Where is this coming from? Making groundedness count in the evaluation of Document VQA models

Armineh Nourbakhsh[1,2], **Siddharth Parekh**[1], **Pranav Shetty**[2], **Zhao Jin**[1], **Sameena Shah**[2], **Carolyn P. Rosé**[1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]J.P. Morgan, New York
anourbak@cs.cmu.edu

Abstract

Document Visual Question Answering (VQA) models have evolved at an impressive rate over the past few years, coming close to or matching human performance on some benchmarks. We argue that common evaluation metrics used by popular benchmarks do not account for the semantic and multimodal groundedness of a model's outputs. As a result, hallucinations and major semantic errors are treated the same way as well-grounded outputs, and the evaluation scores do not reflect the reasoning capabilities of the model. In response, we propose a new evaluation methodology that accounts for the groundedness of predictions with regard to the semantic characteristics of the output as well as the multimodal placement of the output within the input document. Our proposed methodology is parameterized in such a way that users can configure the score according to their preferences. We validate our scoring methodology using human judgment and show its potential impact on existing popular leaderboards. Through extensive analyses, we demonstrate that our proposed method produces scores that are a better indicator of a model's robustness and tends to give higher rewards to better-calibrated answers.

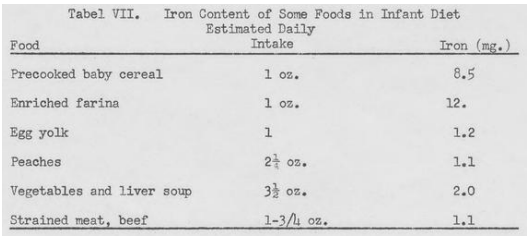
1 Introduction

Visual Question Answering (VQA) over multimodal documents requires joint reasoning over textual, spatial, and visual signals. Several benchmarks have been proposed to measure the performance of SotA models on this task, including single-page and multi-page VQA (Mathew et al., 2021; Tito et al., 2023; Mathew et al., 2022; Van Landeghem et al., 2023; Tito et al., 2021). In these benchmarks, the ground truth answer is expressed as a sequence of tokens and evaluated against the sequence of tokens produced by each model. As such, the evaluation metrics used by these benchmarks focus on the surface similarity between the model output and the ground-truth answer. This misses two key aspects of the model's output: 1) Is it aligned with the expected semantic category? For example, if the ground truth is a number, is the model also producing a number? 2) Can it be located within the input document? In other words, is the model hallucinating a response or is it generating something based on the document (even if it is wrong)? Grounded responses help determine the provenance of the model output and verify its accuracy.

Figure 1 and Table 1 illustrate this using an example from the DocVQA benchmark (Mathew et al., 2021), which uses Normalized Levenshtein Distance as its evaluation metric (Levenshtein, 1966). Given the two excerpts from an image document in Figure 1, two questions are listed in Table 1. The first question, "How many mgs of iron is in enriched farina?", requires the model to reason over a tabular structure and produce the answer "12". If the model produces "26" as the answer, it will be rewarded by a score of 0.5 because "26" shares one digit with the ground-truth answer, "12". In contrast, if the model produces "8.5" as the answer, it will not be rewarded, as there is no overlap with the ground-truth. This is potentially problematic, as the first answer is not mentioned anywhere on the page, and can therefore be considered hallucinatory. The second answer, although inaccurate, captures a number that is present in the table in Figure 1a, and is located on the same column as the ground-truth, potentially signifying some level of tabular reasoning by the model. A more robust evaluation metric would provide a small reward to the second answer, and give the first answer a score of 0.0.

Another question, "How much added iron do premodified infant formulas contain?", requires verbal reasoning over the paragraph in Figure 1b. If a model responds by "up to 12 mgs", it is penalized for its surface dissimilarity to the ground-

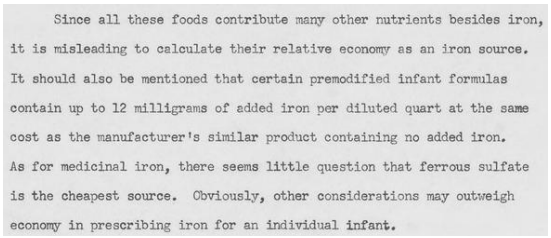
Findings of the Association for Computational Linguistics: NAACL 2025, pages 5341–5361
April 29 - May 4, 2025



Tabel VII. Iron Content of Some Foods in Infant Diet

Food	Estimated Daily Intake	Iron (mg.)
Precooked baby cereal	1 oz.	8.5
Enriched farina	1 oz.	12.
Egg yolk	1	1.2
Peaches	2½ oz.	1.1
Vegetables and liver soup	3½ oz.	2.0
Strained meat, beef	1-3/4 oz.	1.1

(a) Tabular snippet.



Since all these foods contribute many other nutrients besides iron, it is misleading to calculate their relative economy as an iron source. It should also be mentioned that certain premodified infant formulas contain up to 12 milligrams of added iron per diluted quart at the same cost as the manufacturer's similar product containing no added iron. As for medicinal iron, there seems little question that ferrous sulfate is the cheapest source. Obviously, other considerations may outweigh economy in prescribing iron for an individual infant.

(b) Text snippet.

Figure 1: Two excerpts from an image document from the DocVQA dataset (Mathew et al., 2021).

Table 1: Two example questions based on the snippets in Figure 1. The "NLS" column shows the score awarded to hypothetical answers for each question using the NLS metric (Mathew et al., 2021). In "Ours", we show how our proposed score is calculated.

Question	Context	GT Answer	Predicted Answer	NLS	SMuDGE (Ours)						
					Match Score			Grounding Score			Composite Score ($\alpha = 0.25$)
					Text Score	Num Score	Agg.	Horizontal Distance	Vertical Distance	Agg.	
How many mgs of iron is in enriched farina?	Figure 1a	12	26	0.5	-	0.0	0.0	-	-	0.0	0.0
			8.5	0.0	-	0.0	0.0	~ 0.0	0.2	0.02	0.01
How much added iron do premodified infant formulas contain?	Figure 1b	up to 12 milligrams	up to 12 mgs	0.58	0.59	1.0	0.74	0.0	0.0	1.0	0.93
			up to 1z milligrams	0.95	0.94	0.0	0.0	0.0	0.0	1.0	0.75

truth answer, "up to 12 milligrams". In contrast, if the model produces "up to 1z milligrams", it is awarded a higher score because its answer has a larger overlap with the ground-truth. Again, this is problematic, as the second answer misrecognizes a key component of the ground truth (i.e. the number) and as such indicates a completely inaccurate quantity. A more robust evaluation metric should reward a higher score for the first answer than for the second.

In this paper, we propose a new evaluation methodology, which we name Semantics and MUltimodal Document Grounded Evaluation (SMuDGE). SMuDGE addresses the above issues by grounding the similarity score in the expected output type (i.e. numeric, textual, or hybrid answers). We also add a new component—a multimodal grounding score that determines whether the model's output is located within the input document, and where it is located in relation to the ground-truth. In the Document AI literature, this form of multimodal grounding is also referred to as localization (Karatzas et al., 2015).

As Nourbakhsh et al. (2024) argued, grounding is an important requirement (and challenge) for the operationalization of Document VQA models especially in enterprise domains. Nevertheless it is difficult to determine how much grounding might matter to one downstream application versus another. Therefore, we design our evaluation approach to accommodate different settings by allowing users to set the preferred weights for each component.

Concretely, our study makes the following contributions to the field:

1. We propose a new evaluation framework (SMuDGE) that accounts for the groundedness of outputs and the semantic type of the output. We design SMuDGE to be configurable and easy to tune for downstream applications.

2. Using SMuDGE, we re-evaluate the performance of SotA models on four common Document VQA benchmarks, and analyze the impact of grounding on the ranking of each leaderboard.

3. We perform a detailed analysis of the types of questions and answers most impacted by grounding-sensitive criteria, and propose a configurable setting that allows the downstream users of each model to tune the evaluation to their needs.

4. Our analyses show that SMuDGE produces scores better aligned with human preferences.

5. We experimentally demonstrate that better-grounded generation is associated with better

calibrated outputs.

6. Lastly, our analyses show that SMuDGE rewards models that are more robust to variations in tasks and datasets.

2 Background

In recent years, generative multimodal models have made major strides in Visual Question Answering over image documents. As an example, as of January 2025, the top-performing model on the DocVQA leaderboard is within 2 points of human performance[1].

A key challenge of generative models is that their output is difficult to ground within the input document (Zmigrod et al., 2024b) (this is also known as the challenge of localization (Karatzas et al., 2015)). Localization is a common requirement in many real-world applications, especially in enterprise domains where maintaining a proper lineage of data is crucial from a governance perspective (Nourbakhsh et al., 2024). Since generative models produce sequences that are sampled from their vocabulary, they are not guaranteed to generate answers that are based on the input, unless forced to do so via grounded decoding (e.g. as in (Qian et al., 2024)). This in turn makes it difficult to detect hallucinations, establish the provenance of the model's generations, or measure the reliability of its outputs, all of which limit the applicability of such models in many enterprise domains (Nourbakhsh et al., 2024).

This problem is compounded by the fact that most popular Document VQA benchmarks do not account for grounding in their evaluation criteria. A common metric used by these benchmarks is Average Normalized Levenshtein Similarity (ANLS), as proposed by Mathew et al. (2021), which measures the similarity between the ground truth and predicted answers based on their edit distance. As an example, the words 'apple' and 'app1e' have a Levenshtein distance of 1, a normalized Levenshtein distance of 0.2, and an NLS of $1-0.2=0.8$. If the score for a ground-truth/prediction pair is below a predefined threshold (typically set to 0.5 (Biten et al., 2019; Tito et al., 2023; Mathew et al., 2022; Peer et al., 2024)), the score is flattened to zero, otherwise the raw similarity score is used.

The flexibility that the NLS metric provides allows the benchmarks to handle minor errors such as character misspellings resulting from poor Optical Character Recognition, without over-penalizing the models. Contrast this with a metric that relies on n-gram overlap metrics, or cosine similarity of distributed representations. Such metrics might consider "apple" and "app1e" to be very dissimilar words, given a single-character difference between them (where the the letter "l" has been replaced by the digit "1"). Nevertheless, relying solely on surface similarity carries other risks for robust evaluation: 1) Surface similarity does not account for how a small change in the characterization of an answer can impact its meaning (e.g. changing a single digit in a number can change its value by a large magnitude). 2) Surface similarity cannot distinguish between answers that can be traced back to the input document, and those that can result from hallucination.

More recently, some studies have noted the shortcomings of common evaluation metrics in the field of multimodal document understanding, and proposed alternatives (Zmigrod et al., 2024a). Most notably, Peer et al. (2024) proposed ANLS*, a data-type-aware metric that can be used for single or multi-piece extraction and QA over documents. While it addresses many challenges of the ANLS metric, ANLS* is not designed to capture the multimodal groundedness of model outputs. In contrast, we focus on the challenge of measuring groundedness for extractive VQA over documents, where correct answers are guaranteed to be expressed in the input. We propose a configurable evaluation method that not only accounts for the groundedness of predictions, but also incorporates the semantic type of the output, similar to ANLS*. To the best of our knowledge, this is the first study that examines the impact of groundedness in evaluating Document VQA models. The following section describes our proposed approach in detail.

3 Proposed methodology

To measure the impact of groundedness in Document VQA performance, we develop a composite score to rate the output of each model. To ensure that the score can be applied to all models and benchmarks, we assume access to four objects only: 1) The question. 2) The ground truth answer. 3) The answer provided by the model. 4) A dictionary of words and corresponding bounding box coordinates extracted from the input document. This dictionary can be obtained by applying any OCR

[1] https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=1

tool to the document, though the quality of character recognition often differs between different providers. Most benchmarks provide this dictionary as part of their data release.

In the next two subsections, we describe how we calculate two subscores: 1) The multimodal grounding score addresses the question of whether the predicted answer can be located within the input document, and if so, where it is located with respect to the ground truth answer. 2) The type-aware surface similarity score evaluates the predicted answer based on its type, i.e. numeric, textual, or hybrid.

3.1 Multimodal groundedness

Given a question q_i, a ground truth answer t_i, and a predicted answer a_i, we develop a score g_i that places a_i within the originating document (composed of words $w_1, w_2, \cdots, w_N$ and corresponding bounding boxes $b_1, b_2, \cdots, b_N$) and measures its distance to t_i. We do this in two steps:

Locating the predicted answer (a_i) and the ground truth (t_i) within the document. To locate t_i within the document, we find a continuous sequence of words $w_k, w_{k+1}, \cdots, w_{k+n}$ that matches t_i.[2] If no such segment is found (say, due to OCR errors), then we find a sequence that has the highest Normalized Levenshtein Similarity (NLS) to t_i. We name this sequence $\mathbf{w}_{t_i}$ and the corresponding bounding box $\mathbf{b}_{t_i}$, which is calculated by merging $b_k, b_{k+1}, \cdots, b_{k+n}$[3]. Similarly, we find the sequence $\mathbf{w}_{a_i}$ and the corresponding bounding box $\mathbf{b}_{a_i}$ by placing a_i within the document. Note that a_i is not guaranteed to be found on the page, for instance in case of hallucinations. If we can't find a $\mathbf{w}_{a_i}$ such that $\text{NLS}(a_i, \text{concat}(\mathbf{w}_{a_i})) > 0.3$[4], then we define $\mathbf{b}_{a_i}$ as:

$$[\mathbf{b}_{t_i}^{\texttt{left}}, \mathbf{b}_{t_i}^{\texttt{top}}, -\texttt{width}_i - \mathbf{b}_{t_i}^{\texttt{right}}, -\texttt{height}_i - \mathbf{b}_{t_i}^{\texttt{bottom}}] \quad (1)$$

where $\mathbf{b}_{\cdot}^{\texttt{left}}, \mathbf{b}_{\cdot}^{\texttt{top}}, \mathbf{b}_{\cdot}^{\texttt{right}}, \mathbf{b}_{\cdot}^{\texttt{bottom}}$ indicate the four coordinates of the bounding box $\mathbf{b}_{\cdot}$ and $\texttt{width}_i, \texttt{height}_i$ indicate the width and height of the page, respectively. In other words, we use the bounding box of the ground-truth answer t_i and mirror its bottom right corner in the negative space. This ensures that the distance between $\mathbf{b}_{t_i}$ and $\mathbf{b}_{a_i}$ is measured as 1 (see below).

Measuring the distance. Next, we measure d_i, the distance between $\mathbf{b}_{a_i}$ and $\mathbf{b}_{t_i}$. We do this by first finding the centroid of each bounding box, and then measuring the Normalized Manhattan Distance (NMD) between the centroids.[5] In other words:

$$d_i = \quad (2)$$
$$\left|\frac{\mathbf{b}_{t_i}^{\texttt{right}}}{2\times\texttt{width}_i} - \frac{\mathbf{b}_{t_i}^{\texttt{left}}}{2\times\texttt{width}_i} - \frac{\mathbf{b}_{a_i}^{\texttt{right}}}{2\times\texttt{width}_i} + \frac{\mathbf{b}_{a_i}^{\texttt{left}}}{2\times\texttt{width}_i}\right| +$$
$$\left|\frac{\mathbf{b}_{t_i}^{\texttt{bottom}}}{2\times\texttt{height}_i} - \frac{\mathbf{b}_{t_i}^{\texttt{top}}}{2\times\texttt{height}_i} - \frac{\mathbf{b}_{a_i}^{\texttt{bottom}}}{2\times\texttt{height}_i} + \frac{\mathbf{b}_{a_i}^{\texttt{top}}}{2\times\texttt{height}_i}\right|$$

If the predicted answer a_i cannot be located within the document, the formulation presented in Equation 1 yields $d_i = 1$. Note that $0 \leq d_i \leq 1$.

Finally, we calculate the grounding score g_i by applying an exponential decay function to d_i: $g_i = e^{\frac{-d_i}{1-d_i}}$. Note that the score rewards cases where $\mathbf{b}_{t_i}$ and $\mathbf{b}_{a_i}$ are close, or horizontally/vertically aligned (due to lower Manhattan Distance) with the reward dropping exponentially with distance. The exponential decay function was demonstrated to best represent positional information in unimodal text in Chi et al. (2022), and extended to multimodal documents in Wang et al. (2023).

3.2 Type-aware surface similarity

To measure m_i, the surface match score between t_i and a_i, we follow the below criteria:

1. If t_i is textual[6], we use the NLS metric.

2. If t_i is numeric, we use a binary score that indicates whether the predicted answer matches the ground truth exactly. We allow some flexibility in the match, for example numbers scaled by 100, thousand, million, or billion are considered a match. This is to account for different expressions of percentages, basis points, financial metrics, etc.

3. If t_i is composed of both textual and numeric characters, we first create substrings num_{a_i},

[2] Note that a multimodal document is a 2-D artifact, and therefore a "continuous sequence" can extend in multiple directions, depending on the reading order of the page. Most commercial OCR packages such as `Textract` segment each page based on semantic information, e.g. an address block is presented as one segment, even if it contains multiple lines. We therefore rely on the segments provided by these packages to determine continuity. In the absence of such information, a graph representation of the document can be used as a proxy. In Appendix A, we provide an algorithm that can be used to ground the sequence using this graph representation.

[3] See Appendix B.1 for additional details.

[4] See Appendix B.5 for more information on how this threshold was selected.

[5] See Appendix B.2 for a discussion of alternative distance normalization methods.

[6] See Appendix B.3 for additional details.

str_{a_i}, num_{t_i}, and str_{t_i} by extracting the numeric and non-numeric characters of a_i and t_i, respectively. Next, we calculate the number-based and text-based scores for each substring according to the above criteria. The final score is a weighted harmonic mean of the two sub-scores: $\frac{11}{\frac{10}{\text{num\_score}_i} + \frac{1}{\text{str\_score}_i}}$.[7] Note that the model has to get the numeric part of the answer correctly to be rewarded higher.

3.3 Composite metric

Given the mutimodal grounding score g_i and type-aware match score m_i, we propose the following composite score parameterized by α:

$$s_i = \alpha m_i + (1 - \alpha) g_i \tag{3}$$

Note that $\alpha = 0$ yields the grounding score and $\alpha = 1$ yields the type-aware match score. The configurability of the α parameter allows users to tune it on a validation set of their choice, or, as we will show in Section 5.3, to optimize it such that it rewards well-calibrated outputs.

4 Experiments

Given the composite score proposed in Section 3.3, we investigate the impact of groundedness on four prominent Document VQA benchmarks.

DocVQA (Mathew et al., 2021) is a visual question-answering (VQA) dataset designed specifically for document images. It contains over 12,000 document images sourced from scanned business forms, reports, and invoices, among others. The dataset is structured with over 50,000 question-answer pairs, and questions are broken down into 9 categories, indicating the context of the correct answer (e.g. "Free_text", "Layout", "Figure/Diagram", etc.). This breakdown is not available for the text collection. Therefore we determine the type of each question using GPT-4o (gpt)[8]. Next we remove questions in the "Yes/No" category to filter potentially abstractive questions. This results in 5,130 questions in the final dataset.

InfographicVQA. (Mathew et al., 2022) is a dataset aimed at visual question answering over complex infographic documents. The dataset includes over 5,000 infographic images and over 30,000 questions that require reasoning over text, charts, and images embedded within the infographic. We filter multi-piece answers from the test collection, resulting in 3,272 samples.

MP-DocVQA (Tito et al., 2023) focuses on multi-page documents. It consists of over 46,000 question-answer pairs from 6,000 multi-page documents. We use 5,019 questions in the test set.

DUDE (Van Landeghem et al., 2023) is a document understanding dataset focused on structured documents such as forms, invoices, and tables. It includes around 5,000 documents and 41,000 question-answer pairs. We limit the test collection to single-piece extractive questions, resulting in 2,552 samples.

For each sample in each dataset, we calculate the NLS as well as the composite score, with α set to increments of 0.05 in the $[0, 1]$ range.

5 Analysis

Throughout most of our experiments, we set $\alpha = 0.25$, as it proves optimal based on the calibration analysis provided in Section 5.3. Since α is optimized on the DUDE dataset, we have not included this dataset in any of the analyses that use this optimal value for α.

5.1 Leaderboard analysis

We first analyze how SMuDGE can affect the rankings produced by Document VQA benchmarks. Figure 2 illustrates this using the top 10 models[9] on the DocVQA leaderboard. The leftmost column of the figure shows the original ANLS-based ranking[10]. The second column shows how the ranking changes if we switch to SMuDGE with $\alpha = 0.25$. As the figure shows, human performance and QWen2-VL (Wang et al., 2024) remain stable, but all other models move by at least one position on the leaderboard. The middle segment of the figure shows how the models would rank based on the type of question. Certain question types such as "Figure/Diagram" and "Table/List" offer little volatility, but for questions that fall under "Handwritten" or "Other", the volatility is higher.[11]

[7] See Appendix B.4 for additional details.

[8] Please see Appendix C for details.

[9] As of September 2024.

[10] Note that our ANLS-based rankings could be slightly different from the leaderboard, since we have filtered the questions per Section 4.

[11] An example of a question classified as "Other" is: "What does GCC stand for?" requiring the model to infer that an acronym mentioned on one part of a page is related to an entity mentioned on a different part. This category of questions constitutes about 0.2% of the DocVQA dataset, and can be considered negligible.

The middle segment of the figure also shows that some models such as SMoLA-PaLI-X (Wu et al., 2024) are better at answering questions based on "Free_text" contexts, whereas they struggle with "Table/List" questions compared to other models.

The right segment of the figure shows the rerankings broken down by the type of answer. As expected, textual answers offer the closest ranking to the original one produced by ANLS, whereas numeric and hybrid answers perturb the ranking of the leaderboard. Notably, humans remain the top performer for textual and hybrid answers, but fall behind two other models in the numeric category. This can be attributed to the human tendency to rephrase certain entities such as numbers and dates. For example, in Question #3027, the ground truth answer "(16.1%)" is rephrased as "-16.1%" by human respondents, and for question #3290, "1,700" is modified as "about 1,700".

Figure 3 shows the correlation between rankings produced by ANLS and by our composite score with $\alpha = 0.25$. Following Alzahrani et al. (2024), we calculate the correlation based on a two-tailed Kendall's τ analysis. Note that the y-axis on Figure 3 begins at 0.70. As the figure shows, questions with textual answers are the least affected by switching to our score, but numeric and hybrid answers impact the ranking by a larger margin. This is expected as the text-only version of our score is the closest to ANLS. Of the three benchmarks shown in the figure, InfographicVQA is most affected by our score, whereas DocVQA and MP-DocVQA retain a strong correlation with their original rankings. As evidenced by Figure 2, this strong correlation does not indicate a stable leaderboard, but one where the models move by $\pm d$, where d is a small number.

5.2 Question type analysis

Figure 4 shows the correlation between our composite score and the original ranking of the DocVQA leaderboard for each question type. As expected, moving from small values of α (weighing groundedness more that type-aware similarity) to large values (weighing type-aware similarity more than groundedness), moves the rankings closer to the original ANLS ranking. This is especially true of the "Free_text" category, where our score comes closest to ANLS. Once again, "Other" is the outlier category, which can be safely ignored due to its small sample size. The remaining categories show a similar trend, further establishing that groundedness is not accounted for in ANLS-based rankings.

5.3 Association with calibration

The DUDE dataset provides the confidence scores produced by each model (when available). This enables the benchmark to report Expected Calibration Errors (ECE) (Pakdaman Naeini et al., 2015), indicating if the models are wrongfully over or underconfident about the accuracy of their output. We use this metric to determine whether our proposed score can account for accuracy through calibratedness. To do this, we map the score at various α's against the calibration error of each model, and calculate the Pearson-R correlation between the two. The results are displayed in Figure 5. As the figure shows, at small values of α (focusing on groundedness), there is a negative correlation with ECE, indicating that a higher score is correlated with a lower ECE. As α increases and the score shifts towards surface similarity, the association moves towards positive, crossing 0 around $\alpha = 0.5$. This trend can be observed for all categories of questions except "Textual" questions, which enforce surface similarity at all α values. The optimal value for α, which minimizes the correlation with ECE across most categories lies at around $\alpha = 0.25$.

5.4 Association with robustness

Next, we inspect the association between SMuDGE and the robustness of a given model. Robustness is not a formally defined term in the Document VQA field, but can be interpreted as a model's consistent performance across different settings, benchmarks, and sample types. Therefore, we define robustness as the volatility[12] of a model's ranking when evaluated on various subsets of questions (e.g. textual, numeric, hybrid, or all questions at once). We plot this volatility against the volatility of a model's scores, using the DocVQA, MP-DocVQA, and InfographicVQA benchmarks. Figure 6 shows the results using ANLS as well as SMuDGE with $\alpha = 0.25$. Each dot represents one model, with red dots representing models evaluated using ANLS, and blue dots representing models evaluated by SMuDGE. As the regression lines in the figure show, both approaches maintain a positive trend between the volatility in scores and rankings. In other words, models with stable rankings tend to have stable scores as well. However, the positive trend is stronger for our score compared to ANLS, with a small but statistically significant regression coefficient of 0.58 (compared to ANLS's 0.33).

[12] See Appendix B.6 for additional details.

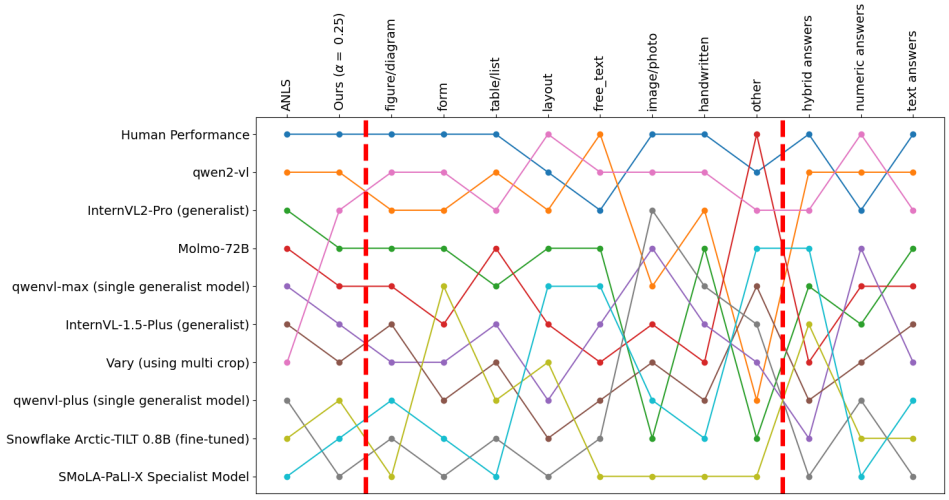


Figure 2: The rankings of the top 10 models on the DocVQA leaderboard, before and after applying our composite score with $\alpha = 0.25$. Left segment: Rankings based on ANLS versus our score. Middle segment: Our rankings broken down by question type. Right segment: Our rankings broken down by answer type.

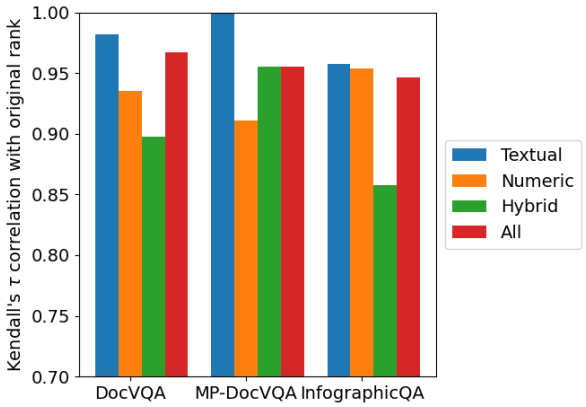


Figure 3: The correlation between the rankings produced by our method (with $\alpha = 0.25$) and the original ANLS-based ranking, broken down by the type of answer. All τ values are significant at $p \ll 0.05$.

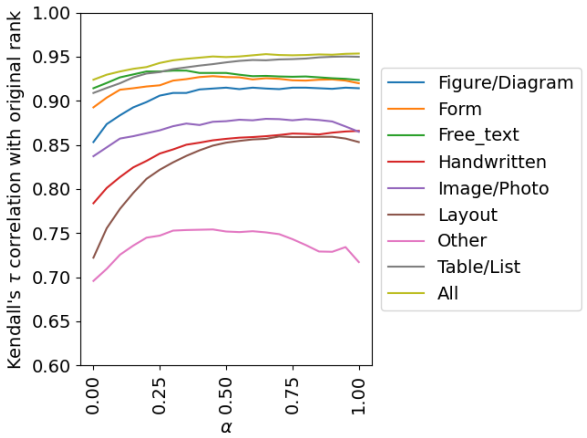


Figure 4: Kendall's τ rank correlation with the original DocVQA leaderboard, broken down by question types. All τ values are significant at $p \ll 0.05$.

Next, to present a qualitative view of how our score can reward robust models, we calculate a robustness score for each model in the DocVQA benchmark. To do this, we scale a model's rank volatility by its median rank. This ensures that if a model is stable across rankings, it receives a high robustness score, unless it is a generally poor performing model (e.g. a model that comes last in all rankings). Table 3 lists the top-5 models identified using this technique. The ANLS-based models reflect the default ranking of the DocVQA leaderboard, with Humans leading the group, followed by Large MLMs such as QWen2-VL (Wang et al., 2024) and InternVL2-Pro/InternVL-1.5 (Chen et al., 2024).

In contrast, our score produces a ranking that includes a Small MLM, namely, Arctic-TILT (Borchmann et al., 2024). As of October 2024, this model is ranked 11 on the DocVQA leaderboard, above all other Small MLMs and a few Large MLMs. In addition, it is ranked 1st on the MP-DocVQA and DUDE leaderboards. No other models listed in the ANLS column show the same level of cross-benchmark robustness. Similarly, Molmo-72B (Deitke et al., 2024) is 4th on the InfographicVQA benchmark. The strong cross-benchmark rankings indicate that our method can generate rankings that reward robust models.

Table 2: Five samples from the human preference study, showing cases where the human judges preferred our score, NLS, or neither scores. In the latter case, the human judges preferred equal scores for Models A and B.

Dataset	Question	GT	Model A	Model B	Human pick
DocVQA	What is the vitamin A requirement (in I.U.) for a 'lactating' mother ?	"1,000 i.u. plus basic requirements"	"basic requirements"	"1,000"	SMuDGE
			NLS: 0.54	NLS: 0.0	
			Ours: 0.0	Ours: 0.62	
MP-DocVQA	What is the day and date of Meeting?	"thursday 22 october"	"thursday"	"saturday 24 october"	SMuDGE
			NLS: 0.0	NLS: 0.74	
			Ours: 0.81	Ours: 0.25	
InfographicVQA	Which age group uses social media the most?	"18-29 year olds"	"18-29 group"	"18-24 year olds"	SMuDGE
			NLS: 0.53	NLS: 0.93	
			Ours: 0.98	Ours: 0.0	
DocVQA	What is the date of the letter?	"august 1, 1983"	"The date of the letter is August 1, 1983."	"August 1983"	Neither
			NLS: 0.0	NLS: 0.78	
			Ours: 0.97	Ours: 0.0	
InfographicVQA	What is the estimated number (in billions) of social media users around the globe by 2019?	"2.72"	"#infographic"	"2. 72"	ANLS
			NLS: 0.0	NLS: 0.8	
			Ours: 0.0	Ours: 0.0	

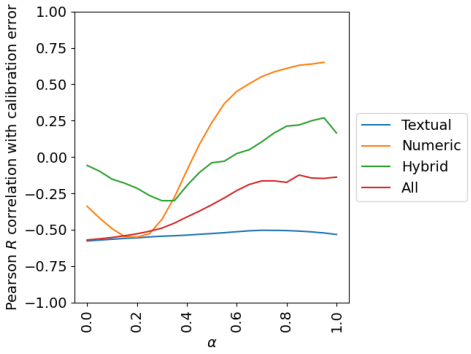


Figure 5: Pearson R correlation with the calibration error of models based on the DUDE leaderboard, broken down by answer type.

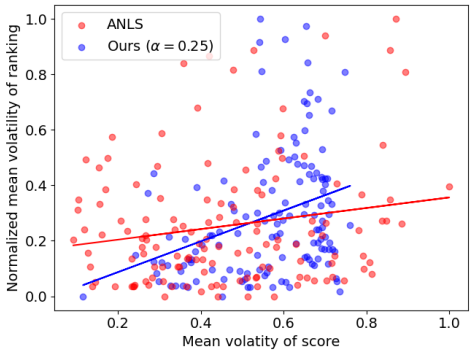


Figure 6: The mean volatility of each model's score versus its ranking. Red dots represent ANLS scores and blue dots represent SMuDGE with $\alpha = 0.25$.

Table 3: Top-5 models based on robustness rankings produced by ANLS versus our score (with $\alpha = 0.25$).

	ANLS		SMuDGE
1	Human	1	Human
2	QWen2-VL	2	QWen2-VL
3	InternVL2-Pro	3	InternVL2-Pro
4	QWenVL-Max	4	Molmo-72B
5	InternVL-1.5-Plus	5	Snowflake Arctic-TILT

5.5 Human evaluation

We used human judgment to assess the validity of our scores compared to ANLS. To do this, we used data from three benchmarks: DocVQA, MP-DocVQA, and InfographicVQA. In each benchmark, we sampled questions and a pair of answers produced by two models, indicated by model A and model B (different models could be selected for each sample). We limited the samples to cases where model A's NLS score was higher than B, but SMuDGE scored B higher than A, or vice versa. We sampled up to 100 such question-answers triplets from each benchmark[13]. Three researchers were presented with these triplets, as well as the ground truth answer, and asked which model they thought should be scored higher. The annotations produced a mean Cohen's κ of 0.82, indicating a high level of agreement. We filtered the annotations to those on which at least two annotators agreed. This resulted in 28 samples for DocVQA, 86 samples for MP-DocVQA, and 66 samples for InfographicVQA.

Figure 7 shows the annotators' agreement rates with NLS versus our score. The "Neither" bucket indicates that the annotators believed the models should have been scored equally. As the figure shows, annotators agreed with SMuDGE in the majority of cases across all three benchmarks, indicating that our approach is better aligned with human judgment. We observe that InfographicVQA, which yielded the highest rate of agreement with NLS, contains the largest number of misspelled

[13] Some datasets had fewer qualifying triplets.

numbers, as seen in the last row of Table 7. This could be a result of the complex layout and design of infographics.

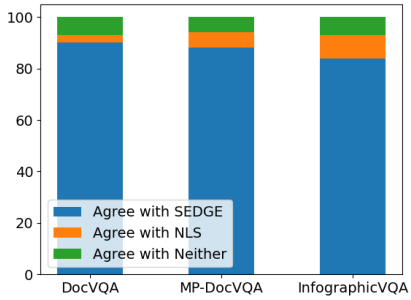


Figure 7: Human preference for pairwise rankings produced by NLS versus SMuDGE (with $\alpha = 0.25$).

6 Conclusion

In this study, we showed how popular evaluation metrics such as ANLS can miss important nuances when used to analyze Document VQA models. Instead, we proposed SMuDGE, a new metric that is sensitive to the groundedness of the models' outputs. Through extensive analyses, we showed how SMuDGE is better aligned with human judgement as well as the calibratedness of the models. Our analyses also showed that rankings produced by SMuDGE were better indicators of a model's robustness across question types and in different benchmarks. Our studies demonstrate the importance of groundedness in the performance and assessment of Document VQA models. We hope that in addition to presenting a new evaluation method, our study inspires researchers to develop better grounded Document VQA models.

7 Limitations

The analyses performed in this paper were all conducted on single-span, extractive answers. To extend the grounding mechanism to multi-span answers, the matching algorithm would need to handle an arbitrary number of partitions, unless the benchmark specifically identifies each span in its test set annotations. Determining groundedness on abstractive questions is a challenging task that is outside of the scope of this study.

The methodology proposed in this study does not account for semantic categories that go beyond textual/numeric/hybrid forms, such as currencies, dates, timestamps, etc. each of which come with nuances that can be mishandled by solely considering surface similarity.

Since the α parameter was tuned on the DUDE dataset, it was excluded from some of the other analyses. The remaining benchmarks (DocVQA, MP-DocVQA, and InfographicVQA) are all released as part of the same suite of tasks, with the first two datasets being based on the same collection of documents. This can lead to biases in the analyses that are based solely on these three benchmarks. However, none of these benchmarks provided access to the confidence scores produced by the models, and therefore could not be used to tune α.

Lastly, the grounding algorithm mentioned in Section 3.1 relies on the accuracy of the reading order of each page, as presented in the OCR output. As Zhang et al. (2023) point out, this can often be noisy or misleading. Appendix A offers an alternative, more generalizable, yet slower solution using a walk over the β-skeleton presentation of each page.

8 Acknowledgements

Armineh Nourbakhsh, Pranav Shetty, and Sameena Shah's work is supported by JPMorgan Chase & Co. This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-09-15.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and

Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. ICDAR 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE.

Łukasz Borchmann, Michał Pietruszka, Wojciech Jaśkowski, Dawid Jurkiewicz, Piotr Halama, Paweł Józiak, Łukasz Garncarek, Paweł Liskowski, Karolina Szyndler, Andrzej Gretkowski, et al. 2024. Arctic-TILT. Business document understanding at sub-billion scale. *arXiv preprint arXiv:2408.04632*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. 2022. Kerple: Kernelized relative positional embedding for length extrapolation. *ArXiv*, abs/2205.09921.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, James Park, Reza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE.

Chen-Yu Lee, Chun-Liang Li, Chu Wang, Renshen Wang, Yasuhisa Fujii, Siyang Qin, Ashok Popat, and Tomas Pfister. 2021. ROPE: Reading order equivariant positional encoding for graph-based document information extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 314–321, Online. Association for Computational Linguistics.

V Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10, pages 707–710.

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

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. DocVQA: A Dataset for VQA on Document Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Armineh Nourbakhsh, Sameena Shah, and Carolyn Rose. 2024. Towards a new research agenda for multimodal enterprise document understanding: What are we missing? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14610–14622, Bangkok, Thailand. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

David Peer, Philemon Schöpf, Volckmar Nebendahl, Alexander Rietzler, and Sebastian Stabinger. 2024. ANLS*–A universal document processing metric for generative large language models. *arXiv preprint arXiv:2402.03848*.

Hongjin Qian, Zheng Liu, Kelong Mao, Yujia Zhou, and Zhicheng Dou. 2024. Grounding language model with chunking-free in-context retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1311, Bangkok, Thailand. Association for Computational Linguistics.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens,

and Tomasz Stanislawek. 2023. Document Understanding Dataset and Evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19528–19540.

Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Kang Gu, and Sameena Shah. 2023. DocGraphLM: Documental graph language model for information extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1944–1948, New York, NY, USA. Association for Computing Machinery.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. 2024. Omni-SMoLA: Boosting generalist multimodal models with soft mixture of low-rank experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14205–14215.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. *arXiv preprint arXiv:2310.11016*.

Ran Zmigrod, Zhiqiang Ma, Armineh Nourbakhsh, and Sameena Shah. 2024a. TreeForm: End-to-end annotation and evaluation for form document parsing. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 1–11, St. Julians, Malta. Association for Computational Linguistics.

Ran Zmigrod, Pranav Shetty, Mathieu Sibue, Zhiqiang Ma, Armineh Nourbakhsh, Xiaomo Liu, and Manuela Veloso. 2024b. "What is the value of {templates}?" Rethinking document information extraction datasets for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13162–13185, Miami, Florida, USA. Association for Computational Linguistics.

A A β-skeleton based grounding algorithm

Algorithm 1 describes a possible way to ground a model output O within a page, without apriori access to the reading order. First, a page is represented by a β-skeleton graph, similar to Lee et al. (2021). Next, the first and last tokens of O are matched to the page by finding all nodes (i.e. tokens) on the graph that have a Levenshtein similarity to the first and last token, beyond a threshold T. Lastly, all possible paths between such nodes are found, and the path with the highest NLS to O is selected as the matching path.

A threshold can be set on the score of the best matching path S, below which the path is considered a mismatch and therefore no effective matches are found on the page, e.g. in cases when the model has hallucinated the output.

This algorithm ensures that any path that is matched to O is within a contiguous 2-D walk on the page, without the need for information related to reading order. A major downside of this algorithm is its quartic time complexity, which can be improved by caching partial paths. Nevertheless, we decided to use a simpler algorithm that relies on the reading order provided by OCR tools.

B Additional experimental details

B.1 Merging bounding boxes

A sequence of bounding boxes can be merged by finding the left-most, top-most, right-most, and bottom-most corners in the sequence in order to create a new bounding box. If all bounding boxes in the sequence form a contiguous segment, merging them would yield their union. However, if the bounding boxes are in disparate locations, this simple merging algorithm will not yield their union, and will cover additional areas. As an example, if a ground truth answer spans two lines, covering the second half of one line and the first half of the next, the merging algorithm will create a bounding box that covers both lines in full. Despite this limitation, we use this algorithm because we are only interested in measuring the distance between the resulting bounding boxes based on their centroids.

B.2 Normalizing the distance

Given the ground truth bounding box $\mathbf{b}_{t_i}$ and the predicted bounding box $\mathbf{b}_{a_i}$, our goal is to measure the distance between the centroids of the bounding boxes. In our proposed formulation, this distance is normalized by the width and height of the page, namely $\texttt{width}_i$ and $\texttt{height}_i$. This is not the only possible option for normalizing the distance. For example, the distance can be normalized by the width/height of the ground truth bounding box $\mathbf{b}_{t_i}$, or the average size of the ground truth and predicted bounding boxes $\mathbf{b}_{t_i}$ and $\mathbf{b}_{a_i}$.

Each option offers advantages and disadvantages, which we will demonstrate using examples. For simplicity, we will suppose that the height of all bounding boxes is similar, and focus on width only.

Normalizing the bounding boxes by the width

Algorithm 1 β-skeleton walk for placing a sequence of tokens within a page.

// β-skeleton representation of a page
Input: $G = (N, V)$
// Matching target: a sequence of tokens
Input: $O = o_1, o_2, \cdots, o_n$
// Threshold for token similarity
Input: T
// Best path on the graph that matches O
Output: P
// The similarity of the best path to O
Output: S
// Create empty indices of all possible paths over the graph, starting from o_1 ending in o_n.
$p_s \leftarrow \{\}$
$p_e \leftarrow \{\}$
for $i \in \{1, \ldots, |N|\}$ **do**
  $s_{i1} = \text{NLS}(N_i, o_1)$
  $s_{in} = \text{NLS}(N_i, o_n)$
  if $s_{i1} > T$ **then**
    append(p_s, n_i)
  end if
  if $s_{in} > T$ **then**
    append(p_e, n_i)
  end if
  // Search all possible paths and select the one with the highest score
  for $p_j \in p_s$ **do**
    for $p_k \in p_e$ **do**
      for path $\in$ paths$(p_j \rightarrow p_k)$ **do**
        if NLS(path, O) $> S$ **then**
          $S \leftarrow \text{NLS}(\text{path}, O)$
          $P \leftarrow$ path
        end if
      end for
    end for
  end for
end for

of $\mathbf{b}_{t_i}$ over-penalizes models that provide short answers compared to the ground truth, and under-penalizes models that provide longer answers compared to ground truth. A real example from the DocVQA dataset is the question "What decides the selection of terms of Committee members?" The ground truth answer is "decided by a lottery", whereas some models may produce "lottery" and some may produced "will be decided by a lottery". We would want the grounding distance to be consistently low for these variations. But once normalized by the width of ground truth, the first model will be over-penalized and the second model will be under-penalized.

An alternative is to normalize the widths by the average widths of $\mathbf{b}_{t_i}$ and $\mathbf{b}_{a_i}$. While this formulation does not suffer from sensitivity to the variety of sizes, it does ignore the sizes of the bounding boxes relative to the size of the page. For example, consider a page with width 1000. On this page, these two scenarios produce the same distance of 1:

Scenario A: $\mathbf{b}_{t_i}$ spans $[0 - 500]$ and $\mathbf{b}_{a_i}$ spans $[500 - 1000]$. The average width is 500. The centroids are at 250 and 750, respectively. Therefore the centroids are at a raw distance of 500 and a normalized distance of 1.

Scenario B: $\mathbf{b}_{t_i}$ spans $[0 - 50]$ and $\mathbf{b}_{a_i}$ spans $[50 - 100]$. The average width is 50. The centroids are at 25 and 75, respectively. Therefore the centroids are at a raw distance of 50 and a normalized distance of 1.

On the one hand, it can be argued that it is fair for the distances to be equal, as the bounding boxes are adjacent in both scenarios. On the other hand, it can be argued that the distance should scale with the width of the bounding boxes compared to the width of the page, because locating a small bounding box on a large page is more difficult than a large bounding box on a small page.

In contrast, given that the width of the page is 1000, our proposed method would produce a distance of 0.5 for Scenario A and 0.05 for Scenario B. This is an interpretable metric, as it indicates that the two bounding boxes are within 50% of the width of the page in Scenario A, and 5% in Scenario B. It can of course be argued that our formulation is too sensitive to the scale of the page. Our proposed score is indeed not perfect, but we consider it to be an interpretable "layout-agnostic" metric that can be easily calculated across all samples.

We thank our reviewer for suggesting these alternative options.

B.3 Determining the semantic type of the predicted answer

To classify a string of characters s as numeric, textual, or hybrid, we follow the below algorithm:

1. If every character in s is a digit, then s is numeric.
2. If every character in s is alphabetical, then s is textual.

3. Otherwise s is hybrid.

Note that this simple algorithm renders a large portion of strings such as "1,700" or "(8)" as hybrid. This is not detrimental to SMuDGE, as it still favors the accuracy of numbers against non-numeric characters by a factor of 10 to 1 (see Appendix B.4).

Note that hybrid strings are split into numeric sequences and non-numeric sequences, e.g. "1,700" is split into "1700" and "," and each part is evaluated separately before being combined in the weighted harmonic mean.

B.4 Tuning the weights for the numeric score and the text score

Setting the weight of num_score$_i$ to 1 would mean that the numeric and text components of an answer would have equal importance, which is indeed not valid. For example if the ground truth is "12 milligrams", then the answers "2 milligrams" and "12 milligram" should not receive equal scores, as the former is quantitatively incorrect, but the latter has a simple typo. On the other hand, setting a very high weight for num_score$_i$ can be problematic. For example if the ground truth is "12 mgs" and the predicted answer is "12 ms", we would need to properly penalize the text component, because "ms" stands for "milliseconds" and not "milligrams".

Therefore we tuned the weight of num_score$_i$ against str_score$_i$ by testing values in the set $\{1, 10, 100, 1000\}$. The tuning was performed on a subsample of 100 hybrid answers from the DocVQA validation set, and validated by three human annotators. Each annotator was presented with answer/ground-truth pairs and the four variations of the score calculated using the four values in $\{1, 10, 100, 1000\}$. The annotators were asked to select the score best representing the similarity between the predicted answer and the ground-truth answer. Annotators most frequently selected the score produced by a weight of 10. On average, each annotator selected this weight 86% of the times. For 73 samples on which the three annotator agreed, they selected this weight 96% of the times.

B.5 Setting the similarity threshold

It is common practice in the field of Document VQA to set a threshold for NLS (Biten et al., 2019; Mathew et al., 2021; Tito et al., 2023; Mathew et al., 2022; Peer et al., 2024). This is done to determine whether a match can be reasonably expected, or whether any similarity is coincidental (e.g. the NLS between "dog" and "giraffe" is larger than 0 as they share the letter "g", but the two are entirely different tokens). Following Biten et al. (2019), most studies have set the threshold to 0.5. Given that this was not justified by any validation study in Biten et al. (2019), we instead conducted our own tuning exercise using the validation dataset described in Appendix B.4. Three human annotators performed a binary classification on 100 pairs of predicted answers and most similar spans from the corresponding documents. Each pair was tagged as a "match" or a "mismatch", indicating whether the predicted answer referred to the same span (perhaps with slight changes in spelling). The NLS value of 0.3 yielded the most optimal threshold for distinguishing between matching and mismatched pairs, predicting a mismatch with an F1 of 0.94.

B.6 Calculating volatility

We use the standard definition of volatility as scaled standard deviation:

$$\text{vol}([x_1, \cdots, x_T]) = \text{std}([x_1, \cdots, x_T])\sqrt{T} \quad (4)$$

C Determining the types of questions in DocVQA

To determine the type of each question, we passed the following information to GPT-4o: 1) The document image. 2) The question. 3) The ground truth answer, as provided by the dataset. 4) A prompt, asking the model to determine the context from which the answer was extracted.

You can see an example prompt below:

Question: What is the extension number?
Answer: 5177
The above question was answered based on the document attached. What do you think best describes the context from which the answer was extracted? Select one of the below options. Simply return the correct option without any explanation.

1. **Figure/Diagram**
2. **Form**
3. **Table/List**
4. **Layout**
5. **Free_text**
6. **Image/Photo**
7. **Handwritten**
8. **Yes/No question**
9. **Other**

The experiment ran on September 7th, 2024. The agreement rate with the DocVQA validation set was 69.5%.

D Extended leaderboard analysis

Figures 8 to 10 show the reranking analysis for MP-DocVQA, InfographicVQA, and DUDE benchmarks, respectively. As with Figure 2, our composite score has been calculated with $\alpha = 0.25$.

E Extended question type analysis for DocVQA

Figure 11 shows how the top 10 models on the DocVQA leaderboard would be reranked if our score was used to evaluate them, broken down by question types.

F Answer type analysis for DocVQA

Figure 12 shows how the top 10 models on the DocVQA leaderboard would be reranked if our score was used to evaluate them, broken down by answer types.

G Answer type analysis for MP-DocVQA

Figure 13 shows how the top 10 models on the MP-DocVQA leaderboard would be reranked if our score was used to evaluate them, broken down by answer types.

H Answer type analysis for InfographicVQA

Figure 14 shows how the top 10 models on the InfographicVQA leaderboard would be reranked if our score was used to evaluate them, broken down by answer types.

I Answer type analysis for DUDE

Figure 15 shows how the top 10 models on the DUDE leaderboard would be reranked if our score was used to evaluate them, broken down by answer types.

J Correlation between answer types and original ranking

Figure 16 shows the correlation between the ranking of each leaderboard and the ranking produced by SMuDGE at various various for α, broken down by the type of answer.

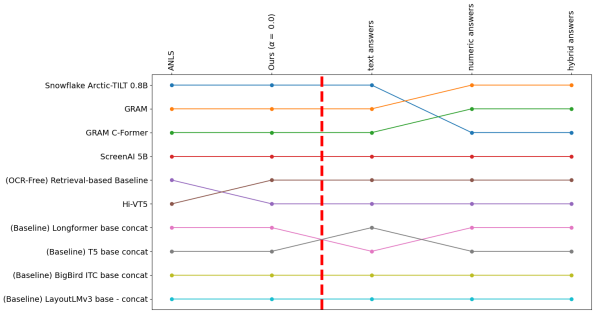


Figure 8: MP-DocVQA leaderboard.

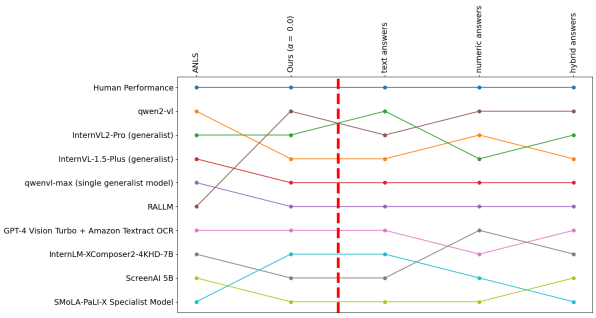


Figure 9: InfographicVQA leaderboard.

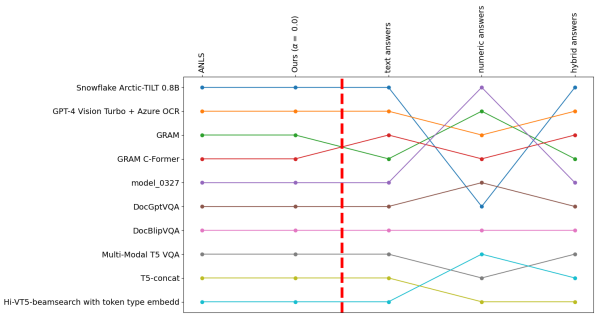
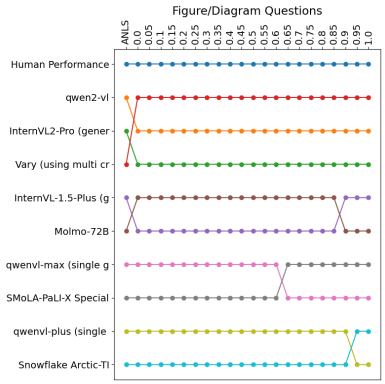
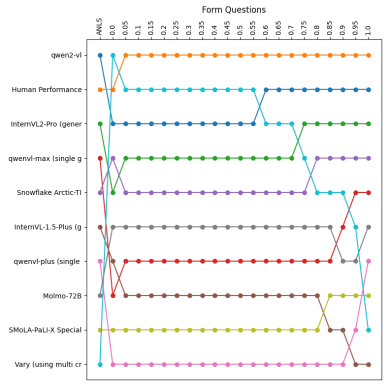


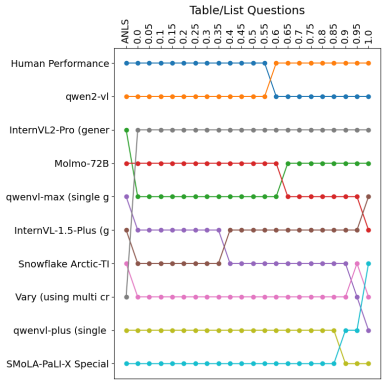
Figure 10: DUDE leaderboard.



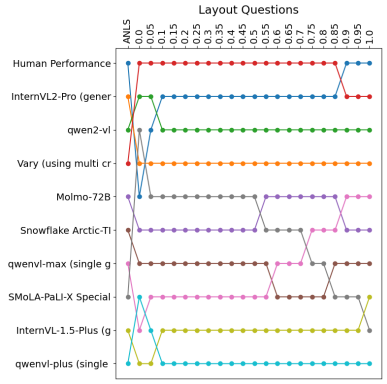
(a) Figure/Diagram



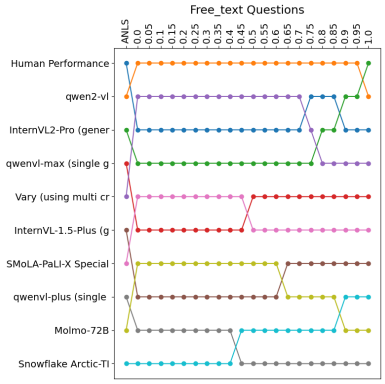
(b) Form



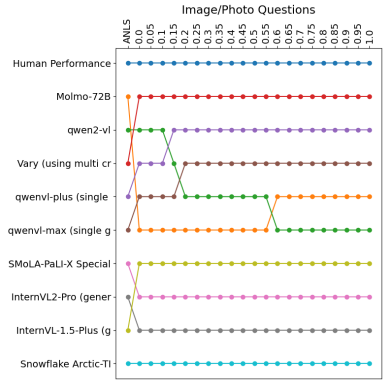
(c) Table/List



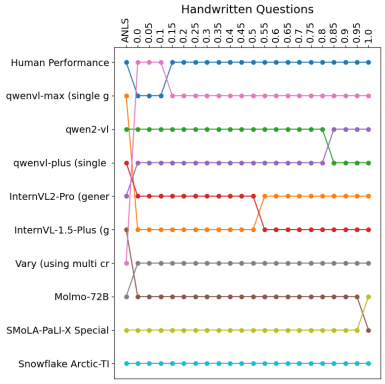
(d) Layout



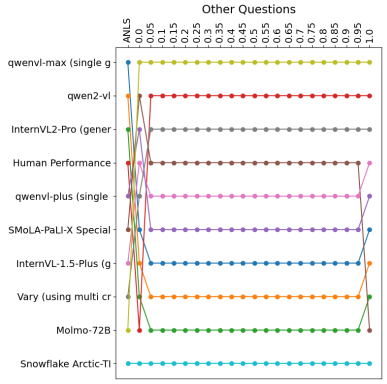
(e) Free text



(f) Image/Photo

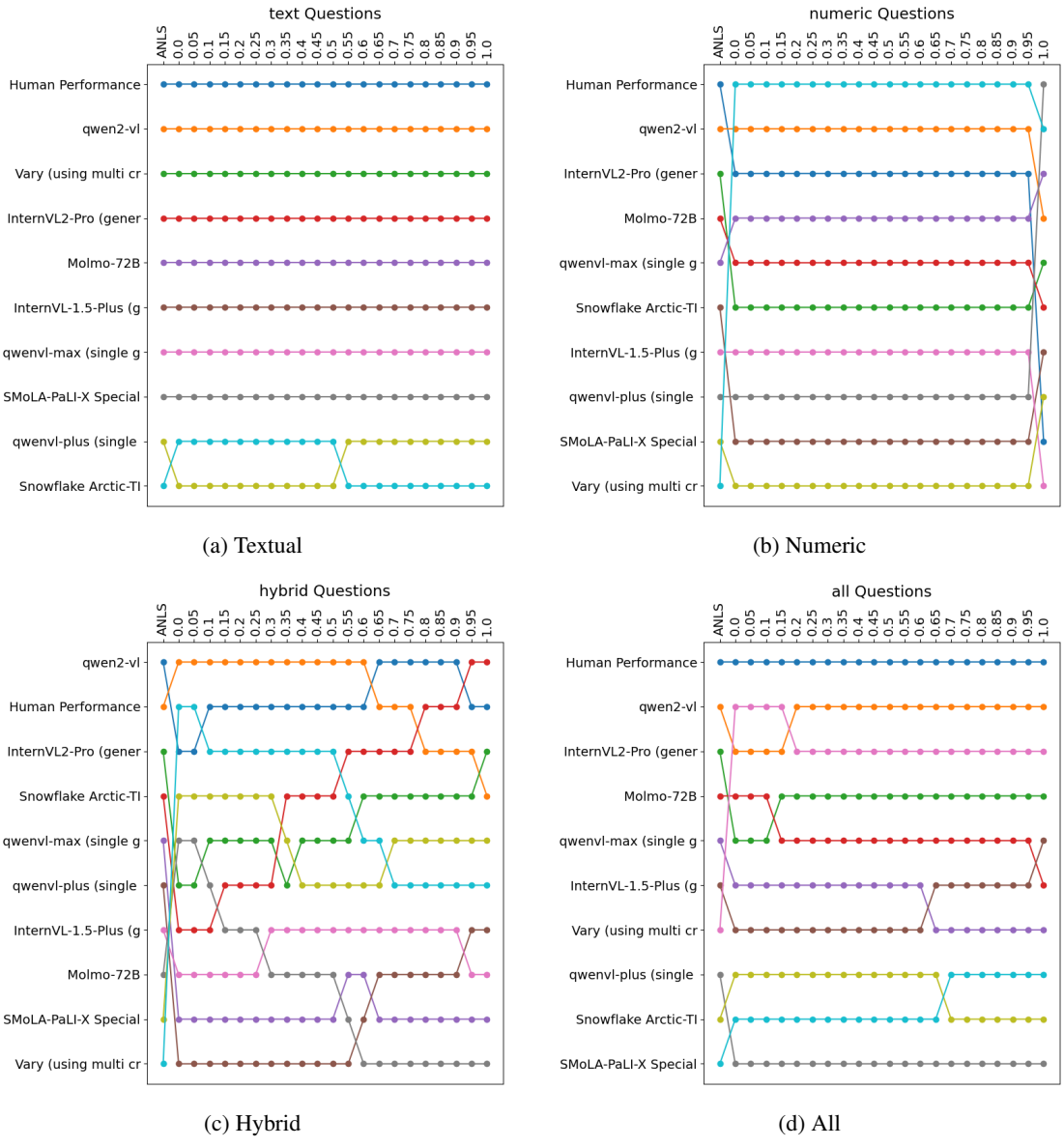


(g) Handwritten



(h) Other

Figure 11: The impact of our score on the ranking of the top 10 models on the DocVQA benchmark, broken down by question type.



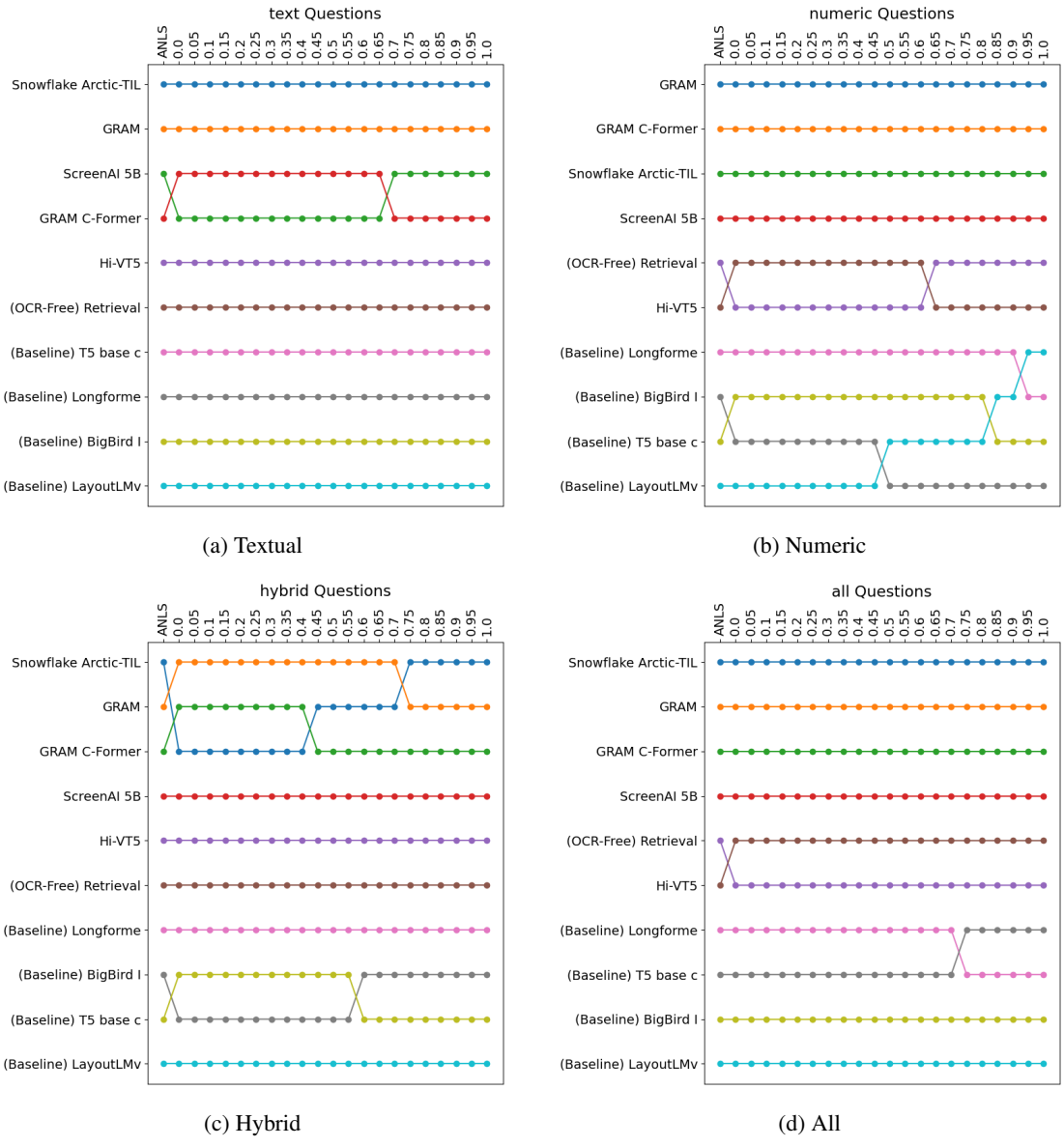
(a) Textual

(b) Numeric

(c) Hybrid

(d) All

Figure 12: The impact of our score on the ranking of the top 10 models on the DocVQA benchmark, broken down by answer type.



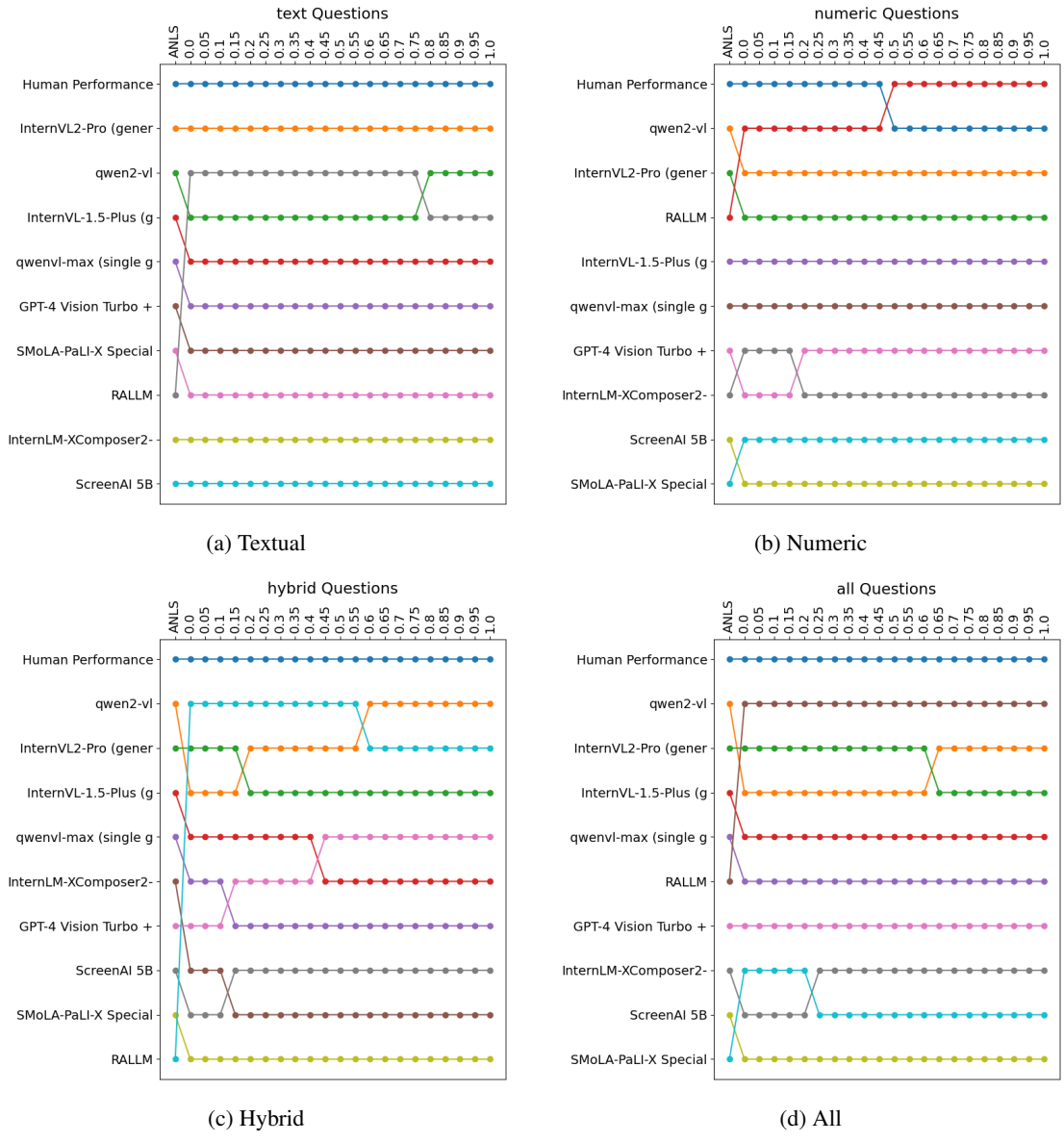
(a) Textual

(b) Numeric

(c) Hybrid

(d) All

Figure 13: The impact of our score on the ranking of the top 10 models on the MP-DocVQA benchmark, broken down by answer type.



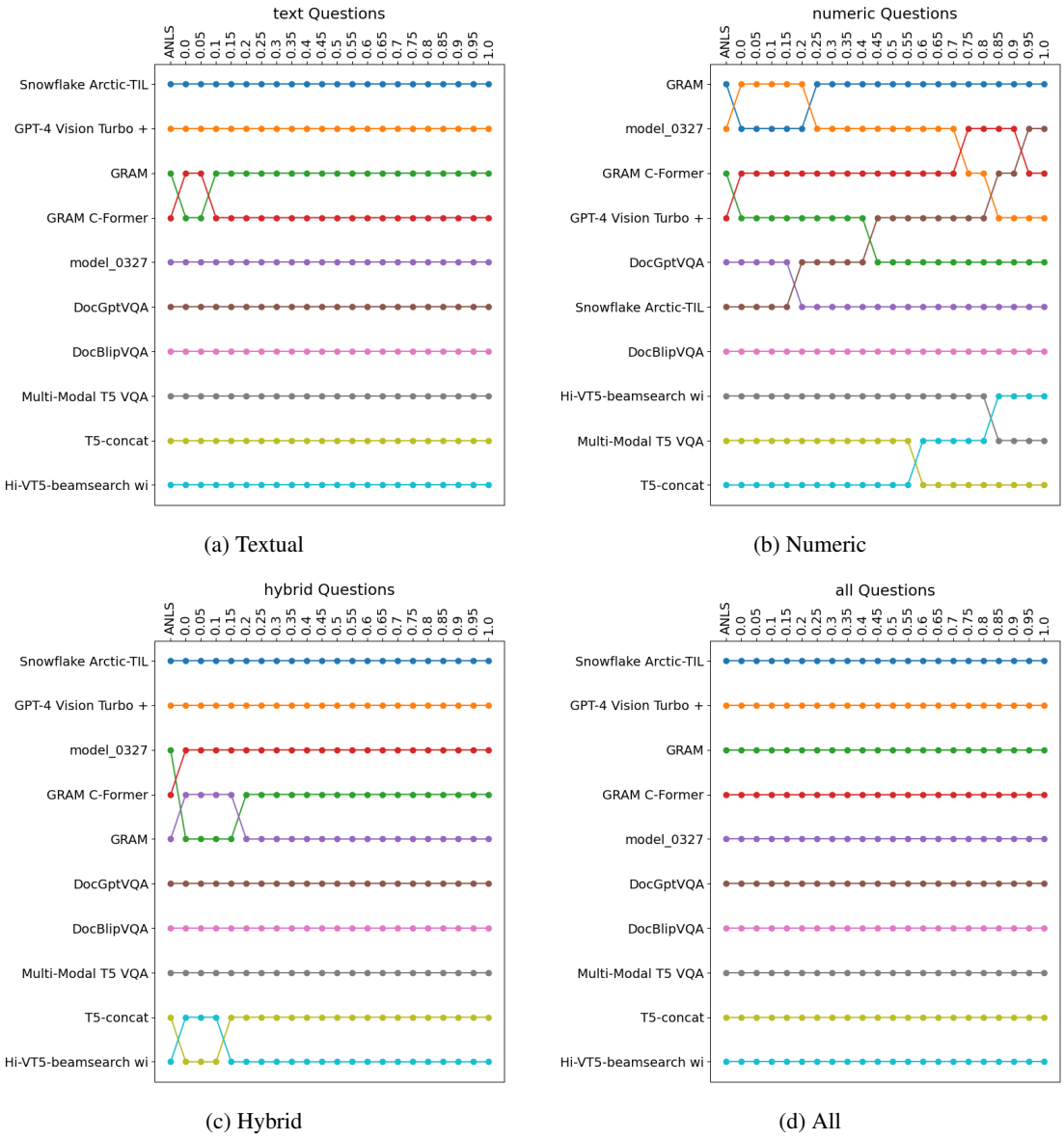
(a) Textual

(b) Numeric

(c) Hybrid

(d) All

Figure 14: The impact of our score on the ranking of the top 10 models on the InfographicVQA benchmark, broken down by answer type.



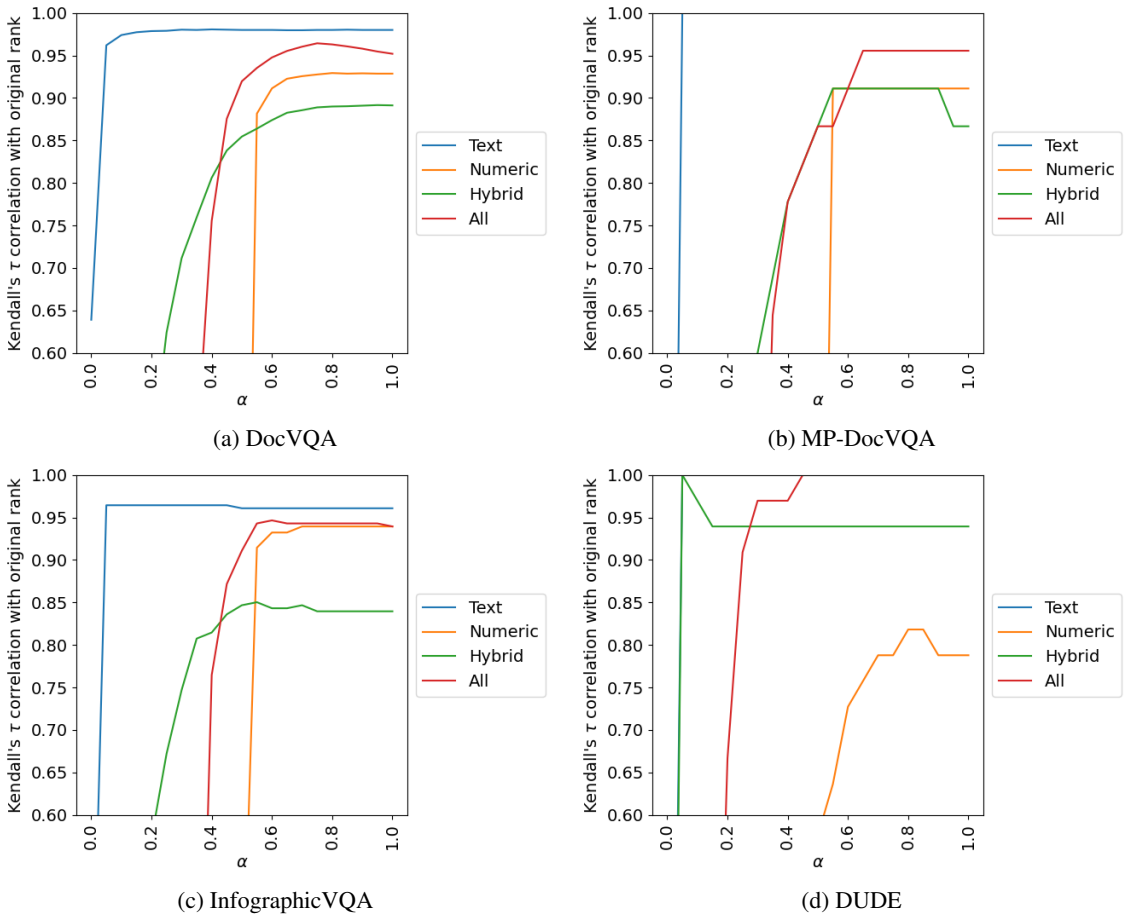
(a) Textual

(b) Numeric

(c) Hybrid

(d) All

Figure 15: The impact of our score on the ranking of the top 10 models on the DUDE benchmark, broken down by answer type.



(a) DocVQA

(b) MP-DocVQA

(c) InfographicVQA

(d) DUDE

Figure 16: Kendall's τ correlation between different α settings and the original ranking of each benchmark, broken down by the type of answers.