Pitfalls in Evaluating Automated Building-Plan Review

Rohan Movva
Stanford University

Abstract

Automated review of building-permit drawings is being sold to municipal permit offices, and open-ended review of architectural document sets has begun to be benchmarked. Labelling such data requires licensed reviewers, so a natural shortcut is to generate drawings and inject known violations. We report that this shortcut fails, and how. Injected defects have consequences under other code provisions, so the drawing contains defects the answer key does not name and a model reporting them is scored wrong for being right. Rederiving labels by evaluating every rule against the generated geometry showed that our supposedly compliant baseline was never compliant: one deterministic template property fails in all 22 cases, including all six intended clean controls, so the benchmark contained no defect-free case. That single property supplies 22 of 48 labelled defects, and it is one observation replicated, not many defects. Separately, on a real 154-sheet municipal permit set, we measure how much of a sheet survives rasterisation: rendering a 42×30 inch sheet to 1024 pixels recovers 0.6 % of its text and 1536 pixels recovers 6.0 %, scored against the PDF's own vector text layer. The low-resolution end is underwritten by geometry rather than by the reader: 9.94 pt median text at 1536 pixels across a 42-inch sheet is 3.5 pixels of cap height. A pilot detection experiment on the generated benchmark found one between-condition effect that survives correction, and it concerns output volume rather than accuracy: supplying the governing code text reduced findings per case by -3.09, in 18 of 22 cases. We make no claim about a deployed system, no claim of agreement with municipal reviewers, and no claim of generalisation beyond one jurisdiction.

1 Introduction

A plans examiner receives a permit submittal, often more than a hundred large-format sheets spanning architectural, structural, mechanical, electrical, plumbing and energy disciplines, and must find the defects in it. No question is supplied. The number of defects is unknown. Some are life-safety failures and most are not, and the examiner must produce few enough findings that the applicant can act on them.

This measurement structure is that of free-response detection, in which a reader places an unknown number of confidence-rated marks on each case, rather than that of document question answering. Precision is not well defined without fixing what counts as an opportunity for error, and the human record one would score against is itself partial.

Open-ended review of architectural sets is already benchmarked [6, 18], and this paper does not compete with those resources. It reports what happens when one tries to avoid the cost of expert labelling by constructing ground truth instead, and how much of a production sheet is available to be reasoned over in the first place.

Contributions.

1. Evidence that constructively labelled benchmarks inherit the incompleteness they are built to avoid, with the mechanism and two worked instances from our own generator, and the corollary that a recurring template property is one observation rather than many defects (Section 5).
2. A model-free measurement of text recovery against render resolution on a real municipal permit set, whose low-resolution end rests on cap-height geometry rather than on the reader used (Section 4).
3. Two reporting devices for review tasks whose incumbent human record is partial: precision as an interval over an explicit unadjudicated class, and a matcher self-audit that recovered a false-alarm rate we had overstated fourfold (Sections 3 and 5).

Our reporting of recall against findings per case is an import of free-response ROC analysis from medical imaging [3, 2] and is not offered as a new metric.

2 Related Work

Open-ended review of drawings. RedlineBench scores models on a real drawing set containing 58 known issues across seven categories, penalising false alarms, with a public leaderboard [6]. AEC-Bench releases 196 task instances over intrasheet, intradrawing and intraproject scopes together with its harness, and reports that locating the relevant sheet is the binding difficulty [18]. We claim neither a benchmark contribution over these nor the cross-sheet framing.

Drawing benchmarks and drawing understanding. Recent benchmarks evaluate multimodal models on architectural and construction drawings [11, 9, 12], reporting strong text extraction, weak symbol recognition, and a large gap to expert performance. Each supplies the question. Floor-plan

parsing is well developed on clean raster plans [10, 14, 23] and on native vector CAD [5], extracting elements rather than review decisions.

Automated compliance checking. This literature largely presumes a semantically rich BIM/IFC model exists [20, 25, 1, 24, 17]. Deployment studies report that assumption as the practical obstacle [26], and real submittals are PDF drawing sets.

Resolution and long context. That input resolution governs text reading in multimodal models is established, and is the motivation for tiling and any-resolution encoders [15, 13]. Long or partly irrelevant context can also degrade accuracy [16, 19], which is the alternative explanation for our one surviving between-condition effect.

Free-response detection and selective prediction. Medical imaging established the measurement structure for finding an unknown number of items per case, together with its statistics [3, 2, 4]. Risk-coverage analysis and abstention are studied for single-input, single-label decisions [7, 8, 21, 22].

3 Problem Formulation

Let a case c be a plan set and $V(c)$ the set of defects it contains. A system emits findings $F(c)$, with $|F(c)|$ unknown a priori. A matching relation assigns findings to defects, at most one each way.

Precision as an interval. For real data $V(c)$ is not observed. What is observed is $R(c) \subseteq V(c)$, the subset a human reviewer recorded. Findings outside $R(c)$ have unknown status: scoring them as errors biases precision downward by an unknown amount, and scoring them as discoveries biases it upward. We place them in an unadjudicated class U and report

$$p_{\text{lo}} = \frac{|TP|}{|TP| + |FP| + |U|}, \quad p_{\text{hi}} = \frac{|TP| + |U|}{|TP| + |FP| + |U|}.$$

The interval width is the adjudication debt. It is not a confidence interval and does not shrink with more cases, only with expert labelling.

Burden and clustering. Because $|F(c)|$ is unbounded, recall is only interpretable against a stated review budget, so we report findings per case and recall at a budget K: the fraction of defects surfaced within the K highest-confidence findings. Findings within a plan set are correlated, and so are repeated instances of a generator property across cases, so all paired tests are computed with the case as the unit of analysis.

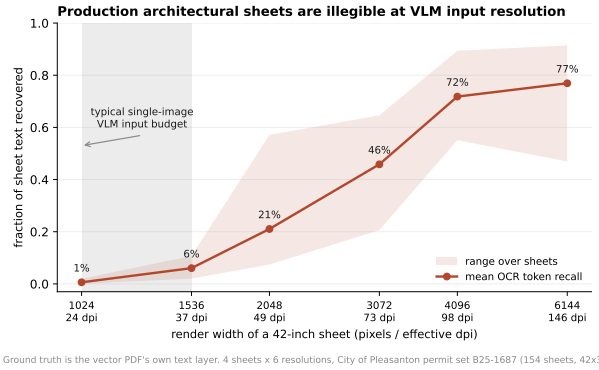

Figure 1: Text recovered from four real production architectural sheets as a function of render resolution, macro-averaged. Ground truth is the vector PDF's own text layer. Sheets are drawn from a 154-sheet municipal permit set (42×30 in). **Shaded band:** min to max across the four sheets.

Design sensitivity. Under a simulation on the benchmark's per-case structure with a symmetric alternative, the design resolves paired recall differences of roughly 0.15 at a base rate of 0.938 and roughly 0.30 on the 19 non-degenerate defects at a base rate of 0.842. Nulls reported below are therefore weak nulls.

4 Text Recovery Against Resolution

The transferable quantity here is geometric, not instrumental. Median drawing text on this set is 9.94 pt over 3384 spans. Rendered at 1536 pixels across a 42-inch sheet, that is 3.5 pixels of cap height. No reader recovers glyphs at that size, so the low-resolution conclusion does not depend on which reader is used.

A vector-native permit PDF carries an exact record of its own text, which lets one instrument that geometry without a model: render at a chosen resolution, OCR with Tesseract, and score recovered tokens against the text layer. Figure 1 gives the curve, macro-averaged over 4 sheets. Recovery is 0.006 at 1024 pixels and 0.060 at 1536. These two points are stable across aggregation choices: micro-averaging gives 0.006 and 0.062, and excluding the atypical sheet discussed below gives 0.008 and 0.074.

The rest of the curve is not stable. One sampled sheet carries only 49 ground-truth tokens, and it dominates the mid-range. Excluding it moves the 2048-pixel point down from 0.211 to 0.090, and the 6144-pixel point up from 0.769 to 0.869. Token precision is also non-monotone and never exceeds 0.62, rising from 0.369 at 1024 pixels to 0.615 at 6144. We therefore report the high-resolution end as a property of this instrument rather than of the sheets.

Consequence, and its limits. An architecture that presents a whole sheet as one low-resolution image discards the text before reasoning begins. Production multimodal APIs already tile internally [15, 13], so this is a domain-specific quantification of a known constraint rather than a new one. Its use is that it gives the constraint a number for construction documents, and it implies that detection failures measured at such resolutions are not evidence about reasoning. Four sheets is a small sample and they were chosen to span disciplines rather than by a random draw over the 154.

5 Constructive Ground Truth and Its Limits

Expert adjudication is the correct way to label review data and is also the bottleneck. A common alternative is to inject known defects into generated artefacts, as industrial anomaly detection does. We built such a generator. It emits residential floor-plan sheets whose defects are departures from numerically stated code thresholds, 5.7 square feet of net clear egress opening, 20 and 24 inch minimum opening dimensions, 36 inch corridors, 32 inch exit doors, 84 inch ceilings and 8 percent glazing, or the absence of a required element.

The shortcut fails. Taking ground truth to be the list of injected defects is wrong, in a way we expect to hold for any constructive benchmark. Reducing a bedroom's egress window to violate the 5.7 square foot rule also drops that room below the 8 percent natural-light minimum. The drawing then contains two defects while the key names one, and a model reporting the second is scored wrong for being correct.

Rederiving labels by mechanically evaluating every rule against the generated geometry raised the label count from 20 to 48 and exposed a second failure. The baseline we had treated as compliant was never compliant: the living room carries a single fixed 20.0 square foot window over a 255 to 302 square foot room, giving 6.6 to 7.8 percent against an 8 percent minimum, for every reachable seed. It fails in all 22 cases, including all six intended clean controls, so the benchmark contained no defect-free case and the hallucination measurement those controls existed to provide was not obtained.

One property is not many defects. That single template property supplies 22 of the 48 labelled defects, and the natural-light family as a whole supplies 29. Counting it once per case inflates every aggregate over the label set and, as Section 6 shows, can manufacture a between-condition difference on its own. Constructive ground truth must be computed from the artefact rather than inherited from the generator's intent, clean controls are clean only with respect to defects the designer considered, and a property that is constant across cases must be counted once.

Reasoning type	sheet only	+ code text
area aggregation†	29/29	22/29
dimension comparison	3/3	3/3
presence or absence of an element	2/2	2/2
note lookup	2/2	2/2
note text	2/2	2/2
cross-sheet consistency	2/2	2/2
callout arithmetic	5/6	5/6
symbol presence	0/2	0/2

Table 1: Detection by reasoning type, 22 cases. † 22 of these 29 are one recurring template property, not 29 defects. Every other row is identical across conditions. Over all 48 items recall is 0.938 and 0.792; over the 19 non-degenerate items it is 0.842 in both.

Matching, and a self-audit. The matcher requires a finding to name both the subject and the failing attribute, and assigns one-to-one. Its subject and attribute patterns were written before any model was run and are unchanged; its exclusion patterns were revised after reading output from the sheet-only condition. An initial version over-fired, scoring a carbon-monoxide-alarm finding as asserting six unrelated families because it contained the tokens "alarm", "garage" and "no". Correcting it moved the reported confirmed-false rate from 2.78 to 0.67 per case on the same nine pilot cases, with the unadjudicated share rising from 39.5 to 64.0 percent, and widened the precision interval from [0.263, 0.688] to [0.247, 0.926]. Honest scoring widened the interval rather than narrowing it.

Two biases remain and are not corrected. The room tie-break applies only when the ground-truth room is a bedroom, so for the living room no room check runs and any finding naming glazing plus a shortfall term is credited to that violation regardless of which room it discusses. This is an upward bias on precisely the family that carries the aggregate difference below. And an exclusion written against one condition's phrasing does not bias two conditions equally, so the claim that exclusions can only lower recall holds for a single number and not for a paired comparison.

6 Detection on the Generated Benchmark

The model is not told that a defect exists, how many there are, or which families are possible. The condition that supplies the code text necessarily discloses the families in scope, which is part of what that condition manipulates.

Where the conditions differ, and where they do not. Over all 48 items recall is 0.938 without the code text and 0.792 with it. All 7 discordant defects belong to the natural-light family and none run the other way. Over the 19 non-degenerate defects the two conditions detect an identical set and miss an identical set, giving 0.842 in both arms with 0

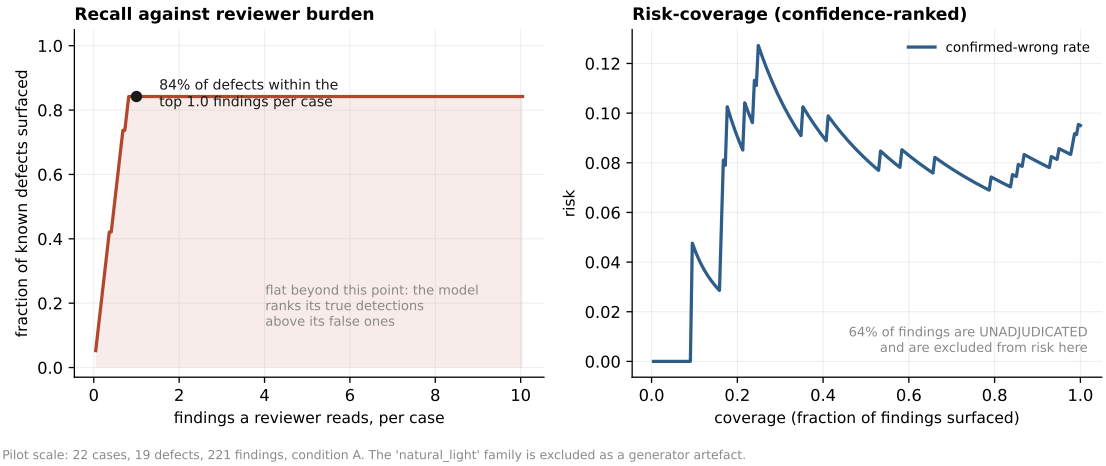

Figure 2: **Left:** defects surfaced against findings a reviewer must read, sheet-only condition. **Right:** risk against coverage under confidence ranking. The natural-light family is excluded from the recall denominator in both panels.

discordant items. The aggregate difference is produced by the recurring template property of Section 5, and it is not a difference in detection. Critical-severity recall is 0.750 in both conditions (9 of 12, Wilson [0.47, 0.91]) and major-severity recall is 1.000 in both, so nothing outside the moderate class moved.

Tests, at the case level. An exact sign test over cases gives 5 cases favouring the sheet-only condition and 0 the other way, $p = 0.0625$, and an exact sign-flip permutation over all 2^{22} assignments of the per-case recall differences gives a mean of -0.1439 at $p = 0.0625$. Neither reaches significance. An item-level McNemar test on the same data gives $p = 0.016$, and a percentile cluster bootstrap gives $p = 0.014$; both treat replicates of one template property as independent observations and undercover here. We report the exact case-level tests as primary.

Under a matched review budget. Truncating both arms to their K highest-confidence findings narrows the aggregate difference but does not remove it: 0.896 against 0.771 at $K = 6$ and 0.917 against 0.792 at $K = 7$, a difference of -0.125 against -0.146 unbounded. Over the non-degenerate defects the two arms are identical at every K. The operating-point control therefore does not explain the difference; the degenerate family does.

The one effect that survives. Supplying the code text reduced output volume. Findings per case fell from 10.04 to 6.96, a mean difference of -3.09, with fewer findings in 18 of 22 cases and more in 3 (exact sign-flip $p = 0.00033$). Enumerating the family of inferential tests this paper reports as $m = 3$ over findings per case, false-alarm share and case-level recall, only findings per case is rejected under Holm-Bonferroni; false-alarm share ($p = 0.191$) and recall ($p = 0.0625$) are not.

This is a shift in output behaviour, not in accuracy, and its cause is not identified. The condition changes prompt content and prompt length together, and no length-matched or scrambled-provision control was run, so content and context dilution cannot be separated [16, 19]. It is also a single run per cell under an uncontrolled sampling distribution, so it is provisional pending the same-condition replicate described in Section 8.

Where detection failed. Symbol-presence detection failed on both cases in both conditions. Each required noticing that a smoke alarm is missing from a sleeping room, and in each the model instead reported that no alarm is shown outside the sleeping area, which is true and is a real requirement. The contrast with the other absence family, which was detected 2/2, is not a difference in reasoning: deleting an egress window also removes a row from the drawn window schedule, a text table, whereas removing a smoke alarm deletes a 6-point circle and a 6.5 pt label from the plan graphic while the general note announcing smoke alarms remains, and that note is what the model is observed to report from.

Precision is reported as the interval [0.217, 0.899] because 63.8 % of the 221 findings are unadjudicated. Confidence ranks matched findings above confirmed-wrong ones, and Figure 2 shows most scoreable defects surfacing in the top findings per case. Calibration is poor, with an expected calibration error of 0.343, so the ordering is usable while the magnitudes are not probabilities.

7 Document Anchoring

We hold 326 findings about a real project, produced by a prior pipeline of unknown model, prompt and configuration, and recovered from a 37-page document. They were not pro-

duced by any system described in this paper. This section demonstrates an audit method on that corpus and does not measure an anchoring rate for automated review generally.

For each finding we extract code sections and dimensioned quantities and test their presence in the real plan set's text layer. 48.8 % are not textually testable. Of the 167 that are, 58.1 % have at least one anchor present, and 21.5 % of all findings have none anywhere in 154 sheets. The findings describe a first submittal and we hold the fourth, so a missing anchor may mean the defect was corrected rather than invented. Anchoring is also not correctness.

8 Limitations

No system. No production plan-review system was accessible to this work and we make no claim about one.

Ground truth. The verified municipal record available to us is a single comment, a planning documentation item rather than a code violation. No licensed professional adjudicated any finding here.

Replication is an unmet precondition, not a caveat. There is one run per condition and case, 22 and 22, with temperature, seed and served model version uncontrolled. Within-condition variance is never estimated, so the intervals reported omit the dominant noise term and no between-condition number here is separable from run-to-run drift. The recall, burden and false-alarm comparisons are all conditional on a replicate that has not been run.

One template. All 22 cases share one seven-room topology from a single generator function, varying only bedroom dimensions, hall width and the injected family, with seven openings fixed in position and type. The benchmark cannot separate detection from familiarity with one template, and this is the same mechanism by which one latent defect became 22 of 48 labelled defects.

Resolution was held constant. Resolution is controllable in the generator but every detection run used a single render at 150 dpi (2401×1531 pixels). Downsampling and tiling between the file-read tool and the encoder are undocumented, and belong with temperature and seed as uncontrolled variables.

Unmeasured choices in our own prompt. The review prompt states the four-digit opening-callout convention, identifies the smoke-alarm symbol, and instructs the model not to invent findings. Each is plausibly worth more than any system variable we studied, and the symbol legend was present in the prompt for the family that failed 0/2. None was ablated.

The false-alarm composition is confounded. 19 of the 32 findings asserting a satisfied family in the code-text condition concern hallway width, against 8 of 21 in the sheet-only condition, and the matcher's pattern for that family includes the threshold token supplied verbatim to that condition. A finding that merely restates a supplied requirement is therefore scored as a false alarm.

Reproducibility of Section 6. The scored artefact stores finding indices only. The per-finding text, the review prompt, the output schema and the agent journals are not in the repository, so precision, calibration and the selective-review curve are not recomputable by a reader. Condition and case are recovered post hoc by pattern match over agent transcripts rather than from a dispatch manifest.

Scale. One jurisdiction, one code edition, one project, one building type.

9 Experiments Not Run

Four, in the order they should be done, none of which is reported above. First, the sheet-only condition replicated on the same 22 cases, to establish the run-to-run noise floor stratified by family; every between-condition statement here is conditional on it. Second, detection at 1024, 1536, 2048 and 2401 pixels on the same cases, which is the experiment that joins Sections 4 and 6 and turns a juxtaposition into a causal claim. Third, a length-matched or scrambled-provision control, separating supplied content from context dilution. Fourth, the model under test transcribing the same four sheets against the same text layer, removing the Tesseract proxy.

10 Conclusion

Constructing ground truth by injecting defects into generated drawings does not escape the incompleteness that motivates it. Injected defects have consequences the key does not name, controls are clean only with respect to defects the designer imagined, and a property that holds in every case is one observation rather than many. In our own benchmark that last failure was enough to manufacture a between-condition difference that dissolves when the property is counted once. Separately, production sheets carry almost none of their text at the resolutions a whole-sheet image affords, which bounds what any downstream reasoning can be measuring. Both problems are measurable, and we have tried to measure them rather than route around them.

A Harness and Reproducibility

Model access. No frontier-model API key was available in the environment where these experiments were run. The model under test is invoked through an agent harness: a fresh agent instance receives the prompt, opens the sheet image with a file-read tool, and returns a structured finding list validated against a JSON schema.

Consequences, which bound every detection number in this paper:

- The serving model version is not pinned in the artefact and can differ from the configured one.
- Temperature and random seed are not controllable. Repeat runs estimate variance; they do not replay.
- Token and cost accounting is available only at workflow granularity.
- One model family only, so no cross-model claim is possible.

This is why the load-bearing result of the paper (§4) is model-free.

B Legibility Measurement

Ground truth is the vector PDF's own text layer, extracted per page with PyMuPDF. Each sampled sheet is rasterised with PyMuPDF at the target dpi and OCR'd with Tesseract 5.3.4. Both the text layer and the OCR output are tokenised by the same regular expression (`[A-Za-z0-9][A-Za-z0-9/.\-']{2,}`), uppercased, and compared as sets; recall is $|GT \cap OCR|/|GT|$.

Sheets were sampled across disciplines (architectural, structural, plumbing) from a 154-sheet permit set; sheets with fewer than 40 ground-truth tokens were skipped. The font-size measurement covers 3,384 spans sampled from eight sheets, converted to rendered cap height as $\text{pt}/72 \times \text{dpi} \times 0.7$.

C PlanLab

Each case is a deterministic function of an integer seed. Geometry is rendered vector-natively through matplotlib and rasterised at a chosen dpi, so resolution is a controlled variable rather than a property of the source.

Twelve violation families are supported, each derived from a numerically stated code threshold or a required element's absence. Every family records its section number and the quoted requirement in `planlab.CODE_RULES`, so each label traces to the sentence it derives from.

Ground truth is computed, not declared. `planlab.evaluate_rules` mechanically evaluates every rule against the generated geometry and returns the violations actually present. This differs from the list of injected violations, and the difference matters: injected defects have code-relevant side effects (§5).

Regeneration.

```
python3 code/harness/build_benchmark.py
```

reproduces the dataset byte-deterministically from seeds 1000–1021.

D Matching

The matcher is fixed before any run. A match requires the finding to name the subject and the failing attribute; assignment is greedy and one-to-one. Per-family exclusion patterns prevent keyword collisions (for example, a carbon-monoxide-alarm finding must not match the smoke-alarm family, and a finding about openings in a garage separation must not match the family concerning the separation's material). Every exclusion makes matching strictly harder and can therefore only lower reported recall.

Strict and lenient variants are reported for the three egress-dimension families. On the sample reported here they agree exactly.

E Statistics

All intervals resample *cases*, not findings, because findings within a plan set are correlated. Comparisons between conditions evaluated on the same cases use a paired cluster bootstrap in which clusters are resampled once and applied to both arms. Wilson intervals are used only where observations are genuinely independent. Families of comparisons carry Holm–Bonferroni correction.

F Artefacts

- `results/RESULTS_LEDGER.csv` — every reported number, with its n, the artefact it came from, and the commit it was computed at.
- `benchmark/planlab_v1/` — cases, ground truth, and the code rules each label derives from.
- `research_clearblock/` — audit, protocol, error analysis, hostile review, and the record of three defects found in our own apparatus.

The real permit set is not distributed: it is NDA-covered, third-party copyrighted professional work. PlanLab is fully releasable.

References

[1] Robert Amor and Johannes Dimyadi. The promise of automated compliance checking. *Developments in the Built Environment 5:100039*, 2021.

[2] Andriy I. Bandos, Howard E. Rockette, Tao Song, and David Gur. Area under the free-response roc curve (froc) and a related summary index. *Biometrics, 65(1), 247-256*, 2009.

[3] Dev P. Chakraborty. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Physics in Medicine and Biology, 51(14)*, 2006.

[4] Dev P. Chakraborty and Hong-Jun Yoon. Jafroc analysis revisited: figure-of-merit considerations for human observer studies. *Proc. SPIE Medical Imaging 7263*, 2009.

[5] Zhiwen Fan, Lingjie Zhu, Honghua Li, Xiaohao Chen, Siyu Zhu, and Ping Tan. Floorplancad: A large-scale cad drawing dataset for panoptic symbol spotting. *ICCV 2021; arXiv:2105.07147*, 2021.

[6] Ben Feicht. Redlinebench: Ai architectural drawing review benchmark, 2026. Public benchmark; a real architectural drawing set with 58 known issues across seven categories, false alarms penalised, with a frontier-model leaderboard.

[7] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *NeurIPS 2017*, 2017.

[8] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. *ICML 2019 (PMLR v97)*, 2019.

[9] Yoonhwa Jung, Junryu Fu, and Mani Golparvar-Fard. Drawingvqa: A real-world benchmark for multi-depth visual-textual reasoning on construction drawings. *CVPR 2026 Findings; arXiv:2607.15418*, 2026.

[10] Ahti Kalervo, Juha Ylioinas, Markus Häikiö, Antti Karhu, and Juho Kannala. Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis. *SCIA 2019 (Scandinavian Conference on Image Analysis), Springer LNCS; arXiv:1904.01920*, 2019.

[11] Aleksei Kondratenko, Mussie Birhane, Houssame E. Hsain, and Guido Maciocci. Aecv-bench: Benchmarking multimodal models on architectural and engineering drawings understanding. *arXiv:2601.04819 (submitted Jan 8, 2026); code at github.com/AECFoundry/AECV-Bench*, 2026.

[12] P. Li, H. Chen, Z. Zheng, J. He, and H. Yang. Archplanvqa: Benchmark and comprehensive evaluation of vision-language models for understanding architectural floor plan cad drawings. *ASCE Journal of Computing in Civil Engineering, vol. 40, no. 5 (DOI 10.1061/JCCEE5.CPENG-7571)*, 2026.

[13] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773, 2024.

[14] Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. Raster-to-vector: Revisiting floorplan transformation. *ICCV 2017, pp. 2214-2222 (DOI 10.1109/ICCV.2017.241)*, 2017.

[15] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. Introduces the AnyRes dynamic high-resolution scheme, which tiles an input image so that a fixed-resolution vision encoder can read finer detail.

[16] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *TACL — Transactions of the ACL (aclanthology 2024.tacl-1.9); arXiv 2307.03172*, 2024.

[17] Soumya Madireddy, Lu Gao, Zia Din, Kinam Kim, Ahmed Senouci, Zhe Han, and Yunpeng Zhang. Large language model-driven code compliance checking in building information modeling. *arXiv:2506.20551 (also Electronics 14(11):2146)*, 2025.

[18] Nomic AI. Aec-bench: A multimodal benchmark for agentic systems in architecture, engineering, and construction. *arXiv preprint*, 2026. 196 task instances over intrasheet, intradrawing and intraproject scopes; harness and data released under Apache 2.

[19] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[20] Wawan Solihin and Charles Eastman. Classification of rules for automated bim rule checking development. *Automation in Construction 53:69-82*, 2015.

[21] Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jaeger. Overcoming common flaws in the evaluation of selective classification systems. *arXiv preprint (2407.01032)*, 2024.

[22] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *TACL 2025 (arXiv 2407.18418)*, 2025.

[23] Zhiliang Zeng, Xianzhi Li, Ying Kin Yu, and Chi-Wing Fu. Deep floor plan recognition using a multi-task network with room-boundary-guided attention. *ICCV 2019; arXiv:1908.11025*, 2019.

[24] Jiansong Zhang. How can chatgpt help in automated building code compliance checking? *Proceedings of the 40th ISARC, Chennai, India, pp. 63-70*, 2023.

[25] Jiansong Zhang and Nora M. El-Gohary. Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking. *Journal of Computing in Civil Engineering 30(2)*, 2016.

[26] Yang Zou, Brian H. W. Guo, Eleni Papadonikolaki, Johannes Dimyadi, and Lei Hou. Lessons learned on adopting automated compliance checking in the aec industry: A global study. *Journal of Management in Engineering 39(4)*, 2023.