LED-BENCH: A BENCHMARK FOR LARGE-FORMAT ENGINEERING DRAWING UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world engineering drawings are often large-format and information-dense. Answering practical engineering questions therefore requires both fine-grained understanding of engineering information and complex reasoning over objects, relations, and evidence distributed across different regions or views. Existing benchmarks cover a range of engineering-drawing understanding tasks involving local regions, individual views, specific elements, or more constrained visual settings. However, systematic evaluation of understanding over realistic, large-format engineering drawings remains limited. To address this gap, we introduce LED-BENCH, a multidisciplinary benchmark designed to evaluate fine-grained understanding and complex reasoning over large-format engineering drawings, spanning the mechanical, architectural, and transportation domains. Evaluation of 11 representative multimodal large language models, including 8 proprietary models and 3 open-weight models, reveals substantial remaining difficulty: GPT-5.6 Sol achieves 52.65% average accuracy, performance varies across engineering domains, and all evaluated models perform worse on questions requiring broader evidence integration. These results highlight fine-grained understanding, evidence integration, and complex reasoning over large-format engineering drawings as persistent challenges for current multimodal models. Code and data will be made publicly available.

1 INTRODUCTION

Engineering drawings communicate technical specifications through standardized graphical conventions used across mechanical engineering, architecture, and construction (International Organization for Standardization, 2020). Beyond geometric shape, they carry dimensions, tolerances, materials, manufacturing requirements, and other information needed to define and realize a design (Quintana et al., 2010). On a complete drawing sheet, this information may be distributed across views, detail regions, tables, notes, and symbols. Answering an engineering question can therefore require reading small visual elements, associating them with the correct objects or views, and integrating evidence from distant parts of the sheet. This setting combines the challenges of visually rich document understanding (Mathew et al., 2021; 2022) with the loss of fine detail that can arise when multimodal models process high-resolution images (Wang et al., 2025).

Recent benchmarks establish important but complementary evaluation settings. MechVQA studies recognition, reasoning, and judging on mechanical drawings, while TriView2CAD evaluates orthographic-projection reasoning for CAD (Kou et al., 2026; Niu et al., 2025). DesignQA evaluates whether multimodal models can interpret and apply requirements from engineering documentation (Doris et al., 2025). In architecture and construction, AECV-Bench evaluates object counting and drawing-grounded document question answering, ArchPlanVQA focuses on architectural floor-plan CAD drawings, and DrawingVQA organizes questions over real construction drawings by professional reasoning depth (Kondratenko et al., 2026; Li et al., 2026; Jung et al., 2026). AEC-Bench further studies tool-using agents over construction document sets (Mankodiya et al., 2026). Building on these directions, a common complete-sheet evaluation spanning mechanical, architectural, and transportation drawings remains useful for studying how models associate and integrate evidence across heterogeneous engineering representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

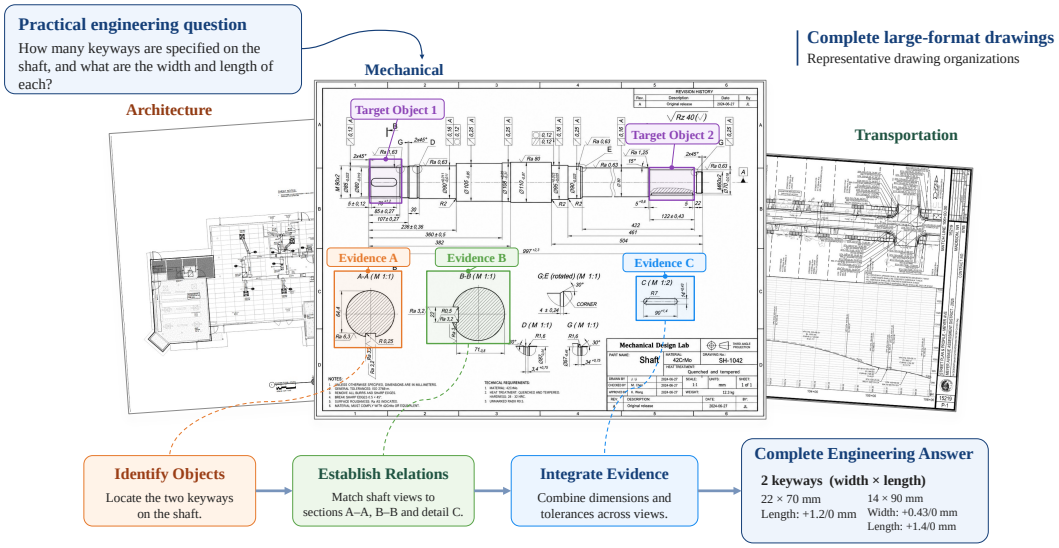


Figure 1: **Overview of LED-BENCH.** The schematic illustrates target localization, object association, and evidence integration over complete engineering drawings. Questions range from localized evidence understanding to cross-region or cross-view reasoning. The transportation domain covers civil-infrastructure drawings such as roads and bridges; the artwork is schematic rather than a montage of benchmark samples.

To address this gap, we introduce *LED-BENCH*, a multidisciplinary benchmark designed to evaluate fine-grained understanding and complex reasoning over large-format engineering drawings, spanning mechanical, architectural, and transportation domains. The benchmark is grounded in practical drawing-reading and engineering-analysis needs, covering both detailed local understanding and more complex cases that require evidence association and integration across different regions or views. These three domains exhibit substantially different organizations of engineering information, from dense multi-view mechanical drawings to spatially distributed architectural layouts and alignment- or network-oriented transportation drawings, providing heterogeneous settings for evaluating the robustness of engineering-drawing understanding and reasoning. Figure 1 illustrates the evaluation setting.

Table 1 positions LED-BENCH relative to representative engineering-drawing benchmarks along the dimensions most relevant to this setting.

LED-BENCH follows an expert-led, model-assisted construction process in which engineering-background annotators define practical question requirements and verify the retained questions and reference answers against the source drawings.

We evaluate 11 representative multimodal large language models, including eight proprietary models and three open-weight models, under a unified three-run evaluation protocol. The strongest model achieves 52.65% average accuracy across the three engineering domains, while performance varies substantially across domains and all evaluated models perform worse on the *Hard* subset. A supplementary diagnostic study further shows that improving access to local visual evidence can improve performance, highlighting persistent challenges in fine-grained understanding, evidence integration, and complex reasoning over large-format engineering drawings.

Our main contributions are threefold:

1. **A multidisciplinary benchmark for large-format engineering drawings.** We introduce LED-BENCH to evaluate fine-grained understanding and complex reasoning over large-format engineering drawings, spanning mechanical, architectural, and transportation domains.
2. **An expert-led, model-assisted benchmark construction process.** We construct practical drawing-reading and engineering-analysis questions through an expert-led process in which

Table 1: Comparison with representative engineering-oriented multimodal benchmarks. A dash indicates that the corresponding work does not report a directly comparable count.

Benchmark	Domain	Images	Questions	Input Type	Task Focus
MechVQA (Kou et al., 2026)	Mechanical	3,281	20,778	Mechanical drawings	Recognition & reasoning
AECV-Bench (Kondratenko et al., 2026)	AEC	21	192	AEC floor plans	OCR & spatial reasoning
ArchPlanVQA (Li et al., 2026)	Architecture	11,477	26,266	Floor-plan CAD	Semantic reasoning
DrawingVQA (Jung et al., 2026)	Construction	33	92	Construction drawings	Expert drawing reasoning
DesignQA (Doris et al., 2025)	Eng. Design	–	1,451	CAD + documents	Requirement compliance
TriView2CAD (Niu et al., 2025)	Civil / Bridge	203,000	–	Three-view drawings	Orthographic reasoning
LED-Bench	**Mech. / Arch. / Trans.**	**107**	**377**	**Complete drawing sheets**	**Fine-grained & complex reasoning**

model assistance supports candidate drafting and refinement, followed by drawing-grounded human verification.

3. **A systematic evaluation of current multimodal models.** We evaluate 11 representative MLLMs under a unified three-run protocol and analyze performance across engineering domains and difficulty levels, together with diagnostic interventions that reveal persistent limitations in visual evidence access and integration.

2 Related Work

Multimodal reasoning and visually rich document understanding. General-purpose multimodal benchmarks have progressively moved beyond object recognition toward knowledge-intensive and compositional reasoning. MMMU (Yue et al., 2024) evaluates expert-level multimodal understanding across a broad range of academic disciplines, while MM-Vet (Yu et al., 2024) examines the integration of multiple visual and linguistic capabilities. A closely related line of work studies visually rich documents, where answering questions depends not only on recognizing text but also on interpreting layout, graphical structure, and relations among multiple visual elements. DocVQA (Mathew et al., 2021) established document-image question answering as a major evaluation setting; InfographicVQA (Mathew et al., 2022) and ChartQA (Masry et al., 2022) further require models to combine textual, graphical, and numerical evidence. OCRBench (Liu et al., 2024) systematically evaluates text-centric visual capabilities, while MMLongBench-Doc (Ma et al., 2024) extends document reasoning to long-context settings in which relevant evidence may be distributed across multiple pages and heterogeneous document modalities.

Together, these studies show that multimodal document understanding requires substantially more than OCR. Engineering drawings share this need for localization and evidence integration, but introduce an additional representational layer: dimensions, projection conventions, technical symbols, geometric relations, stationing systems, and domain-specific drafting conventions jointly determine the meaning of the visual evidence.

Engineering drawing and design benchmarks. Recent benchmarks evaluate multimodal models directly on engineering representations, but differ substantially in domain, visual unit, and reasoning scope. MechVQA (Kou et al., 2026) studies large-scale mechanical-drawing understanding through fine-grained tasks spanning recognition, reasoning, and judging. DesignQA (Doris et al., 2025) evaluates the interpretation and application of engineering requirements using textual technical documents, CAD-related imagery, and engineering drawings. CReFT-CAD / TriView2CAD (Niu

et al., 2025) focuses on orthographic-projection reasoning and correspondence between multi-view representations and underlying CAD or geometric structure.

In architecture and construction, AECV-Bench (Kondratenko et al., 2026) evaluates drawing-grounded capabilities including text interpretation, object counting, spatial reasoning, and comparison. ArchPlanVQA (Li et al., 2026) studies visual question answering over architectural floor plans, while DrawingVQA (Jung et al., 2026) emphasizes expert-curated reasoning questions grounded in real construction drawings. AEC-Bench (Mankodiya et al., 2026) further extends evaluation toward agentic reasoning over real-world AEC document sets, spanning intra-sheet, intra-drawing, and intra-project tasks.

Taken together, prior benchmarks occupy complementary points along three axes: the engineering domain being represented, the visual unit supplied to the model, and the depth of reasoning required. *LED-BENCH* focuses on their intersection in a complete-sheet setting, covering heterogeneous engineering domains and large-format drawings, with questions ranging from localized evidence understanding to cases in which information must be associated across different regions or views and integrated into a complete engineering answer.

3 LED-BENCH

3.1 Drawing Collection

LED-BENCH is constructed as a reasoning-intensive benchmark. We adopt a purposive, need-driven curation strategy focused on large-format engineering drawings that support fine-grained understanding and complex reasoning. Candidate drawings are collected from publicly accessible educational, technical, and project-document sources, including engineering textbooks and handbooks, public drawing repositories, and released design documents.Annotators inspect each candidate drawing and retain those that support meaningful drawing-reading or engineering-analysis questions involving detailed engineering information, relations among engineering objects, and, where applicable, evidence distributed across different regions or views. A drawing is retained only when it is sufficiently legible for reliable human interpretation, answer-relevant evidence can be grounded unambiguously, and it supports multiple distinct and determinate questions. Incomplete, corrupted, severely blurred, visually duplicated, or insufficiently legible drawings are excluded according to these predefined curation criteria.

3.2 Expert-Led Question Construction

Figure 2 summarizes the expert-led, model-assisted question construction process. For each selected drawing, engineering-background annotators first identify a practical drawing-reading or engineering-analysis need and formulate the corresponding question specification. The specification defines the target engineering object or task, the evidence required from the drawing, the relations or reasoning involved, and the requirements for a complete answer.

Language models are used to assist candidate generation and wording refinement. For each candidate, engineering-background annotators review the question and reference answer against the corresponding source drawing, checking question grounding, target-object association, required evidence, answer correctness and completeness, engineering relevance, ambiguity, and unintended answer cues.

Items with unclear grounding, incorrect or incomplete reference answers, unintended answer cues, insufficient engineering relevance, or substantial semantic redundancy are revised or removed. The retained questions are subsequently divided into *Medium* and *Hard* according to the predefined evidence-structure and reasoning criteria described above, with the distinction reviewed by engineering-background experts with reference to the source drawings and the evidence required to answer each question. Appendix E provides additional details on the authoring instructions and construction process.

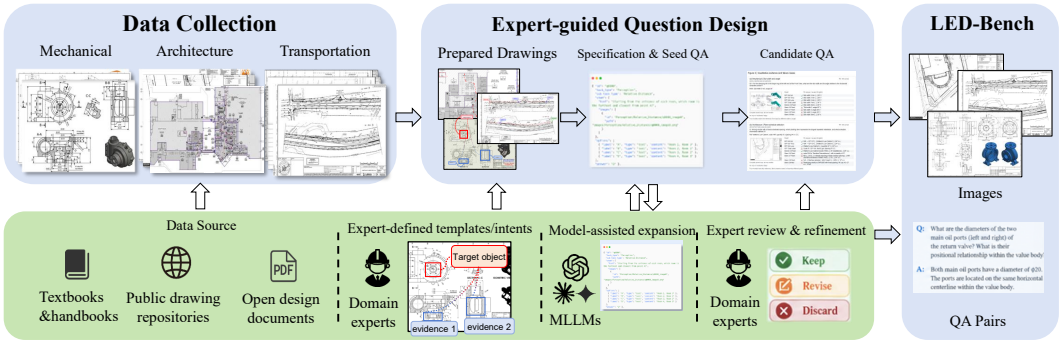


Figure 2: **Expert-led, model-assisted benchmark construction.** Engineering-background annotators define the engineering intent, target objects, required drawing evidence, reasoning requirements, and answer constraints. Language models assist candidate question drafting, answer proposals, and wording refinement under human-defined requirements. Retained questions and reference answers are reviewed against the original complete drawings, and the *Medium*/*Hard* distinction is made according to the predefined difficulty criteria with review by engineering-background experts.

3.3 Dataset Overview

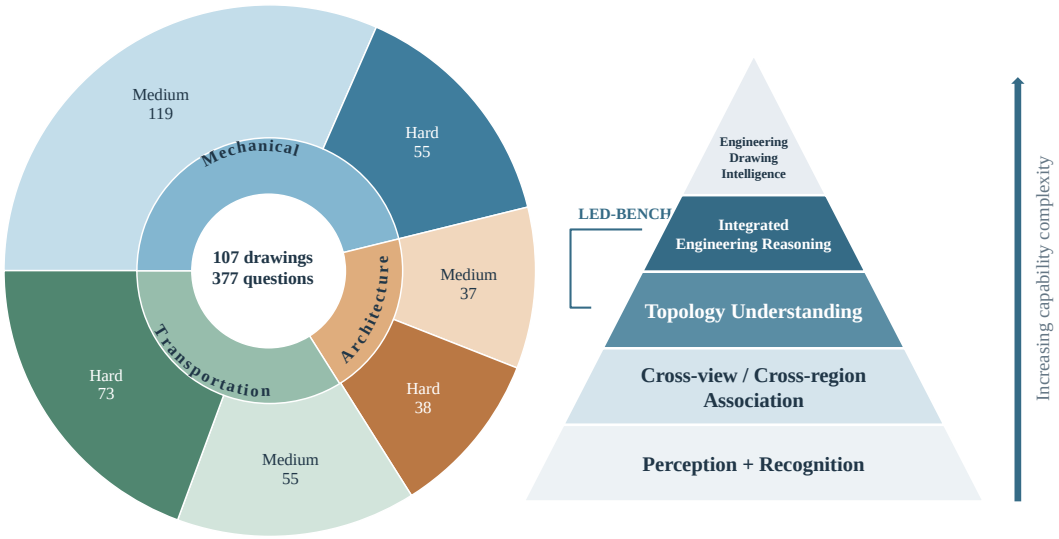
Table 2 summarizes the frozen benchmark. LED-BENCH contains 107 complete engineering drawing sheets and 377 questions: 45 mechanical drawings with 174 questions, 21 architectural drawings with 75 questions, and 41 transportation drawings with 128 questions. Multiple questions may be associated with the same source sheet, but each question is evaluated independently with a fresh model context.

Table 2: **Dataset composition across engineering domains.** Q/Draw. denotes the average number of questions associated with each complete drawing. Medium and Hard are relative difficulty levels within LED-Bench.

Domain	Drawings	Questions	Q/Draw.	Medium	Hard
Mechanical	45	174	3.87	119	55
Architecture	21	75	3.57	37	38
Transportation	41	128	3.12	55	73
Total	**107**	**377**	**3.52**	**211**	**166**

Figure 3 summarizes the domain composition of the benchmark and the capabilities targeted by its questions.

Difficulty levels. LED-BENCH uses two relative difficulty levels, *Medium* and *Hard*. *Medium* questions typically rely on a bounded set of evidence and require operations such as combining several annotations, resolving a local structural or assembly relation, establishing a simple cross-region or cross-view correspondence, or performing a one-step engineering calculation. *Hard* questions require broader or more interdependent evidence, such as associating information across different regions or multiple views, reconstructing spatial or topological relations, tracing ordered dependencies, performing multi-step inference or calculation, or jointly applying multiple engineering constraints. Both levels remain drawing-grounded reasoning tasks; the distinction reflects relative evidence-integration and reasoning demand within LED-BENCH rather than a separation between recognition and reasoning. The *Medium*/*Hard* distinction is made according to these predefined criteria and reviewed by engineering-background experts with reference to the corresponding source drawings and the evidence required to answer each question.



(a) Dataset composition. *(b)* Capability hierarchy.

Figure 3: **Dataset composition and benchmark focus. (a)** Question distribution and *Medium*/*Hard* split across the three domains. **(b)** Conceptual hierarchy of capabilities involved in large-format engineering drawing understanding and reasoning.

Drawing scale. LED-BENCH exhibits substantial variation in image scale across domains, with median long-side lengths of approximately 1,450 pixels for mechanical drawings, 6,800 pixels for architectural drawings, and 3,400 pixels for transportation drawings. This variation reflects the heterogeneous spatial organization of large-format engineering drawings. Appendix H.1 further examines model sensitivity to input rasterization resolution.

4 Evaluation Results

4.1 Evaluation Setup

Testing models. We evaluate 11 representative multimodal large language models (MLLMs), including eight proprietary models and three open-weight models. The proprietary models are GPT-5.6 Sol, GPT-5.6 Luna, GPT-5.5, Gemini-3.5 Flash, Qwen3.8-Max, Qwen3.8-Flash, Qwen3.7-Flash, and Qwen3.6-Flash; the open-weight models are Kimi K2.5, Qwen3.6-27B, and Qwen3.8-27B.

Automatic scoring. LED-BENCH uses strict binary semantic scoring for open-ended answers. Each response is independently evaluated by three LLM judges using the question, reference answer, and required answer content, while the original drawing and evaluated-model identity are hidden from the judges. A response receives a score of 1 only when all indispensable engineering content is correct; semantically equivalent wording, numerical representations, and units are accepted, whereas substantive omissions, incorrect object or relation associations, numerical errors, or contradictions receive 0. The final score is determined by majority vote among the three judges. A blinded human review of 90 sampled responses yields 95.6% agreement with the LLM-majority decisions (Cohen's $\kappa = 0.902$). Further scoring and validation details are provided in Appendix B. Each model is evaluated in three independent runs over the full benchmark. Scores are pooled across the three runs within each domain, without majority voting across model responses. Because the three domains contain different numbers of questions, we use the unweighted domain average $A_{\text{avg}} = \frac{1}{3}(A_{\text{mech}} + A_{\text{arch}} + A_{\text{trans}})$ as the primary metric.

Difficulty-level results are computed in the same way within the *Medium* and *Hard* subsets. Run-level results and stability statistics are reported in Appendix C.

4.2 Overall Performance

Table 3 reports performance across the three engineering domains. The highest domain-avg accuracy is 52.65%, achieved by GPT-5.6 Sol, with 58.81% on mechanical, 48.89% on architectural, and 50.26% on transportation drawings.

Table 3: **Performance across engineering domains.** The best result in each column is shown in bold, and the second-best is underlined.

Model	Mechanical ↑	Architecture ↑	Transportation ↑	Domain Avg
Closed-source MLLMs				
GPT-5.6 Sol	**58.81**	**48.89**	**50.26**	**52.65**
GPT-5.6 Luna	50.38	47.11	44.27	<u>47.26</u>
Qwen3.8-Max	46.36	<u>48.00</u>	<u>46.09</u>	46.82
Gemini-3.5 Flash	<u>57.47</u>	35.11	31.77	41.45
GPT-5.5	41.00	17.78	39.84	32.87
Qwen3.7-Flash	32.57	16.44	27.86	25.63
Qwen3.8-Flash	29.89	16.44	27.34	24.56
Qwen3.6-Flash	18.39	6.22	15.36	13.33
Open-weight MLLMs				
Kimi-K2.5	32.18	15.11	27.08	24.79
Qwen3.8-27B	22.41	20.44	24.22	22.36
Qwen3.6-27B	31.99	6.22	22.92	20.38

Performance also varies substantially across domains. For example, Gemini-3.5 Flash reaches 57.47% on mechanical drawings but 35.11% and 31.77% on architectural and transportation drawings, respectively. In contrast, Qwen3.8-Max remains within a narrower range of 46.09–48.00% across the three domains. These results indicate that aggregate performance alone can obscure substantially different domain-specific behavior.

4.3 Performance across Difficulty Levels

Table 4 compares performance on *Medium* and *Hard* questions using domain-avg accuracy. All evaluated models perform worse on the *Hard* subset. GPT-5.6 Sol decreases from 56.57% to 48.47%, Qwen3.8-Max from 49.97% to 42.09%, and Gemini-3.5 Flash from 47.39% to 34.49%, the largest absolute drop among the evaluated models at 12.90 percentage points.

The consistent performance gap is compatible with the intended distinction between the two difficulty levels, but model performance is not used to define the split. *Hard* questions generally require broader or more interdependent evidence and more involved reasoning than *Medium* questions. Both subsets remain drawing-grounded reasoning tasks, so the distinction should be interpreted as relative evidence-integration and reasoning demand within LED-BENCH rather than a separation between recognition and reasoning.

Table 4: **Domain-macro accuracy (%) across the two difficulty levels.** Δ denotes the absolute decrease from Medium to Hard.

Model	Medium ↑	Hard ↑	Δ
Closed-source MLLMs			
GPT-5.6 Sol	56.57	48.47	8.10
GPT-5.6 Luna	48.59	44.60	3.99
Qwen3.8-Max	49.97	42.09	7.88
Gemini-3.5 Flash	47.39	34.49	12.90
GPT-5.5	34.34	29.36	4.98
Qwen3.7-Flash	27.01	23.91	3.10
Qwen3.8-Flash	27.21	20.63	6.58
Qwen3.6-Flash	16.94	8.57	8.37
Open-weight MLLMs			
Kimi-K2.5	25.98	22.68	3.30
Qwen3.8-27B	25.80	18.04	7.76
Qwen3.6-27B	23.59	16.86	6.73

4.4 Failure Stages and Diagnostic Interventions

We analyze model failures at four conceptual stages: *Local Evidence Understanding*, *Object Association*, *Relation Reconstruction*, and *Integrated Engineering Reasoning*. In the coded error sample summarized in Figure 4, *Relation Reconstruction* is the largest category (47.3%), followed by *Integrated Engineering Reasoning* (23.4%), *Object Association* (17.2%), and *Local Evidence Understanding* (12.1%).

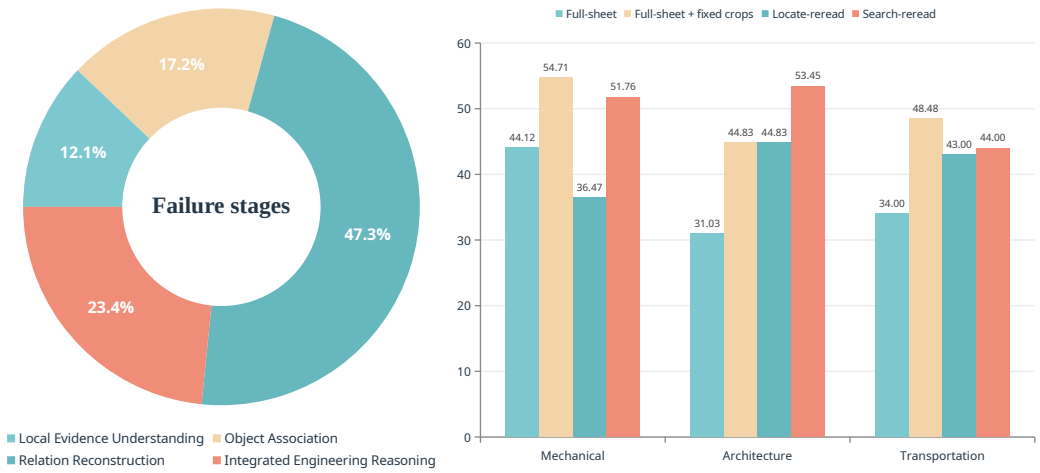
To examine whether improved visual access can alleviate these failures, we conduct a supplementary diagnostic study on 377 questions from 107 drawings, comprising 174 mechanical, 75 architectural, and 128 transportation questions. Four response models—GPT-5.4, Qwen3.8-Flash, Gemini-3.5 Flash, and GPT-5.6 Sol—are evaluated for three runs under a whole-sheet baseline and three visual-access interventions. *Fixed overlapping crops* supplements the full drawing with four predefined overlapping crops. *Locate then reread* first identifies likely target regions and then rereads the corresponding crops. *Search then reread* searches for multiple relevant regions before rereading and integrating the resulting observations.

Table 5: Domain-macro accuracy (%) for the whole-sheet baseline and three visual-access interventions. Δ denotes the absolute change relative to whole-sheet direct answering.

Method	Intended support	Domain Avg ↑	Δ
Whole-sheet direct answering	Baseline	36.38	–
Fixed overlapping crops	Local evidence access	49.34	12.96
Locate then reread	Target localization + rereading	41.43	5.05
Search then reread	Multi-region search + rereading	**49.74**	**13.35**

The three interventions exhibit different cross-domain behavior. Fixed overlapping crops improve accuracy in all three domains and raise the domain-avg score from 36.38% to 49.34%. Search then reread reaches a similar overall level (49.74%) and gives the highest architectural accuracy (53.45%). Locate then reread is more mixed: it improves Architecture and Transportation but decreases Mechanical from 44.12% to 36.47%. These results suggest that improving access to relevant regions can be beneficial, but additional localization or search steps do not yield uniform gains across drawing organizations.

The intervention study is diagnostic rather than a direct causal attribution of individual errors to a single failure stage. Detailed intervention procedures and domain-level results are provided in Appendix F.



(a) Failure-stage distribution **(b)** Accuracy across domains

Figure 4: **Failure stages and domain-level intervention performance.** **(a)** Distribution of the four failure stages in the coded error sample used for the diagnostic analysis. **(b)** Four-model mean accuracy across Mechanical, Architecture, and Transportation for whole-sheet direct answering and three visual-access interventions.

5 Conclusion

LED-BENCH provides a multidisciplinary benchmark for understanding and reasoning over complete engineering drawings across mechanical, architectural, and transportation domains. Under a unified three-run evaluation protocol, current MLLMs exhibit substantial cross-domain variation and consistently lower performance on questions requiring broader and more interdependent evidence. A supplementary diagnostic study further shows that visual-access interventions can improve performance, although their effects vary across methods and engineering domains.

These findings highlight persistent challenges in locating, associating, and integrating engineering evidence across complete drawings. The reported cross-domain differences are descriptive rather than causal, and LED-BENCH is a purposive challenge set rather than an estimate of performance over the natural distribution of engineering practice. We hope the benchmark supports more reliable evaluation and development of multimodal models for complex engineering drawings.

AI Use Statement

Generative AI tools were used as assistants for candidate question expansion, manuscript preparation. Benchmark intent, question requirements, reference-answer verification, dataset inclusion, and final scientific claims remain subject to human review. Model-based semantic judges are part of the evaluation protocol and are reported separately from the blinded human validation of the scoring procedure.

Ethics Statement

LED-Bench evaluates engineering-drawing understanding and reasoning and is not intended to replace professional engineering review or support autonomous safety-critical decision making. Source drawings originate from educational, technical, and publicly accessible project-document materials with varying redistribution conditions. The release will preserve source and rights metadata and redistribute only assets permitted by their original source terms.

Reproducibility Statement

We freeze the benchmark images, questions, reference answers, *Medium*/*Hard* assignments, and evaluation configuration. Each question is evaluated independently with a fresh context using the

corresponding complete drawing, and the same frozen image bytes are submitted to every evaluated model without client-side resizing or recompression. Evaluation uses no demonstrations, external tools, or web retrieval, and each model is evaluated in three independent runs. Exact answering and judging prompts, scoring and aggregation rules, and run-level stability statistics are provided in the appendix. Identical submitted image bytes standardize the external benchmark input but do not imply identical provider-side visual preprocessing.

REFERENCES

Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *Journal of Computing and Information Science in Engineering*, 25(2):021009, 2025. doi: 10.1115/1.4067333.

International Organization for Standardization. Technical product documentation (TPD)—general principles of representation—part 1: Introduction and fundamental requirements. Technical Report ISO 128-1:2020, International Organization for Standardization, 2020. URL https://www.iso.org/standard/65296.html.

Yoonhwa Jung, Junryu Fu, and Mani Golparvar-Fard. Drawingvqa: A real-world benchmark for multi-depth visual-textual reasoning on construction drawings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 2121–2130, June 2026. URL https://arxiv.org/abs/2607.15418.

Aleksei Kondratenko, Mussie Birhane, Houssame E. Hsain, and Guido Maciocci. Aecv-bench: Benchmarking multimodal models on architectural and engineering drawings understanding. *arXiv preprint arXiv:2601.04819*, 2026. doi: 10.48550/arXiv.2601.04819.

Qian Kou, Xiaofeng Shi, Yulin Li, Xiaosong Qiu, Xinyang Wang, Hua Zhou, and Dongxing Cao. Mechvqa: Benchmarking and enhancing multimodal llms on comprehensive mechanical drawing understanding. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.30794.

Pengyang Li, Hainan Chen, Zhuang Zheng, Juntao He, and Hai Yang. Archplanvqa: Benchmark and comprehensive evaluation of vision-language models for understanding architectural floor plan cad drawings. *Journal of Computing in Civil Engineering*, 40(5):04026060, 2026. doi: 10.1061/JCCEE5.CPENG-7571.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. doi: 10.1007/s11432-024-4235-6.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95963–96010, 2024. doi: 10.52202/079017-3041. Datasets and Benchmarks Track.

Harsh Mankodiya, Chase Gallik, Theodoros Galanos, and Andriy Mulyar. Aec-bench: A multimodal benchmark for agentic systems in architecture, engineering, and construction. *arXiv preprint arXiv:2603.29199*, 2026. doi: 10.48550/arXiv.2603.29199.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

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

Ke Niu, Zhuofan Chen, Haiyang Yu, Yuwen Chen, Teng Fu, Mengyang Zhao, Bin Li, and Xiangyang Xue. Creft-cad: Boosting orthographic projection reasoning for cad via reinforcement fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 92007–92029, 2025. doi: 10.52202/085713-2765.

Virgilio Quintana, Louis Rivest, Robert Pellerin, Frédérick Venne, and Fawzi Kheddouci. Will model-based definition replace engineering drawings throughout the product lifecycle? a global perspective from aerospace industry. *Computers in Industry*, 61(5):497–508, 2010. doi: 10.1016/j.compind.2010.01.005.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025. doi: 10.1609/aaai.v39i8.32852.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

A Evaluation Protocol Details

PDF sources are rendered to frozen 200-DPI PNGs for the main baseline; native raster images are used directly. The image hash defines the actual submitted input. Each question receives a fresh request with the complete image and no retrieval tools, demonstrations, or reference answer. The frozen client does not resize or recompress the input. Provider-side image processing is not observable.

Its wrapper sends `max_completion_tokens=32768` and normally requests `response_format=json_object`. Temperature, top-p, seed, reasoning effort, and image detail are omitted: their effective values are provider defaults, not confirmed zeros or fixed seeds. A JSON-format fallback may be used after an interface error and must be identified from the attempt record. The client requests a short final answer rather than a chain of thought. The token cap may include provider reasoning tokens and is not a guarantee of 32,768 visible output tokens.

R1–R3 were collected with the frozen request template and repeated calls, not majority selection across answers. The original wrapper used pass-level deadlines of 240/600/600 seconds with deferred retries; later recovery schedules are versioned in the run artifacts. A universal 120-second timeout with two immediate retries does not describe these executions. Successful answers are retained; technical retries are not triggered by low judged correctness. Retry details, parsing failures, returned model IDs, and finish reasons must be read from per-attempt records.

The eleven displayed main-table API aliases are `gpt-5.6-sol`, `gpt-5.6-luna`, `gpt-chat-latest`, `gemini-3.5-flash`, `qwen3.8-max`, `qwen3.8-flash`, `qwen3.7-flash`, `qwen3.6-flash`, `kimi-k2.5`, `qwen3.8-27b`, and `qwen3.6-27b`. GPT-5.5 is the manuscript label for the gateway alias `gpt-chat-latest`; alias identity alone is not an immutable provider-version guarantee. GPT-5.4 is an additional response model only in the supplementary method study.

A.1 Exact answering prompt

```
You are an expert engineering-drawing reader. Use only evidence visible
↪  in the attached complete drawing.

Discipline: {domain_label}.
Answer only from evidence visible in the attached complete drawing.
Do not use external standards or assumptions to fill missing information.
Do not describe your reasoning. Preserve identifiers, specifications,
↪  numbers, units, tolerances, directions, object-value bindings, and
↪  necessary order.
Answer in the same language as the question. Include every necessary
↪  answer field.

Question: {question}

Return exactly one JSON object with no Markdown:
{"answer":"...","answerability":"answerable|partial|insufficient|ambiguo⌋
↪  us","confidence":0.0}
```

The parser accepts a JSON object, including one within a code fence, and can recover a decodable object containing an `answer` field. Empty, truncated, or unusable responses are retained as technical or response failures. Confidence and answerability fields are stored but do not weight correctness.

A.2 Exact binary judge prompt

Each judge receives only the question, full reference answer, mandatory components, and final candidate answer. Candidate-model identity and the source image are absent. In the frozen scoring configuration, the component list contains the full reference as one item; no independently human-decomposed checklist is implied. Scores must be integer 0 or 1 with a nonempty reason.

```
You are a strict evaluator for an engineering-drawing question-answering
↪  benchmark. Judge only whether the candidate answer is fully correct
↪  with respect to the provided reference answer and mandatory
↪  components. Do not infer the candidate model identity. Return JSON
↪  only.

Judge whether the candidate answer is fully correct.

Question:
{question}

Reference answer:
{reference_answer}

Mandatory components:
{mandatory_components}

Candidate answer:
{candidate_answer}

Rules:
- Score 1 only if the candidate is semantically correct and ALL mandatory
↪  components are satisfied.
- Score 0 if any mandatory component is missing or wrong.
- Score 0 for a wrong engineering object/instance, swapped values, wrong
↪  relation/direction/identifier, or a material contradiction.
- Equivalent wording, standard engineering synonyms, and exactly
↪  equivalent unit conversions are acceptable.
- Do not require extra reasoning or details that are not mandatory.
- Use binary scoring only. Do not output partial credit.

Return JSON only:
{"score": 0 or 1, "reason": "brief explanation"}
```

The judge prompt and aggregation rule are frozen independently of the evaluated-model identity. Judge-panel and run-manifest identifiers are treated as part of the evaluation metadata, and historical judge configurations are not treated as interchangeable.

Three valid votes use majority decision. Two valid agreeing votes also yield a score; conflicting pairs or a single valid vote remain NA. Missing terminal candidate answers are zeros, whereas unresolved judge decisions are excluded from the scoring denominator. The reference-only judge evaluates answer equivalence and cannot independently verify an incorrect reference against the drawing. A validation of this scorer is consequently not a validation of the dataset gold answers.

A.3 Scoring and aggregation formulas

For question i in run r, each of the three judges returns a binary decision $s_{i,r}^{(j)} \in \{0,1\}$ when valid. The final response score is

$$s_{i,r} = \begin{cases} \text{majority}\{s_{i,r}^{(j)}\}_{j=1}^{3}, & \text{three valid votes,} \\ v, & \text{two valid votes, both equal to } v, \\ \text{NA}, & \text{otherwise.} \end{cases} \tag{1}$$

Missing terminal candidate answers are scored as 0 under the frozen policy, whereas unresolved judge outcomes remain NA and are excluded from the denominator.

For domain d containing N_d questions, pooled domain accuracy over the three independent runs is

$$A_d = \frac{\sum_{i \in d} \sum_{r=1}^{3} z_{i,r} s_{i,r}}{\sum_{i \in d} \sum_{r=1}^{3} z_{i,r}}, \qquad z_{i,r} = \mathbb{I}[s_{i,r} \neq \text{NA}]. \tag{2}$$

The primary benchmark metric is the unweighted domain avg-average

$$A_{\text{avg}} = \frac{1}{3} \left(A_{\text{mech}} + A_{\text{arch}} + A_{\text{trans}} \right). \tag{3}$$

Difficulty-level results are computed analogously by restricting the sums in Eq. equation 2 to the corresponding *Medium* or *Hard* subset within each domain and then averaging the resulting domain accuracies with equal weight.

B Human Validation of the LLM Scoring Protocol

We assess the reliability of the automatic majority-vote scorer using a blinded human review of 90 model responses, with 30 responses sampled from each of the Mechanical, Architecture, and Transportation subsets. The review set was constructed from stored R2/R3 responses using a fixed random seed. Within each domain, sampled questions were distinct, sampling did not use the LLM-judge label, and candidate model/run assignments were approximately balanced. During review, the candidate-model identity, run identifier, individual judge votes, and majority decision were hidden. The human reviewer saw the question, reference answer, and candidate response and assigned a binary correctness label under the same complete-answer criterion as the automatic scorer. Cohen's κ is computed over binary human labels; no sampled response was marked unjudgeable.

Table 6: Blinded human validation of the LLM-based majority-vote scoring protocol. Agreement is the fraction of responses for which the human binary label matches the three-judge majority decision.

Subset	N	**Agreement (%)**	**Cohen's κ**
Overall	90	95.6	0.902
Mechanical	30	93.3	0.815
Architecture	30	96.7	0.932
Transportation	30	96.7	0.930
Unanimous judge votes	71	97.2	0.934
2:1 split judge votes	19	89.5	0.791

Across the 90 reviewed responses, 86 receive the same binary decision from the human reviewer and the LLM-majority scorer. Agreement is higher when all three judges agree than when the vote is split 2:1.

C Run-to-Run Stability

To characterize repeated-run variability under the same primary aggregation used in the main evaluation, we compute domain-avg accuracy separately for each of the three runs. Within each run, accuracy is first computed independently on Mechanical, Architecture, and Transportation, and the three domain accuracies are then averaged with equal weight. Table 7 reports the resulting run-wise domain-avg scores and their arithmetic mean and sample standard deviation across the three runs ($n = 3$). The standard deviation is used only as a descriptive stability statistic and should not be interpreted as a confidence interval. No majority vote is applied across runs.

Table 8 gives the corresponding domain-level stability statistics underlying the avg scores. For each model and domain, the reported mean is computed from the three run-wise domain accuracies and matches the pooled domain accuracy in Table 3; the Domain avg column is therefore identical to the final column of Table **??**.

Across the eleven reported models, domain-avg run-to-run sample standard deviations range from 0.31 to 3.03 percentage points. Domain-level variability is larger for some model–domain combinations, especially in Architecture and Transportation, which further motivates interpreting small score differences descriptively rather than as statistically significant model separations.

Table 7: Run-to-run stability under the primary domain avg metric. R1–R3 report domain avg(%) for each independent run; the final column reports the arithmetic mean and sample standard deviation across the three runs.

Model	R1	R2	R3	Mean ± SD
		Closed-source MLLMs		
GPT-5.6 Sol	54.15	53.79	50.02	52.65 ± 2.29
GPT-5.6 Luna	50.02	47.72	44.02	47.26 ± 3.03
Qwen3.8-Max	46.88	47.09	46.48	46.82 ± 0.31
Gemini-3.5 Flash	39.45	44.02	40.88	41.45 ± 2.34
GPT-5.5	32.98	33.28	32.35	32.87 ± 0.47
Qwen3.7-Flash	25.64	26.77	24.46	25.63 ± 1.15
Qwen3.8-Flash	26.52	25.11	22.04	24.56 ± 2.29
Qwen3.6-Flash	13.68	12.99	13.32	13.33 ± 0.35
		Open-weight MLLMs		
Kimi-K2.5	23.70	24.99	25.68	24.79 ± 1.00
Qwen3.8-27B	22.85	22.66	21.57	22.36 ± 0.69
Qwen3.6-27B	22.80	19.88	18.45	20.38 ± 2.22

Table 8: Domain-level run-to-run stability. Values are mean ± sample standard deviation of the three run-wise accuracies (%).

Model	Mechanical	Architecture	Transportation	Domain avg
		Closed-source MLLMs		
GPT-5.6 Sol	58.81 ± 0.88	48.89 ± 4.07	50.26 ± 4.30	52.65 ± 2.29
GPT-5.6 Luna	50.38 ± 1.76	47.11 ± 4.29	44.27 ± 6.64	47.26 ± 3.03
Qwen3.8-Max	46.36 ± 2.02	48.00 ± 2.67	46.09 ± 1.56	46.82 ± 0.31
Gemini-3.5 Flash	57.47 ± 1.00	35.11 ± 4.29	31.77 ± 3.52	41.45 ± 2.34
GPT-5.5	41.00 ± 1.85	17.78 ± 4.07	39.84 ± 1.35	32.87 ± 0.47
Qwen3.7-Flash	32.57 ± 1.20	16.44 ± 3.36	27.86 ± 1.97	25.63 ± 1.15
Qwen3.8-Flash	29.89 ± 0.57	16.44 ± 5.39	27.34 ± 2.07	24.56 ± 2.29
Qwen3.6-Flash	18.39 ± 1.52	6.22 ± 2.04	15.36 ± 1.19	13.33 ± 0.35
		Open-weight MLLMs		
Kimi-K2.5	32.18 ± 1.15	15.11 ± 3.36	27.08 ± 1.19	24.79 ± 1.00
Qwen3.8-27B	22.41 ± 2.63	20.44 ± 2.78	24.22 ± 1.56	22.36 ± 0.69
Qwen3.6-27B	31.99 ± 3.27	6.22 ± 2.78	22.92 ± 2.39	20.38 ± 2.22

D Operational Interpretation of the Failure Stages

These stages are used to organize qualitative error coding and intervention design in Section 4.4. The main text reports the aggregate distribution of the coded failure stages and uses them as a diagnostic analysis framework.

E Question Authoring and Quality-Control Details

Archived authoring materials. Candidate construction was iterative. The archived development prompts ask a model to propose questions, candidate answers, evidence dependencies, and wording refinements from a complete engineering drawing. These prompts are authoring aids rather than autonomous inclusion rules. Final inclusion decisions are made through drawing-grounded human review.

Operational construction criteria. For each candidate, annotators identify a practical information need, the target engineering object or task, the drawing evidence required to answer it, the relations

Table 9: Operational interpretation of the four conceptual failure stages used in the diagnostic analysis.

Stage	Operational interpretation
Local Evidence Understanding	Relevant local text, numbers, identifiers, specifications, symbols, or geometric cues are read or interpreted incorrectly.
Object Association	Evidence is plausibly read but associated with the wrong engineering object, instance, space, component, annotation target, or view-specific counterpart.
Relation Reconstruction	Relevant objects are identified, but connectivity, adjacency, path, direction, order, continuity, scope, datum, or cross-region/view correspondence is reconstructed incorrectly.
Integrated Engineering Reasoning	Upstream evidence is substantially available, but the final answer fails during evidence combination, multi-step derivation, arithmetic or unit handling, aggregation, constraint application, or answer completion.

or inference involved, and the requirements for a complete answer. Candidates are revised or removed when grounding is unclear, the reference answer is incorrect or incomplete, wording creates unintended answer cues, engineering relevance is weak, or substantial semantic redundancy remains.

Difficulty distinction. The *Medium*/*Hard* distinction is based on evidence structure and reasoning demand rather than model accuracy. Medium questions typically use a bounded evidence set and relatively direct combinations or calculations. Hard questions require broader or more interdependent evidence, such as cross-region or cross-view association, spatial or topological reconstruction, ordered-dependency tracing, multi-step derivation, or joint application of multiple engineering constraints. Engineering-background experts review this distinction with reference to the source drawing and the evidence required to answer each question.

F Supplementary Method Study: Execution and Results

F.1 Inputs and intervention procedures

The supplementary study compares whole-sheet direct answering with three visual-access interventions on the same 377 questions and answer requirements used in the aligned diagnostic analysis. Reference answers are withheld during inference. Each model–question–method combination is evaluated in three independently requested runs, without voting across runs or selecting the best response.

Table 10: Visual-access intervention inputs and normal deployment calls per response. Failures and fallback paths may change the realized count.

Method	Input / procedure	Calls
Whole-sheet direct answering	Complete drawing and question; answer directly from the full visual input	1
Fixed overlapping crops	Complete drawing plus four predefined overlapping crops in one request	1
Locate then reread	Locate at most two likely target regions, reread the corresponding crops, and answer using the full drawing and reread observations	3
Search then reread	Search for multiple relevant regions, reread up to two selected regions, and integrate the resulting observations	3

Fixed overlapping crops cover 0.55×0.55 of the normalized image, with adjacent overlap 0.10. Their boxes are $(0, 0, 0.55, 0.55)$, $(0.45, 0, 1, 0.55)$, $(0, 0.45, 0.55, 1)$, and $(0.45, 0.45, 1, 1)$. Crops preserve native pixels and are stored losslessly without rescaling. The locate-then-reread and search-then-reread settings use crops produced by their respective localization stages.

Table 11: Four-model mean accuracy (%) for the whole-sheet baseline and three visual-access interventions. Domain Macro is the unweighted mean over Mechanical, Architecture, and Transportation.

Method	Mechanical	Architecture	Transportation	Domain Macro
Whole-sheet direct answering	44.12	31.03	34.00	36.38
Fixed overlapping crops	**54.71**	44.83	**48.48**	49.34
Locate then reread	36.47	44.83	43.00	41.43
Search then reread	51.76	**53.45**	44.00	**49.74**

F.2 Domain-level results

Fixed overlapping crops provide the largest point estimate in Mechanical and Transportation, while search then reread performs best in Architecture and has the highest domain-macro point estimate. Locate then reread improves Architecture and Transportation over whole-sheet direct answering but is lower in Mechanical. The results therefore support a benefit from improved visual access in several settings while also showing that additional localization steps are not uniformly beneficial across domains. The study is diagnostic and does not establish a unique causal mechanism for each improvement or failure.

G Limitations and Additional Discussion

Cross-domain comparisons are descriptive. Mechanical, architectural, and transportation drawings differ in drafting conventions, visual scale, language, source composition, and the relations emphasized by their questions. Domain-level scores therefore characterize model behavior under heterogeneous engineering representations but do not isolate the causal effect of any single domain factor.

Finite numbers of unique drawings. LED-Bench contains 107 complete sheets, with multiple questions originating from some drawings. Within-sheet questions can share visual and structural difficulty, and the architectural subset contains only 21 unique drawings. Small domain-level or inter-model differences should therefore be interpreted cautiously. The supplementary intervention study reports descriptive point estimates for method comparisons and does not provide significance tests for the main model ranking.

Strict scoring depends on reference and judge quality. Complete-answer scoring is sensitive to reference completeness and semantic scoring boundaries. Three separately called judges reduce dependence on any single judge, but their errors need not be statistically independent.

Potential source contamination. Some source drawings come from publicly accessible educational, technical, and project-document materials and may have been encountered during model pretraining. The benchmark questions and answer requirements are newly constructed, reducing direct QA memorization but not eliminating possible prior exposure to the underlying source documents.

Purposive challenge-set scope. LED-Bench intentionally selects reasoning-rich complete drawings and questions rather than sampling the natural frequency distribution of engineering practice. Scores should therefore be interpreted as capability measurements on a curated challenge setting, not as estimates of expected performance on arbitrary engineering documents.

H Drawing Scale and Visual Input Considerations

The three domains differ substantially in raster dimensions because their source sheets follow different drawing conventions and spatial organizations. For native mechanical rasters and the frozen 200-DPI architectural/transportation renders, the median long-side lengths are approximately 1,450 pixels for mechanical drawings, 6,800 pixels for architectural drawings, and 3,400 pixels for transportation drawings. Raster size is not treated as a proxy for reasoning difficulty: compact mechanical sheets

may contain dense multi-view geometry and technical annotations, while larger architectural and transportation sheets often distribute evidence over wider spatial extents.

These dimensions describe the submitted benchmark images before provider-side processing. Supplying identical frozen image files standardizes the external input, but different model APIs may apply different resizing, tiling, tokenization, or visual encoding procedures. Identical submitted bytes therefore do not imply identical effective visual resolution across models.

H.1 Exploratory Rasterization Results

Figure 5 summarizes an exploratory rasterization study on architectural and transportation PDF-source drawings. We compare multiple client-side rendering resolutions and report the strict accuracy among adjudicated answers for four response models. The results are model dependent and not strictly monotonic: higher input resolution is generally helpful, especially for GPT-5.6 Sol and Gemini 3.5 Flash, but some intermediate settings outperform nearby resolutions for individual models.

These observations support the practical importance of rasterization choices while not establishing a universally optimal DPI. We therefore retain 200 DPI as the fixed client-side rendering baseline for the main benchmark, rather than as an experimentally established optimum. Native mechanical rasters are outside this PDF-rasterization comparison, and provider-side resizing or visual encoding remains uncontrolled, so identical submitted image files still do not imply identical effective visual resolution across model APIs.

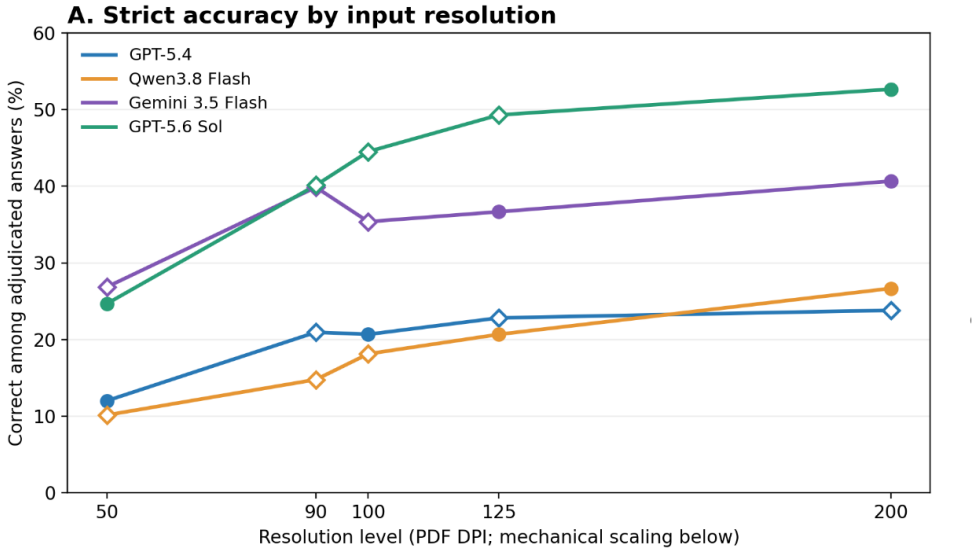


Figure 5: **Exploratory rasterization study.** Strict accuracy among adjudicated answers as a function of client-side PDF rasterization resolution for four representative response models. The study is exploratory and should not be interpreted as identifying a universally optimal DPI setting.