VISION-LANGUAGE MODELS ON SPECIFICATION-GOVERNED INSTRUMENT APPROACH CHARTS: WHICH FAILURES, AND WHICH SUPPLIED INFORMATION CHANGES THEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Converting engineering drawings into structured data requires recovering every specification-defined instance and its attributes, and recent vision-language models are a candidate solution. On instrument approach charts, they omit instances or fill in wrong attributes. Scoring whole records as right or wrong cannot tell an unformed instance from a nonconforming field set or a misread value, nor which information to supply. From charting specifications and chart evidence, we define the instances and attributes each chart requires; this shared semantic reference scores outputs and yields the required instances (instance list), the fields to fill (field list), and reading rules. We evaluate Luna on 862 charts and Luna, Qwen, and K3 on a common 171-chart subset. Most strict failures are required instances that never form a record; with the instance list, matching is almost fully restored, nearly all newly correct records are previously unmatched targets recovered whole, and matched instances show almost no net change. Unformed instances account for 85.28% and 92.16% of Luna's and Qwen's strict failures; with the list supplied and identities still self-written, matching rises from 45.7%, 12.2%, and 56.7% to 98.8%, 73.9%, and 98.9%. The field list mainly aligns output fields with the scored reference fields; its effect on unsupplied values depends on response design. The reference makes each failure type separately measurable, turning an accuracy gain into an account of which information changed what.

1 INTRODUCTION

Automated checking and reuse of engineering drawings require more than recognizing text and numbers: the drawing must be converted into a complete set of meaningful records. Instrument approach charts provide a concrete case. Their graphical and textual elements jointly describe flight procedures, so connecting chart content to navigation data requires recovering the objects, relationships, and constraints expressed in the drawing.

For a vision-language model (VLM), a correctly read altitude is therefore only part of the task. It must be associated with the appropriate object and applicability condition. Likewise, a segment needs directed endpoints and an ordered position as well as its distance or course. Extraction can fail either to form a required record or to complete its attributes correctly.

Aggregate accuracy combines these outcomes: even a large gain does not reveal whether an intervention helped recover instances, select applicable fields, or determine values. Public charting specifications and chart evidence supply a common reference within a declared scope: a determination-basis ledger links target attributes to evidence, rules, and dependencies, so supplying endpoint names and listing requested fields address different requirements of the same targets.

We evaluate Luna on 862 FAA instrument approach charts and Qwen and K3 on a 171-chart subset, with three kinds of reference-generated information: an instance list (required instances and identities, which the model must still write), a field list (each entry's reference-scored fields), and two reading rules.

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.

Under this task contract, most strict failures are expected instances that never form a matchable record (Luna 85.28%, Qwen 92.16%); when models receive the instance list but must still write the identities themselves, matching is almost fully restored (45.7→98.8%, 12.2→73.9%, and 56.7→98.9% across three deployments), and the net strict-correct gains of +842, +641, and +672 records come almost entirely from previously unmatched targets that are now matched and entirely correct (+873, +661, +670), whereas the net change within instances matched under both conditions is only −20, −2, and +4. We make three contributions. First, a specification-grounded shared semantic reference generates both the scoring targets and three kinds of supplied information: instance lists, field lists, and reading rules. Second, a target-level accounting of strict-correct gains shows that, in all three deployments, the gains concentrate in previously unmatched targets. Third, a field-level interpretation of whole-record gains shows that the field list mainly changes the agreement between the output field set and the reference-scored field set, and that its effect on unsupplied readings depends on the response design.

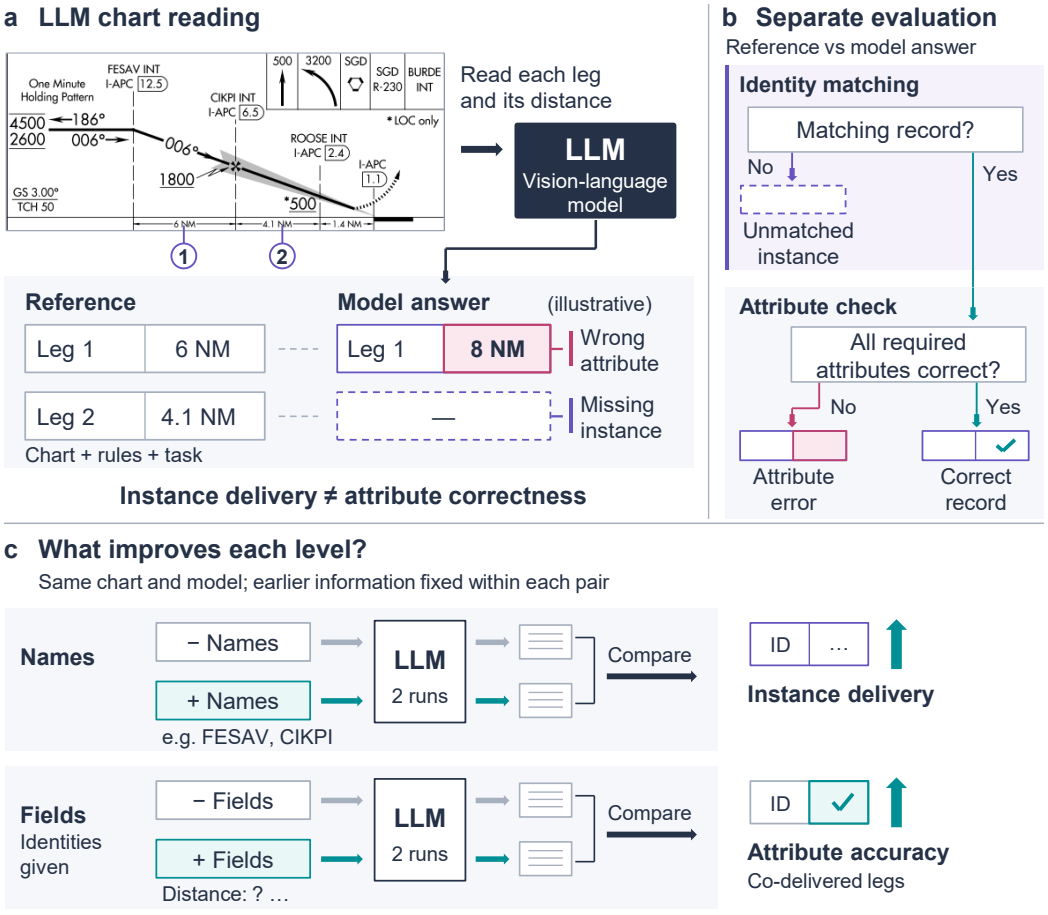


Figure 1: Two levels of chart reading and information supply. (a) A real chart excerpt with two illustrative output errors: a wrong distance and an unmatched instance. (b) Each expected instance is checked for identity matching and then attribute correctness. (c) Two paired comparisons within cumulative information supply: adding names and, later, unfilled fields. Arrows summarize observed responses in the layered study, not measurements on the illustrated chart.

2 Related Work

Structured extraction benchmarks make different choices about completeness. Enginuity matches parts by item number, with missing parts affecting description scores; DesignQA asks compliance questions using specification text and drawings (Kumar et al., 2026; Doris et al., 2024). Other evaluations distinguish absent fields from null values, count omitted keys, or separate key and value

scores (Ferguson et al., 2026; Zhang et al., 2026; Tenckhoff et al., 2026). We build on this distinction; our separation of the output field set from values belongs to the same family as separate key and value scoring. The difference is that we first define a complete instance set within a declared task scope, link its attributes to their determination evidence, and then vary these parts with information generated from the same reference.

Multimodal diagnostics supply reference intermediate information, including keypoints, recognized text, chart tables, or graph nodes (Luo et al., 2021; Masry et al., 2022; Guo et al., 2026); TopoAgent, for example, supplies gold node lists on the same diagrams and measures only the remaining edges. Information can also vary systematically: MathVerse changes where problem information is presented; SciVerse supplies theorem names before equations and application instructions; GroundBench separates information bundles and includes image-removal controls (Zhang et al., 2024; Guo et al., 2025; Sattigeri, 2026). These studies report overall changes after information is supplied. Because our supplied information and scoring targets come from the same instance reference, a change can be traced, target by target, to previously unmatched instances, to the attributes of matched instances, or to the output field set.

Task structure has been studied through operation counts and program subtrees associated with success and failure (Hudson & Manning, 2019; Eyzaguirre et al., 2025). Rule-following evaluations vary the availability of rules and facts, including supplied reference rules, graded rule contexts, and explicit rule structures (Zhou et al., 2025; Sun et al., 2024; Lu et al., 2026). Here, recorded evidence and dependencies define what information remains under each condition. Whether those accounts also yield numerical difficulty is a separate empirical question; Section 6.3 reports the result.

3 Required Instances and the Determination-Basis Ledger

3.1 Instances, Identities, and Attributes

Our shared semantic reference is defined jointly by public specifications, chart evidence, and the declared extraction scope: it specifies which instances must be recovered and what determines each attribute, and the same reference generates both the scoring targets and the three kinds of supplied information. We study profile and minimums information in U.S. Federal Aviation Administration (FAA) instrument landing system or localizer (ILS/LOC) approach charts.

We call each complete record to be recovered an instance. Outputs comprise five families: approach variants, fix roles, ordered segments, vertical constraints, and published minima. An identity specifies which object an instance denotes within its procedure and applicability conditions (its scope); a segment's ordinal position and directed endpoints form its identity, while its distance, course, and vertical mode are attributes. Strict evaluation scores each complete instance of the latter four families: 14,446 targets on the 862 charts and 2,741 on the nested 171 (Appendix A.1).

3.2 The Ledger and the Scope of Its Checks

The determination-basis ledger links each target attribute to evidence bindings, rule identifiers, a conclusion, and upstream dependencies. Its 17,347 records, compiled from structured reference annotations, provide one derivation per field in an acyclic dependency structure. The ledger separately records which instances exist and their identities, which fields each instance carries, and what determines each value; these three kinds of records generate the three kinds of information in Section 4. Figure 2 illustrates one fix-role record.

Two annotators produced and cross-reviewed the chart annotations, and 862 of the 929 corpus charts form the nonprobability evaluation sample. Construction checks rejected 99.6% of injected mutations, and the 13 of 20 charts comparable with external coded procedures showed no structural contradiction; these checks cover construction consistency and a limited external subset, not a record-by-record review of the reference (Appendix B). Reproduction materials comprise the ledger, rule catalogue, scorer, feature tables, and provenance documentation.

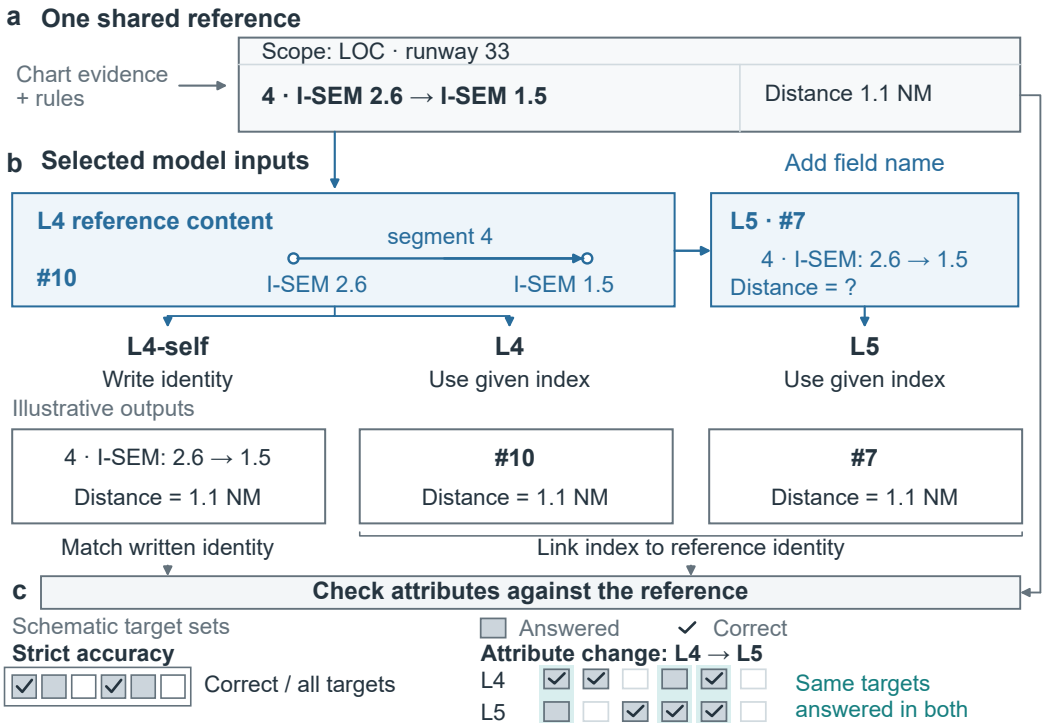


Figure 2: A reference fix-role record. The marked name and Maltese cross bind to the same fix; rule W-12 supplies the final approach fix (FAF) role. The upstream localizer (LOC) procedure scope belongs to the record. Arrows show recorded evidence and dependency links, illustrated here for one chart.

4 Three Kinds of Information and Gain Accounting

4.1 Supplied Information and Contrasts

The same reference directly generates three kinds of information. The instance list lists the chart's procedure scope, fix names, segment order, and all entry identities, retaining existing field placeholders but no values; it is verbatim the L4 reference content of the layered conditions, not merely a list of identities. The field list names the reference-scored fields of each entry, without values. The reading rules are two value conventions: take the underlined altitude attached to a fix's vertical line, and take the visibility outside the parentheses in the minima. Supplying information does not guarantee its correct use; this is what the contrasts test.

The instance-list contrast compares two conditions: free generation with the two reading rules added, and the same task with the instance list inserted before the final response instruction while models must still write complete identities in the free-output grammar (L4-self); the scorer does not recover identities from row numbers. Numbered answers on the same content (L4) test the form of expression. The field contrast is L4→L5: both supply all identities, and L5 additionally lists the requested fields of each entry. The reading-rule contrast is free generation versus free generation with the reading rules.

Layered conditions retain earlier information: scorer-associated identities begin with segments at L3 and extend to all entries at L4; at L5, unscored auxiliary entries with no requested fields change from numbered answers to entries requiring no answer. Appendix C.2 lists all twelve conditions, their identity sources, and the part each contrast targets.

4.2 Gain Accounting, Field Diagnostics, and the Secondary Test

The net gain after supplying the instance list decomposes exactly into targets newly matched with all attributes correct, minus previously correct targets no longer matched, plus the net change within

instances matched under both conditions; this is a counting identity over saved outputs, not a mediation estimate. The field diagnostics fix 787 segments and 931 distance and course fields and also compare complete records after symmetrically deleting reference-unscored fields. A separate fixed-candidate-field design on 36 charts has both conditions fill every candidate field, in line or JSON format, and changes only whether the model is told which fields carry reference annotations.

A stronger secondary hypothesis, registered before layered calls, asks whether structural accounts of enumeration, evidence scale and dispersion, conditions, and dependency depth form an instance-level difficulty scale; the gain accounting and field diagnostics are retrospective (Appendices C.1 and D.5). Section 6 reports these results in turn.

5 Experimental Design and Evaluation

5.1 Samples and deployments

Luna is evaluated on 862 charts with 14,446 targets; a nested 171-chart, 106-airport subset with 2,741 targets supports the Qwen comparisons, the layered Luna study, and the instance-list contrast for all three deployments.

The 171-chart subset is a nested tier fixed before any model run: the corpus was split by whole airport groups with a fixed seed, tiers were taken in airport-group order, and later exclusions removed only charts whose reference did not close or that served in a canary run, never for complexity or performance (Appendix B). The field diagnostics use all 787 segments and 931 numerical fields on the 171 charts, with a sensitivity check on the 89 charts whose number of answered rows does not change between conditions. The fixed-candidate-field design uses 36 of these charts (544 targets per condition, 210 scored numerical fields), and the prompt-language check a separate set of 36.

Luna (gpt-5.6-luna) runs server-side with medium reasoning and no exposed temperature, output limit, or seed; Qwen3.5-9B runs locally without thinking at temperature 0.7; K3 uses low-effort thinking. In the instance-list contrast, Luna's and Qwen's free-plus-reading-rules and numbered conditions reuse earlier outputs; K3's three conditions, with 171 outputs each, were collected contemporaneously. Appendix C.2 specifies prompts and decoding.

5.2 Output events and denominators

Identity match rate is the fraction of targets whose model-written identity combination matches; with supplied identities, entry response rate counts targets answered by row number. Strict accuracy requires a matched or answered instance with all attributes correct, over all targets; paired attribute comparisons use instances matched or answered under both conditions (Appendix A.2).

Strict failures can accordingly be described in three parts: no identity-matched record; a matched record whose output field set disagrees with the reference-scored field set; and an agreeing field set with a wrong value. Because the reference annotates only scored fields, a printed but unannotated field, such as the 172° course printed beside the 4.3 NM distance on chart 0196's AWONE→ROCCK segment, whose reference scores only the distance, counts as extra; field-set agreement therefore means agreement with the reference-scored field set, not physical applicability (Appendix A.3).

5.3 Calls and statistics

Each chart and condition receives one generation. L4-self inserts the verbatim L4 content and requires the supplied identities to be written in full; call counts, the complete and control frames, C1 (explicit identity conventions), and S-free (schema-constrained output) are specified in Appendix C.2. Intervals are pointwise 95% intervals from 10,000 airport-cluster bootstrap replicates and exclude repeated-generation variation and reference error (Appendix C.3). Appendices D.1 and A.3 document the mismatch anatomy and the corrected parser.

6 Results: Where Correct Records Come From, and What Whole-Record Gains Consist Of

6.1 Strict-correct gains concentrate on previously unmatched targets

Under free generation, most strict failures are required instances with no matchable record: 85.28% [83.52, 86.78] of Luna's 9,345 and 92.16% [90.67, 93.51] of Qwen's 2,602 lack an identity-matched output record (Figure 3a). This is an output event, not a cause, and most of these targets are not merely written differently: relaxed scope, instance, and naming forms find a key-compatible record for 26.95% [23.22, 30.73] of Luna's unmatched targets, against a random-pairing baseline of 3.40% (Figure 3b; Appendix D.1).

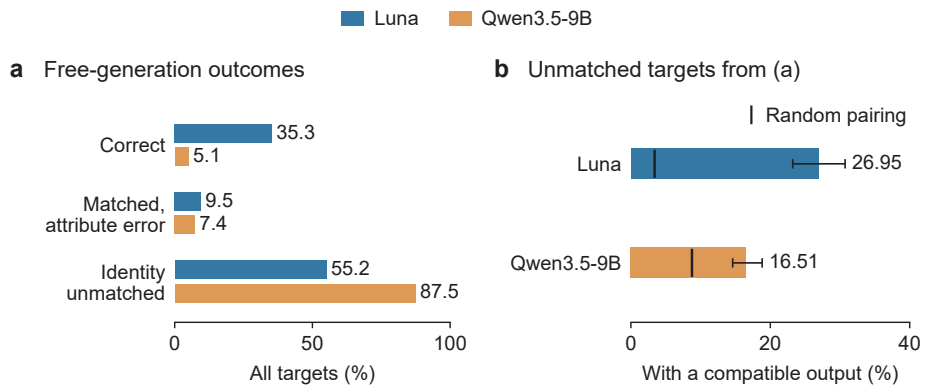


Figure 3: Free-generation outcomes and key compatibility. (a) Luna: 14,446 targets on 862 charts; Qwen: 2,741 targets on 171 charts. (b) Among each model's unmatched targets, the proportion with a compatible output under relaxed scope, instance and naming rules; ticks denote random-pairing means. Official scores are unchanged. Error bars are 95% airport-cluster bootstrap intervals (Appendix D.1).

Supplying the instance list while models still write complete identities largely restores matching and sharply raises strictly correct records in all three deployments (Table 1). The list gives identities but no values, and the scorer never recovers identities from row numbers. Identity match rates rise from 45.71% to 98.80% for Luna, from 12.22% to 73.91% for Qwen, and from 56.69% to 98.94% for K3.

Table 1. Strictly correct records and net-gain composition on 171 charts (2,741 targets, 106 airports). The last three columns sum to the count change; Δ in points with 95% paired airport-cluster bootstrap intervals.

Deployment	Free+rules, n (%)	L4-self, n (%)	Δ pp [95% CI]	New match, correct	Correct→ unmatched	Net in common
Luna	1,116 (40.72)	1,958 (71.43)	+30.72 [26.00, 35.32]	+873	−11	−20
Qwen	158 (5.76)	799 (29.15)	+23.39 [20.91, 26.12]	+661	−18	−2
K3	1,348 (49.18)	2,020 (73.70)	+24.52 [21.64, 27.65]	+670	−2	+4

Nearly all of the net gain comes from previously unmatched targets now matched with the whole record correct. Target by target, it equals these targets, minus formerly correct targets no longer matched, plus the net change within common matches: 873 − 11 − 20 = 842 for Luna, 661 − 18 − 2 = 641 for Qwen, and 670 − 2 + 4 = 672 for K3 (Table 1). This counting identity is not a mediation estimate; value accuracy within common matches changes with intervals all including zero (Appendix D.2). Numbered answers add no common benefit (−2.26, +1.06, and −1.13 points), so the gain does not rely on scorer-recovered identities.

Restored matching does not make whole records correct. Under the list condition, strict segment accuracy is 24.27% for Luna, 12.33% for Qwen, and 26.94% for K3, although all 787 of K3's segments are matched; the remaining gap lies in field sets and values.

6.2 WHOLE-RECORD GAINS REQUIRE A FIELD-LEVEL INTERPRETATION

The field-list contrast (L4→L5) is read on two fixed denominators across all 171 charts. Whole segments are scored on the 787 segments with reference values, answered or not; numerical readings are scored on 931 distance and course fields, with missing and invalid answers counted as failures (Appendix D.3).

The field list raises whole-record accuracy far more than it raises readings (Table 2). Strictly correct whole segments rise by 67.34 points for Luna and 31.89 for Qwen; after reference-unscored fields are deleted from both conditions' answers, the differences shrink to 6.10 and 17.53, and correct numerical readings change by 0.32 and 2.58. Luna's post-deletion gain still excludes zero, and Qwen's numerical interval allows a gain of about 8 points, so neither is read as null.

Table 2. Field diagnostics on fixed denominators, all 171 charts (L4→L5). Deletion removes only reference-unscored keys from both answers and is not an official score; 95% paired airport-cluster bootstrap intervals.

Model	Measure	L4→L5	Difference [95% CI]
Luna	Strictly correct whole segment	191→721/787	+67.34 [62.63, 71.59]
Luna	After deleting reference-unscored fields from both	685→733/787	+6.10 [2.08, 10.43]
Luna	Correct distance and course fields	875→878/931	+0.32 [−2.59, 3.44]
Qwen	Strictly correct whole segment	66→317/787	+31.89 [26.42, 38.43]
Qwen	After deleting reference-unscored fields from both	207→345/787	+17.53 [12.13, 24.26]
Qwen	Correct distance and course fields	425→449/931	+2.58 [−2.62, 8.31]

The whole-record gain mainly reflects agreement between output field sets and reference-scored field sets. Deletion changes only which existing keys are compared and adds, changes, or rematches nothing, so the sharp narrowing it produces locates the gain in field-set agreement, although it does not measure the causal share of field selection in generation. Response obligations also differ on 82 charts, whose entries without reference-scored values must be answered in L4 but are exempt in L5; on the other 89 charts, strict gains remain 65.75 and 33.11 points, but this post hoc subset cannot fully remove the design difference (Appendix D.3).

The effect of field information on readings depends on the response design. In a fixed-candidate-field design that varies only whether models are told which fields carry reference values, Luna's correct readings on the 210 numerical fields shared with the layered design rise from 166 to 191 (line-by-line) and from 162 to 194 (JSON), averaging 13.57 [5.84, 22.36] points, whereas the same fields go from 202 to 193 in the layered design. Qwen's corresponding effect is −3.33 [−13.28, 6.81], while JSON over line-by-line output raises its matching by 55.06 points. Neither whole-record nor fixed-field accuracy is a pure reading measure.

6.3 SUPPORTING CONTRASTS: READING RULES, RESIDUAL VALUE ERRORS, INTERFACES, CUMULATIVE LAYERS, AND THE DIFFICULTY SCALE

Different supporting information changes different outcomes, and residual value errors differ greatly between models (Appendix D.4). Two reading rules raise Luna's value accuracy on 875 jointly delivered instances by 11.54 [7.34, 15.17] points, while its identity matching changes by +2.33 [−2.39, 6.61]. Structured output barely changes free-generation matching (+0.22 and +0.18 points), whereas stating four identity conventions raises Luna's from 40.50% to 59.03%. In the cumulative layers, fix names mainly change matching and the field list mainly changes attributes (Figure 4); with both supplied, values stay wrong on 8.5% (Luna) and 59.0% (Qwen) of 698 commonly answered segments. The registered difficulty scale is not supported: only 23 of 427 within-layer coefficient records meet its improvement criterion (Appendix D.5).

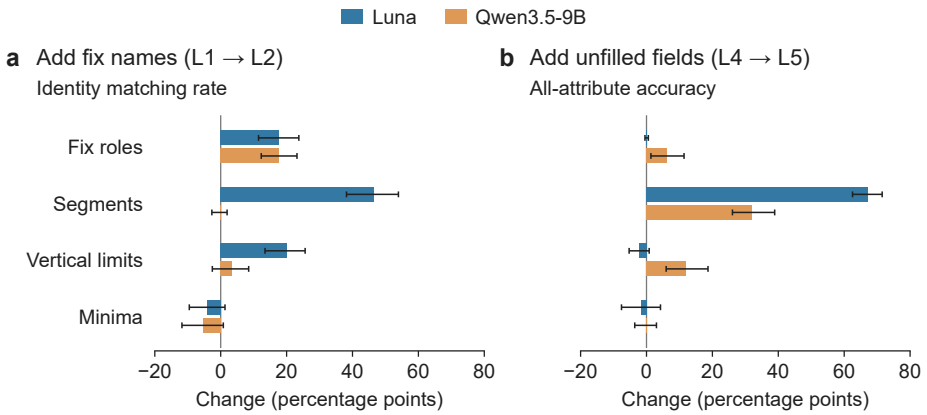


Figure 4: Conditional responses to added information on 171 charts per model. (a) L1→L2 changes identity matching over all targets in each family. (b) L4→L5 changes the joint correctness of applicable fields and their values among targets answered in both arms within each model. Earlier information is retained. Error bars are 95% airport-cluster bootstrap intervals; full contrasts and direct model differences appear in Appendix D.4.

7 Discussion and Limitations

7.1 Correct-record gains can be traced to their sources

Because the shared semantic reference generates both the scoring targets and the supplied information, accuracy gains can be traced to their sources target by target: to previously unmatched targets that become entirely correct rather than to matchable identity strings alone, and to field-set agreement rather than to readings. Since the inputs derive from the reference, such gains do not show autonomous identity inference. Wherever public specifications define which instances a drawing must contain and what determines each field, the construction can be reused; what transfers is the diagnostic procedure of defining targets, fixing denominators, and accounting for gains, not the tested models' behavior or an autonomous extraction system validated on other drawing domains.

7.2 Scope of the conclusions

Both the instance list and the field list are generated from the reference, so they describe supply from an ideal upstream stage; we did not test what recall and precision an imperfect upstream enumerator, whether built on OCR, a detector, or another model, would need to obtain the same gains. The evidence concerns the selected FAA chart scope and a nonprobability sample. Two-person annotation, construction checks, and limited external comparison provide the reference basis described in Appendix B; they do not substitute for a comprehensive independent review, and one recorded derivation per field represents a particular route through the evidence.

Luna and Qwen reuse earlier control outputs (Luna's backend weights cannot be verified); K3's three conditions were collected contemporaneously, which does not resolve cross-model comparability. Each chart and condition has a single generation, and the intervals exclude repeated-generation variation and reference error. The same-day repeat condition (row R in Appendix C.2; readings in Appendix D.4) is the study's only internal check on generation variability, while K3's contemporaneously collected conditions are unaffected by differences in collection time.

Common-matched results concern a response-selected intersection; the field diagnostics fix their denominators but are retrospective and cover two models. The fixed-candidate-field design, which holds response obligations equal, covers only 36 charts; extending it to all 171 charts is the direct next test of how field lists affect readings. The main experiments use Chinese prompts on English charts; in the fixed 36-chart English-prompt check, the field list's improvement on segments persists in both models, whereas the improvement of names on fix roles is unstable (Appendix D.4). These results characterize the tested model deployments and input conditions.

8 CONCLUSION

Under this task contract, most strict failures are expected instances that never form a matchable record (Luna 85.28%, Qwen 92.16%); when models receive the instance list but must still write the identities themselves, matching is almost fully restored (45.7→98.8%, 12.2→73.9%, and 56.7→98.9% across three deployments), and the net strict-correct gains of +842, +641, and +672 records come almost entirely from previously unmatched targets that are now matched and entirely correct (+873, +661, +670), whereas the net change within instances matched under both conditions is only −20, −2, and +4. This accounting is possible because a specification-linked shared semantic reference generates both the scoring targets and the supplied information; the same reference shows that the whole-record gain from field lists comes mainly from agreement between the output field set and the reference-scored field set, with an effect on unsupplied readings that depends on the response design. Extraction evaluations should fix their denominators and account for the sources of gains target by target, rather than report only the rise and fall of aggregate scores.

AI USE STATEMENT

AI tools assisted with editing LaTex manuscript text and polish writing.

ETHICS STATEMENT

The study uses publicly published FAA instrument approach charts and involves no human subjects or personal data. The results characterize the tested deployments under the stated conditions and are not evidence that any model is suitable for operational flight-procedure or navigation-data use.

REPRODUCIBILITY STATEMENT

Targets, identity keys, denominators, and the parser are specified in Sections 3 and 5.2 and Appendix A; reference construction and checks in Appendix B; every condition, prompt design, rendering, and decoding setting in Appendix C.2; and the bootstrap procedure, seeds, and multiplicity handling in Appendix C.3. Each chart and condition has one generation, and one deployment is reached through a service channel that does not expose its settings, so its outputs may not be exactly regenerable.

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, 2024. URL https://arxiv.org/abs/2404.07917v2. arXiv:2404.07917v2.

Cristobal Eyzaguirre, Igor Vasiljevic, Achal Dave, Jiajun Wu, Rares Andrei Ambrus, Thomas Kollar, Juan Carlos Niebles, and Pavel Tokmakov. Understanding complexity in VideoQA via visual program generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15613–15636. PMLR, 2025. URL https://proceedings.mlr.press/v267/eyzaguirre25a.html.

Nick Ferguson, Josh Pennington, Narek Beghian, Aravind Mohan, Douwe Kiela, Sheshansh Agrawal, and Thien Hang Nguyen. ExtractBench: A benchmark and evaluation methodology for complex structured extraction, 2026. URL https://arxiv.org/abs/2602.12247v1. arXiv:2602.12247v1.

Bangwei Guo, Xujiang Zhao, Yanchi Liu, Wei Cheng, Shengyu Chen, Dongyue Li, Masaharu Morimoto, Takayuki Kuroda, Dimitris Metaxas, and Haifeng Chen. TopoAgent: A structure-aware perception-to-reasoning framework for diagram-to-graph topology extraction with large vision-language models, 2026. URL https://arxiv.org/abs/2608.28701v1. arXiv:2608.28701v1.

Ziyu Guo, Ray Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. SciVerse: Unveiling the knowledge comprehension and visual reasoning of LMMs on multi-modal scientific problems, 2025. URL https://arxiv.org/abs/2503.10627v1. arXiv:2503.10627v1, submitted 13 March 2025; arXiv comments field: "Initially released in September 2024. Project page: this https URL".

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506v3. arXiv:1902.09506v3.

Abhishek Kumar, Isha Motiyani, Tilak Kasturi, Ethan Seefried, Prahitha Movva, and Tirthankar Ghosal. Enginuity: A dataset and benchmark for vision-language understanding of engineering diagrams, 2026. URL https://arxiv.org/abs/2606.03410v1. arXiv:2606.03410v1.

Zhifeng Lu, Dianyuan Wang, Yuhu Shang, and Zhenbo Xu. RuleSafe-VL: Evaluating rule-conditioned decision reasoning in vision-language content moderation, 2026. URL https://arxiv.org/abs/2605.07760v1. arXiv:2605.07760v1, submitted 8 May 2026; author name spelled "Dianyun Wang" in the paper body.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. ChartOCR: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1917–1925, 2021. URL https://www.microsoft.com/en-us/research/uploads/prod/2020/12/WACV_2021_ChartOCR.pdf.

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. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Sarthak Sattigeri. GroundBench: A factorized, counterfactual benchmark for locating VLM affordance failures, 2026. URL https://arxiv.org/abs/2609.13308v1. arXiv:2609.13308v1, submitted 10 September 2026.

Wangtao Sun, Chenxiang Zhang, XueYou Zhang, Xuanqing Yu, Ziyang Huang, Pei Chen, Haotian Xu, Shizhu He, Jun Zhao, and Kang Liu. Beyond instruction following: Evaluating inferential rule following of large language models, 2024. URL https://arxiv.org/abs/2407.08440v4. arXiv:2407.08440v4, 17 October 2024; the four rule-supply settings cited here were located in v3, 16 October 2024.

Sönke Tenckhoff, Mario Koddenbrock, and Erik Rodner. LLMStructBench: Benchmarking large language model structured data extraction, 2026. URL https://arxiv.org/abs/2602.14743v1. arXiv:2602.14743v1, submitted 16 February 2026.

Boyang Zhang, Adrian Lyjak, Eli Stewart, Zhaoqi Li, and Simon Suo. ExtractBench: A benchmark for schema-guided enterprise document extraction, 2026. URL https://arxiv.org/abs/2607.29677v2. arXiv:2607.29677v2, 5 August 2026; first submitted 31 July 2026. Distinct work from arXiv:2602.12247 of the same benchmark name.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624v2. arXiv:2403.14624v2, 18 August 2024; first submitted 21 March 2024; published at ECCV 2024, per the arXiv comments field "Accepted by ECCV 2024, 46 Pages, Benchmark Project Page: this https URL".

Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. RuleArena: A benchmark for rule-guided reasoning with LLMs in real-world scenarios. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 550–572. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.27. URL https://aclanthology.org/2025.acl-long.27/. Mechanism located in the arXiv preprint arXiv:2412.08972v2.

A TASK CONTRACT AND SCORING

A.1 SCORED TARGETS

This section completes the definition of targets and denominators in Section 3.1. Targets concern profile and published-minimums information in FAA ILS and LOC approach charts. A joint identity key contains the family, approach type, normalized runway end, applicability-condition set, and the family-specific identity fields in Table A1; evaluated attribute values are not identity fields. Approach variants form a fifth output family but contribute no scored targets. From 17,347 determination records, scoring excludes 1,901 variant records, 952 segments with no target value, and 48 unauthorized minimums, leaving 14,446 targets in the 862-chart domain; the nested 171-chart domain contains 2,741. Each target is a whole instance, not a separate attribute cell.

Table A1. Scored families. Counts are for the nested 862-chart and 171-chart domains.

Family	Identity within scope	Evaluated attributes	Targets: 862 / 171
Fix role	Fix name	Role	2,940 / 585
Ordered segment	Ordinal, start fix, end fix	Distance, course, vertical mode, where declared	4,212 / 787
Vertical constraint	Segment or fix	Constraint type, altitude above mean sea level	4,351 / 777
Published minimum	Aircraft-category set	Minimum type, altitude, height, visibility	2,943 / 592

The base prompt requests the identity fields but does not fully specify four boundaries; condition C1 states the conventions in Table A2 explicitly.

Table A2. Identity conventions stated in condition C1.

Aspect	Explicit convention
Segment chain	Use all fixes on the variant's approach track; end at the final approach fix for ILS and the missed approach point for LOC; number consecutive segments from 1.
Applicability	Use the conditions of conditional minimums row groups; a chart-wide equipment suffix is not itself a variant condition.
Minimums grouping	One instance per printed row, with the set of aircraft categories spanned by that row; neither split nor merge rows.
Fix-name suffix	Keep printed INT; use identifier plus facility type for navaids and station plus distance for DME points; use one spelling per fix.

A.2 OUTPUT EVENTS, DENOMINATORS, AND THE DESCRIPTION OF FAILURES

This section gives operational definitions for the output events, denominators, and failure description in Section 5.2. Free output has three pipe-separated parts (family; scope and identity; values), with semicolon-separated fields; numbered frame output supplies values for listed entries. Within a joint identity key, predictions are consumed one-to-one, preferring an unused record with identical values. When the model writes identities, a target is identity-matched if an output record has the same key. When identities are supplied, the scorer associates the supplied identity with the numbered answer, and the target counts as an answered entry; this does not measure whether the model found and named the instance itself. A bare "?" counts as an answer without a value; "X" and a missing numbered row do not. Strict correctness requires a match or answer with every evaluated attribute correct, over all targets; conditional attribute accuracy uses matched or answered instances; attribute contrasts between two conditions use only targets matched or answered in both.

Strict failures are described in three parts, a device for organizing results that does not assume the parts respond independently to interventions. For the segment with identity (1, FESAV, CIKPI) and reference distance 6 NM: no output record with that identity is an instance without a matched record; a record with this identity and 6 NM that also writes a course field the reference does not score has an output field set that differs from the reference's scored field set; a record whose field set agrees but gives 8 NM has a value error. Layered conditions accept all permitted attributes of the family, with

the same acceptance set for scored targets in L4 and L5, so extra permitted attributes enter strict comparison; L5 also allows unscored entries without requested fields to be omitted, so L4→L5 includes a change in response obligations. The reference annotates values only for scored fields, and a field printed on the chart but not annotated in the reference (for example, the AWONE→ROCCK segment on chart 0196 prints both 172° and 4.3 NM, while the reference annotates only the distance) counts as an extra field. Agreement of field sets therefore means agreement with the reference's scored field set, not a judgment of physical applicability.

A.3 Parser and reference version

This section identifies which readings depend on the parser version noted in Section 5.3. All readings use one parser implementation. It accepts brackets containing multiple semicolon-separated conditions, as illustrated in the prompt, and parses a numbered line outside the supplied frame with the free grammar after removing its number; the remainder must still satisfy the free grammar. Rows with missing identity fields or unknown families form no scored record. Compared with an implementation that accepts only the strict grammar, two readings change materially: Luna's overall identity-matching difference for L0→L1 is 3.87 [0.45, 7.58] points, whereas its interval includes zero under the strict grammar; Qwen's matched counts in L0, L1, and L2 are 359, 422, and 518, and are clearly lower under the strict grammar. The directions of the other main readings are the same under both implementations. Appendix B states the reference version used for all readings. Predictions were registered before the first layered call, and readings under the current parser are not untouched preregistered analyses.

B Reference sources and scope of checks

This section states the coverage of the reference construction and checks in Section 3.2. The 862 evaluation charts were selected nonprobabilistically from 929 charts, mainly from publication cycle 2606; 67 charts were excluded before any model call because they failed reference checks or had unresolved ink loss.

The 171-chart subset was frozen on 2026-07-28, before any model run. The corpus was split by whole airport group (seed 20260728) into 100 development and 829 test charts; nested prefixes T1 (58) $\subset$ T2 (113) $\subset$ T3 (171) follow airport-group order in the test split. A gold-closure pass removed only charts whose reference did not close (9 + 9) and three canary-run charts, without redrawing. Main tables use the largest tier all deployments completed; no chart was selected or dropped on complexity or performance. All outputs date from September 2026.

Two annotators produced and cross-reviewed the profile, minimums, and final-approach-fix-to-missed-approach-point annotations, and programs compiled the target records and their acyclic derivations, with exactly one derivation per field. Mutation validation rejected 99.6% of injected errors. Evidence pointers aligned for 99.6% of records. Of 251 localizer missed-approach-point records, 244 agreed with a separately built reference for the same charts. All 62 rule citations passed text, page, and section checks. Of 20 externally coded procedures, 13 had comparable reference facilities and no structural contradiction, and 7 were not comparable. The reference contains one source-binding error, for BOGGA in the LOC procedure on chart 0037: the chart prints a minimum altitude of 2,300 ft, but the reference bound the 1,600 ft missed-approach instruction; all readings use the corrected reference, with target identities, matching rules, and denominators unchanged. These checks cover construction consistency and a limited external subset; they do not yield a semantic accuracy for the whole reference.

C Three kinds of information, conditions, and statistics

C.1 Three kinds of determination basis and three kinds of information

This section states where the three kinds of information in Section 4.1 come from. The determination-basis ledger records three kinds of basis separately for each scored target. An enumeration account records candidates, instance quotas, whether names are printed, and excess same-key and variant-membership instances, answering which instances should exist and which each one is.

The same account records which fields of each instance carry scored reference values. A determination record keeps one derivation per field, linking evidence bindings, a rule identifier, the conclusion, and upstream dependencies. The three kinds of information are generated from these three parts. The instance list gives the chart's procedure scopes, fix names, segment order, and all entry identities, verbatim as in layered condition L4 and retaining existing field placeholders, without values. The field list gives the reference-scored field names for each entry, without values. The reading rules are two general conventions from the value bases: take the underbarred altitude attached to the fix's vertical line, and take the visibility outside the parentheses in the minimums. A supplied conclusion removes the corresponding requirement in the ledger, but a zero remaining requirement describes the input state and does not guarantee a correct model output.

C.2 Conditions and deployments

This section lists every condition used in Sections 4.1 and 5.1 (Table C1). Layered conditions accumulate earlier information, so adjacent contrasts measure an addition given what precedes it; they do not separate information, identity authorship, and interface factorially. All conditions use Chinese prompts and English charts. Charts are rendered at 200 dpi (1076×1650 pixels) and uploaded without further cropping or resizing. The image-free condition omits the image and states its absence while keeping the frame text; the distractor frame adds fictitious entries and reorders rows.

Table C1. Conditions. M: model-written identity; S: supplied identity associated by the scorer after an answer.

Condition	Added information	Identity source	Deployments and charts
Free	Family definitions and free-output grammar	M	Luna 862, 171; Qwen 171
Free + reading rules	Free plus the two reading rules	M	Luna, Qwen, K3, 171 each
L0	Common interface, no chart-specific information	M	Luna, Qwen 171
L1	L0 plus procedure scopes	M	Luna, Qwen 171
L2	L1 plus unordered fix names	M	Luna, Qwen 171
L3	L2 plus ordered segment chain	S for segments; M otherwise	Luna 171
L4	L3 plus remaining entry identities	S	Luna, Qwen, K3, 171 each
L5	L4 plus requested, unfilled fields	S	Luna, Qwen 171
L6	L5 plus disambiguating values	S	Luna 171
L3′	L4 without segment order; anonymous segment rows	M for segments; S otherwise	Luna 171
L4′	L3′ plus separately listed order	M for segments; S otherwise	Luna 171
R	Same-day repeat (Luna L6, Qwen L5)	S	Luna, Qwen 171
L4-self	The verbatim L4 material, with full identities written by the model	M	Luna, Qwen, K3, 171 each
C1	Free plus the four conventions in Table A2 (prompt grows from 2,957 to 4,824 characters)	M	Luna, Qwen 171
S-free	Schema-constrained structured free output	M	Luna, Qwen 171
Complete frame (with image, image-free, distractor)	Identities, requested fields, some disambiguating values, and reading rules	S	Luna 862; Qwen 171
Fixed-protocol eight cells	Complete instance catalog, all candidate fields required; reference scoring-field catalog absent or present × full identity or row identifier × lines or JSON	M for full identity; S for row identifier	Luna, Qwen 36

Luna is gpt-5.6-luna called through a coding-agent channel with medium reasoning and fast service; the channel does not expose temperature, top-p, output limit, or seed. Qwen is Qwen/Qwen3.5-9B

served locally with vLLM, thinking disabled, temperature 0.7, top-p 0.8, top-k 20, min-p 0, presence penalty 1.5, repetition penalty 1, and up to 4,096 new tokens; all its 171-chart batches use the same weight snapshot and settings, and no seed is sent. K3 requests model k3 with thinking enabled at low effort and a 24,576-token output limit, with no temperature or seed sent. Each chart and condition has one formal output; Luna's layered study uses 2,052 calls, and K3's three conditions total 513 outputs. Model comparisons concern these deployments and prompts.

C.3 Intervals and analysis status

This section states what the intervals in Section 5.3 cover. Main intervals are percentile intervals from 10,000 airport-cluster resamples, weighted by target counts; cross-model differences reuse airport weights across models, and attribute contrasts keep each model's jointly answered subset. Seeds are 20260910 for the complete frame, 20260919 for Luna's layered study, 20260921 for the mismatch anatomy, 20260922 for cross-model differences, 20260923 for the instruction contrasts, 20260924 for Luna's and Qwen's instance-list contrasts, 20260925 for jointly answered checks and the field diagnostics, 20260926 for K3's three conditions, the fixed-protocol eight cells, and absolute language rates, and 20260925 for language contrast differences. All reported 95% intervals are pointwise. Registered directional hypotheses additionally receive Benjamini–Hochberg false-discovery-rate control at 0.10, but the collection of contrasts has no familywise error guarantee. A prespecified five-point bound is used for equivalence and repeat stability: Luna's same-day repeat changes output coverage by +0.11 [−0.54, 0.78] and paired attribute accuracy by −0.63 [−2.33, 0.92], while Qwen's L5 repeat coverage interval, −2.01 [−5.42, 1.68], crosses the bound. Intervals reflect only the airport composition of the selected corpus, not reference error or repeated-generation variation; the mismatch anatomy, field diagnostics, gain accounting, and retrospective feature associations are exploratory.

D Results

D.1 Composition of free-generation failures and key compatibility

This section gives the full readings (Table D1) behind the failure composition and Figure 3 in Section 6.1. Under free generation, identity mismatch dominates strict failures for both models; among matched instances, all attributes are correct for 78.76% under Luna and 40.52% (139/343) under Qwen. To see what unmatched targets contain, we search unused outputs within the same chart and family under nested relaxations: scope; segment ordinal and category overlap; removal of type words from a fixed 24-word list; and separator and trailing-zero normalization. The broadest rule also allows one matching segment endpoint and drops category agreement. One-to-one assignment minimizes compatibility cost without reading scored values, and attributes are compared afterwards; a random-link baseline repeats the procedure twenty times with donor charts from different airports.

Table D1. Free-generation failure composition and key-compatibility diagnostics. Compatibility shares use each model's unmatched targets as denominator; brackets are 95% intervals.

Model	Strict failures	Unmatched (share of strict failures)	Naming-compatible	Random baseline	Broadest rule	Counterpart attributes identical
Luna (862 charts, 14,446 targets)	9,345	7,969 (85.28% [83.52, 86.78])	2,148 (26.95% [23.22, 30.73])	3.40%	3,061 (38.41% [33.67, 42.83])	1,252 (15.71% [13.24, 18.30])
Qwen (171 charts, 2,741 targets)	2,602	2,398 (92.16% [90.67, 93.51])	396 (16.51% [14.60, 18.83])	8.76%	614 (25.60% [23.18, 28.36])	100 (4.17% [3.11, 5.44])

Of Luna's 4,908 targets outside the broadest rule, 477 have no unused output in that chart and family, 3,972 have unused outputs with incompatible keys, and 459 have a compatible output assigned to another target; many-to-one assignment raises counterpart coverage to 44.17%. These are output categories defined by key rules and assignment, not shares of semantic omission, wrong decomposition, hallucination, or formatting.

D.2 INSTANCE LIST: GAIN COMPOSITION IN THREE DEPLOYMENTS

This section expands Table 1 in Section 6.1. Luna, Qwen, and K3 are compared on the same 171 charts, 106 airports, and 2,741 targets per condition, between free generation with reading rules and L4-self; both conditions require the model to write full identities. The net gain in strictly correct targets splits exactly into three terms: targets unmatched under free generation that become matched and fully correct; targets correct under free generation that become unmatched; and the net change among targets matched in both conditions. The companion rows in Table D2 give the change in value accuracy on instances matched in both conditions.

Table D2. Composition of the net gain in strictly correct targets. For each deployment, the first row gives the composition and the second the companion reading on jointly matched instances; brackets are 95% intervals from paired airport resampling.

Deployment	Reading	Free → self	Newly matched and correct	Correct under free, unmatched under self	Net change among jointly matched	Net gain or difference
Luna	Strictly correct	1,116 → 1,958	+873	−11	−20	+842; +30.72 [26.00, 35.32]
Luna	Jointly matched, 1,242	1,105 → 1,085				−1.61 [−3.79, 0.85]
Qwen	Strictly correct	158 → 799	+661	−18	−2	+641; +23.39 [20.91, 26.12]
Qwen	Jointly matched, 305	140 → 138				−0.66 [−6.91, 5.54]
K3	Strictly correct	1,348 → 2,020	+670	−2	+4	+672; +24.52 [21.64, 27.65]
K3	Jointly matched, 1,552	1,346 → 1,350				+0.26 [−0.38, 1.04]

In all three deployments, nearly all of the net gain comes from newly matched and fully correct targets, and the intervals for the change in value accuracy on jointly matched instances include zero. Of the 12 deployment-by-family changes on jointly matched instances, only Qwen's fix-role family has an interval excluding zero (−16.67 [−33.90, −2.86], n = 48), a small sample without multiplicity adjustment; K3's jointly matched families are 295 roles (all correct in both conditions), 326 segments (−1.23 [−3.41, 1.11]), 401 vertical constraints (+2.00 [−0.35, 4.57]), and 530 minimums (all correct in both). Once instance coverage is high, attribute determination for segments remains a substantial problem (Table D3).

Table D3. Identity matching and the segment family, all on 171 charts; the segment family has 787 targets.

Deployment	Identity matching, free → self (of 2,741)	Matching difference [95%]	Segment matching, free → self	Segment strictly correct, free → self	Segment strict difference [95%]
Luna	1,253 → 2,708 (45.71% → 98.80%)	+53.08 [47.93, 58.14]	165 → 783	76 → 191	+14.61 [11.08, 18.47]
Qwen	335 → 2,026 (12.22% → 73.91%)	+61.69 [56.92, 66.23]	45 → 554	14 → 97	+10.55 [7.82, 13.62]
K3	1,554 → 2,712 (56.69% → 98.94%)	+42.25 [37.97, 46.98]	326 → 787	146 → 212	+8.39 [5.65, 11.42]

With the same material, answering by row number (L4) gives no shared additional strict-accuracy gain in the three deployments (Table D4). Qwen's entry-response rate rises by 19.19 [14.65, 23.98] points under row numbers, but answering a supplied entry and matching a model-written identity are different events.

Table D4. Row-number answering versus model-written identities with the same material. Jointly answered targets are matched or answered in both conditions.

Deployment	Strictly correct, self → number	Strict difference [95%]	Jointly answered n	Correct among jointly answered, self → number	Value difference [95%]
Luna	1,958 → 1,896	−2.26 [−3.96, −0.68]	2,669	1,929 → 1,875	−2.02 [−3.56, −0.45]
Qwen	799 → 828	+1.06 [−1.99, 4.13]	1,887	741 → 595	−7.74 [−10.61, −4.96]
K3	2,020 → 1,989	−1.13 [−2.23, −0.17]	2,667	2,001 → 1,981	−0.75 [−1.83, 0.19]

The gain composition is a counting identity on saved outputs, not a mediation estimate. For Luna and Qwen, the free-with-reading-rules and L4 outputs were collected at a different time from L4-self (Luna's backend weights cannot be verified), whereas K3's three conditions were collected together; intervals reflect airport composition only (seed 20260924 for Luna and Qwen, 20260926 for K3).

D.3 Field list: fixed denominators, three chart sets, and two answer designs

This section gives the definitions, subsets, and alternative answer design behind Table 2 in Section 6.2. The field diagnostics use saved L4 and L5 answers on all 171 charts, with fixed denominators of 787 segments carrying reference values (another 164 segments without reference values are never scored) and 931 distance and course cells (719 distances and 212 courses). Missing rows, refusals, missing fields, and incorrect values count as failures; vertical mode is specified as descend by the common prompt and is not counted among the numerical cells. Symmetric removal deletes only field keys the reference does not score from both conditions' existing value dictionaries. It adds no values, changes no values, and leaves matching unchanged; it diagnoses which part of the scoring difference is due to field sets and is not an official score. The difference remaining after removal lies mainly in fields that should have been written but were not: Qwen writes 0 of 207 vertical-mode fields correctly in L4 and 193 in L5 (+93.24 [89.67, 96.55]), and Luna 158 and 206 (+23.19 [17.33, 29.63]).

Two subsets check whether the conclusion depends on chart selection (Table D5). The 89-chart subset drops 82 charts containing entries with no scored reference value: such entries are numbered and require answers in L4 but are listed as exempt rows in L5, changing the number of answer rows and subsequent numbering. The 78-chart subset further drops 11 charts with disambiguation slots; all disambiguation slots lie in vertical constraints and minimums, none in segments, and L4 and L5 do not display disambiguating values, so this step is unnecessary for the segment analysis and the 78-chart set serves only as a historical check.

Table D5. Field diagnostics on three chart sets, L4→L5. Differences in percentage points; brackets are 95% intervals from paired resampling of 59–106 airport clusters.

Chart set	Model	Complete-record strict correctness	Complete record after symmetric removal	Distance and course
171 charts (106 airports; 787 segments; 931 cells)	Luna	191 → 721, +67.34 [62.63, 71.59]	685 → 733, +6.10 [2.08, 10.43]	875 → 878, +0.32 [−2.59, 3.44]
	Qwen	66 → 317, +31.89 [26.42, 38.43]	207 → 345, +17.53 [12.13, 24.26]	425 → 449, +2.58 [−2.62, 8.31]
89 charts (68 airports; 438 segments; 531 cells)	Luna	116 → 404, +65.75 [59.19, 72.07]	389 → 408, +4.34 [−0.59, 9.32]	507 → 501, −1.13 [−4.66, 2.32]
	Qwen	45 → 190, +33.11 [24.34, 41.77]	131 → 204, +16.67 [8.37, 25.05]	263 → 269, +1.13 [−7.93, 9.90]
78 charts (59 airports; 387 segments; 478 cells)	Luna	110 → 359, +64.34 [58.13, 70.27]	348 → 363, +3.88 [−1.08, 9.30]	460 → 454, −1.26 [−4.53, 2.01]
	Qwen	45 → 174, +33.33 [23.98, 42.82]	125 → 184, +15.25 [6.05, 24.55]	254 → 256, +0.42 [−9.61, 10.06]

The alternative design uses fixed candidate fields: both conditions provide the complete instance catalog and require full identities and every candidate field, and they differ only in whether the model is told which segment fields carry reference annotations; unannotated values are not scored. This design shares 36 charts, 170 segments, and 210 numerical cells (161 distances and 49 courses) with the layered design above, with identical field keys, reference values, image bytes, and recorded model name and reasoning settings. On these 210 cells, Luna's correct readings go from 202 to 193 in the layered design. In the fixed-candidate design, Luna's correct readings go from 166 to 191 with lines and from 162 to 194 with JSON, a field-catalog effect of +13.57 [5.84, 22.36]; the effect on segment strict correctness is +10.88 [3.71, 19.40], and on strict correctness over all 544 targets +2.67 [−1.74, 7.33] (Table D6). About half of this numerical gain comes from "?" turning into correct values and half from incorrect values turning correct, with a few cells turning from correct to incorrect (Table D7). Qwen shows no detectable effect in the fixed-candidate design: −3.33 [−13.28, 6.81] on numerical cells, −6.18 [−14.60, 2.61] on segment strict correctness, and −3.49 [−9.68, 2.77] on all targets.

Table D6. Luna's counts under the two answer designs on the same 36 charts and 210 numerical cells.

Design	Condition	Segment complete records correct (of 170)	Numerical cells correct (of 210)	Numerical cells "?"	Numerical cells incorrect
Layered (numbered, fields chosen by the model)	L4	49	202	—	8
Layered (numbered, fields chosen by the model)	L5	153	193	—	17
Fixed candidate fields	No catalog, lines	130	166	20	24
Fixed candidate fields	No catalog, JSON	126	162	21	27
Fixed candidate fields	Catalog, lines	141	191	0	19
Fixed candidate fields	Catalog, JSON	152	194	4	12

Table D7. Paired transitions of Luna's numerical cells in the fixed-candidate design (no catalog → catalog; 210 cells per syntax).

Syntax	Correct→ correct	Correct→ incorrect	"?"→correct	"?"→ "?"	"?"→incorrect	Incorrect→ correct	Incorrect→ incorrect
Lines	158	8	19	0	1	14	10
JSON	161	1	14	4	3	19	8

The two designs also differ in identity form, rule wording, scoring requirements, and collection date, so the difference between them cannot be attributed to any single change; each cell has one generation, and intervals exclude generation variability.

D.4 Supporting contrasts

This section gives the readings for the supporting contrasts in Section 6.3, one paragraph each.

Reading rules and remaining value errors. Adding only the two reading rules produces no distinguishable change in identity matching for either model; Luna's value accuracy on instances delivered in both conditions rises clearly, while Qwen's moves in the same direction on a small sample with an interval including zero (Table D8). With both instance identities and fields supplied, value errors remain on the same targets answered in all four model-condition cells, and their level differs sharply between models.

Table D8. Reading rules and remaining value errors. Jointly delivered or answered targets are matched or answered in the relevant conditions.

Reading	Denominator	Luna	Qwen
Reading rules: identity-matching difference	2,741	+2.33 [−2.39, 6.61]	−0.44 [−2.55, 1.66]
Reading rules: value-accuracy difference, jointly delivered	875 / 163	+11.54 [7.34, 15.17]	+5.52 [−1.36, 12.14]
Reading rules: value-accuracy difference, minimums	359 / 63	+27.02 [17.65, 35.65]	+7.94 [−1.72, 18.18]
Field list L4→L5: joint field-and-value accuracy, segments answered in all four cells	698	+66.48 [61.46, 71.12]	+31.95 [25.98, 38.68]
Still incorrect after L5: segments answered in all four cells	698	8.5%	59.0%
Still incorrect after L5: minimums answered in all four cells	521	15.7%	92.9%

The model difference in the field row is −34.53 [−43.20, −24.83], so the segment magnitude gap does not arise from different jointly answered denominators; no interval is computed for the difference in remaining errors, which is compared only by level.

Interface and complete frame. Under free generation, stating the four identity conventions (C1) raises identity matching only for Luna, and switching to schema-constrained structured output (S-free) barely changes matching for either model. Within the four full-identity cells of the fixed protocol with the complete instance catalog, JSON versus lines changes results substantially only for Qwen, mainly through whole-document grammar validity (Qwen's valid line documents are 13/36 and 16/36, versus 36/36 and 32/36 for JSON). In the row-identifier cells, 1,910 of Qwen's 1,923 parseable records write full identity key-value pairs instead of the requested row identifier, and no target is strictly correct; correctly filled controls for all 144 tasks retain 2,176/2,176 targets, so the scoring interface drops no valid identifier answer. The complete frame supplies identities, requested fields, some disambiguating values, and reading rules together; strict accuracy falls sharply without the image (Table D9).

Table D9. Interface contrasts and the complete frame.

Reading	Luna	Qwen
C1: free-generation identity matching (2,741)	40.50% → 59.03%, +18.53 [14.02, 22.59]	12.66% → 11.42%, −1.24 [−3.35, 0.73]
S-free: free-generation identity matching (2,741)	40.50% → 40.72%, +0.22 [−3.58, 3.71]	12.66% → 12.84%, +0.18 [−2.41, 2.55]
Fixed protocol, full identity: JSON versus lines, strict correctness (544)	+3.03 [−1.95, 8.04]	+23.16 [16.03, 30.28]
Fixed protocol, full identity: JSON versus lines, identity matching (544)	0 (543 in all four cells)	+55.06 [43.57, 66.01]
Fixed protocol, row-identifier cells: strictly correct (of 544)	263, 172, 290, 152	0, 0, 0, 0
Complete frame, strict accuracy: free → frame with image	35.31% → 90.57% (862 charts)	5.07% → 53.12% (171 charts)
Complete frame without image	15.03%	21.78%
Distractor frame (true targets)	87.37%	47.21%

C1 changes content and length together, and S-free changes container, state representation, extra-field handling, and parsing route together; the row-identifier result is specific to this prompt structure, where full identity key-value pairs in the catalog may cue copying.

Twelve cumulative conditions. Among Luna's 15 pairs, adding unordered fix names (L1→L2) mainly raises output coverage, while adding requested fields (L4→L5) and the reading rules (free → free with reading rules) mainly raise attribute accuracy on the joint subset (Table D10). The name effect on fix-role matching is similar in the two models (Luna 17.78 [11.48, 23.72], Qwen 17.61 [12.30, 23.15], difference −0.17 [−7.97, 7.91]); on ordered segments it appears only for Luna (46.51 [38.18, 53.94] versus −0.13 [−2.68, 1.94], difference −46.63 [−53.74, −38.72]); on vertical constraints it is 20.08 [13.45, 25.63] for Luna and 3.35 [−2.54, 8.49] for Qwen (difference −16.73 [−23.42, −9.39]). Both lower responses come with concurrent changes: L4→L4′ changes segment coverage by −19.82 [−26.58, −12.55] points, with all 158 downward flips answered under different identity keys as identity authorship, presentation, and parsing change together; in L6, the 98 fully supplied vertical-constraint targets show a response-rate change of −19.23 [−30.80, −7.40] points, and all of them involve identity-key collisions.

Table D10. Luna's fifteen pairs (later minus earlier, percentage points). Output coverage uses 2,741 targets and denotes identity matching or entry response according to the identity source; attribute differences use the listed joint subset.

Pair	Output coverage Δ [95%]	Joint subset n	Attribute correctness Δ [95%]
Free → Free + reading rules	2.33 [−2.39, 6.61]	875	11.54 [7.34, 15.17]
Free + reading rules → L0	−0.44 [−5.58, 4.99]	890	−1.69 [−4.17, 1.08]
L0 → L1	3.87 [0.45, 7.58]	942	0.11 [−2.40, 2.73]
L1 → L2	21.96 [17.03, 26.24]	1,118	0.45 [−2.24, 3.17]
L2 → L3	11.49 [7.58, 15.51]	1,732	−2.25 [−4.09, −0.52]
L3 → L4	15.72 [13.71, 18.02]	2,249	0.58 [−1.22, 2.55]
L4 → L5	0.47 [−0.23, 1.25]	2,673	18.82 [16.35, 21.10]
L5 → L6	0.07 [−0.83, 0.88]	2,685	2.01 [0.04, 3.98]
L6 → R	0.11 [−0.54, 0.78]	2,688	−0.63 [−2.33, 0.92]
L2 → L3′	11.24 [6.65, 16.14]	1,634	−2.57 [−4.67, −0.40]
L3′ → L4′	10.76 [7.50, 14.20]	2,144	0.47 [−2.05, 2.96]
L4 → L4′	−5.22 [−7.40, −2.94]	2,516	0.68 [−1.73, 2.91]
L3 → L3′	−0.26 [−3.41, 3.52]	1,803	0.28 [−1.92, 2.72]
Free + reading rules → L6	53.16 [48.56, 57.78]	1,249	3.12 [0.52, 5.49]
Free → L6	55.49 [51.44, 59.52]	1,186	14.84 [10.93, 18.30]

Each transition in the cumulative design inherits earlier information, and entry response is near ceiling at L4→L5 (Luna 96.78%–99.83% across families, Qwen 93.10% overall at L4), so unchanged coverage does not show that the field list has no effect on matching.

Prompt language. Conclusions concern Chinese prompts. On a separate fixed set of 36 charts with 522 targets per cell, both models produce English-prompt outputs for L1, L2, L4, and L5, compared with the Chinese outputs for the same charts and conditions. Luna's English-minus-Chinese strict-accuracy differences are +3.45, +6.32, +0.19, and −1.15 points, all with intervals including zero; Qwen's L4 rises by 6.13 [1.17, 11.11] points, while its L2 identity matching falls from 108/522 to 44/522. The field list's improvement on segments holds in both languages (Luna +63.57 in Chinese and +66.43 in English; Qwen +22.31 and +38.57), whereas the name gain in Qwen's fix-role matching drops from +25.00 to +5.65 (language difference −19.35 [−36.07, −3.91]). Chinese and English outputs were collected at different times; this check shows how contrast directions and absolute rates respond to prompt language, not that English is generally better.

D.5 THE UNIFIED DIFFICULTY SCALE IS NOT SUPPORTED

This section gives the readings for the secondary test in Section 6.3. The stronger hypothesis holds that the ledger's structural accounts—premise and dependency counts, chart regions and symbol channels, distinct rules on retained branches, the longest retained chain, and the enumeration account—form an instance-level difficulty scale; these accounts describe recorded evidence routes, not demonstrated minima over all valid derivations. The registered predictions are that reductions in remaining requirements explain improvements within levels and that fully supplied targets should not perform worse. Only 23 of 427 within-level coefficient entries meet the registered improvement rule, with direction and magnitude varying across families; adding enumeration features retrospectively raises Luna's area under the curve for identity mismatch by 0.0080 [0.0041, 0.0117], but the fix-role and minimums estimates are negative with intervals including zero. That some information supplies are effective does not establish the scale.