Assembly completeness and pose recovery in a robotic hand benchmark

Sunnyday Technologies

Abstract

Automated CAD assembly requires both complete component coverage and correct spatial relationships. These properties can be obscured by a plausible overall appearance. This exploratory HandBench study examined six retained Astra and Grok Bot attempts to reconstruct one robotic-hand assembly containing 29 component designs and 226 occurrences, under a target budget of 30 minutes and 100 execution-tool calls. Five returns were gradeable. Astra Medium and Extra high matched 12 and 20 reference placements; each of three supervised Grok Bot returns matched only the coordinate-frame anchor. Two returns contained all 226 occurrences despite limited reference-pose agreement. A known-correct control matched 226/226. Direct renders and an audit of returned calculation artifacts distinguish component coverage, placement agreement and reproducibility. The observations do not establish a model ranking: configurations, execution controls and evidence completeness differed, and alternative mechanically plausible poses were not assessed. The study provides a descriptive baseline and motivates reporting geometry, calculations and visual evidence together.

Introduction

CAD assembly involves relationships between parts, including their relative positions and permitted movement. Learning-based joint prediction has been studied in systems such as JoinABLe, which uses parametric CAD information to assemble pairs of parts [1]. Reconstructing a complete arrangement from component geometry presents a related but distinct evaluation problem: a system can return every required part without recovering the specified assembly pose.

HandBench examines that distinction using Pollen Robotics' four-fingered Amazing Hand [2]. The present case targets a single authored static snapshot. Its purpose is to retain an initial baseline for later system updates, with separate records of supplied components, matched placements and execution quality. It does not measure physical dexterity or assembly motion. Explicit computation records are also relevant: reproducibility depends on the availability of code and data, not solely on a final output [3].

Results

Coverage and reference agreement

The retained series comprises one Astra attempt at each of three reasoning settings and three supervised Grok Bot attempts. All six outcomes are reported; none was replaced. Five native CAD returns were graded. Astra Low lacked sufficient retained execution evidence and remains ungraded rather than receiving a geometric zero.

Configuration	Supplied / 226	Matched / 226	Execution qualification
Astra Low (Lite)	Unavailable	Ungraded	Timed out; incomplete evidence
Astra Medium	98	12	Timed out; not execution-qualified owing to call-accounting problems
Astra Extra high	226	20	Timed out; execution qualified
Grok Bot 01	1	1	Supervised; incomplete call accounting
Grok Bot 02	226	1	Supervised; incomplete call accounting
Grok Bot 03	139	1	Supervised; incomplete call accounting

Table 1. Descriptive outcomes for the retained pilot. Execution qualification concerns run control, not assembly correctness. The denominator includes missing occurrences. Component-level counts accompany the manuscript as source data.

Astra Extra high and Grok Bot 02 each supplied the complete inventory, but matched 20 and one placements, respectively. The common-frame anchor contributed one match to every graded return. Astra Medium and Extra high therefore matched 11 and 19 additional placements; no Grok return matched a non-anchor placement. These additional counts explain the original score and do not replace it. The known-correct native control matched 226/226.

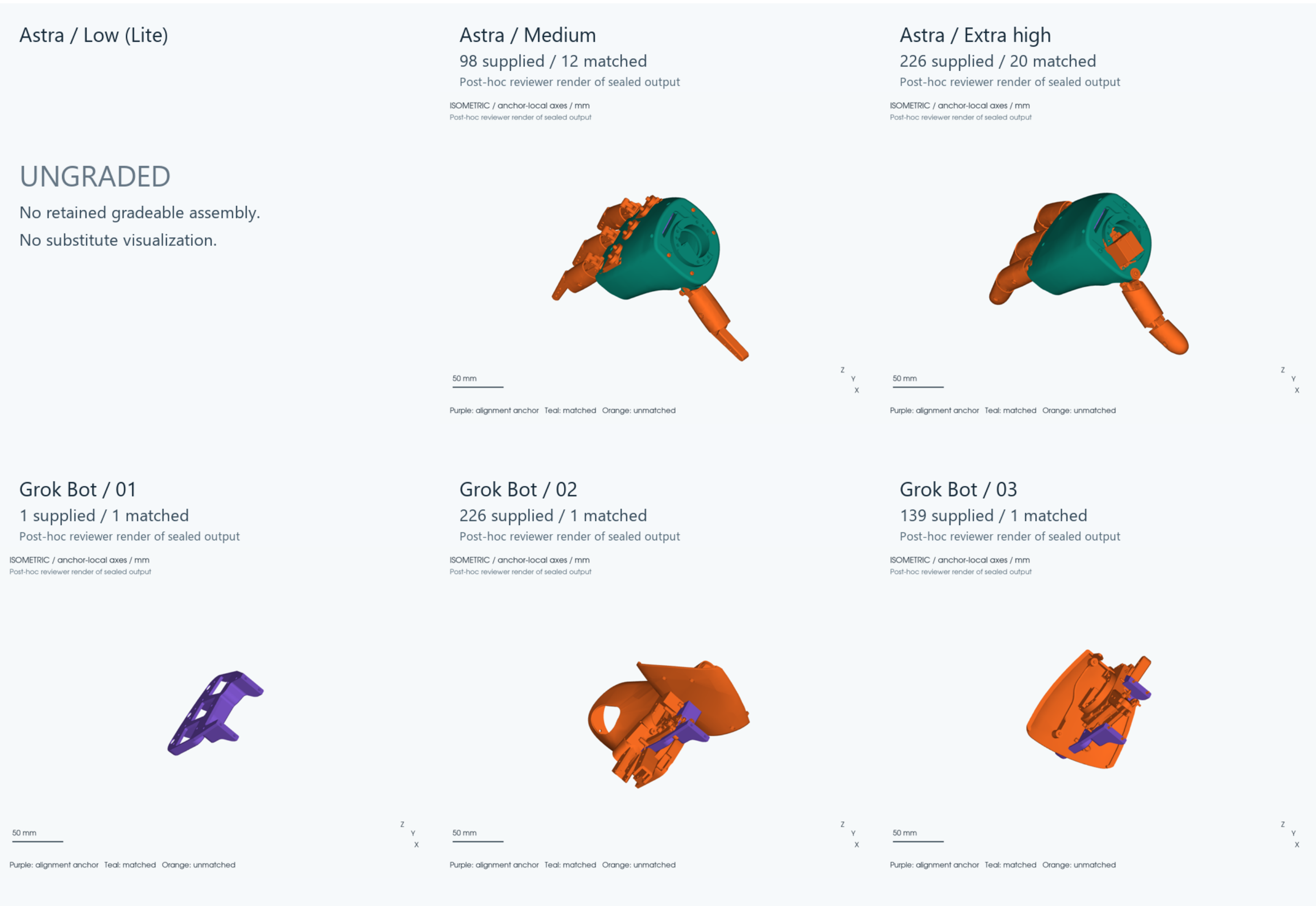


Figure 1. Post-hoc renders of the actual retained assemblies. Purple marks the alignment anchor, teal other matched placements and orange unmatched placements. Camera direction and magnification are shared; each return is centered separately. Orange denotes disagreement with the reference snapshot, not mechanical invalidity. No component was individually repositioned or repaired for display.

Visual and calculation evidence

The Astra returns had recognizable hand-like arrangements, whereas Grok Bot 02 formed a more compact grouping (Extended Data Figures 1-5). These are visual observations, not collision measurements. A separate post-hoc inspection found three pairs of identical components in Grok Bot 02 sharing exactly repeated native placements. They were required copies, not excess inventory; this observation did not alter the score.

Inspection without executing candidate code found that returned artifacts did not consistently include the procedures and dependencies needed to reproduce the placement process. Astra Medium's export script depended on a placement file absent from its returned folder. Extra high returned no build script. Grok Bot 01 and 02 supplied short stubs; Grok Bot 03 described a registration procedure, but its returned code only loaded the final placement record and counted entries. This audit concerns collected artifacts rather than every computation performed during a run. Build code and images were optional in the original task, so these omissions are not retroactive failures.

Export fidelity is a separate outcome

Native placement and STEP export fidelity measured different properties. The known-correct control passed placement measurement but had 199 passing STEP comparisons, 26 discrepancies and one inconclusive mass comparison. Consequently, export discrepancies alone cannot establish model-caused damage. The primary placement results are independent of these export diagnostics. Future representations require an adapter that preserves the geometry, units, identities and placements needed by the chosen metric.

Discussion

The retained outputs demonstrate a distinction between inventory completeness and reference-pose agreement within this case. The figures also show why apparent assembly quality cannot substitute for measurement: large matched surfaces can dominate a view even when many small components remain unmatched. Each required occurrence receives one unit of placement credit, irrespective of size or visibility.

The principal limitation is interpretive. A static snapshot is narrower than a set of mechanically valid configurations. The study did not establish that every withheld joint angle was uniquely recoverable from the supplied geometry. An unmatched placement therefore cannot, by itself, diagnose incorrect mating, nonfunctionality or the cause of an error. Joint connectivity, clearance, motion and strength require separately validated measurements.

The series is also too limited and heterogeneous for a model or reasoning-effort ranking. There was one Astra observation per setting, its controllers changed between attempts, and Grok's backend identity and complete call accounting were unavailable. Fresh Grok conversations shared account storage without verified isolation. No controlled higher-budget or Ultra comparison was performed. The low counts cannot be attributed causally to reasoning settings or resource restrictions.

The practical contribution is a documented baseline with inspectable outputs, not evidence of general robotic-hand capability. Future comparisons should retain the task and primary metric, disclose changes in models and execution tools, and preserve unsuccessful outcomes. Additional tasks and repeated runs would be needed to assess repeatability and broader applicability.

Methods

Task and execution

The frozen worker kit contained 29 component definitions and 226 required occurrences drawn from the reviewed Amazing Hand case. An occurrence denotes one required copy. Astra and Grok received identical component inputs and inventories, without the solved reference placements. The design was publicly available upstream; withholding the benchmark reference does not establish absence from model training data.

The target limits were 1,800 seconds and 100 execution-tool calls. Astra's time cap included a five-second shutdown reserve; collection and grading followed output freezing. Astra used the requested Low (menu Lite), Medium and Extra high settings. Grok used a managed-default Bot configuration with operator-supervised stopping and an instructed call ceiling. Hard remote enforcement and complete Grok call accounting were not established. A tool call may perform many numerical operations; it is not equivalent to a solver iteration, conversational turn or token count.

Geometry and placement checks

The native evaluator checked inventory identities, valid rigid transforms and preserved component geometry. Actual CAD transforms had to agree with the declared transforms within an elementwise tolerance of `1e-9`; scoring used the actual CAD placements. Local-geometry preservation was checked through corresponding serialized component data with bounded floating-point differences. All five graded returns passed the relevant native-geometry checks.

For submitted transforms `S_i`, reference transforms `R_j` and anchor `a`, comparison used:

```
submitted relative transform = inverse(S_a) × S_i
reference relative transform = inverse(R_a) × R_j
```

This removes a common rigid movement of the assembly. The eligible anchor becomes identity in both frames and therefore does not demonstrate an inferred relationship between different components. Copies of the same component were assigned to reference slots through maximum-cardinality one-to-one matching.

The method, `marb.amazing_native_sampled_coverage.v1`, uses each template's fixed `Shape.Center()` reference point and the maximum sampled point-to-boundary distance in each direction. A pair qualified when its reference-point separation and both directional maxima were at most **0.100001 mm**: a nominal 0.1 mm cutoff plus an explicit 0.000001 mm numerical guard. The reference point was not treated as a validated physical center of mass. Raw distances were compared without rounding; no retained result depended on guard-band credit.

The reviewed template sample set covered 1,109 faces with 5,125 points across the 29 definitions. Sampling used all vertices and up to three strictly interior points per trimmed face, with a bounded fallback for initially missed faces. Each sample's minimum distance to the opposite CAD boundary was measured, and the maximum was taken separately in each direction. A shortcut assigned zero deviation to accepted identical templates whose relative matrices agreed within `1e-12` elementwise. These are finite sampled checks, not a full-surface bound or manufacturing tolerance. Retained pair-distance records cover candidates passing the initial proximity filter and cannot support a complete error-distance distribution for unmatched parts.

Controls and descriptive analysis

The known-correct native control tested recognition of the accepted assembly. Additional controls covered global rigid movement, interchangeable copies, symmetry, displacement, geometry alteration and disagreement between declared and actual transforms. These controls support the specified digital measurement, not the unique inferability of the target or physical functionality.

The unit of observation was an attempt on this single assembly. Its 226 occurrences are coupled parts and were not treated as independent experimental replicates. Analysis was descriptive: no significance tests, population confidence intervals or estimated model effects were calculated. The unavailable Low outcome was not imputed. Anchor breakdowns, repeated-placement inspection, returned-code audit and reviewer renders were post-hoc analyses of unchanged results.

Inspectable calculations and images

A rigid transform specifies rotation and translation, but does not alone reproduce how the placement was selected. For example, one actual Grok Bot 02 placement was:

```
T = [ 0.067638  -0.230287   0.970769  -33.216116 ]
    [ 0.517321  -0.823889  -0.231488  -26.659567 ]
    [ 0.853115   0.517856   0.063406  -38.708297 ]
    [ 0          0          0           1        ]

For an illustrative local point p = (10, 0, 0) mm:
world point = rotation × p + translation
            ≈ (-32.539737, -21.486361, -30.177151) mm
```

Values are rounded for display; full-precision coordinates were used for the check. The test point is illustrative rather than an identified physical feature. This is candidate output, not a target transform, and the occurrence did not match the reference.

Figures were rendered from frozen native solids after the runs. Input file fingerprints were checked against retained grading records. Each return received one inverse-anchor viewing transform, preserving all relative placements. The overview shares magnification across runs; each three-view plate is fitted separately. CadQuery 2.7 produced display meshes using a tolerance argument of 0.07 in relative-deflection mode and an angular argument of 0.2 radians; existing triangulations may be reused. This is not an absolute mesh-error bound. Display meshes were separate from the CAD-boundary scoring procedure.

Future cohorts will require geometry, complete calculation dependencies, three documented assembly views and verification records linked to the same frozen output. This is a prospective reporting requirement; automated collection enforcement is not yet integrated. Candidate-generated images consume the candidate budget. Post-freeze reviewer rendering must not change placements or provide feedback to an active candidate.

Data availability

The accompanying source-data table provides required, supplied and matched counts for all component definitions across the five graded returns. All six assembly figures are included. Underlying CAD returns, evaluation records and reference placements are retained separately and are not openly deposited. Access requires review of source rights and benchmark-confidentiality restrictions; this limits independent reproduction from the reader package alone.

Code availability

Evaluation, rendering and analysis code are retained internally and have not been deposited in a public release. The method and numerical example are described here; the package does not claim to provide a complete executable reproduction environment. The candidate-code omissions reported above remain unresolved.

Competing interests and AI assistance

Sunnyday Technologies develops CADCLAW and operates MARB, the software and benchmark associated with this study. AI tools assisted software development, analysis scripting, manuscript preparation and internal review. The assembly figures are direct CAD renders. Internal AI review does not constitute external journal peer review.

References

1. Willis, K. D. D. et al. JoinABLe: Learning Bottom-Up Assembly of Parametric CAD Joints. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15849-15860 (2022). Publisher record.

2. Pollen Robotics. *AmazingHand: code and model to control the AH!* Project repository. The benchmark uses its separately frozen case inputs.

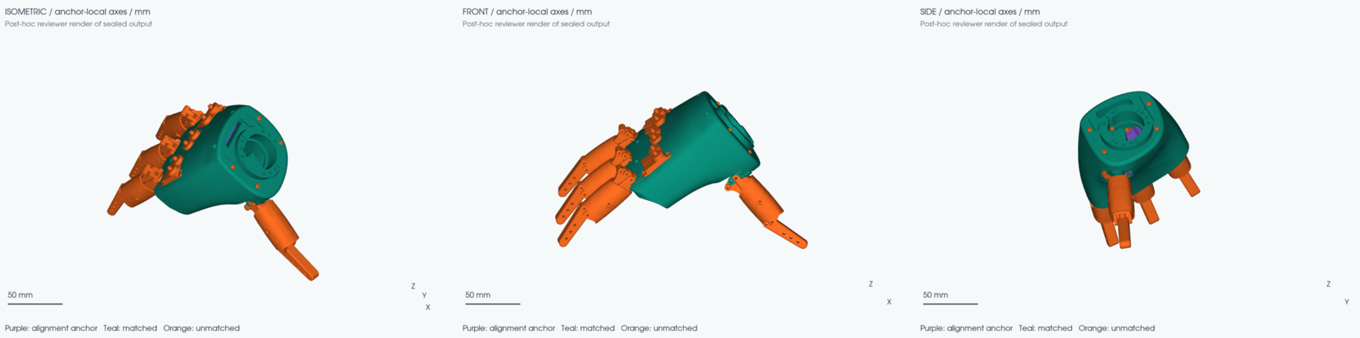
3. Pineau, J. et al. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research* **22**, 1-20 (2021). Journal record.

Extended assembly views

Extended Data Figure 1 | Astra Medium

Astra / Medium

98/226 supplied | 12/226 matched | 11 beyond the anchor



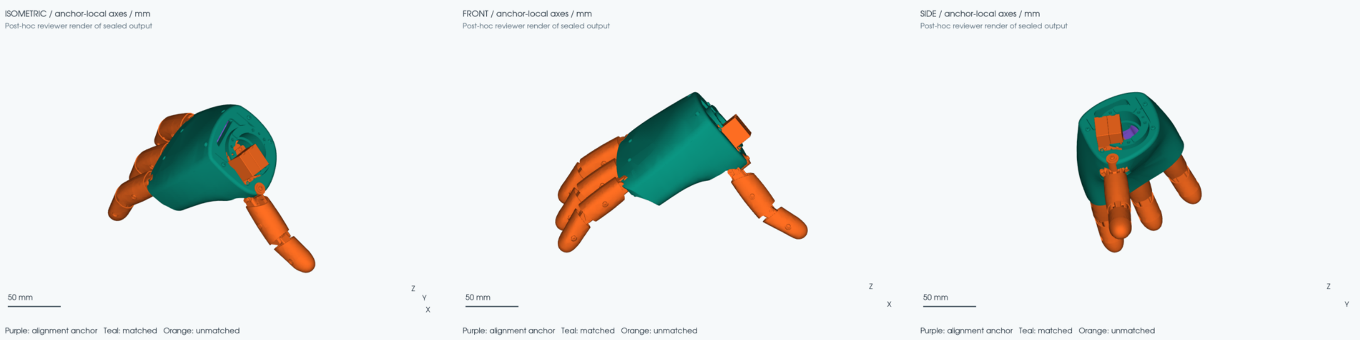
Post-hoc reviewer render of sealed output

98 occurrences supplied; 12 matched, including the anchor. Another 128 occurrences are absent. Directions are defined in the anchor-local frame, not anatomically. Scale bars are in millimetres. All extended figures use the color convention in Figure 1 and identify their post-hoc origin.

Extended Data Figure 2 | Astra Extra high

Astra / Extra high

226/226 supplied | 20/226 matched | 19 beyond the anchor



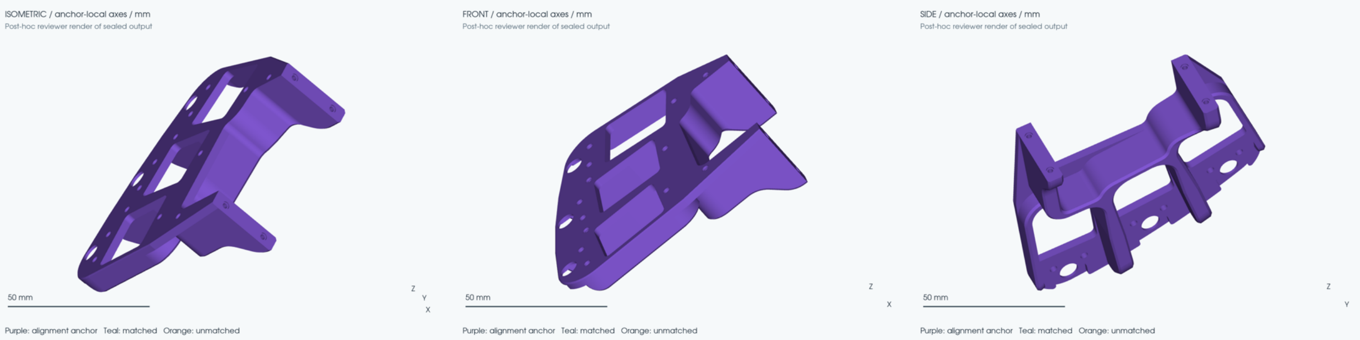
Post-hoc reviewer render of sealed output

All 226 occurrences supplied; 20 matched. Large teal surfaces do not receive more credit than small matched parts. Each occurrence contributes at most one match.

Extended Data Figure 3 | Grok Bot 01

Grok Bot / 01

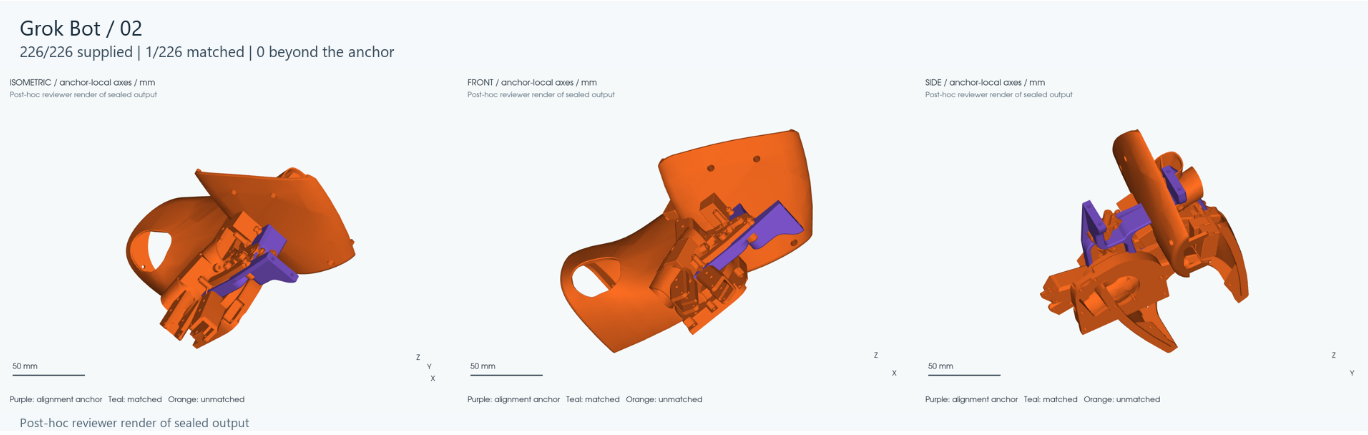
1/226 supplied | 1/226 matched | 0 beyond the anchor



Post-hoc reviewer render of sealed output

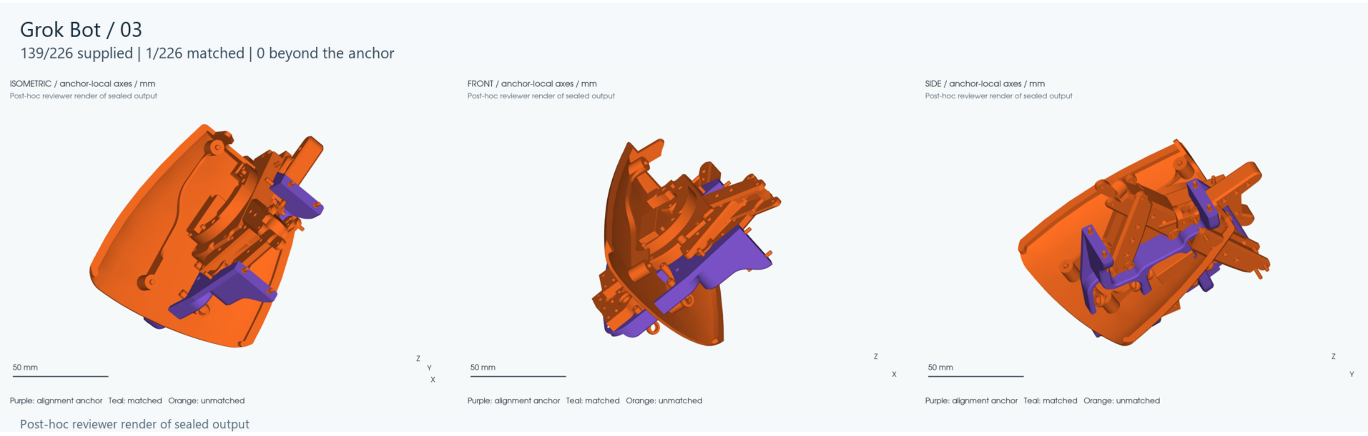
Only the alignment component was returned. Its match supplies no evidence of another recovered component relationship.

Extended Data Figure 4 | Grok Bot 02



All 226 occurrences supplied; only the anchor matched. The compact grouping is visible in three directions. No collision or mechanical-function claim is inferred from these views.

Extended Data Figure 5 | Grok Bot 03



139 occurrences supplied; one matched and 87 absent. The arrangement differs from Grok Bot 02 despite the same matched count.