WORKCRAFT-ARENA: HOW CAN MODELS TURN HETEROGENEOUS WORKSPACES TO WELL-DESIGNED WORK ARTIFACTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models are increasingly serving as work assistants, turning heterogeneous workspace materials into slides, HTML pages, posters, and documents. Evaluating these artifacts requires more than checking task completion, because long-horizon creation can lose evidence between understanding and delivery. Additionally, existing aesthetic assessments lack an exploration of the professional delivery requirements and diverse design tastes of work artifacts. We introduce **WorkCraft-Arena**, a benchmark of 100 tasks across 10 professional work roles and six artifact types. We propose a unified design agent harness that records all creation processes from material understanding to self-revision for complete evaluation. Our arena has established the first comprehensive aesthetic evaluation system for work assistant artifacts, which separates faithfulness from aesthetics and further distinguishes **AestheticPass**, an absolute professional-readiness judgment, from **AestheticElo**, a pairwise measure of relative design preference. We conduct experiments across six frontier models and observe that faithfulness and aesthetics are almost orthogonal on the Pareto front. AestheticPass and AestheticElo will present different performance in the system. The results also show that factual grounding, design taste, and delivery reliability diagnose distinct capabilities for models. WorkCraft-Arena therefore provides both a credible measurement system for artifact aesthetics and diagnostic evidence for developing multimodal models and agent harnesses that produce reliable, improvable professional work.

1 INTRODUCTION

As foundation large language models grow more capable Kuang et al. (2025); Zhang et al. (2026b); Naveed et al. (2025) and agent harnesses mature Zhao et al. (2024); Kuang et al. (2026); Luo et al. (2025); Guo et al. (2026), agents are moving from research prototypes into real knowledge-work workflows Yao et al. (2026); Chen et al. (2026). They must interpret papers, reports, datasheets, code, and visual assets, then synthesize evidence and operate professional tools to deliver a work artifact such as slides, webpages, or documents. Users ultimately depend less on an intermediate response and more on a file they can inspect, revise, present, and use for real work. The natural unit of evaluation is therefore shifting from an isolated operation to a *deliverable work artifact*.

A successful artifact must be faithful to its workspace and turn that information into a clear, professional, and persuasive visual communication. Visual hierarchy, typography, colour, information density, and audience fit determine whether correct content can actually be understood and trusted by audience (Team et al., 2026). Yet aesthetic quality is unusually difficult to measure. Legibility, sufficient contrast, and freedom from overlap admit professional consensus, whereas composition, typographic voice, pacing, and visual storytelling permit multiple valid designs. Reference reproduction penalizes legitimate alternatives, while an open VLM score conflates compliance with preference, which can hide bland "all acceptable" outcomes. This motivates the question: **How can the creative organisation and aesthetic judgement of multimodal work agents be measured consistently and credibly while preserving the diversity of valid designs?**

Existing benchmarks address parts of this problem. For work artifacts, reference-based HTML page generation tasks provide precise structural or pixel evaluation targets but close the design space

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

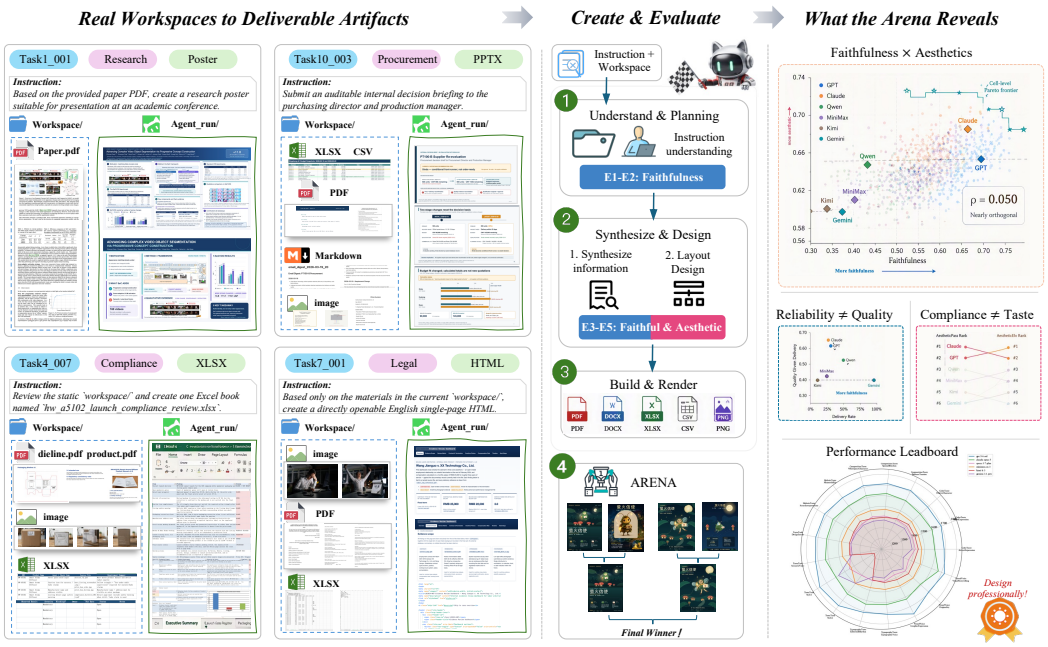

Figure 1. Overview of our WorkCraft-Arena, including workspace cases, evaluation procedures, and insights

(Si et al., 2024; Xiao et al., 2024; 2025). Office and slide benchmarks enrich the product types and evaluate content coverage and partial visual quality, yet aesthetics is usually a coarse sub-score rather than an independent measurement target (Ge et al., 2025; Huang et al., 2025; Gandhi et al., 2026). On the other hand, image generation and graphic-design benchmarks that focus on aesthetic evaluation probe composition, colour, typography, and creativity, but do not capture editable multi-page structure, information evidence provenance, or audience-specific functional aesthetics (Li et al., 2026; Deganutti et al., 2026; Hopkins et al., 2026). As work assistant agents evolve, more long-horizontal benchmarking is being conducted in work environments, primarily evaluating verifiable operational success or the overall score of the final design(Zhang et al., 2026c; Ding et al., 2026; ByteDance Seed et al., 2026). Therefore, a mature benchmark that can simultaneously measure the evidentiary faithfulness of the deliverable, professional design readiness, the open-ended design aesthetic, and the quality improvements or losses during the creation process is under exploration.

To address these, we introduce **WorkCraft-Arena** (Figure 1), comprising 10 work roles and six artifact types—PPTX, HTML, XLSX, poster, DOCX, and Markdown. Each task pairs a task instruction with a heterogeneous workspace and hidden evidence, constraint, and design annotations. To better evaluate the intermediate processes and final deliverables, we propose a design agent harness that supports creation while recording all processes from constraint and workspace understanding to synthesis, design, delivery, and revision. Under this process-aware setup, we aim to explore ***what capabilities are needed of a next-generation multimodal large language model in a real workspace and with high aesthetic expectations?*** And ***What agent-harness can organise these capabilities into artifacts that are deliverable, auditable, and continuously improvable?***

Therefore, WorkCraft-Arena evaluate faithfulness and aesthetics separately and decomposes aesthetics into two complementary protocols. **AestheticPass** evaluates one artifact with a professional delivery bar, while **AestheticElo** compares relative design preference with pairwise comparison in an arena. A task-specific criterion DAG and the multi-stage records then connect scores to failure modes like missing objects, incorrect attributes, planning drift, and ineffective revision, turning the benchmark from a static leaderboard into evidence for improving both models and harnesses. We conduct experiments across 6 frontier models, observing that faithfulness and aesthetics are nearly orthogonal. Professional readiness and design preference also reorder systems, while delivery reliability differs from quality on delivery. These results begin to expose distinct bottlenecks in grounding, synthesis, design, execution, and revision of models, and show why future agent harnesses need process-visible, renderer-aware, and delivery-aware infrastructure. Our contributions are:

- We introduce **WorkCraft-Anera**, which extends multimodal-agent evaluation to editable and renderable professional artifacts, with a unified agent harness making both final delivery and intermediate creation processes reproducible and diagnosable.
- We establish **A credible foundational measurement system for work-artifact aesthetics,** which separates faithfulness from aesthetics and decouples professional aesthetic consensus from open-ended aesthetic taste.
- Experiments provide actionable evidence and insights for the next generation of multimodal models and agent harnesses that there is an urgent need for more diverse and richer visual design styles, as well as a more auditable and visually-feedback-revised creation system.

2 Related Work

Artifact creation and reference reproduction. HTML design benchmarks progressed from static screenshot reproduction to interaction, editing, and repair (Si et al., 2024; Xiao et al., 2024; 2025). Slide and Office benchmarks further introduced instruction-driven creation, layout understanding, editing, and task-specific criteria (Ge et al., 2025; Huang et al., 2025; Ofengenden et al., 2025; Gandhi et al., 2026; Microsoft Research, 2026). These settings establish that content, structure, and visual planning are distinct skills, but reference similarity penalizes coherent alternatives, while reference-free aesthetics is typically a coarse sub-score rather than an independent target.

Table 1. Comparison of different benchmarks. **Work art.**: workspace artifact; **Faith.**: faithfulness; **Prof. aes.**: professional aesthetics; **Open taste**: Aesthetic taste. ✓ addressed, (✓) partially addressed, ✗ not addressed.

Benchmark family and representative work	Work art.	Faith.	Prof. aes.	Open taste	Process
Reference reproduction (Si et al., 2024; Xiao et al., 2025)	(✓)	(✓)	(✓)	✗	✗
Office/artifact creation(Ge et al., 2025; Gandhi et al., 2026)	✓	✓	(✓)	✗	(✓)
Image/design aesthetics (Deganutti et al., 2026; Hopkins et al., 2026)	(✓)	(✓)	✓	(✓)	✗
Workspace agents(Ding et al., 2026; ByteDance Seed et al., 2026)	(✓)	✓	✗	✗	(✓)
WorkCraft-Arena	✓	✓	✓	✓	✓

Aesthetic and creative evaluation. Recent image and design benchmarks move beyond generic aesthetics toward professional layout, typography, real-world fidelity, and creative generation (Li et al., 2026; Deganutti et al., 2026). Human Creativity Benchmark further treats disagreement as meaningful divergence (Hopkins et al., 2026), while construction of fine-grained rubric indicators improves visual measurement (Ban et al., 2026). Their primary unit, however, remains a canvas or image; they do not jointly assess editable multi-page structure, evidence provenance, medium-specific functionality, professional compliance, and relative taste in workspace deliverables.

Long-horizon workspace agents. Agent benchmarks increasingly evaluate realistic websites, stateful environments, native runtimes, and specialized professional workflows (Zhang et al., 2026c; ClawMark Team, 2026; Ding et al., 2026; ByteDance Seed et al., 2026). They make tool use and terminal execution more realistic and often retain trajectories for auditing. Yet their goals are usually verifiable operations or environment states, and process traces primarily support debugging rather than scoring intermediate creative competence. WorkCraft-Arena connects these lines by treating open-ended artifacts as the evaluation object and jointly measuring evidence fidelity, professional aesthetics, design preference, and the process that produces them, as comparison shown in Table 1.

3 WorkCraft-Arena Construction

In Figure 2, we first construct professional tasks with heterogeneous workspaces and human annotated structured evidence and design. Second, a design agent harness supports diverse format

creation while preserving intermediate process from material understanding to final render. Third, an evaluation suite across seven creation stages report faithfulness and aesthetics separately.

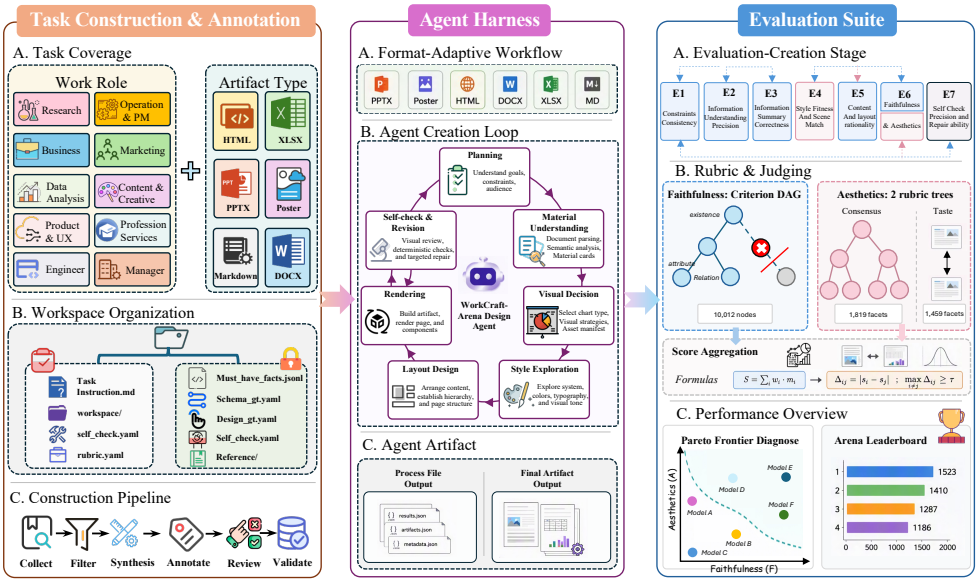

Figure 2. Overview of WorkCraft-Arena with data construction, agent harness, and evaluation system.

3.1 Task Definition

WorkCraft-Arena treats each task as an complete professional work assignment from a practical work instruction and heterogeneous workspace to one audience-ready artifact. We formalize task as

$$\mathcal{T}_i = \big(\mathcal{I}_i, \mathcal{W}_i, \mathcal{G}_i, \mathcal{R}_i\big), \tag{1}$$

where $\mathcal{I}_i$ is an instruction specifying role, audience, and delivery requirements; $\mathcal{W}_i$ is a multimodal workspace; $\mathcal{G}_i$ contains hidden annotations; and $\mathcal{R}_i$ contains hidden reference artifacts. Only $\mathcal{I}_i$ and $\mathcal{W}_i$ are visible during agent inference. The required output is one editable artifact in six formats—PPTX, poster, HTML, DOCX, XLSX, or Markdown. The 100 tasks span ten role domains: **research and knowledge work** (scientific evidence), **business strategy and analysis** (recommendations), **data analysis and decision support** (quantitative synthesis), **product and user experience** (user-centred rationale), **engineering and technical development** (technical communication), **operations and project management** (plans and risks), **marketing and sales** (grounded persuasion), **content and creative work** (narrative and hierarchy), **professional services and expert consulting** (specialist judgement), and **management, finance, and legal**. Appendix A.1 gives full scopes.

3.2 Data Construction

Data Collection and Filtering. We draw candidates from ClawMark (ClawMark Team, 2026), WildClawBench (Ding et al., 2026), Pi-Bench (Zhang et al., 2026a), ClawArena (Ji et al., 2026), and OSWorld (Yuan et al., 2026). First, we filter and classify candidates corresponding to each work role. Then, we remove tasks with online interaction dependencies, information leaks, insufficient evidence, and tasks without final deliverables (merely performing operations as a work assistant). Next, we further supplement the workspace with professional documentation like GDP.pdf (Garre et al., 2026) from different work role domains to enrich the diversity of workspaces. Annotators retain the existing task objectives and evidentiary dependencies for subsequent new task generation. Appendix A.2 details the filtering, rewriting, and resynthesis strategies for each data source.

Data Annotation and Validation. First, a complete annotation template is manually created for each product type. Based on the workspace and work role, we assign an appropriate artifact type and generate a work creation task instruction. Then, workspace materials are organized, and key information and the source location of each piece of fact are annotated, accompanied by numerical

extraction, estimation, and calculation. The content layout of product design is further annotated with the appropriate audience, and reference artifacts are established. Finally, automated annotation validation and manual verification are completed. During inference, the evaluated model only sees the `instruction.md` and `workspace/`. Other constraints, key facts, acceptable designs, and scoring criteria are only used in the evaluation phase. Three annotators validate the task semantics, evidence, design patterns, and rendering results, while disputes in more than 10% of fields trigger an extra annotator's adjudication. Appendix A.3 details the annotation and validation process.

Dataset Statistics. For **input material**, WorkCraft-Arena contains diverse 2,041 input files spanning 19 file types, an average of 20.4 files per task. For **role domains and artifact types**, WorkCraft-Arena comprises 100 tasks (79 English and 21 Chinese), balanced at 9–12 per work role domain, and 12–21 per artifact type. This coverage limits the influence of any single profession or output format and requires evaluated systems to accommodate different audience conventions and format-specific design requirements. Full distributions are in Appendix A.4.

3.3 WorkCraft-Arena Agent Harness

We propose an agent harness capable of creating six types of artifacts and documenting the entire process, which comprises: **(1) workspace material understanding**, **(2) artifact creation**, **(3) format-adaptive production**, and **(4) render-based self-check**. First, deterministic parsers and a material-understanding sub-agent extract text, tables, structure, visual evidence, and provenance into material cards and an asset manifest, while retaining links to the original files. The main agent then iteratively interprets the brief, selects evidence, plans content and design, invokes tools, and records constraints, grounding, organization, and design decisions as persistent process evidence.

A shared controller and completion protocol enable comprehensive design for diverse artifacts, while product adapters expose format-native construction and validation for PPTX, poster, HTML, DOCX, XLSX, and Markdown. Adapters handle low-level package, formula, chart, structure, and rendering requirements but never choose the evidence or visual solution. Finally, deterministic checks and rendered previews expose failures invisible in source code. The evaluated model visually diagnoses the first-pass draft and performs a focused revision. The entire creation process is fully recorded for subsequent evaluation to diagnose where the model has shortcomings. No hidden annotations or references are exposed during inference. Implementation details are in Appendix B.

3.4 WorkCraft-Arena Evaluation Suite

Professional creation is multi-stage, multi-constraint, and admits multiple valid designs. Our suite therefore follows three principles. **Process diagnosis** scores intermediate competencies separately so that, for example, extracting a fact and preserving it in the final artifact are different. **Orthogonal reporting** separates verifiable faithfulness from aesthetics. **Task-specific measurement** derives criteria from each workspace, artifact type, and annotated design intent rather than applying a generic natural-image aesthetic scorer. Therefore, we set a process-level metric system aligned with the creation process. The harness persists evidence along seven aligned stages, so we design different faithful and aesthetic metrics for each stage. E1 retains constraints and intent; E2 extracts facts and provenance; E3 groups evidence and selects visual forms; E4 realizes a scenario-appropriate visual design; E5 plans content structure; E6 renders and evaluates the artifact through faithfulness, professional readiness, and taste; and E7 measures whether revision improves the artifact. Appendix C.1 gives the complete stage-to-evidence mapping, metric roles, and formulas.

3.4.1 Faithfulness: A Dependency-Aware Criterion Graph

Faithfulness asks whether constraints and workspace evidence are understood, extracted, synthesized, and preserved. Process-side records support deterministic, but final artifacts require criteria over heterogeneous structural and artifact types. We therefore generate a task-specific **dependency-aware criterion DAG** and score it, yielding `E6.FaithfulnessChecklist`.

DAG Structure and Construction. As shown in Figure 3, the DAG has *existence* roots for required objects, *attribute* children for their content and formatting, and *relation* nodes for spatial or organizational structure. These nodes are materialized from human-annotated task-specific schema, fact, numeric, and rubric. We generate 1,405 existence, 7,869 attribute, and 738 relation nodes.

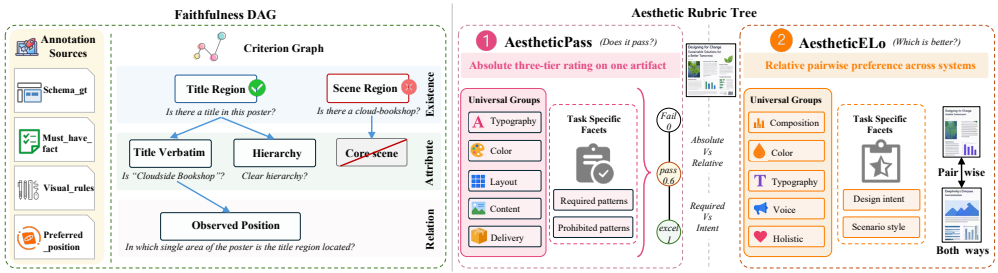

Figure 3. Illustration of the criterion DAG and two aesthetic trees derived for the same poster task.

Judging and gating. Each node is scored using a closed-vocabulary question with a VLM. For relational nodes, the evaluator does not directly receive guiding hints (whether A is to the right of B), but instead directly reads the position from the artifact and compares it with an acceptable set. In score aggregation, descendant nodes with missing existential parent nodes are assigned zero. FaithfulnessChecklist $=$ resolved_yes$/|V_m|$. More details are in Appendix C.2.

3.4.2 Aesthetics: Evaluation of Aesthetic Consensus and Aesthetic Taste

Work-artifact aesthetics mixes professional consensus such as legibility, alignment, and integrity, with open-ended preference in composition, typography, and visual voice. A single score would conflate failure to meet a delivery baseline with legitimate differences in taste. For every task we derive two rubric facet trees (Figure 3) from 5 groups of metrics of a general aesthetic pool. **AestheticPass** asks whether one artifact meets professional requirements, while **AestheticElo** asks which same-task artifact shows stronger visual voice. Full definitions are in Appendix C.3.

AestheticPass score. For each task, an AestheticPass tree is generated based on a pool of general aesthetic metrics, including 5 groups and several facets. In WorkCraft-Arena, we generate **1,919** facets. Next, VLM scores each facets as $\ell \in \{\text{FAIL}, \text{PASS}, \text{EXCELLENT}\}$ or N/Awith a reason, and then maps it to $\varphi(0) = 0.0, \varphi(1) = 0.6, \varphi(2) = 1.0, \varphi(\text{N/A}) = \text{None}$. The pass tier receives 0.6 because it denotes meeting the professional baseline. Facets are averaged within groups. Judge selection and reliability are evaluated in Section 3.4.4 and Appendix D.1.

AestheticElo score. For aesthetic design preferences, AestheticElo generates **1,459** taste facets. Instead of giving an absolute score, VLM determines the winner by comparing them pairwise, where their left and right orders are swapped when comparing in each facet. A tie is declared if the two assessments differ. Bradley–Terry then fits system strength $\pi_s > 0$,

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j}, \qquad \text{Elo}_i = 1500 + 400 \log_{10} \pi_i. \tag{2}$$

The Elo algorithm relies on the participant list and refits when systems change. Therefore, we design an append-only verdict ledger, a symmetric prior on played pairs, and connectivity check. When adding new pairs and evaluation models, new scores and rankings can be quickly refitted. Details are in the appendix C.1. Beyond the two trees, the aesthetics also includes 13 objective aesthetic metrics, and their measurements are given in Appendix E.1.

3.4.3 Score Aggregation: From Metrics to Stage Scores and Two Axes

Stage metrics form a gate tree, and Appendix C.1 details each attribution, threshold, and formula:

$$s = \max(0, \text{mean}(\text{CRITICAL}) - \lambda[1 - \text{mean}(\text{SUPPORT})]) \,. \tag{3}$$

Key leaves define basic creative ability, while support leaves are soft penalties for specific deliverables. Therefore, failure at a key node causes the entire stage to collapse. Faithfulness is the average of nodes for E1-E3, E5, and E6 faithful checks, while Aesthetics is the average of E4, E6 aesthetic pass, and objective aesthetics. AestheticElo remains independent because it is a model-level ranking.

3.4.4 Reproducibility and Meta-Evaluation of the Evaluation System

Before conducting large-scale scoring, we calibrated the VLM judge against three professional designers. We sample 30 anonymous works from five models, covering all artifact types, and select

Table 2. Human meta-evaluation. *Un.* denotes unanimous-case accuracy, *LOO* and *soft* indicate acceptable consistency on human disagreement cases, *all* and *Self-c* indicate pairwise consistency and self-consistency.

Rater	AesthenticPass				AestheticElo		Stability
	un.	LOO	soft	MAE↓	un.	all	Self-c.
GEMINI-3-PRO	**0.855**	**0.663**	**0.932**	**0.494**	**0.652**	**0.571**	**0.851**
CLAUDE-O4.8	0.689	0.566	0.912	0.518	0.568	0.530	0.764
Designer mean	*1.000*	0.536	0.593	0.559	*1.000*	0.537	0.841

nine candidate VLMs with designers to anonymously score the same product render pages and scoring criteria. Task A scored the AesthenticPass levels, while Task B compared taste, with 15% of the works being tested repeatedly in reverse order. Full meta-evaluation design details and complete results are reported in Appendix D.1. Table 2 retains the tests most relevant to the final decision, including self-consistency, model-human agreement, overall pairwise consistency, and order stability. By comparing all candidate models and model combinations, we select GEMINI-3-PRO-PREVIEW, the only candidate that passes the self-consistency test, shows no self-scoring bias, and demonstrates good consistency with humans. It outperforms the designer-set baseline in AesthenticPass and consistency across all paired preference, thus enabling stable and scalable VLM judge.

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS.

All evaluated models use the same interface, model-independent harness instructions, tool protocols, and completion criteria. For each task, the same evaluated model is used across material understanding, artifact creation, and visual self-check, without assistance from external models. We apply fixed interaction, token, timeout, and retry budgets across all runs, while recording execution traces and validation results for reproducibility. Detailed configurations are summarized in Appendix G.1.

4.2 MAIN RESULTS

Table 3 reports not only the two headline axes and arena ranking, but also the seven-stage process and metrics that localise why systems with similar totals fail differently. Figure 4 visualises two aesthetic rankings of the delivered artifacts. We can observe that the top two are separated on taste but almost not on faithfulness. GPT-5-6-SOL and CLAUDE-OPUS-5 differ by 0.001 on the faithfulness axis, while the gap from second to third is 0.114. However, GPT-5-6-SOL wins the taste axis by 97.6, while CLAUDE-OPUS-5 wins the aesthetic consensus axis. Second, all candidates fall within 0.619–0.690 in AesthenticPass, whereas the faithfulness range is **0.361**. We can conclude that models can produce plausible artifact, but making its content correct is not. **Finally, AesthenticElo and faithfulness do not co-order.** MINIMAX-M-3 ranks fourth on faithfulness but third on taste Elo. In addition, we find that product delivery does not perfectly correspond to final product quality. There are models with low delivery rates but higher delivery quality. Including failed delivery entries the final score also changes the model ranking, with more analysis in the Appendix G.2.

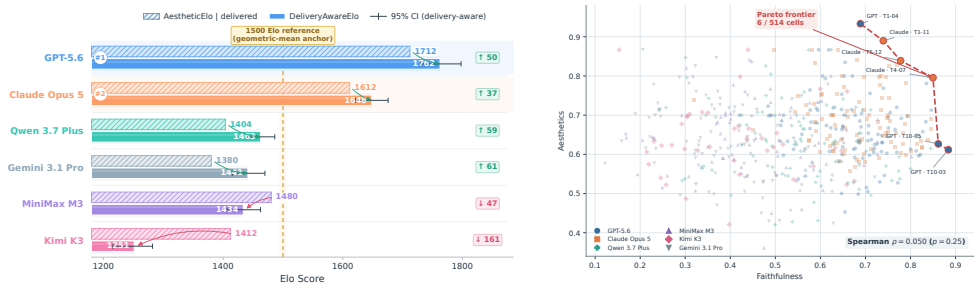

Figure 4. AestheticElo and DeliveryAwareElo as paired bars, and Faithfulness–aesthetics Pareto frontier.

Table 3. Main results. Taste Elo is conditional on both systems delivering; delivery-aware Elo uses one balanced task-level game per pair and counts non-delivery as a loss.

Metric	GPT-5.6-Sol	Claude Opus 5	Qwen 3.7 Plus	MiniMax M3	Kimi K3	Gemini 3.1 Pro
Headline outcomes						
Faithfulness ↑	**0.673**	0.672	0.558	0.416	0.312	0.390
Faithfulness | delivered ↑	0.673	**0.686**	0.558	0.459	0.399	0.394
Aesthetics / AestheticPass ↑	0.661	**0.690**	0.633	0.634	0.621	0.619
Text rendering ↑	0.637	**0.687**	0.509	0.351	0.286	0.344
Taste Elo ↑	**1714.1**	1616.5	1401.6	1477.4	1414.5	1376.0
Delivery-aware Elo ↑	**1762.5**	1648.4	1462.8	1433.6	1251.3	1441.4
Pairwise win rate ↑	**0.798**	0.665	0.336	0.456	0.371	0.299
Delivery rate ↑	**1.000**	0.960	**1.000**	0.770	0.420	0.990
Seven-stage process profile						
E1 Constraint understanding	0.675	**0.716**	0.557	0.428	0.450	0.325
E2 Information extraction	0.521	**0.607**	0.357	0.213	0.176	0.130
E3 Synthesis & visualization	0.627	**0.716**	0.430	0.234	0.200	0.141
E4 Style planning	**0.665**	0.663	0.542	0.545	0.540	0.499
E5 Page planning	**0.617**	0.557	0.468	0.447	0.369	0.302
E6 Final artifact	**0.660**	0.643	0.618	0.428	0.232	0.586
Faithfulness-DAG diagnostics						
Existence pass rate ↑	0.741	**0.789**	0.722	0.735	0.702	0.674
Attribute pass rate ↑	0.709	**0.758**	0.620	0.524	0.556	0.525
Relation pass rate ↑	0.582	0.574	**0.627**	0.570	0.615	0.601
Gated descendants ↓	0.147	**0.107**	0.154	0.128	0.167	0.187
Aesthetic and revision diagnostics						
Delivery readiness ↑	**0.527**	0.522	0.440	0.401	0.420	0.473
Excellent-tier share ↑	0.111	0.111	0.002	0.095	**0.113**	0.065
Edit restraint ↑	0.739	**0.767**	0.580	0.307	0.288	0.632
Self-check net gain ↑	0.478	0.479	**0.482**	0.439	0.455	0.475

Faithfulness and aesthetics are nearly independent. Figure 4 shows all tasks and marks every non-dominated point explicitly. **Across all results the two axes are statistically indistinguishable from independent** ($\rho = 0.050$, $p = 0.25$). This is the strongest single piece of evidence for reporting them as two axes rather than one score. **The system-level Pareto front contains exactly one point**. CLAUDE-OPUS-5 is not worse than any other system on either axis. At the task level only six of 514 points are non-dominated, all belonging to the top two systems and they trace a short arc from $(0.884, 0.612)$ to $(0.689, 0.933)$. In conclusion, a faithfulness–aesthetics trade-off exists. We further validates that AestheticPass and AestheticElo measure different facets and analyze different faithfulness and aesthetic difficulty of different artifact types, as detailed in Appendix G.4 and G.3

Stage-level diagnosis. Stage records reveal that final artifacts and explicit process traces can diverge. As shown in Figure 5, CLAUDE-OPUS-5 is the only system with negative process–artifact fact drift, whereas other systems often deliver facts absent from their recorded extraction. More stage-level results and analysis in Appendix G.6. Based on the faithfulness DAG, and Appendix G.5 provides insights into failure modes of different models at the object, attribute, and relationship in different stages.

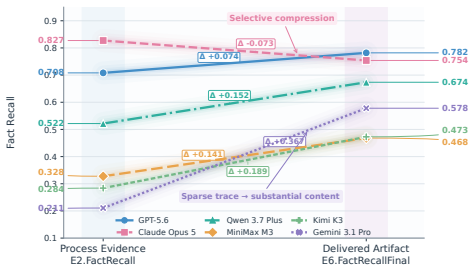

Figure 5. Process–level dissociation in fact recall.

Self-check and revision. Table 4 separates whether revision improves the artifact from whether the agent edits and avoids prohibited changes. Although most systems execute a self-check, only

28% of revisions improve the artifact and 68% make it worse, which indicates that existing models struggle to effectively repair by visual feedback. More strict self-diff ablation are in Appendix G.7.

Table 4. Self-check outcomes comparing artifacts before and after revision.

System	Edit restraint	Patch locality	Prohibited	Net gain	Improved	Worse	Unchanged
GPT-5-6-SOL	0.739 (98)	0.848 (91)	0.945 (85)	0.478 (98)	26	70	2
CLAUDE-OPUS-5	**0.767** (93)	**0.911** (91)	0.947 (84)	0.479 (93)	27	66	0
QWEN-3-7-PLUS	0.580 (97)	0.686 (51)	0.909 (85)	**0.482** (97)	29	64	4
MINIMAX-M-3	0.307 (46)	0.177 (1)	0.921 (66)	0.439 (46)	10	34	2
KIMI-K-3	0.288 (23)	N/A (0)	**0.955** (38)	0.455 (23)	7	15	1
GEMINI-3-1-PRO	0.632 (99)	0.841 (8)	0.900 (84)	0.475 (99)	30	63	6
All systems	N/A	N/A	N/A	**0.473**	**129 (28%)**	**312 (68%)**	**15 (3%)**

Artifact-level cases. Concrete cases in Figure 6 shows why one aggregate ranking is insufficient. Similar total scores can arise from strong grounding with modest polish, widespread non-delivery, or reliable delivery of structurally weaker outputs. Panels (b)–(d) separate selective high-quality delivery, widespread execution failure, and structurally weaker artifacts. Together they connect quantitative leaves to recognizable artifact-level behavior and distinguish end-to-end task utility.

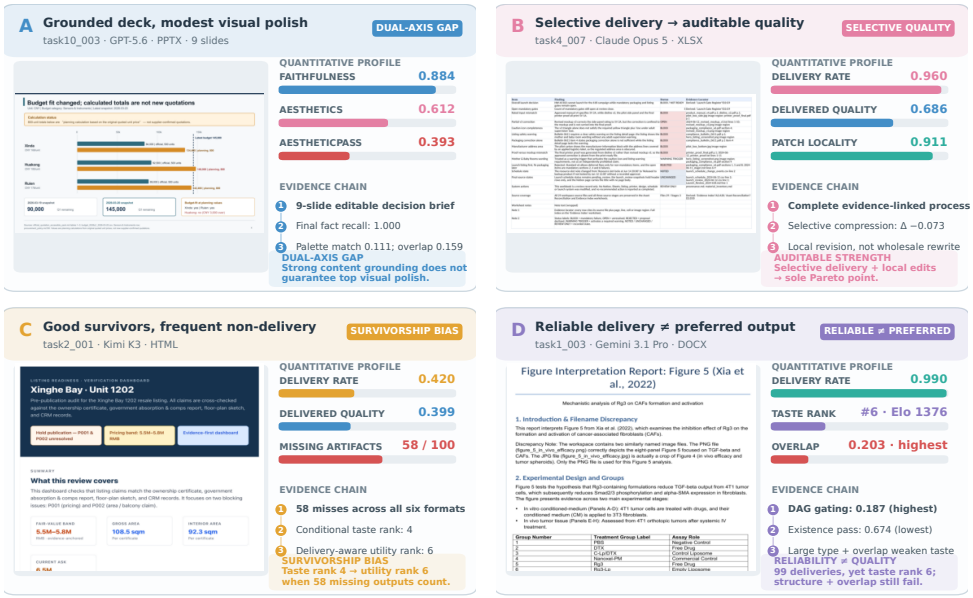

Figure 6. Four concrete cases, illustrated with the actual rendered artifacts.

5 CONCLUSION

We introduce **WorkCraft-Arena**, a 100-task arena for transforming heterogeneous workspace into faithful, professionally designed, and deliverable artifacts. To achieve a unified aesthetic evaluation, it separates factual faithfulness from aesthetics and further distinguishes AestheticPass, an absolute professional bar, from AestheticElo, relative preference among valid designs. The process-level diagnostics further reveal distinct bottlenecks of multimodal language models in understanding, synthesis, design, delivery, and revision: delivery reliability differs from conditional quality, and most measurable self-revisions regress. For better next-generation agent, a shared format-adaptive harness, renderer-aware checks, and versioned records make these capabilities improvable. WorkCraft-Arena thus moves productivity-agent evaluation beyond task completion, which comprehensively locates existing models and agents in delivering reliable products with excellent design aesthetics.

AI USE STATEMENT

LLM-based AI assistants are used for limited writing support, including language polishing, word choice correction, grammar checking, and generating plotting code (e.g., `matplotlib` scripts) for figures from author-specified data and layout instructions. They are not used to originate scientific claims, experimental results, or citations. The specific details of using LLMs to assist humans in data annotation have been explained in the main text and Appendix A.3. The authors verify all technical content, experiments, and manuscript decisions and take full responsibility for the paper.

ETHICS STATEMENT

We introduce the novel WorkCraft-Arena, incorporating a thorough description of data collection, annotation, and filtration processes. We emphasize that the dataset's creation adheres strictly to ethical guidelines. Great care has been taken to uphold ethical standards in the dataset, employing anonymization, desensitization, and data cleaning. The samples pose no risk to public welfare. For all data sourced from open resources, we adhere to open-source licenses for our data usage to ensure compliance with ethical and legal standards. Hence, the innovative research directions and tasks proposed are ethically harmless to society.

REFERENCES

Yuanhao Ban, Tong Xie, Sohyun An, Yunqi Hong, Evan Frick, I-Hung Hsu, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Arena-t2i hard: Benchmarking and improving faithfulness with dependency-aware checklist. *arXiv preprint arXiv:2606.31711*, 2026.

ByteDance Seed, M-A-P, and Humanlaya. Workflow-gym: Towards long-horizon evaluation of computer-use agentic tasks in real-world professional fields. *arXiv preprint arXiv:2606.11042*, 2026.

Tingyang Chen, Shuo Lu, Kang Zhao, Weicheng Meng, Hanlin Teng, Tianhao Li, Chao Li, Xule Liu, Jian Liang, Zhizhong Zhang, et al. Harnessx: A composable, adaptive, and evolvable agent harness foundry. *arXiv preprint arXiv:2606.14249*, 2026.

ClawMark Team. Clawmark: A living-world benchmark for multi-turn, multi-day, multimodal coworker agents. *arXiv preprint arXiv:2604.23781*, 2026.

Adrienne Deganutti, Elad Hirsch, Haonan Zhu, Jaejung Seol, and Purvanshi Mehta. Graphic-design-bench: A comprehensive benchmark for evaluating ai on graphic design tasks. *arXiv preprint arXiv:2604.04192*, 2026.

Shuangrui Ding, Xuanlang Dai, Long Xing, Shengyuan Ding, Ziyu Liu, Jingyi Yang, Penghui Yang, Zhixiong Zhang, Xilin Wei, Xinyu Fang, Yubo Ma, Haodong Duan, Jing Shao, Jiaqi Wang, Dahua Lin, Kai Chen, and Yuhang Zang. Wildclawbench: A benchmark for real-world, long-horizon agent evaluation. *arXiv preprint arXiv:2605.10912*, 2026.

Apurva Gandhi, Vishwas Suryanarayanan, Raja Hasnain Anwar, Firoz Shaik, Shubhang Desai, Thong Q. Nguyen, Muhammad Taqi Raza, Vishal Chowdhary, and Graham Neubig. Ppt-eval: A benchmark for computer-use agents on powerpoint tasks. *arXiv preprint arXiv:2606.31154*, 2026.

Suhaas Garre, Emily Ritchie, Sushant Mehta, and Edwin Chen. Gdp. pdf: Benchmarking grounded multimodal reasoning over professional pdf documents. *arXiv preprint arXiv:2607.11192*, 2026.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, and Trevor Darrell. Autopresent: Designing structured visuals from scratch. *arXiv preprint arXiv:2501.00912*, 2025.

Xinshuai Guo, Jiayi Kuang, Linyue Pan, Yinghui Li, Yangning Li, Hai-Tao Zheng, Ying Shen, Di Yin, and Xing Sun. Evoconfig: Self-evolving multi-agent systems for efficient autonomous environment configuration. *arXiv preprint arXiv:2601.16489*, 2026.

Aspen Hopkins, Allison Nulty, Alexandria Minetti, Anoop Pakki, and Angad Singh. The human creativity benchmark. *arXiv preprint arXiv:2606.30561*, 2026.

Zheng Huang, Xukai Liu, Tianyu Hu, Kai Zhang, and Ye Liu. Pptbench: Towards holistic evaluation of large language models for powerpoint layout and design understanding. *arXiv preprint arXiv:2512.02624*, 2025.

Haonian Ji, Kaiwen Xiong, Siwei Han, Peng Xia, Shi Qiu, Yiyang Zhou, Jiaqi Liu, Jinlong Li, Bingzhou Li, Zeyu Zheng, et al. Clawarena: Benchmarking ai agents in evolving information environments. *arXiv preprint arXiv:2604.04202*, 2026.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Comput. Surv.*, 57(8), March 2025. ISSN 0360-0300. doi: 10.1145/3711680. URL https://doi.org/10.1145/3711680.

Jiayi Kuang, Yinghui Li, Xin Zhang, Yangning Li, Xing Sun, Ying Shen, Philip Yu, et al. Process-level trajectory evaluation for environment configuration in software engineering agents. In *International Conference on Learning Representations*, volume 2026, pp. 113832–113855, 2026.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, and Chenfei Wu. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026.

Haohao Luo, Jiayi Kuang, Wei Liu, Ying Shen, Jian Luan, and Yang Deng. Browsing like human: A multimodal web agent with experiential fast-and-slow thinking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14232–14251, 2025.

Microsoft Research. Mind the gap: Can frontier llms pass a standardized office proficiency exam? *arXiv preprint arXiv:2606.10956*, 2026.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

Michael Ofengenden, Yunze Man, Ziqi Pang, and Yu-Xiong Wang. Pptarena: A benchmark for agentic powerpoint editing. *arXiv preprint arXiv:2512.03042*, 2025.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. 2024.

Tencent WorkBuddy Bench Team, Siqi Cai, Shaopeng Chen, Xiang Fei, Yong Mao, Zihan Xu, Zhiheng Lyu, Zhijian Shao, Yuchen Shi, Shuwen Zhang, et al. Tencent workbuddy bench: A multi-domain coding-agent benchmark with contamination-resistant task construction. *arXiv preprint arXiv:2607.20911*, 2026.

Jingyu Xiao, Yuxuan Wan, Yintong Huo, Zixin Wang, Xinyi Xu, Wenxuan Wang, Zhiyao Xu, Yuhang Wang, and Michael R. Lyu. Interaction2code: Benchmarking mllm-based interactive webpage code generation from interactive prototyping. *arXiv preprint arXiv:2411.03292*, 2024.

Jingyu Xiao, Ming Wang, Man Ho Lam, Yuxuan Wan, Junliang Liu, Yintong Huo, and Michael R. Lyu. Designbench: A comprehensive benchmark for mllm-based front-end code generation. *arXiv preprint arXiv:2506.06251*, 2025.

Yilun Yao, Xinyu Tan, Chao-Hsuan Liu, Yaoming Li, Zhengyang Wang, Wenhan Yu, Zhewen Tan, Yuxuan Tian, Guangxiang Zhao, Lin Sun, et al. Harness-bench: Measuring harness effects across models in realistic agent workflows. *arXiv preprint arXiv:2605.27922*, 2026.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, et al. Osworld2. 0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026.

Haoran Zhang, Luxin Xu, Zhilin Wang, Runquan Gui, Shunkai Zhang, Haodi Lei, Zihao He, Bingsu He, Chicheng Qin, Tong Zhu, et al. pi-bench: Evaluating proactive personal assistant agents in long-horizon workflows. *arXiv preprint arXiv:2605.14678*, 2026a.

Yongheng Zhang, Ziang Liu, Jiaxuan Zhu, Shuai Wang, Xiangqi Chen, Haojing Huang, Jiayi Kuang, Siyu Chen, Ao Shen, Hao Wu, et al. From chatbot to digital colleague: The paradigm shift toward persistent autonomous ai. *arXiv preprint arXiv:2606.14502*, 2026b.

Yuxuan Zhang, Yubo Wang, Yipeng Zhu, Penghui Du, Junwen Miao, Xuan Lu, Wendong Xu, Yunzhuo Hao, Ping Nie, Liang Chen, Dongfu Jiang, Wenhu Chen, and Kelsey R. Allen. Clawbench: Can ai agents complete everyday online tasks? *arXiv preprint arXiv:2604.08523*, 2026c.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

A Benchmark Construction and Annotation

A.1 Professional Role-Domain Taxonomy

WorkCraft-Arena uses ten role domains to vary source evidence, professional norms, target audiences, and artifact conventions while retaining the common requirement of grounded, audience-ready synthesis.

- **Research and knowledge work.** Literature reviews, manuscript drafts, result presentations, and academic posters for researchers and students, emphasizing evidence selection and scientific framing.
- **Business strategy and analysis.** Industry studies, competitor analyses, strategy decks, and decision briefs that turn dispersed signals into structured arguments and recommendations.
- **Data analysis and decision support.** Reports, dashboards, review slides, and workbooks that require reliable aggregation, comparison, calculation, and visualization.
- **Product and user experience.** Product requirements, feature proposals, UX summaries, and launch pages connecting user needs, constraints, and design rationale.
- **Engineering and technical development.** Documentation, benchmark summaries, project sites, and design reviews that communicate technical evidence without losing fidelity.
- **Operations and project management.** Status reports, review decks, process summaries, and trackers organizing plans, risks, milestones, blockers, and actions.
- **Marketing and sales.** Solution decks, campaign pages, posters, and customer collateral balancing factual content with persuasive, audience-aware presentation.
- **Content and creative work.** Infographics, editorial pages, branded reports, and communication assets stressing hierarchy, narrative flow, and coherent style.
- **Professional services and expert consulting.** Advisory reports, due-diligence reviews, compliance briefs, and specialist summaries grounded in evidence and uncertainty.
- **Management, finance, legal, and human resources.** Executive briefings, financial workbooks, legal memoranda, and HR reports requiring traceable reasoning for high-stakes decisions.

A.2 Data Sources and Filtering

We draw candidate scenarios and materials from three source families. For long-horizon scenarios from ClawMark (ClawMark Team, 2026) and WildClawBench (Ding et al., 2026), three expert annotators reviewed roles, workspaces, interaction requirements, and outcomes. They retained instances aligned with the role taxonomy and converted action-oriented objectives—such as messaging, scheduling, or updating a system—into self-contained artifacts that preserve the underlying analysis, planning, or communication task. Scenario information available only in a simulator or

task script was transferred only when necessary to understand the assignment; evidence recoverable from workspace files remained for the evaluated agent to discover.

For interaction-centred environments from Pi-Bench (Zhang et al., 2026a), ClawArena (Ji et al., 2026), and OSWorld V2 (Yuan et al., 2026), we treated local files as candidate evidence rather than inheriting the original action sequence. Annotators prioritized material sets containing substantive local content, multiple evidence-bearing files or modalities, and enough information for analysis, synthesis, and visual organization. A new role- and audience-specific creation brief was then written from what could be verified in the selected workspace, rather than from the source benchmark's interface operations.

For the professional PDFs collected in `GDP.pdf`, annotators indexed each document by topic, role domain, document type, principal content, usable evidence, and potential artifact form. Coherent groups of two to five documents were assembled without reusing a document across tasks. We allowed both within-domain combinations—for example, several records supporting one committee report—and cross-domain combinations supporting risk assessments, management presentations over successive workflow stages, computational workbooks, or visual communication of policy and benefits. A group was accepted only when its documents supported one plausible work context, common audience, and verifiable deliverable, and the instruction was written after coherence check.

All candidates passed a unified filter. We removed tasks dependent on unavailable services or changing online state, tasks dominated by mechanical interface operations, workspaces lacking sufficient evidence, and requirements that could not converge to one professional artifact. We also removed evaluation metadata and incidental logs, checked every file for relevance and accessibility, and rejected unsupported or mutually contradictory requirements. Accepted instances were rewritten as static, self-contained workspaces while preserving the substantive objective and evidence dependencies of the source scenario.

A.3 Annotation Package and Quality Control

Each accepted task follows seven steps: scenario collection; role-domain and artifact assignment; material curation; instruction and constraint drafting; structured ground-truth and reference construction; independent review and adjudication; and final packaging and versioning.

Templates and model-assisted drafting. Before model assistance, human annotators define the purpose, semantics, admissible values, and source-locator conventions of every field for each artifact type. Tasks sharing an output format use the same field schema, with format-specific fields only where medium conventions differ. Human-approved seed tasks span different role domains and workspace compositions and establish the expected evidence granularity, structural detail, and format-specific quality bar. Each `instruction.md` specifies the role, audience, objective, available materials, output format, and delivery requirements, but deliberately omits facts, comparisons, image observations, and derived values recoverable from the workspace. Instructions, annotations, and references follow the workspace's predominant language.

A multimodal model then drafts structured annotations under a prompt containing the benchmark definition, task instruction and workspace, field-level guidelines, the fixed artifact-type schema, and several approved examples of the same format. The draft must preserve field names and nesting, ground factual entries in source materials, and maintain identifiers across files. Human annotators inspect the full workspace and revise every field, verifying claims, numerical values, evidence locators, and cross-file consistency; the model is only a drafting aid, and humans retain authority over task meaning and evidential correctness.

Complementary ground-truth layers. The hidden files are deliberately non-redundant. `constraints.yaml` translates the brief into verifiable delivery and source-use requirements. `must_have_facts.jsonl` links required claims to precise workspace locations, whereas `numeric_gt.jsonl` records quantities that must be parsed or derived from demanding sources such as images, PDFs, tables, charts, and spreadsheets. A value stated directly in the brief is treated as a constraint; a task-relevant value available in ordinary source text is recorded as a fact; neither is duplicated as numeric ground truth.

Table 5. Layered organization of a WorkCraft-Arena task package.

Package layer	Files	Availability
Generation input	`instruction.md`; `workspace/`	Model-visible
Task metadata and scoring	`meta.yaml`; `rubric.yaml`	Hidden
Revision guidance	`self_check_public.yaml`	Revision stage
Evidence ground truth	`constraints.yaml`; `must_have_facts.jsonl`; `numeric_gt.jsonl`	Hidden
Artifact-planning ground truth	`viz_type_gt.json`; `schema_gt.yaml`; `design_gt.yaml`	Hidden
Validation references	`self_check_gt.yaml`; `reference_artifact/`	Hidden

Artifact-planning annotations describe a valid solution space at three levels. `viz_type_gt.json` associates evidence groups with preferred, acceptable, or disallowed visual forms; `schema_gt.yaml` defines an effective content organization and where important evidence should appear; and `design_gt.yaml` captures scenario- and format-appropriate hierarchy, typography, colour, density, and visual language without prescribing one layout. `self_check_gt.yaml` records task-specific diagnostic checks and prohibited failures, while `self_check_public.yaml` exposes actionable revision guidance without hidden answers. Finally, `rubric.yaml` connects instruction following, evidence fidelity, organization, design, and usability to task-specific criteria. Reference artifacts provide one feasible integration of these requirements but remain hidden during generation.

Every package is independently reviewed. Semantic validation checks alignment among the workspace, brief, and intended artifact and detects answer leakage. Evidence validation retraces claims, values, units, tolerances, and calculations to their sources. Structural validation parses all YAML, JSON, and JSONL files and checks schemas, identifiers, and references. Native-file and render validation checks constraints, evidence coverage, language, readability, clipping, overlap, malformed tables, and broken assets. Disagreement affecting more than 10% of fields triggers a third-annotator adjudication.

A.4 Dataset Statistics

Role domains and artifact types . WorkCraft-Arena comprises 100 tasks spanning the ten role domains in our taxonomy and six target artifact types—PPTX, poster, HTML, DOCX, XLSX, and Markdown (Figure 7). The role-domain distribution is tightly balanced, with 9–12 tasks per domain, while every artifact format is represented by 12–21 tasks. This coverage limits the influence of any single profession or output format on aggregate performance and requires evaluated systems to accommodate different evidence structures, audience conventions, and format-specific design requirements.

Input material diversity. Each task's workspace brings together heterogeneous source materials rather than a single document to summarize. Across the 100 workspaces, WorkCraft-Arena contains 2,041 input files spanning 19 file types, an average of 20.4 files per task. Images and PDFs are the most common sources, followed by Python files and Markdown documents; the collection also includes structured and tabular records such as CSV, JSON, JSONL, YAML and XLSX, as well as plain-text files, office documents, HTML, logs, notebooks and databases. This mixture requires systems to locate and combine complementary evidence across files and modalities when constructing a deliverable. In addition to this diversity of input formats, WorkCraft-Arena is bilingual, comprising 79 English tasks and 21 Chinese tasks. For each task, the instruction, annotations, and reference artifacts are aligned with the predominant language of its workspace.

B Agent Harness Implementation

Material understanding. The material-understanding sub-agent combines deterministic parsing with model-based semantic analysis. Parsers preserve exact text, numerical values, tables, page loca-

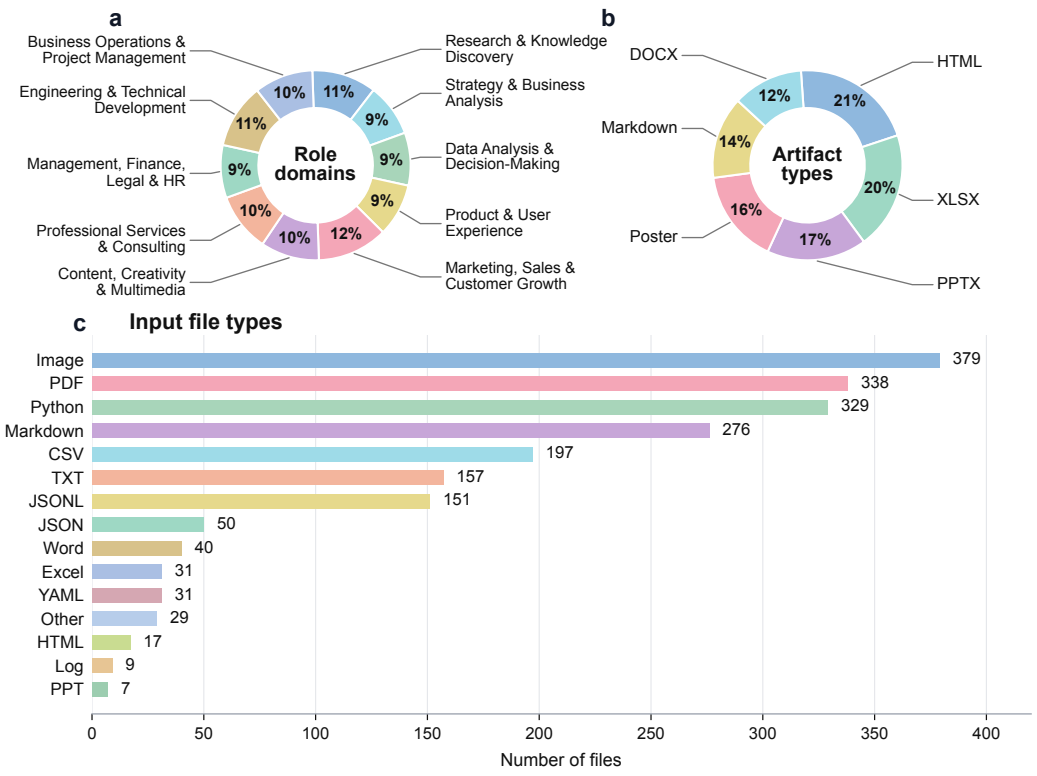

Figure 7. WorkCraft-Arena statistics: 100 tasks across ten role domains and six target artifact types, with heterogeneous input files across the task workspaces.

tions, metadata, and file relationships; rendered-page or region inspection recovers visually encoded evidence unavailable from extraction. The resulting material cards record each source's purpose, key evidence, and locators, while an asset manifest indexes reusable figures, images, and tables with provenance and possible uses. These structures form a compact, traceable map rather than replacing the original files, which remain accessible whenever a claim requires verification. The same evaluated model powers this stage and the main agent; specialization comes only from prompts, tools, and responsibilities.

Iterative artifact creation. The main agent receives the brief, original workspace, material cards, and asset manifest and operates an iterative tool-use loop. It may inspect sources, update artifact code or native files, invoke production tools, and revise earlier decisions when new evidence or rendering feedback reveals a defect. Lightweight records persist the audience, objective, constraints, selected evidence, section/slide/sheet organization, design decisions, and their source grounding. These records provide external state over long trajectories and become the process evidence evaluated in E1–E5. Tool observations, recoverable errors, and validation reports return to the same loop. Completion requires a valid candidate artifact and executed checks, but the harness never exposes private annotations or chooses the narrative, evidence, or visual solution.

Format-adaptive production. A product-adapter architecture reconciles format-specific production with controlled comparison. All six artifact types share the controller, workspace representation, interaction protocol, and completion conditions; each adapter supplies only format-specific prompts, tools, intermediate representations, construction procedures, and checks. PPTX and poster workflows emphasize spatial composition, image preparation, canvas dimensions, and rendering; DOCX and XLSX workflows create editable headings, tables, formulas, charts, and sheet layouts; HTML manages responsive structure and local assets; Markdown preserves portable source organization. Checks cover package and export integrity, formulas and chart references, document structure, miss-

ing assets, and malformed output. Adapter support remains below task semantics and implements model decisions rather than filling content templates.

Render-based self-check. Deterministic checks first detect invalid packages, broken references, missing assets, malformed structures, and formula errors. The first-pass artifact is then rendered as page, slide, sheet, canvas, or browser previews so the same model can detect clipped text, unreadable tables, broken layouts, weak hierarchy, and imbalanced content. The model returns structured issues with observations and repair suggestions, inspects the preserved first-pass artifact and creation context, and performs a focused revision instead of restarting the task. Relevant checks are rerun, and both versions and their context are retained, enabling E7 to distinguish effective repair, no change, and regression.

C Metric Definitions and Score Construction

C.1 The Full Metric Set

Table 6 expands the compact E1–E7 stage, stage-organized specification containing scored and diagnostic metrics, formulas, roles, and axis attribution. Throughout Table 6, C denotes constraint sets, F required facts, T_{final} the text segments extracted from the delivered artifact, N and K the annotated numeric values and derived quantities, G fact groups, R required figures and tables, P and $\hat{P}$ the declared palette and the dominant colours measured from the render, Z layout zones, V_m the measured criterion nodes, $\mathcal{G}$ rubric groups, and J_f the set of judges returning a valid verdict on facet f. The map φ converts a three-tier verdict into a score (Section 3.4.2) and ΔE denotes CIEDE2000 colour difference. The subscript gt refers to the annotation side and run to the agent's process record. Semantic set intersection $A \cap B$ denotes cross-lingual matching by meaning rather than by string equality.

Table 6. Complete E1–E7 metric specification. CRIT. and SUPP. enter stage scores; DIAG. is persisted but unscored. F and A denote the two reported axes, and A* denotes an arena-relative quantity.

Metric	Definition or computation	Role	Axis
E1 Constraint understanding *did the model recover the brief and preserve it?*			
`ConstraintMatch`	Share of annotated hard constraints recovered in the agent's own constraint record: $\lvert C_{gt} \cap C_{run}\rvert / \lvert C_{gt}\rvert$, matched cross-lingually by meaning.	crit.	F
`ConstraintDrift`	Whether intake constraints survive delivery: $1 - \max(0, \text{ConstraintMatch} - \text{ConstraintHitFinal})$.	crit.	F
`IntentRecovery`	Semantic recall of implicit intent tags; retained diagnostically because paraphrase makes it noisier than hard constraints.	diag.	—
E2 Fact and value extraction *was workspace evidence recovered correctly?*			
`FactRecall`	Share of required facts recovered from source materials, $\lvert F_{gt} \cap F_{run}\rvert / \lvert F_{gt}\rvert$; existence gate.	crit.	F
`NumericAcc`	Share of annotated values within per-item tolerance: $\lvert N\rvert^{-1} \sum_i \mathbf{1}[\lvert \hat{v}_i - v_i\rvert \le \tau_i]$, default $\tau_i = 0.05\lvert v_i\rvert$.	crit.	F
`CalculationAcc`	The same test over quantities that must be derived rather than copied; unverifiable rows are skipped.	supp.	F
`EvidenceLocatorAcc`	Agreement in granularity and value between evidence locators of matched facts; locator formats are not comparable enough to score.	diag.	—
E3 Synthesis and visualization *were facts grouped and represented suitably?*			
`FactGroupMatch`	Share of annotated fact groups with at least 60% member coverage: $\lvert G\rvert^{-1} \sum_g \mathbf{1}[\text{cov}(g) \ge 0.6]$.	crit.	F
`RequiredArtifact Coverage`	Share of required charts and tables present, $\lvert R_{present}\rvert / \lvert R\rvert$, detected by fact coverage, title match, or chart-type hit.	crit.	F

continued on next page

Table 6 (continued)

Metric	Definition or computation	Role	Axis
`VizTypeMatch`	Suitability of each visual form, $\lvert G\rvert^{-1}\sum_g s(g)$, with $s \in \{1, 0.7, 0.49, 0\}$ for preferred, acceptable, unlisted, and disallowed.	supp.	F
E4 Style planning *did the visual identity fit the scene and reach the render?*			
`StyleFitScene`, `StyleConsistency`, `StyleExploration`	Three-tier judgements of scenario fit, cross-page coherence, and movement beyond a default template, mapped by $\varphi(\ell)$.	crit.	A
`ColorPaletteMatch`	Share of declared colours recovered by perceptual distance: $\lvert P\rvert^{-1}\sum_{c\in P}\mathbf{1}[\min_{c'\in\hat{P}}\Delta E(c,c') \le 12]$.	supp.	A
`PaletteSizeCompliance`	Whether salient colours stay within budget m: 1 if $n \le m$, else $\max(0, 1-(n-m)/m)$.	supp.	A
`AccentAreaRatio`	Whether accent-colour area lies in the declared interval, with linear decay by violation.	supp.	A
`ObjDominantColorCount`, `ObjFontFamilyCount`, `ObjCrossPageConsistency`	Pixel evidence that the declared identity was realized; ≤ 4 colours, one font family, and consistency ≥ 0.85 are optimal; the last is N/A for one page.	supp.	A
`ContrastRatio`, `ObjHueVariance`, `ObjDeltaE00Color`, `ObjKMeansDominantColors`	Raw or redundant colour observations retained to audit extraction and calibration.	diag.	—
E5 Page planning *was the information architecture complete and balanced?*			
`SchemaCoverage`	Share of required layout zones present in the plan: $\lvert Z_{gt}\cap Z_{plan}\rvert/\lvert Z_{gt}\rvert$, matched by meaning.	crit.	F
`LayoutEntropy`	Diversity of planned page templates, $\min(1, H/H^*)$ with $H=-\sum_k p_k\log_2 p_k$, exposing one repeated template.	crit.	F
`FactDensity`	Facts per page relative to the annotated target, $\min(1,(n_{fact}/n_{page})/d^*)$, exposing hollow plans.	crit.	F
`ImageTextBalance`	Whether figures per thousand words lie in $[0.5, 12]$, with proportional decay outside and zero for no figure.	crit.	F
`LayoutFidelity`, `PlanCreativity`	Recall of recommended layouts and content-tailored planning; diagnostic because neither is a hard faithfulness requirement.	diag.	—
E6 Final artifact: delivery and faithfulness *did the file open, obey the brief, and preserve evidence?*			
`RenderSuccess`	$\mathbf{1}$[opens $\wedge$ (pixels render $\vee$ text extracts)]; existence gate.	crit.	F
`ConstraintHitFinal`	n_{pass}/n_{check} over final aspect ratio, resolution, length, page, and figure constraints.	crit.	F
`FactRecallFinal`	$\lvert F_{gt}\cap T_{final}\rvert/\lvert F_{gt}\rvert$; existence gate, with $F_{run}\setminus T_{final}$ reported as fact drift.	crit.	F
`FaithfulnessChecklist`	$\sum_{v\in V_m}\mathrm{resolved}(v)/\lvert V_m\rvert$ after DAG gating.	crit.	F
`ObjImageTextAlignment`	Piecewise semantic similarity: ≥ 0.30 aligned and ≤ 0.10 unrelated.	supp.	F
E6 Final artifact: aesthetics and arena *did it meet the professional bar, and was it preferred?*			
`AestheticPass`	$\lvert\mathcal{G}\rvert^{-1}\sum_g \lvert g\rvert^{-1}\sum_{f\in g}\lvert J_f\rvert^{-1}\sum_{j\in J_f}\varphi(\ell_{jf})$.	crit.	A
`AestheticElo`	$1500 + 400\log_{10}\hat{\pi}$ from bidirectional same-task preference outcomes.	crit.	A*
`DeliveryAwareElo`	Same map over one task-level game per pair, counting one-sided non-delivery as a loss and two-sided non-delivery as a tie.	arena	A*
`ObjWCAGContrast`, `ObjOCRFontSize`, `ObjWhitespaceRate`, `ObjGridAlignRate`, `ObjLayoutOverlap`	Optimal at contrast ≥ 7, body type ≥ 12 pt, whitespace 25–45%, alignment $\ge 95\%$, and overlap $\le 1\%$.	supp.	A
`ObjLayoutElements`, `ObjVlmCrossCheck`	Region counts and formula–judge gaps retained for instrument diagnosis.	diag.	—
E7 Self-check and iteration *did revision repair defects without uncontrolled rewriting?*			

continued on next page

Table 6 (continued)

Metric	Definition or computation	Role	Axis
`ProhibitedHit`	$1 - n_{hit}/n_{prohibited}$.	crit.	—
`SelfCheckNetGain`	$\text{clip}(0.5 + (S_{after} - S_{before})/2, 0, 1)$.	crit.	—
`ExtraneousEditPenalty`	$1 - c$ for measured changed fraction c.	supp.	—
`PatchLocality`	Full credit when a declared local patch changes at most one third, decaying to zero as the changed fraction approaches one; N/A for self-diffs.	supp.	—

Roles, gates, and artifact-specific penalties. CRITICAL leaves define the principal competence of a stage, SUPPORT leaves can only reduce it through Equation 3, and DIAGNOSTIC leaves never enter a total. Existence-gate failures set their stage to zero, while N/A leaves the denominator and a stage with no measurable critical leaf remains N/A. The support penalty is medium-aware: aesthetic optional items use $\lambda = 0.40$ for posters and HTML pages, where visual design is central, and $\lambda = 0.20$ for DOCX and XLSX, where stricter penalties would over-weight decoration relative to document and computational correctness. Effective task thresholds resolve in the order specified in Appendix E.2.

C.2 Measurement Details of the Faithfulness Criterion DAG

Node materialization and audit records. Existence roots are generated from required zones in `schema_gt` and required visualizations in `viz_type_gt`. Attribute nodes draw from `must_have_facts`, `numeric_gt.jsonl`, `calculation_gt.jsonl`, and faithfulness-oriented items in `rubric.yaml` and `self_check_gt.yaml`. Relation nodes draw from schema relations, `html_tag`, `heading_level`, `layout_id`, and explicit `relations` blocks. Every materialized node stores its criterion, judge instruction, parents, annotation source, criticality, raw verdict, resolved score, and any gate that suppressed it. The 100 task DAGs contain a median of 101 nodes (39–173) and require two to seven vision calls per task under batching.

Provenance distribution. The 10,012 nodes materialized by the pinned instrument over 100 tasks are distributed by annotation source as follows: 2,202 from `must_have_facts`, 1,912 from `numeric_gt`, 1,446 from `rubric.yaml`, 929 from `calculation_gt`, and 1,294 from `schema_gt.zones` together with `required_zones`; the remainder come from `self_check_gt.output_checklist`, `visual_rules`, `minimum_quality_bar`, and `content_balance`. Every node records its source, and the report aggregates pass rates by source. This answers a question the aggregate cannot: whether failures concentrate in the automatically derived structure or in the human-written acceptance criteria; and if a given source is satisfied by every system, it is not discriminative and should be redesigned. Coverage of the relation layer depends on the artifact type: 245 nodes for HTML, 228 for Markdown, 163 for slide decks, and 102 for posters, with all 68 tasks of these four types covered. The 32 word-processing and spreadsheet tasks have no relation layer by design, since these artifacts follow a linear reading order and absolute position is not a property the author controls.

Closed-vocabulary position criteria. Position criteria are phrased as "report the actual position" rather than "confirm the expected position", which departs from common practice and therefore warrants explanation. Asking directly whether "this region is near the top header" embeds the expected answer in the question and thus constitutes a leading question. More seriously, when the annotation is disjunctive—such as `left_column_or_top_left`—any layout satisfies one of its branches and the node is unfalsifiable in principle. The revised procedure is as follows: the judge selects exactly one of ten cells—a 3×3 grid plus the whole canvas as background—and answers only where the region actually is; offline, the evaluator translates the annotation into the set of admissible cells and decides whether the observation falls inside it. Every node records the position the judge actually observed, the admissible set, the chance level $chance_level = |accept|/10$, and the specificity $1 - chance_level$. The relation layer can therefore report a strict pass rate (non-disjunctive annotations) and a permissive pass rate (disjunctive annotations) separately, using the mean chance level as the layer's own noise floor: a pass rate near the noise floor indicates that the annotation does not genuinely constrain the layout. The pinned instrument contains 159 closed-

vocabulary nodes, of which 61 derive from disjunctive annotations, with a corpus-wide mean chance level of 0.351. As a safety rule, if any branch of an annotation cannot be mapped onto the grid we do not construct a closed-vocabulary node at all—otherwise the admissible set would be stricter than the annotation itself and would fabricate failures—but fall back to the yes/no form and flag the node for re-annotation; no such node currently exists in the corpus.

Batched querying and fault tolerance. The two question types are queried in separate groups, because mixing two answer schemas in a single reply induces format drift. Each group is batched at no more than 24 nodes per call, with the output budget set to $max_tokens = \min(2400,\ 200 + 60\,b)$ for a batch of size b, so that the reply cannot be truncated: a truncated reply treated as "all remaining nodes fail" would record an instrument fault as a failure of the system under test. When an entire call fails—through a network error, a parse failure, or an exhausted budget—that batch produces no verdict; after one retry, any node still missing remains unmeasured. Answers outside the vocabulary are recorded as unanswered, and identifiers that appear in a reply without having been asked are discarded. Gating is then resolved breadth-first and the pass rate computed.

C.3 Full Facet Definitions of the Two Aesthetic Trees

Tree generation and composition. Both trees use the same two-level group–facet structure but share no facet. For each task, the generator first prunes a universal pool by artifact type, then adds one task-specific group from human annotations. AestheticPass contributes 12–14 universal facets per applicable artifact and **1,919 facets** overall: 1,440 universal and 479 task-specific. Its task-specific generation prioritizes required visual elements and whether they are competently designed, followed by prohibited visual failures. AestheticElo contributes 11–13 universal facets per artifact and **1,459 facets** overall: 1,240 universal and 219 task-specific. Its task-specific facets derive from scenario, design intent, design principles, and stylistic preference; they ask which result is more compelling while explicitly stating that literal reference copying is unnecessary. This construction keeps professional compliance and open-ended taste as independent signals.

Consensus tree (AestheticPass), universal pool. `TypographyCompliance` contains `Legibility` (is all text legible and free of garbled, overflowing, or clipped characters), `TypeScale` (do headings and body text form a clear, consistent multi-level type scale), and `FontCompliance` (are font families internally consistent and appropriate for the use case). `ColorAccessibility` contains `PaletteDiscipline` (is the palette controlled rather than noisy or garish) and `ContrastReadability` (is text-to-background contrast comfortably readable throughout). `LayoutIntegrity` contains `AlignmentGrid` (are elements aligned to a coherent grid), `VisualHierarchy` (is the information hierarchy unambiguous and easy to scan), and `ElementOverlap` (are all elements free of overlap, clipping, and off-canvas cut-off). `ContentIntegrity` contains `InfographicQuality` (are charts properly labelled with axes, units, and legends where needed; applicable to posters, slide decks, HTML pages, and spreadsheets), `ImageTextIntegration` (are images and text integrated without orphaned images or broken wrapping; applicable to posters, slide decks, HTML pages, and word-processing documents), `DecorativeRestraint` (does decoration support rather than obscure the information), and `NoHallucination` (is the artifact free of broken images, placeholders, fake logos, lorem text, or visibly invented data). `DeliveryReadiness` contains `Completeness` (does the design feel finished rather than partially styled) and `BaselinePolish` (does the artifact meet a professional baseline for delivery in this scenario).

Taste tree (AestheticElo), universal pool. `CompositionTaste` contains `SpatialRhythm` (which design has the more compelling spatial rhythm and pacing), `BalanceTension` (which uses balance, asymmetry, and visual tension more effectively), and `WhitespaceCharacter` (which uses whitespace more intentionally). `ColorTaste` contains `PaletteCharacter` (which has the more distinctive and coherent colour character), `MoodExpression` (which colour treatment expresses the intended mood more compellingly), and `AccentExpression` (which uses accent colour with more effective expressive emphasis). `TypographyTaste` contains `TypographicVoice` (which has the stronger and more appropriate typographic personality) and `EditorialRhythm` (which has the more engaging editorial rhythm across headings and body text). `VisualVoice` contains `Originality` (which feels more original

and less like a generic template), `VisualStorytelling` (which communicates the content through a more compelling visual story; applicable to posters, slide decks, HTML pages, and word-processing documents), and `GraphicExpression` (which integrates imagery, motifs, or graphics more expressively; applicable to posters, slide decks, HTML pages, and spreadsheets). `HolisticTaste` contains `Memorability` (which is more memorable as a complete visual direction) and `SceneInterpretation` (which interprets the scenario and audience more compellingly while remaining appropriate).

C.4 Bradley–Terry Fitting and the Match Ledger

Bradley–Terry strengths are identifiable only up to a multiplicative constant, so their Elo mapping depends on the systems included in the fit. We therefore treat every roster as a versioned arena: ratings are refitted from the complete ledger when a system is added and are not compared numerically across different rosters.

Each valid facet verdict is appended as one auditable row containing task and system identifiers, both presentation orders, winner or tie, judge rationale, render-engine metadata, and digests of the compared images. Historical rows are never overwritten. Adding a system requires only its new pairings before refitting over all rows, so the marginal judging cost is linear in the existing roster size while earlier evidence remains reproducible.

Plain minorization–maximization can silently retain the previous strength when a system has no wins, producing a finite-looking number without meaningful convergence. We prevent this with a symmetric prior of half a virtual draw on pairs that were actually played; no prior is added between systems that never met. We also compute connected components and refuse to emit mutually comparable ratings when the comparison graph is disconnected. Bootstrap resampling produces each system's 95% interval and the probability that one system outranks another.

D Judge Calibration and Human Reliability

D.1 Panel Aggregation and Reliability Reporting

We evaluate nine candidate judges: Gemini-3-Pro-Preview, Gemini-3.1-Flash-Lite-Preview, Gemini-3.5-Flash-Lite, Claude-Opus-4.8, Claude-Sonnet-4.6, Claude-Sonnet-5, GPT-5.5, GPT-5.4, and GPT-5.4-Mini. Table 7 reports the complete matched-protocol comparison. Task A uses 217 items rated by all three designers (64 unanimous and 153 split); Task B uses 340 triply judged comparisons (132 unanimous). Pairwise all/decisive agreement is computed on the 396 comparisons shared with each model judge.

Table 7. Complete judge meta-evaluation. A:un./LOO/soft/MAE/ρ evaluate AestheticPass; B:un./all/dec./tie evaluate AestheticElo; Self-c. is order-swapped consistency. Human unanimous values are definitional.

Judge	A:un.	A:LOO	A:soft	A:MAE↓	A:ρ	B:un.	B:all	B:dec.	B:tie	Self-c.
Gemini-3-Pro	0.855	0.663	0.932	0.494	0.541	0.652	0.571	0.681	0.225	0.851
GPT-5.5	0.694	0.545	0.925	0.523	0.468	0.576	0.500	0.649	0.308	0.764
Claude-Opus-4.8	0.689	0.566	0.912	0.518	0.626	0.568	0.530	0.691	0.391	0.764
Gemini-3.1-Flash	0.683	0.550	0.808	0.693	0.487	0.598	0.535	0.703	0.356	0.675
Gemini-3.5-Flash	0.677	0.524	0.853	0.637	0.451	0.432	0.480	0.693	0.614	0.482
Claude-Sonnet-4.6	0.651	0.527	0.893	0.552	0.621	0.561	0.510	0.651	0.250	0.815
Claude-Sonnet-5	0.500	0.454	0.926	0.590	0.440	0.530	0.505	0.688	0.376	0.758
GPT-5.4	0.459	0.384	0.907	0.616	0.339	0.515	0.472	0.647	0.356	0.721
GPT-5.4-Mini	0.200	0.298	0.913	0.678	0.252	0.311	0.417	0.706	0.710	0.367
Designer A	*1.000*	0.567	0.719	0.523	0.577	*1.000*	0.502	0.730	0.359	0.816
Designer B	*1.000*	0.433	0.490	0.576	0.475	*1.000*	0.543	0.696	0.013	0.868
Designer C	*1.000*	0.608	0.569	0.578	0.493	*1.000*	0.568	0.694	0.076	0.839
Designer mean	*1.000*	0.536	0.593	0.559	0.515	*1.000*	0.537	0.706	0.149	0.841

What each meta-evaluation measure is designed to reject. The ten measures of Table 7 are not redundant: each was introduced to close a specific way in which a judge can look adequate while being unusable. **A:un.** rejects erring where humans see no ambiguity, which is why we treat it

and not the average measures as the admission criterion. **A:LOO** rejects *inflation by asymmetric protocol*: scoring a model against a three-rater mean while scoring a designer against a single other designer lets averaging cancel individual noise, making the model's target easier, and on this data that asymmetry alone raises the model's apparent accuracy by roughly 0.13. **A:soft** rejects answers no professional endorses—given designer tiers $\{0, 1, 1\}$, both 0 and 1 are defensible and 2 is not—but it **must never be read alone**, because the middle tier and the "tie" option are choices a designer may also make, so a judge that abstains frequently scores spuriously high on it: GPT-5.4-MINI reaches 0.913 on A:soft against 0.200 on unanimous items. **A:MAE** rejects the case in which comparable hit rates mask a systematic lean towards one end of the scale, which exact-match scoring cannot detect. **A:ρ** rejects roughly correct tiers that nevertheless order the artifacts wrongly, and is the measure closest to the published quantity. **B:un.** is the pairwise analogue of A:un. **Self-c.** rejects a judge that is reading presentation order rather than design, and because the human value is computed from the swapped duplicates with exactly the same formula, it is the only strictly like-for-like column in the table. For A:ρ, the number of the ten facets reaching $p < 0.05$ is six for the pinned judge and six for each designer.

Panel search. On the taste axis the single pinned judge attains 0.652 on the unanimously decided comparisons, above all thirteen pair and triple combinations, whose best triple reaches 0.568. The mechanism is visible in the abstention rate, which rises from 0.098 for the single judge to 0.402 for pairs: an even-sized panel cannot break a tie among its own members, so member disagreement becomes abstention, and an abstention contributes nothing to a Bradley–Terry fit. An odd-sized panel can break ties but averages the strongest member's correct verdict with weaker members' incorrect ones. On the consensus axis the best triple improves leave-one-task-out rank correlation over the single judge by 0.061, which is below its own across-task standard deviation of 0.185, and it is worse on the unanimous items (0.762 against 0.855).

D.2 Human Evaluation

Before any model is compared against designers, the designers themselves have to be characterised, because the reliability of the human reference bounds every agreement figure computed against it. Three professional designers independently completed the rating package described in Section 3.4.4, returning 780 tier ratings for Task A (44 marked not applicable, 736 usable) and 1,346 pairwise verdicts for Task B. We report their internal consistency only; the absolute quality scores they assigned to individual systems are not used, since the purpose here is to establish a measurement floor rather than a leaderboard.

Table 8. Human reliability of the evaluation instrument. (a) Pairwise inter-designer agreement. "Decisive" restricts Task B to comparisons on which neither designer answered tie; "any tie" is the share on which at least one did. (b) Individual scale use and repeatability. "Extreme" is the share of Task A ratings at Fail or Excellent; self-consistency is measured on swapped duplicates.

(a) Inter-designer agreement

		Task A: tiers			Task B: pairs		
Pair	n_A	**Exact**	**Within 1**	n_B	**All**	**Decisive**	**Any tie**
A–B	255	0.518	0.965	396	0.477	**0.731**	0.361
A–C	221	**0.570**	0.891	340	0.527	0.728	0.329
B–C	217	0.415	0.954	340	**0.609**	0.660	**0.082**
Mean	N/A	**0.501**	**0.937**	N/A	**0.538**	**0.706**	N/A

(b) Individual scale use and repeatability

Designer	**Task A usable**	**Extreme tiers**	**Task B verdicts**	**Tie rate**	**Swapped pairs**	**Self-consistency**
A	260	0.72	472	0.359	76	0.816
B	255	0.40	472	**0.013**	76	**0.868**
C	221	0.93	471	0.076	62	0.839
Mean	245.3	0.68	471.7	**0.149**	71.3	**0.841**

Table 8 reports the central result of this subsection: three professional designers assign the *identical* tier to the same artifact only 50.1% of the time (Krippendorff's $\alpha = 0.479$), yet agree to within one tier 93.7% of the time. On the pairwise task, agreement is 53.8% overall but 70.6% once both sides commit to a winner. The difference is driven by heterogeneous abstention: the mean individual tie rate is 14.9%, while 33.2% of the 340 triply rated comparisons contain at least one tie. Read together, the two tasks show that **the disagreement is concentrated in how strongly to express a judgement, not in which artifact is better**. The per-designer breakdown makes the same point from the other direction. Their use of the three-tier scale differs sharply: designer C places 93% of ratings at the extremes and only 7% in the middle tier, whereas designer B places 60% in the middle tier and 40% at the extremes; tie rates on Task B range from 0.013 to 0.359. Consequently the pair whose scale usage differs most (B–C) shows the lowest exact agreement of the three pairs (0.415) while showing near-ceiling within-one-tier agreement (0.954).

Finally, the 15% of Task B pairs that recurred with the two sides swapped give the noise floor of the taste axis. Asked the same comparison a second time in the opposite order, a designer repeats their own earlier verdict **84.1%** of the time (per designer: 0.816, 0.868, 0.839; 62–76 duplicated items each). Since a professional's judgement is itself only this repeatable, no judge model should be expected to exceed 0.841, and any model–human agreement figure must be read against this ceiling rather than against 1.0.

E Objective Measurement and Threshold Calibration

E.1 Measurement and Calibration of the Objective Aesthetic Parameters

Measurement procedure and calibration. Dominant colours are extracted by k-means in Lab space with $k = 12$, followed by merging clusters that are perceptually identical under a CIEDE2000 distance of at most 8, and finally by retaining the salient colours covering at least 2% of the canvas; k-means in RGB space would split dark colours and merge light ones, because Euclidean distance there does not correspond to perceptual difference. The k-means step uses fixed-stride seeding and no random seed, so that re-evaluating the same artifact yields the same value—otherwise the pinning mechanism would be pointless. Hue statistics use only pixels with saturation at least 0.15 and compute a saturation-weighted circular standard deviation, since hue is an angular quantity for which an arithmetic variance would treat red at 0° and 360° as maximally different, and since the hue of a grey pixel is noise. The whitespace rate treats the largest colour cluster as the background and measures the share of pixels perceptually indistinguishable from it; this is an upper bound, since a large flat photograph is also counted, and is therefore reported together with its method rather than as a conclusion. Cross-page consistency takes the symmetric nearest-colour CIEDE2000 between per-page palettes and maps it linearly onto $[0, 1]$, with single-page artifacts recorded as N/A rather than as a perfect score. Overlap is computed per page, because two shapes on different pages cannot overlap and pooling them would dilute the overlap rate of a long deck to nearly zero, making "only one page broke" look clean. Alignment is judged relative to other elements—sharing a left, right, top, or bottom edge or a centre line, within a tolerance of 1% of the canvas—rather than against a hypothetical twelve-column grid, which would penalize a perfectly aligned five-column layout; pages with fewer than three elements are excluded, since the "alignment rate" of two boxes is arithmetic noise.

Formula-versus-judge cross-check. Four objective parameters of E6 ask the same questions as four facets of the consensus tree, only by different means: element overlap corresponds to `LayoutIntegrity.ElementOverlap`, grid alignment to `AlignmentGrid`, WCAG contrast to `ColorAccessibility.ContrastReadability`, and body font size to `TypographyCompliance`. We compare the two per task and record the absolute gap: agreement constitutes convergent-validity evidence for the judge panel, while disagreement pinpoints exactly the facet the panel cannot see. This also explains why the objective parameters are not merged into the consensus tree. Once merged, the formula terms would become part of the object of comparison and this independent validity evidence would vanish.

E.2 Threshold Resolution Order

Metrics involving thresholds resolve them in the order: the task's own declaration, then the global configuration, then the code default. In the pinned instrument, 80 of the 100 tasks carry their own thresholds, covering 14 threshold items in total; `fact_recall_min` appears in 66 tasks, `max_section_ratio` in 35, `numeric_accuracy_min` and `calculation_accuracy_min` in 29 each, `sheet_coverage_min` in 18, and `formula_error_count_max` in 17. Every task-level override is written into the scorecard.

F Reproducibility and Measurement Discipline

F.1 Pinning and Reproducibility of the Evaluation Instrument

The criterion DAG, the two rubric trees, and all effective thresholds are derived at run time from the annotation and the code. During methodological iteration this is an advantage, but for producing final numbers it creates a serious problem: changing one line of the DAG generation code may change the questions of every task, so that yesterday's 0.62 and today's 0.71 are no longer the same measurement; changing one annotation changes that task's criterion list and invalidates per-task comparison across rounds; and neither kind of change leaves any trace in the scorecard.

We introduce a pinning mechanism: before large-scale evaluation, all DAG nodes of every task—including question text, dependencies, annotation provenance, closed vocabularies, and admissible sets—together with all facets of both rubric trees, the effective thresholds and their sources, an annotation digest, and a code digest, are materialized once into an instrument file, and evaluation reads the frozen values instead of deriving them on the fly. The pinned instrument used in this work contains 100 tasks, 10,012 DAG nodes, 1,919 consensus facets, and 1,459 taste facets.

Pinning must further be paired with two easily overlooked components. The first is the *rendering engine*: rendering the same HTML page with a document-flow rasterizer and with a headless browser produces two different measurements rather than two samples of one, so the scorecard records the rendering engine, match records are separated by engine, and the arena refuses to fit when the artifacts on different sides came from different engines. The second is *caching of judge calls*: semantic matching and criterion judging are not fully deterministic even at temperature zero, which is common with mixture-of-experts models and batched serving. Both text calls and vision calls are therefore cached on the triple of model, full message list, and output budget, and the vision cache is kept in a separate bucket from the text cache so that the two can be invalidated independently—re-rendering should discard only vision entries, not the semantic matching results concerning unchanged text.

F.2 The Measurement Discipline Underlying the Framework

All of the specific rules above follow from a single discipline: N/A and 0 are two different things. A score of 0 means "it was measured and the result was poor" and enters the denominator; N/A means "it was not measured" and leaves the denominator. A criterion node the judge did not answer is recorded as N/A rather than 0; an objective parameter that cannot be measured is recorded as N/A rather than filled with a plausible-looking neutral value; and a panel member abstaining on one facet affects only that facet's denominator, not the denominator of the whole tree.

G Additional Experiment Results and Analysis

G.1 Experimental Configuration

All systems use the same interaction, token, timeout, and retry budgets summarized below.

G.2 Delivery Reliability and Quality

Table 10 shows that the faithfulness axis scores a task with no artifact as 0, because failing to deliver is part of the capability being measured. The consequence is that the faithfulness number is

Table 9. Execution configuration used across all Tasks Design evaluations.

Category	Configuration	Value
Agent budget	Main creation loop limit	180 steps
	Revision loop limit	180 steps
Token budget	Main creation max tokens	32,768 tokens
	Material understanding max tokens	12,000 tokens
	Visual self-check max tokens	12,000 tokens
Timeout	Model request	600 s
Retry	Maximum model-request attempts	5

a composition of capability and delivery rate. Table 10 exposes the two components numerically, and Figure 8 visualises both the score penalty and the rank inversion that appears once delivery is conditioned on. Two systems invert once the components are separated. KIMI-K-3 carries the largest delivery penalty (0.086): 58 of 100 tasks produce nothing at all. But on the 42 tasks it does deliver its faithfulness is 0.399, *higher* than GEMINI-3-1-PRO's 0.394, whose delivery rate is 99%. CLAUDE-OPUS-5 abandons four tasks and, on the remaining 96, overtakes GPT-5-6-SOL on *both* axes (0.686 vs 0.673 faithfulness, 0.690 vs 0.661 aesthetics). The same distinction explains the DeliveryAwareElo shift in Figure 4: conditional taste rewards only the successful subset, whereas the balanced end-to-end metric gives every system 490 games and explicitly charges non-delivery. We retain both metrics because the former measures design quality given delivery and the latter measures practical assignment utility.

Table 10. This table shows that Delivery reliability separated from quality conditional on delivery.

System	Delivered	No artifact	Faith. all	Faith. delivered	Penalty	Aesth. delivered
GPT-5-6-SOL	**100/100**	**0**	**0.673**	0.673	**0.000**	0.661
CLAUDE-OPUS-5	96/100	4	0.672	**0.686**	0.014	**0.690**
QWEN-3-7-PLUS	**100/100**	**0**	0.558	0.558	**0.000**	0.633
MINIMAX-M-3	77/100	23	0.416	0.459	0.043	0.634
KIMI-K-3	42/100	58	0.312	0.399	**0.086**	0.621
GEMINI-3-1-PRO	99/100	1	0.390	0.394	0.004	0.619

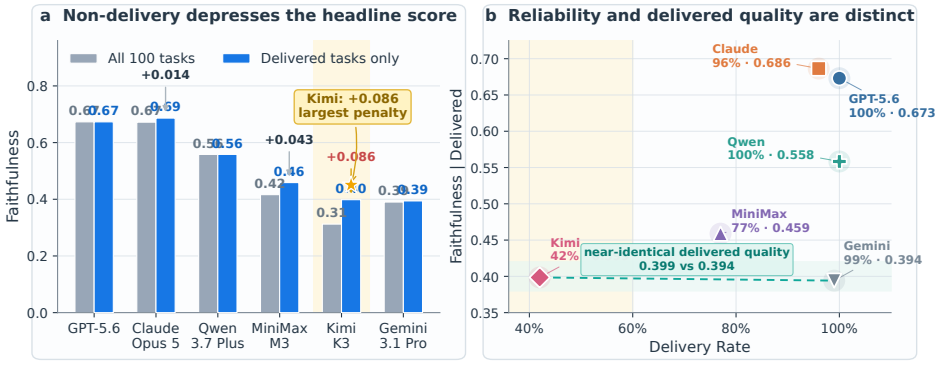

Figure 8. Delivery reliability and delivered quality are distinct. (a) Non-delivery depresses the all-task faithfulness score; labels above the bars show the resulting penalty. (b) A scatter profile maps delivery rate against faithfulness conditional on delivery: KIMI-K-3 and GEMINI-3-1-PRO attain similar delivered quality despite delivery rates of 42% and 99%, respectively.

G.3 Artifact-Family Analysis

For different artifact, Table 11 jointly reports difficulty, cross-system discrimination and pipeline reliability, and Figure 9 exposes the underlying system-by-artifact pattern rather than only its marginal means. We observe that `html` is the easiest family for faithfulness (0.575) and the second hardest for aesthetics. The content drops straight into the DOM, but the resulting layout freedom makes it easy to produce something ugly. On the other hand, `pptx` is hardest overall (lowest aesthetics at 0.585 and lowest E6 at 0.431). That's because slides must respect a per-page information budget *and* cross-page consistency. `poster` has the most extreme ratio (8.02): its aesthetic spread is only 0.055, which indicates that six systems produce posters of near-identical visual quality, while the faithfulness spread is 0.440. `excel` has the *highest* aesthetic mean (0.702) and a low faithfulness mean (0.475), because spreadsheet layout is supplied by the application and the model only fills in content.

Table 11. Artifact-family difficulty, dispersion, and pipeline reliability. "Span" is the best–worst system difference, and ratio is faithfulness span divided by aesthetics span. The last column repeats the within-family dual-axis correlation from Table **??**.

		Mean		Cross-system dispersion			Pipeline			Axis
Family	n	**Faith.**	**Aesth.**	**F span**	**A span**	**Ratio**	**E6**	**E7**	**Delivery**	ρ
DOCX	12	0.438	0.656	0.405	0.136	2.97	0.517	0.593	0.875	0.216
Excel	20	0.475	0.702	0.392	0.083	4.71	0.501	0.546	0.833	0.085
HTML	21	**0.575**	0.593	0.324	0.083	3.89	**0.594**	0.601	0.873	0.258
Markdown	14	0.496	**0.719**	0.312	0.093	3.34	0.585	0.581	**0.905**	0.229
Poster	16	0.509	0.632	**0.440**	**0.055**	**8.02**	0.534	0.648	0.854	0.180
PPTX	17	0.497	0.585	0.393	0.185	2.13	0.431	**0.660**	0.814	**0.390**

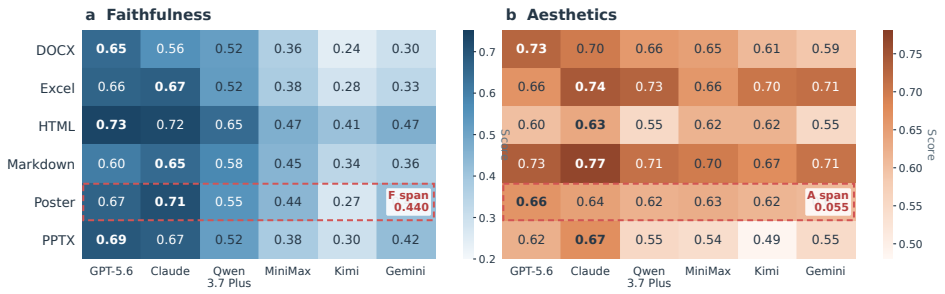

Figure 9. Compact system-by-artifact heatmaps for the two axes. Bold cells are the best system within each family; the highlighted Poster row shows the largest faithfulness span (0.440) but the smallest aesthetics span (0.055).

G.4 AestheticPass versus AestheticElo

Table 12 compares the system ordering induced by the professional consensus rubric with the ordering induced by direct taste preference. Four of six systems change rank between the two protocols. CLAUDE-OPUS-5 is better at passing statable compliance items one at a time, and GPT-5-6-SOL is overwhelmingly preferred when two artifacts for the same task are placed side by side (win rate 0.798). In addition, the consensus axis has a range of **0.070** while the taste axis spans **338 Elo points**, indicating that meeting the professional floor is a low-resolution measurement while taste is what separates these systems.

G.5 Faithfulness-DAG Failure Profiles

Here we discuss the most common loyalty failure patterns. Table 13 separates missing objects, incorrect attributes, and incorrect relations while exposing how much of each task was pruned by gating. The gating rate separates two failure profiles that a single score cannot. CLAUDE-OPUS-5

Table 12. Professional compliance and relative taste induce different rankings (Kendall $\tau = 0.733$). AestheticPass is an absolute three-tier rubric score; AestheticElo is a task-conditional pairwise preference fit. Green marks the leader under each protocol, and $\Delta r = r_{\text{Elo}} - r_{\text{Pass}}$.

	Professional consensus		Relative taste		
System	**AestheticPass**	**Rank**	**AestheticElo**	**Rank**	**Δr**
GPT-5-6-SOL	0.661	2	**1714.1**	**1**	-1
CLAUDE-OPUS-5	**0.690**	**1**	1616.5	2	$+1$
QWEN-3-7-PLUS	0.633	4	1401.6	5	$+1$
MINIMAX-M-3	0.634	3	1477.4	3	0
KIMI-K-3	0.621	5	1414.5	4	-1
GEMINI-3-1-PRO	0.619	6	1376.0	6	0

gates only 10.7% of nodes and leads both the existence and attribute layers, so **its common failures are failures of detail**. GEMINI-3-1-PRO gates the most (18.7%) and has the lowest existence-layer pass rate, which indicates that **it fails to produce the required object at all**. The **relation layer is the one on which all systems are close**, resulting in spatial and organisational relations between elements being a uniform weakness. The **unanswered rate is 0.0000 for every system**, confirming that the vision-call budget was never exhausted and no tail of the DAG was silently left unmeasured.

Table 13. Layer-wise pass rates over *non-gated* nodes. Gating is not an error rate: it is the share of descendants not queried because a parent answered *no*. The final column confirms that no system was truncated by the vision-call budget.

	Pass rate on measured nodes			Measurement path	
System	**Existence**	**Attribute**	**Relation**	**Gated ↓**	**Unanswered ↓**
GPT-5-6-SOL	0.741	0.709	0.582	0.147	0.0000
CLAUDE-OPUS-5	**0.789**	**0.758**	0.574	**0.107**	0.0000
QWEN-3-7-PLUS	0.722	0.620	**0.627**	0.154	0.0000
MINIMAX-M-3	0.735	0.524	0.570	0.128	0.0000
KIMI-K-3	0.702	0.556	0.615	0.167	0.0000
GEMINI-3-1-PRO	0.674	0.525	0.601	**0.187**	0.0000

G.6 STAGE-LEVEL PROCESS AND DELIVERY DIAGNOSIS

We conduct analysis of all seven stage scores. GEMINI-3-1-PRO reaches E6 $= 0.586$ despite E2/E3 of $0.130/0.141$, showing that information absent from its initial process record can appear in later synthesis and presentation. **CLAUDE-OPUS-5 is the only system with a negative gap** (-0.073) and simultaneously the highest `E1.IntentRecovery` (0.893), indicating that it extracts more facts than it finally presents. The systems with near-zero intent recovery fail differently: MINIMAX-M-3 and KIMI-K-3 often omit process records, whereas GEMINI-3-1-PRO misses only one record and delivers on 99% of tasks but still records little checkable intent. Table 14 aligns the fact-recall quantity on the process and delivery sides and reports process observability beside it.

Table 14. Process–delivery dissociation. "Missing" counts cells lacking one or more required process channels.

Metric	**GPT-5.6**	**Claude**	**Qwen**	**MiniMax**	**Kimi**	**Gemini**
E2 Fact recall (process)	0.708	**0.827**	0.522	0.328	0.284	0.211
E6 Fact recall (artifact)	**0.782**	0.754	0.674	0.468	0.473	0.578
Artifact − process, Δ	$+0.074$	-0.073	$+0.152$	$+0.141$	$+0.189$	$+0.367$
E1 Intent recovery	0.800	**0.893**	0.623	0.050	0.034	0.099
Missing-channel cells	2	5	**1**	27	32	**1**
Delivery rate	**1.00**	0.96	**1.00**	0.77	0.42	0.99

G.7 Self-Check and Revision

Strict self-diff ablation. The strict self-diff guard leaves the top four ranks unchanged (Figure 10), including KIMI-K-3 at rank 4. This indicates that the guard corrects an invalid source of credit without manufacturing the main ranking pattern; the only movement is a one-place swap in the bottom pair.

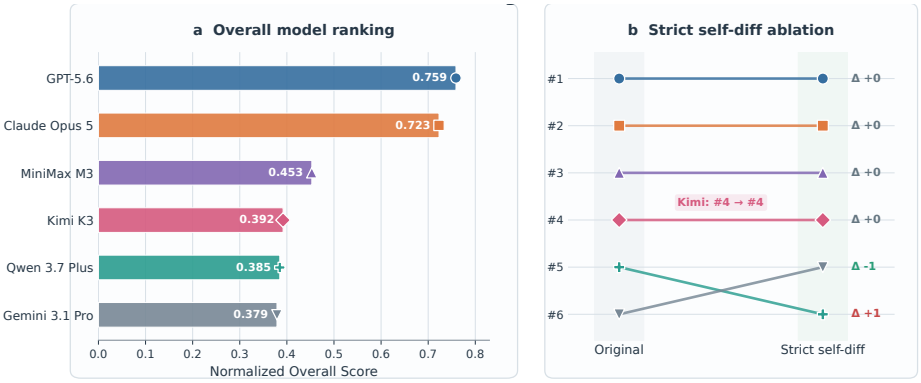

Figure 10. Strict self-diff ablation. The left panel reports the frozen overall ranking; the right panel removes the invalid self-comparisons that would otherwise award perfect edit restraint to a missing final artifact. Model names appear only once on the left, while coloured trajectories identify systems on the right. The top four ranks—including KIMI-K-3 at rank 4—remain unchanged; only the bottom pair swaps.