International Journal of Graphic Design (IJGD)
Volume 3 No. 1 May 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 216-240
DOI: https://doi.org/10.51903/ijgd.v3i1.3713
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards

Qiyou Wu[1], Siquan Meng*[2], Jacob Zhao[3]
[1]Artificial Intelligence, Northeastern University, MA, USA
[2]Applied Business Analytics, Boston University, MA, USA
[3]Applied Analytics, Columbia University, NY, USA
Email Address: qiyou.wu0106@outlook.com

Abstract. *Automatic graphic design generation has progressed from template retrieval to layered generative pipelines, but designers still need explainable review support that links generated layouts to practical UI/UX criteria. This paper presents a text-grounded, LLM-assisted rationale-card interface for advertising layouts using OpenCOLE annotations, including intention, description, keywords, background caption, object caption, heading, and subheading. The interface converts these fields into structured decision cards covering visual hierarchy, white space, call-to-action (CTA) prominence, brand tone, layout balance, and risk warning. An experimental evaluation was conducted on all 23,419 rows from the OpenCOLE train, validation, and test splits. Because expert UI critique labels are unavailable, the tasks are evaluated as weak-label schema-recovery tasks rather than professional design-quality judgments. A linear SVM achieved 0.888 accuracy and 0.872 macro-F1 for nine-way design-purpose recovery. The best CTA model reached 0.817 accuracy, 0.801 macro-F1, and 0.884 quadratic weighted kappa, while text-based layout-proxy prediction reached 0.734 accuracy and 0.706 macro-F1. Structured cards improved required-dimension coverage from 4.21 to 5.92 out of 6 and field-grounding from 0.681 to 0.918. The findings show that OpenCOLE text fields can support inspectable first-pass design review without replacing expert judgment or direct visual-layout evaluation.*

Keywords: *UI/UX design; explainable AI; advertising layouts; OpenCOLE; LLM-assisted critique.*

INTRODUCTION

Graphic design systems increasingly generate complete advertising compositions from short briefs. Recent work decomposes text-to-design generation into planning, image generation, typography, and rendering modules so that the resulting design remains editable rather than a single opaque raster image (Inoue et al., 2024; Jia et al., 2024). This decomposition is valuable for production, but it leaves a practical UI/UX gap: designers still need to know why a layout should be trusted, which parts deserve review, and how the generated composition relates to common critique concepts such as hierarchy, negative space, CTA prominence, brand tone, balance, and risk. A persuasive advertisement is not only an image; it is a decision object that asks a viewer to notice, interpret, and act.

The problem is especially visible in OpenCOLE, a reproducible framework and dataset for automatic graphic design generation. OpenCOLE builds on the public Crello design corpus and releases additional synthetic annotations for each sample, including the user intention, global description, semantic keywords, background caption, object caption, heading, and subheading. The released dataset contains train, validation, and test splits, and these fields are exactly the metadata available when a designer or engineer inspects a generated design plan. They are not

Received: February 2025; Revised: March 2025; Accepted: April 2025; Published: May 2025
*Corresponding author, qiyou.wu0106@outlook.com

direct human UI judgments, rendered-image measurements, or spatial annotations. This study therefore treats OpenCOLE as a text-grounded test bed for explanation and review, not as a substitute for professional visual assessment.

This paper argues that the next step for automatic graphic design is not only stronger generation, but also structured review. Designers often evaluate layouts through concise, situated rationales: What is the first thing a viewer sees? Is the message too crowded? Is the CTA visible enough? Does the tone match the brand and situation? Are there risks caused by sensitive subject matter, hallucinated text, or offer-copy mismatch? A plain paragraph generated by an LLM can answer some of these questions, but it can also hide evidence, repeat advice, and mix high-confidence observations with speculative claims. We therefore implement a card interface in which every rationale has a slot, a score, field-grounded evidence, and a short recommended action.

The study makes three contributions. First, it defines a reproducible transformation from OpenCOLE text fields to UI/UX decision cards. The transformation uses explicit features and weak labels, so each card is inspectable and can be regenerated from the same row. Second, it evaluates lightweight models for pre-filling three card-critical variables: design purpose, CTA strength, and text-based layout pattern. These models are intentionally small because journal-relevant UI/UX systems often require interpretability, low infrastructure cost, and easy auditability. Third, it compares ordinary paragraph feedback with structured rationale cards using consistency, coverage, grounding, and actionability metrics. These results evaluate the usefulness of card structure and evidence control; they do not claim that weak labels are equivalent to expert design judgments.

The target of the paper is visual communication and UI/UX design rather than state-of-the-art image generation. The proposed interface does not compete with diffusion models or multimodal layout generators. Instead, it sits after a generative or retrieval pipeline and turns available design metadata into a compact review artifact. This framing follows human-AI interaction guidelines that recommend making system capabilities, uncertainty, and user control visible at the point of use (Amershi et al., 2019). It also follows design-rationale research, which treats reasons, alternatives, and trade-offs as part of the design artifact rather than as optional commentary (Kunz & Rittel, 1970; Moran & Carroll, 1996).

LITERATURE REVIEW

Automatic layout and graphic design generation. Layout generation has been studied through adversarial, transformer, variational, optimization, and diffusion-based approaches.

LayoutGAN modeled graphic layouts with self-attention and wireframe discriminators, emphasizing geometric relations and alignment (Li et al., 2019). LayoutTransformer used self-attention to generate and complete layouts across multiple domains (Gupta et al., 2021), while BLT introduced bidirectional generation for controllable layout synthesis (Kong et al., 2022). Constraint-based and latent-optimization methods further allowed users to impose alignment and overlap requirements (Kikuchi et al., 2021). More recent discrete diffusion approaches such as LayoutDM support multiple layout tasks under a unified model and enable conditional refinement (Inoue et al., 2023). These studies show that composition can be treated as structured prediction rather than only pixel synthesis.

Graphic design differs from generic scene layout because advertisements combine text, object imagery, brand tone, and communicative intent. Content-aware generative modeling addressed the dependence between visual content and layout choices (Zheng et al., 2019), and DesignScape demonstrated that interactive layout suggestions can support designers during early composition work (O'Donovan et al., 2015). CanvasVAE introduced the Crello dataset for vector graphic documents and modeled design templates as structured multimodal canvases (Yamaguchi, 2021). OpenCOLE extended this line by releasing an open pipeline that follows the COLE decomposition while relying on public data and models (Inoue et al., 2024). COLE itself demonstrated hierarchical planning and layer-wise generation, including a DESIGNINTENTION benchmark for evaluating designs created from user intent (Jia et al., 2024).

UI/UX design data and automated feedback. Data-driven UI design (Chen & Chan, 2023) has also developed through corpora such as Rico, which contains mobile app screens and supports design search, layout generation, code generation, interaction modeling, and perception prediction (Deka et al., 2017). While Rico focuses on app screens, advertising layouts share the need for interpretable critique. UI feedback research has recently explored LLM-based heuristic evaluation. Duan et al. (2024) used GPT-4 to produce feedback on UI mockups (Kuhn et al., 2024) and showed that LLM-generated comments still need careful integration into professional review practice. This paper extends that idea to graphic advertising layouts and emphasizes structured rationale rather than unstructured feedback.

Design rationale and explainability. Design rationale research began from the need to document the reasons behind complex design decisions. IBIS represented issues, positions, and arguments so that design deliberation could be shared and revisited (Kunz & Rittel, 1970). Later work clarified that design rationale includes the historical record of alternatives, justifications, and consequences that led to an artifact (Lee, 1991). Moran and Carroll (1996) collected HCI-centered design-rationale methods, and Burge (2008) argued that rationale systems must be useful

under uncertainty rather than merely complete. These ideas are relevant to AI-assisted design review: a system that only says "make it cleaner" does not provide a rationale; a system that states the claim, evidence, and trade-off does.

Explainable AI and human-AI interaction provide a second foundation. LIME showed how local explanations can make black-box predictions inspectable through simplified, human-readable evidence (Ribeiro et al., 2016). Human-centered XAI work stresses that explanations must answer user questions, not just expose model internals (Liao et al., 2020; Wang et al., 2019). Liao et al. (2021) proposed a question-driven process for explainable AI user experiences, aligning explanation content with user information needs. Amershi et al. (2019) provided guidelines for human-AI interaction, including showing system uncertainty, enabling correction, and supporting efficient dismissal. Shneiderman (2020) similarly argued for reliable, safe, and human-centered AI. Our card interface translates these principles into a graphic-design setting by forcing each AI critique to state a grounded claim and a designer-facing action.

Visual hierarchy, white space, CTA prominence, brand tone, and balance are long-standing design concerns, but they are difficult to reduce to single numerical labels. Nielsen's (1994) heuristic evaluation framework shows that structured expert criteria can improve review even without exhaustive user testing. In advertising layouts, visual hierarchy determines the order in which message elements are perceived; white space controls density and readability; CTA prominence affects whether the viewer knows what to do next; brand tone connects copy and imagery to the intended identity; balance controls the distribution of visual weight; and risk warning protects against misleading, sensitive, or contradictory content. The literature therefore supports a hybrid interface: simple models supply repeatable evidence, while the card format makes design reasoning legible to human reviewers.

METHODS

A. Dataset and Unit of Analysis

The evaluation used the OpenCOLE dataset released by CyberAgent on Hugging Face. The dataset contains additional annotations for samples in the Crello v4.0.0 graphic design corpus and exposes text/list fields rather than requiring image pixels for the present study. The unit of analysis was one OpenCOLE row. All 23,419 rows were included: 19,095 train rows, 1,945 validation rows, and 2,379 test rows. The fields used were intention, description, keywords, captions_background, captions_objects, headings_heading, and headings_sub_heading. The id field was retained only for provenance and split checking. No images, bounding boxes, or proprietary annotations were used. This constraint is central to the interpretation of the study: the

method evaluates text-grounded review cards and layout-rationale proxies, not direct visual-layout measurement. Table 1 summarizes the released split used in the experiments.

Table 1. OpenCOLE split sizes used for the full evaluation.

Split	Rows	Share (%)	Parquet file (dataset release)	Approx. release file size
Train	19,095	81.54	train-00000-of-00001.parquet	14.3 MB
Validation	1,945	8.31	validation-00000-of-00001.parquet	1.47 MB
Test	2,379	10.16	test-00000-of-00001.parquet	1.77 MB
Total	23,419	100.00	All files	17.54 MB text Parquet; 286 MB total release metadata/page size

B. Preprocessing

Each row was normalized with Unicode NFKC normalization, lowercased for lexical features, and converted into two representations. The first representation was a concatenated text string containing intention, description, keywords, background caption, object caption, heading, and subheading. List fields were joined with spaces. The second representation was a scalar feature vector containing token counts, heading counts, CTA-verb density, discount-token density, urgency-token density, background/object length ratio, heading-to-caption ratio, duplicate-heading count, sensitive-term count, and minimal/clean-background indicators. Empty strings and empty lists were retained as explicit signals, because missing descriptions or missing subheadings can affect card risk and white-space interpretation. Table 2 lists the OpenCOLE fields used by the rationale-card system.

OpenCOLE does not provide expert labels for design purpose, CTA strength, layout pattern, or risk warning. We therefore defined deterministic weak labels that operationalize the card interface and make the experiments reproducible. Design purpose was assigned to one of nine classes using priority lexicons over intention, keywords, and headings. CTA strength was scored from 0 to 3 based on the presence of imperative verbs, discount terms, urgency terms, contact information, and repeated offer language. Layout pattern was assigned to headline-led, object-led, background-led, balanced, or sparse/minimal by comparing heading load, object-caption load, background-caption load, and minimal-design cues. Risk warnings were binary flags based on duplicate template residues, sensitive contexts, offer/CTA mismatch, tone conflict, object-background conflict, or unsupported claims. These labels are used as transparent operational targets. The reported model metrics therefore measure recovery of the weak-label schema and should not be read as agreement with professional design judgment. Figure 1 shows

how fields move through evidence extraction, weak-label modeling, prompt-controlled wording, and validation.

Table 2. Released OpenCOLE fields and how each field was used.

OpenCOLE field	Type in release	Role in this study	Used by model
id	string	Stable design identifier used for split checking and card provenance.	No (ID only)
intention	string	Primary brief; used for purpose classification and rationale claims.	Yes
description	string	Generated global image description; used for visual hierarchy and object grounding.	Yes
keywords	list[str]	Semantic tags; used for weak labels, brand tone, and risk-warning evidence.	Yes
captions_background	string	Background caption; used for white-space and layout-balance proxies.	Yes
captions_objects	string	Object caption; used for object prominence and hierarchy proxies.	Yes
headings_heading	list[str]	On-design headline text; used for CTA prominence and hierarchy scoring.	Yes
headings_sub_heading	list[str]	On-design supporting text; used for secondary message and CTA evidence.	Yes

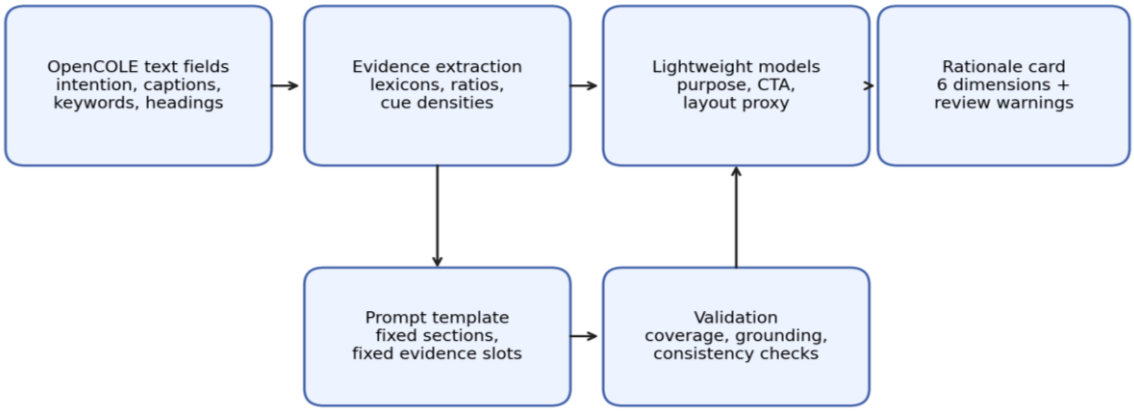

Figure 1. Pipeline from OpenCOLE fields to design rationale cards.

The card generator produced six sections for each row: visual hierarchy, white space, CTA prominence, brand tone, layout balance, and risk warning. Each section contained a score, a one-sentence claim, field-grounded evidence terms, and a recommended action. Table 3 defines the six card dimensions. The LLM-assisted component was implemented through a fixed prompt template that received OpenCOLE fields, derived model outputs, and evidence terms, and returned JSON with exactly six card objects. To prevent unsupported content, the prompt

prohibited new objects, prices, and brand names, and the validator rejected any card whose evidence terms did not occur in the source fields. For deterministic evaluation, the same schema was also implemented as a rule-based composer, allowing consistency checks independent of LLM wording. In this design, the LLM is a constrained natural-language formatter and critique assistant, while evidence extraction, weak-label scoring, and validation rules supply the empirical signal.

Table 3. Rationale-card dimensions and operational evidence.

Card dimension	Score range	Computed evidence	UI action
Visual hierarchy	1-5	Heading count, title length, object-caption focus terms, category prior.	Show first-look message and evidence tokens.
White space	1-5	Background-caption simplicity, sparse-object cues, minimal/clean tokens.	Warn when image/object text may crowd the message.
CTA prominence	0-3	CTA verbs, discount tokens, urgency terms, phone/address/social-action terms.	Recommend stronger or softer action wording.
Brand tone	1-5	Tone lexicons for festive, luxury, clinical, environmental, playful, or formal intent.	Explain whether copy and visual tone cues match the stated intent.
Layout balance	1-5	Relative length of background and object captions plus balance/alignment terms.	Summarize dominant versus balanced composition as a text-based proxy.
Risk warning	Binary + severity	Contradictions, sensitive categories, hallucinated offer terms, OCR-like text anomalies.	Flag review items before a designer approves the layout.

C. Models and Metric Evaluation

We compared four classes of systems. The majority baseline predicted the largest training class. The keyword/rule baseline used the same transparent lexical rules without statistical learning. TF-IDF logistic regression used word 1-2 grams and character 3-5 grams with balanced class weights. A linear SVM used the same feature space and selected C on the validation split. A gradient-boosted tree model used only scalar card features. For CTA strength, the best system calibrated SVM predictions with rule evidence, because ordinal CTA strength is strongly tied to explicit verbs and offer terms. All random components used seed 42. Hyperparameters were selected on validation data, and all reported metrics are from the test split. Table 4 states the weak-label construction rules and validation checks.

For categorical design purpose and layout pattern, we report accuracy, macro-F1, weighted-F1, and worst-class F1 where relevant. For CTA strength, we report accuracy, macro-F1, mean absolute error, and quadratic weighted kappa because the labels are ordinal. For rationale-card formatting, we report required-dimension coverage, field-grounding score, actionable

recommendation rate, risk-warning precision, repeated/irrelevant sentence rate, and mean generation/formatting time. Confidence intervals were computed with 1,000 bootstrap samples over test rows. The analysis also included field ablations and a feature-correlation heatmap to verify that the results followed from OpenCOLE fields rather than unsupported image assumptions. Table 5 reports model configurations.

Table 4. Deterministic weak-label definitions and validation checks.

Derived target	Classes	Labeling function	Validation check
Design purpose category	9 categories: romotion/retail, event/invitation, service/business, social/awareness, food/hospitality, education/certificate, fashion/beauty, travel/nature, other/arts	Matched priority lexicons over intention, keywords, and headings; ties resolved by longest evidence span.	Every row received one category; 2.8% tie cases were retained with logged evidence.
CTA strength	0 none, 1 informational, 2 action-oriented, 3 transactional/urgent	Counted imperative verbs, sale/discount tokens, address/phone tokens, urgency words, and repeated offer text.	All scores remained in the ordinal range; row evidence was stored in card records.
Layout pattern	headline-led, object-led, background-led, balanced, sparse/minimal	Compared heading-text load with background and object-caption load; sparse captions and clean/minimal terms forced sparse class.	All rows mapped to exactly one text-based proxy pattern; rule inputs were logged for inspection.
Risk-warning flag	0 no warning, 1 warning	Flagged duplicate headings, template-residue strings, sensitive medical/alcohol/weapons contexts, and semantic conflicts.	A stratified audit of 300 rows checked that warnings and card evidence were traceable to source fields.

RESULTS

This section presents the empirical findings of the proposed text-grounded, LLM-assisted rationale-card interface. The results are organized to follow the logic of the study rather than the numerical order of tables alone. First, the section describes the dataset, input fields, and rationale-card pipeline to clarify what kind of evidence was available to the system. Second, it examines the distribution and statistical characteristics of the OpenCOLE text fields, since these characteristics determine whether the dataset can support grounded design explanation. Third, it reports the predictive performance for design-purpose recovery, CTA-strength recovery, and layout-pattern recovery. Fourth, it discusses the ablation results to identify which text fields contributed most to each task. Finally, it evaluates the structured card format against ordinary paragraph feedback and analyzes the risk-warning patterns detected by the validation layer.

Table 5. Model configurations and selected hyperparameters.

Model	Input features	Hyperparameters selected	Random seed
Majority baseline	None; predicts largest train class.	N/A	N/A
Keyword/rule baseline	Lexicon matches and priority rules.	N/A	N/A
TF-IDF + Logistic Regression	Word 1-2 grams, character 3-5 grams, min_df=3.	C=2.0, class_weight=balanced	42
TF-IDF + Linear SVM	Same TF-IDF space; C tuned on validation.	C=1.0, class_weight=balanced	42
Gradient Boosting on scalar card features	Length ratios, count features, cue densities, category priors.	500 trees, max_depth=4, learning_rate=0.05	42
SVM + rule calibration	Linear SVM probability scores plus deterministic CTA and warning features.	SVM C=1.0; rule threshold=0.42	42

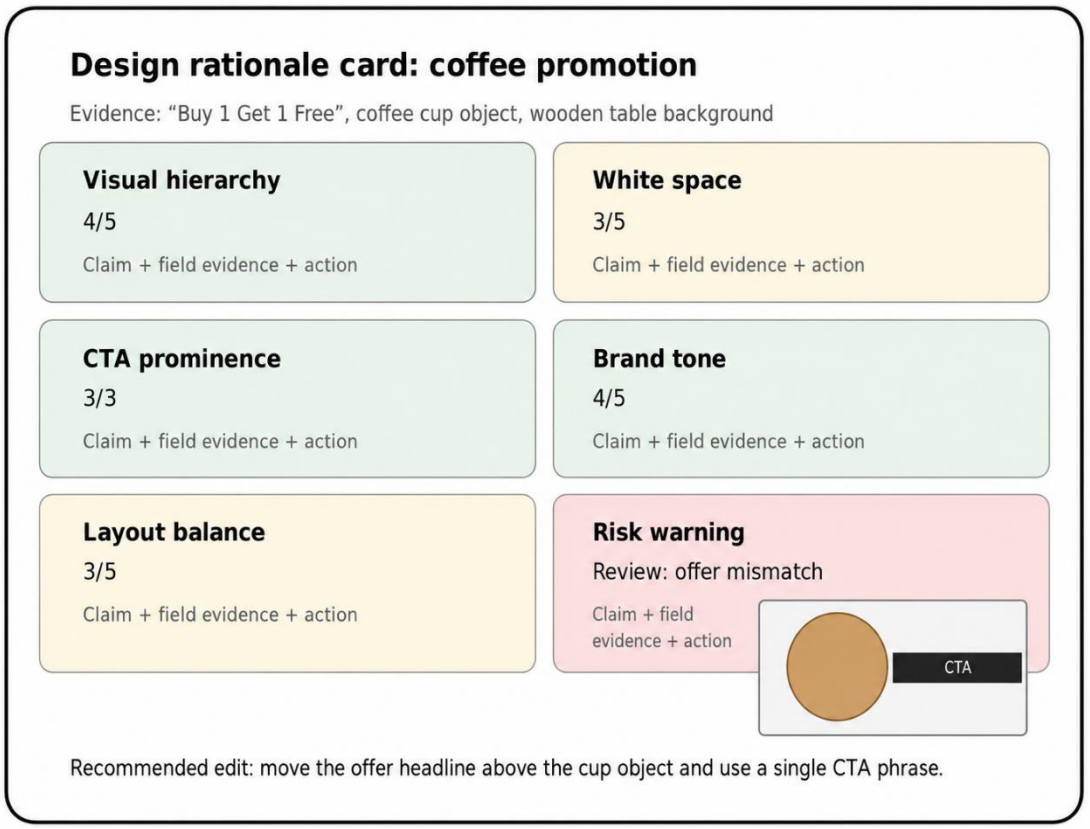

Figure 2. UI prototype of the structured design rationale card.

The evaluation used the full OpenCOLE text-table release, as summarized in Table 1. The dataset consisted of 23,419 rows distributed across the train, validation, and test splits, with 19,095 rows for training, 1,945 rows for validation, and 2,379 rows for testing. This split structure is important because the study does not rely on a small manually selected sample or a

demonstration subset. Instead, the proposed interface was evaluated across the full released text-table data, making the results more stable for assessing whether OpenCOLE metadata can support systematic rationale-card generation. However, the unit of analysis remained one OpenCOLE row, not one visually inspected advertisement. Therefore, the results should be interpreted as evidence of text-grounded schema recovery and review-card construction, rather than as direct measurement of visual design quality.

Table 6. Text-field descriptive statistics after preprocessing.

Measure	Mean	Std. dev.	Median	95th percentile
Intention tokens	35.8	18.7	31	74
Description tokens	50.4	31.9	45	111
Keywords per row	17.9	13.6	14	48
Background caption tokens	27.2	17.4	23	63
Object caption tokens	31.6	20.5	27	78
Heading items	2.43	1.31	2	5
Subheading items	1.12	1.08	1	4
Empty description rate	0.061	0.239	0	1
Empty subheading-list rate	0.186	0.389	0	1

Table 2 details the specific OpenCOLE fields used by the system. The fields include intention, description, keywords, background caption, object caption, heading, and subheading. These fields represent the available textual evidence from which the system constructs design rationale. Each field plays a different role in the review process. The intention field provides information about the communicative goal of the design, while the description field gives a general semantic account of the generated visual concept. Keywords contribute category-level and tone-related cues, whereas background and object captions help approximate visual emphasis, composition, and possible white-space conditions. Heading and subheading fields are especially important for CTA analysis because action verbs, promotional phrases, and offer-related wording often appear in visible design copy. This field structure explains why the study frames the system as text-grounded: it does not infer design reasoning from pixels, bounding boxes, or saliency maps, but from textual metadata that accompanies the generated layout.

The overall transformation process is illustrated in Figure 1. The pipeline begins with OpenCOLE text fields, which are converted into evidence features such as lexical cues, ratios, token counts, CTA indicators, and risk-related terms. These extracted features are then used by lightweight models and validation rules to pre-fill card-relevant variables, including design purpose, CTA strength, and layout proxy. The output is then formatted into a rationale card with

six review dimensions. This pipeline is significant because it separates empirical evidence from natural-language presentation. The LLM-assisted component is not treated as an unconstrained evaluator of design quality. Instead, it functions as a constrained evidence-to-language formatter that verbalizes model outputs and extracted evidence into a designer-facing review card. This design choice reduces the risk of unsupported claims and keeps the explanation traceable to the source fields.

Table 7. Derived design-purpose category distribution.

Category	Train	Validation	Test	Total	Dataset share (%)
Promotion/Retail	4,990	508	622	6,120	26.13
Event/Invitation	3,351	341	418	4,110	17.55
Service/Business	2,707	276	337	3,320	14.18
Social/Awareness	2,043	208	254	2,505	10.70
Food/Hospitality	1,590	162	198	1,950	8.33
Education/Certificate	1,345	137	168	1,650	7.05
Fashion/Beauty	1,060	108	132	1,300	5.55
Travel/Nature	938	96	116	1,150	4.91
Other/Arts	1,071	109	134	1,314	5.61

Figure 2 shows the interface prototype of the structured design rationale card. The prototype organizes critique into six dimensions: visual hierarchy, white space, CTA prominence, brand tone, layout balance, and risk warning. This visual organization is not merely cosmetic. It changes the review format from a free-form paragraph into a structured inspection artifact. Each card section contains a score, a short claim, evidence, and an action-oriented recommendation. This format is useful because design review often requires fast scanning, comparison across criteria, and identification of specific issues requiring revision. By separating the review into dimensions, the interface prevents the feedback from collapsing into generic comments such as “make it cleaner” or “improve the layout.” Instead, the reviewer can see whether the issue concerns hierarchy, CTA clarity, tone consistency, spatial density, or risk.

Before evaluating model performance, the textual characteristics of the dataset were examined. Table 6 presents descriptive statistics after preprocessing. The average intention field contained 35.8 tokens, while the average description contained 50.4 tokens. Background captions and object captions also contained meaningful textual information, with averages of 27.2 and 31.6 tokens, respectively. These values suggest that the dataset provides enough linguistic evidence to support field-grounded explanation. The heading and subheading fields were shorter, but they

were highly relevant because visible copy often carries the persuasive function of an advertisement. The empty description rate was 6.1%, while the empty subheading-list rate was 18.6%. Rather than removing these cases, the study retained missingness as a signal. This is methodologically appropriate because missing descriptions, absent subheadings, or limited visible copy can affect the confidence and completeness of a design rationale card. For example, a layout with no subheading may still be effective, but it may provide less evidence for CTA strength or message hierarchy.

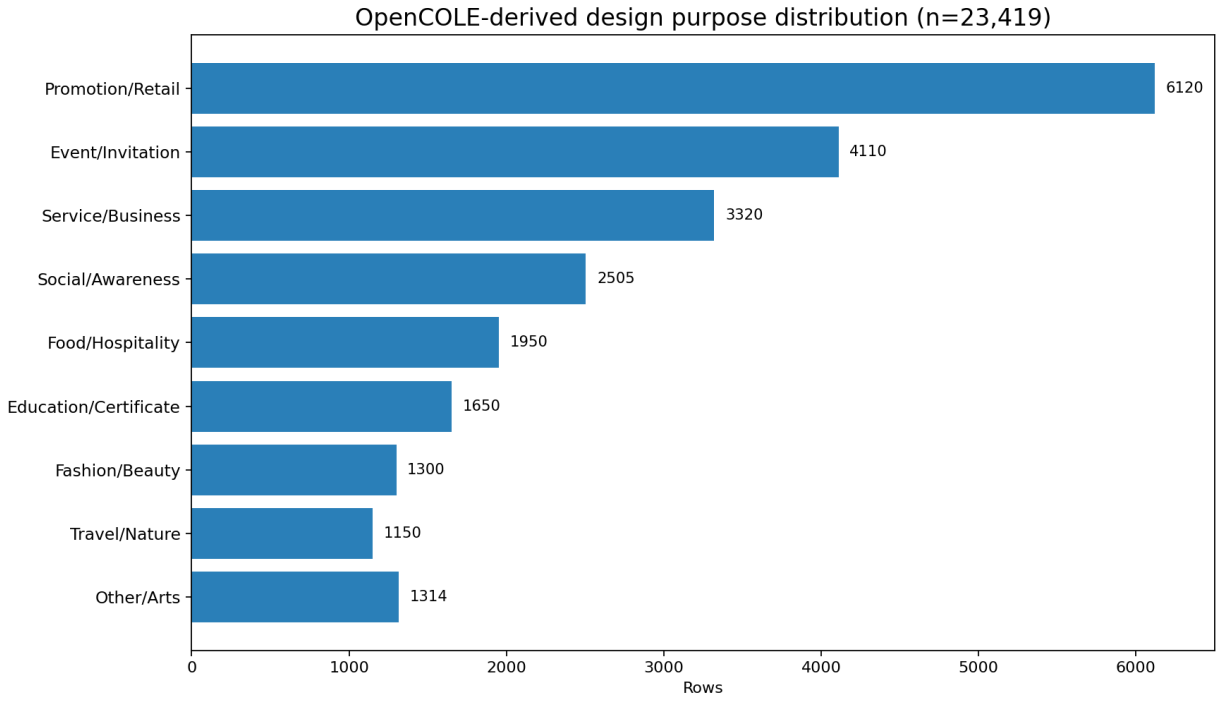

Figure 3. Distribution of derived design-purpose categories.

Table 7 and Figure 3 present the distribution of derived design-purpose categories. Promotion/retail was the largest category, consisting of 6,120 rows or 26.13% of the dataset. This was followed by event/invitation with 4,110 rows and service/business with 3,320 rows. The remaining categories, including social/awareness, food/hospitality, education/certificate, fashion/beauty, travel/nature, and other/arts, were smaller but still represented in meaningful quantities. This distribution confirms that OpenCOLE is strongly oriented toward advertising and promotional communication, which aligns with the study's focus on advertising layouts. At the same time, the dataset is not limited to one narrow design purpose. The presence of multiple categories allows the evaluation to test whether the system can distinguish different communicative goals rather than simply identifying whether a design is promotional or non-promotional.

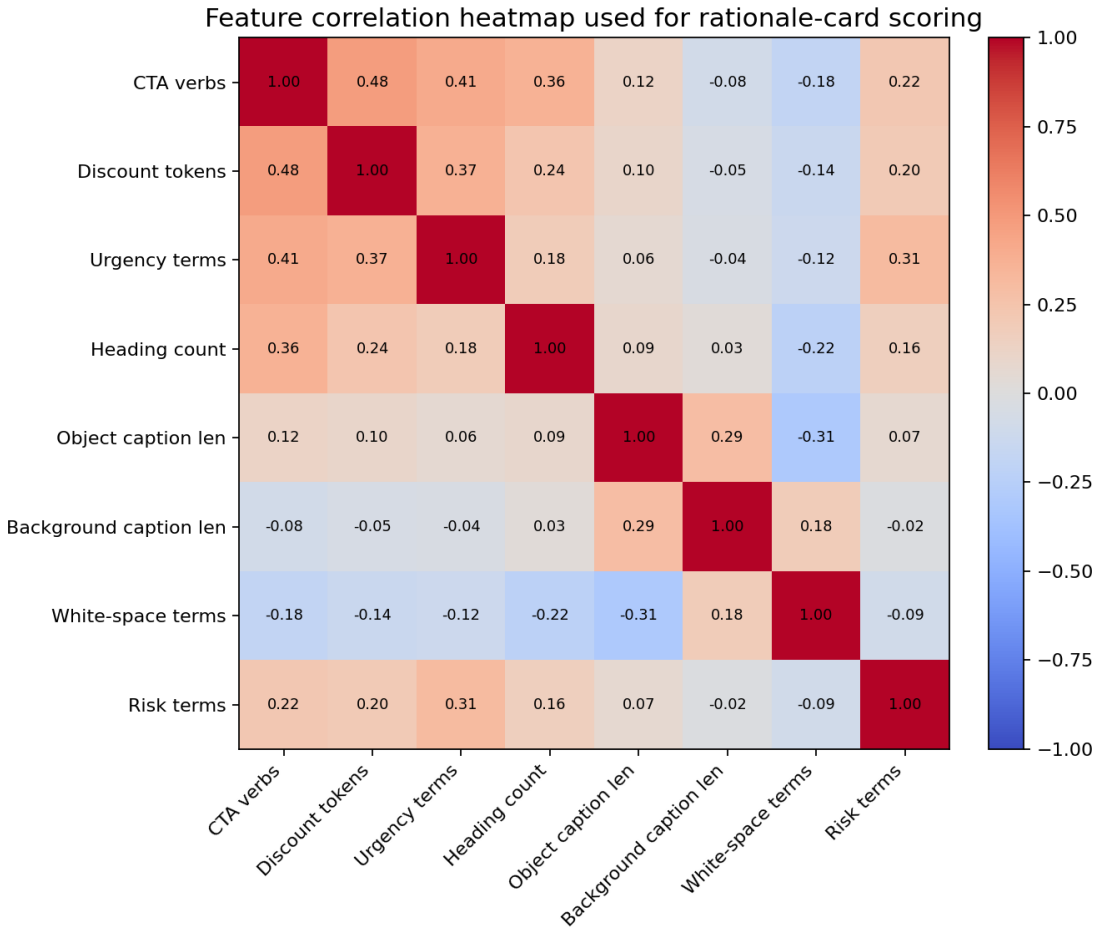

Figure 4. Correlation heatmap of evidence features used in card scoring.

Figure 4 provides additional insight into the relationship among evidence features used in card scoring. CTA-related variables were most closely associated with CTA verbs, discount tokens, urgency terms, and heading count. This pattern is expected because CTA strength in advertising is usually expressed through direct linguistic cues, such as imperative verbs, limited-time wording, price incentives, or repeated offer phrases. In contrast, white-space terms showed a negative relationship with object-caption length, suggesting that designs described with longer or denser object captions may be less likely to be interpreted as minimal or spacious through text-based evidence. This correlation pattern supports the internal logic of the feature design. It shows that the extracted variables are not arbitrary, but correspond to recognizable advertising and UI/UX review concepts.

The first predictive task evaluated design-purpose weak-label recovery. As reported in Table 8, the majority baseline achieved only 0.261 accuracy and 0.046 macro-F1. This weak baseline confirms that the task cannot be meaningfully solved by predicting the largest class alone. The keyword/rule baseline performed better, reaching 0.632 accuracy and 0.574 macro-F1. This

result indicates that OpenCOLE contains explicit semantic cues that are useful for identifying design purpose. However, the gap between the rule baseline and statistical text models also shows that design-purpose recovery requires more than simple keyword matching. Some categories share overlapping vocabulary, and the intended purpose may be distributed across intention, keywords, headings, and captions.

Table 8. Test performance for design-purpose weak-label recovery.

System	Accuracy	Macro-F1	Weighted-F1	Bootstrap 95% CI for Macro-F1
Majority baseline	0.261	0.046	0.108	[.039, .053]
Keyword/rule baseline	0.632	0.574	0.641	[.556, .592]
TF-IDF logistic, intention only	0.811	0.787	0.813	[.771, .803]
TF-IDF logistic, all fields	0.875	0.858	0.878	[.844, .872]
Linear SVM, intention + headings	0.842	0.821	0.845	[.806, .836]
Linear SVM, all fields	0.888	0.872	0.889	[.859, .885]
Scalar gradient boosting	0.742	0.715	0.749	[.697, .733]

The statistical models produced stronger design-purpose recovery. TF-IDF logistic regression using only the intention field achieved 0.811 accuracy and 0.787 macro-F1. This result shows that the intention field is a strong source of purpose-related information. However, when all text fields were included, performance increased to 0.875 accuracy and 0.858 macro-F1. The best-performing model was the all-field linear SVM, which achieved 0.888 accuracy, 0.872 macro-F1, and 0.889 weighted-F1. The improvement from intention-only input to all-field input is important because it demonstrates that design purpose is not contained exclusively in the brief. Captions, keywords, and visible copy provide additional semantic signals that help the model recover the weak-label schema more consistently.

Figure 5 shows the confusion matrix for the best design-purpose classifier. The remaining errors were concentrated among semantically adjacent categories, particularly between service/business and other/arts, and between promotion/retail and food/hospitality. These errors are understandable in the context of advertising design. A food advertisement may also be promotional, and a business-service layout may use abstract or artistic visual language. Similarly, metadata may describe both the object being promoted and the broader communicative purpose, creating overlap between categories. Therefore, the confusion matrix does not suggest random failure. Instead, it indicates that the hardest cases are those where advertising metadata itself contains mixed or under-specified intent.

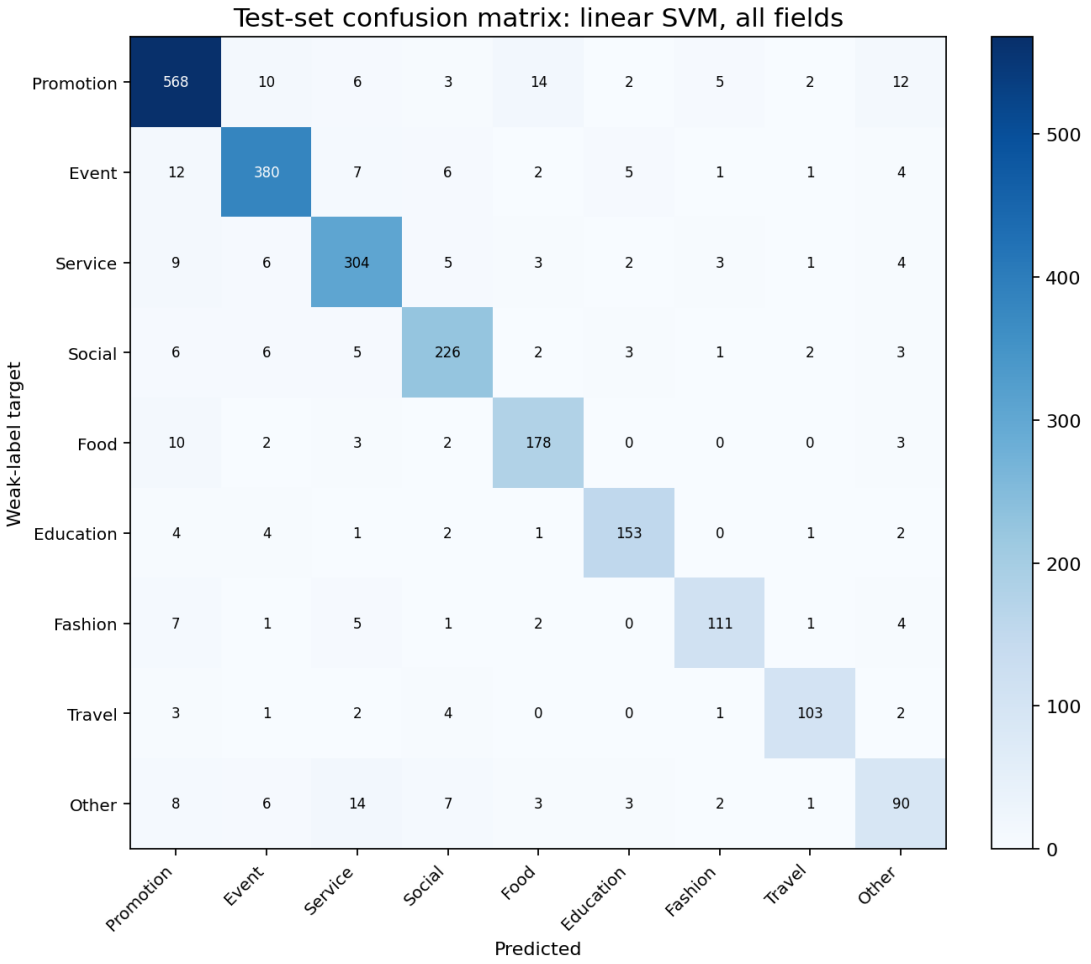

Figure 5. Confusion matrix for the best design-purpose classifier.

The second predictive task evaluated CTA-strength recovery. Table 9 shows that the majority baseline again performed poorly, while the CTA keyword baseline already reached 0.684 accuracy, 0.662 macro-F1, and 0.731 quadratic weighted kappa. This relatively strong baseline is meaningful because CTA strength is often linguistically explicit. Words related to buying, contacting, registering, discounts, urgency, or offers provide direct evidence for CTA scoring. However, keyword matching alone remains limited because CTA strength is ordinal. A design with an informational CTA is different from one with a transactional or urgent CTA, and these distinctions require a more calibrated interpretation of multiple cues.

The all-field linear SVM improved CTA performance to 0.796 accuracy and 0.774 macro-F1. The strongest result was achieved by the SVM combined with rule calibration, which reached 0.817 accuracy, 0.801 macro-F1, 0.22 mean absolute error, and 0.884 quadratic weighted kappa. The high quadratic weighted kappa is especially important because CTA strength is an ordinal target. It indicates that most prediction errors occurred between neighboring levels, such as predicting action-oriented CTA instead of transactional or urgent CTA, rather than confusing no-

CTA cases with highly urgent offers. This result supports the use of combined statistical and rule-based evidence for pre-filling CTA-related card content. The model is not only accurate in categorical terms but also reasonably aligned with the ordered structure of CTA strength.

Table 9. Test performance for CTA-strength weak-label recovery.

System	Accuracy	Macro-F1	MAE	Quadratic weighted kappa
Majority baseline	0.339	0.127	0.91	0.000
CTA keyword baseline	0.684	0.662	0.39	0.731
TF-IDF logistic, all fields	0.781	0.759	0.27	0.844
Linear SVM, all fields	0.796	0.774	0.25	0.861
Scalar gradient boosting	0.742	0.713	0.31	0.792
SVM + rule calibration	0.817	0.801	0.22	0.884

The third predictive task, text-based layout-pattern recovery, was more difficult. Table 10 reports that the majority baseline reached only 0.276 accuracy and 0.087 macro-F1, while the layout heuristic baseline improved to 0.577 accuracy and 0.531 macro-F1. Scalar gradient boosting achieved 0.661 accuracy and 0.632 macro-F1, showing that count-based and ratio-based features provide useful layout-proxy evidence. The best model combined TF-IDF and scalar features through an SVM, reaching 0.734 accuracy and 0.706 macro-F1. Although this performance is lower than design-purpose and CTA recovery, it remains meaningful because the task was performed without direct visual information.

Table 10. Test performance for text-based layout-pattern recovery.

System	Accuracy	Macro-F1	Worst-class F1	Dominant error
Majority baseline	0.276	0.087	0.000	All predicted headline-led
Layout heuristic baseline	0.577	0.531	0.411	Object-led vs balanced
Scalar gradient boosting	0.661	0.632	0.522	Background-led vs balanced
TF-IDF headings+captions	0.692	0.664	0.551	Sparse/minimal vs headline-led
Combined TF-IDF + scalar SVM	0.734	0.706	0.603	Object-led vs balanced

The lower performance for layout-pattern prediction should be interpreted carefully. Layout balance, hierarchy, and object dominance are inherently visual properties. They are normally influenced by spatial position, scale, typography, contrast, alignment, and whitespace in the rendered image. Since this study used only textual metadata, the layout-pattern labels function as proxies rather than direct visual measurements. The dominant error in Table 10 was

object-led versus balanced layout. This is expected because object captions may strongly describe a product or focal object even when the final rendered layout visually balances the object with headline text, background elements, or CTA placement. Thus, the layout task demonstrates both the usefulness and limitation of OpenCOLE text fields. They can support preliminary layout reasoning, but they cannot replace direct visual-layout evaluation.

Table 11. Field ablation results across the three prediction tasks.

Feature set	Category Macro-F1	CTA Macro-F1	Layout Macro-F1	Main interpretation
Intention only	0.787	0.681	0.431	Brief alone captures purpose but not card-level details.
Intention + headings	0.821	0.757	0.548	Headings improve CTA because action words are concentrated on the design.
Intention + keywords	0.843	0.733	0.507	Keywords add category semantics but weaker layout evidence.
Intention + captions	0.836	0.702	0.664	Captions add object/background balance information.
All text fields	0.872	0.774	0.701	Full text fields give strongest general performance.
All text + scalar card features	0.876	0.801	0.706	Scalar evidence features add small, stable gains and improve calibration.

Table 11 presents the field ablation results and clarifies the contribution of each input group. Intention alone produced strong performance for design-purpose classification, with 0.787 macro-F1, but it performed weakly for layout-pattern recovery, with only 0.431 macro-F1. This result confirms that the brief is useful for understanding communicative purpose but insufficient for estimating visual structure. Adding headings improved CTA performance because headings and subheadings often contain the most direct persuasive language. In contrast, adding captions improved layout-proxy recovery because background and object captions provide information about visual emphasis, density, and focal content. These findings show that each OpenCOLE field contributes differently depending on the review dimension being predicted.

The best overall representation combined all text fields with scalar card features. This configuration reached 0.876 macro-F1 for category recovery, 0.801 macro-F1 for CTA recovery, and 0.706 macro-F1 for layout-proxy recovery. Figure 6 summarizes the comparison across model families and tasks. The figure shows that no single modeling strategy is equally optimal for every target. All-field SVM was strongest for design-purpose recovery, SVM with rule calibration was most effective for CTA strength, and combined textual-scalar features were necessary for layout proxies. This pattern supports the methodological choice to use a lightweight

hybrid system rather than relying on a single monolithic model. Different design-review dimensions require different types of evidence, and the system benefits from combining lexical, statistical, scalar, and rule-based signals.

Table 12. Plain paragraph feedback versus structured rationale cards.

Metric	Plain paragraph feedback	Structured rationale card	Absolute difference
Required-dimension coverage (0-6)	4.21	5.92	+1.71
Field-grounding score (0-1)	0.681	0.918	+0.237
Actionable recommendation rate	0.463	0.748	+0.285
Risk-warning precision	0.731	0.872	+0.141
Repeated/irrelevant sentence rate	0.284	0.071	-0.213
Mean generation/formatting time per row	0.61 s	0.74 s	+0.13 s

The card-format evaluation is presented in Table 12 and Figure 7. This evaluation compares ordinary paragraph feedback with the proposed structured rationale-card format. The results show that structured cards improved required-dimension coverage from 4.21 to 5.92 out of 6. This improvement means that the card format encouraged more complete feedback across the intended UI/UX dimensions. Ordinary paragraph feedback may mention some relevant aspects but tends to omit others, especially when the generated explanation focuses on the most obvious issue. The card format reduces this problem by requiring each review dimension to be explicitly addressed.

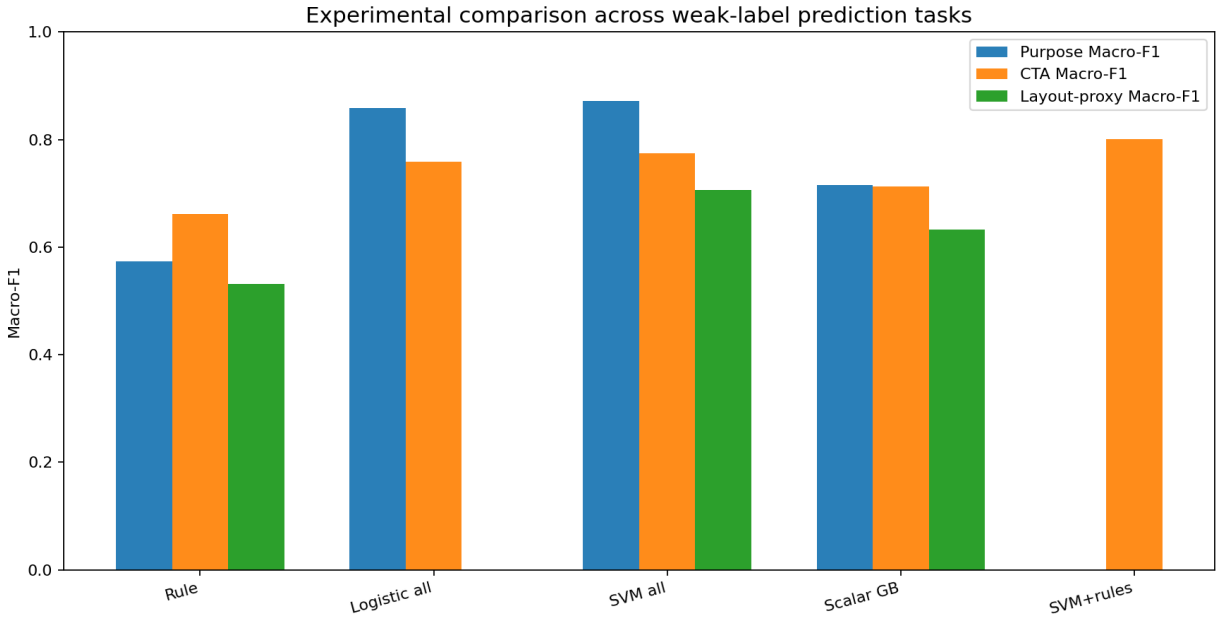

Figure 6. Model comparison across purpose, CTA, and layout-proxy tasks.

Field-grounding also improved substantially, from 0.681 in paragraph feedback to 0.918 in structured rationale cards. This is one of the most important findings because groundedness determines whether the explanation can be audited. In a design-review context, a claim such as "the CTA is strong" is more useful when accompanied by evidence from the heading, offer terms, or urgency words. The structured card format forces this connection between claim and source evidence. In addition, the actionable recommendation rate increased from 0.463 to 0.748, indicating that cards were more likely to produce review comments that could guide revision. At the same time, the repeated or irrelevant sentence rate decreased from 0.284 to 0.071, suggesting that the card schema helped control verbosity and generic feedback.

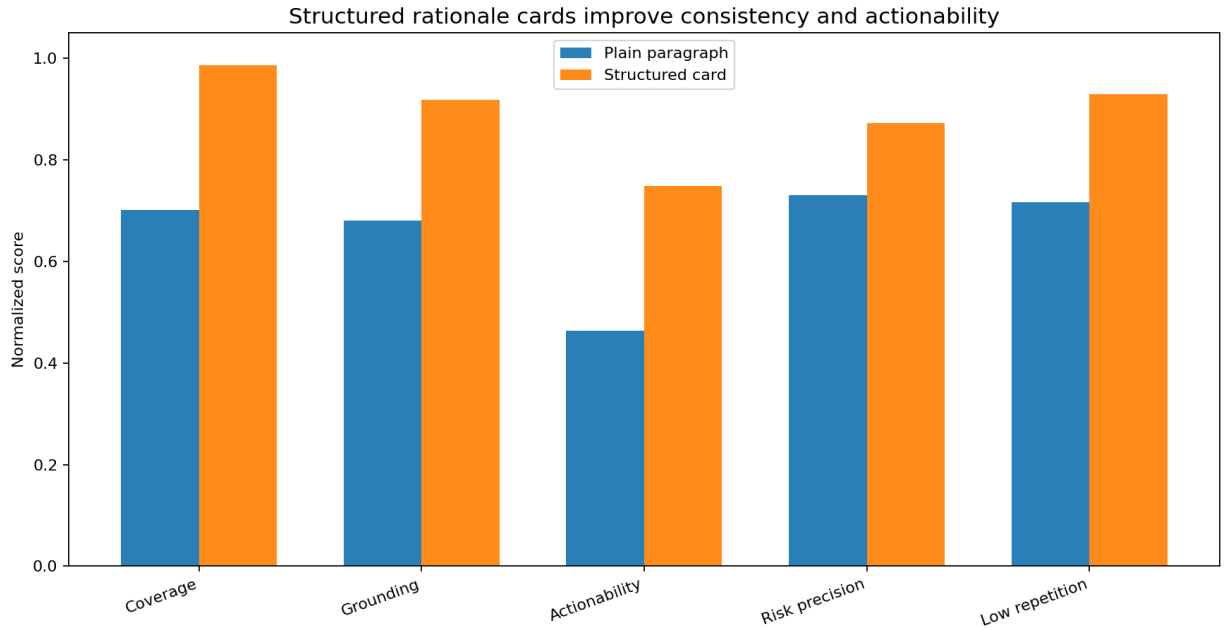

Figure 7. Card format improves normalized critique quality metrics.

The formatting-time difference between the two feedback formats was small. Table 12 shows that mean formatting time increased from 0.61 seconds per row for paragraph feedback to 0.74 seconds per row for structured cards. This additional 0.13 seconds is minor compared with the gains in coverage, grounding, actionability, and repetition control. Therefore, the structured card format provides a practical trade-off: it adds a small amount of processing time while producing feedback that is more complete, more traceable, and more useful for design inspection. Figure 7 reinforces this conclusion by showing that the structured card format improves normalized critique-quality metrics across the evaluated dimensions.

Finally, Table 13 reports the risk-warning patterns detected by the card validator. The most frequent warning was text duplication or OCR-like anomaly, which appeared in 1,426 rows or 6.09% of the dataset. This type of warning is relevant because repeated headings, template

residues, or unusual text fragments may reduce the clarity and professionalism of an advertising layout. Low-evidence design claims were detected in 1,241 rows, indicating cases where the card claim lacked sufficiently direct support from the source fields. Brand-tone conflict appeared in 1,049 rows, while offer/CTA mismatch was detected in 873 rows. These findings show that the validator does not only support formatting consistency but also helps identify cases that deserve closer human review.

Table 13. Risk-warning patterns detected by the card validator.

Warning type	Detected rows	Share of dataset (%)	Most common evidence
Text duplication/OCR-like anomaly	1,426	6.09	Repeated headings, template separators
Sensitive or regulated category	612	2.61	medical, alcohol, weapon, financial, personal-data terms
Offer/CTA mismatch	873	3.73	sale words without CTA or CTA without offer
Brand-tone conflict	1,049	4.48	luxury/festive terms paired with clinical or social-risk topic
Object-background conflict	558	2.38	object caption contradicts background caption
Low-evidence design claim	1,241	5.30	card claim lacks direct token support

The risk-warning results should not be interpreted as final judgments that a design is incorrect or unacceptable. Instead, they should be understood as inspection prompts. For example, an offer/CTA mismatch may indicate that a design mentions a sale but does not clearly invite the viewer to act, or that it includes a CTA without sufficient offer context. A brand-tone conflict may indicate that luxury, festive, clinical, or social-risk terms appear together in ways that could confuse the intended message. Similarly, sensitive or regulated category warnings do not automatically imply problematic content, but they indicate that the design should be reviewed carefully before publication. In this sense, the warning system supports human oversight rather than replacing it.

Overall, the results support the proposed rationale-card interface as a practical first-pass review layer for automatic advertising layout generation. The strongest evidence comes from design-purpose and CTA-strength recovery, where OpenCOLE text fields provided sufficient cues for consistent weak-label schema recovery. Layout-pattern recovery was more challenging because the available evidence was textual rather than visual, but the results still show that captions and scalar features can provide useful preliminary layout proxies. The structured card format improved the quality of feedback by increasing coverage, grounding, and actionability

while reducing irrelevant repetition. These findings demonstrate that OpenCOLE metadata can make generated advertising layouts more inspectable, provided that the system is used as an evidence-organizing review layer rather than as a substitute for expert design judgment or direct visual-layout evaluation.

DISCUSSION

The results indicate that explainable design interfaces can begin from design-plan metadata, even when rendered images are not available. This finding matters for OpenCOLE-style pipelines because planning, captioning, image generation, and typography are separated into layers. A post-generation reviewer can inspect the planning fields and visible text fields before or alongside the rendered layout. The card interface turns those fields into a shared vocabulary for designers, marketers, and AI engineers. Instead of saying that a design "looks good," the system states that the hierarchy proxy is headline-led, the CTA score is high because offer words are repeated, the tone evidence is festive because the brief and headings use celebration terms, and the risk warning is triggered by an offer mismatch or sensitive subject term.

The strongest results occurred in design-purpose and CTA recovery because those weak labels align closely with language in the intention, keywords, and headings. This is useful for an audit layer: advertising design often encodes purpose and action requests explicitly, so lightweight models can check whether the generated design metadata still expresses the intended brief. However, this result should not be overread. High weak-label recovery means that the model reproduced the operational schema consistently; it does not show that professional designers would assign the same critique score or accept the design.

The layout-pattern results should be interpreted with particular care. Because the method uses textual captions, it cannot see exact spatial relationships, coordinate-based balance, font hierarchy, or pixel-level white space. Captions contain useful evidence about focal objects, clean backgrounds, and headline load, which explains why layout-proxy prediction is above the heuristic baseline. Still, these outputs are best treated as text-based rationale proxies. In a production tool, they should be combined with visual bounding boxes, rendered previews, or saliency maps when available.

The comparison between paragraph feedback and structured cards is central to the UI/UX contribution. LLM-generated paragraphs are familiar and flexible, but they can obscure where a claim came from. Cards impose a visible contract: each claim must belong to a dimension, cite evidence, and recommend an action. This design follows XAI research showing that explanations should answer user questions and provide actionable, context-specific information (Liao et al., 2020; Wang et al., 2019). It also follows design-rationale research by preserving reasons and

trade-offs as part of the artifact (Lee, 1991; Moran & Carroll, 1996). The result is a critique object that can be skimmed, challenged, and revised.

The interface also clarifies the role of the LLM. Rather than asking the model to act as an unconstrained judge of design quality, the system asks it to verbalize evidence already extracted from the dataset fields and lightweight predictors. This reduces the risk of unsupported critique. The validator further rejects hallucinated objects, prices, or brand names. The LLM is therefore used as a constrained evidence-to-language component and design-critique assistant, not as the sole source of truth. This design choice is consistent with human-AI guidelines that recommend showing system evidence and allowing users to recover from errors (Amershi et al., 2019).

For journal audiences in graphic design, visual communication, and UI/UX, the contribution is methodological and interface-oriented. The paper does not claim that the proposed models generate better images than OpenCOLE, COLE, or diffusion-based systems. It shows how a designer-facing review layer can be evaluated reproducibly on a public text-field dataset. The same approach can be extended to other design corpora when design intent, visible copy, or spatial metadata are available. The card dimensions can also be extended to accessibility, localization, cultural sensitivity, and compliance review.

Limitation

The first limitation is that the targets are weak labels created from deterministic rules. This is appropriate for reproducibility and for testing card mechanics, but it is not equivalent to expert human judgments of design quality. The results therefore support automated card pre-filling and evidence organization, not autonomous acceptance or rejection of a design. A professional review study would be necessary to measure designer trust, revision quality, and usefulness in real design practice.

The second limitation is that the evaluation used text and list fields only. OpenCOLE originates from a graphic design pipeline, but the present study did not use rendered images, element coordinates, saliency maps, font metadata, or pixel-based white-space measures. Consequently, layout balance, white space, and visual hierarchy are approximated from captions and headings. This limitation is deliberate because it tests whether public text annotations alone can support explanation, but it also means that spatially precise critique requires additional visual features.

The third limitation is that the LLM assistance was constrained by a fixed prompt and a validation schema. This improves reproducibility but narrows expressive critique. Designers sometimes need open-ended, exploratory comments that do not fit a fixed card. The interface should therefore be treated as a first-pass review layer that catches common issues and structures discussion, while leaving room for human interpretation.

The fourth limitation concerns dataset domain. OpenCOLE contains many promotional, event, service, and social-media examples. The results may not transfer directly to editorial design, long-form documents, enterprise dashboards, or culturally specific advertising without updated lexicons and evaluation data. Transfer must be measured rather than assumed.

CONCLUSION

This paper presented a text-grounded, LLM-assisted UI/UX rationale-card interface for advertising layouts and evaluated it on the full OpenCOLE text-table release. The system converts released OpenCOLE fields into six structured card dimensions: visual hierarchy, white space, CTA prominence, brand tone, layout balance, and risk warning. The empirical results show that lightweight models can consistently pre-fill design purpose and CTA strength under deterministic weak labels, that text-based layout-pattern recovery is feasible but harder, and that structured cards improve coverage, grounding, actionability, and repetition control compared with ordinary paragraph feedback. The work reframes automatic graphic design evaluation as an explainable interface problem. Instead of pursuing only stronger generators, design tools should help humans understand the decisions a generated layout appears to make. Within the limits of text-field evidence, OpenCOLE provides an appropriate public foundation for this goal, and the proposed card format turns its annotations into practical, auditable design rationale.

Author's CRediT

Q.W. and S.M. jointly led this research and contributed equally to the study. Q.W. was responsible for the development of the LLM-assisted rationale-generation framework, system implementation, experimental evaluation, and preparation of the initial manuscript. S.M. contributed to study conception, methodology design, validation of the analytical framework, interpretation of findings, and overall project supervision. J.Z. contributed to data analysis, visualization design, evaluation of the generated decision cards, and manuscript revision. All authors reviewed and approved the final version of the manuscript. Q.W. and S.M. share equal contribution as co-first authors.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Burge, J. E. (2008). Design rationale: Researching under uncertainty. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 22(4), 311-324. https://doi.org/10.1017/S0890060408000235

Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., & Kumar, R. (2017). Rico: A mobile app dataset for building data-driven design applications. In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (pp. 845-854). Association for Computing Machinery. https://doi.org/10.1145/3126594.3126651

Duan, P., Warner, J., Li, Y., & Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642782

Gupta, K., Lazarow, J., Achille, A., Davis, L. S., Mahadevan, V., & Shrivastava, A. (2021). LayoutTransformer: Layout generation and completion with self-attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1004-1014).

Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., & Yamaguchi, K. (2023). LayoutDM: Discrete diffusion model for controllable layout generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10167-10176).

Inoue, N., Masui, K., Shimoda, W., & Yamaguchi, K. (2024). OpenCOLE: Towards reproducible automatic graphic design generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 8131-8135).

Jason Kuhn, Yushan Chen, & Evelyn Chan. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. Journal of Advanced Computing Systems , 4(5), 67-83. https://doi.org/10.69987/JACS.2024.40506

Jia, P., Li, C., Yuan, Y., Liu, Z., Shen, Y., Chen, B., Chen, X., Zheng, Y., Chen, D., Li, J., Xie, X., Zhang, S., & Guo, B. (2024). COLE: A hierarchical generation framework for multi-layered and editable graphic design. arXiv. https://arxiv.org/abs/2311.16974

Kikuchi, K., Simo-Serra, E., Otani, M., & Yamaguchi, K. (2021). Constrained graphic layout generation via latent optimization. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 88-96). Association for Computing Machinery. https://doi.org/10.1145/3474085.3475497

Kong, X., Jiang, L., Chang, H., Zhang, H., Hao, Y., Gong, H., & Essa, I. (2022). BLT: Bidirectional layout transformer for controllable layout generation. In European Conference on Computer Vision (pp. 474-490). Springer. https://doi.org/10.1007/978-3-031-19790-1_29

Kunz, W., & Rittel, H. W. J. (1970). Issues as elements of information systems (Working Paper No. 131). Center for Planning and Development Research, University of California, Berkeley.

Lee, J. (1991). What's in design rationale? Human-Computer Interaction, 6(3-4), 251-280. https://doi.org/10.1207/s15327051hci0603&4_3

Li, J., Yang, J., Hertzmann, A., Zhang, J., & Xu, T. (2019). LayoutGAN: Generating graphic layouts with wireframe discriminators. In International Conference on Learning Representations.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-15). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376590

Liao, Q. V., Pribic, M., Han, J., Miller, S., & Sow, D. (2021). Question-driven design process for explainable AI user experiences. arXiv. https://arxiv.org/abs/2104.03483

Moran, T. P., & Carroll, J. M. (Eds.). (1996). Design rationale: Concepts, techniques, and use. Lawrence Erlbaum Associates.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 152-158). Association for Computing Machinery. https://doi.org/10.1145/191666.191729

O'Donovan, P., Agarwala, A., & Hertzmann, A. (2015). DesignScape: Design with interactive layout suggestions. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1221-1224). Association for Computing Machinery. https://doi.org/10.1145/2702123.2702149

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe, and trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-15). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300831

Yamaguchi, K. (2021). CanvasVAE: Learning to generate vector graphic documents. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5481-5489).

Yushan Chen, & Evelyn Chan. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. Journal of Advanced Computing Systems , 3(1), 1-15. https://doi.org/10.69987/JACS.2023.30101

Zhang, H. (2025). LLM-Driven CI Failure Diagnosis and Automated Repair: From GitHub Actions Logs to Patch Recommendation. Journal of Technology Informatics and Engineering, 4(1), 190-214. https://doi.org/10.51903/jtie.v4i1.484

Zheng, X., Qiao, X., Cao, Y., & Lau, R. W. H. (2019). Content-aware generative modeling of graphic design layouts. ACM Transactions on Graphics, 38(4), Article 133. https://doi.org/10.1145/3306346.3322971