A. Screen grabs of Annotation Tool

As mentioned in Section 3.1 in the main paper, annotation process involves three stages. In Figure A.1, Figure A.2 and Figure A.3 we show screen grabs from stage 1, stage 2 and stage 3 of the annotation process respectively.

B. Examples of Question Types

We define 9 question types, based on the kind of reasoning required to answer a question. Question types are assigned at the second stage of the annotation. We discuss the question types in Section 3.2. in the main paper.

Examples for types form, yes/no and layout are shown in Figure B.1. Examples for a question based on a handwritten date in a form (types form and handwritten) are shown in Figure B.2. An example for a question based on information in the form of sentences or paragraphs (type running text) is shown in Figure B.3. Examples for types photograph and table are shown in Figure B.4. An example for a question based on a plot (type figure) is shown in Figure B.5. In all examples a crop of the original image is shown below the original image, for better viewing of the image region where the question is based on.

C. Additional Qualitative Examples

Here we show more qualitative results from our baseline experiments. These results supplement the Results section (Section 5.3) in the main paper.

Remember that BERT [2] question answering model is designed to answer questions asked on sentences or paragraphs of electronic text (reading comprehension). In Figure C.1 we show two examples where the model answers questions outside the scope of reading comprehension style question answering. In Figure C.2 we show examples where the M4C [3] model outperforms the BERT model to answer questions based on text seen on pictures or photographs. Such questions are similar to questions in TextVQA [4] or ST-VQA [1] datasets where M4C model yield state-of-the-art results. In Figure C.3 we show an example where both the models yield inconsistent results when posed with questions of similar nature, highlighting lack of reasoning behind answering. In Figure C.4 we show two examples where both the M4C and BERT model fail to answer questions which require understanding of a figure or a diagram. In Figure C.5 we show how OCR errors have resulted in wrong answers although the models manage to ground the questions correctly.

References

Figure A.1: Annotation stage 1 - Question Answer Collection: Questions and answers are collected for a given document image. Annotator can add up to 10 questions for a document. The document can be skipped if it is not possible to frame questions on it.

Figure A.2: Annotation stage 2 - Data Verification: For each question shown annotators have to (i) enter answer(s) (answer(s) from first stage are not shown) and (ii) Tag the question with one or more question types from the 9 question types shown in a drop-down (question types assigned to a question are shown in green highlight color) or (iii) flag/ignore the question by selecting the check-box corresponding to one of the reasons such as “invalid question”, “Serious lang. issue” etc. (the reasons chosen for flagging a question are shown in red highlight color)
Figure A.3: **Annotation Stage 3: Reviewing answer mismatch cases:** If none of the answers entered in the first stage for a question match with any of the answers entered in the second stage, the question is sent for review in a third stage. This review is handled by the authors and reviewer is allowed to edit question as well answers or add new answers before accepting the question.
Q: Is it an existing item?
Question types: form and yes/no
A: yes

Q: What is the date given at the top left?
Question types: layout
A: 03/17/98

Figure B.1: On the left is a question based on an yes/no check box. On the right, the question seeks for a date given at a particular spatial location — top left of the page.
Q: What is the date written next to RSM approval?

Question types: form and handwritten

A: 3-17-98

Figure B.2: Date is handwritten and it is shown in a key:value format.
Q: If the request needs to be warehoused by RJR, what needs to be done?
Question types: running text
A: write to RJR

Figure B.3: Question is grounded on a sentence.


Q: Whose picture is given?

**Question types:** photograph and layout

A: Dr. Dwayne G. Westfall

Q: What is the average sucrose % for N level 501+?

**Question types:** table

A: 15.9

---

Figure B.4: On the left is a question asking for name of the person in the photograph. To answer the question on the right, one needs to parse the table and pick the value in the appropriate cell.
Q: What is the highest value for “Intake, mg/1000 kcal” plotted on the ‘X’ axis of the graph?

Question types: figure

A: 300

Figure B.5: Question is based on the plot shown at the bottom of the given image, asking for the highest value on the X axis.
Q: What is the total cost for Fat cell size (Mt. Sinai) in the -05 year?

GT: $35,864

M4C best: 4400

BERT best: $35,864

Human: $35,864

Q: What is the first recipe on the page?

GT: hawaiian fruit cake

M4C best: island desserts (continued from cake)

BERT best: hawaiian fruit cake

Human: hawaiian fruit cake

Figure C.1: Examples where BERT QA model [2] answers questions other than ‘running text’ type. On the left is a question based on a table, and for the other question one needs to know the ‘first recipe’ out of the two recipes shown. For the first question the model gets the answer correct except for an extra space, and in case of the second one the predicted answer matches exactly with the ground truth answer.
Figure C.2: How does the M4C [3] model perform on questions based on pictures or photographs. Here we show two examples where the best variant of the M4C model outperform the BERT best model in answering ‘layout’ type questions seeking to read what is written in a logo/pack. The BERT model doesn't make any predictions for the questions.
**Figure C.3: Contrasting results for similar questions.** Here both the questions are based on the table at the bottom of the image. Both questions ask for ‘committee strength’ for a particular meeting (first or last). Both models get the answer right for the first one. But for the question on the right, the models predict same answer as the first one (“6”) while the ground truth is “5”. This suggests that the models’ predictions are not backed by a proper reasoning/grounding in all cases.
Q: What is the position above "vice chairman"?

GT: chairman
M4C best: legal counsel
BERT best: legal counsel
Human: chairman

Q: What is the highest value shown on the vertical axis?

GT: 99.99
M4C best: 50
BERT best: 32
Human: 99.99

Figure C.4: **Understanding figures and diagrams.** In case of the question on the left, one needs to understand an organizational hierarchy diagram. For the second question, one needs to know what a ‘vertical axis’ is, and then find the largest value. Both the models fail to answer the questions.
Q: What is the name of the passenger?
GT: dr. william j. darby
M4C best: larry
BERT best: larry
Human: dr. william j. darby

Q: What is the date present in the memo?
GT: 1/7/77
M4C best: 1 7 77
BERT best: 1 / 7
Human: 1/7/77

Figure C.5: Impact of OCR errors. Here the models are able to ground the questions correctly on the relevant information in the image, but failed to get the answers correct owing to the OCR errors. In case of the question on the left, even the answer entered by the human volunteer is not exactly matching with the ground truth. In case of the second question, OCR has split the date into multiple tokens due to over segmentation, resulting in incorrect answers by both the models.