

A Question Templates with Examples

All question templates for three tasks are listed in Table 1, 2, 3 with the corresponding real question examples. Task A contains 36 question patterns, including 22 Existence type question patterns and 14 Counting type question patterns. For Task B, the Structural Understanding type contains 10 question patterns, and Object Recognition contains 5. Regarding Task C, there are 5 patterns provided for Child Relation Understanding and 10 patterns designed for Parent Relation Understanding, respectively.

Question Pattern	Question Example
Existence Type Question Patterns	
Is there any [E] on the [pos] of this page?	Is there any table on the top of this page?
Can you find any [E] on the [pos] of this page?	Can you find any figure on the right of this page?
On the [pos] of this page, is there a [E]?	On the left of this page, is there a table?
Is it correct that there is no [E] at the [pos]?	Is it correct that there is no figure at the bottom?
When you check the [pos] of this page, can you find any [E]?	When you check the right of this page, can you find any table?
Are there any [E1] are [R] the [E2]?	Are there any figures upper the 'Competition analysis'?
Can you find any [E1] [R] the [E2]?	Can you find any table above the 'Balanced networks'?
Is there a [E1] found [R] the [E2]?	Is there a table found under the 'Competition analysis'?
Is it correct that there is no [E1] [R] the [E2]?	Is it correct that there is no table upper the 'Discussion'?
Confirm if there are any [E1] [R] the [E2]?	Confirm if there are any figures upper the 'Result and Discussion'?
When you check the page, is there any [E1] [R] the [E2]?	When you check the page, is there any table below the 'Results'?
Is there any [E]?	Is there any table?
Are there any [E] on this page?	Are there any figures in this page?
Is there a [E] in this page?	Is there a table on this page?
Can you find a [E] on this page?	Can you find a figure on this page?
When you check this page, can you find any [E]?	When you check this page, can you find any table?
Is there a [E] on this page?	Is there a 'Results' on this page?
Can you find a [E] on this page?	Can you find a 'Discussion' on this page?
Does this page include a [E]?	Does this page include a 'Conclusion'?
Can [E] be found on this page?	Can 'Abstract' be found on this page?
When you check this page, can you find [E]?	When you check this page, can you find 'Introduction'?
Confirm if there is [E] on this page.	Confirm if there is an 'Abstract' on this page.
Counting Type Question Patterns	
How many [E1] are [R] the [E2]?	How many tables are left for the 'Result and Discussion'?
What is the number of [E1] [R] the [E2]?	What is the number of tables below the 'Background & Summary'?
How many [E1] can you find on the [R] of [E2]?	How many figures are upper the 'Discussion'?
Count the number of [E1] on the [R] of [E2].	Count the number of figures below 'Material and methods'.
When you check this page, how many [E1] can you find on the [R] of [E2]?	When you check this page, how many tables can you find on the top of 'Background'?
Can you find [num] [E](s) on the page?	Can you find 2 table(s) in the page?
Does this page include [num] [E](s)	Does this page include 2 figures?
Confirm if there are [num] [E](s) on this page.	Confirm if there are 1 table(s) in this page.
Are there [num] [E](s) on this page?	Are there 3 figure(s) in this page?
Is there only [num] [E](s) on this page?	Is there only 2 table(s) in this page?
How many [E]s on this page?	How many tables in this page?
When you check this page, how many [E]s are on this page?	When you check this page, how many tables are on this page?
What is the number of [E]s on this page?	What is the number of figures on this page?
How many [E]s can be found on this page?	How many figures can be found on this page?

Table 1. Task A question pattern templates with corresponding example questions.

B Human Evaluation Details

We randomly selected 30, 30 and 40 question-answer pairs from Task A, Task B and Task C, respectively and put them with the related document page images or file links in the google forms (An example of Task C can refer to Figure 1). For each task, raters need to check each generated question-answer

Question Pattern	Question Example
Structural Understanding	
What is the [turn] section in this page?	What is the last section in this page?
Can you describe the [turn] section of this page?	Can you describe the first section of this page?
What does the [turn] section include in this page?	What does the last section include in this page?
What is the main contents of the [turn] section in this page?	What is the main contents of the first section in this page?
When you check the [turn] section of this page, what information can you get?	When you check the last section of this page, what information can you get?
What is the [pos] section about?	What is the top section about?
What is the [pos] of the page about?	What is the left of the page about?
What is the topic of [pos] section?	What is the topic of bottom section?
Can you describe the main topic of the [pos] section?	Can you describe the main topic of the right section?
When you check the [pos] of this page, what information can you get?	When you check the bottom of this page, what information can you get?
Object Recognition	
What is the [E] on the [pos] of the page?	What is the table on the top of the page?
What is the [pos] [E] about?	What is the bottom table about?
Can you describe the [E] on the [pos] of the page?	Can you describe the figure on the bottom of the page?
What information does the [pos] [E] contain?	What information does the left figure contain?
When you check the [pos] [E], what information can you get?	When you check the top table, what information can you get?

Table 2. Task B question pattern templates with corresponding example questions.

Question Pattern	Question Example
Child Relation Understanding	
What does the [E] include?	What does the Introduction include?
What is the [E] about?	What is the Competing interests about?
What subsections are in the [E]?	What subsections are in the 2. Clinical Presentation?
What subsections can be found in the [E]?	What subsections can be found in the Materials and methods?
When you check the [E], which subsections are included?	When you check the Methods, which subsections are included?
Parent Relation Understanding	
Which section does describe the [E] ?	Which section does describe the Table 3?
Which section does include the description of the [E]?	Which section does include the description of the Table 2?
Name out the section that include the [E].	Name out the section that include the Table 2.
Where can you find the [E]?	Where can you find the Table 2?
When you search for the description of [E], which sections do you need to check?	When you search for the description of Figure 1, which sections do you need to check?
Which section does include the [E]?	Which section does include the 'Corwin HL et al,2009'?
Which section does cite the [E]?	Which section does cite the 'Wang C et al,2017'?
Where is the [E] cited in the document?	Where is the 'Horner KC et al,2005' cited in the document?
Where can [E] be found in the document?	Where can 'Guan KL et al,1991' be found in the document?
When you search for the citation of [E], which sections can you find it?	When you search for the citation of 'Zhang Z et al,2013', which sections can you find it?

Table 3. Task C question pattern templates with corresponding example questions.

pair together with the attached document page or file to determine whether the question-answer pairs meet the requirements of three aspects, *Relevance*, *Correctness*, *Meaningfulness*. For example, for a given question in Figure 1, "Name out the section that describes Figure 1", raters need to first go through the entire document to check whether the document has Figure 1 and then check which sections provide the description of that figure to compare with the provided answer. Finally, raters are required to determine whether this question will be asked in the real world. We show the detailed definition of each aspect to ensure raters can understand the evaluation metrics of each criterion at the beginning of the questionnaire of each task, as Figure 1 shows.

Evaluation Criteria

PDFVQA Dataset Quality Manually Evaluation Task C

This questionnaire evaluates the quality of automatic question-answer pairs from the PDFVQA dataset. Please select the corresponding levels for each question-answer pair to evaluate the quality of each pair. We will evaluate the question-answer based on the following aspects.

- **Question Relevance:** Do you think the question is directly related to the corresponding document?
 1. Not related to the corresponding document at all
 2. Related to the document
- **Answer Correctness:** Do you think the proper information is extracted from the document as the answer?
 1. Correctly extracted
 2. Incorrectly extracted
- **Question-Answer Pair Meaningfulness:** Do you think the question would be asked in the real world?
 1. Meaningless QA pair (e.g. too general and no one will ask in the real world)
 2. Meaningful and interesting question
- **Question-Answer Coverage:** Do you think that it is necessary to understand the cross-page information in order to answer the question?
 1. Just a single page is enough
 2. Cross-page understanding

Survey Sample

Task C Question 1

Q: Name out the section that describe the Figure 1.
A: Discussion

Please get into this link to check the PDF Files
[PDF File of this Question-Answer Pair](#)

Question Relevance: Do you think the question is directly related to the corresponding document? *

☐ 1. Not related to the corresponding document

☐ 2. Highly related to the corresponding document

Answer-Relevance: Do you think the proper information is extracted from the document as the answer? *

☐ 1. Correctly extracted

☐ 2. Incorrectly extracted

Question-Answer Pair Meaningfulness: Do you think the question would be asked in the real world? *

☐ 1. Meaningless QA pair (e.g. too general and no one will ask in the real world)

☐ 2. Meaningful question (e.g. questions might be asked in the real-world)

Question-Answer Coverage: Do you think that it is necessary to understand the cross-page information in order to answer the question? *

☐ 1. Just a single page is enough

☐ 2. Cross-page understanding (actual document understanding is required)

URL to PDF File/Image




Fig. 1. A human evaluation sample with evaluation criteria of Task C. Task A and B have a similar style as Task C.

C Additional Dataset Analysis

C.1 Distribution of Question Length

We show the distribution of question length of each task in Figure 2. The average question length for Task A, B and C are 25, 10 and 15, respectively.

D Baseline Details

D.1 Baseline Descriptions

- **M4C** [2] applies the multimodal transformer, which takes into the question embedding, OCR token embedding and image object features as inputs, and iteratively decodes the answers over the combined answer space of OCR tokens and the fixed answer list.
- **VisualBERT** [4] is a pretrained vision-and-language model that passes the sequence of text and object region embeddings to a transformer to get the integrated vision-and-language representations.
- **LXMERT** [5] applies three transformer encoders to encode the text embeddings, object region embeddings and the cross-modality learning between texts and image features.
- **ViLT** [3] operates linear projection over image patches to get a sequence of image patch representations and input to the transformer encoder together with the text embeddings to get a pretrained vision-and-language model.

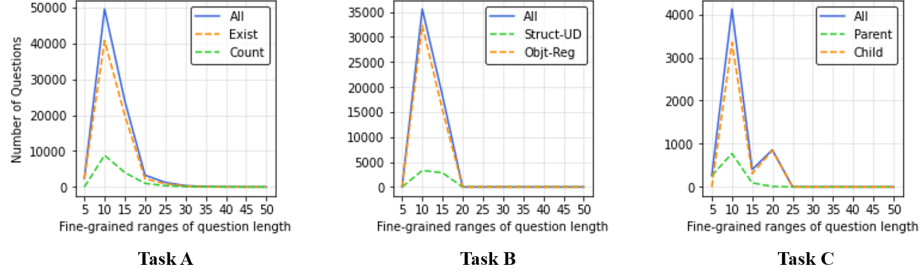


Fig. 2. Question length distribution of Task A, B and C

- **BERT** [1] is a pretrained language model that applies the structure of a multi-layer bidirectional transformer encoder. We used only the textual features from document pages as the inputs.
- **LayoutLMv2** [6] is a pre-trained model to operate on the position and textual features of document elements and generate the integrated representations that can be used for downstream document-related tasks.

D.2 Baseline Setup

- **M4C** applied multiple transformer layers, learning question embeddings, image object features and OCR token features in the common embedding space and iteratively decoding answer tokens from a fixed answer space or the OCR tokens in the image. The OCR tokens are encoded in rich representations, including the textual embedding of each token, appearance features of the token region on the image, Pyramidal Histogram of Characters (PHOC) features and the location features. We evaluated all three tasks with M4C but slightly modified the inputs and output layer to suit document-based VQA. Firstly, since the number of OCR tokens is much larger in PDF documents than that in real-life scenes, instead of inputting the features of all OCR tokens in the page, we used the BERT [CLS] token features to represent the sequence of textural contents in each document element region and took them together with the question embedding and the visual features of each document element region as the input sequence to the multi-layer transformer. Secondly, in the decoding part, Task B and C, we used the d -dimensional representations for the index numbers of the corresponding document element region in the page and generated the scores through the dynamic pointer network to predict the index number of document element region over the list of document element region index numbers. For applying M4C to Task A, we set fixed answer space as the decode inputs and put the pointer network on top to get a final prediction.
- **BERT**, **LayoutLM2** are used only for Task A and B because the inputs of both models are question and context token level information with the 512 maximum limitations. For multi-page documents, the number of tokens

is normally much higher than 512 tokens, which means those two models can only catch the first-page context information. In this case, we did not select those two models for conducting Task C tests. For both Task A and B, we directly extract 768-dimension [CLS] token embedding and feed it into classifiers for predicting the corresponding answer or object sequential index.

- **VisualBERT**, **LXMERT** can process visual features of document layout elements extracted from pretrained ResNet101-Res5. After we feed those raw object-level visual features and question tokens into those vision-language pretrained models, we extract the enhanced visual representation of document layout elements and feed them into a pointer network to get final scores for predicting corresponding answers for all three tasks.
- **ViLT** is directly applied for conducting Task A and B by using the provided feature extractor and pre-trained
- **ViltForQuestionAnswering** model to predict the corresponding answer based on input questions and image patch features. For addressing task C, we concatenate all document pages into an image pixel matrix and feed into the feature extractor to extract image patch features for feeding forward pass. The outputs pass through a Sigmoid layer instead of the Softmax function adopted by other tasks for backward propagation in the training stage and answer prediction in the inference stage.

E Implementation Detail

Dimension for the visual features of each document element region d_f is 2048. The activation function used in GCN is Tanh. The GCN is trained with AdamW optimizer and 0.0001 learning rate for 10 epochs. Each question token is encoded into a 768-dimension fine-tuned on the BERT-base model. Our model utilized a 6 layers transformer encoder and a 4 layers transformer decoder with 12 heads and 768-dimension hidden size. The maximum numbers for input question tokens and objects (document layout elements) are 50 and 25, respectively, for Task A and B and 50 and 400 for Task C. For a fair comparison, epoch times are selected as 5, 10, and 20 for all Task A, B and C models, respectively. All the experiments are conducted on 51 GB Tesla V100-SXM2 with CUDA 11.2.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
2. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9992–10002 (2020)

3. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
4. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
5. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5100–5111 (2019)
6. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2579–2591 (2021)