International Journal of Graphic Design (IJGD)
Volume 4 No. 1 April 2026
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 266-292
DOI: https://doi.org/10.51903/ijgd.v4i1.3994
*Available online at*: https://journal.stekom.ac.id/index.php/ijgd

# Beyond CTR and VCR: LLM-Assisted Design Evaluation with Eye-Tracking Validation for Graphic Advertising

**Qiwen Zheng*[1], Xiaochen Li[2], Fiona Wang[3]**
[1]Integrated Marketing, New York University, NY, USA
[2]Marketing Analytics, Bentley University, MA, USA
[3]Applied Analytics, Columbia University, NY, USA
Email Address: qiwen.zheng.626@gmail.com

***Abstract***. *Digital advertising design is often judged with delivery or response metrics such as click-through rate (CTR), video-completion rate (VCR), and viewability, although these measures do not describe whether a banner communicates through a clear visual hierarchy. This study evaluates whether large language model (LLM)-assisted scores and lightweight image features can support graphic advertising review while keeping design-quality prediction distinct from measured visual attention. GraphicDesignEvaluation provided 1,200 principle–image rating rows for alignment, overlap, and white space. Under grouped five-fold cross-validation by base design, the GPT-only Ridge model reached R2 = 0.408, RMSE = 1.566, MAE = 1.257, Pearson = 0.639, and Spearman = 0.634. Direct GPT–human agreement was uneven: overlap was strong (Spearman = 0.769), white space was moderate (0.637), and alignment was weaker (0.519). An external construct analysis used 1,000 advertisements from ADD1000 with aggregate human fixation-density maps and matched EAID ratings. A combined gradient and local-contrast proxy modestly exceeded a center prior (CC = 0.457, SIM = 0.562, KL divergence = 0.637), while lower fixation entropy was associated with higher aesthetic ratings (Spearman = -0.319). BannerRequest400 was used only to describe brief, CTA-language, logo, and format constraints because it contains no human design-quality or gaze labels. The findings support a preliminary, human-supervised design-review framework: LLM scores are useful screening evidence, especially for overlap, but they do not replace human creative judgment, eye tracking, or advertising-performance measures.*

***Keywords:*** *graphic design evaluation; advertising banner; LLM-assisted scoring; eye tracking; fixation-density map; visual hierarchy.*

## INTRODUCTION

Advertising measurement has long relied on behavioral or delivery metrics. CTR captures a click, VCR captures video completion, and viewability captures whether an ad had the opportunity to be seen. These metrics are useful for media accountability, yet they are too narrow to describe whether the creative itself is visually readable, balanced, and persuasive. Recent attention-measurement guidance treats attention data as a signal that can predict full-funnel outcomes when it is validated against awareness, CTR, installs, sales lift, or lifetime value (Interactive Advertising Bureau, 2024). IAB Europe (2023) similarly frames attention as a way to move beyond simple exposure by considering whether the user had a realistic opportunity to process the ad. For graphic design journals, however, the more relevant question is not whether attention replaces conversion metrics, but whether attention can be translated into a visual communication framework for evaluating the structure of advertising design.

Campaign-level AI work has already expanded the behavioral layer of advertising measurement (Bai & Wu, 2026; Kuo et al., 2025). LLM-augmented customer representation learning, email-advertising incrementality modeling, uncertainty-aware uplift policies, privacy-robust cookieless incrementality, offline counterfactual evaluation for advertising slots, and

*Received: Januari 2026; Revised: February 2026; Accepted: March 2026; Published: April 2026*
*Corresponding author, qiwen.zheng.626@gmail.com

conservative off-policy bidding/ranking all help decide which audience, channel, creative family, or placement should receive an impression (Bai et al., 2026; S. Lu & Zhou, 2023; Y. Lu et al., 2024; Mu et al., 2023; Mu et al., 2025; Xin, 2026; Ye et al., 2026). Recommendation-oriented work has also examined bias mitigation and faithful natural-language explanations for ranking decisions (Chang et al., 2023). These studies are valuable, but they sit downstream of the rendered ad. They do not answer whether the banner itself has a readable hierarchy, a visible brand mark, a separated CTA, or enough white space to support comprehension.

This study treats design-quality proxies and measured visual attention as related but distinct constructs. Alignment, overlap, white space, readability, visual balance, and CTA-like contrast describe properties of a static design that may shape perceptual priority, but they do not by themselves show where viewers looked. The term visual attention is reserved here for the human fixation-density evidence used in the external analysis. This distinction makes it possible to ask whether inexpensive design proxies contain useful spatial information without equating them with gaze, behavioral response, or campaign outcomes.

This paper therefore reframes the phrase beyond CTR and VCR as a graphic design problem. A banner may attract attention because its focal point is clear, the brand mark is legible, the text does not collide with imagery, the CTA area can be perceived quickly, and white space helps viewers separate primary and secondary information. Eye-tracking research in marketing has shown that brand, text, and pictorial areas compete for limited visual attention (Pieters & Wedel, 2004; Wedel & Pieters, 2008). Visual complexity can also increase stopping power while making comprehension harder if the design becomes cluttered (Pieters et al., 2010). These findings align with design theory: alignment, contrast, proximity, repetition, and spacing are not decorative rules, but mechanisms for guiding visual hierarchy (Lidwell et al., 2010; Lupton, 2010; Williams, 2014).

The central research problem is whether LLM-assisted design scores and inexpensive raster features can predict human judgments of banner quality (Zhou, Wang, & Chang, 2025), and whether the spatial proxies show measurable agreement with human fixation density. Recent multimodal-model work suggests that GPT-based evaluators can assess design principles with reasonable agreement to human raters, although subtle differences remain difficult (Haraguchi et al., 2024). Computational layout research has also formalized alignment, overlap, and white space (Kong et al., 2023; O'Donovan et al., 2014). The present analysis combines these streams while reporting principle-specific weaknesses and the absolute magnitude of fixation-map agreement.

Three research questions guide the paper. RQ1 asks how closely GPT-based design ratings align with human ratings for alignment, overlap, and white space. RQ2 asks whether lightweight

raster features can compete with or complement LLM-assisted scores when predicting human design-quality judgments. RQ3 asks how well training-free gradient and local-contrast maps agree with human fixation-density maps and how attention-related image features relate to human advertising ratings. BannerRequest400 contributes a separate contextual analysis of briefs, CTA language, logos, and aspect-ratio constraints; it is not part of the supervised or fixation validation. Figure 1 summarizes these distinct analytic streams.

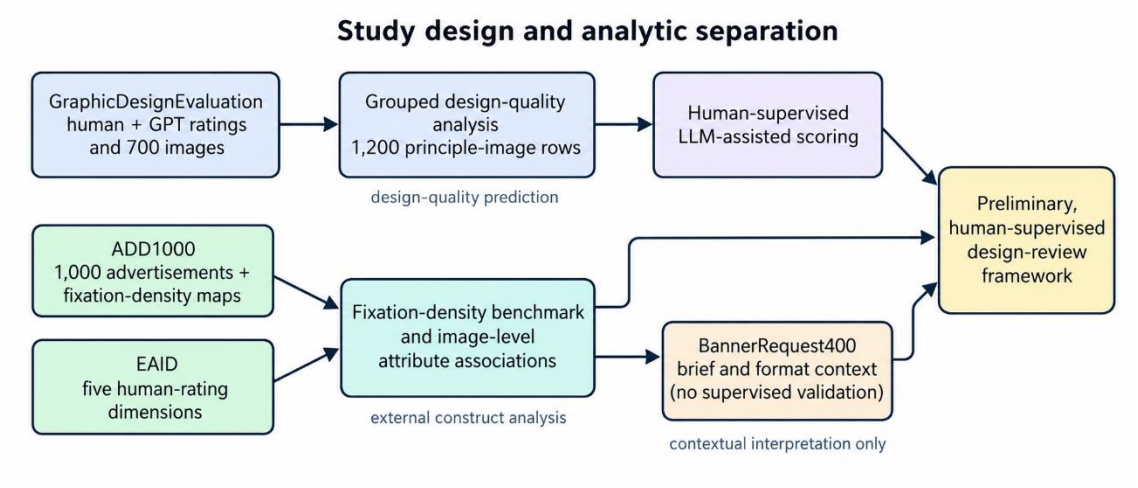


Figure 1. Study design separating human-supervised design-quality prediction, fixation-density validation, affective-rating analysis, and contextual brief analysis.

## LITERATURE REVIEW

Visual attention research provides the theoretical basis for attention-aware design evaluation. Feature integration theory argues that viewers bind visual features into objects through selective attention (Treisman & Gelade, 1980). Saliency models operationalize this idea by estimating how color, intensity, orientation, and local contrast attract early attention (Itti et al., 1998). Subsequent work emphasizes that attention is guided both by bottom-up visual attributes and by task demands (Wolfe & Horowitz, 2004). Load theory further explains why cluttered scenes can impair selective processing when perceptual load is high (Lavie, 2005). In graphic advertising, these theories translate into practical concerns: a banner should make the primary message easy to find, keep brand and CTA elements separable, and avoid unnecessary visual competition. Rosenholtz et al. (2007) show that clutter can be measured from image statistics, which motivates the use of edge density, component count, and entropy as traditional image baselines.

Graphic design literature offers complementary principles that are directly tied to visual communication. Williams (2014) describes alignment and proximity as ways to create order and relationships among elements. Lupton (2010) treats type, spacing, and hierarchy as central to

readable communication. Lidwell et al. (2010) connect universal design principles such as hierarchy, signal-to-noise ratio, and legibility to user perception (Z. Li et al., 2025). Tufte (2001) similarly argues that visual displays should reduce nonessential clutter and support clear comparison. These principles are especially relevant to banner advertising, where a small canvas must combine brand, offer, image, text, and call to action (Solikhan et al., 2026).

The same principles also explain why a purely behavioral metric can misrepresent creative quality. A banner with a strong offer may receive clicks even if the layout is visually weak, and a brand-awareness banner may be visually effective even when CTR is low. Design evaluation therefore needs intermediate evidence about the ad object itself. Alignment indicates whether the layout has a readable grid or whether elements appear accidental. Overlap indicates whether elements compete or obscure one another. White space indicates whether the message has breathing room and whether the viewer can group related items quickly. These principles are observable in image form and meaningful to design reviewers, making them suitable targets for AI-assisted assessment.

Computational design research has attempted to formalize these principles. O'Donovan et al. (2014) proposed energy-based layout models that encode alignment, overlap, and spacing. Kong et al. (2023) extended this tradition by refining graphic designs through design principles and human preference. Layout-generation research has also moved toward controllable, editable, and design-aware systems, including diffusion models for layouts (Inoue et al., 2023) and hierarchical generation frameworks for multi-layered graphic design (Jia et al., 2023). These approaches show that design quality can be modeled computationally, but the evaluation of final designs remains difficult because visual judgment is partly semantic, partly perceptual, and partly context-dependent (Wang, Ren, & Chang, 2025).

Recent LLM-assisted UI/UX work adds an explanation layer to this computational-design tradition. LLM-as-design-critic research treats AI feedback as something that must be aligned with human graphic design judgment, while text-grounded design-rationale cards and visual brief cards translate layout metadata, creative intention, audience, brand tone, CTA, and design risk into reviewer-facing structures (Y. Li et al., 2025; Liu et al., 2025; Tu et al., 2025; Wu et al., 2025). Structured visual brief interfaces and sketch-to-prototype studies further show that generative or vision-language design systems become more usable when free-form intent is converted into editable slots and consistency checks rather than a single opaque score (Y. Chen & Li, 2025; Mu et al., 2026). Evidence-card and recommendation-card interfaces make the same point in adjacent search and e-commerce contexts: a helpful AI review should show why a judgment was made and what evidence supports it (Jin, 2025b; Meng et al., 2026; B. Zhang et al., 2025).

The most relevant design-quality resource is GraphicDesignEvaluation, which compares GPT and human ratings for alignment, overlap, and white space (Haraguchi et al., 2024). ADD1000 provides 1,000 advertising images paired with aggregate fixation-density maps collected from 57 participants, enabling direct spatial comparison with human gaze (Liang et al., 2021). EAID adds five human rating dimensions for the same numbered advertisements: ad liking, emotional response, aesthetic quality, functional quality, and brand liking (Liang et al., 2024). BannerRequest400 contributes complementary design context through 100 logos, 400 banner requests, and 5,200 standard-size specifications (Wang, Shimose, & Takamatsu, 2025). The machine-learning comparison uses Ridge regression as a linear baseline, Random Forests for nonlinear feature interactions (Breiman, 2001), and XGBoost for gradient-boosted trees (T. Chen & Guestrin, 2016), implemented within a scikit-learn pipeline (Pedregosa et al., 2011).

The brief-to-metric translation also benefits from requirements-to-constraints reasoning. Although developed in a healthcare test-fit setting, Sun et al. (2023) demonstrate a relevant pattern: natural-language requirements can be decomposed into checkable constraints before computational evaluation. In the present advertising setting, the same pattern appears when a request such as 'make the logo prominent' is reframed as measurable brand-region visibility, component separation, and CTA hierarchy rather than treated as a vague creative preference.

LLM/VLM scoring is treated as a candidate design-review instrument rather than a substitute for ground truth. Human ratings remain the criterion for design quality, while fixation-density maps provide the criterion for spatial visual attention. Traditional features remain useful as diagnostic evidence because they can identify density, spacing, balance, or collision patterns even when they do not recover the full semantic judgment. The analysis therefore evaluates predictive accuracy, spatial agreement, and interpretability as separate properties.

Reliability is therefore a methodological condition, not only a deployment concern (Zhong et al., 2026). Work on hallucination firewalls, evidence-chain RAG, evidence-calibrated abstention, and multi-hop retrieval shows that AI-generated judgments should be checked for evidence consistency, attribution, and appropriate refusal when evidence is insufficient (Y. Chen & Xu, 2026; Jin, 2025a; C. Li et al., 2024; C. Li et al., 2023; Su et al., 2025; Zhong et al., 2025). Related safety studies on red-teaming, multi-stage refusal, self-verification, scientific claim checking, and privacy/data-integrity risk cards reinforce the separation between a model score and the explanation shown to a human reviewer (Su, Chen, & Qian, 2026; Su, Rao, & Ma, 2026; Zheng & Li, 2024; Zheng et al., 2023; Zheng et al., 2024; Zhou, Li, & Liu, 2025).

## METHODS

The empirical design used four datasets with distinct roles, summarized in Table 1. GraphicDesignEvaluation includes human and GPT ratings on a 1–10 scale for alignment,

overlap, and white space (Haraguchi et al., 2024). Its 100 base designs, three principle-specific perturbation levels, and original versions form 700 unique images; stacking the three rating tasks yields 1,200 principle–image rows rather than 1,200 independent designs. ADD1000 contains 1,000 advertising images and one aggregate fixation-density map per image from an eye-tracking study with 57 participants (Liang et al., 2021). EAID provides five 1–7 human-rating matrices matched by image identifier (Liang et al., 2024). BannerRequest400 includes 400 abstract requests, 100 logos, and 5,200 standard-size specifications (Wang, Shimose, & Takamatsu, 2025). Because BannerRequest400 contains neither judged rendered banners nor human design-quality or gaze labels, it was analyzed only for brief, CTA-language, logo, and size context.

For GraphicDesignEvaluation, raster features were extracted after resizing each image to a maximum long side of 200 pixels. White-space ratio was estimated from a background-color and edge-dilated foreground mask. Edge density, luminance contrast, saturation, colorfulness, and grayscale entropy represented visual density and contrast. Connected components approximated layout elements, supporting component count, largest-component share, component-area inequality, content bounding-box area, and margin statistics. Alignment dispersion measured deviations from common edge and center lines; overlap used pairwise bounding-box intersection; and CTA-like contrast identified button-shaped rectangles in likely action zones. These operational feature groups are summarized in Table 2.

The extraction procedure followed a fixed sequence for every image. The image was converted to RGB, resized with aspect ratio preserved, and converted to grayscale and HSV representations. A foreground mask was estimated by combining background-color distance, edge response, and morphological dilation; this mask defined occupied area, empty area, content bounding box, and margins. Connected components smaller than noise thresholds were removed before component-level features were calculated. The same preprocessed images were used for edge, entropy, balance, and saliency proxies so that feature differences reflected design differences rather than inconsistent scaling. All features were numeric and model-ready after missing values were imputed with training-fold medians.

The raster measures were interpreted as evidence tags rather than independent design labels. High edge density can support a clutter warning but cannot establish conceptual confusion, and low margin can support a spacing warning but cannot establish weak brand hierarchy. This evidence-first interpretation is consistent with design-rationale and risk-card interfaces that expose the reason for a recommendation while preserving human authority over the final review (Jin, 2025b; Su, Rao, & Ma, 2026; Wu et al., 2025; Y. Zhang & Zhang, 2025a; Zhou, Li, & Liu, 2025).

International Journal of Graphic Design (IJGD)
Volume 4 No. 1 April 2026
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 266-292
DOI: https://doi.org/10.51903/ijgd.v4i1.3994
*Available online at*: https://journal.stekom.ac.id/index.php/ijgd

Table 1. Dataset inventory and analytic role.

| Dataset | Content | Analytic unit | n used | Criterion data | Analytic role | Boundary |
|---|---|---|---|---|---|---|
| GraphicDesignEvaluation | 700 unique images; human/GPT ratings | principle–image row | 1,200 rows; 100 base IDs | 1–10 human ratings | grouped design-quality prediction | folds grouped by base ID |
| ADD1000 | ads + aggregate fixation maps | image | 1,000 pairs | fixation density from 57 participants | external spatial validation | no raw scanpaths or dwell time |
| EAID | five rating sheets matched to ADD1000 | image | 1,000; brand n = 667 | 1–7 human ratings | image-level associations | -2 treated as unrecognized brand |
| BannerRequest400 | briefs, logos, 13 size slots | request/specification | 400 briefs; 5,200 slots | none for design quality or gaze | contextual analysis | excluded from prediction and gaze tests |

Table 2. Operational feature groups used in the empirical models.

| Feature group | Variables or evidence | Interpretation |
|---|---|---|
| LLM-assisted score | gpt_avg, gpt_sd | principle-specific rating and run dispersion |
| GDE spacing and structure | white_space_ratio, margins, components, content_bbox_area | occupied area, grouping, and border spacing |
| GDE density and balance | edge_density, entropy, text_area_proxy, center and quadrant balance | clutter and distribution of visual structure |
| GDE alignment, overlap, CTA | alignment_dispersion, bounding-box intersections, CTA-like contrast | supplementary layout evidence |
| ADD1000 proxy maps | gradient, local color contrast, equal-mass combination | training-free spatial predictions |
| ADD1000 fixation evidence | aggregate density, central mass, normalized entropy | human spatial-attention criterion |
| EAID human ratings | ad liking, emotional, aesthetic, functional, brand liking | image-level perceptual and affective associations |
| BannerRequest400 context | brief length, CTA language, logo and size fields | scope and format requirements only |

*Received: Januari 2026; Revised: February 2026; Accepted: March 2026; Published: April 2026*
*Corresponding author, qiwen.zheng.626@gmail.com

International Journal of Graphic Design (IJGD)
Volume 4 No. 1 April 2026
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 266-292
DOI: https://doi.org/10.51903/ijgd.v4i1.3994
*Available online at*: https://journal.stekom.ac.id/index.php/ijgd

Model evaluation used grouped five-fold cross-validation, with image_id as the grouping variable. This split prevents original and perturbed versions of the same design identity from appearing in both training and test folds. Five feature families were compared: a metadata control using principle and perturbation label, a visual-only baseline, a GPT-only model, a hybrid model using GPT and visual features, and a hybrid-plus-metadata diagnostic model. The last family is not a deployment model because perturbation level is an experimental manipulation; it shows how much predictive signal is available when design-degradation metadata is known. Performance was evaluated with RMSE, MAE, R2, Pearson correlation, and Spearman rank correlation. Spearman was emphasized because graphic design review often requires ranking alternatives rather than estimating an exact score.

Each model family used the same folds and the same target variable, which makes the comparisons directly interpretable. Ridge regression used standardized numeric predictors and one-hot encoded categorical variables. Random Forest and XGBoost used the same feature matrix without requiring linearity. The evaluation reports both error metrics and association metrics because they answer different design questions. RMSE and MAE indicate the expected scoring error when a reviewer wants an absolute rating. Pearson correlation indicates linear calibration, while Spearman correlation indicates whether the model ranks better and worse designs in a human-like order. This combination is appropriate for design screening, where ranking a set of alternative banners is often more important than assigning a perfect score to one banner.

The external spatial analysis used the 1,000 ADD1000 advertisement–fixation-map pairs. Each pair was resized together with aspect ratio preserved to a maximum side of 256 pixels. A luminance-gradient proxy used Sobel magnitude after Gaussian smoothing (sigma = 1 pixel). A local color-contrast proxy measured Euclidean RGB distance from a Gaussian local background (sigma = 8 pixels). The combined proxy mixed the separately normalized gradient and local-contrast maps in equal mass. Each proxy was then smoothed at 3% of the shorter resized dimension and normalized to unit mass. A centered Gaussian prior (sigma_x = 0.25W; sigma_y = 0.25H) served as a non-content baseline.

Agreement with aggregate fixation density was measured with pixelwise correlation coefficient (CC), histogram-intersection similarity (SIM), and KL(GT||prediction), where higher CC and SIM and lower KL indicate better agreement. Means and paired differences used 95% confidence intervals from 10,000 image-level bootstrap resamples. Sensitivity analyses repeated the combined map with smoothing at 2% and 4% of the shorter side and within each of eight consecutive 125-image identifier blocks.

*Received: Januari 2026; Revised: February 2026; Accepted: March 2026; Published: April 2026*
*Corresponding author, qiwen.zheng.626@gmail.com

EAID ratings were aggregated to image-level means. Blank cells remained missing, and the Brand Liking value of -2 was excluded because the workbook defines it as an unrecognized brand rather than a low score. Brand associations were restricted to 667 images with at least five recognized-brand ratings. Spearman correlations used 2,000 image-level bootstrap resamples, and Benjamini–Hochberg correction controlled the false-discovery rate across 30 prespecified feature–rating tests. A sensitivity analysis removed fixed differences among the eight 125-image blocks and resampled within blocks.

Table 3. Descriptive ratings by principle and perturbation level.

| principle | perturbation label | n | human mean | human sd | gpt mean | gpt sd | mean white space ratio | mean edge density |
|---|---|---|---|---|---|---|---|---|
| alignment | original | 100 | 8.116 | 1.234 | 6.628 | 1.209 | 0.332 | 0.131 |
| alignment | small | 100 | 5.592 | 1.767 | 6.19 | 1.28 | 0.329 | 0.13 |
| alignment | medium | 100 | 4.846 | 1.876 | 5.746 | 1.464 | 0.331 | 0.128 |
| alignment | large | 100 | 3.71 | 1.651 | 5.04 | 1.239 | 0.335 | 0.124 |
| overlap | original | 100 | 7.122 | 1.102 | 6.64 | 1.226 | 0.332 | 0.131 |
| overlap | small | 100 | 6.14 | 1.7 | 5.798 | 1.536 | 0.312 | 0.141 |
| overlap | medium | 100 | 4.92 | 1.818 | 5.098 | 1.661 | 0.293 | 0.147 |
| overlap | large | 100 | 3.998 | 1.51 | 4.008 | 1.387 | 0.281 | 0.152 |
| white space | original | 100 | 6.962 | 0.948 | 6.33 | 1.033 | 0.332 | 0.131 |
| white space | small | 100 | 6.012 | 1.584 | 5.718 | 1.236 | 0.312 | 0.141 |
| white space | medium | 100 | 4.954 | 1.64 | 5.05 | 1.359 | 0.293 | 0.147 |
| white space | large | 100 | 4.242 | 1.729 | 4.344 | 1.321 | 0.281 | 0.152 |

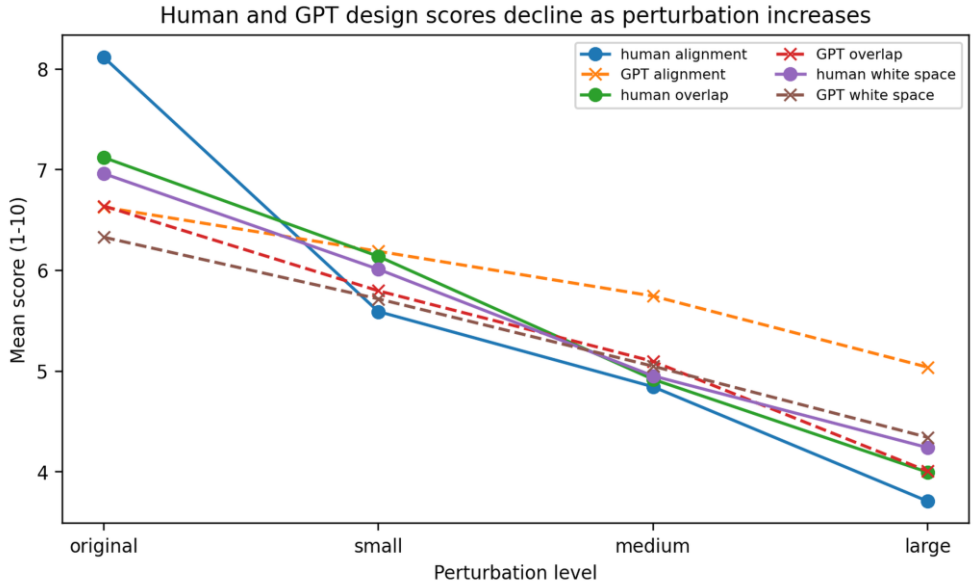


Figure 2. Mean human and GPT ratings by design principle and perturbation level.

## RESULTS

The first finding is that the experimental manipulations created a clear quality gradient. In Table 3 and Figure 2, human alignment scores declined from 8.116 for original banners to 3.710 for large perturbations. Human overlap scores declined from 7.122 to 3.998, and white-space scores declined from 6.962 to 4.242. GPT means followed the same monotonic direction for all

three principles, although the GPT scale was compressed relative to the human scale, especially for alignment. This compression is visible in Figure 3, where GPT scores understate high-quality originals and overstate some heavily perturbed alignment examples.

Table 4. Direct human-GPT agreement by design principle.

| principle | n | human mean | gpt mean | human rater SD mean | GPT run SD mean | RMSE | MAE | R2 | Pearson | Spearman |
|---|---|---|---|---|---|---|---|---|---|---|
| alignment | 400 | 5.566 | 5.901 | 1.566 | 0.669 | 2.003 | 1.647 | 0.244 | 0.524 | 0.519 |
| overlap | 400 | 5.545 | 5.386 | 1.466 | 0.623 | 1.274 | 0.983 | 0.573 | 0.771 | 0.769 |
| white space | 400 | 5.542 | 5.361 | 1.732 | 0.692 | 1.417 | 1.113 | 0.394 | 0.651 | 0.637 |

The experimental manipulations produced a consistent quality gradient. Original designs occupied the high end of the human scale, and larger perturbations produced progressively lower ratings. GPT scores followed the same ordering but used a narrower range, especially for alignment. The compression indicates that GPT scores are more dependable for relative screening than as direct substitutes for human numeric ratings.

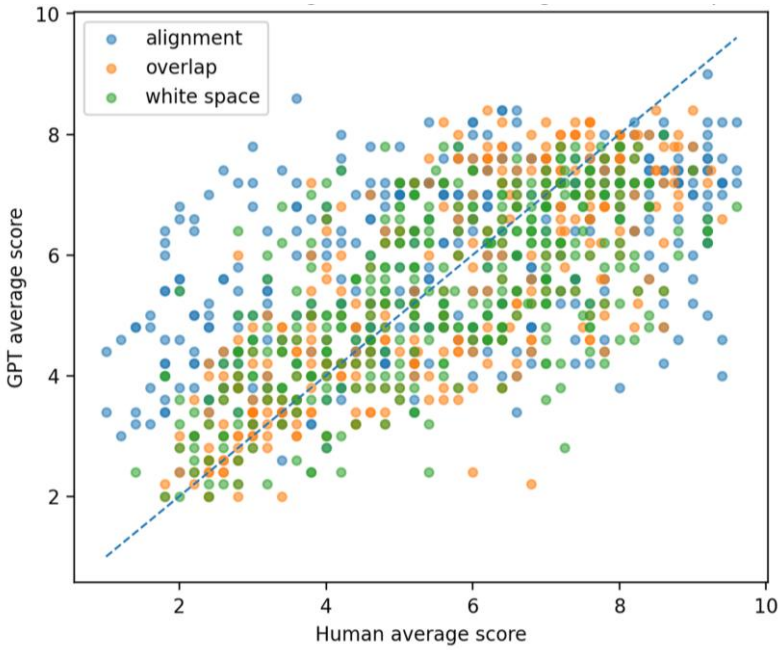


Figure 3. Direct scatterplot of human average scores and GPT average scores.

Table 4 and Figure 3 show substantial variation in direct human–GPT agreement by design principle. Overlap was strongest (Pearson = 0.771; Spearman = 0.769), white space was moderate (Pearson = 0.651; Spearman = 0.637), and alignment was weaker (Pearson = 0.524; Spearman = 0.519). Small alignment defects can depend on layer geometry or exact element coordinates that are difficult to infer from a flattened raster. GPT run dispersion was lower than human-rater

dispersion, showing stable model output without implying equal calibration or sufficient design judgment.

Table 5. Grouped five-fold model comparison for human rating prediction.

| feature family | model | RMSE | MAE | R2 | Pearson | Spearman |
|---|---|---|---|---|---|---|
| metadata_control | Ridge | 1.597 | 1.305 | 0.384 | 0.62 | 0.584 |
| metadata_control | Random Forest | 1.571 | 1.277 | 0.404 | 0.636 | 0.618 |
| metadata_control | XGBoost | 1.571 | 1.277 | 0.405 | 0.636 | 0.618 |
| visual_only | Ridge | 2.073 | 1.72 | -0.037 | 0.199 | 0.209 |
| visual_only | Random Forest | 2.007 | 1.689 | 0.028 | 0.246 | 0.217 |
| visual_only | XGBoost | 2.035 | 1.706 | 0.001 | 0.207 | 0.201 |
| gpt_only | Ridge | 1.566 | 1.257 | 0.408 | 0.639 | 0.634 |
| gpt_only | Random Forest | 1.629 | 1.295 | 0.359 | 0.605 | 0.599 |
| gpt_only | XGBoost | 1.585 | 1.268 | 0.394 | 0.628 | 0.62 |
| hybrid | Ridge | 1.653 | 1.305 | 0.34 | 0.603 | 0.605 |
| hybrid | Random Forest | 1.632 | 1.298 | 0.357 | 0.604 | 0.597 |
| hybrid | XGBoost | 1.625 | 1.295 | 0.362 | 0.608 | 0.601 |
| hybrid_plus_metadata | Ridge | 1.475 | 1.18 | 0.475 | 0.696 | 0.701 |
| hybrid_plus_metadata | Random Forest | 1.402 | 1.117 | 0.525 | 0.725 | 0.724 |
| hybrid_plus_metadata | XGBoost | 1.379 | 1.103 | 0.541 | 0.737 | 0.735 |

The grouped prediction results in Table 5 show that the best visual-only model, Random Forest, reached R2 = 0.028 and Spearman = 0.217. The GPT-only Ridge model was the best internally cross-validated model that did not use perturbation metadata, with RMSE = 1.566, MAE = 1.257, R2 = 0.408, Pearson = 0.639, and Spearman = 0.634. Adding the lightweight visual features did not improve the aggregate result; the best hybrid Spearman was 0.605. The hybrid-plus-metadata XGBoost model reached R2 = 0.541 and Spearman = 0.735, but perturbation severity is an experimental label and this result is therefore diagnostic rather than an operational estimate.

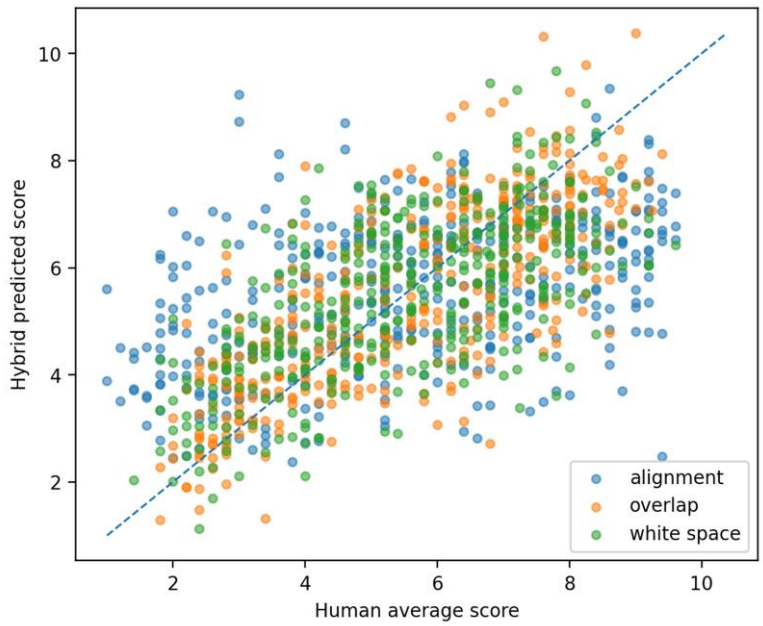


Figure 4. Cross-validated predicted scores versus human average scores for the selected hybrid model.

Table 6. Selected hybrid model performance by design principle.

| principle | n | RMSE | MAE | R2 | Pearson | Spearman |
|---|---|---|---|---|---|---|
| alignment | 400 | 2.111 | 1.719 | 0.161 | 0.434 | 0.441 |
| overlap | 400 | 1.32 | 1.053 | 0.542 | 0.75 | 0.747 |
| white space | 400 | 1.415 | 1.142 | 0.396 | 0.651 | 0.637 |

The gap between GPT-only and visual-only models suggests that human ratings depend on semantic and relational judgments among text, objects, and layout, not only on raster statistics. A practical review output should therefore pair a calibrated score with visible evidence such as collision, insufficient spacing, weak focal hierarchy, or risky CTA microcopy (Y. Li et al., 2025; Tu et al., 2025; Wu et al., 2025).

Table 7. Top predictors in the selected hybrid model.

| feature | importance |
|---|---|
| gpt_avg | 1.364 |
| content_bbox_area | 0.613 |
| edge_density | 0.471 |
| component_count | 0.387 |
| principle_alignment | 0.371 |
| text_area_proxy | 0.337 |
| component_area_gini | 0.258 |
| alignment_dispersion | 0.254 |
| gray_entropy | 0.229 |
| content_area_ratio | 0.211 |
| white_space_ratio | 0.211 |
| margin_min | 0.2 |

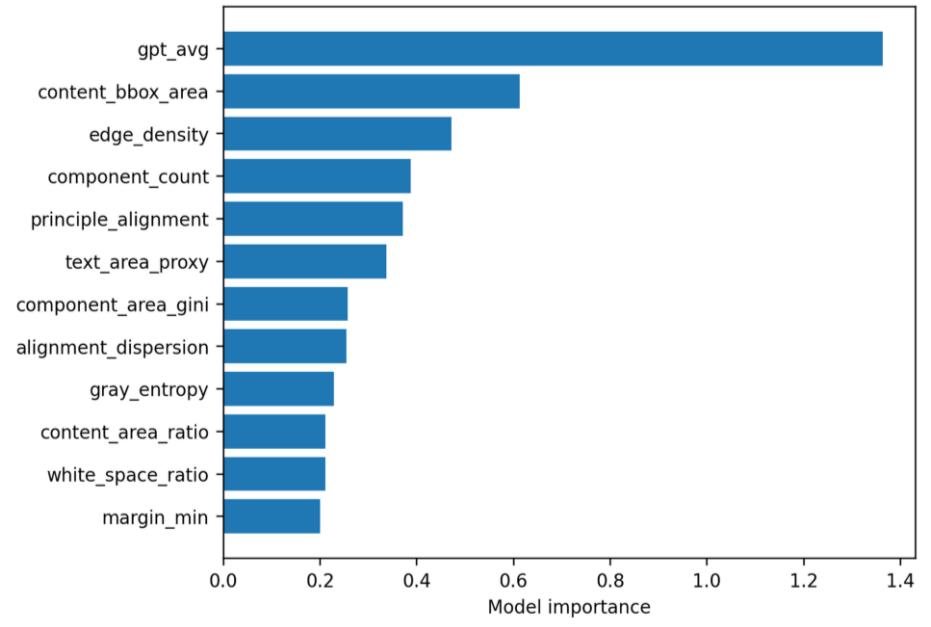


Figure 5. Feature importance profile of the selected hybrid model.

Model complexity did not improve performance automatically. When the GPT score carried most of the predictive signal, linear calibration outperformed the nonlinear alternatives in

the operational feature sets. The visual-only models remain useful as conservative baselines and as sources of interpretable evidence, but their limited predictive accuracy does not support using them as stand-alone design judges.

Table 8. Error analysis by design principle and perturbation level.

| principle | perturbation label | n | MAE | signed error | human mean | pred mean |
|---|---|---|---|---|---|---|
| alignment | original | 100 | 2.11 | -1.934 | 8.116 | 6.182 |
| alignment | small | 100 | 1.523 | 0.217 | 5.592 | 5.809 |
| alignment | medium | 100 | 1.625 | 0.557 | 4.846 | 5.403 |
| alignment | large | 100 | 1.619 | 1.105 | 3.71 | 4.815 |
| overlap | original | 100 | 1.007 | -0.381 | 7.122 | 6.741 |
| overlap | small | 100 | 1.288 | -0.194 | 6.14 | 5.947 |
| overlap | medium | 100 | 1.022 | 0.261 | 4.92 | 5.181 |
| overlap | large | 100 | 0.896 | 0.184 | 3.998 | 4.182 |
| white space | original | 100 | 1.097 | -0.468 | 6.962 | 6.494 |
| white space | small | 100 | 1.29 | -0.118 | 6.012 | 5.894 |
| white space | medium | 100 | 1.117 | 0.2 | 4.954 | 5.154 |
| white space | large | 100 | 1.063 | 0.263 | 4.242 | 4.505 |

Performance differed strongly by principle, as shown in Figure 4 and Table 6. The selected hybrid model predicted overlap best (R2 = 0.542; Spearman = 0.747), white space moderately (R2 = 0.396; Spearman = 0.637), and alignment weakly (R2 = 0.161; Spearman = 0.441). Overlap and spacing violations often create visible density changes, whereas small alignment errors can depend on exact layer coordinates. LLM/VLM judgments can therefore support semantic layout screening, but alignment decisions should retain layer measurements or human inspection.

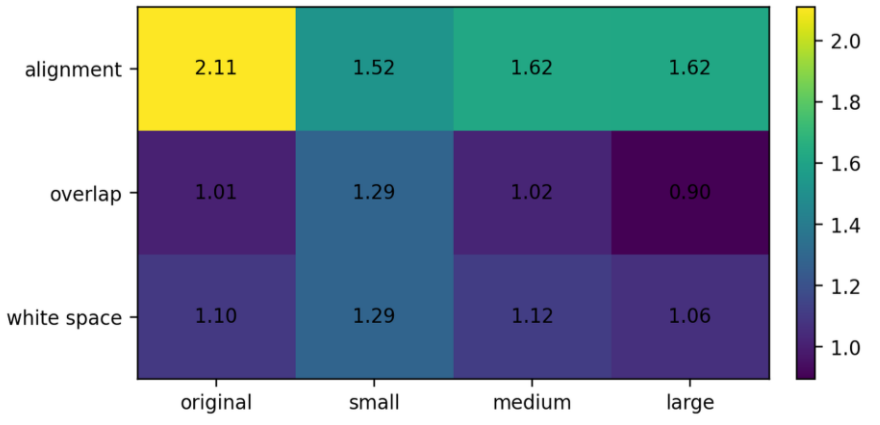


Figure 6. Mean absolute error by principle and perturbation level.

The importance profile in Table 7 and Figure 5 reinforces this interpretation. The GPT average score was the dominant predictor. Among visual features, content bounding-box area, edge density, component count, text-area proxy, component-area inequality, alignment dispersion, entropy, and white-space ratio contributed measurable signal. These variables

correspond to design concerns that human reviewers often discuss in qualitative terms: clutter, readable grouping, spacing, and whether a focal area can be distinguished from background information.

Table 9. Agreement with aggregate ADD1000 fixation-density maps.

| Map | n | CC, mean [95% CI] | SIM, mean [95% CI] | KL(GT‖P), mean [95% CI] |
|---|---|---|---|---|
| Center prior | 1,000 | 0.416 [0.406, 0.427] | 0.542 [0.537, 0.547] | 0.725 [0.709, 0.741] |
| Gradient | 1,000 | 0.460 [0.449, 0.470] | 0.560 [0.555, 0.564] | 0.679 [0.665, 0.694] |
| Local contrast | 1,000 | 0.437 [0.425, 0.448] | 0.558 [0.553, 0.563] | 0.642 [0.629, 0.657] |
| Combined | 1,000 | 0.457 [0.445, 0.467] | 0.562 [0.557, 0.567] | 0.637 [0.623, 0.651] |

Table 8 and Figure 6 locate the largest prediction errors. Original alignment designs had the highest MAE (2.110) because the model predicted them too low, while heavily perturbed alignment designs were predicted too high. Overlap errors were lower and more balanced, including an MAE of 0.896 for large perturbations. White-space errors were moderate across levels. The framework is therefore most credible for screening clear overlap and spacing problems; alignment requires closer geometric or human review.

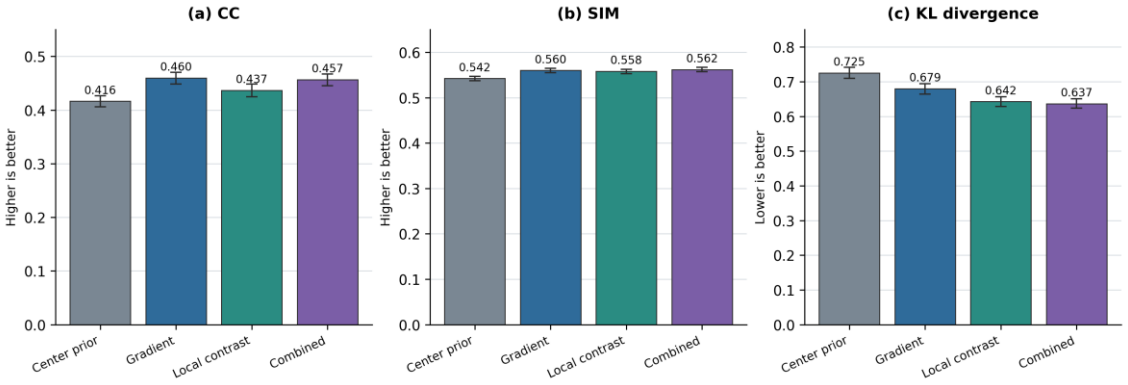


Figure 7. Proxy-map agreement with aggregate ADD1000 fixation density; bars show image-bootstrap 95% confidence intervals.

Table 10. Selected image-level associations with EAID human ratings.

| Image feature | EAID rating | n | Spearman rho [95% CI] | BH-FDR q | Within-block rho |
|---|---|---|---|---|---|
| Fixation entropy | Aesthetic | 1,000 | -0.319 [-0.374, -0.261] | < .001 | -0.318 |
| Gradient energy | Aesthetic | 1,000 | -0.310 [-0.366, -0.253] | < .001 | -0.290 |
| Fixation entropy | Ad liking | 1,000 | -0.292 [-0.348, -0.234] | < .001 | -0.290 |
| Proxy central mass | Aesthetic | 1,000 | 0.278 [0.223, 0.335] | < .001 | 0.265 |
| Gradient energy | Ad liking | 1,000 | -0.249 [-0.309, -0.190] | < .001 | -0.242 |
| Proxy central mass | Ad liking | 1,000 | 0.220 [0.161, 0.279] | < .001 | 0.214 |

ADD1000 provides a direct test of the spatial claims. Table 9 and Figure 7 compare the training-free proxy maps with aggregate human fixation density. The center prior achieved mean CC = 0.416 (95% CI [0.406, 0.427]), SIM = 0.542 [0.537, 0.547], and KL = 0.725 [0.709, 0.741].

The gradient proxy produced the highest CC, 0.460 [0.449, 0.470], while the combined proxy produced the highest SIM, 0.562 [0.557, 0.567], and the lowest KL, 0.637 [0.623, 0.651].

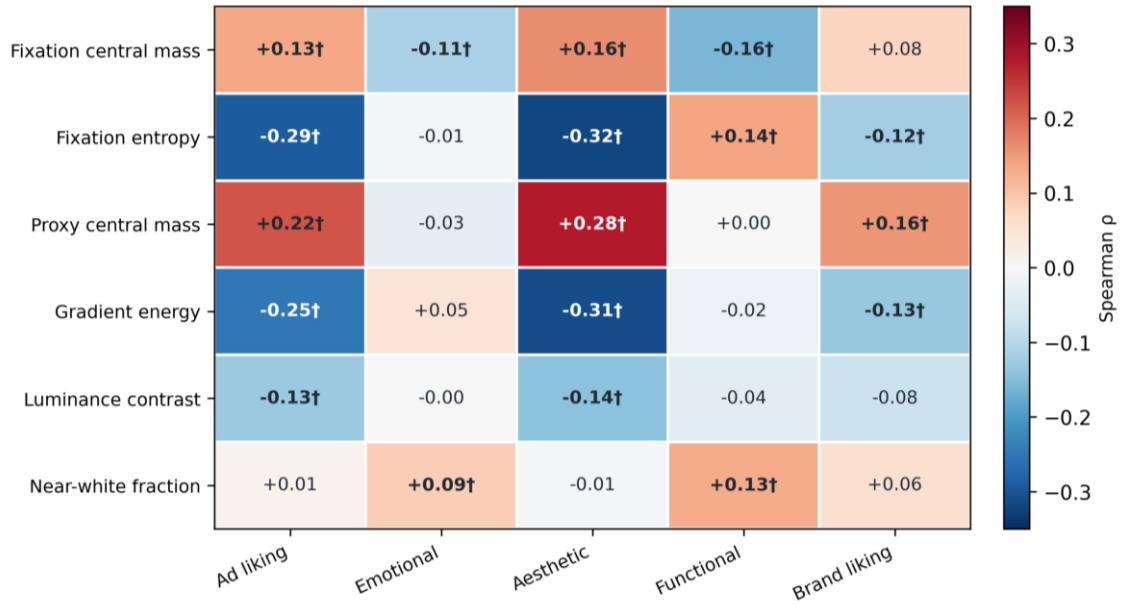


Figure 8. Spearman associations between attention-related image features and EAID human ratings. Dagger symbols indicate BH-FDR q < .05 across 30 tests; Brand Liking uses 667 advertisements with at least five recognized-brand ratings.

Relative to the center prior, the combined proxy improved CC by 0.040 [0.026, 0.054] and SIM by 0.020 [0.014, 0.025] and reduced KL by 0.088 [0.072, 0.105]. The direction of improvement was the same in all eight 125-image blocks and under the 2% and 4% smoothing settings. Combined-proxy central mass also correlated with human-fixation central mass (Spearman = 0.413 [0.356, 0.467]). These results show limited but consistent spatial information beyond center bias; the absolute agreement remains moderate.

Table 11. BannerRequest400 brief and logo summary.

| metric | value |
|---|---|
| Abstract requests | 400 |
| Unique target-audience phrases | 223 |
| Median abstract-request words | 64 |
| Concrete size slots | 5,200 |
| Nonempty concrete specifications | 4,000 |
| Nonempty specification rate | 0.769 |
| Median nonempty-request words | 128 |
| CTA-present rate | 0.75 |
| Standard sizes | 13 |
| Median logo aspect ratio | 2.473 |
| Median logo foreground ratio | 0.287 |

EAID contains 57,000 valid Ad Liking ratings, 44,649 Emotional ratings, 47,922 Aesthetic ratings, 45,567 Functional ratings, and 18,928 recognized-brand ratings. Its 29,198 Brand Liking values coded -2 were treated as unrecognized brands, and blank cells remained missing. Table 10 reports the strongest prespecified associations, while Figure 8 shows the full 30-test pattern. Lower fixation entropy was associated with higher Aesthetic ratings (Spearman = -0.319 [-0.374, -0.261]) and Ad Liking (-0.292 [-0.348, -0.234]). Greater proxy central mass was associated with Aesthetic ratings (0.278 [0.223, 0.335]) and Ad Liking (0.220 [0.161, 0.279]).

Table 12. BannerRequest400 standard-size distribution.

| size | format class | width | height | area | aspect ratio | n | nonempty specs | mean words | CTA rate |
|---|---|---|---|---|---|---|---|---|---|
| 970x250 | rectangle | 970 | 250 | 242,500 | 3.88 | 400 | 400 | 137 | 0.983 |
| 300x600 | vertical skyscraper | 300 | 600 | 180,000 | 0.5 | 400 | 400 | 137 | 0.973 |
| 160x600 | vertical skyscraper | 160 | 600 | 96,000 | 0.267 | 400 | 400 | 134 | 0.983 |
| 336x280 | rectangle | 336 | 280 | 94,080 | 1.2 | 400 | 400 | 132 | 0.978 |
| 970x90 | wide leaderboard | 970 | 90 | 87,300 | 10.778 | 400 | 400 | 135 | 0.97 |
| 300x250 | rectangle | 300 | 250 | 75,000 | 1.2 | 400 | 400 | 131 | 0.98 |
| 120x600 | vertical skyscraper | 120 | 600 | 72,000 | 0.2 | 400 | 400 | 134 | 0.98 |
| 728x90 | wide leaderboard | 728 | 90 | 65,520 | 8.089 | 400 | 400 | 134 | 0.97 |
| 250x250 | rectangle | 250 | 250 | 62,500 | 1 | 400 | 400 | 129 | 0.958 |
| 320x100 | compact mobile | 320 | 100 | 32,000 | 3.2 | 400 | 0 | 0 | 0 |
| 468x60 | wide leaderboard | 468 | 60 | 28,080 | 7.8 | 400 | 400 | 132 | 0.975 |
| 180x150 | compact mobile | 180 | 150 | 27,000 | 1.2 | 400 | 0 | 0 | 0 |
| 300x50 | wide leaderboard | 300 | 50 | 15,000 | 6 | 400 | 0 | 0 | 0 |

The strongest EAID associations remained similar after removing fixed differences among the eight 125-image blocks. They are observational image-level relationships, however, and do not establish that concentrated fixation causes liking or aesthetic quality.

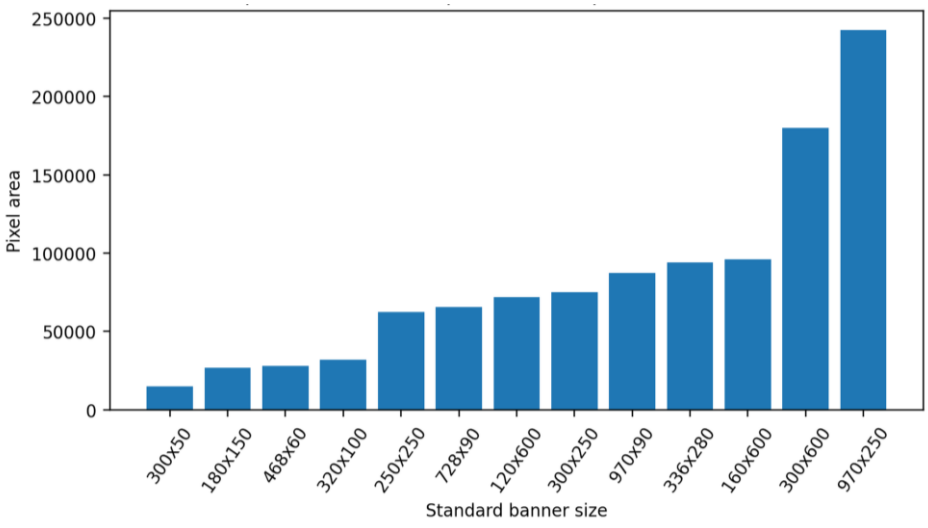


Figure 9. Pixel area across BannerRequest400 standard banner sizes.

BannerRequest400 provides design context rather than outcome validation. Table 11 summarizes its 400 requests and 100 logo records; Table 12 and Figure 9 show the 13 standard size slots. The abstract requests had a median length of 64 words and 223 unique target-audience

phrases. The concrete structure contained 5,200 size slots, including 4,000 nonempty specifications with a median of 128 words. CTA language appeared in 75.0% of all slots. These distributions show why a review framework must handle wide leaderboards, rectangles, vertical skyscrapers, and compact formats. Because the dataset contains requests rather than judged rendered banners, it contributes format and communication requirements but no labels to the prediction or fixation analyses.


Figure 10. Representative high-, low-, over-predicted, and under-predicted cases from grouped cross-validation.

The brief analysis also shows that design criteria vary with canvas shape and communication purpose. A 970x250 leaderboard can separate brand and CTA horizontally, whereas a 160x600 skyscraper requires vertical sequencing. Language-only analysis cannot determine whether the final CTA has enough visual separation, and raster statistics cannot recover the campaign intent in the brief. BannerRequest400 therefore informs the scope of a review interface without serving as evidence that the scoring framework predicts human judgment.

The CTA finding should also be read ethically. A clearly separated button is not automatically a good communication outcome if the microcopy pressures users, hides conditions, or creates a dark pattern. Automated dark-pattern detection from interface microcopy therefore complements visual attention analysis by adding a language-based check on whether persuasive CTA design remains transparent and user-respecting (H. Xu et al., 2025). For product-oriented banners, explainable e-commerce recommendation cards and review-grounded recommendation evaluations suggest a similar extension: link product claims, CTA wording, and visual prominence back to evidence that a reviewer can inspect (Chang et al., 2026; K. Xu et al., 2024; B. Zhang et al., 2025). Figure 10 illustrates the calibration pattern behind the aggregate errors: clearly strong and weak designs can be separated, while some high-quality alignment examples are under-predicted and some weak examples are over-predicted. The cases reinforce the need to show the score together with principle-specific evidence and retain human review for ambiguous layouts.

## CONCLUSION

This study finds that LLM-assisted scores can support human-supervised review of graphic advertising design, but their performance depends strongly on the principle being judged. Overlap showed strong direct agreement with human ratings, white space showed moderate agreement, and alignment remained weaker. The GPT-only Ridge model was the best internally cross-validated operational model, while lightweight visual features were more useful as interpretable evidence than as stand-alone predictors. Perturbation metadata improved a diagnostic model but is unavailable in ordinary design review. The ADD1000 analysis places the attention claim on a more direct empirical basis. Gradient and local-contrast maps consistently outperformed a center prior, yet their agreement with aggregate fixation density was only moderate. EAID associations linked more concentrated fixation and proxy centrality with aesthetic and liking ratings, but these observational relationships do not establish causal advertising effectiveness. Design-quality scores, fixation-density agreement, and CTR, VCR, brand lift, or sales outcomes therefore remain separate layers of evidence.

Campaign-level incrementality, uplift, and off-policy evaluation can test whether reviewed creatives improve response or sales outcomes (Bai & Wu, 2026; Y. Lu et al., 2025), but these methods should be added only after design-quality measures have been calibrated against human judgment (Bai et al., 2026; Y. Lu et al., 2024; Mu et al., 2023; Mu et al., 2025; Ye et al., 2026). BannerRequest400 can inform the brief and format constraints used in that workflow, but it cannot validate the scoring model because it has no human quality labels, gaze measurements, or judged rendered banners. A practical implementation can use three linked layers. First, LLM/VLM scores can screen for principle-specific risks, with lower confidence assigned to alignment. Second, interpretable image evidence can locate density, spacing, component, or focal-balance issues, while fixation-based evidence is used only when gaze data are available. Third, an evidence card (Nie et al., 2026; Y. Zhang & Zhang, 2025b) can distinguish what is supported by the image, the brief, human ratings, or eye tracking, leaving the final creative decision to a reviewer (Jin, 2025a; C. Li et al., 2024; C. Li et al., 2023; Zhong et al., 2025). Future work should add direct labels for readability, CTA clarity, brand recognition, ethical microcopy, and hierarchy; obtain participant-level gaze or area-of-interest measures; and test whether these creative measures predict brand or conversion outcomes. The present evidence supports a preliminary design-review framework, not a replacement for human creative evaluation or real advertising-performance metrics.

## REFERENCES

Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the

Hillstrom E-mail experiment. Journal of Technology Informatics and Engineering, 5(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. International Journal of Electronics and Communications System, 6(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. Journal of Electrical Engineering and Computer Science, 11(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

Chang, X., Ye, T., & Luo, S. (2023). LLM-as-reranker for personalized recommendation: Popularity bias mitigation and faithful natural-language explanations on MovieLens 100K. Journal of Advanced Computing Systems, 3(8), 61–78. https://doi.org/10.69987/JACS.2023.30806

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. Journal of Technology Informatics and Engineering, 4(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490

Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. International Journal of Graphic Design, 4(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Haraguchi, D., Inoue, N., Shimoda, W., Mitani, H., Uchida, S., & Yamaguchi, K. (2024). Can GPTs evaluate graphic design based on design principles? In SIGGRAPH Asia 2024 Technical Communications (Article 5, pp. 1–4). Association for Computing Machinery. https://doi.org/10.1145/3681758.3698010

IAB Europe. (2023). Guide to attention in digital marketing. https://iabeurope.eu/wp-content/uploads/IAB-Europe-Guide-to-Attention-in-Digital-Marketing-25.05-1-1.pdf

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). https://doi.org/10.1109/CVPR52729.2023.00980

Interactive Advertising Bureau. (2024). Attention measurement explainer: Data signal approaches. https://www.iab.com/wp-

content/uploads/2024/08/IAB_Attention_Measurement_Explainer_August_2024.pdf

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259. https://doi.org/10.1109/34.730558

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. (2023). COLE: A hierarchical generation framework for multi-layered and editable graphic design [Preprint]. arXiv. https://arxiv.org/abs/2311.16974

Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. Journal of Technology Informatics and Engineering, 4(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. International Journal of Graphic Design, 3(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

Kong, W., Jiang, Z., Sun, S., Guo, Z., Cui, W., Liu, T., Lou, J., & Zhang, D. (2023). Aesthetics++: Refining graphic designs by exploring design principles and human preference. IEEE Transactions on Visualization and Computer Graphics, 29(6), 3093–3104. https://doi.org/10.1109/TVCG.2022.3151617

Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. Journal of Technology Informatics and Engineering, 4(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

Lavie, N. (2005). Distracted and confused? Selective attention under load. Trends in Cognitive Sciences, 9(2), 75–82. https://doi.org/10.1016/j.tics.2004.12.004

Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. Journal of Advanced Computing Systems, 4(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

Li, C., Su, W., & Zhang, E. (2023). Lightweight hallucination firewall for enterprise LLM applications: Evidence consistency, self-checking, and small-model detection on TruthfulQA. Journal of Advanced Computing Systems, 3(1), 49–65. https://doi.org/10.69987/JACS.2023.30104

Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. International Journal of Graphic Design, 3(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. International Journal of Graphic Design, 3(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

Liang, S., Liu, R., & Qian, J. (2021). Fixation prediction for advertising images: Dataset and benchmark. Journal of Visual Communication and Image Representation, 81, 103356. https://doi.org/10.1016/j.jvcir.2021.103356

Liang, S., Liu, R., & Qian, J. (2024). EAID: An eye-tracking based advertising image dataset with personalized affective tags. In B. Sheng, L. Bi, J. Kim, N. Magnenat-Thalmann, & D. Thalmann (Eds.), Advances in computer graphics: CGI 2023 (Lecture Notes in Computer Science, Vol. 14495, pp. 282–294). Springer. https://doi.org/10.1007/978-3-031-50069-5_24

Lidwell, W., Holden, K., & Butler, J. (2010). Universal principles of design (2nd ed.). Rockport Publishers.

Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. International Journal of Graphic Design, 3(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

Lu, S., & Zhou, D. (2023). LLM-augmented customer representation learning for next-purchase prediction in online retail. Journal of Advanced Computing Systems, 3(3), 50–64. https://doi.org/10.69987/JACS.2023.30305

Lu, Y., Mu, J., & Tran, T. (2024). Uncertainty-aware uplift modeling for safer marketing targeting: Conformal prediction and Bayesian calibration with LCB policies. Journal of Advanced Computing Systems, 4(5), 84–101. https://doi.org/10.69987/JACS.2024.40507

Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. Journal of Technology Informatics and Engineering, 4(3), 493–520. https://doi.org/10.51903/jtie.v4i3.466

Lupton, E. (2010). Thinking with type: A critical guide for designers, writers, editors, & students (2nd ed.). Princeton Architectural Press.

Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. Journal of Technology Informatics and Engineering, 5(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. International Journal of Graphic Design, 4(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience-creative-channel policies. Journal of Advanced Computing Systems, 3(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset

(Small). Journal of Technology Informatics and Engineering, 4(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. International Journal of Graphic Design, 4(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

O'Donovan, P., Agarwala, A., & Hertzmann, A. (2014). Learning layouts for single-page graphic designs. IEEE Transactions on Visualization and Computer Graphics, 20(8), 1200–1213. https://doi.org/10.1109/TVCG.2014.48

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Pieters, R., & Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. Journal of Marketing, 68(2), 36–50. https://doi.org/10.1509/jmkg.68.2.36.27794

Pieters, R., Wedel, M., & Batra, R. (2010). The stopping power of advertising: Measures and effects of visual complexity. Journal of Marketing, 74(5), 48–60. https://doi.org/10.1509/jmkg.74.5.48

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. Journal of Vision, 7(2), Article 17. https://doi.org/10.1167/7.2.17

Solikhan, M., Priyadi, A., MacDonald, D., & Noramai, A. (2026). Generative AI as a Live Design Mentor: A Mixed-Reality Approach to Graphic Design Education. International Journal of Graphic Design, 4(1), 123–140. https://doi.org/10.51903/IJGD.V4I1.2375

Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. Journal of Technology Informatics and Engineering, 5(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. Journal of Technology Informatics and Engineering, 4(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. International Journal of Graphic Design, 4(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

Sun, J., Xiao, X., & Dong, H. (2023). From tell-to-design to healthcare test-fit constraint checking: An LLM-compatible requirements-to-constraints framework. Journal

of Advanced Computing Systems, 3(11), 52–67. https://doi.org/10.69987/JACS.2023.31105

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. International Journal of Graphic Design, 3(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

Tufte, E. R. (2001). The visual display of quantitative information (2nd ed.). Graphics Press.

Wang, H., Ren, Y., & Chang, X. (2025). Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD. Journal of Technology Informatics and Engineering, 4(2), 425–446. https://doi.org/10.51903/jtie.v4i2.523

Wang, H., Shimose, Y., & Takamatsu, S. (2025). BannerAgency: Advertising banner design with multimodal LLM agents. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 4304–4329). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.214

Wedel, M., & Pieters, R. (2008). A review of eye-tracking research in marketing. In N. K. Malhotra (Ed.), Review of marketing research (Vol. 4, pp. 123–147). Emerald Group Publishing. https://doi.org/10.1108/S1548-6435(2008)0000004009

Williams, R. (2014). The non-designer's design book (4th ed.). Peachpit Press.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience, 5, 495–501. https://doi.org/10.1038/nrn1411

Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. International Journal of Graphic Design, 3(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

Xin, Q. (2026). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. Journal of Information and Technology, 14(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern Dataset. Journal of Technology Informatics and Engineering, 4(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent classification and personalized recommendation of e-commerce products based on machine

learning. In Proceedings of the 6th International Conference on Computing and Data Science (pp. 143–149). https://doi.org/10.54254/2755-2721/64/20241365

Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). Journal of Technology Informatics and Engineering, 5(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. International Journal of Graphic Design, 3(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. Journal of Technology Informatics and Engineering, 4(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. International Journal of Graphic Design, 3(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. Journal of Advanced Computing Systems, 4(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. Journal of Advanced Computing Systems, 3(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. Journal of Advanced Computing Systems, 4(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. Journal of Technology Informatics and Engineering, 4(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. Journal of Technology Informatics and Engineering, 5(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. International Journal of Graphic Design, 3(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

Zhou, B., Wang, H., & Chang, X. (2025). Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens. Journal of Technology Informatics and Engineering, 4(2), 447–463. https://doi.org/10.51903/jtie.v4i2.522