

Most of a Document-VQA Score Is Answer Formatting

Srinivas Harish
New York University
srinivas.harish@nyu.edu

Abstract

DocVQA and InfographicVQA are scored with average normalized Levenshtein similarity (ANLS), and a DocVQA number is read as a measure of how well a model reads a page, but the `lmms-eval` and `VLMEvalKit` harnesses that compute it append a one-line terse-answer suffix to every question. ANLS divides edit distance by the length of the longer string, so a correct answer inside a sentence more than twice its length falls below the 0.5 threshold and scores zero, and removing only the suffix moved Qwen2.5-VL-3B’s DocVQA ANLS from 94.2 to 7.5. On that model ANLS given a short answer (within two tokens of the longest gold) stayed between 93.9 and 95.7 across three suffixes whose scores spanned 94.2 to 65.3, so the reported score followed the share of short answers more than the share that contained the gold. Two suffixes that `lmms-eval` defines for the same task differed by 28.8 points, and under one of them 79% of the answers ANLS scored zero contained the gold answer, all 305 of which a judge with negated and two-candidate controls accepted as the committed answer. On a 317-question subset, four hosted models lost between 34.4 and 89.5 points without the suffix (Gemini 3.8 Flash fell from 92.0 to 2.5), and with the suffix in place, deleting Markdown emphasis markers raised Claude Opus 5.5 by 10.9 points. On Qwen2.5-VL-3B a length-gated span scorer moved 13.3 points across the five suffixes where ANLS moved 86.6, and it scored a submission of the page’s whole OCR at 1.3, where best-subsequence exact match and the learned PEDANTS classifier gave 93.0 and 98.6. On ChartQA the suffix was worth 28.4 points of relaxed accuracy to Qwen2.5-VL-3B. We release the scorer and recommend reporting the prompt suffix and output post-processing with every document-VQA score.

1 Introduction

DocVQA (Mathew et al., 2021) is scored with ANLS, and a score of 94.2 is taken to mean that a model extracts the requested value from a scanned page almost every time, which is how the Qwen2.5-VL and Qwen3-VL technical reports use it (Bai et al., 2025b,a). Such numbers are computed with evaluation harnesses, the libraries that run a benchmark end to end, which append to every question a suffix that the reports’ results tables do not show. When we removed only that suffix, Qwen2.5-VL-3B’s DocVQA ANLS moved from 94.2 to 7.5, although 86.9% of the answers that ANLS then scored zero still contained the gold answer. In `lmms-eval` (Zhang et al., 2025) a DocVQA question receives “\nAnswer the question using a single word or phrase.” by default and “Answer:” when the model name passed to the harness is `qwen_vl`; `VLMEvalKit` (Duan et al., 2024) appends the first string to every DocVQA question, and the Qwen3-VL report gives the same string as its DocVQA prompt in an appendix. Which suffix a model saw therefore depends on the harness and on how the model is registered in it (files and commits in Appendix B).

Without such a suffix, a correct answer from Qwen2.5-VL-3B or -7B was usually scored zero. ANLS takes the normalized Levenshtein similarity between the prediction and the nearest gold answer and zeroes anything below 0.5, and it normalizes by the length of the longer of the two strings, so a prediction that places the right answer inside a sentence is penalized in proportion to the sentence: *The date mentioned in the letter is 1/8/93.* against the gold *1/8/93* scores 0.14 and is set to zero. The suffix keeps predictions within the length ANLS tolerates, and its effect is included in every DocVQA score these harnesses report.

We measured that effect by changing the suffix and nothing else, on Qwen2.5-VL-3B, Qwen2.5-

VL-7B and SmolVLM run locally and on GPT-6 Sol, Gemini 3.8 Flash, Claude Opus 5.5 and Qwen3.8-Max through an API, and scored the stored generations with ANLS, containment and the length-gated span scorer, and Qwen2.5-VL-3B’s with all seven scorers of Table 5.

1. On Qwen2.5-VL-3B, removing the suffix cost 86.6 ANLS points and 13.3 points of length-gated span score on the same generations (the span scorer looks for the gold inside predictions at most m tokens longer than it, §3), and the two suffixes `lmms-eval` defines for DocVQA differed by 28.8 points on this model (§4).
2. Without a suffix 86.9% and 87.8% of the Qwen2.5-VL models’ ANLS zeros contained the gold, and a judge checked on negated and two-candidate controls found that at least 97.7% of the contained zeros in every condition we audited committed to the gold as the answer (§4.1). ANLS given a short answer (within two tokens of the longest gold) stayed between 93.9 and 95.7 across three suffixes, so the reported score followed the share of short answers (§4.3).
3. The hosted models lost between 34.4 and 89.5 ANLS points without the suffix, and deleting Markdown emphasis markers raised Claude Opus 5.5 by 10.9 points with it (§4.2); a trailing full stop cost SmolVLM 52.5 points of exact match (§6).
4. We release a length-gated span scorer whose gate we set against a page-dump attack (the page’s whole OCR submitted as the prediction); it scored that attack at 1.3, where best-subsequence exact match and PEDANTS (Li et al., 2024) scored it at 93.0 and 98.6 (§5).

2 ANLS and Whitespace Handling

For a gold set G and a prediction p , let $s = \max_{g \in G} (1 - d(g, p) / \max(|g|, |p|))$ with d the Levenshtein distance; ANLS is s if $s \geq \tau = 0.5$ and 0 otherwise, averaged over questions and reported on a 0 to 100 scale. The denominator grows with the prediction, so a verbose answer is penalized for its length whether or not it is right.

We transcribed the scorer from the reference implementation of the ICDAR Robust Reading Competition (Biten et al., 2019a) and from `lmms-eval`,

whose scoring functions are identical. Both compute the edit distance on lowercased, whitespace-normalized strings and the denominator on the raw strings, so leading or trailing whitespace enlarges the denominator without touching the numerator and can only raise a score, by an amount that depends on the model. Qwen2.5-VL-3B emitted almost no stray whitespace and gained 0.0002 points from this, while SmolVLM prefixed a space to every prediction and gained 1.85.

3 Experimental Setup

Data. We used the public 1,200-question DocVQA sample distributed with its OCR tokens (`nielsr/docvqa_1200_examples`), which the attack in §5 needs. Under the default suffix we measured 94.2 for Qwen2.5-VL-3B and 96.2 for the 7B, within 0.5 points of the 93.9 and 95.7 reported on the test split (Bai et al., 2025b). InfographicVQA validation (Mathew et al., 2022) (2,801 questions) is a second document corpus under the same metric and ChartQA test (Masry et al., 2022) (2,500 questions) a third benchmark with a different one.

Models. Run locally: Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), and SmolVLM-2B (HuggingFaceTB/SmolVLM-Instruct, Marafioti et al., 2025), an Idefics3-architecture model from a different family. Hosted: GPT-6 Sol, Gemini 3.8 Flash, Claude Opus 5.5 and Qwen3.8-Max, queried through the OpenRouter API on 26 September 2026 (Appendix F). The hosted models were compared on the 317 questions among the first 427 of the sample for which all eight runs (four models, with and without the suffix) returned a response; on these questions Qwen2.5-VL-3B scored 93.0 and 6.0 with and without the suffix, against 94.2 and 7.5 on all 1,200.

Decoding. Local models decoded greedily at batch size one (Appendix G), with a cap of 32 new tokens for terse suffixes, as in the `lmms-eval` configuration, and 64 where the prompt invites prose. Greedy decoding with a smaller cap produces a prefix of the same output, so we cut every generation made with the 64-token cap to its first 32 tokens and rescored; Qwen2.5-VL-3B’s no-suffix ANLS moved from 7.5 to 7.6 and the 7B’s from 3.6 to 4.0, and restricting to generations that stopped before the cap gave 7.8 and 4.0 (Appendix C). Hosted models ran at temperature 0 with the same 1,024-

Model	Suffix	tok	ANLS	span	EM	n_0	cg
Qwen2.5-VL-3B	default	3	94.2	95.5	86.0	53	18.9
Qwen2.5-VL-3B	" Answer: "	4	65.3	92.6	58.1	386	79.0
Qwen2.5-VL-3B	"briefly"	3	87.6	89.9	30.6	57	31.6
Qwen2.5-VL-3B	none	11	7.5	82.2	0.9	1057	86.9
Qwen2.5-VL-3B	"sentence"	3	83.6	94.1	62.2	161	71.4
Qwen2.5-VL-7B	default	2	96.2	96.7	90.5	35	8.6
Qwen2.5-VL-7B	none	13	3.6	73.0	0.0	1126	87.8
Qwen2.5-VL-7B	"sentence"	10	15.4	89.5	6.2	953	90.2
SmolVLM-2B	default	2	68.4	67.4	2.8	245	2.0
SmolVLM-2B	none	3	54.9	79.3	21.7	458	48.3

Table 1: DocVQA, $n = 1,200$ questions per row; only the suffix differs within a block. *tok* is the median prediction length in tokens, *span* the length-gated span scorer at $m=20$, *EM* exact match, n_0 the number of predictions ANLS scores exactly zero and *cg* the percentage of those that contain a gold answer.

token limit with and without the suffix, which no generation reached. Generations are stored verbatim, every number in the paper is recomputed from them, and differences are computed from unrounded values.

Scoring. All scorers lowercase and collapse whitespace first. *Exact match* compares the whole prediction with each gold. *Containment* splits both strings on spaces, strips the characters `. , ; : ! ? " ' ( ) [ ]` from both ends of every token, and asks whether the token sequence of any gold occurs contiguously in the prediction’s; *best-subsequence exact match* is containment used as a score. The *span scorer* applies when the prediction has at most m more tokens than the longest gold: it takes the best normalized Levenshtein similarity between any gold and any contiguous run of 1 to $|g_{\max}| + 2$ prediction tokens, $|g_{\max}|$ being the token length of the longest gold, and thresholds it at 0.5 , and longer predictions fall back to whole-string ANLS with edit distance and length both computed on the normalized strings. *Token F1* is the bag-of-words F1 of Rajpurkar et al. (2016) against the best gold, and PEDANTS (Li et al., 2024) is a learned answer-equivalence classifier that takes the question, the gold and the prediction. For hosted models we delete Markdown emphasis markers (`*`, `**`, `__` and backticks) before containment and the span scorer, and report ANLS on the raw output, as a harness computes it.

4 Results

Across the five suffixes Qwen2.5-VL-3B’s ANLS ranged over 86.6 points and its span score over

13.3 (Table 1). Under a paired bootstrap over the same 1,200 items (5,000 replicates, 95% percentile intervals throughout), removing the suffix cost the 3B 86.6 ANLS points [84.9, 88.2] and 13.3 points [11.7, 15.1] of span score, so 73.3 of the ANLS points it lost came from answers the span scorer still credited, answers that contained the gold but did not give it as a bare string.

On Qwen2.5-VL-7B the same change cost 92.6 ANLS points [91.2, 93.8]. The 7B wrote longer no-suffix answers than the 3B, so it lost more, and the two models swapped places: with the suffix the 7B led, 96.2 to 94.2, and without it the 3B led, 7.5 to 3.6. On InfographicVQA validation the suffix was worth 71.9 ANLS points [70.3, 73.4] to the 3B, from 72.5 to 0.7, and 60.3% of its zero-scored no-suffix answers contained the gold against 3.9% under the suffix.

Asked to “*Answer in a complete sentence.*”, an instruction a deployed document assistant might receive, the 7B complied at a median of 10 tokens and scored 15.4 ANLS with a span score of 89.5, while the 3B stayed at three tokens and scored 83.6, 68.2 points above the 7B. “*Answer briefly.*” left the 3B at a median of three tokens and 87.6 ANLS, 6.6 below the default. The “ Answer: ” suffix is the one lmms-eval applies to DocVQA for models passed to it under the name `qwen_v1`, and applied to Qwen2.5-VL-3B it cost 28.8 ANLS points against the default.

4.1 Gold Answers Inside ANLS Zeros

Containment is a property of two strings and does not depend on ANLS or on our scorer, so it checks the ANLS zeros without any other metric (Table 1, columns n_0 and *cg*). Under the default suffix 18.9% of the 3B’s zeros contained the gold, so a zero there usually meant a wrong answer; under “ Answer: ” it was 79.0% and with no suffix 86.9%. For the 7B and SmolVLM the figures were 8.6% and 2.0% under the default and 87.8% and 48.3% without a suffix. Restricted to golds of at least two tokens the 3B’s figures under “ Answer: ” and with no suffix were 79.2% and 86.3%, and a raw-substring definition agreed to within 2 percentage points everywhere (Appendix D). One of the 1,200 questions contains its own gold answer, and excluding it changes no *cg* value by more than 0.2 percentage points.

A prediction can contain the gold without giving it as the answer, as in *it is not 1/8/93, it is 1/9/93*. We gave each of the 2,434 zero-scored predictions

that contained a gold, with its question and gold, to Qwen3.8-27B (4-bit weights, run locally, temperature 0) and asked whether the response commits to the reference as its answer. On 600 control responses built from 200 of the 3B’s no-suffix questions the judge accepted 100.0% of bare gold answers and 0.0% of negated (*The answer is not X; it is Y.*) and two-candidate responses. On the real predictions it accepted 100.0% of the 305 under " Answer:" and 99.0% without a suffix for the 3B, 99.6% for the 7B and 97.7% for SmolVLM, so contradictory or multi-candidate outputs were at most 2.3% of the contained zeros in any of these conditions (prompt in Appendix E).

4.2 Hosted Models

All four hosted models lost ANLS without the suffix, by 34.4 to 89.5 points, a factor of 2.6 between models whose suffix scores were within 17 points of each other (Table 2). Gemini 3.8 Flash fell from 92.0 to 2.5 while its span score fell by 18.7 to 77.0 and 92% of its zeros contained the gold, and GPT-6 Sol, whose median answer grew only from 3 to 4 tokens without the suffix, lost 34.4. The ranking changed without the suffix, GPT-6 Sol going from third of the four to first and Gemini 3.8 Flash from second to last. The span score fell by 4.5 to 18.7 points for the hosted models, and their zeros contained the gold in 64% to 94% of cases.

The hosted models also lost points to Markdown, which appeared in fewer than 1% of the locally run models’ answers. Without the suffix every Gemini 3.8 Flash and Claude Opus 5.5 answer and 56% of Qwen3.8-Max answers contained Markdown bold, as in ****1/8/93****, and ANLS counts each asterisk as an edit. Under the default suffix Claude Opus 5.5 still bolded 35% of its answers and scored 78.1; deleting the emphasis markers raised that to 89.0, a gain of 10.9 points [8.5, 13.5], which moved it from last of the four to 0.1 points below GPT-6 Sol’s unstripped 89.2. That 78.1 was computed with the lmms-eval default suffix and ANLS on the raw output, so this model’s DocVQA number depends on whether the evaluation strips Markdown, which neither harness configuration nor the model reports cited here state for DocVQA.

4.3 ANLS Decomposed by Answer Length

We call a prediction *short* if it is within two tokens of the longest gold, so that the prediction is at most two tokens longer than a gold. Under the default suffix ANLS and the span scorer nearly agreed

(94.2 against 95.5 for the 3B), because 98.9% of predictions under that suffix were short, and the agreement disappeared under the other suffix the same harness defines (65.3 against 92.6 under " Answer:").

ANLS given a short answer (Table 3) stayed between 93.9 and 95.7 across three suffixes whose scores spanned 94.2 to 65.3, and ANLS given a long answer was 6.6 to 9.6. The long answers therefore contributed 0.1, 1.2 and 2.3 points of the reported ANLS under those three suffixes, and on this model the reported number was the share of short answers times a factor between 93.9 and 95.7, plus at most 2.3 points from long answers. The span scorer gave 95.5 to 96.7 on short answers and 76.0 to 84.7 on long ones.

Under “*Answer in a complete sentence.*” the 3B’s median prediction stayed at three tokens, but the share of predictions over five tokens rose from 9.3% to 19.6% and the share that were not short from 1.1% to 12.3%, and ANLS scored those at 9.5, so the 3B lost 10.6 points.

5 A Scorer That Separates Formatting from Content

Asking whether the gold answer appears anywhere in the prediction removes the length penalty, but without a limit on length the span scorer gave 98.4 to a prediction consisting of the page’s whole released OCR (best-subsequence exact match gave 93.0), since the answer is on the page. We therefore gate the span scorer (§3) at m extra tokens and set m by measurement (Table 4). The attack score rose monotonically with m , and the no-suffix span score rose from 69.2 at $m = 12$ to 82.2 at $m = 20$ and 87.0 at $m = 40$. We took $m = 20$, the largest value in the sweep at which the attack scored under 2 points (1.3), and the default-suffix span score had reached its ungated value of 95.5 by $m = 12$.

Table 5 compares the gated scorer with six others on the same five suffixes and the same attack. The three that compare the whole prediction with the gold (ANLS, exact match, token F1) scored the attack at 0 to 4 and ranged over 54.7 to 86.6 points across the five suffix conditions. The three that look for the answer inside the prediction (best-subsequence exact match, PEDANTS, the ungated span scorer) ranged over 0.7 to 6.7 points and scored the attack at 93.0 to 98.6. The gated scorer had a range of 13.3 and an attack score of 1.3, and every other scorer had a range above 50 points or

Model	ANLS		drop without suffix		bold %
	suffix	none	ANLS	span	
Qwen2.5-VL-3B	93.0	6.0	87.0 [83.6, 90.1]	10.4 [7.3, 13.6]	0
GPT-6 Sol	89.2	54.8	34.4 [29.8, 39.2]	4.5 [2.0, 6.9]	13
Gemini 3.8 Flash	92.0	2.5	89.5 [86.4, 92.3]	18.7 [14.3, 23.1]	100
Claude Opus 5.5	78.1	4.1	74.0 [69.9, 77.7]	16.0 [11.4, 20.5]	100
Qwen3.8-Max	94.9	40.4	54.5 [49.2, 59.6]	8.9 [5.2, 12.7]	56

Table 2: DocVQA, the same 317 questions for every row; Qwen2.5-VL-3B (local) is included for reference. *suffix* is the lmms-eval default and *none* no suffix. Drops are paired differences in points with 95% bootstrap intervals. *bold* is the share of no-suffix answers containing Markdown bold.

Suffix	short	ANLS	ANLS given	
	(%)		short	long
lmms-eval default	98.9	94.2	95.1	9.6
"complete sentence"	87.7	83.6	93.9	9.5
" Answer: "	65.9	65.3	95.7	6.6
no suffix	1.9	7.5	74.3	6.2

Table 3: Qwen2.5-VL-3B on DocVQA, $n = 1,200$. Across the top three suffixes ANLS given a short answer barely moved while the reported score fell by 28.8 points. The no-suffix row has 23 short predictions, too few for its conditional to be compared with the others.

Scorer	default	none	range	attack
ANLS	94.2	7.5	86.6	0.0
exact match	86.0	0.9	85.1	0.0
token F1	94.2	39.5	54.7	4.0
best-subsequence EM	90.3	86.9	5.0	93.0
PEDANTS	96.7	96.2	0.7	98.6
span, ungated	95.5	88.8	6.7	98.4
span, gated $m=20$	95.5	82.2	13.3	1.3

Table 5: Qwen2.5-VL-3B on DocVQA. *range* is the spread in points across the five suffixes of Table 1 and *attack* the score of the page-dump attack.

6 Punctuation, Whitespace and Markdown

Adding the suffix raised SmolVLM’s ANLS by 13.5 points and lowered its span score by 11.9 [10.2, 13.8], while on the two Qwen2.5-VL models it raised ANLS by more than 80. Under the default suffix SmolVLM ended every answer with a full stop, ' 1/8/93. ' against the gold '1/8/93', and removing that one character raised its ANLS by 6.1 points and its exact match from 2.8 to 55.3; the same operation changed Qwen2.5-VL-3B’s ANLS by -0.1 points.

7 Relaxed Accuracy on ChartQA

ChartQA (Masry et al., 2022) is scored by a relaxed accuracy inherited from PlotQA (Methani et al., 2020) that admits 5% relative error on numeric answers. Removing the suffix moved Qwen2.5-VL-3B from 83.5 [82.1, 84.9] to 55.2 [53.2, 57.2], a drop of 28.4, while a parser that first extracts the number from the prediction and ignores a percent sign that appears on only one of prediction and gold gave 70.4; 383 of the 2,500 no-suffix answers (15.3%) were wrong under the official parser and right under that one. Without the suffix the median answer was one token and the mean 11.3, and the 38% of answers that included reasoning accounted for the lost points. That parser restored 53.6% of

m	default	no suffix	page-dump attack
0	94.3	7.7	0.0
5	94.9	17.1	0.0
8	95.4	48.0	0.3
12	95.5	69.2	0.6
20	95.5	82.2	1.3
40	95.5	87.0	4.1
ungated	95.5	88.8	98.4

Table 4: Setting the gate, Qwen2.5-VL-3B on DocVQA: span score under two suffixes and under the page-dump attack. Higher is better in the first two columns and lower in the third; we use $m=20$.

an attack score above 90.

Selected on DocVQA and applied unchanged to InfographicVQA, $m = 20$ held the 3B’s no-suffix span score to 45.7 against 67.1 ungated, because no-suffix InfographicVQA predictions exceeded the longest gold by a median of 16 tokens against 8 on DocVQA, and 35.1% of them exceeded the gate against 7.5%. We therefore release the selection procedure with the scorer, which evaluates each m against a page-dump attack built from the target corpus’s OCR and takes the largest m at which the attack stays under 2 points.

the points lost without the suffix, against 93.8% for the ungated span scorer on DocVQA, and the rest were reasoned answers that reached a wrong value.

The official parser divides by one hundred exactly when a string ends in %, separately for prediction and gold, and none of the 2,500 test golds ends in % while 1,915 are numeric, so 40% against the gold 40 has a relative error of 0.99 and is wrong. Qwen2.5-VL-3B ended 1 of its 2,500 answers with a percent sign under the suffix and 17 without it, and neutralizing the rule changed its relaxed accuracy by 0.04 and 0.0 points.

8 Related Work

ANLS (Biten et al., 2019b) was designed for scene-text VQA, where a near-miss was a recognition error, and DocVQA and InfographicVQA adopted it while the leading systems were extractive and kept it once systems became generative. ANLS* (Peer et al., 2024) extends it to structured outputs and keeps compatibility with earlier scores; we hold the metric fixed and vary the suffix, which moved ANLS by up to 92.6 points (Qwen2.5-VL-7B). Nourbakhsh et al. (2025) score whether a DocVQA answer is grounded in the page, which an answer that contains the gold without being grounded in the page would fail.

The closest work is Ye et al. (2026), who audit official errors on six short-answer VQA benchmarks with a semantic judge under the standard terse-answer suffix, find that up to half of the errors on text-rich benchmarks are acceptable answers in the wrong surface form, and show that rewording the prompt flips item-level correctness. We change only the suffix the harness itself appends, which measures the effect on a reported number directly (94.2 to 7.5 on Qwen2.5-VL-3B, 92.0 to 2.5 on Gemini 3.8 Flash), decompose the score by answer length, and give a deterministic scorer tested against an attack. VLMEvalKit’s optional use_verifier path (Duan et al., 2024) replaces string matching with a model judge the user supplies; PEDANTS (Li et al., 2024) makes a similar answer-equivalence judgement with a learned classifier, and it scored the page-dump attack at 98.6.

Sclar et al. (2024) show that semantically equivalent prompt formats move accuracy by margins comparable to differences between models, and Mizrahi et al. (2024) argue that single-template evaluation misstates the state of the art and propose evaluating over a distribution of prompts. Both

place the sensitivity in the model’s response to the prompt; in our data 73.3 of the 86.6 ANLS points the 3B lost without the suffix were credited by the span scorer on the same generations, so most of that change came from how ANLS scored the responses.

9 Conclusion

Removing the suffix that makes DocVQA answers terse cost Qwen2.5-VL-3B 86.6 ANLS points and Gemini 3.8 Flash 89.5, and under that suffix Markdown emphasis markers cost Claude Opus 5.5 10.9 points. Scoring the same generations for whether the answer is present, with a gate against the page-dump attack, reduced Qwen2.5-VL-3B’s range across the five suffixes from 86.6 to 13.3 points. Because removing the suffix moved every hosted model’s score by more than the 17-point spread between their suffix scores, and Markdown handling moved Claude Opus 5.5 by 10.9 points, we recommend reporting the prompt suffix and output post-processing with every document-VQA score.

Limitations

The gate reintroduces length sensitivity in smaller measure. Without a suffix the 7B scored 73.0 gated against the 3B’s 82.2, although their ungated span scores were 89.0 and 88.8 and the shares of all their answers (not only zeros) containing a gold 88.1% and 86.9%; the gap comes from the gate applying to 17.4% of the 7B’s answers against 7.5% of the 3B’s. We selected m on the same DocVQA sample we report it on. The DocVQA results use a 1,200-question sample rather than the full test set, whose labels are available only through the evaluation server.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-VL technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.
- Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluís Gomez, Marçal Rusiñol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas.

- 2019a. ICDAR 2019 competition on scene text visual question answering. In *International Conference on Document Analysis and Recognition (ICDAR)*. ArXiv:1907.00490.
- Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluís Gomez, Marçal Rusiñol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. 2019b. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. ArXiv:1905.13648.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. [VLMEvalKit: An open-source toolkit for evaluating large multi-modality models](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.
- Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Boyd-Graber. 2024. [PEDANTS: Cheap but effective and interpretable answer equivalence](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA. Association for Computational Linguistics.
- Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. 2025. SmolVLM: Redefining small and efficient multimodal models. In *Second Conference on Language Modeling (COLM)*. ArXiv:2504.05299.
- Ahmed Masry, Do Xuan 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*. ArXiv:2203.10244.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2022. InfographicVQA. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. ArXiv:2104.12756.
- Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. ArXiv:2007.00398.
- Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over scientific plots. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. ArXiv:1909.00997.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? a call for multi-prompt LLM evaluation](#). *Transactions of the Association for Computational Linguistics*, 12:933–949.
- Armineh Nourbakhsh, Siddharth Parekh, Pranav Shetty, Zhao Jin, Sameena Shah, and Carolyn Rose. 2025. [Where is this coming from? Making groundedness count in the evaluation of document VQA models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5341–5361. Association for Computational Linguistics.
- 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*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*. ArXiv:2310.11324.
- Guanhua Ye, Niu Jingbin, Yan Li, Meiyu Liang, Zhe Xue, Yingxia Shao, and Yawen Li. 2026. [What does your short-answer VQA score actually measure? Evaluator-dependent instability in multimodal short-answer benchmarks](#). *Preprint*, arXiv:2607.10240.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. [LMMS-Eval: Reality check on the evaluation of large multimodal models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916. Association for Computational Linguistics.

A Reproducibility

All claims about the local models regenerate from the released generations with `python3 scripts/verify_claims.py`; each entry in `claims.json` names the function and the generation file that produce it, and the script exits non-zero if a recorded value fails to reproduce. The matched-cap, scorer, judge, hosted-model and canonicalization analyses regenerate from the scripts and generation files in the release’s analysis directory.

B Harness Sources

```
lmms-eval at commit 42296f8: in
lmms_eval/tasks/docvqa/_default_
template_docvqa.yaml, lines 13–19 set
post_prompt to "\nAnswer the question
```

using a single word or phrase." under default and to " Answer:" under qwen_vl, with max_new_tokens: 32 and temperature: 0; in api/task.py, lines 888–891 select the entry keyed by the model name passed to the harness and otherwise fall back to default. The InfographicVQA template uses the same default string, and the ChartQA task uses "\nAnswer the question with a single word." with the same qwen_vl override. VLMEvalKit at commit 6f03707: in vlmeval/dataset/image_vqa.py, lines 86–91 append "\nAnswer the question using a single word or phrase." to every DocVQA question, and vlmeval/config.py runs Qwen2.5-VL-3B-Instruct with use_custom_prompt=False, so that string is what the model sees. Appendix B.4 of the Qwen3-VL technical report (Bai et al., 2025a) gives <image> {question} Answer the question using a single word or phrase. as the prompt for DocVQA, InfographicVQA (InfoVQA in the report) and ChartQA.

C Matched Token Cap

	at cap	ANLS	cut to 32	stopped
3B, no suffix	3.0	7.5	7.6	7.8
7B, no suffix	8.9	3.6	4.0	4.0

Table 6: ANLS of no-suffix generations (cap 64; *at cap* is the percentage that reached it), rescored after re-tokenizing each with the model’s tokenizer and keeping the first 32 tokens, and restricted to generations that stopped before the cap. Default-suffix ANLS for comparison: 94.2 and 96.2.

D Stratified Containment Analysis

Restricting to golds of at least two tokens gave a contained share of 79.2% under " Answer:" and 86.3% with no suffix; restricting to golds of at least eight characters gave 77.0% and 85.7%. A raw-substring definition agreed with the token definition to within 2 percentage points in every condition. Canonicalizing numbers and dates before matching (thousands separators, currency and percent signs, number words up to twelve, and dates to month/day/two-digit year) raised the contained share by 0.6 to 2.8 percentage points in each condition with more than 150 zero-scored answers, for example \$19,902 million against the gold 19,902 and February 4, 1982 against Feb. 4, 1982; the main text uses no canonicalization.

E Containment Judge

The judge saw: *You are checking answers to questions about a document. Question: {q} Reference answer: {g} Response: {p} Does the response commit to the reference answer as its answer to the question? Formatting, extra words, and restating the question do not matter. Say NO if the response gives a different answer, says the reference is wrong or not the answer, or offers the reference as one of several candidates without choosing it. Reply with exactly one word: YES or NO.* Thinking was disabled and every reply parsed. The controls used 200 randomly chosen no-suffix items from the 3B, with the distractor taken from the gold of another question.

F Hosted Models

OpenRouter identifiers openai/gpt-6-sol, google/gemini-3.8-flash, anthropic/claude-opus-5.5 and qwen/qwen3.8-max-0902; every response reported the requested identifier. Each request carried the page as a PNG and the question with or without the default suffix, at temperature 0, with the lowest reasoning setting the API exposes and a 1,024-token output limit.

G Batch Size

Batched decoding with left padding changes Qwen2.5-VL’s multimodal rotary position indices, and on five of eight spot-checked items it produced truncated prefixes such as '1/8/' for the gold 1/8/93, while every single-item generation ended on an end-of-turn token before the cap. Batched decoding also added stray whitespace to 10.7% of outputs, which would enter the whitespace term of §2, so all local generations use batch size one.