EMNLP 2024
Industry Track

Language, OCR, Form Independent (LOFI) pipeline for Industrial Document Information Extraction

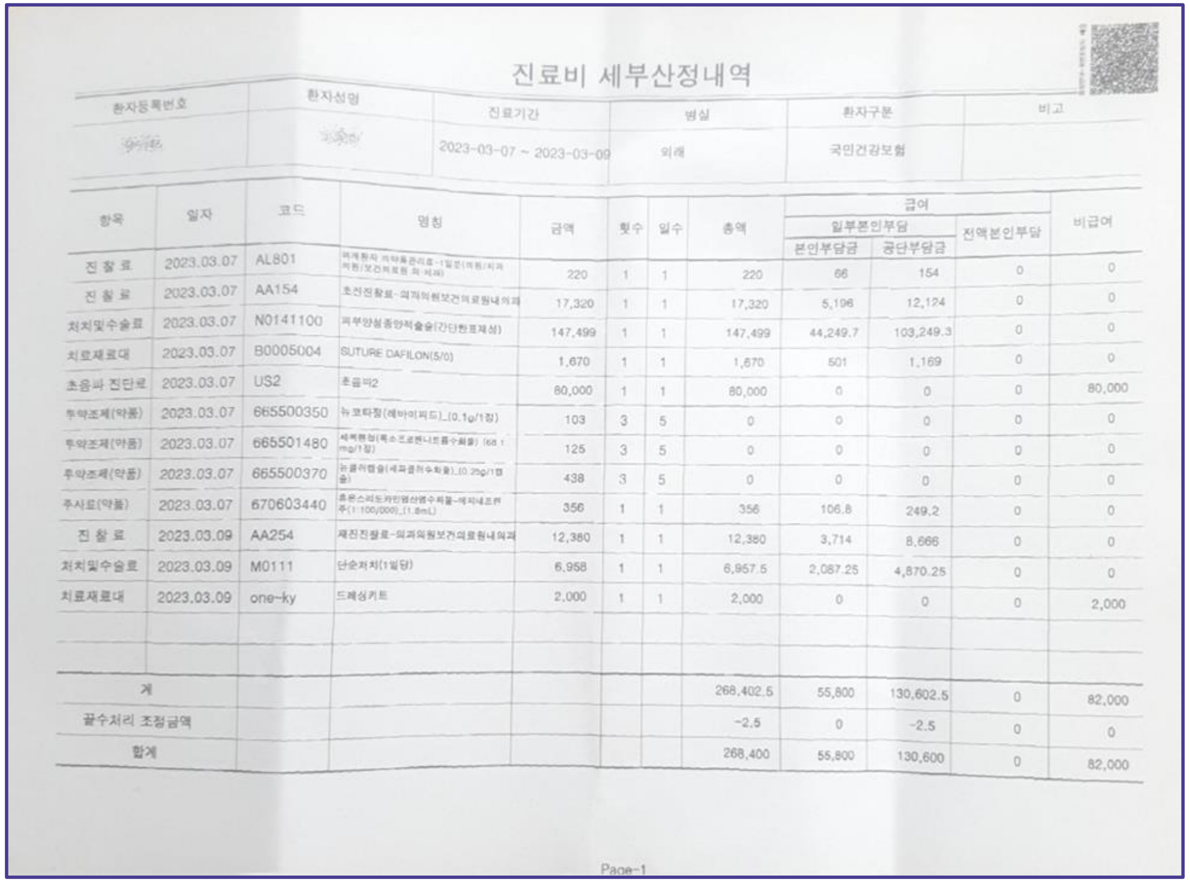
Agile SoDA

Chang Oh Yoon[1,+], Wonbeen Lee[1,+], Seokhwan Jang[1,+] ,
Kyuwon Choi[2,+] , Minsung Jung[2,+] ,
Daewoo Choi[3,*] ,

+AgileSoDA, *Hankuk University of Foreign Studies

1. Introduction

Many industries handle complex documents known as Visually Rich Documents (VRDs).



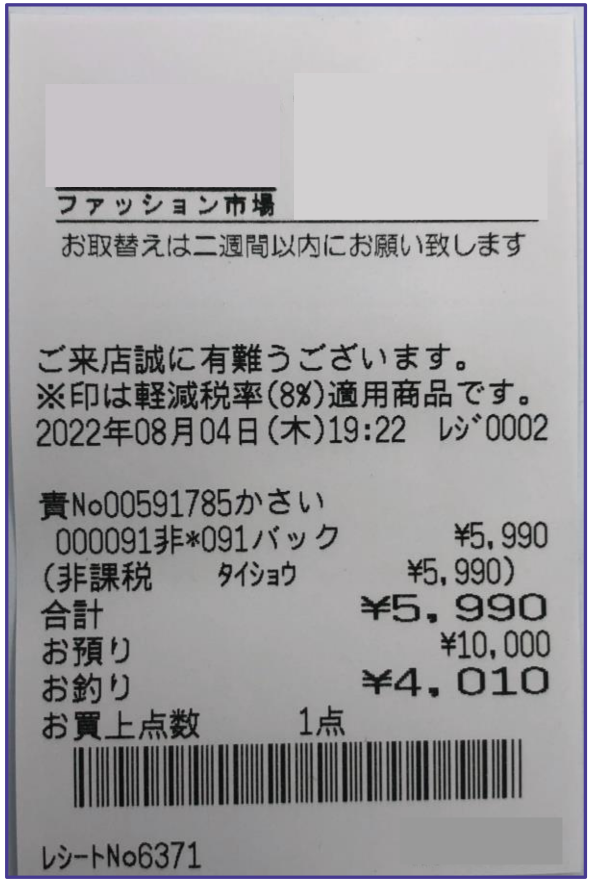
진료비 세부산정내역

환자등록번호	환자성명	진료기간	병실	환자구분	비고
[redacted]	[redacted]	2023-03-07 ~ 2023-03-09	외래	국민건강보험	

항목	일자	코드	명칭	금액	횟수	일수	총액	급여			비급여
								일부본인부담		전액본인부담	
								본인부담금	공단부담금		
진찰료	2023.03.07	AL801	외래환자 의약품관리료-1일분(의원/치과의원/보건의료원 외·외래)	220	1	1	220	66	154	0	0
진찰료	2023.03.07	AA154	초진진찰료-의과의원보건의료원내의과	17,320	1	1	17,320	5,196	12,124	0	0
처치및수술료	2023.03.07	N0141100	피부양성종양적출술(간단한표재성)	147,499	1	1	147,499	44,249.7	103,249.3	0	0
치료재료대	2023.03.07	B0005004	SUTURE DAFILON(5/0)	1,670	1	1	1,670	501	1,169	0	0
초음파 진단료	2023.03.07	US2	초음파2	80,000	1	1	80,000	0	0	0	80,000
투약조제(약품)	2023.03.07	665500350	뉴코타정(레바미피드)_(0.1g/1정)	103	3	5	0	0	0	0	0
투약조제(약품)	2023.03.07	665501480	세팩펜정(록소프로펜나트륨수화물) (68.1 mg/1정)	125	3	5	0	0	0	0	0
투약조제(약품)	2023.03.07	665500370	뉴클러캡슐(세파클러수화물)_(0.25g/1캡슐)	438	3	5	0	0	0	0	0
주사료(약품)	2023.03.07	670603440	휴온스리도카인염산염수화물-에피네프린주(1:100,000)_(1.8mL)	356	1	1	356	106.8	249.2	0	0
진찰료	2023.03.09	AA254	재진진찰료-의과의원보건의료원내의과	12,380	1	1	12,380	3,714	8,666	0	0
처치및수술료	2023.03.09	M0111	단순처치(1일당)	6,958	1	1	6,957.5	2,087.25	4,870.25	0	0
치료재료대	2023.03.09	one-ky	드레싱키트	2,000	1	1	2,000	0	0	0	2,000
계							268,402.5	55,800	130,602.5	0	82,000
끝수처리 조정금액							-2.5	0	-2.5	0	0
합계							268,400	55,800	130,600	0	82,000

Page-1

Koream medical bills



[redacted]

[redacted]

ファッション市場

お取替えは二週間以内にお願い致します

ご来店誠に有難うございます。
※印は軽減税率(8%)適用商品です。
2022年08月04日(木)19:22　レジ゛0002

責No00591785かさい

000091非*091バック	¥5,990
(非課税　タイショウ	¥5,990)
合計	¥5,990
お預り	¥10,000
お釣り	¥4,010
お買上点数	1点

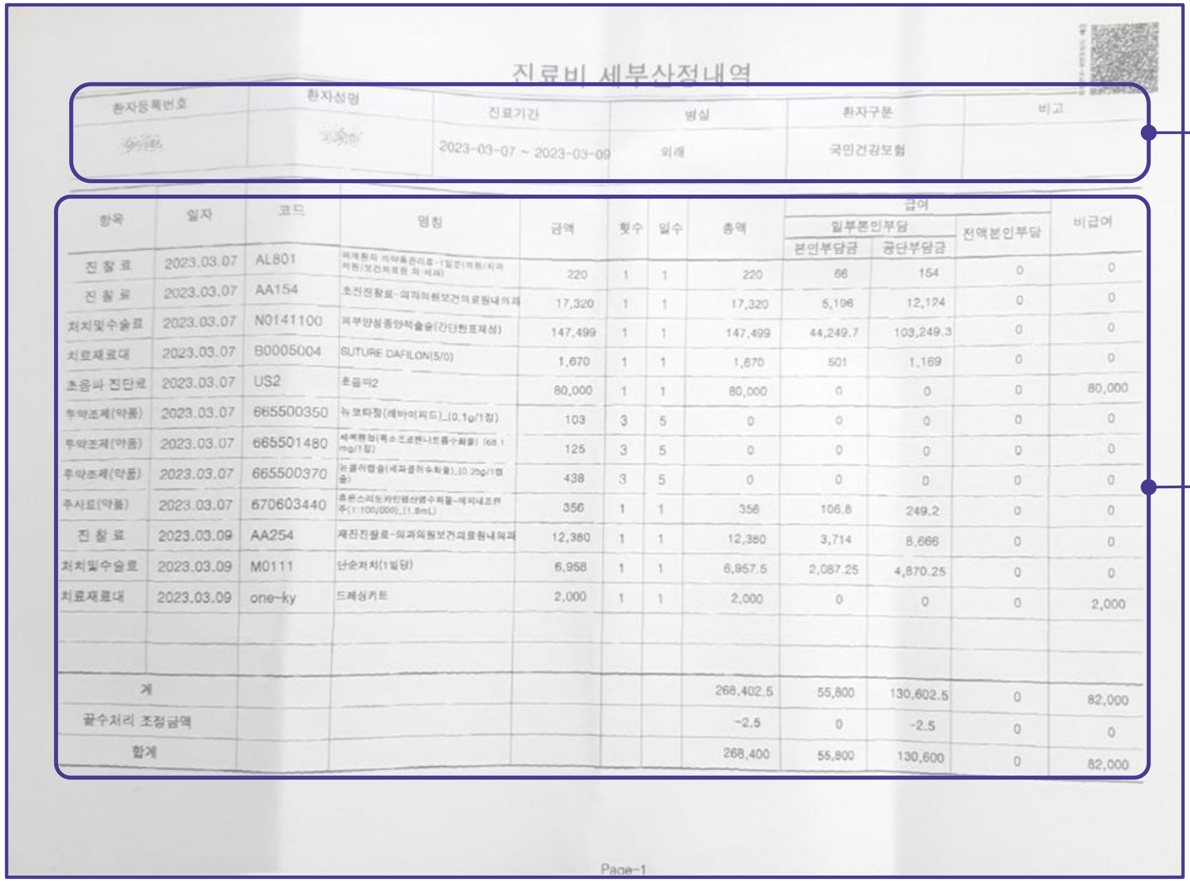
レシートNo6371

[redacted]

Japanese receipts

1. Introduction

In real-world industry, we should consider a process of SER (Semantic Entity Recognition) to automate workflows.



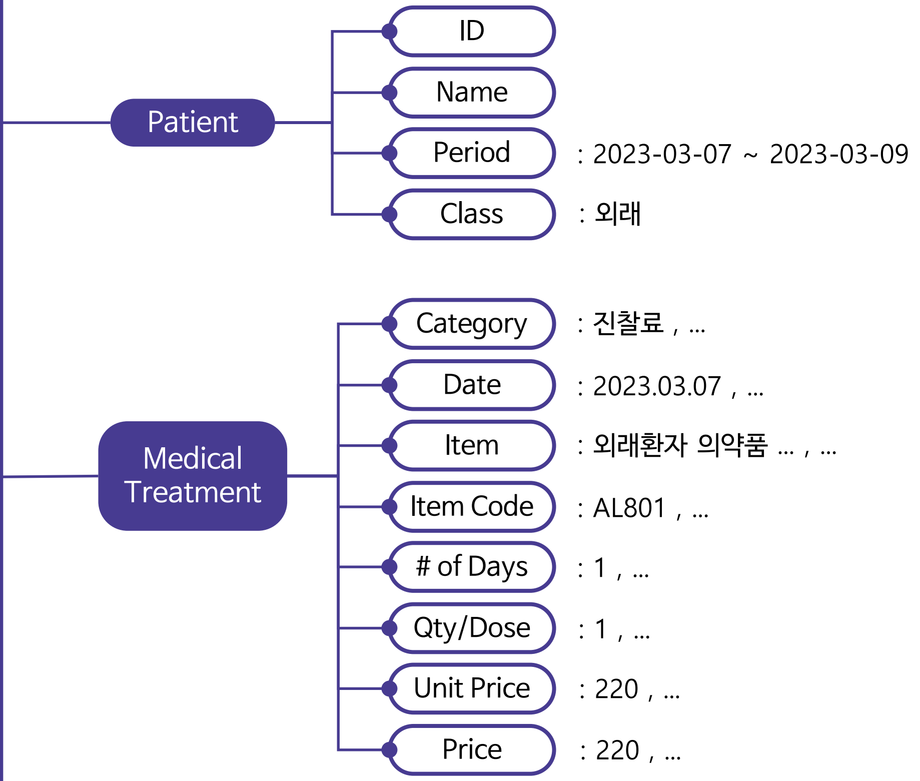
진료비 세부산정내역

환자등록번호	환자성명	진료기간	병실	환자구분	비고
[redacted]	[redacted]	2023-03-07 ~ 2023-03-09	외래	국민건강보험	

항목	일자	코드	명칭	금액	횟수	일수	총액	급여			비급여
								일부본인부담		전액본인부담	
								본인부담금	공단부담금		
진 찰 료	2023.03.07	AL801	외래환자 의약품관리료-1일분(의원/치과의원/보건의료원 외래)	220	1	1	220	66	154	0	0
진 찰 료	2023.03.07	AA154	초진진찰료-의과의원보건의료원내의과	17,320	1	1	17,320	5,196	12,124	0	0
처치및수술료	2023.03.07	N0141100	피부양성종양적출술(간단한표재성)	147,499	1	1	147,499	44,249.7	103,249.3	0	0
치료재료대	2023.03.07	B0005004	SUTURE DAFILON(5/0)	1,670	1	1	1,670	501	1,169	0	0
초음파 진단료	2023.03.07	US2	초음파2	80,000	1	1	80,000	0	0	0	80,000
투약조제(약품)	2023.03.07	665500350	뉴코타정(레바미피드)_(0.1g/1정)	103	3	5	0	0	0	0	0
투약조제(약품)	2023.03.07	665501480	세퍽펜정(록소프로펜나트륨수화물)_(68.1mg/1정)	125	3	5	0	0	0	0	0
투약조제(약품)	2023.03.07	665500370	뉴클러캡슐(세파클러수화물)_(0.25g/1캡슐)	438	3	5	0	0	0	0	0
주사료(약품)	2023.03.07	670603440	휴온스리도카인염산염수화물-에피네프린주(1:100,000)_(1.8mL)	356	1	1	356	106.8	249.2	0	0
진 찰 료	2023.03.09	AA254	재진진찰료-의과의원보건의료원내의과	12,380	1	1	12,380	3,714	8,666	0	0
처치및수술료	2023.03.09	M0111	단순처치(1일당)	6,958	1	1	6,957.5	2,087.25	4,870.25	0	0
치료재료대	2023.03.09	one-ky	드레싱키트	2,000	1	1	2,000	0	0	0	2,000
계							268,402.5	55,800	130,602.5	0	82,000
끝수처리 조정금액							-2.5	0	-2.5	0	0
합계							268,400	55,800	130,600	0	82,000

Page-1

Koream medical bills



Patient

ID

Name

Period : 2023-03-07 ~ 2023-03-09

Class : 외래

Medical Treatment

Category : 진찰료 , ...

Date : 2023.03.07 , ...

Item : 외래환자 의약품 ... , ...

Item Code : AL801 , ...

# of Days : 1 , ...

Qty/Dose : 1 , ...

Unit Price : 220 , ...

Price : 220 , ...

2. Challenges

To address the automation demands of the industry, we face three main challenges.

01

Low Resource Language

- There are limited VRD datasets available for Low-Resource Languages.
- No pre-trained models exist for these languages.
- This scarcity hinders the creation of advanced language models.

02

OCR Dependency

- SER has limitations due to OCR engine output.
- OCR results are typically at the word level, not entity level.
- Additional processing (splitting or combining) may be needed for accurate semantic entities.

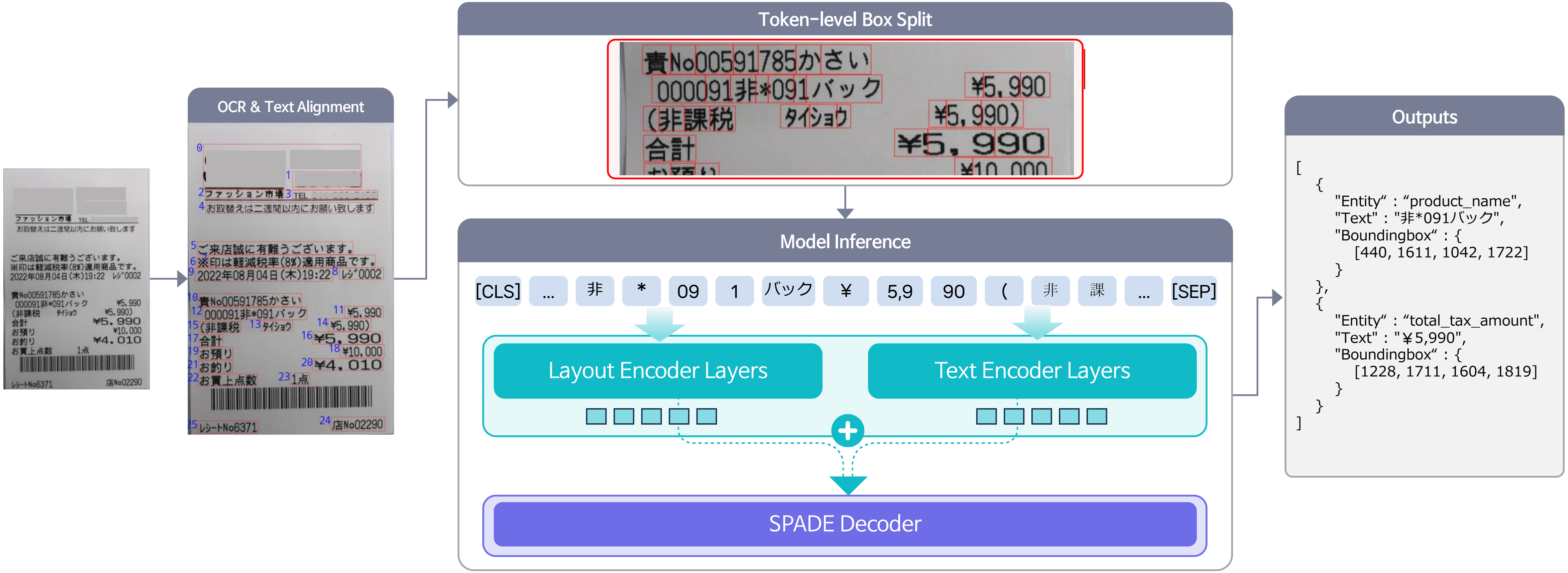
03

Form Diversity

- Industry documents pose challenges for information extraction due to custom formats.
- Even standardized forms have variations in formatting, such as custom medical report templates.
- Image distortions or rotations can alter a document's structure and further complicate extraction.

3. Language, OCR, Form Independent (LOFI) pipeline

So, we present a **L**anguage, **O**CR, and **F**orm **I**ndependent pipeline, named **LOFI pipeline**.

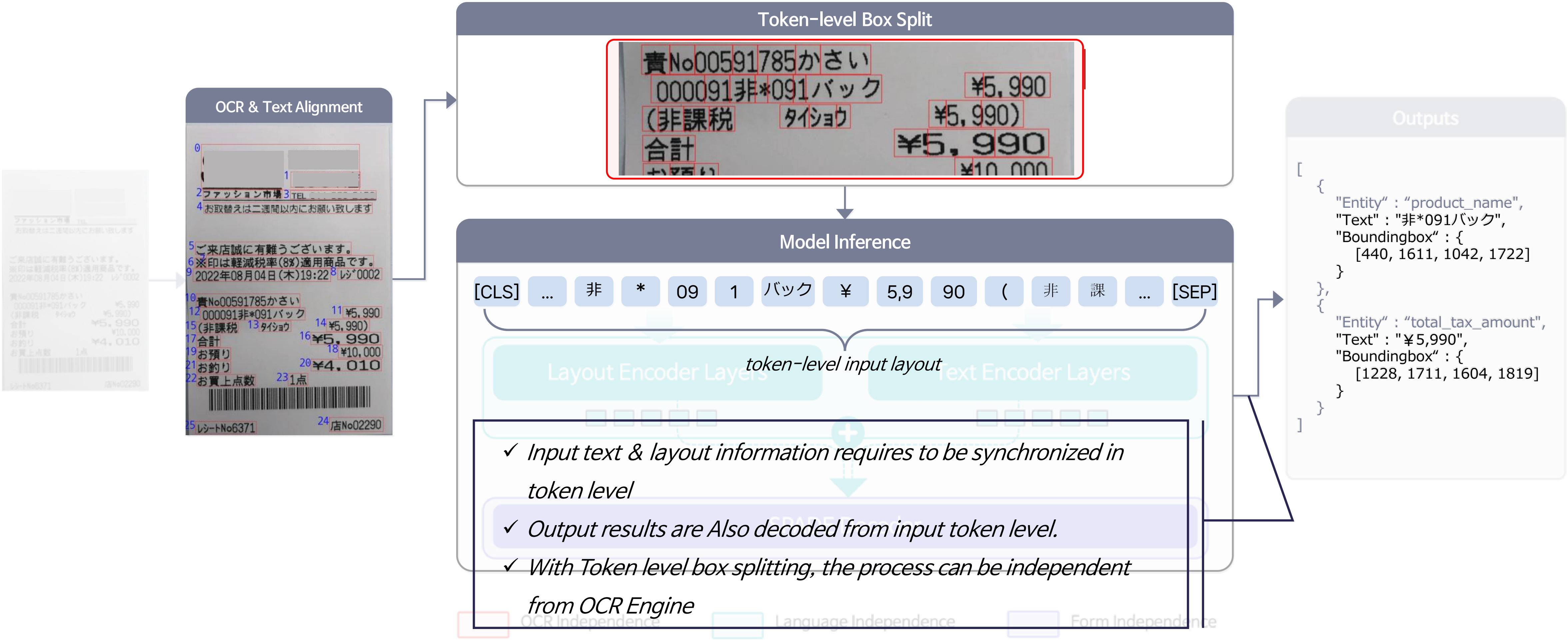


```
[
  {
    "Entity“ : “product_name",
    "Text" : "非*091バック",
    "Boundingbox“ : {
      [440, 1611, 1042, 1722]
    }
  },
  {
    "Entity“ : “total_tax_amount",
    "Text" : "￥5,990",
    "Boundingbox“ : {
      [1228, 1711, 1604, 1819]
    }
  }
]
```

🟥	OCR Independence
🟦	Language Independence
🟪	Form Independence

3. Language, OCR, Form Independent(LOFI) pipeline

We constructed a token-level box splitting to standardize bounding box ranges from various OCR engines.

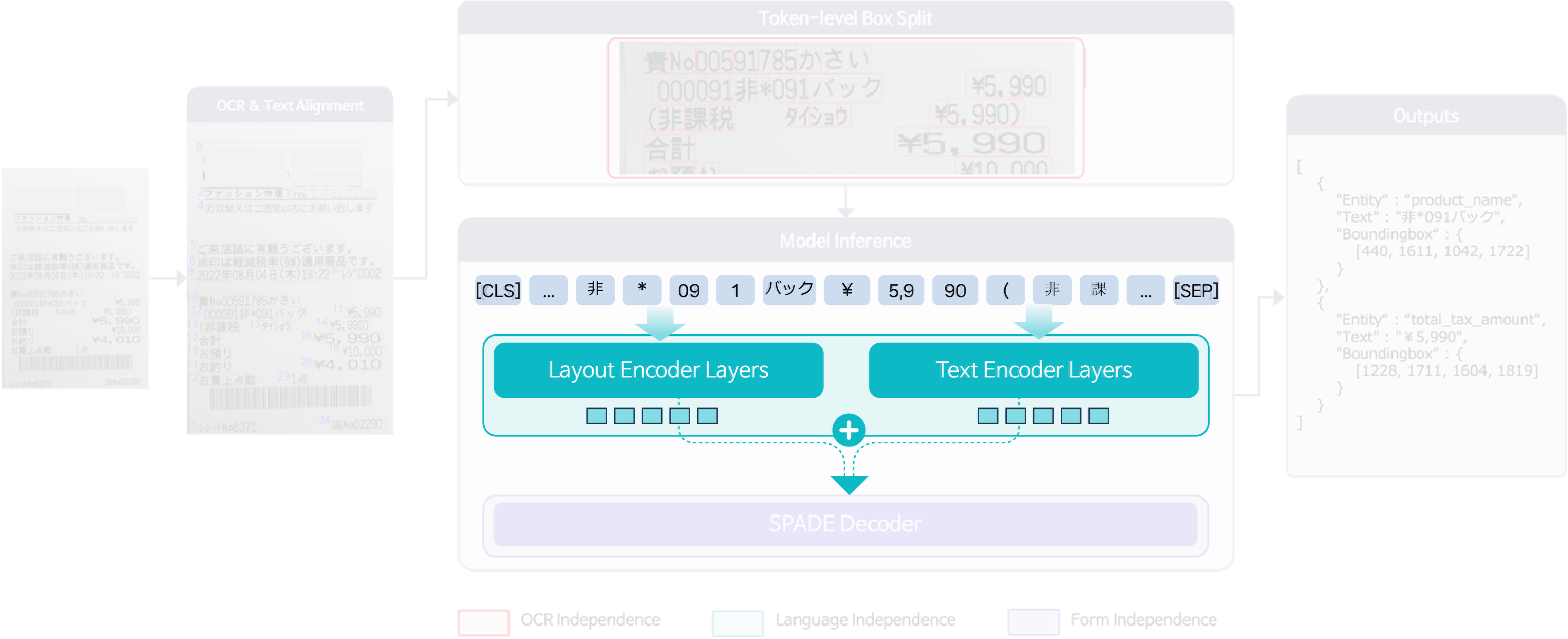


```
[
   {
      "Entity“ : “product_name",
      "Text" : "非*091バック",
      "Boundingbox“ : {
         [440, 1611, 1042, 1722]
      }
   },
   {
      "Entity“ : “total_tax_amount",
      "Text" : "￥5,990",
      "Boundingbox“ : {
         [1228, 1711, 1604, 1819]
      }
   }
]
```

- ✓ *Input text & layout information requires to be synchronized in token level*
- ✓ *Output results are Also decoded from input token level.*
- ✓ *With Token level box splitting, the process can be independent from OCR Engine*

3. Language, OCR, Form Independent(LOFI) pipeline

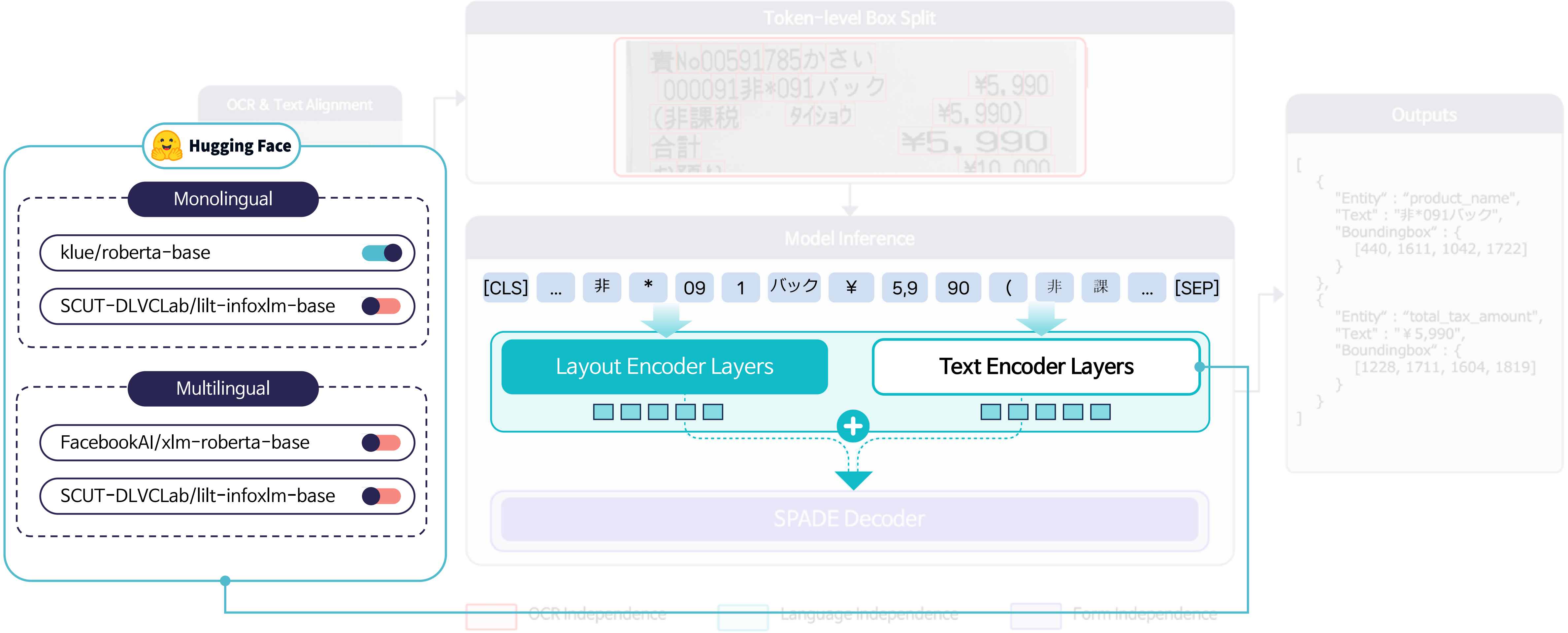
We implemented a language flexible multimodal model for Low-Resource Language(LRL).



```
[
   {
      "Entity" : "product_name",
      "Text" : "非*091バック",
      "Boundingbox" : {
         [440, 1611, 1042, 1722]
      }
   },
   {
      "Entity" : "total_tax_amount",
      "Text" : "¥5,990",
      "Boundingbox" : {
         [1228, 1711, 1604, 1819]
      }
   }
]
```

3. Language, OCR, Form Independent (LOFI) pipeline

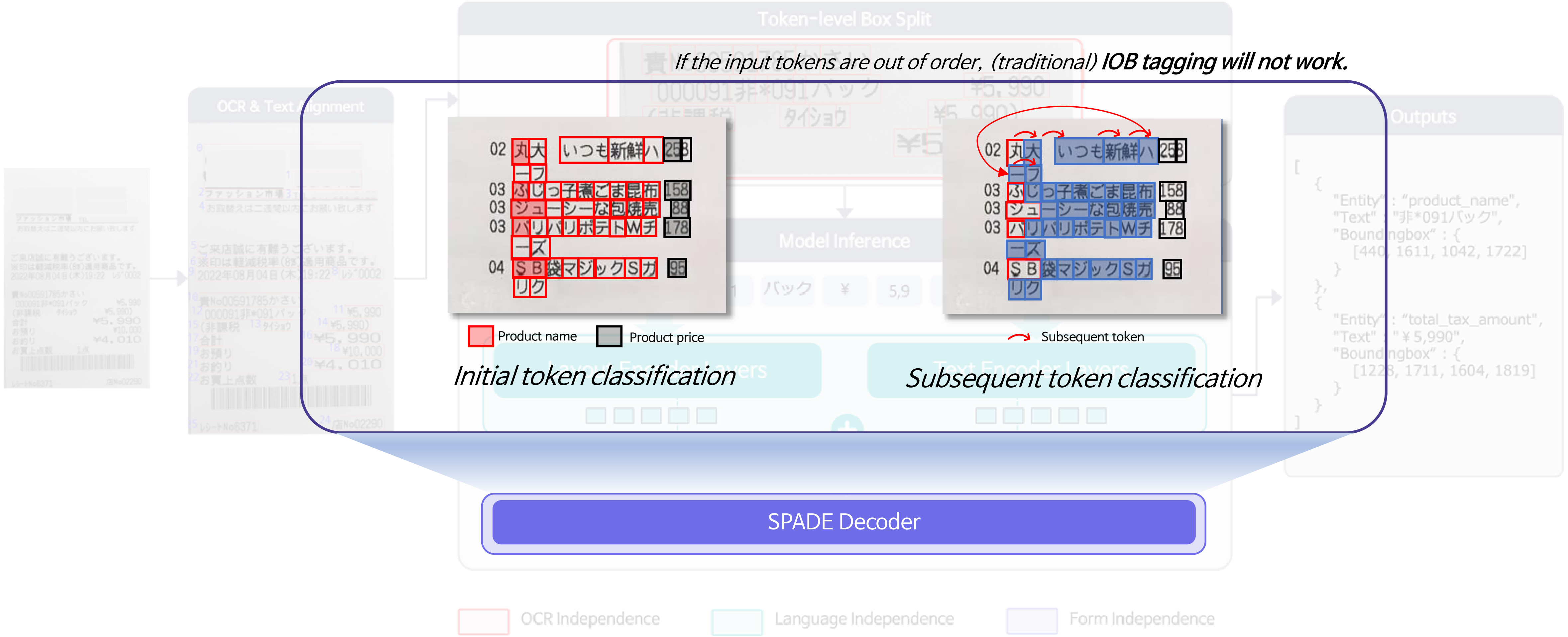
We implemented a language flexible multimodal model for Low-Resource Language (LRL).



```
[
  {
    "Entity" : "product_name",
    "Text" : "非*091バック",
    "Boundingbox" : {
      [440, 1611, 1042, 1722]
    }
  },
  {
    "Entity" : "total_tax_amount",
    "Text" : "￥5,990",
    "Boundingbox" : {
      [1228, 1711, 1604, 1819]
    }
  }
]
```

3. Language, OCR, Form Independent (LOFI) pipeline

We added SPADE decoder for operating independently of document formats and layouts.

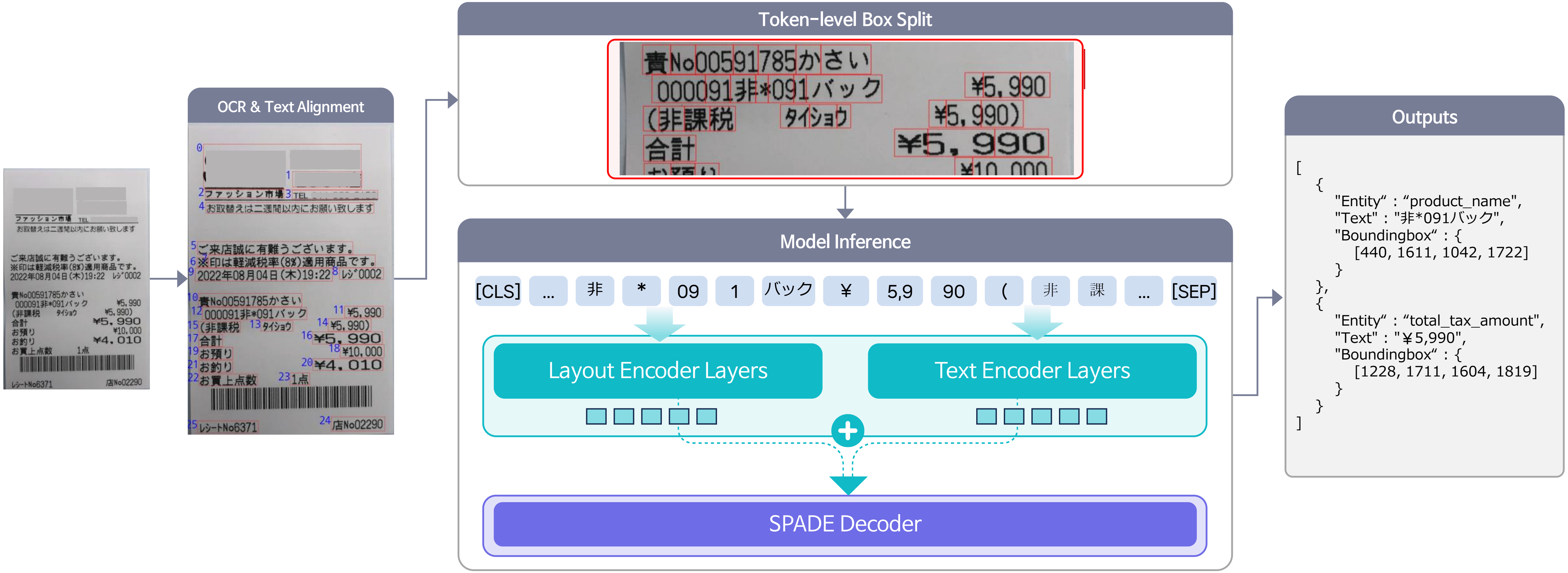


If the input tokens are out of order, (traditional) ***IOB tagging will not work.***

Initial token classification

Subsequent token classification

3. Language, OCR, Form Independent(LOFI) pipeline



OCR & Text Alignment

0 [redacted]
1 [redacted]
2 ファッション市場 3 TEL [redacted]
4 お取替えは二週間以内にお願い致します
5 ご来店誠に有難うございます。
6 7 ※印は軽減税率(8%)適用商品です。
9 2022年08月04日(木)19:22 8 ﾚｼﾞ0002
10 責No00591785かさい
12 000091非*091バック 11 ¥5,990
15 (非課税 13 ﾀｲｼｮｳ 14 ¥5,990)
17 合計 16 ¥5,990
19 お預り 18 ¥10,000
21 お釣り 20 ¥4,010
22 お買上点数 23 1点
25 ﾚｼｰﾄNo6371 24 店No02290

Token-level Box Split

Outputs

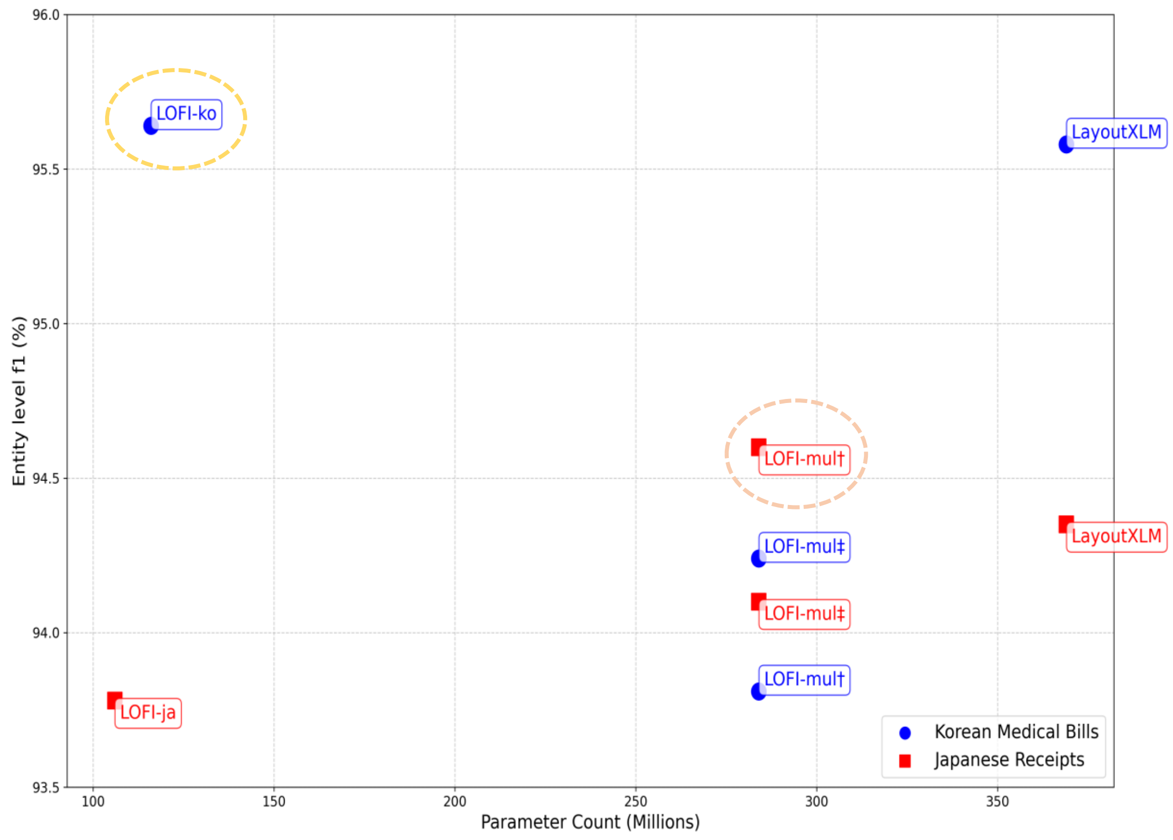
```
[
  {
    "Entity“ : “product_name",
    "Text" : "非*091バック",
    "Boundingbox“ : {
      [440, 1611, 1042, 1722]
    }
  },
  {
    "Entity“ : “total_tax_amount",
    "Text" : "￥5,990",
    "Boundingbox“ : {
      [1228, 1711, 1604, 1819]
    }
  }
]
```

🟥 OCR Independence 🟦 Language Independence 🟪 Form Independence

4. Experiments

LOFI shows better performance than LayoutXLM on Korean medical bills and Japanese receipts, also demonstrating efficiency in terms of parameters and computational resources.

Number of model parameters and entity-level F1 score



LRL Documents

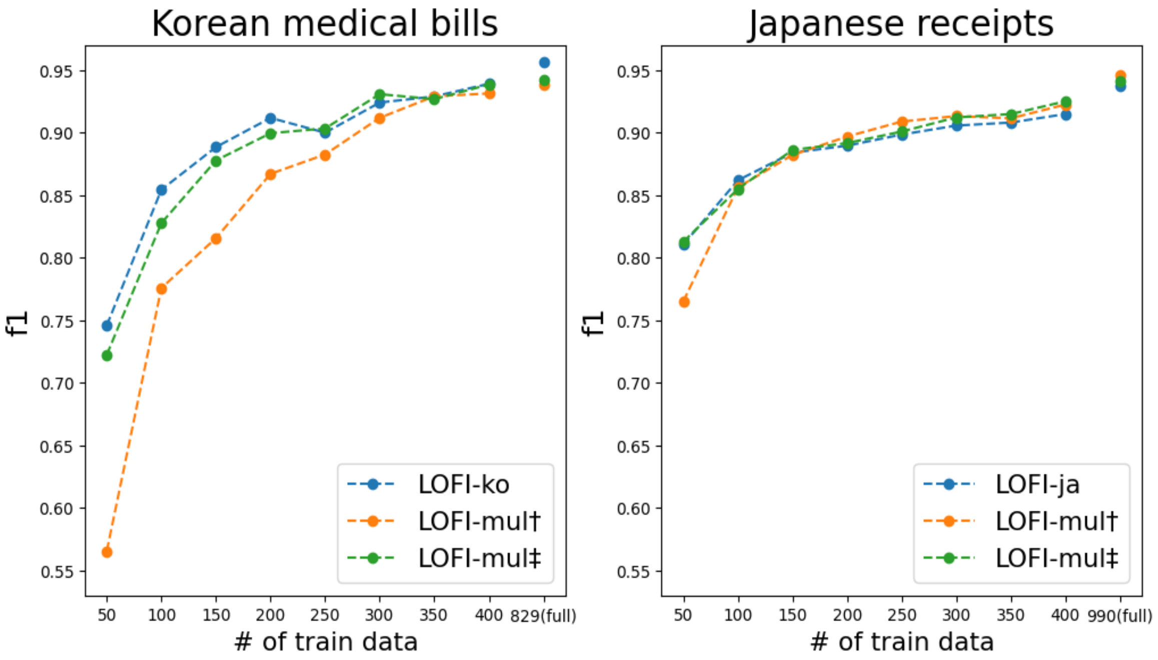
Name	Pretrained Language	Encoder	Params	Korean Medical Bills	Japanese Receipts
LayoutXLM	Multilingual	LayoutXLM-base	369M	95.58%	94.35%
LOFI-mul †	Multilingual	InfoXLM-base + lilt-only-base	284M	93.81%	94.60%
LOFI-mul‡	Multilingual	XLMRoBERTa-base + lilt-only-base	284M	94.24%	94.10%
LOFI-ko	Korean	RoBERTa-base + lilt-only-base	116M	95.64%	-
LOFI-ja	Japanese	RoBERTa-base + lilt-only-base	106M	-	93.78%

English Documents

Name	Pretrained Language	Params	FUNSD	CORD
LayoutLM	English	160M	79.27%	94.72%
LayoutLMv2	English	200M	82.76%	94.95%
LayoutLMv3	English	133M	79.38%	96.80%
BROS	English	110M	83.05%	95.73%
LOFI-en	English	131M	78.99%	96.39%

4. Experiments

Experiment results showing performance variations with different training data sizes used in fine-tuning.



- Through experiments, it was suggested that at least 300–400 training data is required to achieve satisfactory performance.
- The amount of training data required may vary by language; using fewer than 200 training documents resulted in a 5% difference in performance compared to using the full training dataset.

Contributions

- Constructed a flexible pipeline structure, LOFI (Language, OCR, Form independent Extraction) ***to account for multiple challenges in industrial data extraction.***
- The LOFI pipeline demonstrates satisfactory performance on Korean and Japanese ***datasets without additional pre-training.***
- ***Empirical evidence on industrial applicability*** of the LOFI pipeline by successfully implementing it in insurance claim processing and tax handling operations.

Future research

- ***Data augmentation*** techniques to enhance the robustness of the LOFI pipeline.
- ***Efficient annotation methods*** to reduce the annotation burden in SER tasks.
- ***Improved decoder architectures*** to handle complex document challenges and diversify AI capabilities for business scenarios.

Thank you !