Efficient End-to-End Visual Document Understanding with Rationale Distillation

Anonymous ACL submission

Abstract

Understanding visually situated language requires interpreting complex layouts of textual and visual elements. Pre-processing tools, such as optical character recognition (OCR), can map document image inputs to textual tokens, then large language models (LLMs) can reason over text. However, such methods have high computational and engineering complexity. Can small pretrained image-to-text models accurately understand visual documents through similar recognition and reasoning steps instead? We propose Rationale Distillation (RD), which incorporates the outputs of OCR tools, LLMs, and larger multimodal models as intermediate "rationales", and trains a small student model to predict both rationales and answers. On three visual document understanding benchmarks representing infographics, scanned documents, and figures, our PIX2STRUCT (282M parameters) student model finetuned with RD outperforms the base model by 4-5% absolute accuracy with only 1% higher computational cost.

1 Introduction

Information in the digital world is conveyed through text integrated with visual elements, such as complex layouts, figures, and illustrations. Answering user questions based on such visual documents requires models to recognize and connect text and layout to the user need.

While pretrained image-to-text multimodal models have demonstrated strong performance on visual document understanding (VDU) by directly mapping pixel-level input document images to answers corresponding to user queries (Kim et al., 2022; Lee et al., 2023; Chen et al., 2023b,c), state-of-the-art approaches benefit from the use of external tools. Tools include OCR systems (Chen et al., 2023b; Powalski et al., 2021; Huang et al., 2022), structured table source extraction (Liu et al., 2023a), and LLMs reasoning over extracted information and the user query (Liu et al., 2023a; Perot et al., 2023). Additional tools such as image captioning, object classification, and search engines have been used for other multimodal tasks (Yang et al., 2023; Zhang et al., 2023). However, the accuracy gains from these external components come at the cost of decreased computational efficiency and increased engineering complexity.

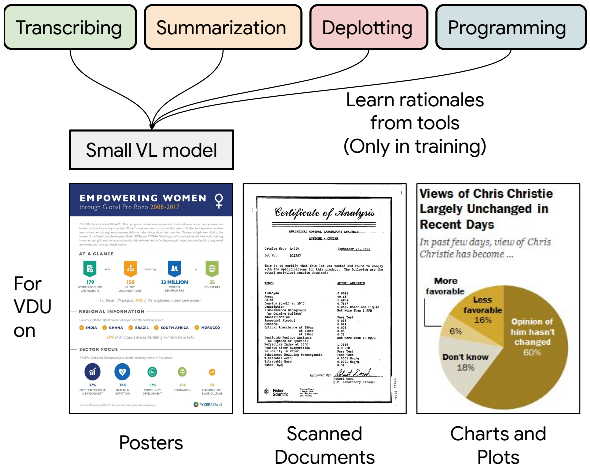

Figure 1: We synthesise the ability of recognizing and summarizing text, deplotting structured plots, and program generation into one small model, and perform efficient rationale-based visual document understanding.

In this work, we ask whether we can achieve high accuracy and efficiency by teaching a smaller model to learn from short rationales generated by external tools and expensive LLMs (see Figure 1). We use a small student image-to-text model to perform VDU tasks by decomposing them into rationale prediction and answer steps, predicting the rationale and answer in sequence. The "rationale" can be any intermediate textual information that helps answer a question correctly: for instance, it could be a subset of relevant text from the image as well as layout, structured information, and reasoning (see Figure 2).

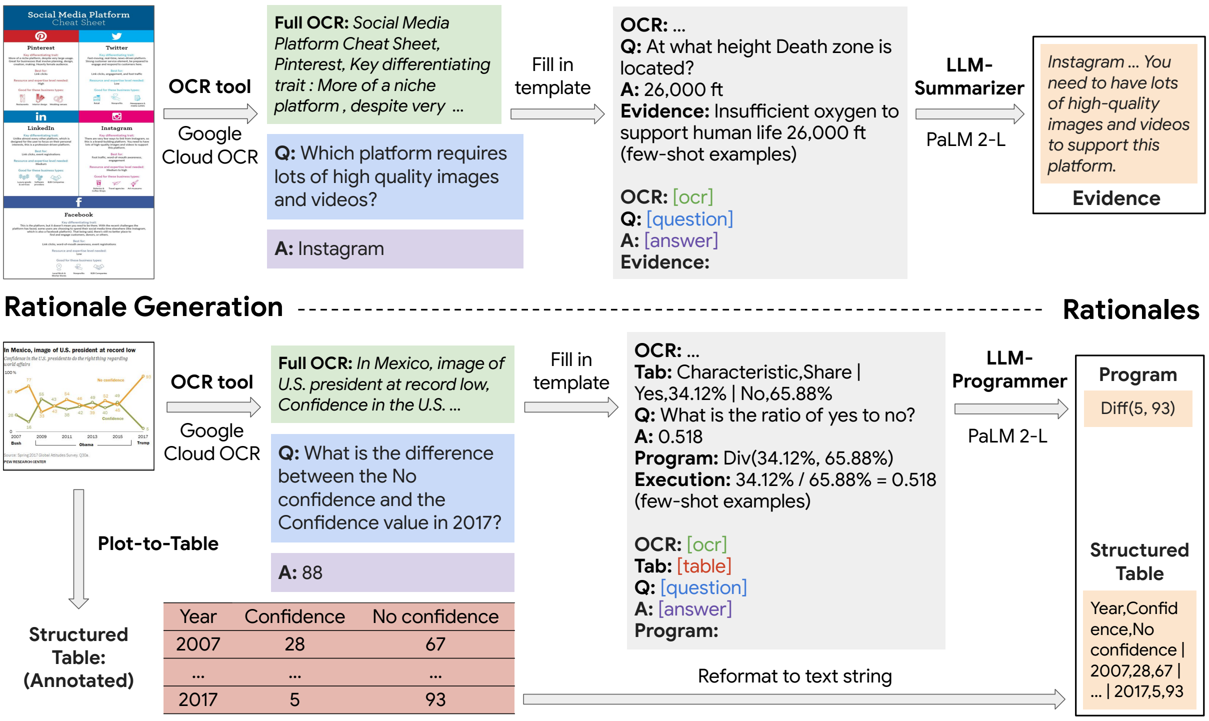

Figure 2: For training examples, we first generate the full OCR of each image with Google Cloud OCR. Depending on the dataset, we either use LLM-Summarizer (few-shot prompted PaLM 2-L) to generate text evidence (top), or use LLM-Programmer (also PaLM 2-L) to generate a program based on both the OCR and available structured table source for the image (bottom).

The training data for VDU tasks of interest does not generally contain annotated "rationales." It is also not known what types of sufficiently succinct rationales, even if available, would be useful for a small image-to-text model. We take inspiration from related works on chain-of-thought distillation (Shridhar et al., 2023; Zhang et al., 2023) for text and multimodal tasks, borrowing techniques and adding novel components to address the specific challenges within the visual document understanding domain. We use chains of tools at training time to derive short rationales representing salient subtasks of the problem—recognizing text and layout, and deriving programs to encode numerical reasoning. To increase the quantity and validity of example rationales, and the student's robustness to incorrect predictions, we design data augmentation schemes and DAGGER-style (Ross et al., 2011) loss functions, which improves the student's ability to benefit from intermediate predictions.

Our method takes advantage of task decomposition and reasoning, but offers the following advantages over other tool-using models:

- No OCR or other external tools used during inference, reducing engineering complexity.
- Only a short, query-dependent rationale is predicted versus longer structures typically extracted by external tools, saving computation.
- Computation is increased by only about 1% (in FLOPS) compared to models that predict the answer directly.

We conduct experiments on three VDU benchmarks: InfoVQA (Mathew et al., 2022), DocVQA (Mathew et al., 2021), and ChartQA (Masry et al., 2022). We show accuracy improvements over models that predict answers directly. For models based on PIX2STRUCT-Base (282M parameters), improvements are 4.0 and 4.6 points in ANLS on InfographicVQA and DocVQA respectively, and 3.3 / 7.7 points in relaxed accuracy on ChartQA's augmented and human sets, with similar improvements for larger PIX2STRUCT models (1.3B parameters).

2 Task definition

In VDU, a model is given an image I and user question q, and predicts text answer a. We focus on training a single small image-to-text model with parameters θ for this task. Prior work in VDU trains such models by maximizing the training data log-likelihood according to an image-to-text (or image+text-to-text) model that directly

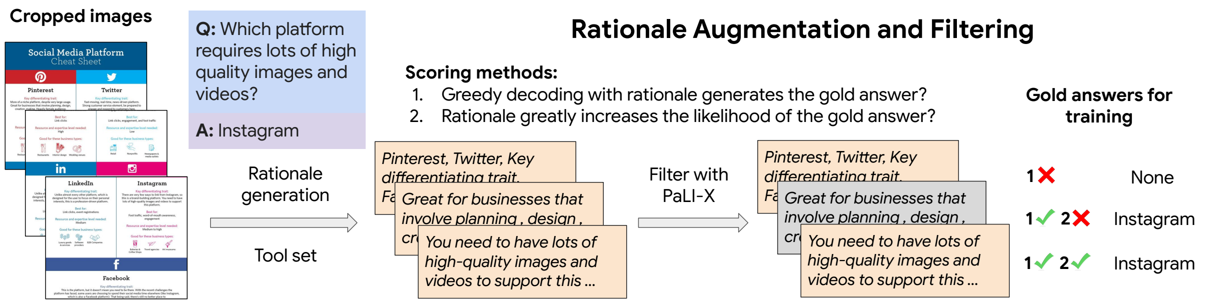

Figure 3: We first crop along the longer edge of the image to create multiple smaller square images. We generate rationales using the appropriate subset of tools (OCR, LLM-Summarizer, LLM-Programmer, Plot-to-Table) on these images, then categorize the examples and rationales with Multimodal-Verifier (PaLI-X).

generates the text answer a given the input and makes predictions through greedy decoding, *i.e.*, $\hat{a} = \arg\max p_\theta(a \mid q, I)$ (Kim et al., 2022; Lee et al., 2023; Chen et al., 2023b; Wang et al., 2022).

We assume that external tools such as OCR systems, LLMs, larger image-to-text models, or structured input image source information may be available at training time, but not inference time. We use such tools and metadata to derive rationales r paired with training input-output examples (I, q, a), and train a small student image-to-text model to predict rationales r as an intermediate reasoning step, before predicting the answer a.

3 Rationale Distillation

We propose Rationale Distillation (RD), which distills rationales from a predefined set of tools, and trains a student model to predict the relevant rationales before predicting the answer.

Rationales are sequences of text tokens of relevant information to arrive at the answer. We consider two kinds of rationales r: natural language text evidence derived from the output of an OCR system; and tabular representation of charts in the input image concatenated with simple custom programs with predefined operations. The tools we leverage are: an OCR tool (Google Cloud OCR); Plot-to-Table, a converter that converts charts or plots to structured tables; a LLM-Summarizer (designed by us), which summarizes OCR text to evidences relevant to the question using a prompted PaLM 2-L model (Anil et al., 2023); a LLM-Programmer (also designed by us and based on PaLM 2-L), which generates simple programs for numerical reasoning tasks; and a Multimodal-Verifier based on PaLI-X (Chen et al., 2023b), which verifies the quality of the rationales. We provide detailed descriptions of these tools in Appendix A. As these tools add heavy computation (for LLMs) or engineering complexity (for OCR), we depend on none of them at inference time.

In this section, we first discuss the process of generating the two types of rationales from tools (§3.1). We then describe a data augmentation scheme for increasing the number of examples with rationales, and making student models more robust to potentially noisy rationales (§3.2). Finally, we discuss training and inference for student models to predict the rationale and the answer (§3.3, 3.4).

3.1 Rationale generation from tools

InfoVQA and DocVQA require a strong ability to recognize text, so we first use the OCR tool to extract the text from the image, then perform 5-shot prompting with LLM-Summarizer to generate question evidence (Figure 2, top). ChartQA focuses on numerical reasoning on charts, so we extract the full OCR text, obtain structured tables using Plot-to-Table,[1] and then prompt LLM-Programmer (using 8 in-context examples) to generate a program to derive the answer. The concatenation of the structured table and the program are then used as the rationale (Figure 2, bottom). Detailed prompt templates are in Appendix E.

3.2 Rationale augmentation and filtering

We aim to enable a small student model to reason over visual documents ranging over diverse formats and complexity. The expensive tools can typically generate high-quality rationales from such data, but it is a significant challenge for a small student model to match the quality of these rationales from a limited amount of training data. To overcome this challenge, we devise a data augmentation approach based on image cropping to greatly enlarge the

[1] We use provided ChartQA structured tables directly.

number of examples available for rationale prediction, and to teach the model to use variable quality rationales in generating the final response.

Algorithm 1 Rationale Filtering

1: **Input:** image I; question q; answer a; rationale r; multimodal verifier with parameter ϕ.
2: **Output:** A tuple (the category of the rationale, the assigned answer used for training).
3: **if** $\arg\max_{\hat{a}} p_\phi(\hat{a} \mid I, q, r) \neq a$ **then**
4:     **return** "irrelevant", "None"
5: **end if**
6: **if** $[p_\phi(a \mid I, q, r)]^\lambda \geq p_\phi(a \mid I, q)$ **then**
7:     **return** "useful", a
8: **else**
9:     **return** "relevant but not useful", a
10: **end if**

Task name	Encoder input	Decoder input	Target output
QRA	I	-	q, r, a
ASR	I	$q, \hat{r}$	a
QRACI	i_j	-	$q, r_j, \bar{a}$
ALRCI	i_j	q, r_j	$\bar{a}$

Table 1: We compute loss on the target output tokens for four student training tasks. Encoder input images have questions q rendered as the header. Rationale r (resp. r_j) is generated by tools on image I (resp. i_j). Rationale $\hat{r}$ is generated by students.

Cropping-based augmentation. We crop the original image along the longer dimension, resulting in multiple square images (Figure 3). To minimize the possibility that the most relevant segment does not fit within any crop, we use a sliding window with adjacent croppings overlapping by half the image size (Algorithm 2, appendix). For an input image I, we obtain k cropped images $i_1, \ldots, i_k$ and generate corresponding rationales for them as detailed above. As an example, in the InfoVQA dataset we observe an average of $k \approx 4$.

Filtering relevant and useful examples. While cropping significantly increases the size of our training dataset, many of the images might not contain information pertaining to the answer, and we may not be able to extract reasonable rationales. Including such examples in our dataset can amplify noise and make the problem more challenging for the student. So we carefully filter the augmented data to extract examples which are useful for rationale and/or answer prediction. We use a powerful Multimodal-Verifier (PaLI-X) with parameter ϕ to design two filters on VDU tasks (Algorithm 1).

(1) The *relevance filter* checks if the cropped image i_j contains information for answering the question by comparing greedy decoding with the rationale as input against the gold answer: $\arg\max_{\hat{a}} p_\phi(\hat{a} \mid i_j, q, r_j) = a, j \in \{1..k\}$ (row #3 of Algorithm 1). For examples failing this filter, we replace the answer a with None in the training data, assuming the cropped image is insufficient to generate the answer. For instance, the first cropped image of Figure 3 does not contain the gold answer "Instagram" and the example falls within the irrelevant category. We still use the rationale r_j for rationale prediction, since it could help distill the tool into the student model.

(2) The *rationale filter* applies to examples that pass the relevance filter, and checks if the probability of the gold answer is sufficiently increased given the rationale (row #6 of Algorithm 1). We use a factor $\lambda = 2$ to avoid small perturbation caused by changing the format of the model prompt by concatenating the rationale. For examples that pass the relevance filter but not the rationale (row #9), we regard the rationale r_j as low-quality, and do not use it for learning rationale prediction. For instance, the second cropped image of Figure 3 contains the gold answer "Instagram", but the tools do not generate a useful rationale.

We classify (i_j, r_j) pairs into three categories (rows #4,7,9 of Algorithm 1), which determine their assigned answer $\bar{a} = a$ or None and the way their rationales are used in training.

Dataset balancing. Most examples fail the relevance filter, and more than half of the ones that pass fail the rationale filter. We subsample the examples with label None (row #4) such that their number $n_{\text{row \#4}} \leq n_{\text{row \#9}} - n_{\text{row \#7}}$.

3.3 Training student models

In Rationale Distillation, we perform multi-task training for the student model, using tasks derived from the original and augmented data annotated with rationales. Tasks differ by their encoder and decoder inputs and decoder outputs (Table 1). We weight four tasks equally (*i.e.*, 0.25 for each), and train on a linear combination of them, with loss defined over the target output.

Distilling the tools directly. This vanilla Question, Rationale and Answer (QRA) distillation setup teaches the model to take in the original image and predict q (which can be read out from the

Model	Method	Dev				Test			
		InfoVQA	DocVQA	ChartQA		InfoVQA	DocVQA	ChartQA	
				aug.	human			aug.	human
Base (282M)	Ans-Only	36.8	72.3	75.9	34.3	38.2	72.1	81.5	30.3
	QID	38.2	75.5	76.2	35.4	39.5	75.7	82.3	32.5
	RD (Ours)	**41.3**	**76.3**	**78.9**	**36.7**	**42.2**	**76.7**	**84.8**	**38.0**
	Oracle	48.1	82.5	84.7	43.1	-	-	-	-
Large (1.3B)	Ans-Only	39.6	76.0	77.3	36.3	40.0	76.6	83.8	35.2
	QID	41.0	77.8	78.5	37.8	41.9	77.9	85.0	35.9
	RD (Ours)	**43.5**	**79.2**	**81.6**	**39.3**	**44.3**	**79.0**	**88.6**	**40.6**
	Oracle	53.5	84.0	85.8	46.5	-	-	-	-
(≥5B)	SOTA	-	-	-	-	62.4†	88.6†	91.0‡	67.6‡

Table 2: PIX2STRUCT-based results on three benchmarks. We show Rationale Distillation consistently outperforms the Ans-Only and QID baselines on both Base and Large models. Results marked by † are from Chen et al. (2023c), and ones marked by ‡ are from Liu et al. (2023a).

image header), r (the intermediate rationale generated by the tools), and then by the answer a.

Robustifying against student rationale errors. To help make the student model robust to its own mistakes, the Answer with Student Rationale (ASR) task provides question q and student generated rationale $\hat{r}$ as decoder input for the student model to predict the answer. To generate such student rationales $\hat{r}$, we use a separately trained PIX2STRUCT-based student model, which learns to predict only rationales.

We sample three student generated rationales for each input example and use them as the low-quality rationales $\hat{r}$. Since the training loss for ASR is only applied to the answer prediction, the RD student is not encouraged to replicate these noisy rationales, but to be able to recover from potential errors and predict the gold answer. We note that other than the difference of a separate student model generating the rationale, this is akin to student-forcing or DAGGER style approaches to structured prediction (Ross et al., 2011).

Leveraging cropped images. In Question, Rationale and Answer on Cropped Images (QRACI), we use cropped images i_j with rationales identified as useful (row #7 of Algorithm 1) or irrelevant (row #4), to learn to predict those rationales and the original answer or `None`, respectively. Answer with Low-quality Rationale on Cropped Images (ALRCI) is similar to ASR, taking cropped images as encoder input and providing low-quality rationales (row #9) in the decoder input.

3.4 Model architecture and inference details

PIX2STRUCT is an encoder-decoder model using a Transformer image encoder for an input image, and a Transformer-based decoder generating text. Following Lee et al. (2023), we render the question q as the header of the image I for visual document understanding tasks and do not provide the question through a textual input channel. We take `<s>` and `<answer>` as separators, and use the following encoding format for the decoder sequence: [*question*] `<s>` [*rationale*] `<answer>` [*answer*]. As the decoder sequence length of PIX2STRUCT is 128 tokens,[2] we trim the sequence before [*answer*] to 108 tokens and leave 20 tokens for the answer.

If the rationale has programs, like in ChartQA, we put both the structured table and the program in the [*rationale*] slot, using the format [*rationale*] = [*table*] `<program>` [*program*]. As the structured table is usually long, we trim the sequence before [program] to 64, leaving 44 tokens for the program.

During inference, we evaluate only on the original, non-cropped images with greedy decoding. To avoid generating answer `None`, we force the model to decode non-`None` after the `answer` token.

Note that student model's intermediate predictions are relatively short. The overall floating-point operations (FLOPs) compared to a baseline model that directly generates answers are increased by less than 1% (see Appendix D for a derivation).

4 Experimental results

We study the impact of rationale distillation across three benchmarks, analyze the contribution of each component of our approach, and the extent to which a single student model can match the capabilities of the external tools and LLMs it learns from.

[2]Defined using PIX2STRUCT's tokenizer.

4.1 Dataset metrics

InfoVQA and DocVQA use the average normalized Levenshtein similarity (ANLS) score as the evaluation measure. ChartQA uses relaxed accuracy (RA) and includes an easier augmented evaluation set and a harder human-generated evaluation set.

4.2 Baselines

PIX2STRUCT We compare with the original PIX2STRUCT fine-tuning approach for both Base and Large models, where the model takes in an image I with the question rendered as a header as encoder input and directly predicts a.

QID Fine-tuning tasks QRA and QRACI predict the question as part of the decoder output. To detect improvements due to reading out the question as an intermediate step, we compare to the question-in-decoder (QID) setup, where the PIX2STRUCT model takes in I in the encoder input and predicts the sequence q, a separated by <answer>.

Oracle To establish an upper bound on performance of the student model if it was able to condition on the tool-generated high-quality rationales, we also compare to an oracle method on the development set. We use the tool generated rationale r during evaluation to get an oracle measure that uses information about the gold answer a.

We also describe other existing VDU approaches and compare RD to them in Table 8 (Appendix B).

4.3 Main results

Table 2 evaluates our rationale distillation (RD) method against baselines.

Overall trends. Overall, RD shows consistent improvements on InfoVQA (4.0 and 4.3 points), DocVQA (4.6 and 2.4 points) and ChartQA-human (7.7 and 5.4 points) test sets for both base and large model variants (respectively) over the PIX2STRUCT baseline. We also see that including the question in the decoder brings benefits across all datasets and variants. Next, we discuss the value of rationale distillation in comparison to this stronger QID baseline.

Textual rationales. Table 2 shows consistent improvements due to OCR and LLM-Summarizer rationales compared to the QID baseline. RD records improvements of 2.7 and 2.4 points on InfoVQA and 1.0 and 1.1 points on DocVQA for base and large variants respectively for the test set.

Method	Dev Set			
	InfoVQA	DocVQA	ChartQA	
			aug.	human
RD	41.3	76.3	78.9	36.7
RD+Voting	41.7	76.6	79.4	37.0

Table 3: RD on PIX2STRUCT-Base with voting during inference. Decoding with voting shows small but consistent improvements across datasets.

Table and program rationales. On ChartQA, we use rationales including Plot-to-Table (underlying tables for charts), as well as programs derived by LLM-Programmer (based on this table and OCR). Using such rationales results in improvements of 2.5 and 3.6 points respectively on base and large variants on the augmented set over the QID baseline. We see even larger improvements: 5.3 and 4.7 points for base and large models, respectively, on the harder human eval set which requires more complex mathematical reasoning.

Accuracy and efficiency trade-off. We show that efficiency and accuracy can be improved at the same time. The performance of the Base model with RD is better than that of the Large model with Ans-Only; the inference FLOPs of the former ($\sim$2.65E+12) are also lower than those of the latter ($\sim$4.63E+12; Appendix D shows a derivation).

On the other hand, PIX2STRUCT Large with RD still shows gaps compared to the SOTA methods — PaLI-3 with OCR (Chen et al., 2023c) on InfoVQA and DocVQA, and a tool use case with deplotting and prompted LLM (Liu et al., 2023a). It is worth noting that these methods use more than 10 times the FLOPs of the PIX2STRUCT Large model and also use more data.

4.4 Analysis

Using Base-sized models, we analyze the impact of the inference method and compare RD to pipelines where external tools and LLMs can be called at inference time. We also ablate the impact of the different tasks designed to drive student model learning and examine the types of questions that benefit most from Rationale Distillation.

Top-n voting in inference. We can naturally apply top-n voting during inference, which is similar to making predictions using self-consistency in chain-of-thought (Wang et al., 2023c). We simply perform beam search decoding with a beam size of $n = 5$ and aggregate the probabilities of the distinct answers appearing in these hypotheses.

#	Task	InfoVQA Dev Set
1	QRA	36.7
2	QID	38.2
3	QRA and ASR	40.1
4	QRA, ASR and QRACI	41.0
5	QRA, ASR and ALRCI	40.5
6	All 4 tasks	41.3

Table 4: We conduct ablation study of different student training task combinations on the InfoVQA dev set: Question, Rationale, Answer (QRA), Answer with Provided Rationale (APR) and analoguous tasks on Cropped Images (CI). We show the importance of both training to predict the gold rationales and training to predict the answer based on the noisy rationales (row # 3), as well as the usefulness of image cropping augmentation (row # 6).

We choose the answer (that is not `None`) with the highest aggregate probability as the final prediction. From Table 3, we see that this leads to small but consistent improvements across datasets, albeit at an increased computation cost.

How much does each of the tasks aid the student in predicting helpful rationales? In Table 4, we tease apart the contribution of each training task. First, we see that a model which uses standard supervised training with QRA (*i.e.*, predicting the question, rationale and answer) performs worse than the QID baseline. This result suggests that it is important to make the student model robust to its own errors and expose it to rationales with varying degrees of relevance to the question.

Augmenting QRA with ASR (training with predicted rationales) results in a gain of about 1.9 points absolute (row #3). The additional image, rationale and answer examples obtained through image cropping and verifier categorization bring further improvements of 1.2 points (row #6).

What is the usefulness of the rationale generated by the student in comparison to external tools? On InfoVQA, we analyze the usefulness of the student-generated rationale in comparison to evidence from the OCR tool and several ways to sub-select fragments of similar length from it including LLM-Summarizer without access to gold answer (based on PaLM 2-L) (Figure 4). The systems are shown (from left to right) in order of increasing computation costs and engineering complexity. All methods except QID are evaluated with Pix2Struct-Base trained with RD, using corresponding rationales as decoder input during inference.

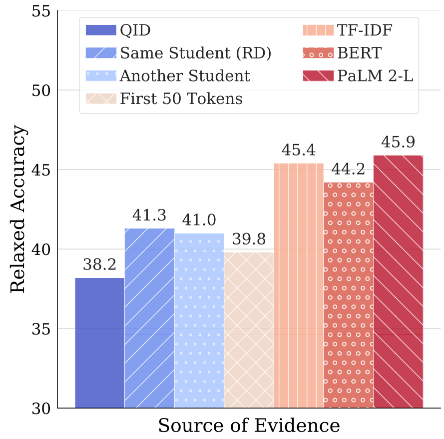

Figure 4: We analyze the usefulness of student generated rationales. The systems are shown in order of increasing engineering complexity. All red bars use a pipeline with Google Cloud OCR during inference. RD trades off between accuracy and efficiency/complexity.

Since OCR outputs can be very long, we experiment with different methods for selecting 50-token segments. The simplest variant truncates the OCR output to the first segment of 50 tokens which results in a small gain (1.6 points) over QID. More complex methods which select segments based on TF-IDF (7.2 points) or BERT-embedding based (6 points) similarity to the question result in larger gains. Finally, the rightmost red bar shows the performance with rationales from few-shot prompted PaLM 2-L. For this experiment, we modified the prompting template for PaLM 2-L, to generate rationales from the OCR without being given the answer. Specifically, we ask PaLM 2-L to predict the evidence first, and the answer next and use only the evidence (and not the answer) from PaLM 2-L as student decoder input. This variant performs the best with a 7.7 point improvement over QID. Overall, these results indicate a significant room for improvement in rationale prediction for student models. We also see that an external OCR tool would still provide benefits at the cost of added computation by the OCR system and, since OCR is relatively efficient, the more significant cost of increased engineering complexity and potential service fees for production solutions.

What if we use an external calculator on the generated programs? Using a calculator – an additional but computationally inexpensive tool – could further enhance the capabilities of our models. For valid programs generated by student models, we use a calculator to carry out computations

Method	Answer type				Evidence					Operation		
	Image span	Question span	Multiple spans	Non span	Table/ List	Textual	Visual object	Figure	Map	Comparison	Arithmetic	Counting
Ans-Only	41.5	43.8	16.6	30.1	33.5	49.7	23.8	36.3	32.6	23.4	40.4	**18.9**
RD+Voting	**46.6**	**46.7**	**18.8**	**30.4**	**40.6**	**57.7**	**28.0**	**37.9**	**36.5**	**28.1**	**41.2**	17.7

Table 5: Breakdown ANLS score on different types of questions and answers from InfoVQA test set. RD benefits questions related to text or table evidence most.

dictated by the programs, and take the output of the calculator to replace model output. For invalid programs, we keep using the model generated answer prediction. We observe that on ChartQA, the calculator use, when combined with voting, leads to further improvements of 0.4 RA on the augmented set and 3.3 RA on the human set.

Breakdown analysis of the improvement. The InfoVQA leaderboard provides a breakdown of model performance over subsets categorized by answer type, evidence type, and question operations. We compare the performance of Ans-Only models (ANLS 38.2) and RD+Voting (ANLS 42.2) in Table 5. We observe large improvements when answers are text spans in the image or in the question. The former type indicates the helpfulness of the intermediate rationales; the latter suggests the helpfulness of decoding the question before answering.

We see a 7.1 points gain when the evidence comes from a table or list, 8 points when the evidence comes from text, which implies the student can extract better rationales from such parts of the images, in comparison to parts with more complex layouts such as figures and maps.

We did not use programs as rationales for InfoVQA, and we do not see large improvement on arithmetic and counting questions. Using programs as parts of the rationales in this and other types of tasks is a promising direction for future work.

5 Related work

Using tools to augment the input in a prediction problem can be seen as using additional reasoning steps, *i.e.*, calling a tool with a set of arguments and integrating its result with the rest of the context. Much prior work on VDU has relied on calling OCR (Tang et al., 2023; Appalaraju et al., 2021; Huang et al., 2022), object detector (Kim et al., 2023), or de-plotting tools (Liu et al., 2023a). Such works have not attempted to recognize text or structured data as an intermediate reasoning step using the same small model.

On the other hand, the specific structure of reasoning chains through prompting LLMs has been shown to have significant impact (Wei et al., 2022; Zhou et al., 2023; Khot et al., 2023; Yao et al., 2023). Distilling these text rationales from large teacher models has been shown successful by chain-of-thought distillation works on NLP benchmarks (Shridhar et al., 2023; Li et al., 2023) and ScienceQA (Zhang et al., 2023; Wang et al., 2023a). Toolformer (Schick et al., 2023) trains smaller language models to call tools. Generic multimodal tool use solutions based on LLMs have also been proposed (Yang et al., 2023). However, these works do not replicate the results of tool output and replace them for efficiency.

We marry the powerful ideas of taking intermediate reasoning steps from tools for accuracy, and distilling to small student models for efficiency, as we have proposed in RD.[3]

6 Conclusions

We showed that the visual document understanding ability of small image-to-text models can be improved by our proposed Rationale Distillation. In RD, we obtain rationales for training examples using external tools and LLMs, and train small end-to-end student models to predict rationales as intermediate reasoning steps. We demonstrated the importance of designing student training tasks that make the model robust to irrelevant rationales.

RD leads to substantial improvements via textual evidence distillation on the text-heavy InfoVQA & DocVQA datasets, and via Plot-to-Table and program distillation on the numerical reasoning-focused ChartQA dataset. Analysis shows the gains transfer to stronger models such as MatCha (Appendix B) and larger Pix2Struct models. Marginalizing over rationales and using a cheap calculator tool at inference time bring additional consistent benefits. Controlled experiments show that RD offers a tradeoff between performance and computational cost/engineering complexity, in comparison to systems relying on tool pipelines.

[3] We overview more related works in Appendix C.

Limitations

Our study shows RD can teach small models to successfully generate and utilize two types of rationales: summarized OCR evidence, and structured table concatenated with a simple program. A broader set of tools, such as object detection, image segmentation and captioning tools, can be further explored as rationales to enhance the ability of visual document understanding.

To use resources sparingly, we evaluate on the PIX2STRUCT series of models up to a size of 1.3B parameters (including the stronger MATCHA model; see Appendix B). In the future, RD could also be evaluated on other more powerful pre-trained models for visual document understanding, such as PaLI-3 (Chen et al., 2023c) or ERNIE-Layout (Peng et al., 2022).

We focus on single-page visual document understanding, and have not explored the potential of RD on multi-page images. Multi-page image problems may have longer-distance dependencies, and require student models to generate more complex rationales as the intermediate reasoning steps.

We inherit the ethical concerns of existing LLMs and multimodal models, such as privacy considerations and potential misuse. Here we use public peer-reviewed datasets to evaluate our method. For use in deployed applications, the data for RD should be constructed with careful data curation. Privacy-sensitive documents which contain personal information, should be excluded from the training data to prevent potential privacy breaches and unintended consequences.

References

Kriti Aggarwal, Aditi Khandelwal, Kumar Tanmay, Owais Khan, Qiang Liu, Monojit Choudhury, Hardik Chauhan, Subhojit Som, Vishrav Chaudhary, and Saurabh Tiwary. 2023. DUBLIN - document understanding by language-image network. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *International Conference on Computer Vision (ICCV)*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. PaLI-X: On scaling up a multilingual vision and language model.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023c. PaLI-3 vision language models: Smaller, faster, stronger. *arXiv*.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. Depth-adaptive transformer. In *International Conference on Learning Representations*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *ICLR*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Donut: Document understanding transformer without OCR. In *ECCV*.

Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. 2023. Visually-situated natural language understanding with contrastive reading model and frozen large language models. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023b. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Winter Conference on Applications of Computer Vision (WACV)*.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Yuhui Cao, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Jiaqi Mu, Hao Zhang, and Nan Hua. 2023. Lmdx: Language model-based document information extraction and localization.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, pages 732–747. Springer.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA. PMLR.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen tau Yih, and Ashish Sabharwal. 2018. Quarel: A dataset and models for answering questions about qualitative relationships. In *AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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)*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *arXiv preprint arXiv:2305.03453*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching llm to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. 2023. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv:2310.05126*.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*.

Wang Zhu, Jesse Thomason, and Robin Jia. 2023. Chain-of-questions training with latent answers for robust multistep question answering. In *EMNLP*.

A Implementation details

A.1 Description of used tools

OCR For all datasets, we begin with calling an off-the-shelf external OCR tool (Google Cloud OCR), which takes the image as input and outputs the full text recognized in the image together with location information (see Figure 2).

LLM-Summarizer OCR outputs can be quite long for some images, and not all text in the input image is directly relevant to a given question. To minimize computation spent on intermediate rationale prediction steps, we employ another powerful tool — a prompted large language model PaLM 2-L (Anil et al., 2023), to generate a significantly shorter span of text (less than 100 tokens), given the question, answer, and the full image OCR text (see Figure 2 top for an example). We sample a single evidence with temperature of 0.1 to obtain these rationales from PaLM 2-L.

Plot-to-Table In addition to relevant text on the screen, some visual document domains and types of problems can benefit from other types of intermediate structure. An example is understanding charts and figures, whose underlying structured source data is not well captured by OCR systems. Such structured source data is available in some datasets, *e.g.*, ChartQA provides structured data tables extracted by ChartOCR (Luo et al., 2021); but they can also be inferred for unannotated images through tools like DePlot (Liu et al., 2023a).

LLM-Programmer For problems involving numerical reasoning, we use a prompted LLM, PaLM 2-L, to generate a simple program capturing common numerical reasoning patterns corresponding to user queries, given the question, answer, the full image OCR text and the structured table (see Figure 2 lower half for an example). The programs are limited to the following formats: `Div(a,b)`; `Mul(a,b)`; `Avg(a list of numbers)`; `Sum(a list of numbers)`; `Diff(a,b)`; `Greater(a,b)`; `Less(a,b)`; `Find(str)`. All programs except `Find(str)` have execution steps in the prompt templates, which explain how to connect the programs to arithmetic and comparison operations. The last program type is applicable if numerical reasoning of the other types is not needed, and has no operation involved. Note that the program rationale is not executed by default, but is only used to guide the model towards the correct answer.

Multimodal-Verifier To determine the helpfulness of the rationale generated by other tools and the relevance of image augmentations, we employ a multi-task trained, large multimodal model PaLI-X 55B (Chen et al., 2023b). We construct the text encoder input in the following format:

[rationale] Answer in en: [question]

The verifier takes in the image I as input to the vision encoder, the question q and the rationale r as input to the text encoder. We use the log-probability of the gold answer (with and without conditioning on the rationale), and the correctness of the predicted answer (through greedy decoding), to define two measures of rationale helpfulness.

A.2 Algorithm for rationale augmentation

Here we list the detailed algorithm for rationale augmentation described in §3.2.

Algorithm 2 Rationale Augmentation via Image Cropping

```
Input: image I; question q; answer a; tools for rationale generation.
Output: a set of cropped images, and a corresponding set of rationales.
Initialize the counter j ← 0, the cropped image set 𝓘 ← ∅ and the rationale set 𝓡 ← ∅.
Get the height h and the width w of the image I.
if h ≥ w then
    while wj < h do
        start ← wj/2
        end ← min(wj/2 + w, h)
        image i_j ← crop [start, end] on the height of I.
        Get rationales r_j for i_j, q, a from tools.
        𝓘 ← 𝓘 ∪ i_j; 𝓡 ← 𝓡 ∪ r_j; j ← j + 1.
    end while
else
    while hj < w do
        start ← hj/2
        end ← min(hj/2 + h, w)
        image i_j ← crop [start, end] on the width of I.
        Get rationales r_j for i_j, q, a from tools.
        𝓘 ← 𝓘 ∪ i_j; 𝓡 ← 𝓡 ∪ r_j; j ← j + 1.
    end while
end if
return 𝓘, 𝓡
```

A.3 Student rationale generation for ASR

For student rationale generation, we cannot directly use the student trained on the whole training set, as it is likely to remember and replicate the tool-generated rationale but this would not be representative of its behavior on unseen data.

On InfoVQA and DocVQA, we split the training data of into 3 folds. We train 3 student models, each takes in 2 folds as the train data and generates student rationale for the remaining fold. On ChartQA, to avoid the distribution shift from the augmented set and human set, we split both augmented set and the human set into 3 holds, in total 6 folds. We train 6 student models, each takes in 5 folds for training and generates student rationale for the remaining fold.

Here, the student models are only trained to generate the question and the rationale, not the answer. The output format of the student models is

[question] <s> [rationale]

For each example, we sample 3 rationales to create the ASR training set.

A.4 Hyper-parameters

Following the setup in Lee et al. (2023), for PIX2STRUCT-Base, we use an input sequence length of 6155 patches for InfoVQA, and 4096 patches for DocVQA and ChartQA. We train with a batch size of 128 for InfoVQA, and 256 for DocVQA and ChartQA, on 32 v3-Google Cloud TPUs.

For PIX2STRUCT-Large, we use an input sequence length of 3072 patches and train with a batch size of 64 for all datasets, , on 64 v3-Google Cloud TPUs.

We train all the model with 10k steps, optimizing using Adafactor (Shazeer and Stern, 2018). The learning rate schedule uses a linear warmup of 1k steps to 0.01, followed by cosine decay to 0. On InfoVQA and DocVQA, we select the model with the best ANLS score on the dev set for evaluation. On ChartQA, we select the model with the best RA on the dev augmented set for test evaluation. We report all the results under a single-run setup.

A.5 Scientific Artifacts and Licenses

We evaluate on three public datasets, InfoVQA, DocVQA and ChartQA, in our experiments. InfoVQA and DocVQA data is shared for non-commercial, research and educational purposes, which aligns with our use. ChartQA is under GNU General Public License v3.0. The questions in all three datasets are in English. We put the statistics of our evaluated datasets in Table 6.

Dataset	Domain	Train	Dev	Test
InfoVQA	Documents	23,946	2,801	3,288
DocVQA	Documents	39,463	5,349	5,188
ChartQA-human	Illustrations	7,398	960	1,250
ChartQA-aug.	Illustrations	20,901	960	1,250

Table 6: Statistics of the datasets we evaluate on.

Method	ChartQA Dev Set		ChartQA Test Set	
	aug.	human	aug.	human
Ans-Only	83.5	40.4	88.5	36.6
QID	84.6	40.2	89.7	37.5
RD	**86.0**	**40.9**	**90.8**	**42.1**

Table 7: We initialize the student model with MATCHA, which has stronger numerical reasoning skills. RD also improves MATCHA for ChartQA.

We finetune public models PIX2STRUCT and MATCHA. They are under Apache License 2.0.

B Additional experimental analysis

B.1 Model ablations

We show that RD also benefits stronger pretrained model such as (Liu et al., 2023b), while decoupling rationale and answer prediction is harmful.

What if we use a stronger pretrained model tailored to math reasoning as in ChartQA? We initialize our student model parameters with MATCHA (Liu et al., 2023b) instead of PIX2STRUCT before finetuning with RD on ChartQA (Table 7). MATCHA is based on PIX2STRUCT-Base but has stronger numerical reasoning and other abilities obtained through additional pretraining on relevant data. We see that RD leads to consistent improvements over stronger MATCHA models specialized for this domain.

Decoupling rationale and answer prediction. RD uses the same student model (with a single set of parameters θ) to predict rationales and answers. In Figure 4, "Another Student" refers to using a student model, with a separate set of parameters, only responsible for rationale prediction. While training separate models for predicting different intermediate steps has been shown beneficial for ScienceQA (Zhang et al., 2023), this configuration results in slightly worse performance on InfoVQA dev set. Moreover, it also adds engineering complexity, storage, and compute.

Selecting appropriate rationales is important. Instead of using a simple customized program, we construct the rationale for ChartQA by structured table concatenated with text evidence. The text evidence describes information in the figure that is relevant to the question and is predicted by PaLM 2-L given the question, answer, structured table, and OCR, but does not specify a program that can be executed to obtain the answer. For example, for the input in the lower half of Figure 2, the text evidence generated by PaLM 2-L in this setting is *"No confidence value in 2017 is 5, confidence value in 2017 is 93"*. The same RD training on evidence-based rationales achieves 83.4 / 33.0 RA on the ChartQA's augmented and human test sets, which is 1.4 / 5.0 points lower than the program-based rationales.

B.2 Comparison to other approaches

We make an additional comparison to other approaches, which may have different setups, such as the use of tools or LLMs at inference time, or the use of additional pretraining, in Table 8. We show that except the powerful pretrained model PaLI-3 (5B parameters), RD is better than other approaches under the setup of pixel-level image-to-text model without the use of external tools at inference time.

UniChart (Masry et al., 2023) is pretrained on chart-specific objectives, but on a larger corpus than MATCHA. The pretraining data is augmented by knowledge distillation from LLMs. Without further pretraining, RD shows better performance on ChartQA, initialized with MATCHA. DUBLIN (Aggarwal et al., 2023) proposes pretraining objectives at four different levels: language, image, document structure, and question-answering. It demonstrates high performance on InfoVQA and DocVQA, at the cost of sacrificing the ability to understand charts. In addition, UReader (Ye et al., 2023) designs a shape-adaptive cropping module to process high-resolution images. It is jointly fine-tuned on multiple VDU tasks with low-rank adaptation approach. Cream (Kim et al., 2023) utilizes contrastive learning to align the visual representation of the image and text representation of OCR and objects (generated from tools). We show that RD is better than or close to Cream even under the setup where Cream uses tools in inference.

UDOP (Tang et al., 2023) uses external OCR tool for text layout information at training and inference time. It is also pretrained on the IIT-CDIP scanned documents corpus, achieving great performance gains on InfoVQA and DocVQA.

B.3 Qualitative analysis

We randomly select 5 examples in the dev set of InfoVQA to illustrate that tool generated rationales extract relevant information from the visual context, which are helpful to answer the question (Table 9).

We also randomly select 20 examples from the dev set of InfoVQA for a qualitatively analysis of student generated rationales (Table 11). The first five examples are for the same inputs as the tool-generated rationale examples. We observe that for 3 examples out of 5, the student generated rationales match the tool generated ones. In the table, we list the student generated and TF-IDF extracted rationales, along with the question and the ground truth answer. We compute the TF-IDF weight for each OCR block in the image, and measure the cosine similarity of the question to these OCR blocks. Starting from the closest OCR block to the question, we gradually add more OCR blocks to the final TF-IDF string until it reaches 50 tokens under PIX2STRUCT tokenizer. Note that this process is also applied to the TF-IDF and BERT embedding analysis in Figure 4.

For more than 50% of the student generated rationales, answers can be inferred from them without looking at the images. Also, 90% of the student generated rationales are relevant to the answer. It is possible for the student model to generate an irrelevant rationale, such as in the last row of Table 11, the student rationale (*27 % fake or empty 28 % inactive 43% good*) is irrelevant to the question (*Who uses the twitterid @Ev?*) as well as the answer (*twitter co-founder evan williams*). This observation verifies the importance of robustifying against student rationale errors during training.

C Extended related works

Here we summarize related research in text only and visual language understanding, focusing on methods using intermediate reasoning steps.

Tool use in visual language understanding Using tools to augment the input in a prediction problem can be seen as using additional reasoning steps of specific type, *i.e.*, calling a tool with a set of arguments and integrating its result with the rest of the context. Much prior work on visual document understanding has relied on an OCR component (Tang et al., 2023; Appalaraju et al., 2021).

Model	Tool-use in inference	Multi-dataset fine-tuning	Prompt LLM in inference	InfoVQA	DocVQA	ChartQA
Donut	✗	✗	✗	21.7	67.5	41.8
Pix2Struct	✗	✗	✗	40.0	76.6	59.5
MatCha	✗	✗	✗	37.2	74.2	64.2
UniChart	✗	✗	✗	-	-	66.3
DUBLIN	✗	✗	✗	43.0	**80.7**	35.2
UReader	✗	✓	✗	42.2	65.4	59.3
Cream-Vicuna7B (w/o tools)	✗	✓	✓	22.1	41.1	50.0
RD (best model)	✗	✗	✗	**44.3**	79.0	**66.5**
PaLI-3 (w/o OCR)	✗	✗	✗	**57.8**	**87.6**	**70.0**
Cream-Vicuna7B (w/ tools)	✓	✓	✓	43.5	79.5	63.0
UDOP	✓	✗	✗	47.4	84.7	60.7
PaLI-3 (w/ OCR)	✓	✗	✗	**62.4**	**88.6**	69.5
DePlot	✓	✗	✓	-	-	**79.3**

Table 8: We compare the best model of RD (Pix2Struct-Large on InfoVQA and DocVQA, MatCha on ChartQA) with other existing approaches, some of them (bottom part) have different setups. We show that except the powerful pretrained model PaLI-3, RD is better than other approaches under the same setup. Red is the best model and blue is the second best.

Question	Tool Generated Rationales	GT Answer
What is the cost of a cup of coffee in Luanda and Tokyo, taken together?	Cost of a Cup Of Coffee (USD), Cost of a Cup Of Coffee (USD), $ 3.80, $ 6.65, $ 3.12, $ 8.29, $	$10.45
What are the points to be kept in mind while reading?	When you read you have to remember a lot of things, like: Characters Main plot Sub-plots.	characters, main plot, sub-plots
What will the diastolic reading be if you have High blood pressure stage 2?	High Blood Pressure (Hypertension) Stage 2, 140 or higher, or, 90 or higher, Hypertensive Crisis, Higher than 180, (Call your doctor immediately), and/or, Higher than	90 or higher
Which country has the lowest count of critical care beds, China, India, or UK?	China, 3.6, India, 2.3.	india
What is the meaning of the symbol "Hearts in Hearts" in Doodles?	Hearts in Hearts, Shy person.	shy person

Table 9: We show five randomly selected examples with tool generated rationales. The rationales are helpful to answer the question.

PaLI-X (Chen et al., 2023b) and the smaller PaLI-3 model (Chen et al., 2023c), which are image-and-text encoders paired with text decoders, achieve strong results both with and without additional OCR input. Since OCR extractions can be very long, *e.g.*, InfoVQA has images with OCR more than 1k tokens, the recognized text often needs to be truncated to a given maximal token length given pretrained model assumed token limits and efficiency considerations. Other architectures are heavily centered on the recognized document text, with examples being TILT (Powalski et al., 2021) and LayoutLM (Huang et al., 2022).

In addition to OCR, de-plotting has been used as a pre-processing step to either augment or entirely replace the input image representation (Liu et al., 2023a). Both object detection and OCR are used as an auxiliary input by Cream (Kim et al., 2023) to augment the vision feature.

Such works have not attempted to recognize text or structured data as an intermediate reasoning step using the same small model, as we have proposed in RD.

Tool use and chain-of-thought distillation Distilling text rationales from large teacher models has been shown successful by chain-of-thought distillation works (Shridhar et al., 2023; Li et al., 2023; Wang et al., 2023b) on NLP benchmarks, such as CommonsenceQA (Talmor et al., 2019) and QuaRel (Tafjord et al., 2018).

MMCoT (Zhang et al., 2023) and T-Sci (Wang et al., 2023a) have utilized annotated or decomposed reasoning chains for improving vision-language reasoning on ScienceQA, which is not representative of the visual document understanding challenges we focus on (*e.g.* text-only models can reach accuracy of over 79% on this benchmark). In addition, these works only distill using our QRA task, which we show is insufficient to teach the student model to produce high-quality rationales and be robust to potential errors. We also use a single small model instead of two different models for ra-

Method	FLOPs
PIX2STRUCT-Base, Ans-only	2.62E+12
PIX2STRUCT-Base, RD	2.65E+12
PIX2STRUCT-Large, Ans-only	4.63E+12
PIX2STRUCT-Large, RD	4.72E+12
PaLI-3, w/o OCR	4.81E+13

Table 10: FLOPs of evaluated approaches. RD only increase the FLOPs of Base model by around 1%, Large model by 2%, and uses less than 10% the FLOPs of the SOTA model.

tionale and answer generation, reducing complexity and engineering cost, and focus on short rationales for efficiency. Finally, we use a broader set of tools instead of just one LLM chain-of-thought tool.

Toolformer (Schick et al., 2023) trains smaller language models to call tools. Generic multimodal tool use solutions based on LLMs have also been proposed (Yang et al., 2023). However, these works do not replicate the results of tool output and replace them for efficiency.

Other related work on text-only models with intermediate reasoning steps Intermediate reasoning in text-only models has been successful through prompting large language models to perform a chain-of-thought (Wei et al., 2022). More traditionally in NLP, smaller models have been shown to be able to successfully learn to generate semantic parses before predicting final answers, including when such parses are not directly annotated in training data (Yih et al., 2016). Decomposing intermediate questions is also known to help small models on multistep text question answering (Zhu et al., 2023). Marginalizing over multiple intermediate rationale possibilities has brought consistent gains (Wang et al., 2023c).

The specific structure of reasoning chains (which can be guided by tailored prompting strategies for LLMs) used has been shown to have significant impact (Zhou et al., 2023; Khot et al., 2023; Yao et al., 2023). In addition to text as intermediate predictions, generating programs has also been shown useful (Chen et al., 2023a).

D Detailed FLOPs analysis

We show that RD only increases the FLOPs of the Base model on InfoVQA by around 1%, those of the Large model by around 2%, and uses around 10% the FLOPs of the SOTA model, as listed in Table 10.

We only consider the computation of transformer blocks of the encoder and the decoder, and ignore the small cost in the last linear layer for token generation. Most of the computation cost is from the attention and feed-forward layers, and we ignore the activation and normalization layers. Notice that matrix multiplication of with dimension $[N, P] \times [P, M]$ uses FLOPs of $NM(2P-1)$; for simplicity, we use $2NMP$ to approximate.

For each self-attention layer, we suppose an input sequence length of d_q, a hidden size of d_h. The query, key, value matrix computation takes $6d_q d_h^2$, the multiplication of these three matrices takes $4d_q^2 d_h$, and the linear transformation towards the output takes $2d_q d_h^2$. The total is $8d_q d_h^2 + 4d_q^2 d_h$.

For each cross-attention layer, we suppose a query input sequence length of d_q, and a key-value input token sequence of d_k. The query, key, value matrix computation takes $2d_q d_h^2 + 4d_k d_h^2$, the multiplication of these three matrices takes $4d_q d_k d_h$, and the linear transformation towards the output takes $2d_q d_h^2$. The total is $4d_q d_h^2 + 4d_k d_h^2 + 4d_q d_k d_h$.

For one feed-forward layer, suppose the sequence length from the attention layer is d_q and the hidden size from the attention layer is d_h and the feed-forward size is d_f, the total computation is $6d_q d_f d_h$ if gated activation is used, otherwise $4d_q d_f d_h$.

Now we derive the formula of FLOPs for encoder-decoder models. We use d_e and d_d to denote the encoder sequence length, and the whole decoder sequence length, respectively. Given the models we discuss here all have same hidden dimension for the encoder and the decoder, we use d_h to denote the hidden size and d_f to denote the feed-forward size. For simplicity, we assume a batch size of 1. The computation cost of each encoder layer, denoted with FCE, is

$$\begin{aligned}\text{FCE}(d_e, d_h, d_f) &= 8d_e d_h^2 + 4d_e^2 d_h + 4d_e d_f d_h \\ &+ 2[\![\text{Gated}]\!] d_e d_f d_h,\end{aligned}$$

where $[\![\text{Gated}]\!]$ is the indicator function on whether the model uses gated activation. Similarly, without caching the past attention matrices, the computation cost of each decoder layer, denoted with $\text{FCD}_{\text{exact}}$, is

$$\begin{aligned}\text{FCD}_{\text{exact}}(d_e, d_d, d_h, d_f) &= 4d_e d_h^2 + \sum_{t=1}^{d_d} 4d_h t^2 \\ &+ (12d_h^2 + 4d_e d_h + 4d_f d_h + 2[\![\text{Gated}]\!] d_f d_h)t.\end{aligned}$$

Notice the query, key, value matrices from the encoder output only have to be constructed once through the decoding time steps.

Instead, if we consider KV-caching and reusing the past attention matrices in the decoding (Pope et al., 2023), we can achieve the following at step t:

- reuse the first $t-1$ rows of the query, key, value matrices;
- reduce the matrix multiplication cost by a factor of t with block matrix computation;
- for both self-attention and cross-attention, we only have to care about the last row of the output matrix.

Given the decoding with caching, we reduce the computation cost of each decoder layer to $\text{FCD}_{\text{approx}}$, written as

$$\text{FCD}_{\text{approx}}(d_e, d_d, d_h, d_f) = 4d_e d_h^2 + \sum_{t=1}^{d_d} 4d_h t$$

$$+ (12d_h^2 + 4d_e d_h + 4d_f d_h + 2[\![\text{Gated}]\!] d_f d_h).$$

This formula matches the one provided by Elbayad et al. (2020). For a N-layer encoder-decoder model, the total computation cost is $N(\text{FCE} + \text{FCD}_{\text{approx}})$ and $N(\text{FCE} + \text{FCD}_{\text{exact}})$ with and without caching, respectively.

Based on the formula derived above, we start to compute FLOPs for specific models. Taking InfoVQA as an example, the student generated rationales have 41.8 tokens on average, the questions have 15.3 tokens on average and the answers have 5.0 tokens on average.

PIX2STRUCT-Base The model has $N = 12$, $d_h = 768$, $d_f = 2048$ and uses gated activation. For InfoVQA, we have $d_e = 6155$, $d_d = 5$ for answer-only generation, and $d_d = 62$ for RD generation (including the question, rationale, and the answer). Without caching, the total FLOPs computation is 2.63E+12 for answer-only generation, and 3.46E+12 for RD generation, resulting in a $\sim 30\%$ increase of computation. With caching, the total FLOPs computation is 2.62E+12 for answer-only generation, and 2.65E+12 for RD generation, resulting in a only $\sim 1\%$ increase of computation.

PIX2STRUCT-Large The model has $N = 18$, $d_h = 1536$, $d_f = 3968$ and uses gated activation. Similarly, for InfoVQA, we have $d_e = 3072$, $d_d = 5$ for answer-only generation, and $d_d = 62$ for RD generation. With caching, the total FLOPs computation is 4.63E+12 for answer-only generation, and 4.72E+12 for RD generation, resulting in a only $\sim 2\%$ increase of computation.

PaLI-3 We also estimate FLOPs for PaLI-3 (Chen et al., 2023c), which is constructed by a 2B ViT-G/14 vision encoder and a 3B UL2 language encoder-decoder.

The vision encoder has $N = 48$, $d_h = 1536$, $d_f = 8192$, and does not use gated activation. For evaluating on InfoVQA, the model uses the resolution of 1064×1064, which has $d_e = 5776$ patches. The FCE formula gives the computation cost of 2.90E+13.

The language encoder-decoder has $N = 24$, $d_h = 1024$, $d_f = 16384$, and uses gated activation. We consider the extra text tokens (15 on average) from the question but not the ones from the OCR input. Hence, we have $d_e \geq 5791$ and $d_d = 5$. With caching, the total computation cost of the UL2 language transformer is at least 1.91E+13.

Combing two parts, the 5B PaLI-3 model uses FLOPs of at least 4.81E+13 on the setup of the InfoVQA task, which is 10 times more than PIX2STRUCT-Large with the RD generation.

E Prompt templates

We list the prompt templates for rationale generation on InfoVQA, DocVQA and ChartQA in Fig. 5, Fig. 6 and Fig. 7, respectively. The former two use 5-shot prompting for LLM-Summarizer and the last uses 8-shot prompting for LLM-Programmer.

Question	Student Generated	TF-IDF	GT Answer
What is the cost of a cup of coffee in Luanda and Tokyo, taken together?	1, LUANDA, Angola, 4,799,432, $ 7,000, $ 3.80, $ 4.93, $ 17.90, $ 0.55, 2, TOKYO, Japan, 12,369,000, $ 4.436,	Cost of Monthly, Cost of a, Rent Luxury Cup Of Coffee, 2 Bedroom (USD), (USD), This new found excess wealth has attracted many, Auxury retailers, hotels and restaurants to full indulgences	$10.45
What are the points to be kept in mind while reading?	READING, LISTENING, DRINKING TEA, TO MUSIC, OR COFFEE, WALKING, 68 %, 61 %, 54 %, 42 %, 68 %, 61 %, 54 %, 42 %, Reduces stress, Just 6 minutes of reading :, 68 %, 61 %, 54 %	READING, Alters your, state of mind, Reading develops theory of mind , which leads to better social connections ., Just 6 minutes of reading :, Why?, • It helps	characters, main plot, sub-plots
What will the diastolic reading be if you have High blood pressure stage 2?	High Blood Pressure, 140 or higher, or, 90 or higher, 2 TRACK LEVELS, American Heart Association., Check. Change. Control., Health care providers can take blood pressure readings and provide recommendations. Check. Change. Control. helps you track your progress in reducing blood pressure., Check., Change.,	High Blood Pressure (Hypertension) Stage 1, High Blood Pressure (Hypertension) Stage 2, Elevated Blood Pressure, BLOOD PRESSURE CATEGORY,	90 or higher
Which country has the lowest count of critical care beds, China, India, or UK?	China, 3.6, India, 2.3	India O, China, The Countries With The Most Critical Care Beds Per Capita, Total number of critical care beds per 100,000 inhabitants in selected countries *, Sources : National	india
What is the meaning of the symbol "Hearts in Hearts" in Doodles?	Hearth in Hearts Shy person	Hearts in Hearts, Shy person, Hearts Romantic person, Hearts with Arrow Daydreamer , idealistic person, What Do Your Doodles Mean?, Doodles can be a window into your inner thoughts	shy person
What was the number of factory workers in the confederate states during the American Civil War?	111K, 70K, 9K, 21K, 9K, 1.7K, 9K, Factories, Factory workers, Miles of Railroad, MILITARY	Factory workers, X, CIVIL WAR, BORDER STATES CONFEDERACY, Prisoners of War, African American :, Native, 178,975 American	111k
How many countries are hosting the 2015 ICC Cricket World Cup?	2 COUNTRIES Australia and New Zealand - hosting the World Cup 2015	COUNTRIES Australia and New Zealand - hosting the World Cup 2015, 3, Teams participating in the World Cup, ICC CRICKET WORLD CUP, 2015, AUSTRALIA	2
Which of these countries is least corrupt - Great Britain, China or Mexico?	GREAT BRITAIN, $ 37,500, RUSSIA $ 18,000, MEXICO $ 35,950, GREAT BRITAIN, $ 37,500	CHINA GREAT 2.6 % BRITAIN, 2.5 %, MEXICO, 35.9 % CHINA S, CORRUPTION INDEX, (OUT OF 100	great britain
How many points did Shaq score in 2000?	49 %, 47 %, 47 %, 13, 22 25 32 33 34 42 44 52, On Tuesday night, Shaquille O'Neal's number 34 will become the 9th retired number raised to the rafters at STAPLES Center. Here's a unique look at the intriguing	POINTS, POINTS, 2000/2001, 1999/2000, fff, 2001, 2002	2,344
How many countries have number of critical care beds less than 5?	United States, Germany, Italy, France, South Korea, Spain, Japan, United Kingdom, United States, 34.7, 29.2, 12.5, 11.6, 10.6, 9.7, Japan, 7.3, 6.6, 6.6, China, 3.6, India, 2.3	The Countries With The Most Critical Care Beds Per Capita, Total number of critical care beds per 100,000 inhabitants in selected countries *, Sources : National Center for Biotechnology Information, Inten	2
What percentage of women find video ads really annoying?	80 %, find video ads really annoying	80 %, find video ads really annoying, % women who watch online video, Majority of women watch online video in the afternoon or evening, 47 % watch video for up to 10 minutes a	80%

In 2009, how many pedestrian men died?	In 2009, 157 Pedestrian Deaths, http://www.nj.gov/njsp/info/fatalace/2009_fatal_crash.pdf, MALE :, 112, FEMALE :, 45, MALEP: 45,	Pedestrian Deaths in Southern New Jersey Look Both Ways Before You Cross, In 2009 , 157 Pedestrian Deaths, Between 2007 and 2009 the highest	112
What percentage of clothing and consumer electronic products of men photographed by mobile shoppers, taken together?	22 %, 22 %, 32 %, 18 %, 4 %, 5 %, 13 %, 2 %, 20 %, 30 %, PRODUCTS PHOTOGRAPHED BY MOBILE SHOPPERS, 15 %, At work, 25 %, 12 %, In the	PRODUCTS PHOTOGRAPHED BY MOBILE SHOPPERS, Consumer Clothing electronics, MEN, WHERE MEN AND WOMEN DO, THEIR MOBILE SHOPPING, TYPES	44%
What is the value of New York Knicks?	NEW YORK KNICKS $ 3.30B	NEW YORK KNICKS $ 3.30B, NEW YORK METS $ 2.00B, NEW YORK GIANTS $ 3.10B, NEW YORK YANKE	$3.30b
How much more is the value of Barcelona FC when compared to Real Madrid ($bn)?	BARCELONA FC $ 3.64B, NEW YORK KNICKS $ 3.30B, LOS ANGELES LAKERS $ 3.00B, CHICAGO BULLS $ 2.60B, GOLDEN STATE WARRIORS $ 2.60B, CHICAGO BULLS $ 2.50B, BRO	REAL MADRID, $ 3.58B, BARCELONA FC, $ 3.64B, A mountain of sponsorship and advertising cash keeps Man U king of the soccer castle, though Barcelona	0.06
Which is the second last tip for staying healthy?	Don't touch your, face, Avoid close contact with someone who's, sick, Clean and disinfect surfaces and objects people frequently touch	Tips for staying healthy, ON, What to do if you feel sick, Stay home, Most people with COVID - 19 have mild to moderate symptoms and can recover at home. Rest up and prevent germs from spreading by staying home	wear a cloth face mask in public
What percent of adults in age group 65+, buy their food based on the 'availability of nutritious food'?	33 %, 28 %, 21 %, 32 %, 11 %, 17 %, 15 %, Making it easier for the 50+ to eat more nutritious foods, i, 56 %, Help find information on fruits & vegetables, Source : AARP Foundation : Food Insecurity	Food Availability, AARP®<unk>, FOUNDATION, A recent AARP Foundation survey of 1,000 low - income adults age 50+ reveals that, in the past 12 months, two in	15%
Who provides statements for the presentencing investigation report?	ANALYSIS OF LEGAL HISTORY, ANALYSIS OF LEGAL HISTORY, OI, Snapshot of the DV Criminal History including, Domestic Incident Report (DIR) history, How many arrests in DV related crimes? Convictions?, • Stalking history, • Protective orders?, • Level of compliance if under supervision before?, • Current release status, • Jail days credited, Domestic Incident Report (DIR) history, •	THINGS TO INCLUDE WHEN CREATING A PRESENTENCING INVESTIGATION REPORT, • Arrest Report / DIR • Depositions Summary of Witness Statements, Review Police report	arresting officer, victim
What happened first; Gaza conflict or Scottish independence?	GAZA CONFLICT August 1 : 64K Peak Shares	GAZA CONFLICT August 1 84K Peak Shares SCOTTISH INDEPENDENCE September 14 35K Peak Shares, CRIMEAN INDEPENDENCE March 17	gaza conflict
Who uses the twitter id @Ev?	27 % fake or empty 28 % inactive 43 % good	Twitter co - founder Evan Williams @Ev, WHOLESALERS, IN DARK CORNERS OF THE INTERNET , THEY PLY TOOLS TO OVERRIDE TWITTER'S RULES, THE	twitter co-founder evan williams

Table 11: We show 20 random selected examples with student generated or TF-IDF extracted rationales. The first 5 examples are the same as in Table 9, where 60% of student generated rationales match the tool generated ones. For more than 50% of the student generated rationales, answers can be inferred from them without looking at the images. 90% of the student generated rationales are relevant to the answer, others are irrelevant.

Please extract the relevant evidence of the QA from the OCR string for the last examples. The evidence should be within 50 tokens.

OCR string from image: H, EVOLUTION OF THE SKATEBOARD, 1940, 1959 1960 1964 1970, 1975, 1980, 1990, 2000, SIDEWALK SURFBOARDS The first skateboards started with wooden boxes , or boards , which kids added roller skate wheels to in the late 40's and early 50's ., ROLLER DERBY SKATEBOARD The Roller Derby Skate Company was the company who coined the name skateboard . They were the first company to mass produce the Roller Derby skateboard . Their factory was in La Mirada , CA. By 1959 , people could purchase . the boards nationwide at Roller Derby arenas ., NASH SHARK In the 1960's , another company by the name of NASH came out with their own skateboards , and they called it the Shark . Today it's known as the Nash Shark Skateboard ., GANDS FIBERFLEX PINTAIL In 1964 , the G & S FiberFlex Pintail was born . It was made by surfers for surfers . G & S stand for Larry Gordan and Floyd Smith . In the 60's, these guys became one of the largest and most succesful skateboard companies ., BANANA BOARD In the mid 1970's , a new board hit the streets . It was called the Banana board . The Banana boards are skinny , flexible boards made out of polypropylene that have ribs on the underside for structural support ., ROAD RIDER CRUSIER In 1975 Road Rider came out with the first ever skateboard that had precision bearings made just for skateboards . This would bring an end to decades of loose ball bearings ., OLD SCHOOL FISHTAIL In the 1980's , skateboards changed for vert skaters . The ideal board to ride vert was the Fishtail deck . People still skated street with these short nosed , wide vert , soft wheeled boards ,, POP SICKLE, POP SICKLE, In the 1990's . skateboarding started focusing more on street skateboarding . Most boards are 7 1/4 to 8 in and 30-32 inches long with a largely symmetrical shape with a relatively narrow width ., The board hasn't changed much from the 90's til now , but the concave may be a little deeper . However , people are starting to ride their own custom shaped boards more and more !
Question: when was nash shark introduced?
Answer: 1960
Evidence: NASH SHARK. In the 1960's, another company by the name of NASH came out with their own skateboards, and they called it t
he Shark. Today it's known as the Nash Shark Skateboard.

… (omit two examples)

OCR string from image: State, Government, Chad Foust FIVE, [great], Reasons to hire me as Art Director , PRESENTATION, [reason : five], +, TENT, years experience, 2001, 2002, 2003, 2004 2005, 01 02 03 04 05, creating beautiful presentation design, for, Community Groups, Direct Marketing Sales (B2B), 2006, 2007, 2008, 2009, 2010, 2011, 07 06, 08, 09, 10, ww, Real Estate Ventures, Non - Profit Sector, Youth Camps O, [reason : four], Motion Graphics, +, FIVE years DIRECTING creative teams, Lower Thirds, Loremipsum dolor sit amet , consectetur ad pisicing elit , sed do eiusmod tempor incidic, 28, 28, 34, videographers, photographers, 19, 21, set, designers, dancers, musicians, graphic designers ,, singers / vocalists, tech personnel, dramatists, [reason : three], 3xtensive public speaking , PRESENTATION , & performance 3xperience ., in small teams of 11, MEDIUM GROUPS OF 350, 3, AND, LARGE CROWDS UP TO, multiple software, [reason : one], [reason : two], proficiencies, 2.898, Yours of profession, experience, Prezi | 1 Keynote 12 ProPresenter | 2 InDesign 2 MediaShout | 3 Illustrator13 After Effects 13 Flash 4, Dreamweaver 6 Photoshop 8 PowerPoint 10 (and many more), M, T, T, M, W, W, Th, I OFFER YOU 133 %, Some give 110 %, Th, T1ME, >> to make, Whatever it takes, the company, successful ,, the client, satisfied , and, the, competition weep ., Integrated skill , knowledge , and demonstrated leadership across >> multiple creative, 3XPERTISE disciplines . 3NTHUSIASM, BONUS QUALIFICATIONS : Video editing and motion graphics ✓ Web design , XHTML , interactive experience, Strong writing skills ✓ Infographic design ✓ Flash , animation * sorry , I'm a terrible photographer, Excitement , Energy , Excellence , Initiative, Chad Foust Art Director & Designer design@chadfoust.com 734.775.2427, © Copyright 2011 Chad Foust / colordrive.net / chadfoust.com
Question: Which is the second biggest category of creative teams Chad Foust has directed?
Answer: dramatists
Evidence: videographers, photographers, set designers, dancers, musicians, graphic designers, singers / vocalists, tech personnel, dramatists.

OCR string from image: וייד, DIY GIFT IDEAS, Tea Wreaths Stripped Umbrellas, This unique wreath is perfect for any tea - lover you know . What you'll need, Two pieces of 12x12ish cardboard, Clothes, pin, Ribbons for hanging, Patterned paper, Hot glue, Turn a blah umbrella into a stylish accessory in no time . What you'll need, An umbrella, Painter's tape, Foam brush, Paint, Leather Pouch A one - of - a - kind gift that only costs $ 15 to make . What you'll need, A pouch template Fabric Scissors, Ruler, Pencil, Ball Head, Screw Studs, Sewing Machine / Thread, Pin Shears, Permanent Paint Marker, Collegiate Scarf Forget the college bookstore - you won't even need to leave home to make this spirited gift . What you'll need, Bull - dog clips, A Scarf, Patch of your choice, Shower Curtain Instagram Cards, Hand - embroided shower curtain will turn any bathroom into a fun and relaxing oasis . What you'll need, Shower Curtain Medium Gauge Yarn, Ruler, Pencil, Disappearing Ink Marker, Scissors, Print special memories you've captured on your Instagram and celebrate cards . What you'll need, Large Yarn Needle with Sharp Point, Photos of your choice, Graph Paper, Printer, Fabric, These key - chains inexpensive stocking stuffers . What you'll need, Fabric, Scraps Medium Weight Iron on Infefacing, Key Rings, Pinking Shears Small Piece of One - sided iron on interfacing Twill tape or grosgrain ribbon, Buttons , felt , for embellishing Thread , sewing stuff, Tie Dye T - Shirts, CUSTOM T - SHIRTS, 1. CHOOSE A COLOR PALETTE , SUCH AS BRIGHT COLORS OR EARTHY MUTED TONES , TO TRANSFORM YOUR PLAIN WHITE TEE ., Custom T - Shirts, 2. BE READY TO DYE WITH RUBBER DISH WASHING GLOVES TO PROTECT YOUR HANDS , A BIG ROD OR SPOON TO STIR WITH , RUBBER BANDS OR STRING TO TIE CLOTHING WITH , AND A BIG HEAT - RESISTANT TUB TO DO THE DYING IN ., 3. BUY INEXPENSIVE ONE - STEP DYE BRANDS AT MANY GROCERY , FABRIC AND CRAFT STORES ., 4. COLOR YOUR FABRIC ALL AT ONCE BY MIXING THE DYE IN VERY HOT WATER IN YOUR TUB AND SUBMERSING YOUR T - SHIRT UNTIL YOU GET A COLOR TWO SHADES DARKER THAN YOU WANT THE FABRIC WILL BE A LIGHTER COLOR WHEN DRY , THEN RINSE IN COLD WATER UNTIL THE WATER SQUEEZED OUT IS CLEAR ., DIRECT TO GARMENT INK JET DIGITAL PRINTING IS FANTASTIC AND COST EFFECTIVE, 2222, 5. DYE YOUR SHIRT A LIGHT COLOR , ADD MORE TIES , AND THEN DYE A DARKER COLOR FOR A MULTI - COLORED LOOK ., Sources :, DRAW PAINT T - SHIRTS, DRAW , WRITE , AND DOODLE DIRECTLY ON YOUR SHIRT WITH SPECIALLY FORMULATED FABRIC MARKERS .,
http://newlyweds.about.com/od/Anniversaries/tp/Diy-Gifts-For-Your-Spouse.html
http://kojo-designs.com/2010/03/kojotutorial-tea-tea-tea-kitchen-wreath/ http://www.styleoholic.com/diy-fashionable-striped-umbre
lla/#sthash.BDh5Kjrs.dpuf, http://www.designlovefest.com/page/4/?s=No+sew, http://www.craftinessisnotoptional.com/2011/06/scrap-yo
ur-stash-guest-post-living-with.html http://www.huffingtonpost.com/2013/12/14/45-diy-gift-ideas_n_4442662.html?utm_hp_ref=diy-gift
-ideas http://www.ehow.com/way_5306117_diy-custom-tshirts.html#ixzz20iQG6KUR, http://www.coastalprintworks.com, Coastal Printworks Museum Quality Screenprinting Coastal Printworks.com
Question: which t-shirt has a smiley drawn on it?
Answer: paint t-shirts
Evidence: DRAW PAINT T-SHIRTS. DRAW, WRITE, DOODLE DIRECTLY ON YOUR SHIRT WITH SPECIALLY FORMULATED FABRIC MARKERS.

OCR string from image: [[ocr]]
Question: [[query]]
Answer: [[answer]]
Evidence:

Figure 5: InfoVQA prompt template.

Please extract one or two sentences within 50 tokens from the OCR string as the evidence to answer the question.

OCR string: B & W, BROWN & WILLIAMSON TOBACCO CORPORATION RESEARCH & DEVELOPMENT, TO :, R. H. Honeycutt, CC :, T.F. Riehl, FROM :,C. J. Cook, DATE :, May 8 , 1995, SUBJECT :, Review of Existing Brainstorming Ideas / 483, INTERNAL CORRESPONDENCE, The major function of the Product Innovation Group is to develop marketable novel products that would be profitable to manufacture and sell . Novel is defined as : of a new kind , or different from anything seen or known before . Innovation is defined as : something new or different introduced ; act of innovating ; introduction of new things or methods . The products may incorporate the latest technologies , materials and know - how available to give then a unique taste or look ., The first task of the Product Innovation Group was to assemble , review and categorize a list of existing brainstorming ideas . Ideas were grouped into two major categories labeled appearance and taste / aroma . These categories are used for novel products that may differ from a visual and / or taste / aroma point of view compared to conventional cigarettes . Other categories include a combination of the above , filters , packaging and brand extensions ., Appearance, This category is used for novel cigarette constructions that yield visually different products with minimal changes in smoke chemistry, • Two cigarettes in one . Multi - plug to build your own cigarette . Switchable menthol or non menthol cigarette.
Question: Who is in cc in this letter?
Answer: T.F. Riehl
Evidence: TO :, R. H. Honeycutt, CC :, T.F. Riehl, FROM :, C. J. Cook.

OCR string: :, Confidential RJRT PR APPROVAL, DATE :, SUBJECT :, 1/8/93 · Lu glas PROPOSED RELEASE DATE :, FOR RELEASE TO : CONTACT : P. CARTER, for response, ROUTE TO I, Home, Peggy Carter, Maura Payne, David Fishel Tom Griscom Diane Barrows, Ed Blackmer, Tow Rucker, Initial, Ace, out, OB7, tus ., TYR, Return to Peggy Carter , PR , 16 Reynolds Building, Date, 1/8/93, Source : https://www.industrydocuments.ucsf.edu/docs/xnbl0037, 51142 3977
Question: what is the date mentioned in this letter?
Answer: 1/8/93
Evidence: DATE :, SUBJECT :, 1/8/93 · Lu glas, Date, 1/8/93

OCR string: DOMESTIC PRODUCT DEVELOPMENT (cont'd .), Project Marlboro, - POL 0330 -1.6 tar / puff - 80mm has been produced and currently is in C.I. for analytical ., - POL 0331 - 1.6 tar / puff - 84mm was produced 6/1/90 . Samples have been submitted to C.I., - Marlboro Double Batch - RL & RCB was produced 6/4/90 . Samples have been submitted for analytical testing ., - POL 3634 - RL Evaporator Upgrade - Scheduling for primary at the M / C has been completed . Fabrication is scheduled for the week of 6/18/90 in Semiworks ., Marlboro Menthol, Marlboro Menthol 80mm and 83mm were subjectively smoked by the Richmond Panel . After further review of the data and specifications , another model of the 83mm with zero ventilation will be made at Semiworks within the next 2-3 weeks ., Bucks, Bucks K.S. Lights and Full Flavor with various aftercut modifications were smoked by the Richmond Panel . Particular models were selected from the group and POL testing will be done on these prototypes ., Miscellaneous, Additional tipping papers of Marlboro Lights have been received and currently are being analyzed for lip release coatings . Cigarettes will be produced and submitted to O / C Panel for evaluation of lip release ., 3 :, Source : https://www.industrydocuments.ucsf.edu/docs/khxj0037, 2022155853
Question: what mm Marlboro Menthol were subjectively smoked by the Richmond Panel
Answer: 80mm and 83mm
Evidence: Marlboro Menthol, Marlboro Menthol 80mm and 83mm were subjectively smoked by the Richmond Panel .

OCR string: SFE - GC were also demonstrated in quantitative measurements of phenolics in woodsmoke analysis . W. T. Foreman (U.S. Geological Survey , CO) extracted the C. , cartridge with SFE to recover pesticides in high yield ., DETERMINATION OF POLAR VOLATILE ORGANICS (PVOC) IN AMBIENT AIR, The polar compounds are those containing hetero - atoms such as nitrogen , sulfur and oxygen . The single most difficult problem in developing protocols for analyzing polar compounds at trace level in air is probably moisture . Sampling of sidestream smoke components shared similar difficulty . The moisture in the ambient air clogged up the cryogenic trap and prevented sample enrichment . The evaporation of water vapor in the source of the mass spectrometer interfered with the high vacuum and the detection of co - eluting compounds . The present EPA TO - 14 method requires the use of Naphion dryer to eliminate water . Unfortunately , the Naphion tube is also permeable to many polar compounds such carbonyls and alcohols . Method TO - 14 with canister sampling is only for nonpolar organic compounds , e.g. aromatics and hydrocarbons ., Source : https://www.industrydocuments.ucsf.edu/docs/qhxj0037, 2022155945
Question: Which hetero-atoms does polar compounds contain?
Answer: nitrogen, sulfur and oxygen.
Evidence: The polar compounds are those containing hetero - atoms such as nitrogen , sulfur and oxygen .

OCR string: CUT TOBACCO :, BLEND :, MT - 768 D BST391 BW - 6071, BEST PROTOTYPE , 327391, LBS AT 12.5 %, SOLID LBS, LBS AT TARGET, STRIPS : FLUE CURED ., 3.681.7, 3,221.5, 3,790.0 @ 15.0 %, BURLEY ..., 1,996.3, (1,746.8), + CASING (S), 2,159.0, 2,540.0 @ 15.0 %, ORIENTAL ..., 1,243.4, 1,088.0, 1,280.0 @ 15.0 %, RECONSTITUTED ., 2,321.7, 2,031.5, 2,390.0 @ 15.0 %, TOTAL STRIPS ..., 9,243.1, 8,500.0, 10,000.0 @ 15.0 %, Source : https://www.industrydocuments.ucsf.edu/docs/lycj0037
Question: What is the LBS AT TARGET of TOTAL STRIPS?
Answer: 10,000.0 @ 15.0 %
Evidence: TOTAL STRIPS ..., 9,243.1, 8,500.0, 10,000.0 @ 15.0 %

OCR string: [[ocr]]
Question: [[query]]
Answer: [[answer]]
Evidence:

Figure 6: DocVQA prompt template.

Please generate the program as the intermediate step to answer the question based on the OCRs and tables. The tables show the layout of the plot, but the numbers may be inaccurate or incomplete. Please check if these numbers appear in the OCR; if not, please ignore them in the tables.

The only available functions of the programs are
Div(a,b); Mul(a,b); Avg(a list of numbers); Sum(a list of numbers); Diff(a,b); Greater(a,b); Less(a,b); Find(str).

OCR: Public Expects Political Division to Persist Level of nation's political division in five years will be ..., Don't, know, More, Same, 36 %, 41 %, Less, 5 % 17 %, Survey conducted Dec. 3-7 , 2014 . PEW RESEARCH CENTER
Table: Entity,Value | loss,517 | Same,41 | More,36 | Less,17
Question: What is the difference in value between Same and sum of More and Less?
Answer: 12
Program: Diff(41, Sum(36, 17))
Execution: Diff(41,(36+17))=Diff(41-53)=|41-53|=12

OCR: T - Series, YouTube Movies, Music, Cocomelon - Nursey Rhymes, PewDiePie, SET India, Gaming, 89.2, Kids Diana Show, 79.1, WWE, Sports, Additional Information, 0, 77.6, 75, 25, 25, 50, 75, 115, 112, 110, 105, 100, 137, 125, 183, 150, 175, 200, 225, Number of subscribers in millions, *, 155, 59, © Statista 2021, Show source
Table: Characteristic,Number of subscribers in millions | T-Series,183.0 | YouTube Movies,137.0 | Music,115.0 | Cocomelon - Nursey Rhymes,112.0 | PewDiePie,110.0 | SET India,105.0 | Gaming,89.2 | Kids Diana Show,79.1 | WWE,77.6 | Sports,75.0
Question: What's the average number of subscribers of the most 3 popular Youtube channels?
Answer: 145
Program: Avg(183.0, 137.0, 115.0)
Execution: (183.0+137.0+115.0)/3=145.0

OCR: Overwhelming Majority of Russians Say Breakup of USSR Was Bad for Russia Do you think the dissolution of the Soviet Union was a good thing or bad thing for Russia ?, Good, thing, 17 %, Don't, Bad, know, 14 %, thing, 69 %, Source : Spring 2015 Global Attitudes survey ., Q34 ., PEW RESEARCH CENTER
Table: Entity,Value | Bad thing,69 | Good thing,17 | Don't know,14
Question: What is the percentage of Don't know in the chart?
Answer: 14
Program: Find(percentage of Don't know)

OCR: In Canada , only a quarter of the public has confidence, in Trump Among Canadians ..., 100 %, Favorable view of the U.S., 72 a 63, 59, 59, 40, 88, 83, 81, 76, 68, 68, 65, 64, 43, 55, 39, 28 Confidence in U.S. president, 25, 22, 0 2002, 2006, 2010, 2014, 2018, Bush, Obama, Trump, Source : Spring 2018 Global Attitudes Survey . Q17a & Q35a ., PEW RESEARCH CENTER
Table: Year,Confidence in U.S. president,view of the U.S. Favorable | 2002,59,72 | 2006,40,59 | 2010,88,68 | 2014,81,64 | 2018,25,39
Question: Is the average of highest and lowest value of green bar greater than 80?
Answer: No
Program: Greater(Avg(72, 39),80)
Execution: Greater((72+39)/2,80)=Greater(55.5,80)=55.5>80? No

OCR: Pakistanis Say It's Important to Educate Both Girls and Boys Education is more important for ..., Boys and girls equally 86 %, 7 % 5 %, 2 %, Don't, Boys, Girls, know, Source : Spring 2014 Global Attitudes, survey ., PEW RESEARCH CENTER
Table: Entity,Value | Boys and girls equally,86 | Girls,5 | Boys Girls,75 | Don't know,2
Question: Take sum of three smallest segment, multiply it 5, is the result greater than largest segment?
Answer: No
Program: Greater(Mul(Sum(7, 5, 2), 5), 86)
Execution: Greater(Mul(7+5+2,5),86)=Greater(Mul(14,5),86)=Greater(14*5,86)=Greater(70,86)=70>86? No

OCR: Americans Give China Mostly Negative Ratings, U.S. views of China, 80 %, 43, 52, 35, 0, 2005, Unfavorable, 55, 54, 51, 52, 50, 49, 42, 42, 40, 40, 39 39, 38, 36, 36, 37, 38, 35, 29, 2007, 2009, Source : Spring 2015 Global Attitudes survey . Q12b ., PEW RESEARCH CENTER, Favorable, 2011, 2013, 2015
Table: Year,Favorable,Unfavorable | 2005,0,35 | 2007,5250,39 | 2009,50,38 | 2011,49,36 | 2013,35,52 | 2015,38,54
Question: How many values are below 40 in Unfavorable graph?
Answer: 6
Program: Find(count of values below 40 in Unfavorable graph)

OCR: How often people interact with people of other races , ethnicities varies widely, % who say they race or ethnicity, interact with people of a different, Never / Rarely, Occasionally /, Frequently, India, 27 %, 66 %, South Africa, 34, 66, Venezuela, 40, 60, Lebanon, 40, 57, Colombia, 46, 53, Jordan, 48, 51, Kenya, 48, 51, Tunisia, 59, 40, Philippines, 61, 38, Vietnam, 64, 33, Mexico, 69, 30, Note : Don't know responses not shown . Source : Mobile Technology and Its Social Impact Survey 2018 ., Q38b ., " Attitudes Toward Diversity in 11 Emerging Economies ", PEW RESEARCH CENTER
Table: Entity,Never/Rarely,Occasionally) Frequently | Mexico,69,30.0 | Philippines,61,38.0 | Kenya,48,nan | Jordan,48,51.0 | Colombia,46,53.0 | Lebanon,40,nan | Venezuela,40,60.0 | South Africa,34,66.0 | India,27,66.0
Question: Is the median of the green bar smaller than the median of the blue bar?
Answer: No
Program: Less(51, 48)
Execution: 51<48? No

OCR: ·, No 65.88 %, -, *, Yes 34.12 %, <, 99, di, Additional Information, © Statista 2021, Show source
Table: Characteristic,Share of respondents | Yes,34.12% | No,65.88%
Question: What is the ratio of yes to no?
Answer: 0.518
Program: Div(34.12%, 65.88%)
Execution: 34.12%/65.88%=0.518

OCR: [[ocr]]
Table: [[table]]
Question: [[query]]
Answer: [[answer]]
Program:

Figure 7: ChartQA prompt template.