

Article

Prompt Complexity, Overprompting, and the Honesty of Error in LLM-Based Mechanical Drawing Interpretation: A Structured Empirical Evaluation

Petros Pistofidis , Anastasia Petrou and Pantelis N. Botsaris 

Department of Production and Management Engineering, School of Engineering, Democritus University of Thrace, 67100 Xanthi, Greece; anaspetr9@gmail.com (A.P.); panmpots@pme.duth.gr (P.N.B.)

* Correspondence: pistofid@pme.duth.gr

Abstract

Large language models (LLMs) are increasingly used to interpret mechanical engineering drawings, yet the main risk is a wrong answer that reads as correct. We report an exploratory empirical study of how prompt design and evaluation shape the reliability of this interpretation, in two connected stages over the same twenty drawings. Stage 1 runs one prose prompting protocol on two model families and finds that progressive enrichment does not improve accuracy; beyond a threshold, it raises confident fabrication. Stage 2 compares six structured prompt architectures at thirty repetitions per cell, producing 3540 generations scored under four judge conditions. Across both stages, identification accuracy does not improve, while the rate of claims the drawing does not support varies by nearly an order of magnitude with prompt architecture. Only a judge that sees the drawing can measure it; a text-only judge inflates confident fabrication three to four times, making evaluator modality a first-order variable. Without a ground-truth answer, an image-only check can test only whether claims are visible; it recovers about half of the reference flags, enough for human-in-the-loop triage. A judge from another model family reproduces the ranking. We distill five practical recommendations for this setting.

Keywords: large language models; mechanical engineering drawings; prompt engineering; overprompting; hallucination; vision-language models; LLM-as-a-judge; honesty of error

1. Introduction

Multimodal large language models (LLMs) are entering engineering documentation workflows: digitizing legacy drawing archives, supporting technical training, drafting inspection and maintenance records, and extracting structured data from schematics on demand. Mechanical engineering drawings sit at the center of these workflows. They are dense, convention-governed artifacts, built on orthographic projection, standardized symbols, and referenced norms. They carry the design intent that downstream manufacturing, procurement, and maintenance decisions depend on. A model that reads such drawings reliably would remove a real bottleneck in engineering practice. A model that reads them unreliably, and does so fluently, creates a new and less visible one.

The hazard, we argue, is not the wrong answer as such, but the confidently wrong answer. An interpretation that says a component might be some form of coupling, though the view is ambiguous, fails loudly. As such, a reader can question it, request another view, or escalate. An interpretation that names the wrong component and supports it with an



Academic Editor: Riccardo Caponetto

Received: 30 July 2026

Revised: 7 October 2026

Accepted: 9 October 2026

Published: 14 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

invented standard and exact dimensions fails silently. It carries the surface markers of expertise, so a reader is likely to trust it, and the error propagates into whatever depends on it. Research on hallucination in vision-language systems has mostly measured how often models are wrong. In an engineering deployment, what matters at least as much is how they are wrong: whether their failures announce themselves or impersonate correct answers. Two systems with the same error rate can sit at opposite ends of this reliability spectrum, and standard accuracy metrics cannot tell them apart.

Two kinds of remedy are available. One works on the model itself: a higher-resolution visual backbone, fine-tuning on drafting convention rather than on general imagery, or a parser that turns the drawing into explicit geometry before a language model reads it. Each buys grounding at the price of training or a parsing stack. None is available to a team that consumes a hosted model through an interface. The other remedy works on everything that surrounds the model: how the request is phrased, what shape the answer is allowed to take and what the evaluator is allowed to see. That one is immediate, and it is the one most teams actually hold.

Prompt engineering is the lever practitioners reach for first, and for visual interpretation in specialized domains its guidance is thin and partly contradictory. Practitioner intuition says that more context, more role framing, and more instruction should help, yet there is growing evidence that beyond some threshold additional prompt scaffolding degrades output quality, a phenomenon described as overprompting. At the same time, the vision-language literature offers several structural interventions, each individually promising but untested in this domain: schema-constrained extraction [1,2], staged pipelines whose known failure is the describe-then-generate bottleneck [3], image-grounded verification, correction and visual grounding [4,5], visual chain-of-thought designs [6–8], and few-shot visual exemplars [9]. What none of these strands establishes is which structural choices govern the safety of errors, rather than the accuracy of answers, when an LLM interprets mechanical drawings.

There is a second gap and it concerns measurement itself. Studies that use an LLM as an automated judge (for validation and review) worry chiefly about the judge's identity, in particular whether a model favors outputs of its own family [10]. Far less attention has been paid to what the judge is allowed to see. Whether it sees the drawing or only text about it and whether it also sees the reference answer or must work without one are design variables. They bear directly on what an evaluation can detect, especially for fabrication, which is by definition a claim unsupported by the visual evidence. The deployment version of this question is pressing: a fielded interpretation system has no answer key, so if reliability cannot be monitored without one, it cannot be monitored at all.

This article addresses both gaps with a single exploratory empirical study of LLM interpretation of mechanical engineering drawings. The investigation is organized in two connected stages over a fixed set of twenty drawings. Stage 1 is a prose-prompting baseline. It comprises a pilot evaluation, that crossed three levels of prompt complexity with three levels of information depth on two OpenAI models, and a cross-model validation that repeats the same protocol on a current model (Claude Opus 4.7) at two reasoning-effort levels. Stage 2 is a structured-prompting program motivated by what Stage 1 reveals. It compares six structured architectures at thirty repetitions per cell, producing 3540 generations, and evaluates them under four judge conditions that vary what the evaluator sees. These are a multimodal reference judge, a text-only ablation, an image-only condition with no ground truth, and a cross-family judge from a different model vendor (Google). In total, the study rests on close to twelve thousand scored judgments.

The study addresses five research questions:

- RQ1. Does progressively enriching a prose prompt with more structure and more context improve interpretation, either in accuracy or in the honesty of its errors? This tests the common assumption that more prompt scaffolding produces a better reading of the drawing.
- RQ2. Do structured prompt architectures differ in the safety of their errors rather than in their accuracy? Two architectures can identify components equally well and still differ in how often a wrong answer asserts invented specifics.
- RQ3. Which properties of the evaluator (judge) determine whether that difference can be measured at all? An evaluator that cannot see the drawing may be unable to separate an honest wrong answer from a fabricated one.
- RQ4. Can confident fabrication be detected at deployment time, when no reference answer exists? A deployed system holds the drawing but no answer key, so this decides whether reliability can be monitored in practice.
- RQ5. Are the findings robust to the model family of the judge? This tests whether the results are an artifact of a model scoring outputs from its own family.

The remainder of the article is organized as follows. Section 2 reviews the related work on prompt design, hallucination in vision-language models, and LLM-based evaluation. Section 3 sets out the two-stage methodology, the materials, the error taxonomy, and the four judge conditions. Section 4 reports the results in the order the pipeline produced them. Section 5 discusses what the findings mean for prompt design, for evaluation methodology, and for deployment, and states the limitations. Section 6 concludes.

2. Background and Related Work

This section reviews the three strands the study builds on: prompt design and overprompting (Section 2.1), hallucination in multimodal models and its measurement (Section 2.2), and machine interpretation of engineering drawings (Section 2.3). Each strand closes on what it leaves unmeasured, and Section 2.3 ends with the gaps that motivate the research questions.

2.1. Prompt Engineering, Complexity, and Overprompting

Prompt engineering is now a documented discipline, with a toolkit of clear instruction, role framing, delimiters, worked examples, and reasoning scaffolds [11]. Chain-of-thought exemplars more than double arithmetic solve rates at sufficient model scale [12], and the effect extends to the zero-shot setting [13]. The same seminal work already touches on this article's theme: some correct answers rode on flawed reasoning chains, so answer-level accuracy has masked reasoning-level reliability from the start.

Later work shows that prompt effects are large, unstable, and model-dependent. Equivalent paraphrases of one query swing a model's win rate by up to 45 points, and the worst prompt differs from model to model [14]. An identical prompt can give inconsistent answers across sessions [15]. With multimodal input, merely reordering image and text shifts accuracy, with no optimum common to all models [16]. Structured output formats lift weak models by over twenty points and leave mature ones almost unchanged [17]. Across 462 controlled experiments, phrasing changes the size of accuracy but not which architecture wins [18].

The most consequential pattern is overprompting: enrichment beyond a point stops helping and begins to hurt. Accuracy rises with prompt specificity, plateaus, and dips at maximal detail [19], and more few-shot examples help up to a small optimum and then degrade accuracy across seven models [20]. No configuration is best everywhere. Single-task and multitask prompts trade places across model families [21], role enrichment

leaves some error types unsolved under every strategy [22], and prompts that quadruple document accuracy still leave confident errors in the highest confidence band [23].

Where added structure constrains rather than informs, it harms: chain-of-thought scaffolding hurts entity recognition [24], stacked prompt classes degrade conceptual design [25], and closed instruction layers reshape rather than improve output [26]. Across 162 personas and nine models, an expert persona brings no significant gain [27]. This literature measures, almost without exception, accuracy. Whether the same manipulations change the character of the failures is rarely asked.

2.2. Hallucination and Its Measurement in Multimodal Models

That question is the subject of the second strand. Hallucination is provably inevitable for any computable language model, so prompting can relocate it but not remove it, and safety-critical use needs external guardrails [28]. It has two sources, the prompt and the model: vague prompts nearly double the rate of chain-of-thought prompts, some models stay prone under every prompt, and structured reasoning can produce a longer wrong answer instead of an admission of uncertainty [29]. The visual side shows the same pattern. A multi-step protocol raises medical visual question answering by up to nineteen points while its reasoning traces justify wrong answers in fluent, terminologically correct language [30]. Multi-agent visual reasoning still emits well-formatted answers that silently omit required elements [31], confident mislabeling worsens with scene complexity [32], and an unconstrained prompt drives fabrication toward its ceiling [33]. Heavier scaffolding changes how a model fails, and its confident failures are the ones that survive.

A parallel literature proposes structural interventions, and they define the design space tested here. Pipelines that describe first and generate later pass invented detail downstream when the later stage loses the image [3]. Image-grounded verification and visual grounding re-attach claims to visual evidence [4,5]. Visual chain of thought imposes observation before classification [6–8], schema constraints bound where a model may elaborate [1,2], and few-shot exemplars import task structure by demonstration [9]. Each has been validated alone, on different tasks and metrics. None has been compared with the others on the same materials, or with an honest error separated from a confident fabrication as the outcome.

Whether that separation can be made depends on the evaluation, which in this setting is itself performed by a language model: a second model, the judge, scores each interpretation. The known risks concern who the judge is and how it is prompted. Models amplify their own biases when they judge their own outputs [10], and the wording of the judge's prompt shifts its agreement with human ground truth [34]. Judges from two families can disagree almost completely, with biases that reverse sign across metrics [35], and language-model raters are reliable for ranking outputs but not for absolute judgments [36]. The scoring metric alone can move measured correctness by tens of points [37]; one response is to make perception explicit, so that an error can be traced to vision or to language [38]. Re-running one model on a hazard analysis over a week gives stable surface success but volatile analytical quality [39]. None of this work varies what the judge can see: the drawing or only the text, the reference answer or none. For fabrication, a claim the drawing does not support, these choices decide whether the error can be detected at all, and whether reliability can be monitored at deployment, where no reference answer exists.

2.3. Interpreting Engineering Drawings with Vision-Language Models

Machine reading of technical drawings predates language models. Detection-based systems remain standard for symbols, reaching 95% accuracy across sixteen classes on single-line electrical diagrams [40], and an optical-character pipeline extracts dimensions

and tolerances from production drawings at about 90% precision [41]. Neuro-symbolic designs make every answer inspectable through a scene graph [42]. A recent survey finds vision-language approaches strongest overall, names hallucination as a central unresolved challenge, and does not cover mechanical engineering drawings [43].

Language models have since been applied across the domain, usually one intervention at a time: prompt engineering for phase diagrams [44] and hydrological visual tasks [45], and question answering over piping and instrumentation diagrams, where accuracy climbs as prompt context is enriched [46]. A survey of large models for computer-aided design flags inconsistent evaluation, with hallucination tracked unevenly [47].

The closest work addresses engineering drawings directly and already gestures at our theme. On manufacturing feature recognition, the accuracy leader and the hallucination leader are different models [48]. On tolerancing extraction from real drawings, frontier models scored zero-shot hallucinate at 41% and 56%, more as drawing density grows [49]. For single components, text-only input outperforms multimodal input on average in computer-aided design generation [50], and a retrieval system for mechanical components swings from 38% to 90% accuracy with configuration and prompting alone [51]. A broad evaluation across engineering design documents patterned failures and argues that a useful model should signal uncertainty rather than, in the authors' words, make up a response that sounds correct [52]. These studies vary the model or the task and score hallucination as an aggregate rate. None takes the kind of individual errors, hedged or fabricated, as the primary outcome at the level of prompt architecture.

Three gaps follow. First, prompt-design research measures accuracy and hallucination research measures error rates; whether a wrong answer asserts unsupported detail is an outcome in neither, although confident wrong answers appear in both. Second, the structural interventions of Section 2.2 have never been compared under one protocol with fabrication as the outcome, and mechanical engineering drawings have hosted no such comparison. Third, the evaluation literature asks who judges and with what wording, but not what the judge can see, so the roles of judge modality and ground-truth access, and monitoring without a reference, remain untested.

3. Materials and Methods

This section sets out the study's methodology. The task under study is specific: the interpretation of mechanical engineering drawings by a large language model. Throughout, the pipeline has two operations: (i) a subject model first interprets each drawing and names the component it sees, and (ii) an independent evaluation step, which we call the judge, then scores that interpretation, both for correctness and for fabrication. The study's concern is reliability: not the accuracy of the answer alone, but whether a wrong answer stays open to correction or asserts details that the drawing does not support. We frame the work as an exploratory study in two connected stages that share this pipeline. Stage 1 is a prose-prompting baseline that asks whether progressively enriching a prompt improves interpretation. Stage 2 moves to structured prompting and pursues reliability along the two operations of the pipeline: a generation axis that compares six architectures, J1 to J6, to isolate the levers that govern fabrication and an evaluation axis that interrogates the judge itself.

Both stages interpret the same 20 drawings. Stage 1 uses three subject models from two families, and Stage 2 holds one of them, Claude Opus 4.7, fixed. They also share one error taxonomy, so that findings stay comparable across the progression. The design maps onto the five research questions of Section 1. Stage 1 answers RQ1. Within Stage 2, the generation axis answers RQ2, while the evaluation axis answers the remaining three: judge modality for RQ3, ground-truth access for RQ4, and judge model family for RQ5. Figure 1

summarizes the full pipeline. The prompts of both stages and of every judge, the human validation, and the inferential statistics are given in full in the Supplementary Materials.

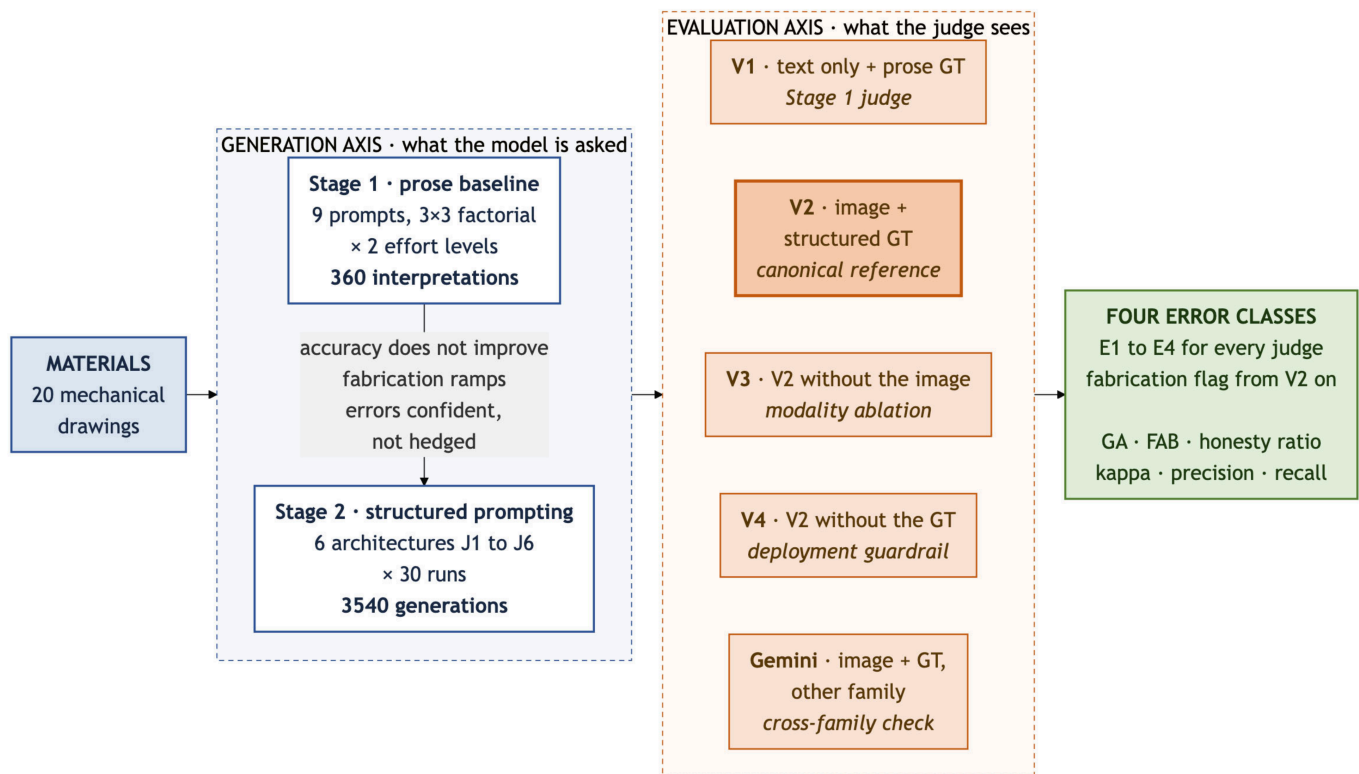


Figure 1. Overview of the two-stage methodology: the generation axis (Stage 1 prose and Stage 2 structured architectures) and the evaluation axis (five judge conditions). The canonical reference judge V2 is highlighted.

3.1. Materials: The Drawing Set

The interpretation task uses a fixed set of 20 mechanical component drawings. They are drawn from Narayana, Kannaiah, and Venkata Reddy's *Machine Drawing* [53], a standard undergraduate engineering reference. Selection was purposive, and it followed four criteria. First, the set gives graduated coverage across component families of increasing design complexity. Second, it favors drawings with two or more orthographic views, so that correct interpretation needs genuine visual observation rather than textual inference. Third, it is limited to single parts or simple assemblies of no more than four components. This keeps prompt effects separate from raw task difficulty. Fourth, it favors industrially common components, which are more likely to appear in model training data. This has a cost: a model may recognize a familiar component rather than read the drawing in front of it.

The set spans three component families of increasing design complexity, and the three are the source reference's own. They are Chapters 5, 6, and 7 of that reference, presented there in that order. Threaded fasteners account for 11 drawings, namely bolts, nuts, screws, and studs (Class I). Keys, wedges, and pin joints account for 5 drawings (Class II). Shaft couplings account for 4 drawings (Class III), and these are assemblies of several components with complex spatial relationships. We placed each family in one of the three classes by weighing five properties: how strongly the design parameters are coupled to one another, the number of parts in the assembly, the number of functions and connections the component carries, its symmetry, and how readily it can be analyzed and manufactured. The classes are ordinal. They rank the three families and do not measure the distance between them. The scale describes the engineering design of the components, not the difficulty a vision-language model has in reading their drawings.

The same 20 drawings are interpreted in both stages, so the materials act as a controlled constant across the methodology (Table 1). The few-shot architecture in Stage 2 reserves two drawings as exemplars, so it is evaluated on 18 (see Section 3.5.2).

Table 1. The 20 mechanical component drawings by family. Drawing identifiers follow the numbering of the source reference [53]. Class I is the least complex.

Component Family	Drawings	N	Design Complexity Class
Threaded fasteners	5.11, 5.15a, 5.15b, 5.18, 5.19, 5.20, 5.21, 5.22, 5.22b, 5.28, 5.29	11	I
Keys, wedges, and pin joints	6.3, 6.4, 6.5, 6.13, 6.15	5	II
Shaft couplings	7.3, 7.4, 7.8, 7.11	4	III

3.2. Error Taxonomy: The E1 to E4 Scheme

Both stages report every response in the same four error classes. Stage 2 derives them from the three axes set out below. The Stage 1 judge assigns them with the earlier rubric of the pilot, so every comparison across the two stages is made with the Stage 2 judge on both sides. The taxonomy uses three independent axes. A fixed rule then maps them to one of four error classes. The first axis is Category (binary). It asks whether the response names the correct broad component family, such as a bolt, a key, or a shaft coupling. The second axis is Subtype (ternary). It asks how well the response identifies the specific subtype, at the level of full, partial, or wrong. The third axis is Fabrication (binary). It asks whether the response makes specific factual claims that the available evidence does not support. Such claims include named standards (for example DIN or ISO), exact dimensions and named geometric features. Hedged statements such as “appears to be” do not count as fabrication. General engineering vocabulary, used without unsupported specifics, does not count either. A deterministic rule maps the three axes to four classes, with no borderline adjudication (Table 2).

Table 2. Deterministic derivation of the E-class from the three axes. The four rows are mutually exclusive and exhaustive, so every response maps to exactly one class.

Category	Subtype	Fabrication	E-Class
right	full	any	E1
right	partial or wrong	any	E2
wrong	any	none	E3
wrong	any	present	E4

The first two axes set the identification quality. E1 is a full identification, while E2 is a partial one, where the category is right but the subtype is not. E2 therefore covers two cases, a partial subtype and a wrong subtype within the right category. They are pooled to keep one scale across both stages and are also reported separately. The fabrication axis matters only when the category is wrong. There it separates two very different failures. E3, which we call an honest error, is a wrong identification that asserts no unsupported specific detail, either because it stays general or because it hedges. E4, which we call a confident fabrication, is a wrong identification that states at least one unsupported specific without hedging.

The rule has one further consequence. The fabrication axis is recorded for every response, including correct ones. When identification is correct, fabricated secondary detail does not downgrade E1 to E2. It is instead tracked as a separate flag and reported as its own rate. That rate is the broader safety measure. E4 is its most severe case, not its only harmful

one: a correct identification that carries an invented standard or dimension is also a risk, and the fabrication rate counts it. What counts as adequate support for the fabrication axis depends on what the judge can see. Each stage therefore states its own grounding for this axis (Sections 3.4.3 and 3.5.3).

The labels of this scheme describe the text of a response, not the state of the model that wrote it. Each is operational. An honest error is a wrong answer without an unhedged, unsupported claim. A confident fabrication is a wrong answer with at least one. A safer architecture is one with a lower fabrication rate. None of these measures what a model believes or how certain it is internally. The study observes only what the model writes, and every statement about honesty or confidence in this article is a statement about that text. The fabrication axis also scores two properties jointly: a claim counts only if it is both unsupported and unhedged. A hedged claim that the drawing does not support is therefore not counted, and neither is a categorical claim that the drawing does support. We score the conjunction because it is the case that misleads a reader, but the two properties are not measured separately.

3.3. Metrics and Statistical Analysis

Both stages are summarized with the same four interpretation metrics, all computed from the error classes of Section 3.2, with their formulas collected in Table 3. General accuracy is the share of responses that name the correct component category, whether fully (E1) or partially (E2); it is the standard identification-quality measure. The fabrication flag rate is the share of responses that assert a specific claim the evidence does not support; it is tracked even when the identification is correct, which makes it the study's primary safety signal. Confident fabrication (E4) is the share of responses that are both wrong-category and fabricated, the most severe class: a wrong identification asserted with invented specifics. The honesty ratio is, among wrong-category answers, the share that stay honest and hedged (E3) rather than fabricated (E4). A higher ratio means that more of the wrong answers carry no unsupported claim. Because it conditions on wrong answers, it is independent of raw accuracy.

Table 3. The interpretation metrics and their formulas. N is the number of scored responses in the cell.

Metric	Formula
General accuracy (GA)	$(E1 + E2)/N$
Fabrication rate (FAB)	$\text{flagged}/N$
Confident fabrication (E4)	$E4/N$
Honesty ratio	$E3/(E3 + E4)$

Stage 2 adds one family of measures for comparing judges. Because its four judge conditions score the same generations, one judge serves as the reference and another is scored on how well it recovers the reference's fabrication verdicts, using Cohen's kappa together with precision and recall.

Comparisons respect the structure of the design. Stage 1 is a single run of twenty drawings per cell, where one item moves a cell by five points, so its tests pool cells before comparing. The pilot half uses exact Fisher tests and Cochran–Armitage trend tests. The second half retains per-item labels, so it uses Cochran's Q across configurations and exact McNemar tests between judges. In Stage 2, the same drawings pass through every architecture and the thirty runs of a cell are repeated observations of one drawing under one prompt rather than thirty independent trials. Runs are therefore averaged to one value per drawing and architecture before any test. The architectures are then compared with the Friedman test, which treats the drawings as blocks. Individual pairs are compared

only where a specific contrast is claimed, by a paired permutation test on drawing-level differences, with Holm correction across the fifteen pairs. Every interval is a drawing-level bootstrap over 10,000 resamples, so the size of the drawing set is carried into every interval. Because the absence of a significant difference is not evidence of equivalence, accuracy is additionally assessed by two one-sided tests at a margin of five points. That margin was fixed in advance, on the grounds that a smaller difference would not change which architecture a practitioner deploys.

3.4. Stage 1: Prose Baseline and the Overprompting Ceiling

Stage 1 is the prose baseline of the investigation, and it asks whether progressively enriching a prose prompt improves the interpretation of a drawing.

3.4.1. Models and Generation Steps

Stage 1 applies one prose-prompting protocol to two model families, run as two matched halves (Table 4). The first half is the pilot: it interprets the 20 drawings with two OpenAI models, GPT-4o (2024) as a direct-inference model and o3-mini (2025) as a reasoning-augmented model that favors stepwise deliberation. The second half is a cross-model validation. It repeats the identical protocol with a current model, Claude Opus 4.7 (2026), at two effort levels: low as a direct-inference analogue and xhigh as a reasoning-augmented one. Each half therefore pairs a direct-inference subject with a reasoning-augmented one and produces 360 interpretations (9 prompts $\times$ 20 drawings $\times$ 2 effort levels). Each half is scored by its own text-only judge. The second half changes model family, model generation, and reasoning mode at once. It can therefore show whether the prompt effects of the first half recur, but not which of those changes causes a difference between the halves. Prompt effects are estimated within each half, and contrasts between the halves are descriptive.

Table 4. The two matched halves of Stage 1.

Half	Subject Models	V1 Judge (Text Only)
Pilot	GPT-4o (direct-inference), o3-mini (reasoning-augmented)	o4-mini
Cross-model validation	Claude Opus 4.7 at low (direct-inference) and xhigh (reasoning-augmented)	Opus 4.7 (medium)

3.4.2. Prompt Design and the Two Axes Strategy

The prompts form a 3×3 factorial over two axes and the two axes deliberately enrich different things. Structural complexity governs how the model is told to behave: how much role framing, governance instruction, and organization the prompt carries, without adding any fact about the drawing. Information depth governs what the model is told about the task: how much of the answer space the prompt narrows before the model looks. Separating them matters, because a prompt can grow longer in either direction, and the two directions are not the same intervention. Table 5 sets out the escalation along each axis. Crossing the two axes gives nine configurations, from the minimal A to the maximally scaffolded C2. This grid lets us ask whether more scaffolding and more context improve interpretation, and where the overprompting threshold lies.

3.4.3. Evaluation with Two Judge Conditions

Evaluation in Stage 1 compares each interpretation against a reference description of the component, the study's first ground truth. For every drawing, an expert in mechanical

engineering and design wrote a precise prose description stating the component's type, category, and functional role. Scoring maps the comparison of the model's answer with this reference onto the three axes and four classes of Section 3.2.

Table 5. The two prompt axes of Stage 1 and the escalation along each. Crossing them gives the nine configurations A0 to C2.

Axis	Level	What the Prompt Adds	What the Level Is Designed to Test
Structural complexity	A	A minimal, direct instruction. Nothing else.	The baseline: what the model does when it is simply asked.
	B	A full professional role and an explicit instruction to report only what is observed.	Whether persona framing and an observational guardrail change the reading.
	C	An organized structure of Role, Context, Task, and Guidelines.	Whether a fully sectioned, scaffolded prompt beats plain instruction.
Information depth	0	Nothing about the drawing.	The baseline: the answer space stays fully open.
	1	The statement that the views belong to a single component.	Whether one grounding fact about the materials helps, while leaving the answer space open.
	2	A constraint to answer with one of three given component categories.	Whether closing the answer space helps or forces a commitment.

Each half is scored by its own text-only judge V1, which never sees the drawing: it receives the response and the prose ground truth (o4-mini for the pilot, Opus 4.7 for the second half). V1 keeps the rubric of the pilot. It assigns the same four classes from two criteria rather than three axes: whether the answer identifies the component, and whether its terminology is correct. Its E4 is a wrong identification whose terminology looks correct and so misleads the reader. Its E3 is a wrong identification that the terminology does not disguise. V1 therefore has no fabrication axis. Blind to the image, it can judge whether the wording fits the prose ground truth, never whether the drawing supports a claim. The three-axis rubric begins with V2. The Opus interpretations are then re-scored with the multimodal Stage 2 judge V2, which both grounds the fabrication axis in the drawing and places both stages on the same E1 to E4 scale, so the move from prose to structured prompting can be compared cleanly. V1 and the Stage 2 judges are set out together in Section 3.5.3.

3.5. Stage 2: Structured Architectures and Multimodal Evaluation

Stage 2 is the structured program that Stage 1 motivates. The model no longer answers in free prose, but fills a structured JSON schema, and its output is scored against a structured ground truth of the same form. The question is which structural choices govern the honesty of the errors rather than the accuracy of the answers, with the practical aim of minimizing confident fabrication at no cost to accuracy.

3.5.1. Model and Generation Setup

Stage 2 keeps the subject model fixed at Claude Opus 4.7 and varies the prompt architecture instead. Each architecture interprets the 20 drawings. We repeat every drawing 30 times per architecture and this produces 3540 generations (6 architectures $\times$ 20 drawings

× 30 runs, with 18 drawings for the exemplar architecture J6). The repetition is the main change from Stage 1, which used one run per cell. It lets us estimate distributions and test differences between architectures.

3.5.2. Prompt Design and the Six Architectures

The six architectures, J1 to J6, are not six arbitrary prompts. Each isolates one of three design levers: where the image is available in the pipeline, how tightly the output schema is constrained, and whether the prompt supplies worked examples. With the subject model and the drawings held fixed, changing one lever at a time lets us attribute any difference in the safety of the errors to that lever rather than to the model.

The architectures fall into three families, one per lever (Table 6). The schema family, J1 and J2, isolates the schema: both keep the image present in a single inference call and differ only in the JSON they must fill. J1 uses broad, loosely anchored fields, so the model can report its full interpretation and elaborate freely. J2 mirrors the structured ground truth exactly, a tighter contract that bounds where the model may add detail. Comparing the two isolates the effect of constraining the output form while leaving the identification itself free. The reasoning family, J3, J4, and J5, isolates where the image travels through a multi-step pipeline. J3 is a two-stage pipeline: it describes the drawing in free prose in a first call, then fills the schema in a second call that no longer sees the image, so any detail invented in the description is carried forward as fact. J4 keeps the same two-stage shape but retains the image in the second call and verifies each claim against it, closing that gap. J5 is a single-stage visual chain of thought: it records named observations first and then classifies, with the image present throughout, so no claim is made without the drawing in view. The exemplar family, J6, adds two worked examples to a single call, isolating whether demonstration alone improves the reading (few-shot approach). Each family thus varies one lever while holding the others fixed: image placement across J3 to J5, schema tightness between J1 and J2, and the presence of exemplars in J6 (Table 6).

Table 6. The six structured prompt architectures of Stage 2.

Arch.	Family	Design Orientation	Calls	Image Access	Literature Basis
J1	schema	one call, broad anchored fields	1	yes	schema-constrained extraction [1,2]
J2	schema	one call, mirrors the ground truth schema	1	yes	schema-constrained extraction [2]
J3	reasoning	two calls, describe then generate, blind second call	2	first call only	describe-then-generate bottleneck [3]
J4	reasoning	two calls, image kept and claims verified	2	both calls	evidence anchoring and post-hoc correction [4,5]
J5	reasoning	one call, observe then classify	1	yes	visual chain of thought [6–8]
J6	exemplar	one call with two worked examples	1	yes	few-shot visual exemplars [9]

3.5.3. Evaluation with Four Judge Conditions

For Stage 2, we developed a new, structured ground truth with eight fields, and Table 7 lists them with a worked example. This schema separates what the component is, recorded in the category, subtype, and function fields, from what is merely visible on the drawing, recorded in the geometric features and views fields. That separation is what lets a judge tell

a wrong identification from a fabricated detail. We built each entry from the drawing and from standard mechanical design references. The same expert who wrote the prose ground truth designed and produced the structured ground truth against the drawing, delivering the version that was used for scoring.

Table 7. The eight fields of the structured ground truth, with a worked example for drawing 7.3, a split muff coupling.

Field	Value for Drawing 7.3
Component category	shaft_coupling
Component subtype	split_muff_coupling
Primary function	Rigidly couples two shafts via a split muff clamped by bolts and secured with a sunk key.
Key geometric features	split_muff, four_through_bolts, nuts, through_holes, sunk_key, section_partial
Views shown	isometric, front, side_left, section_partial
Standards referenced	(none)
Assembly type	multi_part_assembly
Prose description	Split muff coupling with split coupling sleeve, which is tightened with four through bolts with nuts in through holes and secured with a sunk key. The coupling is shown in isometric view with partial section, front view with partial section, and side view in section.

Table 8 sets out the four Stage 2 judge conditions. The judges differ in what evidence they see, which changes how the fabrication axis is grounded. V2, V3, and the cross-family judge apply the same three-axis rubric and the same automated validator. V4 applies only the fabrication axis, because without a reference answer the identification axes cannot be scored at all. In greater detail, the canonical judge V2 holds the image, the ground truth, and the response and asks whether each claim is visible in the drawing (it is the reference for all Stage 2 results). The other three conditions each change one input to isolate a variable: V3 removes the image (the modality ablation), V4 removes the ground truth (the deployment guardrail, which returns only a grounding verdict and is compared with V2 on the fabrication flag alone) and the cross-family check re-applies the V2 rubric with gemini-2.5-pro on a matched 590-generation subsample. Unlike V2, V4 also counts a functional or categorical claim that contradicts the drawn geometry as ungrounded, since it has no separate identification axis. Its prompt was chosen from three variants on a pilot of 118 of the 3540 generations, and two of its worked examples describe drawings in the set.

3.6. Human-in-the-Loop Validation

Two key roles in this study are performed by people, not models. The first is the construction of the ground truth: for every drawing, an expert in mechanical engineering and design wrote the prose reference of Stage 1 and produced the structured JSON reference used from V2 onward. The reference against which every interpretation is scored is human, not model-generated.

Table 8. The five evaluator conditions used in the study.

Judge	Sees	Fabrication Grounded On	Role	Judge Model
V1	prose ground truth, response	none; terminology checked against the prose ground truth	Stage 1 judge	o4-mini; Opus 4.7, medium
V2	image, ground truth, response	visible in the drawing	canonical reference	Opus 4.7, xhigh
V3	ground truth, response	compatibility with the ground truth class	modality ablation	Opus 4.7, xhigh
V4	image, response	visible in the drawing	deployment guardrail	Opus 4.7, xhigh
Gemini	image, ground truth, response	visible in the drawing	cross-family check	gemini-2.5-pro

The second is a validation of the judge. Using one language model to score another may raise the obvious concern that the scoring may be unreliable. Nevertheless, this tiered approach mirrors how modern AI-augmented systems are deployed in practice. One model and prompting layer performs the primary task, interpretation, advice, or decision support, and a second one validates, verifies, and cross-checks that output to form a guardrail around the service.

To test how far the labels of V1 hold up, two of the authors re-scored samples of the Stage 1 results with the rubric of V1. One sample covered the OpenAI pilot cases ($n = 117$), almost all produced by o3-mini and evaluated by o4-mini. The other covered the Anthropic second half ($n = 120$), produced and evaluated by Opus 4.7. Both samples were built the same way. Cases were selected by the class the judge had given them, a set number from each class and more from the two error classes. Every class therefore carried enough items to be reported on its own, with one exception. Under V1 the second half holds almost no honest errors: the judge recorded a single one across all 360 responses of that half. That one case is in the sample, so the honest error class of the second half is reported as 1 of 1 and carries no rate. In both runs, the expert had the drawing, the reference description and the model's interpretation, and settled on a class before consulting the judge's label. The judge itself had the reference and the interpretation only. No response was rated by both authors, so these rates describe agreement between a human expert and the judge, not between the two humans. The consistency of the human reference itself is therefore not measured.

Agreement on the exact class was 76.1% in the pilot half and 79.2% in the second half, a Cohen's kappa of 0.66 and 0.69 (0.74 and 0.75 with linear weights). It was highest where the judge had scored a response correct (E1, 81.8% and 93.3%) and lowest on the confident fabrication class (E4, 65.0% and 69.5%). The two remaining classes fall between those bounds. Agreement on partial identification (E2) was 77.8% and 83.3%. Agreement on honest error (E3) was 77.2% in the pilot half. The second half holds only the single case noted above. The classes can also be grouped in pairs, E1 with E2 for a component identified and E3 with E4 for one not identified. At that level, the human and the judge agree on 90.6% of items in the pilot half and 87.5% in the second half. By the usual reading of kappa, agreement is substantial overall. It is lowest on E4. Of the responses the judge called E4, the human assigned another class to 7 of 20 in the pilot half and to 18 of 59 in the second half, where 15 of the 18 moved to a correct or partial identification. That is the blind spot of a judge without the drawing, and Stage 2 addresses it by giving its reference judge the drawing. This review covers V1 only. V2 was not re-scored by a person, since full human review of 3540 generations was not feasible. V2 rests instead on two supports: it holds the evidence the fabrication axis is defined over, and a judge from another model family reproduces its architecture ranking.

4. Results

This section reports the results in the order of the pipeline set out in Section 3: the Stage 1 prose baseline, the transition to structured prompting with the generation axis of Stage 2 and the three evaluation-axis studies.

4.1. The Overprompting Threshold

Stage 1 asks whether progressively enriching a prose prompt improves drawing interpretation (RQ1). We report the pilot and its cross-model validation together, all four subject conditions scored by the text-only judge V1 of each half (Figure 2), and close with what the image-grounded re-scoring of the second half adds (Figure 3).

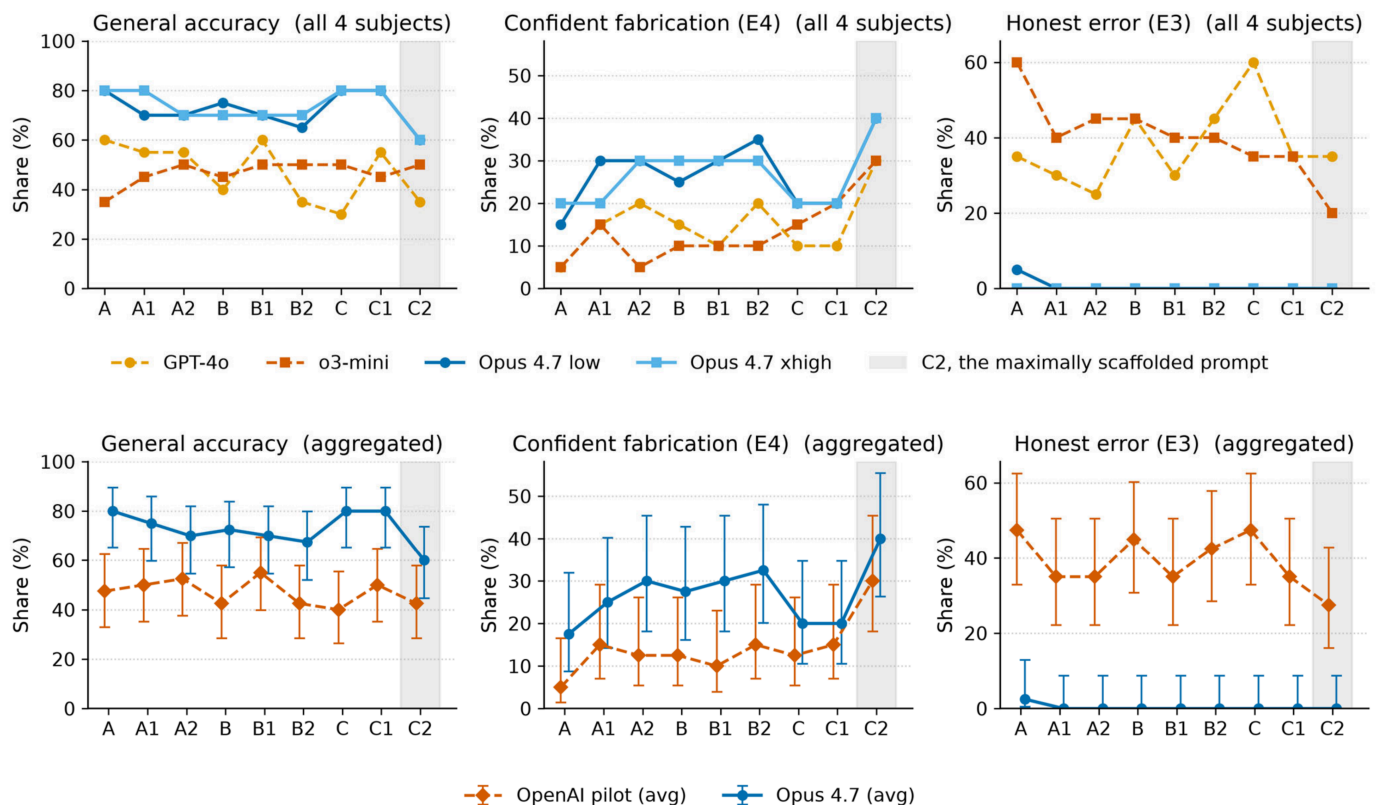


Figure 2. Stage 1 E-class metrics under the text-only judge V1, across the nine prompt configurations for the two OpenAI pilot models and the two Opus 4.7 effort levels. Error bars are 95% Wilson intervals, on the aggregated row only. The per-subject row is a single run of twenty.

The enrichment path applies the two levers that practitioner guidance most often recommends: structural complexity, from the minimal instruction A to the full role, context, task, and instructions template C, and information depth, from no added payload to a rich domain payload. Stage 1 measures what each lever actually buys. The overprompting threshold recurs across model generations. Confident fabrication (E4) is lowest at or near the minimal prompt and reaches its maximum at C2 in all four conditions. The pilot models raise their confident-error rate six-fold between the minimal and the maximally scaffolded cell, from 5% to 30% of responses (Fisher exact $p = 0.006$), and the Opus 4.7 subject peaks at the same cell, at 40%.

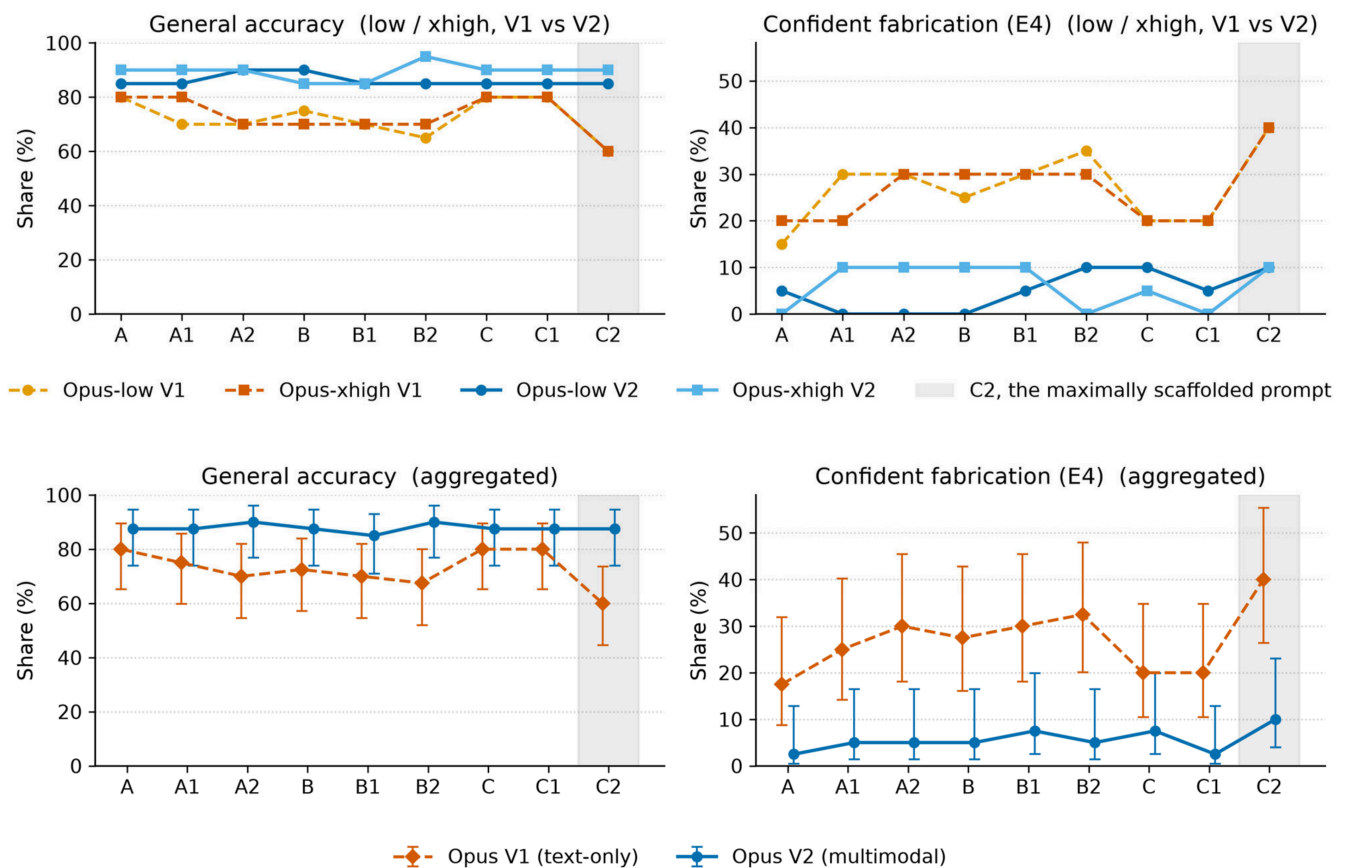


Figure 3. The 360 Opus 4.7 responses scored by the text-only judge V1 and the multimodal judge V2, across the nine prompt configurations. Error bars are 95% Wilson intervals, on the aggregated row only. The per-subject row is a single run of twenty.

Which lever does the damage, however, depends on the model generation. For the 2024 direct-inference pilot model (GPT-4o), structure alone is costly: its accuracy falls steadily as the template grows from A to C, with no payload added, from 60% to 30%. At twenty drawings per cell this is a trend rather than a demonstrated effect ($p = 0.06$). For the 2026 subject of the second half (Opus 4.7), the pattern inverts. The full structured template with no payload behaves like the minimal prompt: at C and C1 the subject matches its own accuracy and confident-error levels at A and A1 ($p = 0.85$ and $p = 1.00$). What degrades the newer model is information depth. As the payload deepens, its accuracy declines and its confident-error rate climbs from about one fifth to about one third of responses (trend $p = 0.04$ for accuracy and $p = 0.03$ for confident error). The collapse at C2 comes from the two levers combined rather than from structure alone. The overprompting threshold therefore does not sit at a fixed place. It differs between the two halves: the newer model tolerates structural scaffolding that visibly harmed its predecessor and remains vulnerable to deep information payloads.

The Opus 4.7 subject is also substantially more accurate. Under V1, it identifies the component correctly in about 73% of responses, against about 47% for the pilot models, a gap of roughly 26 points (Table 9). The two halves are scored by different V1 instruments, o4-mini and Opus 4.7, so the gap is indicative rather than exact. Its size and its consistency across all nine prompts describe the two models as tested.

Table 9. Stage 1 aggregate metrics by subject and judge, under the text-only judge V1 and, for the Opus 4.7 half, the multimodal judge V2.

Subject	Judge	GA (%)	E3 (%)	E4 (%)
GPT-4o (pilot, direct)	V1	47.2	37.8	15.0
o3-mini (pilot, reasoning)	V1	46.7	40.0	13.3
Opus 4.7 low (2nd half, direct)	V1	72.2	0.6	27.2
Opus 4.7 xhigh (2nd half, reasoning)	V1	73.3	0.0	26.7
Opus 4.7 (image-grounded re-score)	V2	87.8	6.7	5.6

The error styles separate the two halves even more sharply than the accuracy levels. Roughly four in ten of the pilot responses are honest errors (E3), wrong answers whose wording does not disguise the error. That share itself responds to prompting. On the minimal prompt, six in ten of the reasoning pilot model's responses are honest errors; by C2 this has fallen to two in ten, replaced by answers stated with confidence, right and wrong alike. The Opus 4.7 subject almost never produces one. Across its 360 responses, V1 records a single honest error, so its responses are either correct or wrong with correct-looking terminology, and the safety question concentrates entirely in E4. Read across the two halves, the space of honest error is far narrower in the second, and prompt enrichment spends what is left of it.

The closing step of Stage 1 re-scores the same 360 Opus 4.7 responses with the multimodal judge V2 and the two judges turn out to disagree in one direction only (Figure 3). Every response V1 accepted as correct, V2 accepts as well. The reverse never occurs; V2 (with the drawing in hand) restores 54 responses that V1 rejected and removes from E4 most of the responses V1 had placed there, while never placing there one that V1 had not. Confident fabrication falls from 26.9% to 5.6% of the same responses, and accuracy rises from 72.8% to 87.8%. The corrections concentrate where enrichment is heaviest: the richer the prompt, the more elaborated the response, and the more often a text-only reading takes elaboration for invention. The image-grounded judge also reveals what enrichment does buy. With accuracy essentially flat across the grid (Cochran $Q(8) = 4.00$, $p = 0.86$), the share of responses that carry unsupported specifics rises steadily with structural complexity, from about a quarter under the minimal template to nearly half under the full one (26.7% to 47.5%, trend $p < 0.001$). Depth pushes the model toward confidently wrong identifications; structure pushes it toward correct identifications wrapped in unsupported detail. Fabrication, not accuracy, is where prompt design acts, and measuring it reliably requires a judge that can look at the drawing. This is why V2 is the reference judge of Stage 2.

Human review of the same responses points the same way, though less far. Two of the authors re-scored stratified samples of the Stage 1 results against the same rubric, 117 of the pilot responses and 120 of the Opus 4.7 ones (Section 3.6). The disagreements are not spread evenly, and where they fall says more than how many there are. The numbers that follow are shares of the judge's own verdicts, not of the disagreements. Take first the responses the judge had scored E1 or E2. In the pilot half, 8 of 40 were revised (20.0%), and in the second half 7 of 60 (11.7%). Most of these were adjustments of degree between the two classes. Only 2 of the 40 in the pilot half were reclassified as incorrect (5.0%), and none of the 60 in the second half. The responses the judge had scored E3 or E4 behave differently. There 20 of 77 were revised in the pilot half (26.0%) and 18 of 60 in the second half (30.0%). Much of that movement crosses the line between the two groups. On review, 9 of the 77 (11.7%) and 15 of the 60 (25.0%) proved to be correct or partly correct identifications after all. A verdict of correct from this judge is therefore rarely overturned, while a verdict of incorrect is the one that human review most often revises. The revision runs in one direction, out of

the error classes and into identification. With honest error all but absent from this half, a failed identification had almost nowhere to go but confident fabrication, and human review indicates that about a quarter of that class was not fabrication at all.

Stage 1 therefore hands three facts to Stage 2. Prompt enrichment does not buy identification accuracy under either judge. It does change the failure profile and for the Opus 4.7 subject it concentrates failure in E4. And that class can only be measured reliably by an image-grounded judge, which Stage 2 adopts as its reference.

4.2. Structured Prompting: Safety by Design

Moving from prose to structure, accuracy on the two sides shows no detectable difference, if anything a shade lower under the rigid schema. What the transition changes is the character of the errors. Confident fabrication is nearly halved, and the overall fabrication flag rate falls by more than half. The errors do not become rarer so much as they change kind: honest error (E3) roughly doubles, and the honesty ratio, the share of wrong answers without an unsupported claim, rises from about half to about four in five (Table 10, Figure 4). The structured forms are wrong about as often, but fewer of their wrong answers carry invented detail. Structuring the output does not raise accuracy; it lowers the fabrication rate.

Table 10. The prose-to-structured transition under the judge V2: prose pools the nine Stage 1 configurations; structured pools J1 to J6 at 30 runs per cell.

Group	N	E1 (%)	E2 (%)	E3 (%)	E4 (%)	GA (%)	FAB (%)	Honesty Ratio (%)
Prose (Stage 1)	360	61.9	25.8	6.7	5.6	87.8	36.4	54.5
Structured (Stage 2)	3540	59.5	24.9	12.4	3.2	84.4	16.1	79.5
Change		−2.4	−0.9	+5.7	−2.4	−3.4	−20.3	+25.0

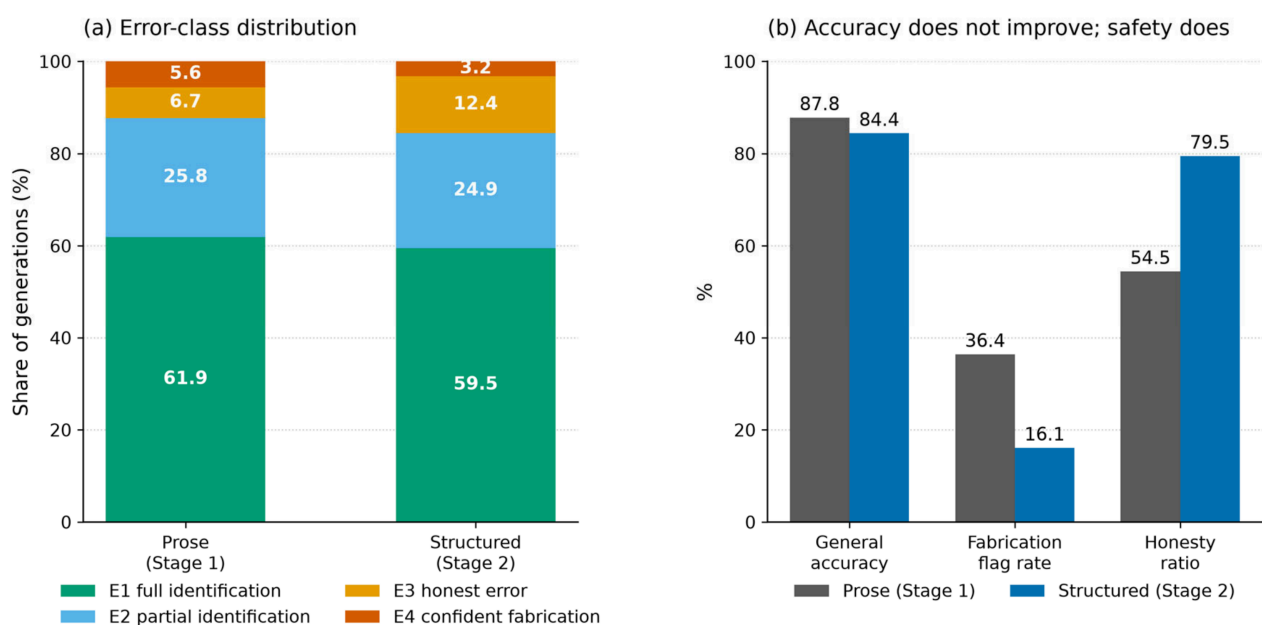


Figure 4. The prose-to-structured transition under the multimodal judge V2, same subject model on both sides: (a) error-class distribution; (b) accuracy, fabrication rate, and honesty ratio.

Across thirty runs per drawing, the six architectures are indistinguishable on accuracy and far apart on safety (Table 11, Figure 5). Their accuracy clusters within about five points, from roughly 82% for the few-shot design J6 at the low end to roughly 87% at the high end, and the six do not differ (Friedman $\chi^2(5) = 2.55$, $p = 0.77$). No pairwise comparison

approaches significance even before correcting for multiplicity (smallest $p = 0.38$). Because the absence of a difference is not equivalence, we tested for equivalence directly: at a five-point margin it holds for six of the fifteen pairs, and the largest difference observed is 4.2 points, with a 90% upper confidence bound of 10.3 points.

Table 11. The six structured architectures under V2 (30 runs per cell), ordered by fabrication flag rate; $N = 600$ (J6: 540). The numeric columns are shown as %.

Arch.	Family	E1	E2	E3	E4	GA	FAB	Honesty Ratio
J5 visual chain of thought	reasoning	59.0	24.8	15.3	0.8	83.8	3.5	95.0
J4 constrained two-stage	reasoning	58.0	25.0	15.0	2.0	83.0	7.5	88.2
J2 strict-mirror schema	schema	64.8	22.3	8.0	4.8	87.2	14.3	62.5
J6 few-shot exemplars	exemplar	54.3	27.6	15.2	3.0	81.9	15.0	83.5
J1 anchored-broader schema	schema	58.3	26.8	10.2	4.7	85.2	26.3	68.5
J3 naive two-stage (blind)	reasoning	62.3	23.0	11.0	3.7	85.3	29.7	74.8

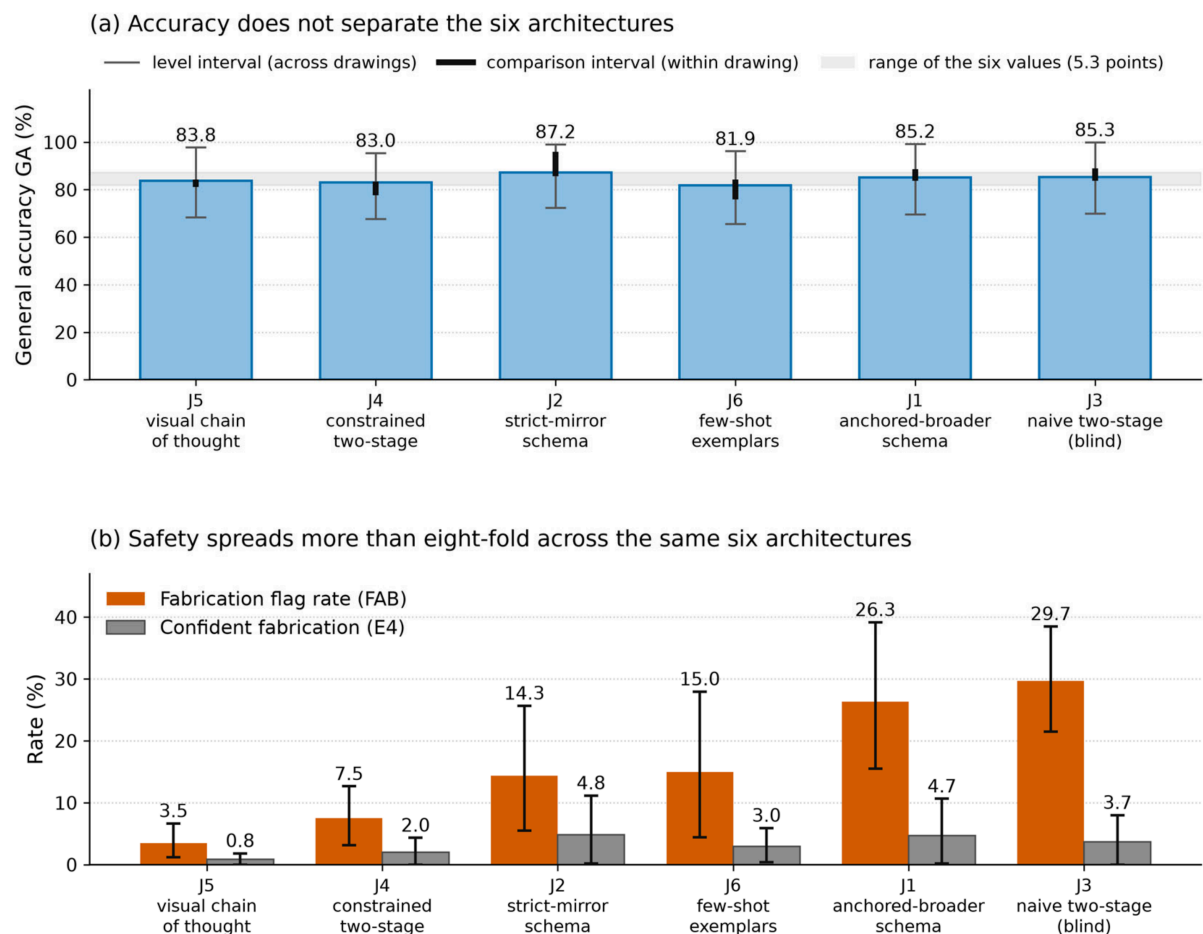


Figure 5. The six structured architectures under the multimodal judge V2 (30 runs per cell), ordered safest to riskiest: (a) general accuracy; (b) fabrication and confident-fabrication (E4) rates. Bars in (a) carry two 95% bootstrap intervals, the outer for the level and the inner for the comparison. Panel (b) shows level intervals only.

The flatness has a simple structure behind it. On nine of the eighteen drawings scored under all six architectures, every architecture is correct on every run, and the differences that do exist are concentrated in a few hard drawings where no architecture leads consistently. Within the partial class E2, a partial subtype is more common than a wrong one in every

architecture: 17.8% against 7.1% of all responses, most narrowly under the permissive schema J1 (15.8% against 11.0%).

On safety, the same six designs separate by nearly an order of magnitude. The fabrication flag rate runs from about one response in thirty for the safest design, the visual chain of thought J5, to almost one in three for the riskiest, the two-stage pipeline J3, an eight-fold spread that is significant in the paired design (Friedman $\chi^2(5) = 36.9$, $p < 0.001$; J3 against J5, 26.2 points, 95% CI 18.8 to 34.3). The safe end is occupied by the designs that keep the model looking at the drawing (J5, J4); the risky end by the designs that let it write and analyze freely (J3, J1). Most of this fabrication sits on answers that identify the component. Of the 569 flagged responses, 457 (80%) are E1 or E2, from 66% of the flags for J2 to 88% for J3, and E4 accounts for the rest. The ranking on fabrication is therefore driven mainly by invented detail attached to correct or partial identifications, which is why the fabrication rate, recorded for every response, is the primary safety measure.

Read against the design table, three levers account for the ordering. The first and strongest is to keep the image in view at every step that makes a claim. The two-stage pipeline J3, whose second call writes the final answer without seeing the drawing, is the worst fabricator. Its constrained sibling J4, identical in staging but with the image retained and its claims checked in the second call, cuts fabrication several-fold. The single-stage visual chain of thought J5, which never loses the image and forces named observations before it commits to a category, is the safest cell of the study. The second lever is to constrain the output: between the two single-call schema designs, the strict mirror J2 nearly halves the fabrication of the broad, permissive J1, because the permissive field set invites the elaboration where fabrication lives. The third lever is a negative result: adding worked exemplars does not buy safety. The few-shot design J6 lands mid-pack on fabrication and carries the lowest accuracy of the six, a sign of over-anchoring to the examples rather than of a safety mechanism.

The picture is consistent. The architectures that hold the drawing in view and require the model to describe what it sees before it names it produce the fewest unsupported claims, at no cost in accuracy. On this task and with this model, the fabrication rate depended on how the pipeline was built.

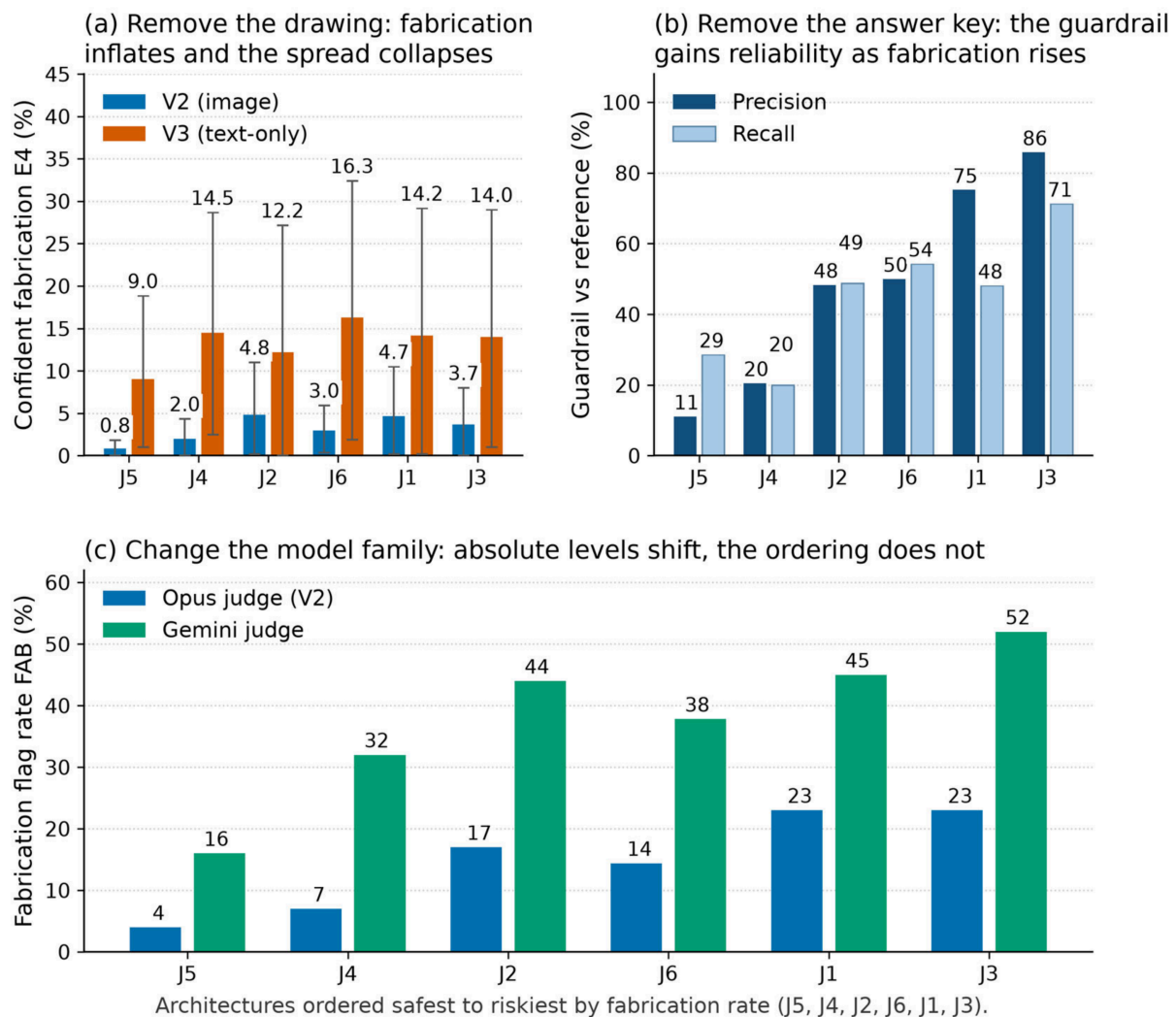
4.3. The Evaluator Axis

Three ablations isolate what each input to the evaluator is worth: remove the drawing, remove the answer key, or change the model family. They converge on one asymmetry, which the rest of this section documents: Identification requires the answer key, while grounding requires only the drawing.

Removing the drawing leaves identification almost untouched and destroys the safety axis. Under V3, the judge still recovers the component category and the exact error class within a point or two of the reference (Table 12). What collapses is the separation between an honest error and a confident invention: honest error shrinks by more than half, confident fabrication inflates several-fold, and the wide safety spread that set the architectures apart under V2 compresses from nearly six-fold to under two-fold (Figure 6a). Under the image-grounded judge the six architectures differ sharply in fabrication rate (Friedman $\chi^2(5) = 36.9$, $p < 0.001$); under the text-only judge that difference is no longer detectable ($\chi^2(5) = 8.8$, $p = 0.12$). The cause is structural rather than a matter of judgment. With the fabrication axis reduced to compatibility with the ground-truth class, a response that names the wrong component is very nearly forced into a fabrication verdict. Most of what a text-only judge reports as fabrication is an artifact of its own blindness rather than a property of the response.

Table 12. The modality ablation on the same 3540 generations.

Arch.	E3 V2 (%)	E3 V3 (%)	E4 V2 (%)	E4 V3 (%)	Honesty V2 (%)	Honesty V3 (%)
J1	10.2	2.2	4.7	14.2	68.5	13.4
J2	8.0	3.5	4.8	12.2	62.5	22.3
J3	11.0	0.8	3.7	14.0	74.8	5.4
J4	15.0	3.0	2.0	14.5	88.2	17.1
J5	15.3	6.0	0.8	9.0	95.0	40.0
J6	15.2	4.8	3.0	16.3	83.5	22.7

**Figure 6.** The evaluator axis, three judge ablations: (a) removing the drawing (V2 vs. text-only V3); (b) removing the ground truth (image-only guardrail V4); (c) changing the model family (V2 vs. Gemini). Error bars in (a) are 95% bootstrap intervals across drawings. The V2 to V3 comparison is paired and therefore tighter.

Removing the answer key costs the identification axes outright and leaves the grounding axis intact, so the guardrail and the reference can be compared on the fabrication flag alone. On that flag they agree on 86% of the 3540 responses. The reference flags 569 and the guardrail 522, almost the same volume, and the two coincide on 304 (Table 13). Overall kappa is 0.48, a moderate agreement. In aggregate, it approximates the reference rather than replacing it. The guardrail's reliability is not fixed, but rises with the amount of fabrication there is to find. Ranked by fabrication rate, the architectures fall in exactly that order on kappa agreement with the reference and on the precision of its flags (Table 13, Figure 6b).

This is how any detector behaves when its target is rare and it fixes the tool's place. The lowest precision, 11% on J5, has two further causes. The guardrail also counts a claim that contradicts the drawn geometry as ungrounded, so it flags wrong identifications that V2 records on the category axis instead: of the 48 J5 flags that V2 does not share, 21 fall on E3 responses and 14 on E2. And 38 of the 48 come from three drawings, 5.20, 5.28, and 5.22b. Counted against any problem V2 records, a fabrication or an identification short of E1, 76% of its J5 flags fall on a response with such a problem, against 43% for flags placed at random. The guardrail earns its keep on pipelines that fabricate often enough to be a genuine concern and adds little to pipelines that are already safe, where little is left to catch.

Table 13. The image-only guardrail V4 against the reference judge V2 on the fabrication flag, by architecture and overall; rows ordered safest to riskiest.

Arch.	N	V2 Flags	V4 Flags	Both	Recall (%)	Precision (%)	Kappa
J5	600	21	54	6	29	11	0.12
J4	600	45	44	9	20	20	0.14
J2	600	86	87	42	49	48	0.40
J6	540	81	88	44	54	50	0.43
J1	600	158	101	76	48	75	0.48
J3	600	178	148	127	71	86	0.70
All	3540	569	522	304	53	58	0.48

Changing the model family rather than the evidence leaves the ordering intact. An independent judge applying the same image-grounded rubric to a matched subsample places all six architectures in the identical order by fabrication rate, the visual chain of thought safest and the naive two-stage pipeline riskiest (Figure 6c). Absolute levels do not transfer. The outside judge (Gemini) is a consistently stricter detector, across every architecture. The safety ordering reflects how the pipelines are built; the absolute fabrication rate depends partly on who is judging.

5. Discussion

The study put two questions to prompt design at once: whether it improves the accuracy of drawing interpretation and whether it changes the character of the errors that remain. Across two stages, nine prose configurations, six structured architectures, and nearly twelve thousand scored judgments, the two answers separate cleanly. Accuracy barely moved. Prompt enrichment did not raise it, structured output did not raise it, and no architecture raised it. What moved, by nearly an order of magnitude, was the rate of unsupported claims: whether an answer, right or wrong, asserts specifics that the drawing does not show. The rest of the evidence follows from that separation. That rate is measurable only by an evaluator that can see the drawing. It can be partially monitored at deployment without any answer key and the architecture ranking it produces survives a change of judge model family. Table 14 sets those answers against the five research questions we have introduced.

5.1. Separating Honest and Confident Errors

The conceptual core of the study is the distinction between E3 and E4. The two classes are identical on identification: in both, the model has assigned the component to the wrong broad category. They are separated by manner. An E3 response hedges. It says, in effect, *“this appears to be some kind of coupling, though the type is hard to determine from this view”*. An E4 response asserts. It says, *“this is a hydraulic cylinder with an internal seal, per ISO*

6020-2" [54], when no part of the drawing supports any of it. Both miss the answer. Only one misleads.

Table 14. The five research questions and the evidence.

RQ	Question	Answer
RQ1	Does prose enrichment improve interpretation, in accuracy or in honesty of error?	No to either. Accuracy does not improve and fabrication rises with enrichment.
RQ2	Do structured architectures differ in safety rather than accuracy?	Yes. Accuracy does not separate the architectures, while the fabrication rate spreads eight-fold.
RQ3	Which evaluator properties determine whether the safety signal is measurable?	Modality. Without the image, honest error is reclassified as fabrication and the differentiation washes out.
RQ4	Can the signal be recovered at deployment, without ground truth?	Partially. An image-only check approximates the reference and gains reliability as fabrication rises.
RQ5	Are the findings robust across judge model families?	The ranking, yes, but the absolute level, no.

The asymmetry between these failures is operational, not aesthetic. An honest error is a known unknown. A reader, a reviewer, or a downstream check can act on hedged language: request another view, escalate to a person, decline to proceed. A confident fabrication is an unknown unknown. It carries the surface markers of expertise, a named standard, an exact dimension, a specific feature, so a reader is likely to trust it, and the error propagates into whatever depends on it. In engineering interpretation, the costs are not symmetric. A cautious wrong reading costs a clarification. A confident wrong reading can drive a wrong part, a wrong specification, or a wrong load assumption. The dangerous output is not the wrong answer; it is the wrong answer wearing the uniform of a right one.

This is why two systems with the same error rate are not equally safe, and why the study reports the honesty ratio alongside accuracy. A system whose failures fall into E4 has lost that signal, and no abstention policy can protect its user, because the wording of its answers no longer distinguishes right from wrong. Keeping errors honest is therefore a precondition for automation a reader can rely on, and it is independent of raw accuracy.

The empirical point of the study is that, in this setting, this precondition responded to design. Accuracy stayed within a few points everywhere; the honesty ratio ranged from 62.5% to 95.0% across architectures that differ only in prompt structure. The fabrication rate was set largely in the prompt, at no accuracy cost, by design choices that Section 5.3 turns into recommendations.

Stage 1 reproduced a known phenomenon: the most heavily engineered prompt produced the worst outcomes, and the degradation concentrated at the cell that stacks full role scaffolding on a closed answer constraint (C2). Measuring fabrication directly, rather than inferring it from a weighted score, lets us say more precisely what went wrong, because the factorial design of Section 3.4.2 happens to scale three different kinds of enrichment, and they did not fail together.

The first kind is integrative grounding information. The depth-1 hint states a fact about the materials, that the views belong to one component, and it targets a real failure mode, fragmented per-view analysis, while leaving the answer space open. This enrichment was benign to helpful everywhere, and it never raised confident error. The second kind is persona and governance scaffolding. The move from the minimal prompt to the full role structure added instructions about how to behave, not information about the drawing. Its effect was inconsistent and, for the strongest model, mildly negative. The third kind

is the closed answer set. The depth-2 constraint removes freedom, instructing the model to commit to one of three categories. This is where confident error concentrated: within the C row, E4 roughly doubles at the step from C1 to C2, and the pilot models show the same signature. The closed set converts hesitation on ambiguous drawings into forced, specific, wrong commitments. What the literature calls overprompting was, in this design, over-constraining the answer space while burying the model in behavioral instruction.

Stage 2 sharpens this reading, because at first sight it seems to contradict it. The strict-mirror schema J2, the most constrained output format in the study, nearly halved fabrication relative to the permissive schema J1. Constraint hurt in Stage 1 and helped in Stage 2. The resolution is that the two constraints act on different objects. The depth-2 constraint closes the content of the answer: it forces a commitment about what the part is, regardless of evidence. The J2 schema closes the form of the answer: it bounds where the model may elaborate, and elaboration is where fabricated specifics live, while leaving the identification itself free. The loose schema J1 shows the mirror image: freedom to elaborate is an invitation to fabricate (26.3%). The rule that survives both stages is not “less prompting” or “more structure.” It is: constrain the format, keep the interpretive space open, and ground every claim in the image.

These observations sit naturally in the emerging framing of context engineering, which shifts attention from how much instruction a prompt carries to which context an LLM system is given, in what form, at which step. Our results are a domain-specific instance of that shift with a safety twist: the context that mattered was the image itself, kept present at every claim-making step, and the enrichment that mattered was the kind that sharpens a decision boundary rather than the kind that decorates a role.

These architecture-level results confirm, in a domain where they had not been tested, several findings from adjacent work. The failure of the blind two-stage design reproduces the describe-then-generate bottleneck reported for vision-language pipelines [3]. The restoring of the image with claim verification in the second call, in the spirit of image-grounded post hoc correction [4], removes most of it. The visual chain of thought transfers the observe-then-classify discipline [6–8] to engineering drawings, where its benefit proves to be honesty rather than accuracy. Schema-constrained extraction [1,2] is supported in its safety reading. Few-shot exemplars, effective in fine-grained visual classification [9], did not transfer as a safety mechanism here.

5.2. The Evaluator Is Part of the System

The modality ablation carries a methodological conclusion that we expect to hold beyond this study, although it is tested here on one task: what an evaluation measures depends on what the evaluator can see, and for fabrication the dependence is first-order. A large fraction of what a text-only evaluation would report as fabrication is, in this precise sense, constructed by the evaluator rather than committed by the model.

It is tempting to read the stricter judge as the more objective one, and the temptation should be resisted. The text-only judge is not stricter; it is miscalibrated in both directions at once. It is too harsh on honest wrong-category answers, which it reclassifies wholesale as fabrication because a wrong answer is, by construction, semantically incompatible with the reference class. And it is too lenient on plausible invented specifics, because a fabricated standard or dimension that is consistent with the reference class passes a compatibility test even though nothing in the drawing supports it. That second failure is exactly the dangerous one in engineering. The multimodal judge is not infallible either, and we do not present it as truth. We use it as the better-grounded instrument for one structural reason: it holds the evidence that the construct is defined over. Fabrication means asserting what the

drawing does not show; a judge that cannot see the drawing is not measuring a stricter version of that property, it is measuring a different property.

The prose format of Stage 1 makes that miscalibration worse. A prose interpretation names the component and then elaborates on it, so once the identification is wrong the elaboration describes the wrong component, and the response doubles down rather than hedging. Under a text-only judge, one wrong identification carries the whole response into E4. Human review shows the cost: in the second half, a quarter of the responses the judge placed outside correct identification, almost all of them in E4, proved on review to be correct or partly correct readings. A hedged answer is the opposite case, because its uncertainty is legible in the text, while a confident fabrication carries no such marker. The Opus 4.7 subject rarely hedges, answering in a declarative manner, and only one of its 360 responses was scored E3. This is what focused the study on the safety of errors, and on a judge that can recognize hedging in whatever form it takes.

The honest error class is the one worth protecting, because an error that announces itself can be caught by a guardrail while a confident fabrication offers nothing to catch. Stage 2 responds by constraining the output into a fixed schema and by giving the judge the drawing, both aimed at making that distinction recoverable.

The deployment reading of the ablation is the uncomfortable one. The text-only judge is not a strawman; it is a stand-in for every downstream consumer that receives the interpretation text without the drawing, a database, a report, a second model in a pipeline, a human reading an extracted table. Whatever cannot be distinguished by V3, cannot be distinguished by them. The fabrications that only the multimodal judge catches are precisely the fabrications that a text-consuming pipeline will trust. Nor can the subject write its way out of the problem: the architecture whose safety depended on in-text evidence pairing (J4) lost its entire advantage under the blind judge, while only explicit hedging (J5) remained partially legible in text. Calibrated language helps a text-only consumer a little. It does not substitute for re-attaching the image at validation.

This also answers the standard concern about LLM-as-judge evaluation, that a judge may favor outputs of its own family: an independent family reproduces the entire safety ranking while shifting all absolute levels. The finding we have not, to our knowledge, seen stated elsewhere, and the one we would most like others to test, is that, in this study, judge modality mattered more than judge identity.

5.3. Deployment Monitoring Without the Answer Key

The obvious objection to everything above is that the canonical judge holds the answer key, and deployment does not. Section 4.3 gives the reply: correctness is a relation between the response and the reference, while fabrication is a relation between the response and the drawing, which the deployed system already holds as its input. The two judgments have different evidence requirements, and the dangerous one is the cheaper one.

The guardrail experiment tests exactly this decomposition. An image-grounded check holding no ground truth recovers a usable part of the reference's verdict and still misses about half of it, because a confident wrong reading that is visually consistent with the drawing gives an image-only checker nothing to catch. Both halves of this result matter. The first means that a reference-free monitor is feasible as triage: it flags 43.8% of the E4 responses against 7.2% of the E1 responses, and it agrees best with the reference on the high-fabrication architectures where monitoring is most needed. The second half means the monitor is a triage instrument, not a safeguard, and its blind spot is not fixable by better prompting, because it is an information limit, not a calibration failure.

We suggest reading this as a policy rather than as a metric. A deployed interpretation system cannot triage on accuracy, because the answer is unknown at inference

time. It can triage on grounding, because the image is known. The policy is: generate with a low-fabrication architecture; run the image-grounded check on the output; pass what is grounded; escalate to a human what is confident but ungrounded. Under the study's numbers, the safest architecture leaves a confident-fabrication rate of 0.8% (5 of its 600 responses), of which the guardrail flags 3, too few to give a rate. The residual unmonitored surface is small, and more important, it is characterized: it consists of visually plausible misreadings, which is precisely the case a human reviewer should own. The blind spot of the guardrail is not an argument against the guardrail; it is the quantified argument for keeping the human in the loop on high-stakes drawings.

Table 15 condenses the results into five practical recommendations for LLM interpretation of mechanical drawings. They are recommendations for the tested setting, not general principles: the study is exploratory, with one domain, twenty drawings, three subject models in Stage 1 and one in Stage 2, and the recommendations inherit those bounds.

Table 15. Five practical recommendations for LLM interpretation of mechanical drawings, within the tested setting.

#	Recommendation	Evidence in This Study
1	Select the prompt architecture for safety, not accuracy. Accuracy does not separate the architectures, while the fabrication rate does.	Friedman $p = 0.77$, $df = 5$; no pairwise comparison significant (smallest $p = 0.38$); fabrication spreads 3.5% to 29.7% across the same architectures.
2	Keep the image present at every step that makes a claim. A pipeline stage that loses the image launders fabrication.	Blind two-stage J3 is the worst fabricator (29.7%) and restoring the image in the second call (J4) cuts it to 7.5%.
3	Constrain the output format and keep the answer space open. Loose schemas invite elaboration, while closed answer sets force commitment.	Strict schema J2 nearly halves fabrication against loose J1 (14.3% vs. 26.3%) and the closed-set C2 cell doubles confident error in Stage 1.
4	Prefer disciplined single-pass observation: name what is visible, then classify, with the image throughout.	J5 is the safest cell of the study (FAB 3.5%, E4 0.8%, honesty 95.0%) and the safest under both judge families.
5	Validate against the image, not the text alone. Treat a reference-free image check as first-pass triage, with a human behind it.	Text-only scoring inflates E4 3 to 4× and washes out the safety signal. The image-only guardrail approximates the reference, and the precision of its flags rises with the fabrication rate.

5.4. Limitations and Future Work

The study is exploratory, and its boundaries should be stated plainly.

The materials are twenty purposively selected drawings from one textbook, in one domain, mechanical components. They are single parts and small assemblies drawn from a teaching reference, so the set does not test higher visual density, full production assemblies, other drawing standards, or scanned and degraded originals. The conclusions are drawn for mechanical drawing interpretation; we make no claim that they extend to image interpretation in general and the domain scoping of Section 3 applies to every finding here. Confident fabrication is rare in Stage 2, at 0.8% to 4.8%, and the pairwise differences in that class are not individually significant. The judges are language models. The Stage 1 judge was compared with expert human scoring (Section 3.6), with one expert per item and weaker agreement on E4. The Stage 2 reference judge is supported by a cross-family check, which confirms its ranking rather than its absolute levels. All of it is reproducible:

the drawings, prompts, model versions and judge conditions are fully specified, allowing every reported number to be recomputed.

Three further limits follow from the design. The drawings show common components that models may have seen in training, which can raise accuracy but cannot create the differences between architectures, since all are compared on the same drawings. The fabrication axis scores unsupported and unhedged claims jointly, and nothing is measured about a model's internal state. The guardrail prompt was developed on part of the same generations.

Stage 2 holds the model fixed and varies only what surrounds it. That is a deliberate scope and it is not the only axis available. Model-level interventions bear on the same failure. A higher-resolution visual backbone would raise the ceiling on what can be resolved in a dense drawing, domain-specific instruction tuning would align a model with drafting convention rather than with general imagery, and a neuro-symbolic parser would replace inference over pixels with explicit geometry wherever the drawing is machine-readable. Each of these requires training or a parsing stack. We expect them to compose with the levers reported here rather than to replace them, since an architecture that keeps the image in view and forces observation before commitment should remain the safer way to consume whatever backbone sits underneath. Testing that composition is a natural next step where model training is available.

Specifically, future work could follow three axes. On generation, a second subject family and a broader artifact range, assemblies, schematics, and drawings, would test how far the architecture findings carry. On evaluation, increasing the human validation sample and extending the modality ablation to other multimodal tasks, would test the claim that evaluator modality is first-order. On deployment, the natural next step is the agentic one: embedding the image-grounded guardrail in a working review pipeline, where flagged interpretations route to an engineer and the routing thresholds are tuned on operational data. Such an agent would turn the triage policy of Section 5.3 from a proposal into a measured system.

6. Conclusions

This study asked whether prompt design governs the accuracy of LLM interpretation of mechanical drawings, the safety of its errors, or both. In the tested setting, accuracy did not respond to prompting. The rate at which answers asserted details the drawing does not support did respond, by nearly an order of magnitude, and so did the share of wrong answers that stayed free of such claims. For this task, prompt design acted on the safety of the errors rather than on the accuracy of the answers.

The contribution is a comparative framework, and its parts differ in kind. Several findings confirm, in a domain where they had not been tested, relationships reported elsewhere: overprompting, the describe-then-generate bottleneck, and the benefit of observation before classification and of a constrained output schema. Three are new. The study makes the separation of honest from fabricated errors the outcome of a prompt comparison, through a three-axis scoring scheme. It shows that what the judge can see, not only who the judge is, decides whether fabrication can be measured: a text-only judge inflates confident fabrication three to four times. And it gives a measured account of what an image-only check can and cannot catch without a reference answer.

For practice, every element of the resulting recipe is a prompting or pipeline decision: generate with single-pass observation under a constrained schema, check the output against the image, and send confident but ungrounded interpretations to a person. In this study, that recipe left a confident-fabrication rate of 0.8%, and across Stage 2 the image-only check flagged 49 of the 112 confident fabrications. Both figures are bounded by one domain,

twenty textbook drawings, and one subject model in Stage 2. Testing them on other model families, drawing types and production data is the natural next step.

A model that says a part might be a coupling when it is a bolted joint has failed in a way a reader can catch. A model that names the wrong part with a standard and a bore diameter has failed in a way a reader may not. The two should not be averaged into one error rate. In this study, the distance between them was set largely by choices in the practitioner's hands, not by the model alone.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/technologies14100666/s1>: Section S1, the nine prose prompt configurations of Stage 1; Section S2, the six structured architectures J1 to J6 of Stage 2; Section S3, the judge prompts for the five evaluator conditions; Section S4, the human validation of the Stage 1 judge in full; Section S5, the inferential statistics in full; Table S1: Human agreement with the Stage 1 judge (V1), by the class the judge assigned; Table S2: Stage 2 pairwise comparisons of general accuracy under the multimodal judge V2; Table S3: Stage 2 pairwise comparisons of the fabrication flag rate under the multimodal judge V2; Table S4: Stage 2 pairwise comparisons of confident fabrication (E4) under the multimodal judge V2; Table S5: Equivalence testing on general accuracy; Table S6: Index of every inferential test cited in the article.

Author Contributions: Conceptualization, P.P. and P.N.B.; methodology, P.P. and A.P.; software, P.P.; validation, P.P., A.P. and P.N.B.; formal analysis, P.P.; investigation, P.P. and A.P.; data curation, P.P. and A.P.; writing—original draft preparation, P.P.; writing—review and editing, P.P., A.P. and P.N.B.; visualization, P.P.; supervision, P.N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The prompts, the structured ground truth, the raw model generations, and all judge scores are available from the corresponding author upon request.

Acknowledgments: Generative artificial intelligence was a core element of this study. Because the research investigates large language models directly, such models carried out the interpretation of the drawings and the evaluation of those interpretations. In preparing the manuscript, the authors additionally used AI-based assistance for data analysis, and, in a reviewing capacity, for language editing, grammar improvement, and correction. All these uses were carried out under continuous human direction, with the authors designing, supervising, and jointly editing the research and the manuscript throughout. The authors have reviewed all content and take full responsibility for it.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large language model
GA	General accuracy
GT	Ground truth
JSON	JavaScript Object Notation
CoT	Chain of thought
FAB	Fabrication rate
MLLM	Multimodal large language model
VLM	Vision-language model
GD&T	Geometric dimensioning and tolerancing
RQ	Research question

References

1. Mortezaagha, P.; Shaw, J.; Sun, B.; Rahgozar, A. From chaos to clarity: Schema-constrained AI for auditable biomedical evidence extraction from full-text PDFs. *BMC Med. Res. Methodol.* **2026**, *26*, 119. [\[CrossRef\]](#)
2. Ravikumar, R.; Ibrahim, N.; Batista-Navarro, R. Lost in Formatting: How Output Formats Skew LLM Performance on Information Extraction. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2026; pp. 5498–5513. [\[CrossRef\]](#)
3. Kodathala, S.V.; Vunnam, R. The Describe-Then-Generate Bottleneck: How VLM Descriptions Alter Image Generation Outcomes. *arXiv* **2025**, arXiv:2509.18179. [\[CrossRef\]](#)
4. Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *Sci. China Inf. Sci.* **2024**, *67*, 220105. [\[CrossRef\]](#)
5. Tang, W.; Sun, Y.; Gu, Q.; Li, Z. Visual Position Prompt for MLLM Based Visual Grounding. *IEEE Trans. Multimed.* **2026**, *28*, 3739–3754. [\[CrossRef\]](#)
6. Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Sun, Z.; Gutfreund, D.; Gan, C. Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning. *AAAI* **2024**, *38*, 1254–1262. [\[CrossRef\]](#)
7. Yu, B.; Shen, T.; Chen, L. STA-CoT: Structured Target-Centric Agentic Chain-of-Thought for Consistent Multi-Image Geological Reasoning. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 25426–25444. [\[CrossRef\]](#)
8. Huang, O.; Nobre, C. ViStruct: Simulating Expert-Like Reasoning Through Task Decomposition and Visual Attention. In *Proceedings of the 2025 IEEE Visualization and Visual Analytics (VIS)*; IEEE: Piscataway, NJ, USA, 2025; pp. 216–220. [\[CrossRef\]](#)
9. Lamm, B.; Keuper, J. A Visual RAG Pipeline for Few-Shot Fine-Grained Product Classification. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2025; pp. 2089–2098. [\[CrossRef\]](#)
10. Xu, W.; Zhu, G.; Zhao, X.; Pan, L.; Li, L.; Wang, W. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 15474–15492. [\[CrossRef\]](#)
11. Chen, B.; Zhang, Z.; Langrené, N.; Zhu, S. Unleashing the potential of prompt engineering for large language models. *Patterns* **2025**, *6*, 101260. [\[CrossRef\]](#)
12. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; pp. 24824–24837.
13. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models Are Zero-Shot Reasoners. In *Proceedings of the Advances in Neural Information Processing Systems 35*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; pp. 22199–22213. [\[CrossRef\]](#)
14. Cao, B.; Cai, D.; Zhang, Z.; Zou, Y.; Lam, W. On the Worst Prompt Performance of Large Language Models. In *Proceedings of the Advances in Neural Information Processing Systems 37*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; pp. 69022–69042. [\[CrossRef\]](#)
15. Shahriar, S.; Lund, B.D.; Mannuru, N.R.; Arshad, M.A.; Hayawi, K.; Bevara, R.V.K.; Mannuru, A.; Batool, L. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Appl. Sci.* **2024**, *14*, 7782. [\[CrossRef\]](#)
16. Wardle, G.; Sušnjak, T. Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks. *BDCC* **2025**, *9*, 149. [\[CrossRef\]](#)
17. Pawlik, L. How the Choice of LLM and Prompt Engineering Affects Chatbot Effectiveness. *Electronics* **2025**, *14*, 888. [\[CrossRef\]](#)
18. Cooper, A.; Kato, K.; Shih, C.-H.; Yamane, H.; Vinken, K.; Takemoto, K.; Sunagawa, T.; Yeh, H.-W.; Yamanaka, J.; Mason, I.; et al. Rethinking VLMs and LLMs for image classification. *Sci. Rep.* **2025**, *15*, 19692. [\[CrossRef\]](#)
19. Zi, Y.; Menon, H.; Guha, A. More Than a Score: Probing the Impact of Prompt Specificity on LLM Code Generation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*; The Asian Federation of Natural Language Processing and The Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 2380–2402. [\[CrossRef\]](#)
20. Tang, Y.; Tuncel, D.; Koerner, C.; Runkler, T. The Few-shot Dilemma: Over-prompting Large Language Models. In *Proceedings of the 2025 3rd International Conference on Foundation and Large Language Models (FLLM)*; IEEE: Piscataway, NJ, USA, 2025; pp. 134–141. [\[CrossRef\]](#)
21. Gozzi, M.; Di Maio, F. Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts. *Electronics* **2024**, *13*, 4712. [\[CrossRef\]](#)
22. Aloufi, N.; Aljuhani, A. Empirical Evaluation of Prompting Strategies for Python Syntax Error Detection with LLMs. *Appl. Sci.* **2025**, *15*, 9223. [\[CrossRef\]](#)

23. Chen, L.-C.; Weng, H.-T.; Pardeshi, M.S.; Chen, C.-M.; Sheu, R.-K.; Pai, K.-C. Evaluation of Prompt Engineering on the Performance of a Large Language Model in Document Information Extraction. *Electronics* **2025**, *14*, 2145. [\[CrossRef\]](#)
24. Xiao, Y.; Zou, J.; Yang, Q. Advancing Few-Shot Named Entity Recognition with Large Language Model. *Appl. Sci.* **2025**, *15*, 3838. [\[CrossRef\]](#)
25. Tian, Y.; Liu, A.; Dai, Y.; Nagato, K.; Nakao, M. Systematic synthesis of design prompts for large language models in conceptual design. *CIRP Ann. Manuf. Technol.* **2024**, *73*, 85–88. [\[CrossRef\]](#)
26. Zhen, H.; Shi, Y.; Huang, Y.; Yang, J.J.; Liu, N. Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference. *Computers* **2024**, *13*, 232. [\[CrossRef\]](#)
27. Zheng, M.; Pei, J.; Logeswaran, L.; Lee, M.; Jurgens, D. When “A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 15126–15154. [\[CrossRef\]](#)
28. Xu, Z.; Jain, S.; Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models. *arXiv* **2024**, arXiv:2401.11817.
29. Dang, A.-H.; Tran, V.; Nguyen, L.-M. Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Front. Artif. Intell.* **2025**, *8*, 1622292. [\[CrossRef\]](#)
30. Faria, F.T.J.; Baniata, L.H.; Choi, A.; Kang, S. Analyzing Diagnostic Reasoning of Vision–Language Models via Zero-Shot Chain-of-Thought Prompting in Medical Visual Question Answering. *Mathematics* **2025**, *13*, 2322. [\[CrossRef\]](#)
31. Elhenawy, M.; Abutahoun, A.; Alhadidi, T.I.; Jaber, A.; Ashqar, H.I.; Jaradat, S.; Abdelhay, A.; Glaser, S.; Rakotonirainy, A. Visual Reasoning and Multi-Agent Approach in Multimodal Large Language Models (MLLMs): Solving TSP and mTSP Combinatorial Challenges. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1894–1921. [\[CrossRef\]](#)
32. Ju, H.; Park, I.; Nalcakan, Y.; Jin, Y.; Yeo, S.; Kim, S. Exploring the Limits of Large Language Models’ Ability to Distinguish Between Objects. *Appl. Sci.* **2025**, *15*, 4620. [\[CrossRef\]](#)
33. Liu, S.; Wen, T.; Pattamatta, A.S.L.S.; Srolovitz, D.J. A prompt-engineered large language model, deep learning workflow for materials classification. *Mater. Today* **2024**, *80*, 240–249. [\[CrossRef\]](#)
34. Choi, J. Efficient Prompt Optimization for Relevance Evaluation via LLM-Based Confusion Matrix Feedback. *Appl. Sci.* **2025**, *15*, 5198. [\[CrossRef\]](#)
35. Anghel, C.; Anghel, A.A.; Pecheanu, E.; Craciun, M.V.; Cocu, A.; Niculita, C. PEARL: A Rubric-Driven Multi-Metric Framework for LLM Evaluation. *Information* **2025**, *16*, 926. [\[CrossRef\]](#)
36. Wang, Y.; Huang, J.; Du, L.; Guo, Y.; Liu, Y.; Wang, R. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Comput. Educ. Artif. Intell.* **2025**, *9*, 100481. [\[CrossRef\]](#)
37. Shah, M.H.; Cuayáhuil, H. Systematic Analysis of Vision–Language Models for Medical Visual Question Answering. *Multimodal Technol. Interact.* **2026**, *10*, 16. [\[CrossRef\]](#)
38. Wang, J.; Seng, K.P.; Shen, Y.; Ang, L.-M.; Huang, D. Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering. *Electronics* **2024**, *13*, 2273. [\[CrossRef\]](#)
39. Dokas, I.M. From hallucinations to hazards: Benchmarking LLMs for hazard analysis in safety-critical systems. *Saf. Sci.* **2026**, *194*, 107056. [\[CrossRef\]](#)
40. Bhanbhro, H.; Kwang Hooi, Y.; Kusakunniran, W.; Amur, Z.H. A Symbol Recognition System for Single-Line Diagrams Developed Using a Deep-Learning Approach. *Appl. Sci.* **2023**, *13*, 8816. [\[CrossRef\]](#)
41. Villena Toro, J.; Wiberg, A.; Tarkian, M. Optical character recognition on engineering drawings to achieve automation in production quality control. *Front. Manuf. Technol.* **2023**, *3*, 1154132. [\[CrossRef\]](#)
42. Moon, J. Symmetric Graph-Based Visual Question Answering Using Neuro-Symbolic Approach. *Symmetry* **2023**, *15*, 1713. [\[CrossRef\]](#)
43. Bray, N.; Hempel, M.; Boeding, M.; Sharif, H. Decoding Technical Diagrams: A Survey of AI Methods for Image Content Extraction and Understanding. *Information* **2026**, *17*, 165. [\[CrossRef\]](#)
44. Zha, Y.; Li, Y.; Lu, X.-G. Enhancing Large Language Model Comprehension of Material Phase Diagrams through Prompt Engineering and Benchmark Datasets. *Mathematics* **2024**, *12*, 3141. [\[CrossRef\]](#)
45. Kadiyala, L.A.; Mermer, O.; Samuel, D.J.; Sermet, Y.; Demir, I. The Implementation of Multimodal Large Language Models for Hydrological Applications: A Comparative Study of GPT-4 Vision, Gemini, LLaVa, and Multimodal-GPT. *Hydrology* **2024**, *11*, 148. [\[CrossRef\]](#)
46. Gupta, M.; Wei, C.; Czerniawski, T.; Eiris, R. PIDQA—Question Answering on Piping and Instrumentation Diagrams. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 39. [\[CrossRef\]](#)
47. Zhang, L.; Le, B.; Akhtar, N.; Lam, S.-K.; Ngo, T. Large Language Models for Computer-Aided Design: A Survey. *ACM Comput. Surv.* **2026**, *58*, 1–39. [\[CrossRef\]](#)
48. Khan, M.T.; Chen, L.; Ng, Y.H.; Feng, W.; Tan, N.Y.J.; Moon, S.K. Leveraging Vision-Language Models for Manufacturing Feature Recognition in Computer-Aided Designs. *J. Comput. Inf. Sci. Eng.* **2025**, *25*, 104501. [\[CrossRef\]](#)

49. Khan, M.T.; Chen, L.; Ng, Y.H.; Feng, W.; Tan, N.Y.J.; Moon, S.K. Fine-Tuning Vision-Language Model for Automated Engineering Drawing Information Extraction. In *Proceedings of the 9th International Conference on Innovation in Artificial Intelligence; Lecture Notes in Electrical Engineering*; Springer: Singapore, 2026; pp. 411–423. [[CrossRef](#)]
50. Li, X.; Sun, Y.; Sha, Z. LLM4CAD: Multimodal Large Language Models for Three-Dimensional Computer-Aided Design Generation. *J. Comput. Inf. Sci. Eng.* **2025**, *25*, 021005. [[CrossRef](#)]
51. Li, S.; Corney, J. MechRAG: A multimodal large language model for mechanical engineering. *Commun. Eng.* **2025**, *4*, 187. [[CrossRef](#)]
52. Picard, C.; Edwards, K.M.; Doris, A.C.; Man, B.; Giannone, G.; Alam, M.F.; Ahmed, F. From concept to manufacturing: Evaluating vision-language models for engineering design. *Artif. Intell. Rev.* **2025**, *58*, 288. [[CrossRef](#)]
53. Narayana, K.L.; Kannaiah, P.; Venkata Reddy, K. *Machine Drawing*, 3rd ed.; New Age International: New Delhi, India, 2006.
54. *ISO 6020-2:2015*; Hydraulic fluid power—Mounting dimensions for single rod cylinders, 16 MPa (160 bar) series—Part 2: Compact series. International Organization for Standardization: Geneva, Switzerland, 2015.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.