the marginal coverage is at

Table 3 Statistics of labels

Label[1]	Average set size	Marginal coverage
Prod_price_v	1.13	99.51%
Date_v	1.16	98.41%
Prod_item_v	1.17	99.28%
Tel_v	1.19	99.33%
Date_k	1.21	98.96%
Tel_k	1.24	99.31%
Prod_qty_k	1.25	97.73%
Prod_item_k	1.28	96.19%
Store_adr_v	1.29	98.57%
Prod_price_k	1.31	92.59%
O	1.38	98.09%
Subtotal_k	1.44	95.09%
Store_nm_v	1.45	96.28%
Tax_k	1.47	96.94%
Time_v	1.49	97.83%
Total_v	1.51	98.43%
Subtotal_v	1.54	94.01%
Tips_k	1.59	90.62%
Tax_v	1.68	97.8%
Time_k	1.69	96.72%
Prod_qty_v	1.72	99.47%
Total_k	2.07	95.67%
Tips_v	2.24	93.1%

"*_k" denote keyword labels, "*_v" denote value labels

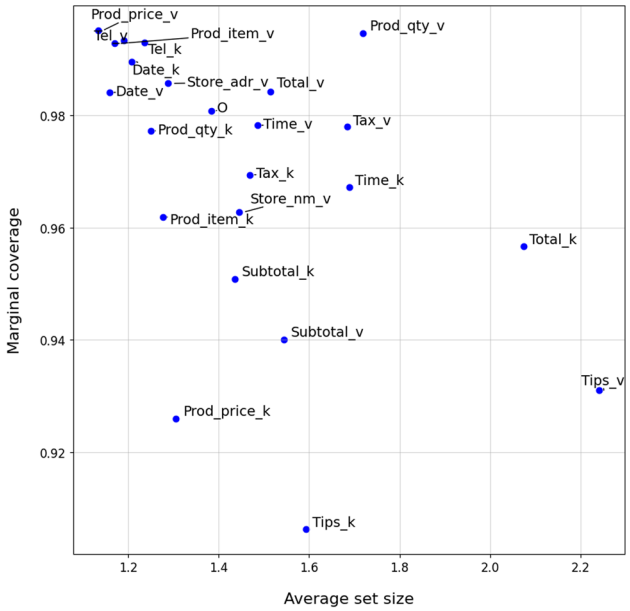

Fig. 4 Relationship between average set size and marginal coverage

least 98%, which indicates high model confidence and accuracy across the majority of labels. Note that the choice of $\alpha = 0.02$ only guarantees an overall marginal coverage of at least 98% and not for each individual label. Therefore, some labels like *Tips_v* show lower individual marginal coverages. Labels such as *Prod_price_v, Prod_item_v* and *Tel_v* perform particularly well, with very small set sizes (around 1.1–1.2) and coverage >99%. Conversely, a few labels, especially *Tips_v, Tips_k, Prod_price_k,* and *Subtotal_v*, show larger average set sizes and at the same time relatively low marginal coverage (around 90–94%), suggesting these are more challenging for the KIE model. Labels with a structured format like prices, dates, phone numbers and quantities tend to perform best. They have clear patterns, low ambiguity and strong contextual cues. Keywords, on the other hand, tend to be more challenging due to their generic wording, shorter length and semantic overlap (e.g., "Total", "Amount"), which can increase model uncertainty.

Figure 4 illustrates the relationship between the average prediction set size and marginal coverage for each label. The figure exhibits a clear clustering behavior: a dense group of labels appears in the top-left region, characterized by small prediction sets and high marginal coverage. This region corresponds to "safe" fields for which the model is both confident and statistically reliable. These safe fields are predominantly well-structured entities, such as *Prod_price_v*, which achieves the most favorable trade-off between the two metrics. Its position confirms that structured monetary fields can be extracted with high confidence and minimal ambiguity, making them well-suited for fully automated processing. In contrast, a small number of labels appear as outliers in the bottom-right region of the plot, where prediction sets are larger and marginal coverage is lower. Notably, the unstructured fields *Tips_v* and *Tips_k* fall into this category. Their position indicates that the model is frequently uncertain while simultaneously failing to provide strong coverage guarantees, marking these fields as "risky" from an extraction perspective. One plausible explanation for this behavior is the relatively rare and inconsistent occurrence of tips on receipts, which limits the amount of representative training data. As a result, the model struggles to distinguish tip-related monetary amounts from other prices and/or often assigns them to the *Others* class. Overall, the clustering visually confirms that model reliability is field-dependent. Importantly, this visualization provides a practical guideline for deployment: labels in the top-left cluster can be processed automatically with high confidence, whereas outliers in the bottom-right region are potential candidates for targeted human review.

4.2.3 Document-level Analysis

Figure 5 shows a boxplot of the average prediction set size per document. The distribution is tightly centered around 1.38 with a narrow interquartile range, confirming consis-

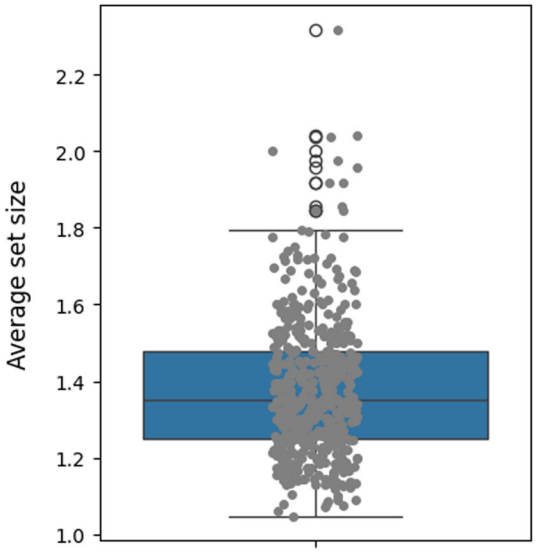

Fig. 5 Average set size per document

tent model behavior across the dataset. Most documents are processed with uniformly small prediction sets. A small number of outliers (average set size $>$ 2.2) reveal documents for which the model is globally more uncertain. A further inspection shows that these correspond to documents with low-quality scans, rare layouts and/or receipts containing handwritten notes overlapping printed text. In such cases, OCR errors and ambiguous transcriptions lead to larger prediction sets across the document. These outliers reflect document-level issues rather than isolated word-level errors and therefore indicate cases requiring human review. Despite this, the average coverage per document remains high at 98.14% (standard deviation 0.03). While 41.44% of documents contain at least one word-level error, the system maintains strong statistical guarantees overall while enabling reliable detection of atypical documents.

Figure 6 illustrates how CP reveals limitations of DL–based KIE models in realistic document layouts. The figure shows an exemplary receipt with bounding boxes annotated by CP sets. Well-structured fields such as the telephone number and date-time entries yield small and/or singleton prediction sets, indicating high confidence. In contrast, layout-ambiguous regions such as the address block produce larger sets containing multiple plausible labels. The behavior differs across document regions. In the itemized section, product descriptions are assigned mostly small prediction sets, indicating high confidence. Similarly, product prices consistently include *Prod_price_v* within prediction sets of size at most two, reflecting reliable extraction. This confidence is likely due to the highly regular receipt layout, where prices typically appear right-aligned and vertically stacked, providing strong spatial cues. In contrast, the summary region at the bottom of the receipt exhibits substantially

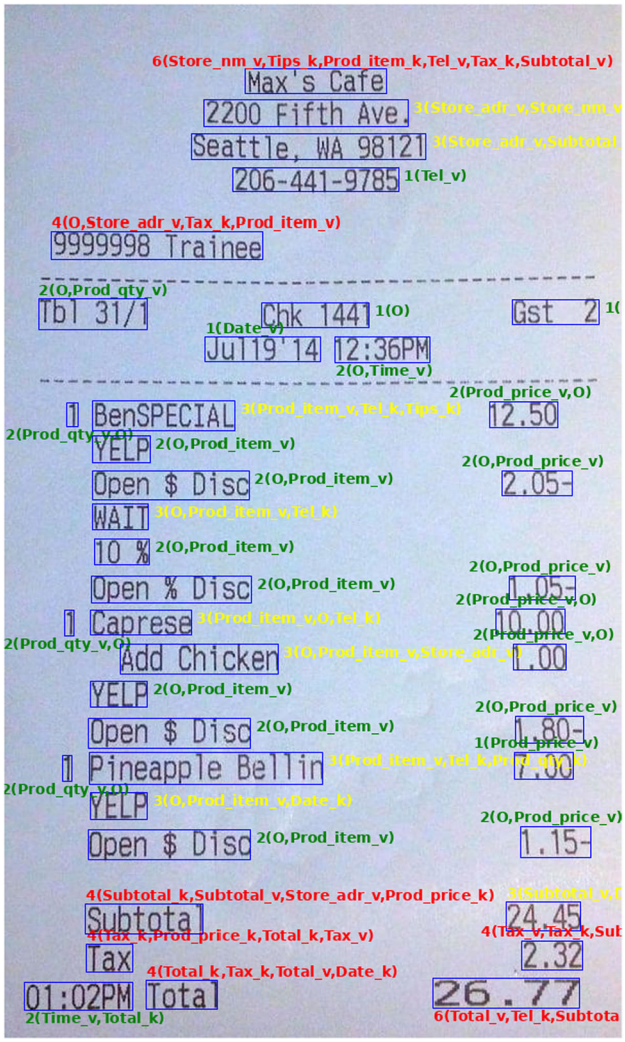

Fig. 6 Receipt with annotated prediction sets

larger and more ambiguous prediction sets, highlighting the model's difficulty in distinguishing semantically related aggregate fields that share similar patterns. Overall, the visualization demonstrates how CP provides an interpretable, document-level analysis of model behavior, highlighting layout-dependent uncertainty patterns that are not captured by aggregate accuracy metrics.

4.3 Ablation study

In the following, we analyze what impact different choices for α have on the resulting prediction sets. To this end, Fig. 7 shows the histograms of set sizes for three different values of α. The histogram for α=0.02 is therefore the one from Fig. 3. As can be seen, and as expected, the distribution of set sizes shifts increasingly to the right as the value of α decreases. When very high statistical guarantees such as 99.5% coverage are required, the KIE model produces prediction sets

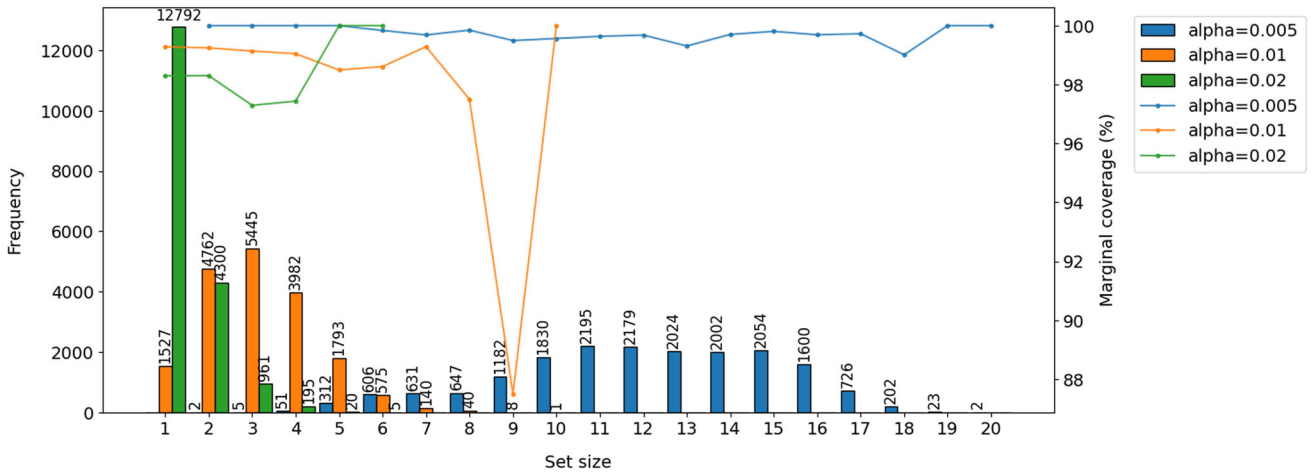

Fig. 7 Impact of alpha values on set sizes

containing mostly between 9 and 16 predicted labels. At the same time, the single-set ratio in this case is 0%, i.e., there was no entity where the model was confident enough to produce a prediction set containing a single label. In the case of α=0.005, the average width of the prediction set is 12.12, which is significantly higher than the reported average of 1.38 regarding α=0.02. Nonetheless, with respect to the field *Prod_qty_v*, the model produces relatively small prediction sets in this scenario with an average set size of 6.35. In the case of α=0.02, however, this specific field was one of those with the highest average set sizes. This suggests that there are no clear correlations between the results for the individual fields across different α values. There are also differences regarding the marginal coverage achieved for each set size. In general, the behavior for α=0.005 is relatively consistent. In the other two cases, however, there are some outliers. For example, with α=0.01, in rare cases where the prediction set contains nine elements, the marginal coverage drops to 87.5%. Therefore, the model is not only uncertain, but it is also unable to detect the correct label.

When choosing α, a trade-off must be made between statistical guarantees and the resulting overhead from large prediction sets. While a very high guarantee of 99.5% might be desirable in real-world settings where overlooked falsely extracted values can result in severe follow-up costs, managing such large prediction sets becomes increasingly complex. Conversely, one might prioritize a high single-set ratio, which however usually comes at the expense of lower statistical guarantees. Based on the findings, α=0.01 provides a reasonable balance between those two factors while still achieving a single-set ratio of 8.36%.

5 Discussion

5.1 General

An inspection of prediction patterns reveals that KIE models sometimes rely more heavily on spatial layout than on textual or visual content. For example, *Time_k* and *Time_v* often co-occur in prediction sets, suggesting that the model defaults to selecting nearby tokens regardless of their semantic differences. Likewise, monetary fields like *Prod_price_v*, *Subtotal_v* and *Total_v* are frequently confused, likely because these values appear in similar table positions on receipts.

The average width of prediction sets for values and keywords are nearly identical: 1.464 for the former and 1.455 for the latter. This indicates that the model's raw uncertainty is not substantially different between the two types. However, marginal coverage shows slight discrepancies: labels for values achieve an average coverage of about 97.67%, whereas keywords only reach about 95.98%. This suggests that, despite similar set sizes, the model more reliably includes the correct label when predicting structured numeric or formatted entities as is the case for the receipts. Keywords, whose wording, font styles and spatial placement vary, are therefore more prone to occasional coverage failures. Examining co-occurrences shows that *Others* often appear alongside product item values and total values, reflecting the fact that miscellaneous text (e.g., restaurant names, store location info) appears adjacent to structured fields. The model is relatively "safer" when classifying strictly formatted entities, whereas free-form text tokens fall into *Others*, resulting in broader uncertainty but decent coverage nonethe-

less. The discrepancy between point prediction results and CP results highlights that high average F1 scores exaggerate end-to-end reliability, as localized errors can still disrupt entire document extractions. CP thus exposes unseen risks by quantifying uncertainty at the token level, revealing that real-world KIE pipelines should incorporate CP-based confidence filtering or human verification to mitigate these blind spots. Together, these analyses underscore that, despite achieving strong token-level performance, advanced KIE systems still show keyword ambiguity and coverage gaps that standard metrics fail to capture.

Real-time feasibility. The proposed framework is implemented as a lightweight post-hoc wrapper around the underlying KIE model and therefore introduces negligible additional computational overhead. The calibration step is carried out entirely offline on a held-out dataset and does not involve any additional training or model optimization. It consists solely of running forward passes on the calibration data to compute nonconformity scores and extracting the corresponding quantile thresholds. At deployment time, inference requires only a single standard forward pass of the base model, followed by simple threshold comparisons on the output logits to construct prediction sets. No additional model evaluations, sampling procedures or gradient-based computations are required. As a result, the dominant computational cost in both calibration and real-time settings remains the KIE backbone itself, making the proposed approach feasible even for resource-constrained workflows. In such cases, the same CP methodology could be applied to more lightweight backbone models, enabling a flexible trade-off between model capacity and computational efficiency without modifying the CP framework.

5.2 Implications for real-world applications

Although average token-level coverage is high, indicating that CP provides strong statistical guarantees, this does not imply complete document-level correctness. As shown earlier, 41.44% of documents contain at least one token whose true label is not included in the prediction set. This highlights a fundamental challenge in KIE: strong average token-level performance can still translate into frequent partial failures at the document level. In practice, this means that even well-calibrated models may require substantial manual review when full field-level correctness is required. In document processing tasks such as invoice or receipt analysis, a single mislabeled token can lead to erroneous extractions, particularly for critical fields such as totals, dates, or identifiers. These findings emphasize the need to complement CP with field-level and document-level aggregation strategies that better align with operational requirements.

More broadly, CP enables hybrid processing pipelines in which uncertainty directly drives downstream decision-making rather than serving solely as a post hoc signal. Extracted entities often trigger business actions, such as initiating payments or matching transactions. By outputting prediction sets instead of single labels, CP allows conditional handling based on uncertainty. For example, singleton prediction sets with high marginal coverage can be processed automatically, while larger sets can be routed for human review.

Depending on organizational risk tolerance, such workflows may include:

- Automatically processing fields with singleton prediction sets.
- Routing multi-label prediction sets to human-in-the-loop validation.
- Prioritizing documents based on the number of uncertain fields.
- Selecting fields for manual review based on uncertainty thresholds, informed by the relationship between set size and marginal coverage (Sect. 4.2.2).

This approach enables risk-aware automation, where CP's statistical guarantees are directly leveraged to balance reliability and resource allocation in large-scale document processing pipelines.

A further aspect to consider is that standard CP assumes exchangeability of data, which can be violated in real-world document pipelines where layouts evolve. While this assumption holds for our study due to usage of the WildReceipt dataset, production deployments may require mechanisms to maintain validity under distribution shift. In practice, this can be addressed through periodic recalibration using recent data [5] or so called Adaptive Conformal Inference that dynamically adjusts the risk level alpha, ensuring coverage remains stable as the document distribution changes [4].

5.3 Explainability and uncertainty

The proposed framework operates on the outputs of DL models and therefore inherits their limitations. We argue that this dependence is not a restriction, but rather a strength of CP. The proposed method treats the underlying KIE system as a black box and does not rely on architectural assumptions, internal representations or training procedures. As a result, the validity guarantees of CP are model-agnostic and hold for any underlying predictor, independent of whether the backbone is LayoutLM, ERNIE or another KIE model. It is also important to distinguish the scope of CP from Explainable AI (XAI) methods. XAI techniques aim to answer the question of *why* a model produced a specific prediction, typi-

cally through feature attribution. While such explanations are valuable for interpretability, they do not provide statistical guarantees about the correctness of individual predictions. A prediction can be fully explainable yet confidently incorrect. In contrast, CP addresses a fundamentally different question: whether a given prediction can be trusted. By producing prediction sets with guaranteed marginal coverage, CP provides a formal, distribution-free notion of reliability that is not available through standard XAI methods. From an operational perspective, this distinction is critical. In automated document processing pipelines, downstream systems often require binary and/or risk-aware decisions rather than post-hoc explanations. For example, it must be decided whether to automatically process the extracted invoice total value or route it for manual verification. CP enables such critical decisions by transforming point predictions into calibrated uncertainty-aware outputs, allowing systems to condition actions on explicit confidence guarantees. This capability is complementary to explainability: CP governs when to trust a model, while XAI explains why a model behaved as it did.

5.4 Limitations

The findings are subject to certain limitations. Firstly, the dataset is annotated on segment-level as mentioned in Sect. 3.2. In this regard, it might be worthwhile investigating more fine-granular datasets in order to see if different observations could be made. Secondly, all fields to be extracted were treated equally during the experiments. In real-world settings, however, fields may be prioritized. For example, it might be less problematic if the keyword of the telephone number was incorrectly extracted compared to the receipt total amount. Thirdly, the chosen CP method assumes exchangeability, and while marginal coverage is guaranteed, no additional calibration across individual labels was performed. This means that certain fields may be consistently over- or under-covered without adjustment. Statistical guarantees for every single label can be obtained by Class-Conditional CP methods [1], however we used the Split CP setup. Also, while our CP methodology remains theoretically valid, the selected calibration data can make the estimated thresholds noisy for rare labels. The coverage for corresponding entities may fluctuate, highlight the challenge of robust calibration under resource-constrained conditions. Last, but not least, we only analyzed sequence-based approaches such as LayoutLM. To obtain a more generalizable view of the uncertainty of KIE methods, experiments should be conducted on other methods, such as graph-based KIE systems.

5.5 Future work

A document processing workflow typically concludes with the extracted values being fed into downstream systems, such as ERP. Therefore, a post-processing step is necessary to convert the predicted labels into extracted text. This involves, among others, a selection of textual information for each relevant field. Based on KIE model predictions, multiple candidates may exist for a given label. To extend our approach, we plan to integrate a candidate selection procedure, for example by using Mixed-Integer Linear Programming (MILP), to select final extracted values from the obtained prediction sets. This enables a structured and comprehensible decision-making process for the extraction candidates.

Since LLM-based KIE methods have become increasingly popular in recent times, it seems worthwhile to investigate CP methods in conjunction with LLMs for KIE tasks. In particular since applications of LLMs suffer from hallucinations and intransparent model decisions, measuring the uncertainty for LLMs represents a key challenge moving forward. Possible CP methods are proposed by [17].

Another aspect of future work is integrating the CP methods into a validation tool for real-world workflows. To this end, color-coded or otherwise highlighted words based on the KIE model's uncertainty given a user-defined α value, can be integrated (similar to Fig. 6). This allows human validators to focus their attention on uncertain predictions and could significantly improve the efficiency of the validation process.

Future work could also be done on how insights from our analyses can be transferred into future KIE research. As the findings show, corresponding models show different behavior with respect to different types of fields to extract. The results also suggest that textual and positional modalities play different roles in the predictions in some cases. A more in-depth analysis on how this can be used to optimize technical aspects of KIE systems can be worthwhile. This could include, for example, the attention mechanisms in transformer models, or the way how different input modalities are fused and provided to the model.

6 Related work

KIE methods fall into three main categories based on document representation. Graph-based approaches model the elements (e.g., words, characters) as nodes with flexible edge definitions [11, 16]. Grid-based methods overlay structured grids (e.g., at character level) using rectilinear links, with values derived from overlapping content [2, 9]. Sequence-based techniques linearize documents while preserving layout and visual cues, processed via sequence labeling [15, 20, 22].

In general, KIE literature pays little attention to the uncertainty of model predictions. The closest related work is proposed by [10], more specifically a method to obtain confidence scores for KIE where invoices with low scores are

forwarded for manual review. The system uses multiple Convolutional Neural Networks which act as feature extractors, followed by a binary classifier to predict whether the KIE system could extract all fields correctly or not. A limitation of the approach is that it requires separate training for the models and the confidence scorer.

Besides, work exists on applying CP methods to different sequence-based transformer models [3, 6, 13]. However, the authors apply these techniques to different tasks such as paraphrase detection or sentiment analysis. As mentioned, KIE has not yet been considered as an application scenario.

7 Conclusion

In this paper, we studied the application of CP to KIE as a principled approach to UQ and model analysis. While state-of-the-art KIE models achieve strong token-level performance, our results demonstrate a substantial gap between conventional evaluation metrics and real-world reliability. Through the analysis of prediction set sizes, label-wise coverage and document-level uncertainty, we showed that CP provides statistically rigorous confidence guarantees while simultaneously exposing structural weaknesses in KIE models. These include confusion between semantically similar labels, sensitivity to spatial layout and reduced reliability for rare or ambiguous fields. Such insights are essential for developing KIE systems that are not only accurate but also trustworthy in operational business workflows. Finally, we discussed the practical implications of CP for deployment. High token-level coverage alone does not remove the need for human validation. Instead, CP enables risk-aware automation strategies by guiding selective human-in-the-loop review and prioritization based on aggregated uncertainty signals. This facilitates more robust document processing pipelines that balance automation efficiency with accuracy and trust.

Author Contributions Alexander Rombach: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Writing - Revision, Visualization Nijat Mehdiyev: Conceptualization, Methodology, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Writing - Revision, Supervision.

Funding: Open Access funding enabled and organized by Projekt DEAL. This research was funded in part by the Federal Ministry of Education and Research (BMBF) under grant number 01IS23064 (Project Taxas). Besides, there are no related financial or non-financial interests.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Angelopoulos, A.N., Bates, S.: A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. (2022). arXiv: 2107.07511
2. Denk, T.I, Reisswig, C.: BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding. Workshop on Document Intelligence at NeurIPS (2019) arXiv: 1909.04948
3. Dey, N., Ding, J., Ferrell, J., et al.: Conformal prediction for text infilling and part-of-speech prediction. The New England J. Stat. Data Sci. **1**(1), 69–83 (2023). https://doi.org/10.51387/22-NEJSDS8
4. Gibbs, I., Candes, E.: Adaptive conformal inference under distribution shift. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., et al. (eds) Advances in Neural Information Processing Systems, vol 34. Curran Associates Inc, pp 1660–1672. (2021). https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf
5. Gibbs, I/, Candès, E.: Conformal inference for online prediction with arbitrary distribution shifts. J. Mach. Learn. Res. 25(1) (2024)
6. Giovannotti, P., Gammerman, A.: Transformer-based conformal predictors for paraphrase detection. In: Carlsson, L., Luo, Z., Cherubin, G., et al. (eds) Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research, vol 152. PMLR, pp 243–265 (2021), https://proceedings.mlr.press/v152/giovannotti21a.html
7. Houy, C., Hamberg, M., Fettke, P.: Robotic process automation in public administrations. In: Räckers, M., Halsbenning, S., Rätz, D., et al. (eds.) Digitalisierung Von Staat Und Verwaltung, pp. 62–74. Gesellschaft für Informatik e.V, Bonn (2019)
8. Huang, Z., Chen, K., He, J., et al.: ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp 1516–1520 (2019) https://doi.org/10.1109/ICDAR.2019.00244
9. Katti, A.R., Reisswig, C., Guder, C., et al.: Chargrid: Towards understanding 2D documents. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 4459–4469 (2018). https://doi.org/10.18653/v1/d18-1476
10. Kivimäki, J., Lebedev, A., Nurminen, J.K.: Failure Prediction in 2D Document Information Extraction with Calibrated Confidence Scores. In: Proceedings - International Computer Software and Applications Conference, vol 2023-June. IEEE, pp 193–202 (2023). https://doi.org/10.1109/COMPSAC57700.2023.00033
11. Liu, X., Gao, F., Zhang, Q., et al.: Graph convolution for multimodal information extraction from visually rich documents. In: NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies - Proceedings of the Conference, vol 2. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 32–39 (2019). https://doi.org/10.18653/v1/n19-2005
12. Majumder, B.P., Potti, N., Tata, S., et al.: Representation Learning for Information Extraction from Form-like Documents. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 6495–6504 (2020). https://doi.org/10.18653/v1/2020.acl-main.580
13. Maltoudoglou, L., Paisios, A., Papadopoulos, H.: BERT-based conformal predictor for sentiment analysis. In: Gammerman, A., Vovk, V., Luo, Z., et al. (eds) Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research, vol 128. PMLR, pp 269–284 (2020)
14. Nourbakhsh, A., Shah, S., Rose, C.: Towards a new research agenda for multimodal enterprise document understanding: What are we missing? In: Ku LW, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics ACL 2024. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 14610–14622 (2024). https://doi.org/10.18653/v1/2024.findings-acl.870
15. Peng, Q., Pan, Y., Wang, W., et al.: ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 3744–3756 (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.274
16. Qian, Y., Santus, E., Jin, Z., et al.: GraphIE: A Graph-Based Framework for Information Extraction. In: NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 751–76 (2019) https://doi.org/10.18653/v1/N19-1082, arXiv:1810.13083
17. Quach, V., Fisch, A., Schuster, T., et al.: Conformal Language Modeling. In: The Twelfth International Conference on Learning Representations, (2024). https://openreview.net/forum?id=pzUhfQ74c5
18. Rombach, A.M., Fettke, P.: Deep learning based key information extraction from business documents: systematic literature review. ACM Comput. Surv. **58**(2), 1–37 (2025). https://doi.org/10.1145/3749369
19. Rombach, A.M., Lahann, J., Niesen, T., et al.: Utilizing deep learning for field-level information extraction from german real estate tax notices. J. Emerg. Tech. Accounting **22**(1), 101–118 (2025). https://doi.org/10.2308/JETA-2023-028
20. Sassioui, A., Benouini, R., El Ouargui, Y., et al.: Visually-Rich Document Understanding: Concepts, Taxonomy and Challenges. In: Proceedings - 10th International Conference on Wireless Networks and Mobile Communications, WINCOM 2023. IEEE, pp 1–7 (2023). https://doi.org/10.1109/WINCOM59760.2023.10322990
21. Sun, H., Kuang, Z., Yue, X., et al.: Spatial Dual-Modality Graph Reasoning for Key Information Extraction. (2021). arXiv:2103.14470
22. Xu, Y., Li, M., Cui, L., et al.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp 1192–1200 (2020). https://doi.org/10.1145/3394486.3403172, arXiv: 1912.13318
23. Xu, Y., Xu, Y., Lv, T., et al.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: ACL-IJCNLP 2021–59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 2579–2591(2021) https://doi.org/10.18653/v1/2021.acl-long.201 arXiv:2012.14740