

LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding

Supplementary Material

A. Dataset Details

A.1. Dataset Description

Details of the publicly available document understanding datasets used in layout-aware pre-training and layout-aware SFT are as follows:

RVL-CDIP [14] is a document classification dataset that consists of 400,000 grayscale document images in 16 classes. The 16 classes encompass diverse document types such as letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, and memo. To enhance the diversity of the data in training, sampling is performed within each class. The text and layout information are extracted by Tesseract OCR.

DocILE [65] is a large dataset for key information localization and extraction. It consists of 6,680 annotated business documents. And there are 932k unlabeled documents and 100k synthetically generated documents in the dataset for unsupervised pre-training. It also provides both word-level text and corresponding locations for the document images in the dataset by using DocTR OCR.

PubLayNet [62] is a document layout analysis dataset that covers typical document layout elements such as text, title, list, figure, and table. It consists of over 360K PDF documents sourced from articles on PubMed Central. By parsing the PDF page using PDFMiner and matching the layout with the XML representation, annotations for both page layout and text information are generated.

DocBank [24] is a document layout analysis dataset that contains 500K document pages with fine-grained token-level annotations for document layout analysis. These documents are sourced from articles on arXiv.org, covering various areas such as Physics, Mathematics, Computer Science, and others. PDFPlumber, which is a PDF parser built on PDFMiner, is used to extract text lines and non-text elements with their bounding boxes.

DocLayNet [40] is a human-annotated document layout segmentation dataset containing 80863 manually annotated PDF documents with 11 complex layout classes from diverse data sources including Finance, Science, Patents, Tenders, Law texts, and Manuals. DocLayNet also provides original PDF pages, parsed PDF text, and text-cell coordinates.

PubTabNet [63] is a large-scale table recognition dataset containing over 568K tables from scientific publications.

It contains heterogeneous tables in both image and HTML format. To support more diverse model designs, it also provides the position (bounding box) of table cells.

FeTaQA [34] is a dataset consisting of 10K Wikipedia-based table question answering pairs along with tabular data extracted from webpages. These question answering pairs require complex reasoning and integration of table information. The tabular data supplied can be rendered into an HTML document, facilitating document understanding tasks. FeTaQA also provides annotations of highlighted cells, which are table regions corresponding to question answering pairs, and these annotations are used to construct the intermediate inference information for LayoutCoT.

A.2. Layout-aware Pre-training Data Construction

This section describes the data construction process for layout-aware pre-training tasks used to train LayoutLLM, including the preprocessing of document data, the prompt templates for prompting GPT (GPT-3.5 Turbo), and the instruction templates utilized during training.

As mentioned in Sec. 4.1, GPT (GPT-3.5 Turbo) is used to generate detailed descriptions of documents for the document-level pre-training task, Document Dense Description (DDD). Fig. 6 illustrates the data generation process for the DDD task and the instruction templates for pre-training DDD. As Fig. 6 (a) shows, the document is represented as layout text. Considering training efficiency and to prevent excessively long generated content, a requirement that the number of words generated should be less than 500 is in the prompt. During pre-training, the instructions that are shown in Fig. 6 (b) are randomly sampled as question instructions for DDD. Fig. 7 shows the instruction templates for the document-level pre-training task: the Text and Layout Reconstruction (TLR). The TLR task aims to reconstruct the complete text and layout information of a document and present it in a specific format, such as the prescribed “<{box}, {text} >” format, JSON format, or Markdown format.

For region-level pre-training tasks, pre-training is conducted by transforming document layout annotations and table annotations from the original dataset into instructions, as illustrated in the templates shown in Fig. 8.

The segment-level pre-training tasks are all self-supervised. Utilizing the original text and layout information, the input involves masking text or coordinates, or performing geometric-related calculations based on coordinates to obtain corresponding self-supervised targets. The

obtained self-supervised targets are transformed into instructions in Fig. 9 for pre-training.

A.3. Layout-aware SFT Data Construction

This section describes the process of constructing layout-aware SFT data. Fig. 10 presents an overall construction process of the layout-aware SFT data from Algorithm 1, including the document representation, QA&Text CoT Generation, LayoutCoT Generation, and Document Images Generation.

For the document representation, image documents (D_I) and text documents (D_M) are from publicly available document understanding datasets. As mentioned in Sec. 4.1, the HTML document (D_H in Algorithm 1) is from GPT’s free generation. Fig. 11 shows an example of the GPT’s free generation pipeline for D_H . To generate diverse HTML documents through GPT, some images and captions are randomly sampled from the LAION-5B dataset, which is a large-scale dataset consisting of more than 5B image-text pairs, as the topic and figure source of the generated D_H . A list of document types is provided for randomly sampling a document type as a constraint when prompting GPT. The detailed prompt is illustrated in Fig. 11. For the QA&Text CoT Generation, Fig. 12 shows the prompts and a specific example used for generating QA and the corresponding Text CoT (\mathcal{T}_c) based on document representation (\mathcal{R}) using GPT (GPT-3.5 Turbo). In addition, the \mathcal{D}_M includes the QA pairs and intermediate inference information process, thereby directly reusing the QA and organizing \mathcal{T}_c using a rule-based method. Then, as shown in the Fig. 10 (3), the step 1 & 3 in \mathcal{T}_c are used as the step 1 & 3 in \mathcal{L}_c . To construct the step 2 (relevant area concentration) of \mathcal{L}_c , the union bounding-box of all located relevant sentences in \mathcal{T}_c are taken as the relevant area. Finally, the text documents (D_M) and generated HTML documents (D_H) are converted to images.

B. Training Setup

The encoder weight of LayoutLLM is initialized from the LayoutLMv3-large [17] which is a widely-used document pre-trained model. And the LLM backbone is initialized from Vicuna-7B-v1.5 [61]. To further validate the adaptability of the LayoutLLM to different LLMs, llama2-7B-chat [50] is also employed, initializing the LayoutLLM’s LLM backbone for experiments. Other parameters are randomly initialized. d_0 is 1024 and d_1 is 4096. Following LayoutLMv3 [17], the maximum tokens of the document are set to 512 for training. The maximum instruction tokens are set to 2048 for training. The AdamW optimizer is used for both pre-training and SFT.

During pre-training, the LLM is frozen, and the parameters of the two projectors and document pre-trained model encoder are updated. Only 1 epoch is pre-trained. The

learning rate is set to $1e-4$ with a cosine scheduler. The weight decay is set to 0.0001. The warmup ratio is 0.03. The max grad norm is set to 1.0. The batch size is 32 for each GPU.

During SFT, both LLM and two projectors are fine-tuned while keeping the document pre-trained model encoder frozen. The SFT stage contains 3 epochs. The learning rate is set to $2e-5$ with a cosine scheduler. The weight decay is set to 0. The warmup ratio is 0.03. The max grad norm is set to 1.0. The batch size is 8 for each GPU.

C. Evaluation Setup

This section introduces the construction of test data for visual information extraction using a QA-based approach. To prompt LLMs/MLLMs for zero-shot VIE, annotations in VIE datasets are transformed into question answering format (QA for VIE).

As mentioned in Sec. 4.3, the FUNSD [19] dataset is a widely used VIE dataset that is annotated in entity-level headers, questions, answers, and others, along with entity linking annotations. As shown in Fig. 13, for question-answer annotations with linking in the FUNSD, the question answering pairs of evaluation are constructed by asking the values of the linking keys. To ensure an absolute answer of an evaluation question to avoid ambiguity, cases where one entity links to multiple entities in the FUNSD dataset are filtered out.

Different from the FUNSD dataset, the VIE annotations in the CORD [37] and SROIE [18] datasets can be directly transformed to {"entity type": "entity text"} as shown in Fig. 14. Therefore, for entity annotations in CORD and SROIE, the question&answer is constructed by directly asking the entity content. Similar to the filter rules conducted on the FUNSD dataset, cases where a specific entity type appears multiple times in the document are filtered out.

In order to avoid the influence of randomness in generation on the evaluation, sampling methods are not used for any of the LLMs/MLLMs during testing. The beam search with a beam size of 5 is employed for generation across all models.

CONT#	Dec 13, 19	3326990	Mod#	Ver# 1 (Last #)	DDS CONT#	0
REP	KATZ RADIO				C/P/E	/ / 8183
TO	WPEG-FM (Charlotte-Gastonia-Rock Hill, NC-SC)					
FM	JESSICA LAVORERIO,					
OFF	PHILADELPHIA				SALESPERSON FAX#	
AGY	Katz Media Group					
ADDR	125 West 6th Street 3rd Floor				PH #	202-965-5060
	New York, NY 10019					
BYR	Helen Hanratty					
ADV	TOM STEYER FOR PRESIDENT					
PDT	2020					
FLT	Dec 17, 19 - Dec 24, 19					

* REP ORDER COMMENT *

** 12/13/2019 1:34:00 PM: THIS IS A KATZ MEDIA GROUP ORDER. ALL BILLING SHOULD BE SENT TO KATZ MEDIA GROUP 125 W 5TH ST, NY, NY 10019. KATZ MEDIA GROUP IS NOT LIABLE FOR PAYMENT.

** 12/13/2019 1:34:00 PM: THIS IS A NEW ISSUROPOLITICAL ORDER. PLEASE NOTE THERE IS A 24 HOUR CANCELLATION POLICY ON ALL ISSUE/POLITICAL ORDERS. PLEASE CONFIRM IN THE SYSTEM.

STEPHANIE.DAVY@KATZMEDIA.COM 215-577-4231, THANK YOU!

MC	LN	DAYS	TIME	LEN	EFFECTIVE DATES	# OF WKS	NPW	RATE	TOT SPTS
		ELIGHT_1							
	1.1	TUWTFM	6A - 10A	60	12/17/2019 - 12/23/2019	1W	8	\$136.00	8
	1.2	TUWTFM	10A - 3P	60	12/17/2019 - 12/23/2019	1W	8	\$90.00	8
	1.3	TUWTFM	3P - 7P	60	12/17/2019 - 12/23/2019	1W	8	\$136.00	8
	1.4S.	6A - 7P	60	12/21/2019 - 12/21/2019	1W	4	\$60.00	4
	1.5S	6A - 7P	60	12/22/2019 - 12/22/2019	1W	4	\$60.00	4
		** WEEKLY FLIGHT TOTALS **							
						32		\$3,320.00	

Dec 19					
SPOTS	32				
CASH	3320.00				
TRADE	0.00				
NSL	0.00				
TOTAL	3320.00				

					TOTAL
SPOTS					32
CASH					3,320.00
TRADE					0.00
NSL					0.00
TOTAL					3,320.00

→ {document layout text}

Representing the document as layout text using the OCR / PDF parsing results



Prompt Example:
 You are asked to Describe the following document image in detail. The generated detailed description must be relevant to the text content of the document.

Document:
 {document layout text}

detailed description (Less than 500 words):

↓ Generating Document Dense Description

GPT(gpt-3.5-turbo) Answer: The document appears to be an order or invoice for advertising services. It is dated December 13, 2019, and includes various details about the order.

At the top of the document, there are several sections providing information about the contract number (CONT#) and modification number (Mod#). The current version number is also mentioned. Next to it is the DDS CONT# and it is listed as 0. Following that, there is a section labeled REP, which indicates the representative for the order is Katz Radio.

The next section is marked as C/P/E, which likely stands for Client/Program/Episode, with a value of 1/8183. This could signify the specific client, program, and episode for which the advertising is intended.

The "TO" section states that the advertising is intended for WPEG-FM, a radio station located in Chanotte-Gastonia-Rock Hill, NC-SC area.

Under the "FM" section, the name Jessica Lavorerio is mentioned, possibly as a contact person for the advertising order.

The "OFF" section indicates that the order is originating from Philadelphia. It also mentions a salesperson and a fax number.

The "AGY" section states that the order is placed through Katz Media Group, an advertising agency.

The "ADDR" section provides the address of Katz Media Group as 125 West 5th Street, 3rd Floor, New York, NY 10019. A phone number is also listed as 202-965-5060.

The order is placed by a person named Helen Hanratty, as mentioned in the "BYR" section.

The "ADV" section indicates that the advertisement is for Tom Steyer for President, most likely referring to a political campaign.

The "PDT" section states the year as 2020, suggesting that the advertising campaign is planned for that year.

The "FLT" section provides the flight dates for the advertising campaign, from December 17, 2019, to December 24, 2019.

Below this, there are a couple of order comments regarding billing and cancellation policies, which should be taken note of.

Next, there is a table that includes details about the ad spots that will be aired. The table includes information such as the days of the week, time slots, length of the ad spots, and the effective dates. It also mentions the number of spots per week (NPW), the rate per spot, and the total cost.

The document then provides a summary of the flight totals, indicating a total of 32 spots and a total cost of \$3,320.00.

Towards the end of the document, there is a breakdown of the payment, with \$3,320.00 paid in cash and no trade or NSL (Non-Standard Limitation) payments.

Overall, this document appears to be an order or invoice for advertising services placed by Katz Media Group on behalf of Tom Steyer's presidential campaign. It includes details about the flight dates, ad spots, and payment information.

(a) Process for generating Document Dense Description

Question Templates

Provide a detailed description of the given document image.
 Describe the following document image in detail.
 Characterize the document image using a well-detailed description.
 ...

(b) Question Templates of Document Dense Description Instruction

Figure 6. The data construction of Document Dense Description (DDD) task for document-level pre-training. (a) Given a document, generating its dense description as the instruction answer using its layout text with the help of GPT; (b) The shown question templates are randomly sampled as DDD instructions.

Input: Document Image + OCR / PDF Parsing Results

FROM: KATZ MEDIA GROUP
 TO: WFLA-TV (Charlotte-Orlando-Raleigh-Hill-NC-BC)
 FROM: KATZ MEDIA GROUP
 TO: WFLA-TV (Charlotte-Orlando-Raleigh-Hill-NC-BC)
 FROM: KATZ MEDIA GROUP
 TO: WFLA-TV (Charlotte-Orlando-Raleigh-Hill-NC-BC)

LN	DAYS	TIME	LEN	EFFECTIVE DATES	# OF WKS	NRW	RATE	TOT SPOTS
1.1	TUWTFM	6A-7A	60	12/17/2019 - 12/22/2019	1	1	\$150.00	1
1.2	TUWTFM	6A-7A	60	12/17/2019 - 12/22/2019	1	1	\$150.00	1
1.3	TUWTFM	6A-7A	60	12/17/2019 - 12/22/2019	1	1	\$150.00	1
1.4	TUWTFM	6A-7A	60	12/17/2019 - 12/22/2019	1	1	\$150.00	1
1.5	TUWTFM	6A-7A	60	12/17/2019 - 12/22/2019	1	1	\$150.00	1

SPOTS: 5
 GROSS: 750.00
 TRADE: 0.00
 NET: 750.00

OCR / PDF Parsing Results:

```

[
  {
    'box': [155, 92, 217, 111],
    'text': 'Dec 13, 19',
    'words': [...]
  },
  {
    'box': [156, 108, 353, 128],
    'text': '33524090 Mod# Ver# 1 (Last=)'
  },
  ...
]
  
```

Instruction Templates and Examples

Q: Can you detect and recognize the text in the document image and display the results using the format of <{position},{text}>?

It contains <{bbox},{text}>...

e.g., This contains <[40, 18, 57, 22],Dec 13, 19>, <[41, 21, 93, 26],33524090 Mod# Ver# 1 (Last=)>...

:A

Q: Please use markdown to describe the location and text content of the detected text in the document image.

| position | text | \n| ----- | ---- | \n| {{bbox}}|{{text}}| \n ...

e.g., | position | text | \n| ----- | ---- | \n|[40, 18, 57, 22] | Dec 13, 19 | \n|[41, 21, 93, 26] | 33524090 Mod# Ver# 1 (Last=) | \n...

:A

Q: Please use JSON formatting to describe the OCR result of the document image.

[{ "bbox" :{bbox}, "text" :{text}}...]

e.g., [{ "box" : [40, 18, 57, 22], "text" : "Dec 13, 19" }, { "box" : [41, 21, 93, 26], "text" : "33524090 Mod# Ver# 1 (Last=)" } ...]

:A

...

Figure 7. The data construction of Text and Layout Reconstruction (TLR) task for document-level pre-training. Given a document image with OCR or PDF parsing results, constructing TLR instructions by reconstructing the complete text and layout information of the document in a specific format.

74 The Open Ophthalmology Journal 2016, Volume 2

Table 1. Reduction in Progression and Change in Intraocular Pressure (IOP) in Relation to the Type of Orbital Decompression Procedure Carried Out

Decompression Procedure	Mean Volume of Fat Excised (ml)	Mean Reduction in Progression (mm)	Mean Preoperative IOP (mmHg)	Mean IOP at 1 Week after Surgery (mmHg)	Mean IOP at 3 Months after Surgery (mmHg)
1 will bone fat excision only (n=10)	2.0 ± 0.36	2.40 ± 0.8	14.56 ± 5.5	13.61 ± 4	13.76 ± 9
2 will bone fat excision (n=20)	1.5 ± 0.8	4.4 ± 1.1	18.26 ± 8	16.66 ± 0	17.16 ± 7
3 will bone fat excision (n=20)	3.6 ± 0.5	5.2 ± 2.4	17.66 ± 7	17.36 ± 7	17.46 ± 2

REFRACTIVE DATA
The majority of eyes (72%, 13 eyes) were myopic (spherical equivalent, range from -3.00 D to -6.25 D) and did not show a significant change following decompression. There were no significant differences found between preoperative and post-operative keratometry readings (K1, p=0.20; K2, p=0.06; keratometry axis (p=0.43), astigmatism reduction (A1, p=0.00; A2, p=0.70) or the axis of astigmatism reduction (p=0.42) at 3 months follow up (Table 2).

DISCUSSION
In this study we have evaluated the changes in IOP and refraction following orbital decompression. Unlike the previous studies we have used Topcon to measure the IOP as it is technically easier to ensure that central corneal IOP is recorded in up and down gaze in addition to primary gaze and is sufficiently consistent and accurate when compared to the Goldmann tonometer [9-11]. As our study group included patients undergoing intracanal fat excision only, as well as two and three wall bone decompression, we have also studied the IOP change in these subgroups individually. All our patients had a preoperative IOP of less than 21mmHg and did not demonstrate a significant reduction in IOP following orbital decompression. There was however a mean in this cohort towards less of a change in IOP following surgery when pre-operative IOP was not significantly raised in comparison to congestive cases with "tight orbits". We used the follow eye as a control to compare the changes in IOP at 1 week following orbital decompression, but could not find a significant difference in IOP between the operated and the unoperated eye. The mean amount of orbital fat resected by Robert et al. [6] was 6.64±3.6ml, whereas we had resected 1.76±0.6ml of orbital fat. The low volume of fat resected in

REDUCTION IN PROGRESSION
There was a significant reduction in mean progression from 23.46 ± 2.6mm preoperatively to 19.74 ± 1.2mm at 3 months postoperatively (p<0.001). The fat excision only group demonstrated a mean reduction in progression of 2.6 (3mm), two-wall bone + fat excision group achieved 4.1 (1mm) reduction and three-wall bone + fat excision group achieved 5.2 (2mm) reduction in progression at 3 months follow up.

IOP DATA
The preoperative IOP was less than 21mm in all our patients with an average of 17.62±2.2mmHg. Mean postoperative IOP at 1 week was 16.58 ± 0.9mmHg and the mean post-operative IOP at 3 months was 16.32±2.0mmHg. There was no statistically significant change in IOP between baseline and 1 week (p=0.2) and baseline (p=0.3) postoperatively, respectively. A reduction in IOP was observed in 8 eyes, no change in IOP was found in 8 eyes and an increase in IOP from baseline was observed in 4 eyes at 3 months post-operatively. The increase in IOP recorded in 4 eyes post-operatively was not clinically significant (1 mmHg increase in IOP recorded in 3 eyes and 2 mmHg increase in 1 eye).
The mean IOP in the contra lateral unoperated eye was 15.14 ± 1.3mmHg and there was no change at 1 week (15.12±2.2mmHg, p=0.1) postoperatively. There was no significant difference in IOP between the operated and the contra lateral unoperated eye at 1 week (p=0.1) postoperatively.

Table 2. Comparison of Pre and Post Operative Astigmatism and Keratometry

	AR 1	AR 2	AR 3	K1	K2	Axis
Median (range) Preoperative	0.25 (0.15-0.35)	0.30 (0.20-0.40)	0.35 (0.25-0.45)	-0.12 (0.00-0.24)	-0.12 (0.00-0.24)	91.5 (84.0-99.0)
Median (range) Postoperative	0.25 (0.15-0.35)	0.30 (0.20-0.40)	0.35 (0.25-0.45)	-0.12 (0.00-0.24)	-0.12 (0.00-0.24)	91.5 (84.0-99.0)
P/Wilcoxon Signed Rank Test	p=0.99	p=0.97	p=0.42	p=0.20	p=0.06	p=0.63

Instruction Templates and Examples

Q: Where can I find the {layout} in this document?
e.g., Where can I find the table in this document?

A: There are {layout_number} {layout} in the document, located at {layout_locations}.
e.g., There are 2 tables in the document, located at [107, 111, 941, 272] and [106, 788, 941, 925].

Q: What layout type does the position at {layout_location} in this document correspond to? {layout_locations_list}.
e.g., What layout type does the position at [107, 111, 941, 272] in this document correspond to? (title, list, table, figure, formula, page-footer, page-header, etc.)

A: {layout}.
e.g., Table.

...

(a) Instruction Templates & Examples for Document Layout Analysis

Age (years)	Number	SBP (mmHg)	DBP (mmHg)
18-30	811	121.7 ± 3.1	78.6 ± 2.4
31-45	337	127.7 ± 3.4	84.2 ± 2.5
46-60	480	136.1 ± 3.7	90.1 ± 3.4
61-75	290	147.2 ± 4.6	92.3 ± 3.8
76-90	56	150 ± 4.4	94 ± 3.4
91-105	26	152 ± 5.7	100 ± 4.8
Overall	2000	129.6 ± 3.5	84.3 ± 3.4

Instruction Templates and Examples

Q: How many rows and columns does the table have?
The table has {row_number} rows and {column_number} columns.
e.g., The table has 8 rows and 4 columns.

A:

Q: Can you list the values in the {row_or_col_number}? Separate multiple values with ";".
e.g., Can you list the values in the column 2? Separate multiple values with ";".

A: {value_list}
e.g., Number;811;337;480;290;56;26;2000.

Q: What is the cell value in row {row_number} and column {column_number}?
e.g., What is the cell value in column 1 and row 4?

A: {value}
e.g., 46-60.

Q: Can you list the values in the "{column_name}" column?
e.g., Can you list the values in the "Age(years)" column?

A: {value_list}
e.g., Age(years);18-30;31-45;46-60;61-75;76-90;91-105;Overall.

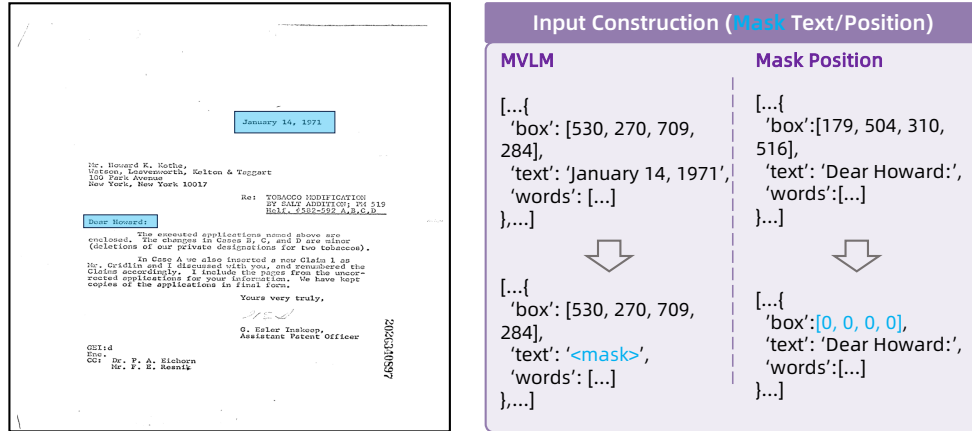
Q: What is the value of cell where column is "{column_name}" and row number is {row_number}?
e.g., What is the value of cell where column is "Number" and row number is 2?

A: {value}
e.g., 811.

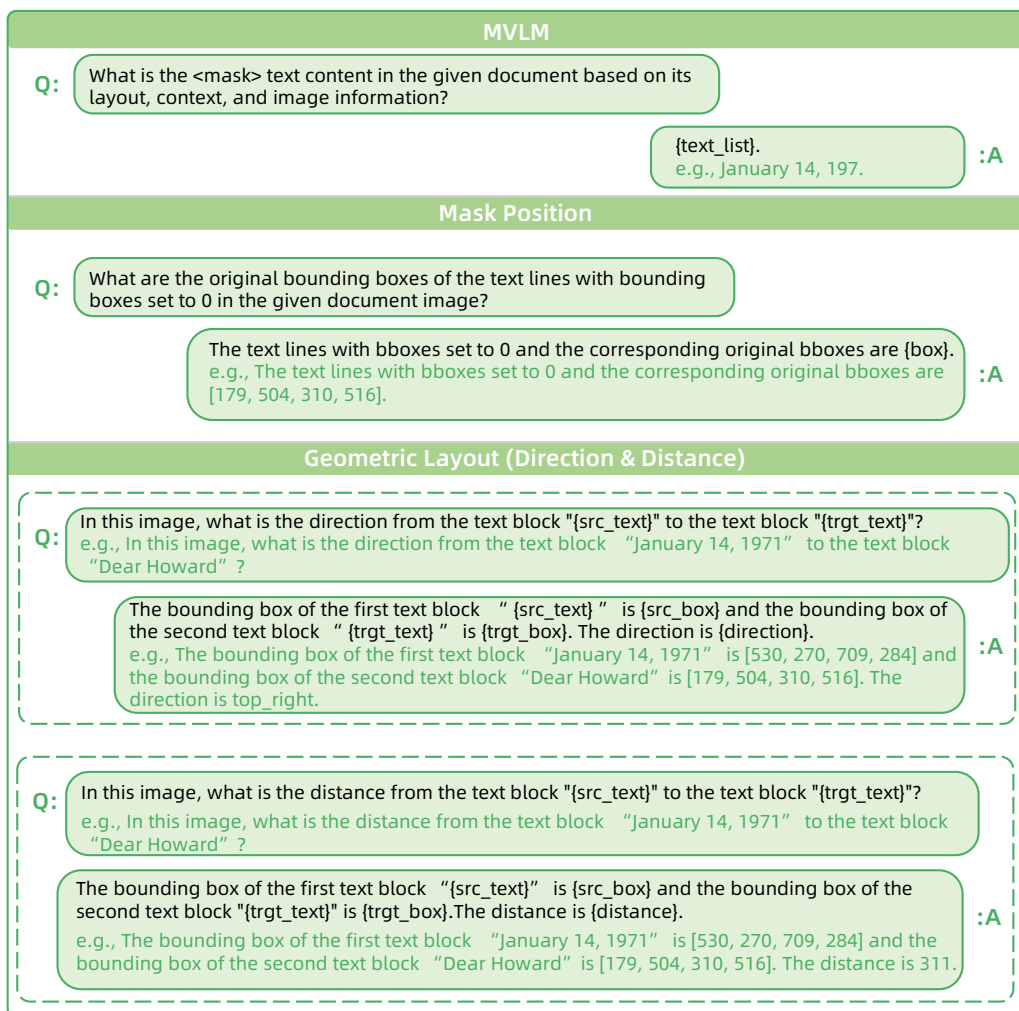
...

(b) Instruction Templates & Examples for Table Understanding

Figure 8. The data construction of Document Layout Analysis task and Table Understanding task for region-level pre-training. (a) Constructing Document Layout Analysis task instructions by transforming layout annotations to the presented templates; (b) Constructing Table Understanding task instructions by transforming table annotations to the presented templates.

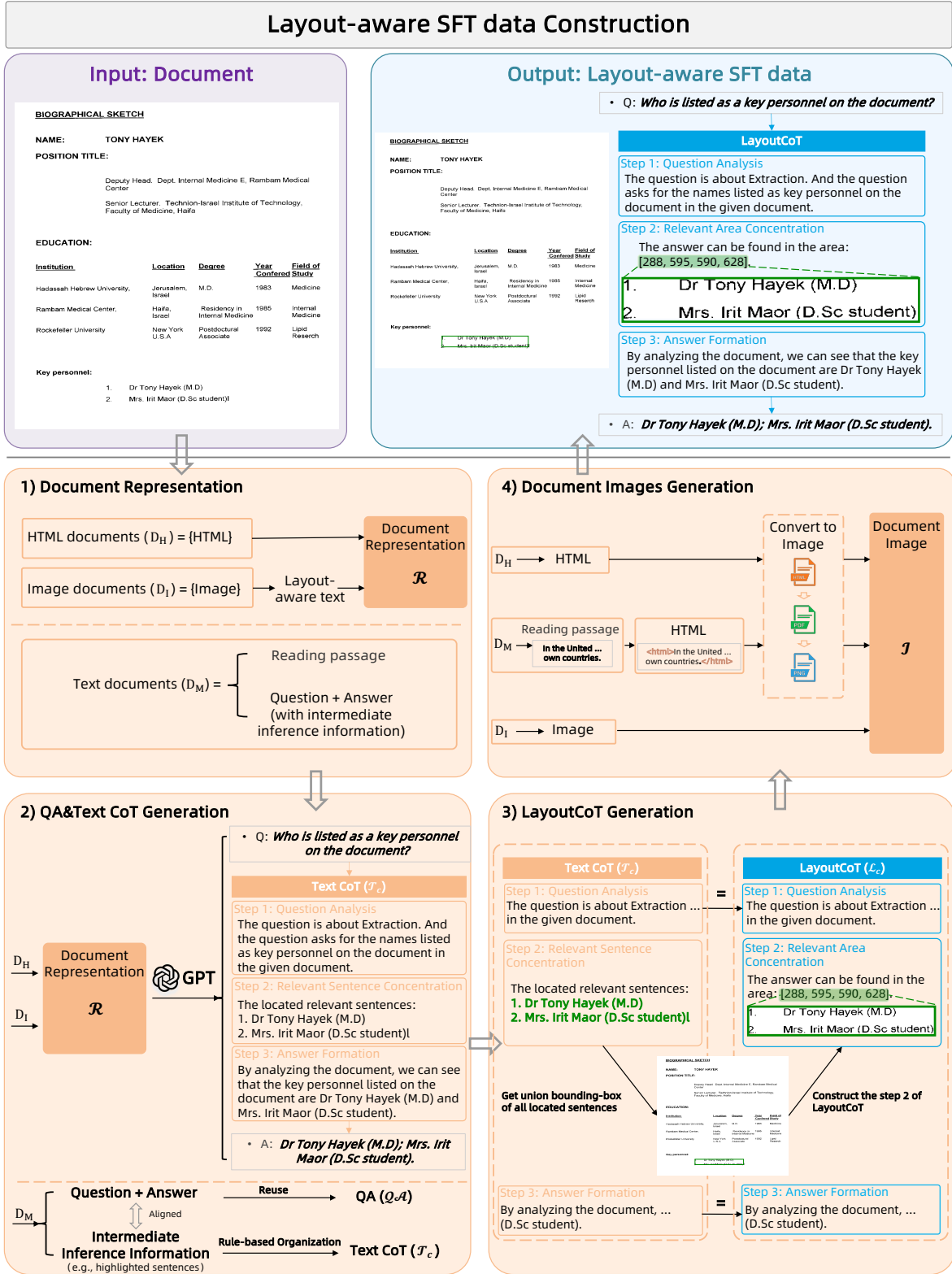


(a) The Input Construction of MVLM & Mask Position tasks



(b) Instruction Templates & Examples for Segment-level Pre-training Tasks

Figure 9. The data construction of MVLM instruction, Mask Position instruction and Geometric Layout instruction tasks for segment-level self-supervised pre-training. (a) For the MVLM and Mask Position instructions, randomly **masking** the text or position coordinates respectively as the inputs for LayoutLLM’s document pre-trained model encoder. The masked information is utilized as the target for self-supervised learning. (b) For the three segment-level pre-training tasks, constructing the instructions for text masking, position masking, and geometric-related tasks based on the presented templates respectively.



1) Document Representation

HTML documents (D_H) = {HTML} → Document Representation

Image documents (D_I) = {Image} → Layout-aware text → Document Representation \mathcal{R}

Text documents (D_M) = { Reading passage, Question + Answer (with intermediate inference information) }

4) Document Images Generation

D_H → HTML → Convert to Image → Document Image

D_M → Reading passage → HTML → Convert to Image → Document Image

D_I → Image → Document Image

2) QA&Text CoT Generation

D_H, D_I → Document Representation \mathcal{R} → GPT → Text CoT (\mathcal{T}_c)

Q: *Who is listed as a key personnel on the document?*

Text CoT (\mathcal{T}_c)

Step 1: Question Analysis
The question is about Extraction. And the question asks for the names listed as key personnel on the document in the given document.

Step 2: Relevant Sentence Concentration
The located relevant sentences:
1. Dr Tony Hayek (M.D)
2. Mrs. Irit Maor (D.Sc student)

Step 3: Answer Formation
By analyzing the document, we can see that the key personnel listed on the document are Dr Tony Hayek (M.D) and Mrs. Irit Maor (D.Sc student).

A: *Dr Tony Hayek (M.D); Mrs. Irit Maor (D.Sc student).*

Question + Answer → Reuse → QA (\mathcal{Q}_A)

Intermediate Inference Information → Rule-based Organization → Text CoT (\mathcal{T}_c)

3) LayoutCoT Generation

Text CoT (\mathcal{T}_c) → LayoutCoT (\mathcal{L}_c)

Text CoT (\mathcal{T}_c)

Step 1: Question Analysis
The question is about Extraction ... in the given document.

Step 2: Relevant Sentence Concentration
The located relevant sentences:
1. Dr Tony Hayek (M.D)
2. Mrs. Irit Maor (D.Sc student)

LayoutCoT (\mathcal{L}_c)

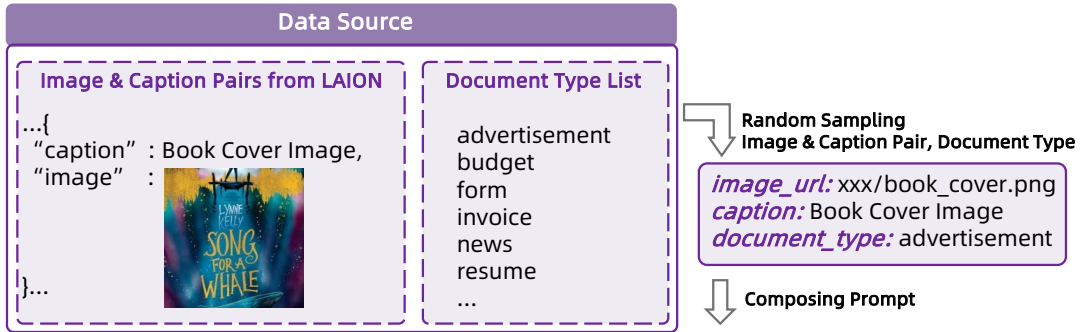
Step 1: Question Analysis
The question is about Extraction ... in the given document.

Step 2: Relevant Area Concentration
The answer can be found in the area:
[288, 595, 590, 628]

Step 3: Answer Formation
By analyzing the document, ... (D.Sc student).

Get union bounding-box of all located sentences → Construct the step 2 of LayoutCoT

Figure 10. Overview of Layout-aware SFT data construction, involving 4 steps: Document Representation, QA & Text CoT Generation, LayoutCoT Generation, and Document Images Generation.



I want you to act as an html programmer using English. I give you the following topic (*caption*), please create a document-like html with some complex tables and a lot of rich paragraphs related to the above topics. And in order to help you make this html better, the html you make also needs to meet the following requests:

Request 1:
The html should add the following all image links(*image_url*) and add alt text to describe the image. Using CSS, you can adjust the size and position of the image to make it fit well with the overall layout.

Request 2:
Please focused on the following topic (*caption*), using CSS styling to adjust the tables in the html that do not have any borderlines but only the header cell (the first row) has a horizontal line, and each cell of the tables is wider than it is tall.

Request 3:
You can layout and beautify the html page using separators, serial numbers, indentation, line breaks, bullet points, etc. And you can add checkboxes and radio buttons that related to the topic (*caption*) to enhance the layout.

Request 4:
You can use CSS styling to change the color, font, and adding borders and shading, etc. And the design will be professional, visually appealing, easy to read and scientifically rigorous, with a clear hierarchy of information and a consistent color scheme.

Final Request:
Overall, the themes of the html should be focused on the following topic: (*caption*), and based on the above topics you should give a appropriate title for the html. You should try your best to make an html (like a "*document_type*"-type document) that meets all the requests mentioned above.




```
<!DOCTYPE html>
<html>
<head>
<title>
  Blog Tour: Song for a Whale – Author Interview
</title>
<style>
  body {
    font-family: Arial, sans-serif;
    background-color: #f2f2f2;
    margin: 20px;
  }
  ...
  ...
```

Rendering Image



Advertisement

Figure 1:



Author Name	Lynne Kelly
Publication Date	February 5, 2019
Genre	Children's Fiction
Publisher	Deborah Books for Young Readers

I recently had the pleasure of interviewing Lynne Kelly, the author of the captivating book "Song for a Whale". The novel tells the story of a young deaf girl who finds a deep connection to the sea as a whale she sings to a language no other whale can hear. It's a heartwarming tale of empathy, communication, and the power of music.

Dear Readers, Lynne shares her inspiration behind the book and her thoughts on the importance of inclusivity in children's literature. She emphasizes the significance of representing characters with diverse backgrounds and abilities, as it allows readers to explore new perspectives and develop empathy.

When asked about her research process, Lynne revealed that she spent a significant amount of time learning about whale songs and marine biology. Her dedication to accuracy shines through in the novel, creating a scientifically rigorous and educational experience for readers.

Key Takeaways from the Interview

- Representation matters: Inclusive stories help children embrace differences and foster empathy.
- Music as a universal language: The power of communication through music transcends barriers.
- Accurate research: The importance of scientific accuracy when incorporating real-world elements into fiction.

Would you like to know more about Song for a Whale?

- Visit the [official website](#) for additional information.
- Check out her [Goodreads page](#) for reviews and ratings.
- Join the discussion on social media using the hashtag #SongforaWhale.

Reader Feedback

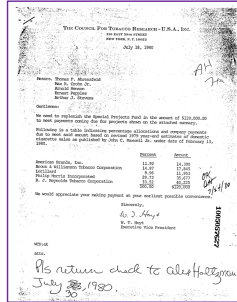
What are your thoughts on the book?

Positive
 Negative

Would you recommend "Song for a Whale" to others?

Yes
 No

Figure 11. An example of the HTML documents D_H generation pipeline using GPT. To generate HTML documents with diverse layout types by GPT, the data source (including image & caption pairs from LAION and document type list) are randomly sampled as the inputs for composing the prompts of D_H generation.



You are asked to generate 10 question & answer pairs on a document image. The generated question must be relevant to the text content of the document. The answer to the question is better to be (but not limited to) the short text spans taken verbatim from the document. This means that the answers comprise a set of contiguous text tokens present in the document.

Document:
`[document layout text / HTML]`

The generated question & answer pairs should be in this format:

```
###SEP
Q: what is the XXX Date ?
A: XXXXX
###SEP
Q: How many XXX ?
A: XX
10 question & answer pairs, please generate:
```



```
###SEP
Q&A 1:
Q: Who is the Executive Vice President?
A: W. T. Hoyt.
###SEP
Q&A 2:
Q: How is the $120,000.00 allocated among the various tobacco companies?
A: The allocation of the $120,000.00 among the various tobacco companies is as follows: - American Brands, Inc.: 11.92% (amount: $14,300) - Brown & Williamson Tobacco Corporation: 14.87% (amount: $17,845) - Lorillard: 9.96% (amount: $11,953) - Philip Morris Incorporated: 29.73% (amount: $35,677) - R. J. Reynolds Tobacco Corporation: 33.52% (amount: $40,225)
...
```

You are tasked with generating Chain-of-thought (CoT) based on an document and a question and answer pair related to the document.

Document:
`[document layout text / HTML]`

10 Q&A pairs:

```
Q&A 1:
Q: Who is the Executive Vice President?
A: W. T. Hoyt.
Q&A 2:
Q: How is the $120,000.00 allocated among the various tobacco companies?
A: The allocation of the $120,000.00 among the various tobacco companies is as follows: - American Brands, Inc.: 11.92% (amount: $14,300) - Brown & Williamson Tobacco Corporation: ...
...
```

The generated CoT should include 3 steps:

Step 1 (question analysis): Analyze the question type (e.g., "table", "counting", "reasoning", "key-value extraction", "layout", "title", etc.), and determine what information is being asked for.

Step 2 (find all the relevant sentences): Find all the relevant sentences from the input document that can inference the answer to the question. Every sentence must be the contiguous original text tokens present in the document.

Step 3 (answer analysis): Provide a clear and comprehensive explanation of the logical process used to obtain the answer.

Notice! The generated question & answer pairs should be in this format:

```
###SEP
Q&A1:
Step 1 (question analysis): question type: XXX. The question is asked for XXX.
Step 2 (find all the relevant sentences): All the relevant sentences are:
1. xxxx
2. xxxx
...
Step 3 (answer analysis): XXX.
```

Please generate corresponding Chain-of-thought (CoT) based on the document separately for the each Q&A (10 Q&A pairs) mentioned above:



```
Q&A1:
Step 1 (question analysis): question type: key-value extraction. The question is asking for the Executive Vice President's name.
Step 2 (find all the relevant sentences): All the relevant sentences are:
1. W. T. Hoyt
2. Executive Vice President
Step 3 (answer analysis): The Executive Vice President mentioned in the letter is W. T. Hoyt.
###SEP
Q&A2:
Step 1 (question analysis): question type: table. The question is asking for the allocation of the $120,000.00 among the various tobacco companies.
Step 2 (find all the relevant sentences): All the relevant sentences are:
1. Following is a table indicating percentage allocations...
...
n. R. J. Reynolds Tobacco Corporation 33.52 40,225
Step 3 (answer analysis): According to the table in the document, the allocation of the $120,000.00 among the various tobacco companies is as follows:
- American Brands, Inc.: 11.92% (amount: $14,300)
- Brown & Williamson Tobacco Corporation: 14.87% (amount: $17,845)
- Lorillard: 9.96% (amount: $11,953)
- Philip Morris Incorporated: 29.73% (amount: $35,677)
- R. J. Reynolds Tobacco Corporation: 33.52% (amount: $40,225)
...
```

Figure 12. An example of QA and corresponding Text CoT (\mathcal{T}_c) generation. Given a document representation (layout text or HTML), generating its QA & Text CoT with the help of GPT.

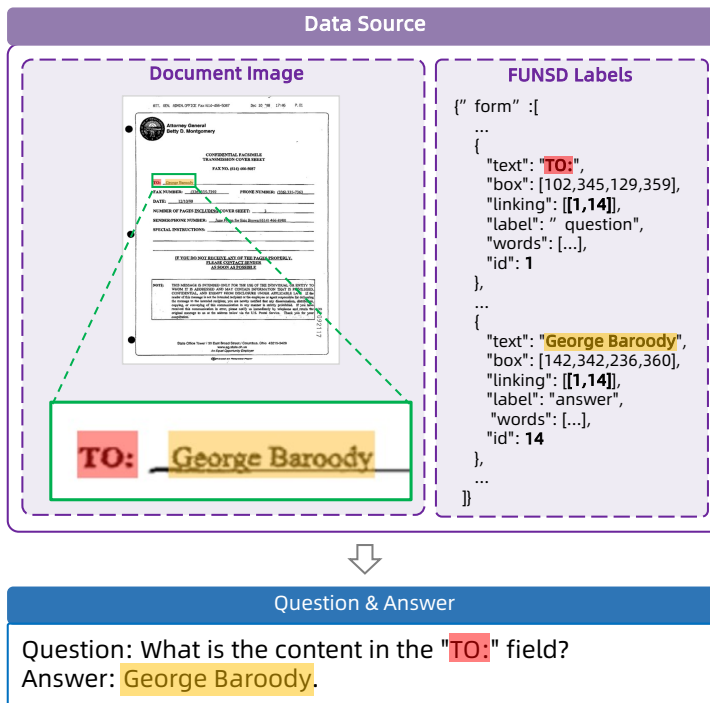


Figure 13. Evaluation data construction example of the QA for VIE task through question-answer with linking annotations in VIE (FUNSD).

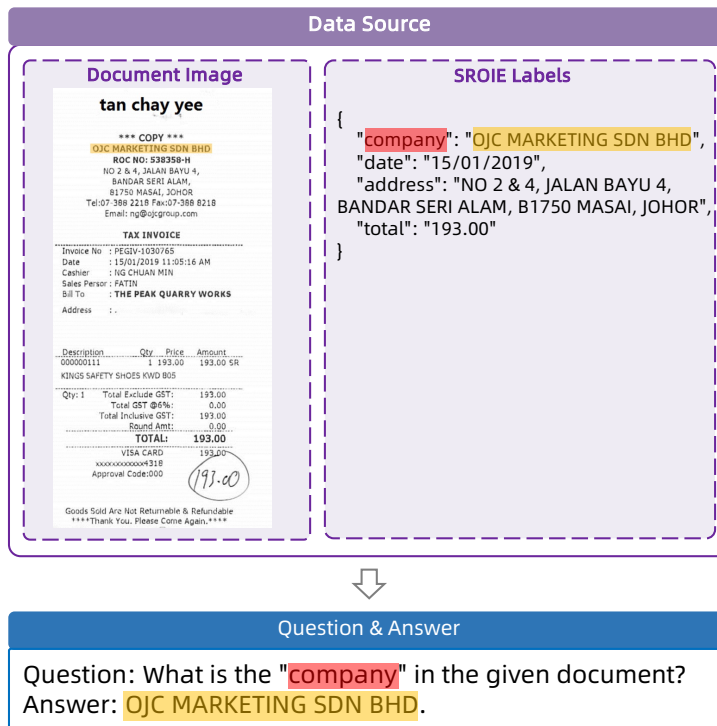


Figure 14. Evaluation data construction example of the QA for VIE task through entity annotations in VIE (SROIE&CORD).