

Documents That Cite Themselves: Reading Forms at Human Level, and Knowing Which Answers to Trust

Madhu

Kansas City · Independent · madhudream.dev

ABSTRACT

A system that reads a document correctly 95 times out of 100 and is equally sure about all 100 has saved nobody any work: the reviewer still checks every value to find the five. We argue that the useful problem in document extraction is not reading but sorting, and we report a system built on that premise. Every value it fills carries a citation, a rectangle on the source page drawn around the words it was read from, and one of three verdicts: accept, accepted after a second reading, or a person should look. On the reading task the system scores 97.53 ANLS on all 5,349 questions of the DocVQA validation set with nothing removed, and 93.56% of its answers are identical to the gold answer; the best published entry on the official test leaderboard scores 97.29 and human performance is 98.11. On our grounding benchmark the composite score rose from 79.1 to 93.5 over 36 kept changes; 93.3% of fields are filled with no person in the loop at 100.00% precision, and 99.5% of citations land on the right ink. None of the gain came from a larger model. We describe the measurements, the experimental discipline behind them, the cost, and the limits. The method is described in a companion paper available on request.

Keywords: document understanding, visual question answering, DocVQA, grounded extraction, citations, selective prediction, trust, evaluation.

1 Introduction

Anyone can rent a model that reads documents at about 94%. That is a commodity, priced by the page. What a business cannot buy off the shelf is accountability: for each value the system fills, whether it may be trusted without a person looking, and if not, where on the page the person should look.

Consider a form with a hundred fields. A reader that is right on 95 of them and equally confident about all 100 has not reduced the reviewer's work, because the reviewer cannot tell the 95 from the 5 without checking all 100. The same reader, if it could say honestly "these 84 are certain, these 16 need your eyes", would have removed 84% of the work. The second system is worth more than the first at the same accuracy. Its value is in the sorting.

This paper reports a system built on that premise. It reads, it points, and it judges. Reading means answering a question about a page. Pointing means returning the rectangle on the page where the answer was read, so that the reviewer's eye lands on the evidence rather than on a blank form. Judging means attaching one of three verdicts to every value. Everything that follows serves one claim: that the verdicts are measured, not asserted.

1.1 Summary of results

- **Reading at human level.** 97.53 ANLS on all 5,349 DocVQA validation questions, nothing removed; 98.35% of answers accepted as matching; 93.56% identical to the gold answer. The best published test-set entry is 97.29; human performance is 98.11 (Section 3).
- **Pointing and judging, measured.** On a grounding benchmark of business documents the composite score rose from 79.1 to 93.5 on development documents and from 79.7 to 90.9 on held-out documents. Seven audits on freshly generated documents were all positive, from +4.7 to +9.0 points (Section 4).
- **In a buyer's terms.** 93.3% of fields filled with no person in the loop at 100.00% precision, zero wrong auto-fills in 1,181 fields; 99.5% of citations on the right ink (Section 4).
- **No larger model.** Every point was earned by engineering around the model and confirmed on questions no change was designed on (Section 5).
- **Cost known to the cent.** Under half a cent per question at the reported accuracy, with every result's cost recorded (Section 6).

1.2 Terms

ANLS, the average normalised Levenshtein similarity, is the standard score for document question answering: full credit for an exact match, partial credit for a near match, nothing for an answer that differs from the gold answer by more than half its characters [2]. *Development*, *holdout* and *second holdout* are fixed, disjoint subsets of the validation questions; a change may be designed on training questions and checked on development, but the holdouts decide. An *audit* is a fixed random sample re-scored on a stated date. A *citation* is a rectangle in page coordinates around the printed words a value was read from. A *verdict* is one of *accept*, *accepted after a second reading*, or *a person should look*.

2 What is measured, and how

2.1 Reading: DocVQA

DocVQA is the public benchmark for answering questions about document images [1]: 5,349 validation questions over scanned forms, letters, tables and reports. We report ANLS over every validation question with nothing removed, and the stricter share of answers identical to the

gold answer. Public comparisons for this benchmark are made on the validation set; the hidden test set is scored only by the organisers.

2.2 Pointing and judging: a grounding benchmark

No public benchmark scores a citation rectangle or a trust verdict, so we built one: generated business documents (invoices, statements, forms, tables) in which every field's value, location and arithmetic relations are known. It yields a composite of two parts, *ground* (does the rectangle cover the right words) and *trust* (was the verdict right), with a wrong auto-fill penalised far more than an unnecessary review. It also yields the two numbers a buyer asks for: the share of fields filled with no person in the loop, and the precision of those fills.

2.3 Discipline

Every change to the system was one experiment: a single change, run against the fixed benchmark, kept or discarded by a rule written down before the first run. For reading, a change was kept only if it improved the development split, improved the 2,865 held-out questions, and the chance that the gain over all 4,365 scored questions was luck fell below ten percent. Every paid model call passes through a cache with a ledger, so any re-run is free and the cost of every result is known. Discarded experiments stay in the log beside the kept ones.

3 Results: reading

Table 1 gives the ANLS score on each split for the 2025 system and the current one. The second holdout was never run under the 2025 system, which is why its earlier value is absent and why the earlier overall figure of about 94.9 covers 3,685 questions rather than all 5,349.

Table 1. ANLS by split. The audit row is a fixed random sample re-scored after the last kept change.

split	questions	2025 system	current
development	1,500	95.42	97.74
holdout	1,201	94.80	98.04
second holdout	1,664	—	96.72
audit sample	984	93.95	98.00
all validation questions	5,349	—	97.53

Over all 5,349 questions, 98.35% of answers are accepted as matching (ANLS of at least 0.5) and 93.56% are identical to the gold answer. For context, the best published entry on the official test leaderboard scores 97.29 (March 2026), and human performance is 98.11 ANLS, or 94.36%

exact match. The system is level with the best published systems and within half a point of people.

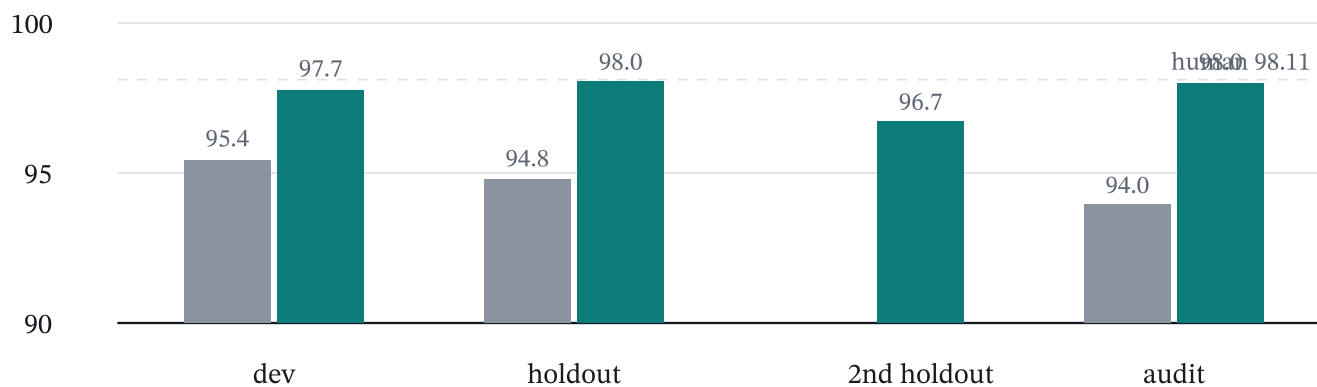


Figure 1. ANLS on each split: the 2025 system (light) against the current one (dark). The dashed line is human performance.

4 Results: pointing and judging

On the grounding benchmark the composite score rose from 79.1 to 93.5 on the development documents and from 79.7 to 90.9 on held-out documents, over 36 kept changes. The two parts moved together: trust from 81.3 to 92.9 and ground from 76.9 to 94.1. Seven audits on freshly generated documents were all positive, from +4.7 to +9.0 points; the latest is +8.5 with a 95% interval of +6.2 to +11.1.

In the terms a buyer uses: 93.3% of fields are filled with no person in the loop at 100.00% precision, that is, zero wrong auto-fills in 1,181 fields, from 83.6% coverage with two wrong auto-fills at the baseline; and 99.5% of citations land on the right ink, six wrong rectangles and none missing, from 106 wrong rectangles at the baseline.

These documents are generated, not collected. The benchmark exists so that each change can be measured before it is tried on real documents, which is the next measurement (Section 7).

5 What the results show

None of the gains came from a larger model. Several larger models, tried as readers, scored below the system rather than above it. The points were earned by the engineering around the model: how the page is presented to it, how more than one reading is obtained and reconciled, how an answer's wording is normalised against the benchmark's convention, and how evidence that costs nothing, such as a date that parses or a total that equals its parts, is turned into

a verdict. Each of those is a measured decision in the companion paper. Here we record the three findings that generalise.

The reader sets the ceiling. Everything built around a reader can only recover points the reader itself nearly had. Spend on the reader first, then on what surrounds it.

Disagreement is a signal, not a failure. When independent readings of the same page differ, the difference is the system telling the reviewer where to look. A trust verdict built on that signal is calibrated by construction, because it is measured on the same questions it is used on.

Design on training questions, confirm on unseen ones. Ideas chosen by inspecting the hardest development questions failed on unseen questions every time they were tried. Rules designed on the training split carried over every time. The holdouts, not the developers, decide what is kept.

6 Cost

At the reported accuracy the reading pipeline costs under half a cent per question, with every call cached so that a repeated question costs nothing. The system records the cost and the latency of every document it processes; a buyer sees the bill per page, not a price list.

7 Limitations

The grounding benchmark is generated, and its development score sits 2.6 points above its holdout score, which suggests some fitting to the development documents; the next measurement is on collected documents. The reading results are on the validation set, as all public comparisons for this benchmark are, and not on the hidden test set. The wording normalisation is fitted to the benchmark's convention of the shortest printed answer and would be re-fitted for a customer's schema.

8 Discussion

Reading documents at human level is now a matter of choosing a strong reader and presenting the page to it well. The remaining value is in what happens outside the model: a citation for every value, a verdict for every value, and the discipline to measure both on questions nobody designed against. This version describes what was measured and why the measurements are credible. It does not describe the readers, the settings, the normalisation rules, the verdict thresholds or the 36 changes to the grounding engine. A companion paper does, with every kept and discarded experiment and the cost ledger, and is available to investors and partners on request.

References

1. M. Mathew, D. Karatzas, C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images. *WACV*, 2021.
2. A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusiñol, E. Valveny, C. V. Jawahar, D. Karatzas. Scene Text Visual Question Answering. *ICCV*, 2019.
3. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei. Scaling Laws for Neural Language Models. arXiv:2001.08361, 2020.

Every number in this paper is read from the project's own logs: the audit log, the experiment logs and the reports. Written October 2026, revision 2.