

Design Intent Extraction from Engineering Drawings: Benchmarking Agentic Pipelines against Frontier Multimodal LLMs

Abstract

Engineering drawings are the source of design intent: they carry the dimensions, tolerances and geometric representation that determine how a part is manufactured. A single misread tolerance or feature can lead to inappropriate downstream manufacturing processes that affect quality and productivity. Existing frontier models lack an accurate understanding of design intent, which leads to inappropriate suggestions for downstream operations. We ask whether a pipeline built from multiple specialised agents recovers design intent more accurately than a single call to a frontier model, and therefore supports better downstream decisions. The agentic pipeline we study first classifies an engineering drawing by process family and part geometry, then routes the drawing to a specialised extractor; we compare it against a single call to a frontier multimodal LLM. Both are evaluated on expert-annotated engineering drawings covering a range of geometry types and complexity levels. We measure two things: how well each system extracts dimensions together with their tolerances, and how well it recovers volumetric attributes. Scoring uses an LLM judge that reads each answer for what it specifies rather than how it is written, and we verify that it grades identical answers identically across systems.

1. Introduction

An engineering drawing is how design intent is communicated to manufacturing. It fixes the dimensions a part must hold, the tolerances it may deviate within, and the geometric controls that constrain form and position. It is the last artefact unambiguously authored by a human, and every manufacturing decision that follows is derived from it.

Reading it automatically is harder than it looks. Dimension text is small and frequently rotated; tolerance and GD&T symbols carry meaning that character recognition alone does not recover; and a value only becomes meaningful once it is bound to the feature it dimensions, which is often drawn in a different view. Drawings also vary enormously in density, in our set, from 3 to 42 annotated dimension rows on a single sheet, so a method that works on a

simple turned part can fail completely on a dense prismatic one.

What makes errors here consequential is that extraction sits at the head of a chain, and nothing downstream can recover from a mistake made at the top. A tolerance read as ± 0.1 mm instead of $(0, +0.1)$ mm may break the geometric structure of the part and make it infeasible to manufacture at all, and each such error cascades. The overall envelope determines the raw stock that must be purchased, and an undersized blank cannot be machined into the part at any cost. The feature inventory, how many holes, pockets and faces, and on which sides, determines the process plan: which operations are required, on which machines, and how many setups the part needs, since features on opposing faces force a refixturing that features on one face do not. Those operations, with their feed and speed parameters, determine cycle time. Cycle time multiplied by machine rate, plus material and setup, determines should-cost. A single misread tolerance therefore does not produce a slightly wrong number at the end; it changes what is being made, and every quantity computed after that point answers a different question.

This is why extraction is the backbone rather than a pre-processing step. Downstream reasoning can be sophisticated and still be wrong for reasons that have nothing to do with its own quality, and a cost estimate cannot be audited without knowing whether the specification it rests on was read correctly. Evaluating how well a system reads a drawing is therefore a prerequisite for evaluating anything built on top of it.

Existing work approaches the task either as detection, localising callouts and reading them with OCR, or as visual question answering over the drawing [2, 4, 7]. Neither produces what a downstream planner needs, which is a structured record of every dimension, the tolerance attached to it, and the feature it belongs to.

Two system designs now compete for this task. The first is direct: give a frontier multimodal LLM the drawing and a schema, and take its answer in one call. The second decomposes the problem, classify what kind of part and process the drawing describes, then route it to an extractor built for that case. The first benefits from the general capability of large models; the second encodes domain structure that a single

prompt cannot. Which design reads design intent more accurately, and where each fails, has not been measured on a common set of drawings under a common scoring rule.

Two properties make that comparison difficult to measure. First, correctness is not string equality. A drawing’s notation admits many equivalent forms, a tolerance may be written symmetrically or as two bounds, a datum letter may or may not be transcribed with the value it qualifies, and a system should not be penalised for the form it chose. Second, and more subtly, a correct answer may be *decomposed* differently: a chamfer annotated $1 \times 45^\circ$ is a single callout on the sheet, but specifies a 1 mm linear leg *and* a 45° angle, and a system that reports these as two items has not made an error. Any comparison across systems that represent their output differently must resolve this before it can score anything.

We study that question. We assemble 47 expert-annotated material-removal drawings spanning five geometry classes and three expert-assigned complexity levels, all in English and dimensioned in millimetres. Against this we run four frontier multimodal LLMs and EIM, a multi-agent pipeline that classifies process family and part geometry before extraction. We evaluate two things: whether a system recovers dimensions together with their tolerances, and whether it recovers volumetric structure rather than transcribing text, the latter indicating whether the model has a structural understanding of the part before proceeding to process planning and raw-stock calculation.

In summary, our contributions are as follows:

- A staged extraction engine. We describe EIM, a multi-agent pipeline that first classifies a drawing’s process family and then its part geometry, and routes it to an extractor specialised for that case, rather than asking one model to read every drawing the same way.
- A comparison study against frontier multimodal LLMs. We evaluate EIM against four frontier models on identical drawings under an identical scoring rule, and characterise where each fails, by geometry class, dimension type and drawing complexity, together with the token cost each design incurs.
- A judge framework for cross-system comparison. Because two systems may express the same specification in different notation and at different granularity, we score with an LLM judge that resolves both before comparing, and we verify that it returns the same verdict for the same content regardless of which system produced it.

2. Related Work

for engineering drawing understanding. Several benchmarks now probe whether multimodal models can read technical drawings, and they agree on where the difficulty lies. DesignQA [2] pairs 1,451 questions with Formula SAE rules, CAD images and drawings, and reports

that current MLLMs struggle in particular to analyse engineering drawings. MechVQA [7] is the largest such resource, with 3,281 mechanical drawings and roughly 21K question–answer pairs across recognition, reasoning and judging tasks, covering GD&T symbols, tolerance values and datums. AECV-Bench [6] evaluates architectural drawings and finds a clear capability gradient: text-centric OCR questions reach 0.95 accuracy while symbol-centric tasks remain at 0.40–0.55, concluding that models act as competent document assistants but lack drawing literacy. BlueprintSymVL [8] isolates symbol recognition, supplying a visual exemplar at query time to remove reliance on pretrained symbol knowledge, and reports degradation under clutter, confusion between visually similar symbols, and a tendency to hallucinate symbols that are not present. Notably, both MechVQA and AECV-Bench score with an LLM judge rather than exact matching, since equivalent answers take different surface forms. These benchmarks establish that symbol-grounded reading is the bottleneck, but all of them score answers to questions or counts of objects, not a structured record of the specification a planner consumes, and each evaluates one system against a reference rather than comparing systems that represent their answers differently.

Structured extraction of dimensions and GD&T. A second line of work targets extraction directly, and converges on a detect-then-parse design. Khan *et al.* [5] fine-tune Florence-2 on 400 expert-annotated drawings for GD&T extraction and report large gains over zero-shot GPT-4o and Claude-3.5-Sonnet, including a 52.4% F1 improvement and a 43.2% reduction in hallucination. The same group scales this to 1,367 drawings annotated across nine categories, pairing YOLOv11-obb localisation with a fine-tuned Donut parser to reach 94% F1, with a case study on downstream tool and process selection [4], and extends it to a three-stage multi-view framework that segments layout, detects oriented annotations, and parses them with separate alphabetical and numerical vision-language models [3]. Related work classifies GD&T symbols with YOLOv8 detectors [1]. This is the closest prior art to our task, and it is strong; but evaluation is performed on localised annotation patches using detection-style precision, recall and hallucination rate, which measures whether individual callouts were found and transcribed rather than whether the drawing’s specification was recovered as a whole. None of these systems is compared against a frontier model reading the complete sheet, and none reports the inference cost each design incurs.

3. Dataset

The benchmark is built from public-domain CAD parts in the STEP format drawn from automotive practice. The set spans flat plates, links, bushes, tubes and angle brackets at one end, through gears, sprockets, pulleys, bored hubs and

formed sheet-metal brackets in the middle, to castings and mouldings at the other: bell housings, engine blocks, turbocharger and pump volutes, intake and exhaust manifolds, lamp and trim bezels, a steering knuckle and an alloy wheel. Parts are assigned to one of three tiers by drafting complexity such as how much drawing work the geometry creates in views, sections, auxiliary views and special callouts. The tiering is explicitly not a judgement of manufacturing difficulty or of part size: a part that is hard to machine may still be trivial to draw, and vice versa.

The tiers are defined as follows:

- Simple: One feature type on a flat or plain-turned body: a flat plate of single thickness, a plain revolve such as a cylinder or sleeve, or a single bend. Features are few and all of the same trivial kind, such as through-holes on flat faces.
- Medium: Several distinct feature types, but the body remains prismatic or rotational, so a single section is enough to show it: keyways, bolt circles, stepped or bored hubs, raised bosses, tapped holes, multi-plane bends, toothed or knurled patterns, or an internal cavity that one section reveals.
- Complex: Any one of the following: true freeform or organic surfacing, as distinct from an axisymmetric revolve; a hollow casting or moulding with internal structure; features on compound or multiple angles requiring auxiliary true views; or a very dense set of features spread across multiple faces.

The tiers are deliberately unbalanced toward the difficult end, at 20 / 30 / 50 percent.

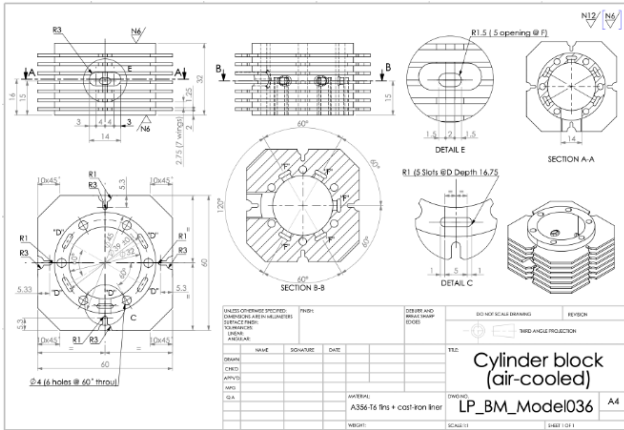


Figure 1. Drawing developed from the CAD model.

3.1. Composition

The benchmark is a set of 47 mechanical part drawings, each paired with an expert answer key, drawn from a planned 50-part collection. Every drawing has a ground-truth record, so the scored set and the extracted set are the

same.

Parts were selected to span manufacturing reality rather than a single process. The set covers sheet metal, high-pressure die castings, machined housings and brackets, turned and revolved parts, gears and sprockets, springs, tubing, injection-moulded trim and forged structural components, in 30 distinct material specifications:

Material matters to the task beyond identification: mass is computed as density $\times$ volume, so a model that reads the material specification incorrectly cannot recover mass even from a perfectly reconstructed geometry.

Drawings are multi-view and multi-sheet, with 257 views in total, a mean of 5.5 per drawing and a range of 3 to 14. The distribution is not the textbook three-view case:

Table 1. View Types

View type	Count	View type	Count
Section	57	Detail	23
Front	48	Bottom	9
Isometric	47	Auxiliary	6
Top	31	Left	4
Right	31	Rear	1

Sections and details together account for 80 views — nearly a third of the set. These are the views that carry internal geometry and enlarged tolerance callouts, and they require a model to associate a dimension with a feature shown at a different scale, in a different frame, elsewhere on the sheet or on another sheet entirely.

3.2. Drawing complexity

Each part carries an expert-assigned complexity label, and the label tracks measurable properties of the sheet rather than being a subjective impression:

Table 2. Drawing Complexity Metrics

Complexity	Parts	Dim rows (mean)	Median	Range	Views (mean)
Simple	10	11.3	9.5	2 – 28	3.6
Medium	15	20.7	17.0	9 – 36	4.7
Complex	22	34.1	32.0	11 – 61	6.8
All	47	25.0	23.0	2 – 61	5.5

A complex drawing carries three times the dimensional content of a simple one and nearly twice the views. The set is deliberately weighted towards the complex end, because that is where extraction is expected to degrade and where a difference between models is most likely to be observable; reporting a single aggregate over a set weighted the other way would understate the difficulty of the drawings that matter in production.

3.3. Dimension and tolerance inventory

The 47 drawings carry 1,175 annotated dimension rows, a mean of 25.0 per drawing, median 23, ranging from 2 on the simplest part to 61 on the densest. Rows are typed:

Table 3. Dimension Row Types

dimType	Rows	Share
Linear	717	61.0%
Diameter	218	18.6%
Radius	136	11.6%
Angular	57	4.9%
Chamfer	15	1.3%
Thread	10	0.9%
(control frame, no size)	22	1.9%

The type distribution is itself a difficulty signal. Linear and Diameter callouts dominate and are the easiest to read; the tail, Angular, Chamfer, Thread, is small but disproportionately error-prone, because each is a compound specification rather than a single number. A chamfer annotated $1 \times 45^\circ$ is one callout on the sheet but specifies a 1 mm leg and a 45° angle, and a thread callout such as $M10 \times 1.5$ |20.83 carries a designator and a depth. The 25 chamfer and thread rows therefore contribute roughly 50 scoreable items, and a per-type breakdown is reported in the results so that strong aggregate performance on Linear callouts cannot mask failure on the compound types.

The 22 rows with no dimension type carry a geometric control frame and no size, a flatness, parallelism or profile control applied to a face. They are the one row category that contributes nothing to the dimension view and everything to the GD&T view.

Tolerances and geometric controls. 89 rows across 17 drawings carry a tolerance or control in the gdAndT field, up to 11 on a single drawing, and a further 39 rows carry a two-sided tolerance written inline in the value. By kind:

Table 4. Tolerances and Geometric Controls

Kind	Rows
Plain size tolerance ($\pm$, $+a/-b$, a range)	54
Flatness	10
ISO fit code (H7, h8, ...)	9
Parallelism	5
Profile of a surface	4
Perpendicularity	4
Circular runout	2
True position	1

Two observations shape how this view is scored. First, GD&T is concentrated rather than uniform: 30 of 47 drawings carry no explicit control at all, so a per-drawing av-

erage over the whole set would be dominated by drawings with nothing to find, and the GD&T view is reported over the 17 drawings that have content. Second, the row count understates the scoreable content by roughly half, because every two-sided tolerance owes both of its bounds as separate items, the expansion is described in the methodology.

3.4. Annotation language

Every drawing in the set is in English and dimensioned in millimetres (47/47 on both counts). The only non-ASCII characters in the annotations are engineering symbols, $\emptyset$ (232 occurrences), $^\circ$ (77), $\pm$ (39) and the depth symbol $\downarrow$ (7), not natural-language text.

This is a deliberate scope decision rather than a coincidence: holding language and unit system fixed isolates drawing interpretation from language handling and from unit conversion, so a failure is attributable to reading the drawing. Multilingual and inch-dimensioned drawings are a separate axis of difficulty and are left to future work.

3.5. The volumetric block: probing structural understanding

Each answer key carries twelve quantities measured from the 3D STEP solid: surface area, volume, mass, the three minimum-oriented bounding-box extents L/W/H, the three principal mass moments of inertia I_1, I_3 about the centroid, and the three principal radii of gyration k_1-k_3 , where $k_i = \sqrt{I_i/m}$. Principal values are used rather than the raw inertia tensor because they are coordinate-frame independent and therefore comparable without first agreeing an origin and a set of axes.

None of these twelve numbers appears anywhere on the drawing. They cannot be transcribed, only computed — which is what makes this block the part of the benchmark that tests whether a model has actually reconstructed the part rather than harvested its text. A model can score well on dimensions by locating and reading callouts one at a time, with no coherent model of the object; it cannot produce a volume or a moment of inertia that way. The volumetric block is the point at which reading a drawing and understanding a part separate, and it is the capability that process planning and raw-stock calculation actually depend on downstream.

Each quantity additionally carries a derivability flag recording whether the 2D drawing alone determines it. The structure of these flags is informative:

Bounding-box extents are derivable on every drawing in the set — they are the overall envelope, and a complete drawing must state it. The remaining nine quantities are fully determined only on simple parts, are partially determined on medium ones, and on complex parts are not determined by the sheet at all: a die-cast housing with unstated draft, fillet and wall-thickness variation cannot be volumet-

Table 5. Derivability of Volumetric Attributes

Complexity	Bounding box L/W/H	The other nine quantities
Simple (10 parts)	Yes	Yes
Medium (15 parts)	Yes	Partial
Complex (22 parts)	Yes	No

rically resolved from its views.

This gives the volumetric view two distinct readings rather than one. The bounding box is a floor every model should clear, and a failure there is a failure of basic spatial consolidation, not a missing-information problem. The other nine quantities form a graded reconstruction test, and on the No cases a wrong answer is partly the drawing’s fault — so those cells measure the quality of a model’s estimate under acknowledged ambiguity, and are reported separately from the derivable ones rather than pooled with them.

One caveat on the flag: it is assigned per complexity class, not judged independently per part, so for the nine non-envelope quantities it carries no information beyond the complexity label. It is a usable stratification but should not be described as an independent annotation.

3.6. Coverage and known gaps

All 47 drawings have a complete drawing-side answer key. Coverage of the twelve solid quantities is partial, at 477 of 564 cells (84.6%):

Table 6. Volumetric Coverage

Volumetric coverage	Parts
All twelve quantities	39
Bounding box only (3/12)	3
None	5

The eight incomplete parts are awaiting STEP files or are assemblies rather than single solids (one is a three-item wheel/rim assembly). Any quantity the reference leaves blank is reported as unscored and is never counted for or against a model, so partial coverage reduces the size of the volumetric comparison without biasing it.

4. Methodology

We compare two designs on identical inputs: a staged multi-agent pipeline, which we call the Engineering Intent Model (EIM), and a single call to a frontier multimodal LLM. Both are scored by the same judge, described in Sec. 4.3.

4.1. The Engineering Intent Model (EIM)

EIM takes an engineering drawing and returns, in addition to the transcribed specification, thirteen mechanical prop-

erties of the solid the drawing describes: density (kg/m^3), surface area (mm^2), volume (mm^3), mass (g), the minimum oriented bounding box $L \geq W \geq H$ (mm), the three principal mass moments of inertia $I_1 \leq I_2 \leq I_3$ about the part centroid ($\text{kg}\cdot\text{mm}^2$), and the three principal radii of gyration k_1, k_2, k_3 (mm).

4.1.1. Reading the sheet

EIM looks at the sheet three times, performing a different job on each pass, at two resolutions. Cropping is local and has no payload limit, so close-ups are always cut from the original; because bounding boxes are expressed on a normalised 0–1000 grid and aspect ratio is preserved, boxes planned on the reduced copy map onto the original unchanged.

The first pass transcribes the whole sheet and rates its own legibility per item. It is not expected to succeed at everything; it is expected to be honest about what it could not read. A planning stage converts those flags into padded, page-numbered boxes, a cropping stage cuts them at full resolution, and a second reading stage reads only the crops, bounded to at most fifteen regions, with an empty region set mapping to an explicitly valid answer. On a clean sheet nothing is cropped, and this pass is nearly free.

A third pass targets the title block specifically, with its own planner and reading stage, because material and stamped mass are the two highest-value fields on a drawing and the first casualties of downsampling. The planner must emit both a tight box and a deliberately oversized one around the same block, and must box separately any other table whose heading could be confused with part mass. The reading stage treats the full sheet as a layout map only and the crops as the sole source of truth, under two governing rules: a field that is not clearly legible returns null and is named in an illegible-fields list rather than being completed from context, and callouts are transcribed character-for-character with no normalisation to an equivalent grade or standard.

4.1.2. From drawing to volumetric understanding

The interpretation stage receives the merged transcription, the authoritative title-block reading and the sheet, and produces geometry without arithmetic. Its output is elicited in a forced order, each field a separate commitment: projection angle and unit; a complete view inventory with each view classified and section and detail views identified as the only place internal geometry exists; the dimension table, with tolerance recorded as structured data in one of seven categories and one row per geometric control frame with its datums in frame order; a verbal description of the three-dimensional shape; a decomposition method; a declared part coordinate frame; and finally an ordered list of primitives drawn from a closed vocabulary of ten shapes, each with an add or subtract role, a quantity, an axis, a centroid

offset and the drawing dimensions it came from.

The computation stage then writes a Python script, executes it, and copies the printed values back without rounding. Six of the thirteen features depend on assembling a symmetric 3×3 inertia tensor with parallel-axis product terms and solving for its eigenvalues—arithmetic a language model reasoning in prose gets wrong in ways that look entirely plausible. Four consistency checks are run and reported: computed mass against any stamped mass on the drawing, the ordering $I_1 \leq I_2 \leq I_3$, each radius of gyration against the part envelope, and computed volume against the bounding-box product.

4.2. Frontier-model baseline

The baseline gives a frontier multimodal LLM the drawing and a single schema-constrained prompt, and takes its answer in one call: no decomposition, no retrieval and no tool use. The prompt asks for the unit, the material as written, the engineering notes, a flat list of every dimensional callout and every geometric tolerance, ordinate positions where the drawing uses ordinate dimensioning, and the twelve derived volumetric quantities; it explicitly forbids a process plan, machining steps, a feature list, a route or a cost estimate, so that the baseline and EIM are asked for the same object. Tolerance notation is pinned to a single form, (+upper, -lower), with symmetric tolerances written out rather than as $\pm x$, so surface-form variation is removed where it can be removed cheaply. Ordinate values are captured one axis per entry and never paired into points, so a sheet carrying four X ordinates and six Y ordinates yields ten entries rather than twenty-four. The full prompt and response schema are reproduced in ??.

and settings. We run four frontier models—Claude Opus 5, Gemini 3.1 Pro, GPT-5.6 and Grok 4.6—through one provider endpoint under identical settings: temperature 0.1, `response_format: json_object`, one call per drawing, and per-call token and cost accounting. Holding the harness fixed means an observed difference between models is attributable to the model rather than to how it was invoked.

. Drawings are supplied as native PDFs through the provider’s document plugin rather than rasterised to an image, so no resolution is lost to a conversion step—which matters on sheets whose smallest dimension text is a few pixels tall at any reasonable raster size. One model family caps total request size below what some sheets encode to; those drawings are downsampled in resolution steps until they fit, and this is recorded.

budget. One property of reasoning models shapes the configuration: on the serving API a single `max_tokens` budget caps reasoning tokens and visible output *together*. Measured on this dataset the visible JSON is 0.9–1.8k tokens while reasoning consumes 12–15k, so a conventional

16k cap starves the answer—one model returned JSON truncated mid-object and another returned an empty message. We therefore set a 32k default budget with a 64k ceiling, and because a budget failure is deterministic, re-issuing the identical request fails identically; each retry doubles the budget rather than repeating the call. A response is diagnosed before it is parsed, on the finish reason and on whether the completion consumed its budget, because a truncated body and an empty body are indistinguishable to a JSON parser but have different causes. The raw response and its token accounting are written out on failure, so an exhausted budget stays visible rather than being recorded as a zero score.

. Every drawing is extracted and only quantities the reference actually provides are scored, so extraction behaviour is observed on the whole set while the comparison stays restricted to annotated content.

4.3. Judge

Both systems are scored by the same judge under one rulebook, so the two sides cannot drift apart. The judge never lets code decide correctness: the surrounding program only counts the judge’s own item lists into precision, recall and F1. It runs GPT-5.6 at temperature 0 with a pinned seed and `response_format: json_object`; the system message carries the entire prompt and the user message carries only the payload, so the instruction text is byte-identical across every part and every system.

a judge rather than string matching. Correctness here is not string equality. The same specification is written many ways— ± 0.1 or $+0.1/-0.1$, M16 or 16, R0.8 or “radius 0.8”, Ø11 THRU or Ø11—and a system should not be penalised for the form it chose. More consequentially, systems differ in *granularity*: a chamfer annotated $1 \times 45^\circ$ is one callout on the sheet but specifies a 1 mm leg *and* a 45° angle, and one system may report it fused where another reports it split. Both are correct, and a comparison across systems that decompose their answers differently has to resolve this before it can count anything.

counts as one item. One rulebook, applied identically to every system, fixes the unit of scoring. *GD&T lives in two places*: a tolerance written inline in a dimension value is a GD&T item exactly as much as a control frame is, so a row whose value reads “4x Ø11 ± 0.1 THRU” under the frame “Positional Ø0.2(M) A B” carries three GD&T items—the frame and both bounds. Ignoring the inline tolerance understates the GD&T denominator by roughly half on this dataset. *Every tolerance splits into its bounds*, symmetric ones included, each matched on its own, so a system that recovers one bound and misses the other is half correct rather than wrong. *A quantity prefix is not a split*: “4x R3” is one radius occurring four times, and a prediction may write it either way. *Compounds split on the ground-*

truth side but may stay fused on the prediction side: a row typed Thread/Linear, and every chamfer, owes two items, and a prediction reporting both numbers inside one entry earns both—though a single number under the wrong type remains an extra plus a missing. *Value and type are coupled*: a correct number carrying the wrong dimension or control type is not correct. *The ground truth’s reasoning field is evidence, never an answer*: it decides whether a row is a chamfer, tells apart two equal values on different faces, and places an item in the right view, but it is never itself scoreable and a prediction is never penalised for carrying no reasoning text. Surface roughness is an engineering note and scores in neither view; a bare datum label carrying no tolerance is in no view at all.

5. Experiments and Results

5.1. Dimension extraction accuracy

Table 7 reports precision, recall and F1 on the dimension view, with the correct / extra / missing counts behind each score. All five systems are scored against the same frozen key of Sec. 4.3: 1222 scoreable dimension items over 47 drawings, so the denominators are identical across columns and the counts can be read against each other directly.

Three things are worth stating. First, the dimension view is close to saturated: four of the five systems sit between 0.951 and 0.967, a spread of sixteen thousandths across a proprietary engine and three frontier models. A benchmark that reported only this number would separate almost nothing. Second, the separation that does exist is entirely in precision, not in reading. Recall is between 0.975 and 0.985 for those four systems — they find essentially every dimension on the sheet — while precision ranges from 0.928 to 0.952, and every position in the ranking is decided by how many items a system emits that the key does not contain. EIM has the highest extra count of the four (93) and the lowest F1 of the four; Gemini has the lowest (60) and the highest F1. Third, Grok is a different failure mode rather than a worse version of the same one: its precision (0.929) is in line with EIM’s, but it omits 131 of the 1222 items, seven times the miss count of the best system, and that alone accounts for its position.

5.2. Error breakdown

Figure 2 breaks the dimension F1 down by two properties of the drawing that the ground truth records independently of any system: how many views the sheet carries, and how many scoreable items the part contains.

5.3. Volumetric accuracy

Table 8 reports the twelve volumetric quantities separately, with the bounding box split into its three extents, the principal moments into I_1 , I_2 , I_3 and the radii of gyration into

System	F1	Precision	Recall	C/E/M
Gemini 3.1 Pro	0.967	0.952	0.981	1199/60/23
Claude Opus 5	0.962	0.941	0.985	1204/76/18
GPT-5.6-sol	0.957	0.932	0.983	1201/87/21
EIM	0.951	0.928	0.975	1192/93/30
Grok 4.6	0.911	0.929	0.893	1091/83/131

Table 7. Extraction accuracy on the dimension view over 47 drawings. C/E/M are correct, extra and missing scoreable items against the frozen key of Sec. 4.3, which holds 1222 items and is identical for every system. Best per column in bold.

k_1 , k_2 , k_3 .

6. Discussions, Limitations and future work

The reconstruction is an approximation. The solid is assembled from a closed vocabulary of ten primitives rather than from a boundary representation. This intermediate representation is therefore not yet suitable for downstream process planning or other industrial applications, which require accuracy comparable to that of a true 3D solid model. Future work will focus on bringing the reconstruction closer to boundary-representation accuracy, so that it can support detailed applications such as manufacturing process planning.

7. Conclusion

We set out to measure whether decomposing engineering drawing understanding into a staged pipeline reads design intent more accurately than a single call to a frontier multimodal model, and to do so under a scoring rule that treats equivalent answers as equivalent regardless of the notation or the granularity in which they are expressed. Two quantities were measured on the same drawings: the recovery of dimensions together with their tolerances, and the recovery of the volumetric structure of the part which is a set of quantities that appear nowhere on the sheet and that can therefore only be reported by a system which has reconstructed the solid the views imply.

References

- [1] L. Bewoor et al. Deep learning-based classifier for geometric dimensioning and tolerancing symbols. *IAES International Journal of Artificial Intelligence*, 14(2):1345, 2025. 2
- [2] Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, 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. 1, 2
- [3] Muhammad Tayyab Khan et al. A multi-stage hybrid framework for automated interpretation of multi-view engineer-

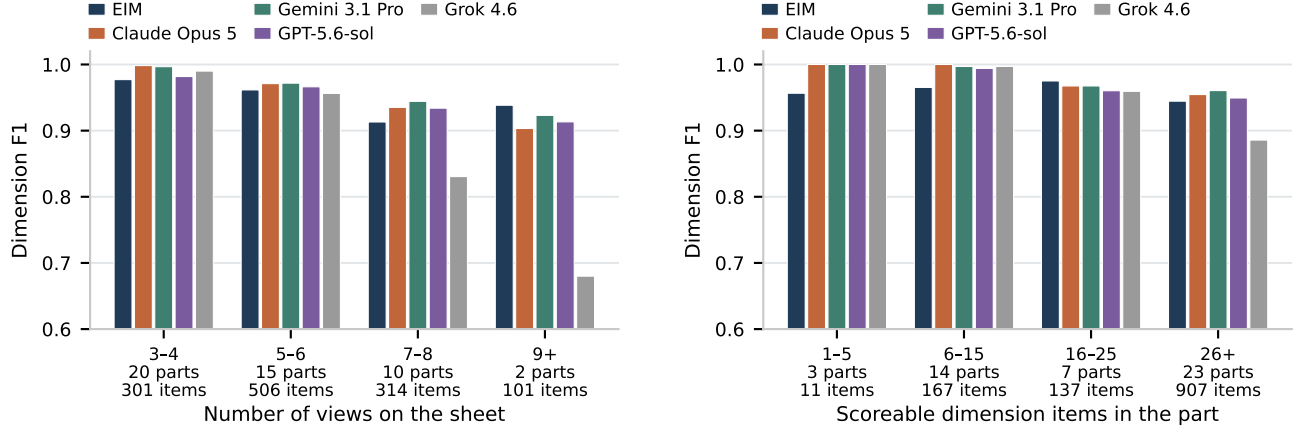


Figure 2. Dimension F1 by drawing property. **Left:** number of views on the sheet, as recorded by the ground truth. **Right:** number of scoreable dimension items in the part, which is the size of the frozen key and therefore identical across systems. Bars are pooled F1 within the bucket; the axis labels give the number of drawings and the number of key items behind each group. The vertical axis is truncated at 0.6.

Quantity	Unit	Parts with GT	Median [%] error (lower is better)					Accepted within 1% (higher is better)				
			EIM	Opus 5	Gemini	GPT-5.6	Grok	EIM	Opus 5	Gemini	GPT-5.6	Grok
Surface area	mm ²	39	7.6	11.4	7.9	7.7	12.1	17.9	12.8	17.9	25.6	23.1
Volume	mm ³	39	8.4	7.6	17.0	5.8	2.9	28.2	35.9	25.6	30.8	35.9
Mass	g	39	10.1	8.2	17.0	9.2	5.2	28.2	33.3	23.1	30.8	33.3
Bounding box L	mm	42	0.0	0.9	10.6	0.9	2.4	57.1	54.8	50.0	59.5	47.6
Bounding box W	mm	42	0.3	3.9	7.0	3.6	5.5	57.1	50.0	52.4	59.5	52.4
Bounding box H	mm	42	0.0	6.7	18.7	4.8	7.2	61.9	59.5	59.5	61.9	57.1
I_1	kg·mm ²	39	17.6	19.5	47.7	22.0	27.5	10.3	15.4	10.3	17.9	15.4
I_2	kg·mm ²	39	24.1	20.0	75.3	27.2	26.3	10.3	12.8	5.1	7.7	10.3
I_3	kg·mm ²	39	20.8	17.2	90.0	19.1	23.4	10.3	12.8	10.3	10.3	12.8
Radius of gyration k_1	mm	39	3.9	5.4	12.6	5.9	6.0	20.5	23.1	12.8	30.8	35.9
Radius of gyration k_2	mm	39	5.1	7.6	6.9	5.9	7.7	12.8	15.4	12.8	23.1	23.1
Radius of gyration k_3	mm	39	5.4	8.1	6.7	5.6	15.1	17.9	20.5	12.8	17.9	17.9
All quantities		477	6.6	8.7	13.9	6.6	11.0	28.3	29.4	24.9	31.9	30.8

Table 8. Volumetric accuracy. *Left block:* median absolute percentage difference from the reference solid, taken directly from the run with no judge adjudication. The median is used rather than the mean because a g·mm² / kg·mm² mislabelling in a few cells moves a mean by orders of magnitude; see text. *Right block:* percentage of drawings on which the judge accepted the value within one percent. “Parts with GT” is the number of drawings carrying a reference value for that quantity, so the denominator differs by row; eight drawings carry no mass-derived reference and five no bounding box, giving 477 scored values per system. Best per column-block in bold.

- ing drawings using vision language models. *arXiv preprint arXiv:2510.21862*, 2025. 2
- [4] Muhammad Tayyab Khan et al. From drawings to decisions: A hybrid vision-language framework for parsing 2D engineering drawings into structured manufacturing knowledge. *Robotics and Computer-Integrated Manufacturing*, 99: 103186, 2026. 1, 2
- [5] Muhammad Tayyab Khan et al. Fine-tuning vision-language model for automated engineering drawing analysis. In *Proceedings of the 9th International Conference on Innovation in Artificial Intelligence (ICIAI)*. Springer Nature, 2026. 2
- [6] Aleksei Kondratenko et al. AECV-Bench: Benchmarking multimodal models on architectural and engineering drawings understanding. *arXiv preprint arXiv:2601.04819*, 2026. 2
- [7] Qian Kou et al. MechVQA: Benchmarking and enhancing multimodal LLMs on comprehensive mechanical drawing understanding. *arXiv preprint arXiv:2605.30794*, 2026. 1, 2
- [8] Vasil Shterianov et al. BlueprintSymVL: A discriminative benchmark for VLM symbol recognition in engineering blueprints. *Results in Engineering*, page 108171, 2025. 2