Leveraging Large Models to Evaluate Novel Content
A Case Study on Advertisement Creativity

Zhaoyi Joey Hou, Adriana Kovashka, Xiang Lorraine Li
Department of Computer Science
University of Pittsburgh
{joey.hou, kovashka, xianglli}@pitt.edu

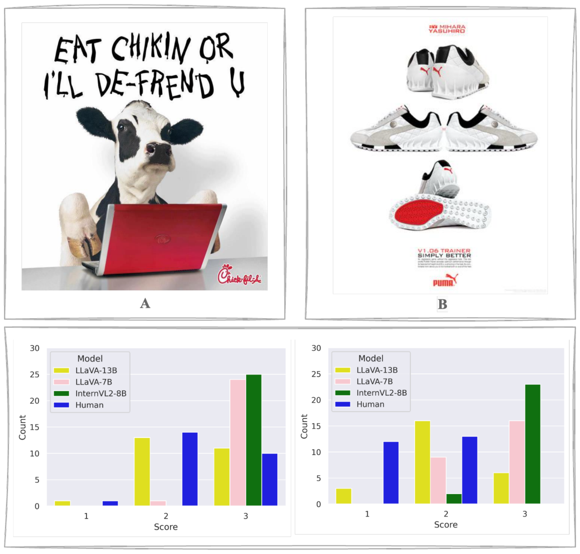
Abstract

Evaluating creativity is challenging, even for humans, not only because of its subjectivity but also because it involves complex cognitive processes. Inspired by work in marketing, we attempt to break down visual advertisement creativity into atypicality and originality. With fine-grained human annotations on these dimensions, we propose a suite of tasks specifically for such a subjective problem. We also evaluate the alignment between state-of-the-art (SoTA) vision language models (VLMs) and humans on our proposed benchmark, demonstrating both the promises and challenges of using VLMs for automatic creativity assessment[1].

1 Introduction

Creativity, the ability to generate novel and high-quality ideas, is a fundamental human cognitive ability. Research in a variety of domains has been devoted to understanding creativity, including psychology focusing on human creativity (Olson et al., 2021; Guilford et al., 2012; Alabbasi et al., 2022), machine learning on machine creativity (Ismayilzada et al., 2024a; Franceschelli and Musolesi, 2024), human-computer interaction (HCI) on the combination of both (Lu et al., 2025a; Chakrabarty et al., 2024a; Porter and Machery, 2024; Marco et al., 2024), and marketing on advertisement creativity (Smith et al., 2007; El-Murad and West, 2004a; Rosengren et al., 2020).

In particular, in the marketing domain, studies have shown positive effects of creative advertisements on consumer behavior, including increased purchase intent and positive brand impression (Sharma, 2012; Terkan, 2014). Therefore, advertisement creators are motivated to develop and evaluate creative advertisement content consistently. Extensive research has been conducted to



	Human Avg. (A / B)	LLaVA-7B (A / B / pairwise)	LLaVA-13B (A / B / pairwise)	InternVL2-8B (A / B / pairwise)	GPT-4v (pairwise)
Creativity	2.40 / 1.56	2.96 / 2.44 / A	2.36 / 2.16 / A	3.00 / 3.00 / A	A
Originality	2.72 / 1.56	2.88 / 2.60 / A	2.72 / 1.88 / A	3.00 / 2.80 / A	A
Atypicality	2.92 / 1.24	3.00 / 1.68 / A	2.40 / 1.52 / A	3.00 / 2.16 / A	A

Figure 1: Top: two advertisements from the dataset; Middle: human and VLM rating distribution on creativity (25 each); Bottom: average human and VLM ratings, and VLM pairwise predictions on creativity, originality, and typicality; Scores are on a scale of 1 to 3, 3 being the best. Both humans and models agree that advertisement **A** is more creative than **B**; however, their score distributions differ substantially, indicating the importance of measuring it.

understand what the general public would consider as creative when it comes to advertisements (El-Murad and West, 2004b; Rosengren et al., 2020; Swee Hoon Ang and Lou, 2014; Smith et al., 2007), many of which consider advertisement creativity to be a combination of "atypicality" (or, synthesis, abnormality, etc.) and "originality" (or, novelty, uniqueness, etc.). Advertisement (A) in Figure 1 is an example that is both atypical and original. It is atypical because cows do not usually type on a computer. The text *"Eat chikin or I'll de-friend U"* (which consists of multiple slang expressions), the

[1]https://github.com/JoeyHou/ads-creativity

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21158–21177

cow, the computer, and the small logo of Chick-fil-A are a rarely-seen combination, given that it is a fast-food advertisement, making it very original. Decoding advertisement creativity with such a complex combination of concepts and ideas requires extensive reasoning, making the evaluation of creativity a challenging task. Unsurprisingly, previous work (Chakrabarty et al., 2024a; Tian et al., 2024) has heavily relied on human evaluation, but human annotators are expensive and often inaccessible.

Recently, foundational models demonstrate impressive performances in other evaluation tasks, such as summarization (Zhong et al., 2022), Long-Form QA (Jiang et al., 2024), and commonsense text generation (Xu et al., 2023), many of which were previously dominated by humans. For creativity evaluation, prior work (Chakrabarty et al., 2024a; Atmakuru et al., 2024; Ismayilzada et al., 2024b) explored the ability of large language models (LLMs) to assess writing creativity. This poses the question of whether we can leverage foundational models to expand automatic evaluation of creativity to multi-modal data, such as visual advertisements, with the help of VLMs.

To this end, we conduct several fine-grained, automatic creativity evaluations for visual advertisements, including creativity, originality, and atypicality. We decompose creativity into atypicality and originality and then collect high-quality human ratings of advertisements in those dimensions, as shown in Figure 1. We experiment with SoTA VLMs to predict these ratings and examine the human-model alignment in both intrinsic (i.e., one image at a time) and pairwise fashion. In contrast to the traditional emphasis on prediction accuracy, we evaluate models' ability to capture the task's subjective nature and to gauge annotator disagreements. We also show that VLMs perform impressively in a pairwise setting, reasonably well in distribution modeling, and less effectively in disagreement. Our results highlight the importance of both high-quality annotations and distribution modeling for subjective tasks. We believe our benchmark and evaluation metrics provide a solid foundation for utilizing VLMs to assist visual content creators.

2 Related Work

Evaluation of Creativity Research in the evaluation of creativity includes cognitive science (Said-Metwaly et al., 2017; Simonton, 2012; James Lloyd-Cox and Bhattacharya, 2022), marketing (El-Murad and West, 2004b; Rosengren et al., 2020; Swee Hoon Ang and Lou, 2014; Smith et al., 2007), creative writing (Skalicky, 2022), human computer interaction (Chakrabarty et al., 2024b), and artificial intelligence (Chakrabarty et al., 2023, 2024a; Atmakuru et al., 2024; Ismayilzada et al., 2024b). There are two consensuses among those works: First, creativity requires both effectiveness (i.e., clearly conveying a message or solving a problem) and divergence (i.e., diverging from the norm or commonly seen objects). We ensure effectiveness via quality check questions in human annotation and mainly focus on modeling divergence with atypicality and originality. Second, the evaluation of creativity is subjective. This motivates our distribution modeling task instead of a majority-label prediction task. Our work is closely related to Smith et al. (2007), which focused on advertisement images and proposed five creativity dimensions, including atypicality and originality. We adopt their creativity decomposition.

Dimension	Questions	Answer
Atypicality	The ad connected usually unrelated objects The ad contained unusual connection The ad brought unusual items together	agree (1), neutral (0), disagree (-1)
Originality	The ad was out of the ordinary The ad broke away from habit-bound and stereotypical thinking The ad was unique	agree (1), neutral (0), disagree (-1)
Creativity	What is the overall level of creativity of this advertisement?	integer (1-5)

Table 1: Questions in Amazon Mechanical Turk; full annotation interface and instructions is in Figure 4.

Automatic Evaluation of Creativity with Foundation Models There has been much recent work that automatically evaluates the creativity of given content with foundational models. For example, Ye et al. (2025); Lu et al. (2025b) evaluate creativity in problem-solving tasks (i.e., math and coding); Chakrabarty et al. (2024a); Ismayilzada et al. (2024b); Lu et al. (2024); Atmakuru et al. (2024) focus on creative writing tasks; Organisciak et al. (2023); Goes et al. (2023); Chen and Ding (2023); Pépin et al. (2024); Zhao et al. (2025) study how to apply psychological tests to evaluate LLM creativity; and Zhang et al. (2025); McLaughlin et al. (2024) study creative question answering. All of these focus on text-only data, while we focus on visual content. One related work that also works with multi-modal data is Zhong et al. (2024), where the focus is on creative humor generation, instead of creative advertisement understanding.

3 Dataset

We use the Pitt Advertisement Dataset (referred to as `Pitt-Ads`) as our basis, where each advertisement image is annotated with its topic (10 topic groups in total), and expected actions from viewers after seeing the advertisement. Additionally, the advertisements are annotated with the category of atypical objects in it (Hussain et al., 2017; Ye et al., 2019)[2]. We sample 100 advertisements and collect fine-grained human creativity annotations (`Creative-100`). We also sample an additional 300 advertisements from the remaining data points for atypicality prediction (`Atypical-300`).

3.1 `Creative-100`

`Creative-100` consists of 100 advertisements, with 10 from each topic group: food, drinks, clothing, non-commercial, automobile, beauty, service, leisure, electronic, and healthcare. For fine-grained creativity evaluation, we include two dimensions of creativity: originality and atypicality, which are the top two indicative aspects of advertisement creativity according to Smith et al. (2007) (more details in Appendix A.2). These two dimensions (originality, atypicality) and the overall creativity combined are the fine-grained creativity label of the advertisements in `Creative-100`.

We utilize Amazon Mechanical Turk (MTurk) to collect the fine-grained annotations. For atypicality and originality, we follow Smith et al. (2007) and record responses about various statements (Table 1). For creativity, we record a 5-scale score and convert it to a 3-scale one, aligning with other dimensions. Creativity is measured on a different scale because we believe there are subtle differences between "very creative" and "somewhat creative". To ensure the annotators actually understand the advertisement, we also include an annotation quality check question, asking annotators to choose the action after seeing a given advertisement (e.g., "I should go to Chick-fil-A" for advertisement A in Figure 1). In this question, five actions are given, with one correct action and four incorrect ones sampled from `Pitt-Ads`. Annotators get 96.88% accuracy on this question, highlighting their accurate understanding of visual advertisements. Detailed annotation interface and instructions are in Figure 4.

Unlike the traditional data annotation process that only collects a few annotations per data point, we choose to collect a larger number of annotations per data point. Thus, the predictive target is not a binary label (e.g., "creative" v.s. "not creative") but a distribution of human ratings (e.g., "10% chooses creative, 50% neutral, 40% not creative"). We believe this would better capture the inherent subjectivity of creativity judgment by showing the diverse perspectives from the annotators. In practice, we collect 25 annotations per advertisement image to approximate the true rating distribution within a certain error rate (McHugh, 2012). Refer to Appendix B.2 for more details on how we choose the exact number of annotations; more dataset construction details are in Appendix B.

3.2 `Atypical-300`

We also randomly sampled 300 advertisements (`Atypical-300`) from `Pitt-Ads` to specifically examine the atypicality dimension. In this subset, 185 instances (62%) contain atypical object(s) and are assigned a positive label for "atypicality,". The remaining 115 cases (38%) are labeled negative. `Atypical-300` is a larger and additional dataset similar to `Creative-100` but with only one dimension, atypicality. Different from `Creative-100`, each advertisement here only has three binary annotations on atypicality. Both Smith et al. (2007) and `Creative-100` (Appendix B.4) show that atypicality has a positive correlation with creativity. Thus, we include this dataset to further study VLM's ability to evaluate advertisement creativity.

4 Experimental Setup

4.1 Models

We experiment with open-sourced VLMs, InternVL2-8B (Chen et al., 2024), LLaVA-7B and 13B (Li et al., 2024), and close-sourced VLMs, GPT4-v (OpenAI et al., 2024). All experiments are done with zero-shot prompting with inference framework vllm [3] (Kwon et al., 2023) . More details are in Appendix C.

4.2 Task Formulation

We define the following three tasks for each dimension (i.e., creativity, originality, and atypicality):

Distribution Modeling is designed to evaluate the model's ability to simulate human group behav-

[2]Defined by the original authors of the dataset as "non-photorealistic" objects within an advertisement, e.g., a farm inside of a cup of ice cream, demonstrating the freshness of milk in the ice cream (Figure 3). An advertisement can feature any number of atypical objects or none at all.

[3]vllm 0.6.1.post1

Dimension	Model	Distribution Modeling		Disagreement	Pairwise		
		Rating Correlation	Distribution Divergence		All	Easy	Hard
		↑ *R (p-value)*	↓ *KL Divergence*	↑ *R (p-value)*	↑ *F1*	↑ *F1*	↑ *F1*
Creativity (`Creative-100`)	LLaVA-7B	0.65 (.00*)	1.01	0.06 (.52)	0.78	0.80	0.75
	LLaVA-13B	0.65 (.00*)	**0.37**	*nan*	0.67	0.69	0.65
	InternVL2-8B	**0.72** (.00*)	1.45	**0.20** (.05)	0.80	0.85	0.75
	GPT-4v	-	-	-0.04 (.72)	**0.97**	**0.98**	**0.96**
Originality (`Creative-100`)	LLaVA-7B	0.76 (.00*)	0.56	0.07 (.49)	0.73	0.80	0.67
	LLaVA-13B	0.71 (.00*)	**0.30**	*nan*	0.67	0.68	0.67
	InternVL2-8B	**0.82** (.00*)	0.57	0.11 (.27)	0.69	0.71	0.68
	GPT-4v	-	-	**0.15** (.13)	**0.87**	**0.93**	**0.85**
Atypicality (`Creative-100`)	LLaVA-7B	0.75 (.00*)	0.40	0.17 (.08)	0.76	0.81	0.72
	LLaVA-13B	0.63 (.00*)	**0.26**	*nan*	0.69	0.68	0.69
	InternVL2-8B	**0.78** (.00*)	0.40	**0.24** (.02*)	0.78	0.81	0.75
	GPT-4v	-	-	-0.07 (.47)	**0.89**	**0.94**	**0.80**
Atypicality (`Atypical-300`)	LLaVA-7B	0.21 (.00*)	0.32	0.01 (.92)	0.79	-	-
	LLaVA-13B	0.17 (.00*)	**0.26**	-0.05 (.43)	0.66	-	-
	InternVL2-8B	**0.23** (.00*)	0.35	**0.02** (.75)	0.77	-	-
	GPT-4v	-	-	-0.00 (.96)	**0.90**	-	-

Table 2: **Bold**: best-performing models; *: statically significant results ($\alpha = 0.05$); *nan*: disagreement predictions are uniform, making correlation test fail; "-" in GPT-4v rows: no distribution modeling task is done due to budget constraint; "-" in Pairwise columns: the classification of easy and hard is not available in `Atypical-300`.

ior when it comes to creativity ratings. In practice, we prompt VLMs multiple times with high temperatures to get the same number of VLM outputs as the number of annotators (more details in Appendix C.3). In this way, we simulate a "group behavior" instead of a single-point judgment about the level of creativity in the advertisement. To evaluate the quality of this simulation, we use two metrics: Spearman's correlation between the average rating from humans and that from VLMs, and the average KL Divergence between the human rating distribution and that from VLMs. These two results are in *Rating Correlation* and *Distribution Divergence* columns in Table 2.

Disagreement Prediction tries to capture the annotator's level of disagreement, which is important in domains like marketing. An advertisement with high creativity ratings and minimal disagreement is desired and could have a more positive impact on the product. In practice, we directly prompt VLMs to predict the level of disagreement (low, middle, or high) for each scoring dimension. We then compute Spearman's correlation between the prediction and the standard deviation of human ratings. This metric studies the level of creativity ambiguity of the advertisements. A very creative advertisement will have a low disagreement rate with a high creativity score. The results are in *Disagreement* column in Table 2.

Pairwise Preference aims at evaluating the model's ability to correctly pick the more creative advertisement out of two advertisements, given that an absolute rating of creativity can be hard when there is no reference. For each scoring dimension, we include all advertisement pairs with average human ratings differences greater than 0.5. For `Creative-100`, we have 938, 2708, and 2631 pairs in creativity, originality, and atypicality; for `Atypical-300`, we sampled 1000 image pairs from 300 images due to constraints in computation resources. The results are evaluated by F1 score and are shown in *Pairwise* column in Table 2.

5 Results

Promising Results in Distribution Modeling For all dimensions in `Creative-100`, the correlations between average human and VLM ratings are both high and statistically significant, with InternVL2-8B being the best-performing model in every dimension. However, the correlations are much lower in `Atypical-300`. We hypothesize this is due to the small annotation size (3 per advertisement) in `Atypical-300`, which can easily be biased by one annotation data point, leading to an overall more noisy distribution. Distribution Divergence also shows promising results, with the lowest divergence achieved by the LLaVA-13B model. Cross-dataset disparity is also much lower, where the KL divergence is similar for the atypicality in both datasets.

It is also worth noting that different distributional metrics result in different best-performing models, with InternVL2-8B tops Rating Correlation and LLaVA-13B wins KL Divergence. Further analysis reveals that LLaVA-13B tends to produce overall lower scores (average 2.19, out of 3) for all three dimensions compared to InternVL2-8B (average 2.31, out of 3). The MTurk human annotators produced a similarly lower score of 1.94 out of 3 on average. We note that KL Divergence focuses more on the absolute difference between two distributions of scores (LLaVA-13B scores are similarly low compared to humans), while the Rating Correlation cares more about the relative ranking of scores. This indicates that these two metrics measure different aspects of the score difference between human and model scorers. Output examples and reasoning text word cloud in Appendix D.2.

Disagreement Prediction Remains Challenging For all scenarios in Disagreement Prediction (except for Atypicality in `Creative-100`, with InternVL2-8B), disagreement predictions have no statistical correlation with human rating standard deviations. For LLaVA-13B, all outputs are "middle", making the correlation result *nan*. Both of these findings suggest that using VLM as a group-opinion synthesizer remains challenging[4]. We believe the main reason for such poor performance is the complexity involved in understanding and predicting disagreement. Understanding disagreement requires not only correctly interpreting the advertisement content, but also understanding how different humans would decode the meaning from different perspectives. Since many advertisements require complex background knowledge (e.g., some require rarely known background knowledge, which might make people disagree highly on whether an advertisement is creative), correctly understanding such necessity is challenging but critical in correctly predicting the disagreement. This process involves multi-hop (or even hierarchical) reasoning, and it remains challenging for VLMs. We hypothesize that adding the related background knowledge of the advertisements, or decomposing the multi-step reasoning needed for the disagreement task, might improve model performance, but we leave this for future work to explore.

Great Performance in Pairwise Preference Results from the pairwise preference task are very impressive, with the best-performing GPT-4v achieving more than 0.9 F1 score. We also further analyze the performance by dividing image pairs into "easy" and "hard" pairs. If the difference between average human ratings of an image pair is higher than the median difference of all image pairs, it is considered an "easy" pair. In other words, those two images have a relatively higher difference in one of the dimensions. As shown in Table 2, all VLMs perform better in easy tasks than hard ones.

Smaller Models' Superior Performance LLaVA-7B and InternVL2-8B consistently outperform LLaVA-13B in Rating Correlation and Pairwise Preference, both of which are ranking-based evaluations. Based on the error analysis (Appendix D.1), we believe this can be explained by ranking tasks requiring higher reasoning capability from the language part of the VLM. The language parts of those two smaller models are Mistral-7B and InternLM2.5-7B-Chat, both having a higher ranking on the HuggingFace Open LLM LeaderBoard[5] in reasoning tasks compared to the language part (Vicuna-13B) of LLaVA-13B.

6 Conclusion

We present a case study of using SoTA VLMs to evaluate creativity in advertisements. Inspired by marketing research, we collect fine-grained human annotations for atypicality, originality, and overall creativity, with enough annotations for every data point to capture the subjective nature of creativity evaluation. We also propose several new tasks, including distribution modeling and disagreement prediction, which specifically test VLMs' ability to deal with subjectivity. We find that SoTA VLMs achieve promising results in the pairwise comparison task while still struggling with disagreement prediction and distribution modeling. Our work opens the opportunity for the automatic assessment of advertisement creativity by providing a benchmark and metrics.

Looking ahead, although GPT-4v achieves impressive results in the pairwise comparison task, we still see smaller, open-source models like LLaVA-7B underperforming. Since smaller models are more commonly used in downstream applications, future work could focus on improving their performance.

[4]We also calculated the correlation between standard deviations of model predictions and that of human ratings, but the correlations are all near zero.

[5]HuggingFace LLM Leaderboard

7 Limitations

One obvious limitation is the size of our dataset. The fine-grained creativity annotation only consists of 100 advertisement images. Two bottlenecks that lead to such a limited number are the budget and annotation quality. Since we want to explore distribution modeling, we need more annotations than typical machine learning tasks, leading to a huge budget requirement. Also, because of the dataset size, we are not able to conduct fine-tuning experiments, as we only have 100 images with high-quality annotations. We will leave exploration of fine-tuning experiments to future work.

Another limitation of the paper is the design of simulating "group behavior" by prompting a VLM 25 times with the same prompt, we recognize the simplicity of how we prompted the VLMs to make predictions. However, creating 25 different prompts for each advertisement could complicate the analysis and results. For example, certain prompts could disproportionately increase the likelihood of a "creative" label compared to others. One interesting approach to generating prompts more meaningfully could involve exploring persona prompts to simulate multiple annotators' behavior. However, we believe that starting with a straightforward approach using the sampling strategy is essential. Therefore, we leave the use of 25 distinct persona prompts for future work.

Finally, due to hardware constraints, we only experiment with VLMs in the 7B to 13B range when much larger models, such as LLaVA-34B, are available. We will leave more extensive prompt tuning and model selections to future work.

Ethical Consideration

As a majority of our annotators are located in the U.S, there are natural biases in our annotation. We have plans to expand the annotation to other platforms (e.g., LabInTheWild Reinecke and Gajos (2015)) where a more diverse set of annotators is available. Future work could also explore alternative prompting approaches to simulate group behavior or conduct a demographic analysis of human annotations, which could check whether VLM holds opinions that are comparable to those of particular groups.

Acknowledgment

We appreciate the valuable feedback and constructive suggestions provided by the anonymous reviewers and researchers at the Pitt NLP group. This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

References

Ahmed M Abdulla Alabbasi, Sue Hyeon Paek, Daehyun Kim, and Bonnie Cramond. 2022. What do educators need to know about the torrance tests of creative thinking: A comprehensive review. *Front. Psychol.*, 13:1000385.

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *arXiv [cs.CL]*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024a. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024b. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, CC '24, page 132–155, New York, NY, USA. Association for Computing Machinery.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Honghua Chen and Nai Ding. 2023. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.*, 67(12):1–18.

Qi Cheng, Michael Boratko, Pranay Kumar Yelugam, Tim O'Gorman, Nalini Singh, Andrew McCallum, and Xiang Li. 2024. Every answer matters: Evaluating commonsense with probabilistic measures. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 493–506, Bangkok, Thailand. Association for Computational Linguistics.

Jaafar El-Murad and Douglas West. 2004a. The definition and measurement of creativity: What do we know? *Journal of Advertising Research*, 44:188–201.

Jaafar El-Murad and Douglas C. West. 2004b. The definition and measurement of creativity: What do we know? *Journal of Advertising Research*, 44(2):188–201.

Giorgio Franceschelli and Mirco Musolesi. 2024. Creativity and machine learning: A survey. *ACM Comput. Surv.*, 56(11).

Fabricio Goes, Marco Volpe, Piotr Sawicki, Marek Grzes, and Jacob Watson. 2023. Pushing GPT's creativity to its limits: Alternative uses and torrance tests. In *14th International Conference on Computational Creativity 2023*.

J P Guilford, Paul R Christensen, Philip R Merrifield, and Robert C Wilson. 2012. Alternate uses. Title of the publication associated with this dataset: PsycTESTS Dataset.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024a. Creativity in AI: Progresses and challenges. *arXiv [cs.AI]*.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024b. Evaluating creative short story generation in humans and large language models. *CoRR*, abs/2411.02316.

Alan Pickering James Lloyd-Cox and Joydeep Bhattacharya. 2022. Evaluating creativity: How idea context and rater personality affect considerations of novelty and usefulness. *Creativity Research Journal*, 34(4):373–390.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2024. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *The Thirteenth International Conference on Learning Representations*.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2025a. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *The Thirteenth International Conference on Learning Representations*.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. 2025b. Benchmarking language model creativity: A case study on code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2776–2794.

Guillermo Marco, Julio Gonzalo, M Teresa Mateo-Girona, and Ramón Del Castillo Santos. 2024. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19654–19670, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Aidan McLaughlin, Anuja Uppuluri, and James Campbell. 2024. AidanBench: Evaluating novel idea generation on open-ended questions. In *Language Gamification - NeurIPS 2024 Workshop*.

Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. 2021. Naming unrelated words predicts creativity. *Proc. Natl. Acad. Sci. U. S. A.*, 118(25):e2022340118.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Think. Skills Creat.*, 49(101356):101356.

Brian Porter and Edouard Machery. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Sci. Rep.*, 14(1):26133.

Antoine Bellemare Pépin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *CoRR*.

Katharina Reinecke and Krzysztof Z. Gajos. 2015. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 1364–1378, New York, NY, USA. Association for Computing Machinery.

Sara Rosengren, Martin Eisend, Scott Koslow, and Micael Dahlen. 2020. A meta-analysis of when and how advertising creativity works. *Journal of Marketing*, 84(6):39–56.

Sameh Said-Metwaly, Wim Van den Noortgate, and Eva Kyndt. 2017. Approaches to measuring creativity: A systematic literature review. *Creativity. Theories – Research - Applications*, 4(2):238–275.

Pooja Sharma. 2012. Advertising effectiveness:" understanding the value of creativity in advertising", a review study in india. *Online Journal of Communication and Media Technologies*, 2(3):1.

Dean Keith Simonton. 2012. Quantifying creativity: can measures span the spectrum? *Dialogues in Clinical Neuroscience*, 14(1):100–104. PMID: 22577309.

Stephen Skalicky. 2022. Liquid gold down the drain: Measuring perceptions of creativity associated with figurative language and play. *Cognitive Semantics*, 8(1):79 – 108.

Robert E. Smith, Scott B. MacKenzie, Xiaojing Yang, Laura M. Buchholz, and William K. Darley. 2007. Modeling the determinants and effects of creativity in advertising. *Marketing Science*, 26(6):819–833.

Yih Hwai Lee Swee Hoon Ang, Siew Meng Leong and Seng Lee Lou. 2014. Necessary but not sufficient: Beyond novelty in advertising creativity. *Journal of Marketing Communications*, 20(3):214–230.

Remziye Terkan. 2014. Importance of creative advertising and marketing according to university students' perspective. *International Review of Management and Marketing*, 4(3):239–246.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. 2024. MacGyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5324, Mexico City, Mexico. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. 2025. Assessing the creativity of llms in proposing novel solutions to mathematical problems. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. 2019. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1308–1323.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*.

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. Assessing and understanding creativity in large language models. *Mach. Intell. Res.*, pages 1–20.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

A Ads Dataset

A.1 Terminology

In the original study (Smith et al., 2007) that proposed the breakdown of creativity, they use the term "synthesis" instead of atypicality. However, they defined "synthesis" as "*...combine, connect, or blend normally unrelated objects or ideas*" which is very similiar to "atypicality". We use the term "atypicality" because that is the term used in the `Pitt-Ads` Dataset and want to keep the terminology aligned.

In terms of their differences, Atypicality is intrinsic to the ad itself, e.g., it contains physically impossible scenes such as people dancing under the ocean; originality is extrinsic and is the result of comparing with other existing ads in the same domain, e.g., a car ad without any image of cars on it is considered original. These two aspects are complementary: some ads are original but not atypical, e.g., a regular car and a typical scenario, but original text that involves a reference to pop culture. The advertisement in Figure 2has an average of 1.56 in atypicality and 2.0 in originality (scale 1 to 3).

A.2 Atypicality and Originality as Selected Dimensions of Creativity

The original marketing research (Smith et al., 2007) conducted an exploratory factor analysis (EFA) on five dimensions of divergence in ads creativity (numbers in parenthesis are loading factors of EFA, higher means more correlated with creativity): originality (0.23), synthesis (0.23), artistic value (0.20), flexibility (0.18), and elaboration (0.15). As mentioned previously, their definition of synthesis is very similar to "atypicality" as defined in `Pitt-Ads` Dataset. Therefore, we keep the highest-scoring two aspects (originality and synthesis/atypicality) in our study.

Figure 2: Example of original but not so atypical advertisement.

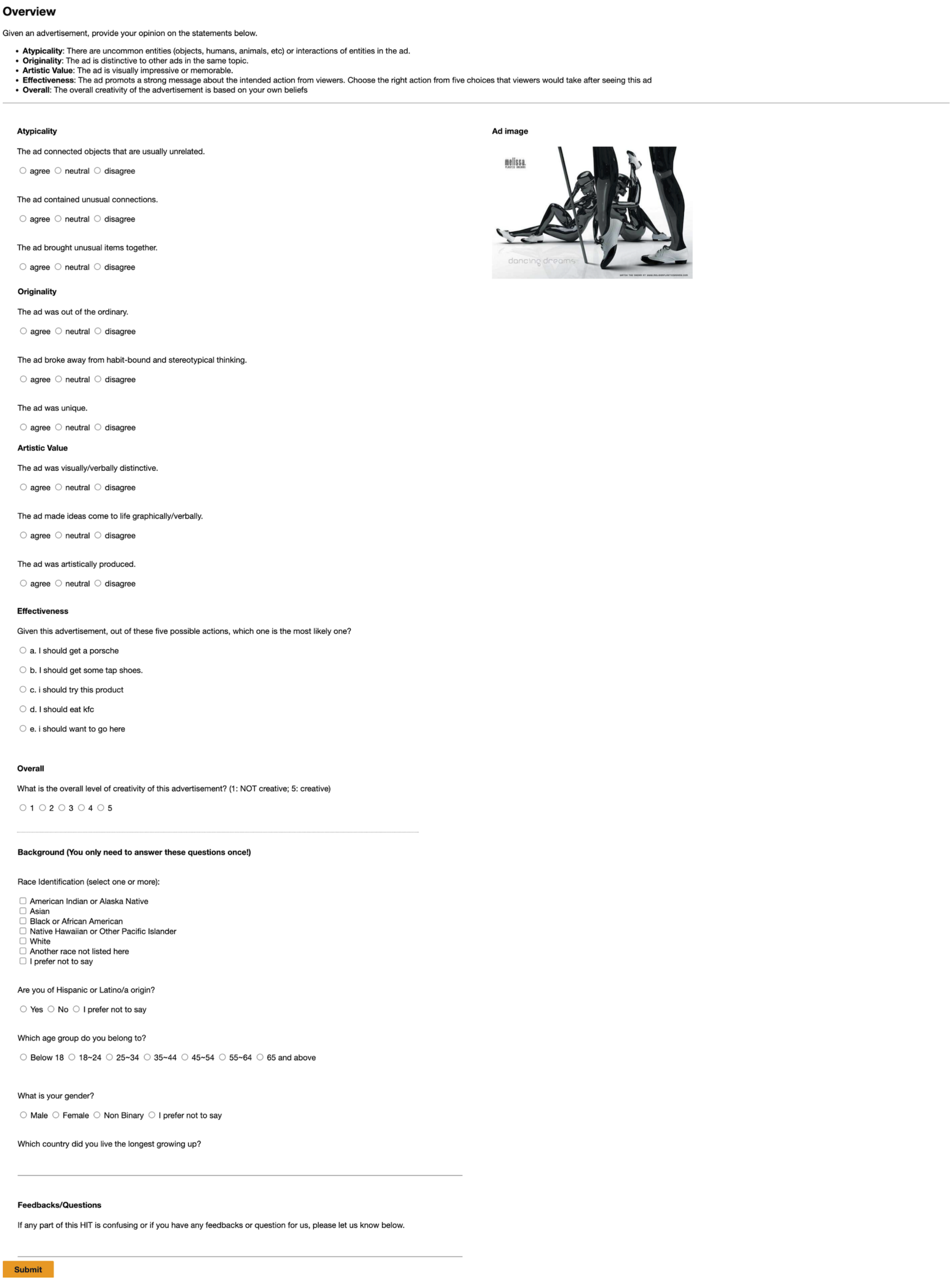
Figure 3: Example of atypical advertisement.

A.3 Sampling Process

The original `Pitt Ads` dataset consists of 38 topics in total (Hussain et al., 2017). We group those topics into 10 "topic groups": clothing, food, automobile, beauty, leisure, electronics, drinks, service, non-commercial, and healthcare. From each topic group, we sampled 10 ads, which makes up the 100 images in `Creativity-100`.

Details about "topic groups"

- food: "Restaurants, cafe, fast food", "Chocolate, cookies, candy, ice cream", "Chips, snacks, nuts, fruit, gum, cereal, yogurt, soups", "Seasoning, condiments, ketchup"
- pet: "Pet food"
- drinks: "Alcohol"", "Coffee, tea"", "Soda, juice, milk, energy drinks, water""
- automobile: "Cars, automobiles"
- electronics: "Electronics"
- service: "Phone, TV and internet service providers", "Financial services , "Security and safety services" , "Software" , "Other services"
- education: "Education"
- beauty: "Beauty products and cosmetics"
- healthcare: "Healthcare and medications"
- clothing: "Clothing and accessories"
- home: "Baby products" , "Cleaning products" , "Home improvements and repairs" , "Home appliances"]
- leisure: "Games and toys" , "Vacation and travel" , "Media and arts" , "Sports equipment and activities", "Gambling"
- shopping: "Shopping"
- non-commercial: "Environment, nature, pollution, wildlife", "Animal rights, animal abuse", "Human rights", "Safety, safe driving, fire safety", "Smoking, alcohol abuse", "Domestic violence", "Self esteem, bullying, cyber bullying", "Political candidates", "Charities"



Overview

Given an advertisement, provide your opinion on the statements below.

- **Atypicality**: There are uncommon entities (objects, humans, animals, etc) or interactions of entities in the ad.
- **Originality**: The ad is distinctive to other ads in the same topic.
- **Artistic Value**: The ad is visually impressive or memorable.
- **Effectiveness**: The ad promotes a strong message about the intended action from viewers. Choose the right action from five choices that viewers would take after seeing this ad
- **Overall**: The overall creativity of the advertisement is based on your own beliefs

Atypicality

The ad connected objects that are usually unrelated.

○ agree ○ neutral ○ disagree

The ad contained unusual connections.

○ agree ○ neutral ○ disagree

The ad brought unusual items together.

○ agree ○ neutral ○ disagree

Originality

The ad was out of the ordinary.

○ agree ○ neutral ○ disagree

The ad broke away from habit-bound and stereotypical thinking.

○ agree ○ neutral ○ disagree

The ad was unique.

○ agree ○ neutral ○ disagree

Artistic Value

The ad was visually/verbally distinctive.

○ agree ○ neutral ○ disagree

The ad made ideas come to life graphically/verbally.

○ agree ○ neutral ○ disagree

The ad was artistically produced.

○ agree ○ neutral ○ disagree

Effectiveness

Given this advertisement, out of these five possible actions, which one is the most likely one?

○ a. I should get a porsche

○ b. I should get some tap shoes.

○ c. i should try this product

○ d. I should eat kfc

○ e. i should want to go here

Overall

What is the overall level of creativity of this advertisement? (1: NOT creative; 5: creative)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Background (You only need to answer these questions once!)

Race Identification (select one or more):

☐ American Indian or Alaska Native
☐ Asian
☐ Black or African American
☐ Native Hawaiian or Other Pacific Islander
☐ White
☐ Another race not listed here
☐ I prefer not to say

Are you of Hispanic or Latino/a origin?

○ Yes ○ No ○ I prefer not to say

Which age group do you belong to?

○ Below 18 ○ 18~24 ○ 25~34 ○ 35~44 ○ 45~54 ○ 55~64 ○ 65 and above

What is your gender?

○ Male ○ Female ○ Non Binary ○ I prefer not to say

Which country did you live the longest growing up?

Feedbacks/Questions

If any part of this HIT is confusing or if you have any feedbacks or question for us, please let us know below.

Submit

Ad image

Figure 4: Amazon Mechanical Turk interface.

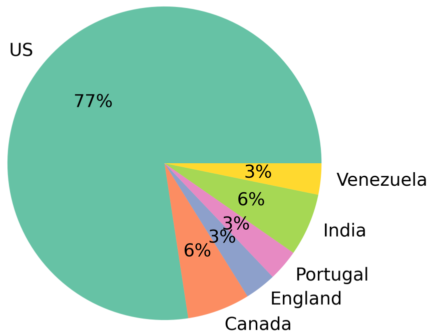


Figure 5: Distribution of workers' response to "In which country did you live the longest time so far?"

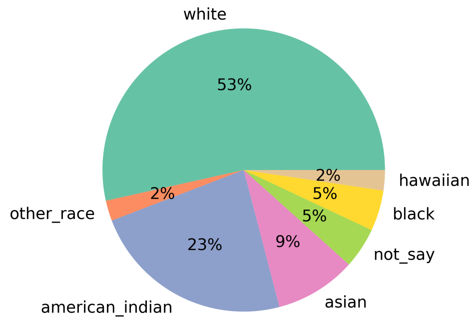


Figure 7: Distribution of workers' response to "Race Identification (select one or more)"

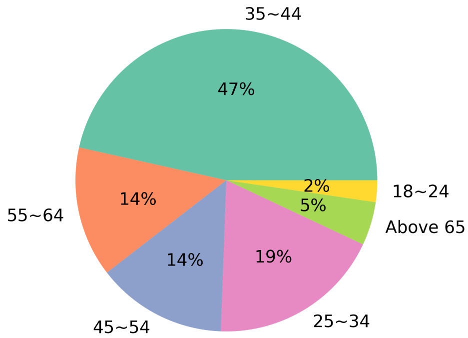


Figure 6: Distribution of workers' response to "What is your age?"

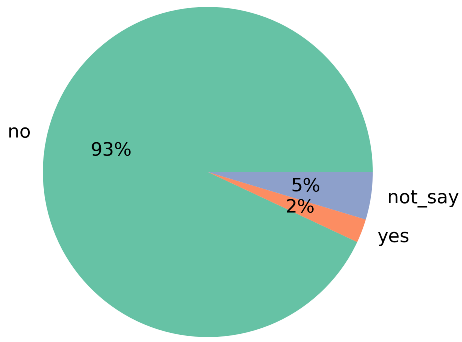


Figure 8: Distribution of workers' response to "Are you of Hispanic or Latino/a origin?"

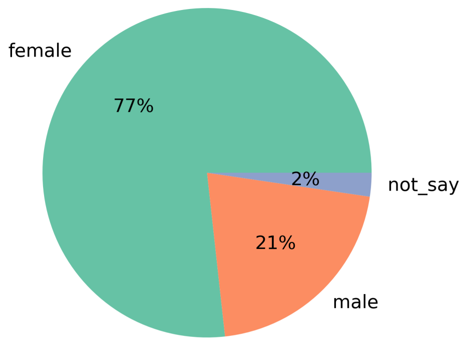


Figure 9: Distribution of workers' response to "What is your gender?"

B Data Collection and Post Processing

B.1 Amazon Mechanical Turk Details

Payment for worker Each HIT receives $0.5 compensation (estimated $15/hour).

Annotation interface See Figure 4 for the annotation interface. Note that there is a section "artistic values". We dropped that section in the later parts of the experiment because 1) it is very subjective and could be further broken down into more fine-grained subcategories, and 2) to keep our focus on atypicality and originality.

In total, 43 workers contributed to our task and finished 2500 HITs. Their background can be found in Figure 5, 6, 7, 8, and 9. As we can see, the annotators are strongly skewed towards the US-based, white, female, middle age group, which should be kept in mind when applying our methodology when it comes to people from another background.

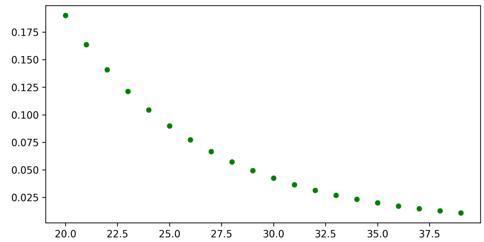


Figure 10: Upper-bound of the error based on calculation.

B.2 Number of Samples for Distribution Task

Following previous works (McHugh, 2012; Cheng et al., 2024), the number of samples required to approximate the real distribution can be calculated as follows:

$$P(D_{KL}(g_{n,k}||f) > \epsilon) \le e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sum_{i=0}^{k-2} k_{i-1}(\frac{e\sqrt{n}}{2\pi})^i\right]$$

c_1 and c_2 are constant values (based on (McHugh, 2012) $c_1 = 2, c_2 = \frac{\pi}{2}$), k is the number of categories in the categorical distribution (in our case, $k = 3$), and n is the number of samples. If we fix the left-hand side to be less than 0.1, we would get n has to be 25 (see Figure 10).

B.3 Label Processing

We process the annotation by first converting the categorical data to numerical values. For atypicality and originality, we code agree, neutral, and disagreement choices as 1, 0, and -1. As there are three subquestions for both atypicality and originality, we simply add up the three scores from each dimension and get one accumulated score for each. For overall creativity, we keep the raw score (an integer number between 1 and 5). Thus each annotation data point consists of three integer scores, corresponding to atypicality, originality, and overall creativity.

We then normalize the score by individual annotators to mitigate the differences in people's rating preferences. In particular, for each score dimension, we group the scores provided by each annotator and standardize them (subtract mean and divide by standard deviation). We then map the standardized score to an integer (1, 2, or 3) by dividing the standardized score interval into three bins.

B.4 Connection between atypicality and creativity

After analyzing the fine-grained creativity data we collected (Sec. 3.1), we find out that the Pearson R correlation between the normalized atypicality and overall creativity score is 0.4017 ($p < 0.01$), a positive correlation (the sample size is 2500: 100 ads with 25 annotations each). Therefore, it makes sense to evaluate the same methodology on data with only atypicality annotation to prove its effectiveness at a larger scale.

C Experiment Details

C.1 Configurations

- Temperature: 0.75 (for distribution prediction) and 0.01 (for disagreement prediction)
- Max New Token: 256
- Model Checkpoint
  - GPT-4: `gpt-4-vision-preview`
  - LLaVa-7B: `llava-v1.6-mistral-7b-hf`
  - LLaVa-13B: `llava-v1.6-vicuna-13b-hf`
  - InternVL2-8B: `InternVL2-8B-MPO`
- Number of pairwise samples (% of label "1")
  - creativity: 938 (46%)
  - atypicality: 2631 (51%)
  - originality: 2708 (50%)
- Device: single card NVIDIA L40S GPU

C.2 Running Time

(Roughly, all tasks combined)

- `Creative-100`
  - GPT4-v: 1 hr
  - LLaVA-7B: 4 hr
  - LLaVA-13B: 5 hr
  - InternVL: 4.5 hr
- `Atypical-300` (atypical data only)
  - GPT4-v: 0.2 hr
  - LLaVA7B: 0.5 hr
  - LLaVA13B: 0.5 hr
  - InternVL: 0.5 hr

Dimension	Model	Distribution Modeling R (*p-value*)	KL-Divergence
Creativity	LLaVA-7B	0.6363 (<.01)	1.0721
	LLaVA-7B	0.6548 (<.01)	1.0127
	LLaVA-7B	0.6392 (<.01)	1.0464
	LLaVA-13B	0.6548 (<.01)	0.3734
	LLaVA-13B	0.6289 (<.01)	0.3933
	LLaVA-13B	0.6257 (<.01)	0.3706
	InternVL2-8B	0.6363 (<.01)	1.4903
	InternVL2-8B	0.6548 (<.01)	1.4542
	InternVL2-8B	0.7153 (<.01)	1.4658
Originality	LLaVA-7B	0.7573 (<.01)	0.5649
	LLaVA-7B	0.7623 (<.01)	0.5946
	LLaVA-7B	0.7781 (<.01)	0.5942
	LLaVA-13B	0.7108 (<.01)	0.2983
	LLaVA-13B	0.6757 (<.01)	0.3297
	LLaVA-13B	0.7262 (<.01)	0.3448
	InternVL2-8B	0.8169 (<.01)	0.5707
	InternVL2-8B	0.8044 (<.01)	0.6203
	InternVL2-8B	0.8124 (<.01)	0.5935
Atypicality	LLaVA-7B	0.7282 (<.01)	0.4226
	LLaVA-7B	0.7470 (<.01)	0.4043
	LLaVA-7B	0.7061 (<.01)	0.4301
	LLaVA-13B	0.6332 (<.01)	0.2617
	LLaVA-13B	0.5732 (<.01)	0.2692
	LLaVA-13B	0.7126 (<.01)	0.2367
	InternVL2-8B	0.7838 (<.01)	0.4041
	InternVL2-8B	0.7821 (<.01)	0.4083
	InternVL2-8B	0.7838 (<.01)	0.4041

Table 3: Results for multiple runs

C.3 Multiple Run Results

For the distribution modeling task, since there is inherent randomness in the modeling process (we set the temperature for LLM generation to 0.75), we run the same setting for 3 times to see if the randomness will have a huge impact. As shown in Table 3, for the same task, same model, the performance is still within reasonable margin. We believe it is safe to believe that the results is not affected by randomness in LLM.

C.4 VLM Prompts - Creativity

Distribution Modeling *How creative is this advertisement? Give your answer that follows these guidelines:*
- Choose your answer from [not creative, neutral, very creative], answer 1 for not creative, 2 for neutral, and 3 for very creative
- Show your reasons and give a final answer (in a single number), in the following format: "reasoning:{reasoning}, answer: {score}"

Disagreement *You are a marketing analyst tasked with predicting the audience's perception of an advertisement's creativity. In next week, 25 individuals will rate this ad's creativity on a scale of 1 to 3. Your goal is to predict the level of disagreement among these ratings. As a reference, ads that are more challenging to understand generally result in higher disagreement, while simpler ads lead to greater consensus.*
Provide your assessment using a single score:
1: Low disagreement (high agreement),
2: Neutral disagreement,
3: High disagreement (low agreement).
Your response should follow this format:
"answer: {score}; explanation: {reasoning}"

Pairwise *Here are two images of advertisement. Which one is more creative? Answer 1 for the one on the left and 2 for the one on the right. Give your answer in the following format: "explanation: {reasoning}; answer: {choice}"*

C.5 VLM Prompts - Atypicality

Distribution Modeling *How atypical is this advertisement? Give your answer that follows these guidelines:*
- Atypical advertisement either connected objects that are usually unrelated, contained unusual connections, or brought unusual items together;
- Choose your answer from [not atypical, neutral, very atypical], answer 1 for not atypical, 2 for neutral, and 3 for very atypical
- Show your reasons and give a final answer (in a single number), in the following format: "reasoning: {reasoning}, answer: {score}; "

Disagreement *You are a marketing expert task with predicting how audiences perceive the atypicality of an advertisement. Atypicality refers to ads that connect objects typically unrelated, feature unusual connections, or bring unexpected items together. In next week, 25 individuals will rate this ad's atypicality on a scale of 1 to 3. Your task is to predict the level of disagreement among their ratings. As a reference, Ads that are more difficult to interpret tend to generate higher disagreement, whereas more straightforward ads lead to greater agreement.*
Provide your answer using a single score:
1: Low disagreement (high agreement),
2: Neutral disagreement,
3: High disagreement (low agreement).

Format your response as follows:
"answer: {score}; explanation: {reasoning}"

Pairwise *Here are two images of advertisement. Which one is more abnormal and unusual? Answer 1 for the one on the left and 2 for the one on the right. Give your answer in the following format: "explanation: {reasoning}; answer: {choice}"*

C.6 VLM Prompts - Originality

Distribution Modeling *How novel is this advertisement? Give your answer that follows these guidelines:*
- Novel advertisement either is out of the ordinary, break away from habit-bound and stereotypical thinking, or is unique;
- Choose your answer from [not novel, neutral, very novel], answer 1 for not novel, 2 for neutral, and 3 for very novel
- Show your reasons and give a final answer (in a single number), in the following format: "reasoning: {reasoning}, answer: {score};"

Disagreement *You are a marketing analyst tasked with predicting how audiences perceive the novelty of an advertisement. Novelty refers to ads that are out of the ordinary, break free from stereotypical thinking, or exhibit uniqueness. In next week, 25 individuals will rate this ad's novelty on a scale of 1 to 3. Your goal is to predict the level of disagreement among their ratings. Ads that are harder to interpret typically result in higher disagreement, while clearer ads lead to more agreement.*
Provide your assessment using a single score:
1: Low disagreement (high agreement),
2: Neutral disagreement,
3: High disagreement (low agreement).
Format your response as follows:
"answer: {score}; explanation: {reasoning}"

Pairwise *Here are two images of advertisement. Which one is more unique compared with other ads in the same product category? Answer 1 for the one on the left and 2 for the one on the right. Give your answer in the following format: : "explanation: {reasoning}; answer: {choice}"*

C.7 `Atypical-300` Prompts (atypicality only)

Distribution Modeling *How atypical is this advertisement? Give your answer that follows these guidelines:*
- Atypical advertisement either connected objects that are usually unrelated, contained unusual connections, or brought unusual items together;
- Choose your answer from [not atypical, neutral, very atypical], answer 0 for not atypical and 1 for very atypical
- Show your reasons and give a final answer (in a single number), in the following format: "reasoning: {reasoning}, answer: {score};"

Disagreement *You are a marketing expert analyzing how audiences perceive the atypicality of an advertisement. Atypicality refers to ads that connect objects typically unrelated, feature unusual connections, or bring unexpected items together. Imagine 25 individuals have rated the ad's atypicality either 0 or 1. Your task is to determine the level of disagreement among their ratings. Ads that are more difficult to interpret tend to generate higher disagreement, whereas more straightforward ads lead to greater agreement.*
Provide your answer using a single score:
0: Low disagreement (high agreement),
1: High disagreement (low agreement).
Format your response as follows:
"answer: {score}; explanation: {reasoning}"

Pairwise *Here are two images of advertisement. Which one is more abnormal and unusual? 1 for the left image and 2 for the right image. Give your answer in the following format: "answer: {number}; explanation: {reasoning}"*

D Output Analysis

D.1 Error Analysis on Pairwise Outputs

Here, we present two error analysis examples on the pairwise task. Detailed analysis are in the caption of each image pair (Figure 11, 12).

D.2 Distribution Modeling Examples

We have three examples with all the scoring metrics; see Figure 13, 14, 15. We have also plotted WordClouds for the *reasoning* part of the output (Figure 18, 21, 24). As shown in those word clouds, the commonly used phrases generally closely correspond to the task definition: 'creative', 'message', 'imagery', 'effective' for creativity task, 'unique', 'novel', 'habit' for originality task, and 'atypical', 'connection', ' unusual' for atypicality task).

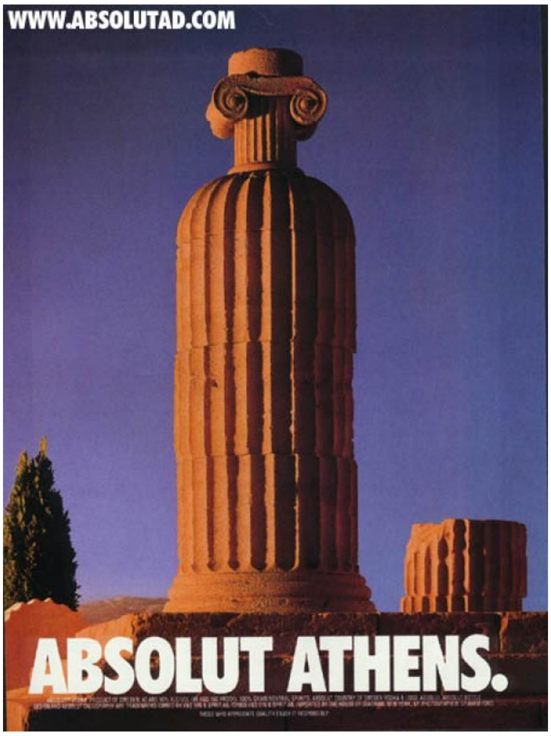


Figure 11: Task: **atypicality**; model outputs are in Table 4

Model	Model Output	Prediction
LLaVA-7B	explanation: The image on the left features a man sitting on a chair with a cup of soda in front of him. The text on the ad is about the risks of consuming sugary drinks and the importance of cutting one's portion sizes. This is a typical advertisement that uses a real-life scenario to convey a message. On the other hand, the image on the right is more unusual and abnormal. It features a large, brown, cylindrical structure that resembles a bottle or a vase. The text on this ad is about the risks of consuming sugary drinks and the importance of cutting one's portion sizes. The presence of the large structure in the ad is unexpected and out of place, making it more unusual and abnormal compared to the first image. answer: 2	right (2)
LLaVA-13B	The image on the left is more abnormal and unusual because it features a person with a significantly larger waistline than the rest of their body, which is not a typical human proportion. **The image on the right shows a large, ancient-looking structure, which is not an unusual sight in itself, but it is presented in a way that suggests it is a part of the advertisement, which is not a common practice for advertising.**	left (1)
InternVL2-8B	explanation: The left advertisement features a person with a large belly, which is a common depiction in media to represent obesity. The use of a person in the image is typical in health campaigns to visually convey the message about portion sizes and health risks. The right advertisement, on the other hand, uses the Parthenon in Athens, an ancient monument, to convey the message about portion sizes. Using a historical or cultural landmark in an advertisement is less common and more unusual compared to using a person. Therefore, the right advertisement is more abnormal and unusual. answer: 2 ***{repeating answer 4 times}***	right (2)
GPT-4v	explanation: The advertisement on the right is more abnormal and unusual because it portrays a well-known architectural column, which is typically straight, as a bottle of Absolut Vodka. This is a surreal and artistic reinterpretation of a cultural landmark, intended to resemble the shape of the iconic Absolut bottle, which is not a normal sight in real-world settings. On the other hand, the advertisement on the left, while addressing a serious health issue regarding portion sizes and diabetes, doesn't alter common objects in an unexpected way but uses real-life elements (a person and soft drink cups) to convey its message. Therefore, the right image is more surreal and unconventional in its approach. answer: 2	right (2)

Table 4: Model output for ads pair in Figure 11. The correct answer is right (2). We can see LLaVA-13B is the only model that gets the answer incorrectly and its explanation is also self-contradicting (the bold part); for InternVL2, the problem is that it always repeat the answer for multiple times, even when it answers correctly.

Figure 12: Task: **originality**; model outputs are in Table 5

Model	Model Output	Prediction
LLaVA-7B	explanation: The left image features a woman with a caption that reads “When I sneeze, I wet my pants. Normal?” This ad is unique because it addresses a common but often embarrassing issue, **using humor to connect with the audience**. The right image, on the other hand, shows a group of women in a bathing suit, with a caption that reads: **Seh die Schönheit! Seh die Schönheit! Seh die Schönheit! Seh die Schönheit! Seh die Schönheit! Seh die Schönheit! Seh die Schönheit!** ***{repeat lot of times}***	left (1)
LLaVA-13B	The left image is more unique compared to the right one. The reason is that the left image features a modern woman with a contemporary hairstyle and clothing, while the right image is an old-fashioned advertisement with a vintage design and illustration style. **The juxtaposition of the modern woman with the vintage advertisement creates a contrast that makes the left image more unique.**	left (1)
InternVL2-8B	explanation: The left advertisement is more unique compared to the right one because it uses a modern approach with a relatable and humorous question about a common pregnancy symptom, which can engage and resonate with potential customers. The use of a smiling woman and a direct question makes it more approachable and relevant to current audiences. In contrast, the right advertisement uses a vintage style and focuses on a specific product, which might not immediately connect with modern viewers as well. **The humor and relatability in the left ad make it stand out in the same product category**. Answer: 1 ***{repeating answer 4 times}***	left (1)
GPT-4v	explanation: The advertisement on the left provides a candid, contemporary approach to discussing health issues, specifically addressing pregnancy-related concerns in a way that openly discusses a commonly experienced, yet often unspoken issue. This upfront and direct style can be considered unique in comparison to typical pregnancy services ads which may traditionally use softer language and imagery. On the right, the **Nivea advertisement** takes a more classic approach, using vintage-style artwork to represent the product in a historical context, **which is creative but not necessarily unique as it aligns with a familiar tradition of period-specific advertising.** This style, though distinctive, may not stand out as much in the skincare product category where vintage and nostalgic themes are frequently employed. answer: 1	left (1)

Table 5: Model output for ads pair in Figure 12. The correct answer is right (2). ALLaVA-13B is the only one that incorrectly combined the visual element in the left image (“modern woman”) and the right image (“vintage advertisement”) together as if they are in the same ad (“the juxtaposition of...”); LLaVA-7B and InternVL2-8B are able to capture the humor in the left ad but still make the mistake. In particular, LLaVA-7B is stuck in consistently generating “Seh die Schönheit!”, which is a description of the left image, without even get a chance to get to the second image. This might be the reason for its failure. Note that GPT-4v is the only model that recognizes the brand name (“Nivea”), highlighting its strong capability in recognizing details.

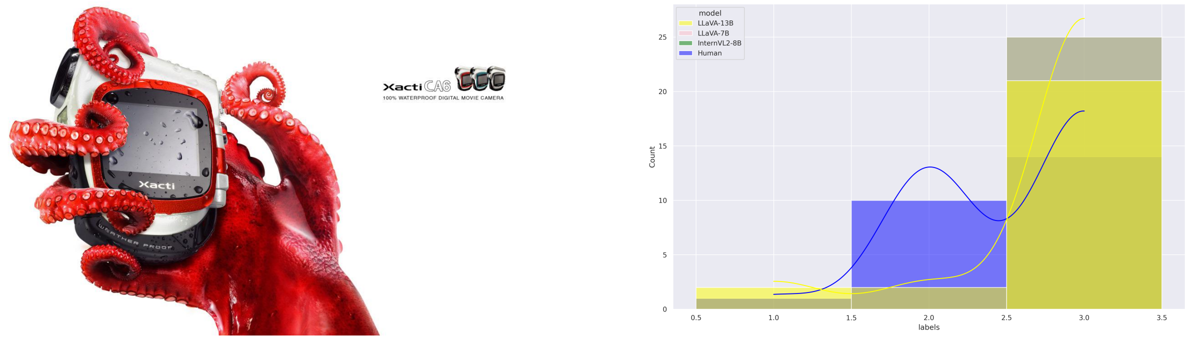


Figure 13: Example (A) and `creativity` predictions by models; complete output in Table 6

Aspect	Human	LLaVA-7B *(KL)*	LLaVA-13B *(KL)*	InternVL *(KL)*
Creativity	2.60	3.00 (1.0054)	2.76 (0.3986)	3.00 (1.0054)
Originality	2.92	2.92 (0.9643)	2.88 (0.3144)	3.00 (1.0054)
Atypicality	2.92	2.88 (0.6030)	2.64 (0.1191)	2.84 (0.2223)

Table 6: Model output (average across 25 runs) and human ratings for Example (A), see ad image and distribution modeling result in Figure 13; *KL* refers to $KL(Human||Model)$

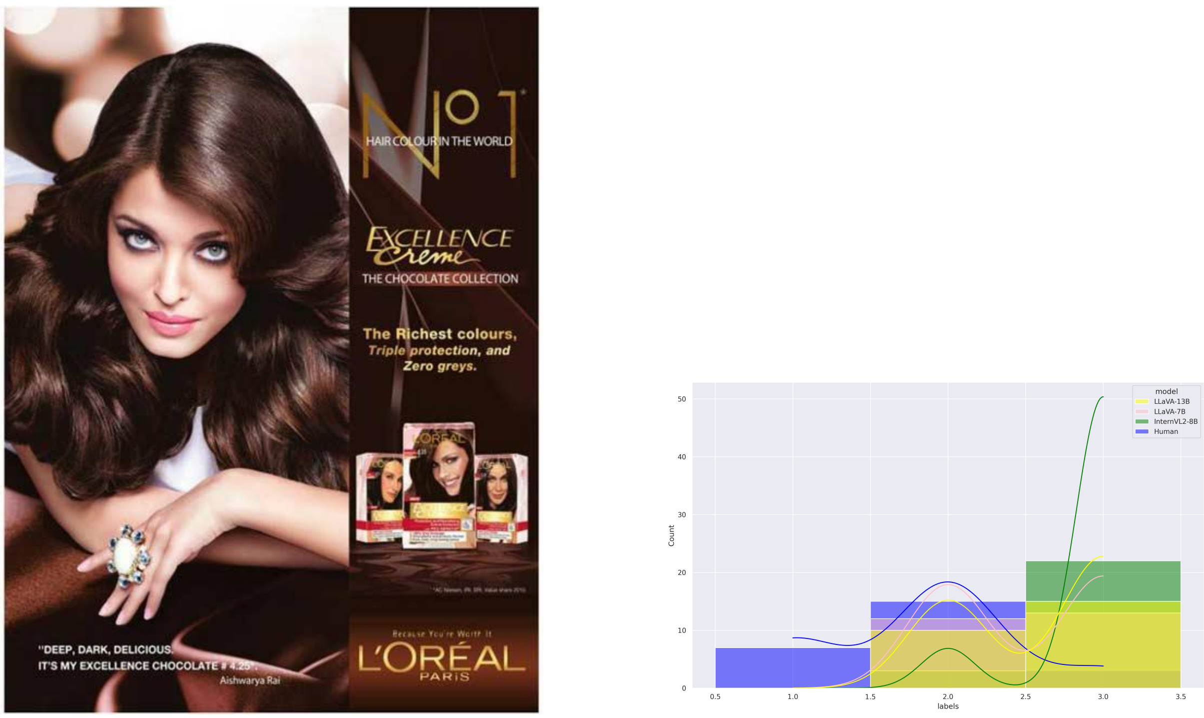


Figure 14: Example (B) and `creativity` predictions by models; complete output in Table 7

Aspect	Human	LLaVA-7B *(KL)*	LLaVA-13B *(KL)*	InternVL *(KL)*
Creativity	2.60	2.52 (0.7701)	2.60 (0.8803)	2.88 (1.6395)
Originality	2.92	2.28 (0.2762)	1.60 (0.0963)	2.12 (0.1791)
Atypicality	2.92	1.76 (0.2560)	1.68 (0.070)	1.36 (0.5113)

Table 7: Model output and human ratings for Example (B), see ad image and distribution modeling result in Figure 14; *KL* refers to $KL(Human||Model)$

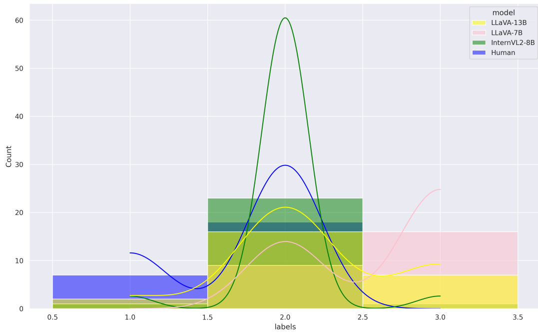


Figure 15: Example (C) and `creativity` predictions by models; complete output in Table 8

Aspect	Human	LLaVA-7B *(KL)*	LLaVA-13B *(KL)*	InternVL *(KL)*
Creativity	2.60	2.64 (1.3395)	2.20 (0.4060)	2.00 (0.3070)
Originality	2.92	2.36 (0.7893)	1.56 (0.0643)	1.60 (0.3263)
Atypicality	2.92	1.80 (0.6703)	1.60 (0.5048)	1.16 (0.8899)

Table 8: Model output and human ratings for Example (C), see ad image and distribution modeling result in Figure 15; *KL* refers to $KL(Human||Model)$

Figure 16: Task: Creativity, Model: LLaVA-7B

Figure 17: Task: Creativity, Model: LLaVA-13B

Figure 18: Task: Creativity, Model: InternVL2-8B

Figure 19: Task: Originality, Model: LLaVA-7B

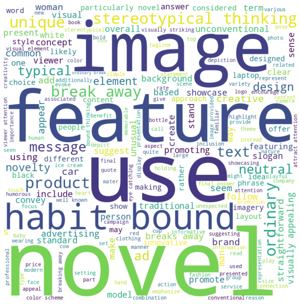


Figure 20: Task: Originality, Model: LLaVA-13B

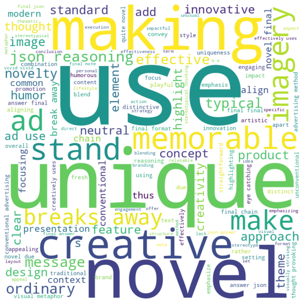


Figure 21: Task: Originality, Model: InternVL2-8B

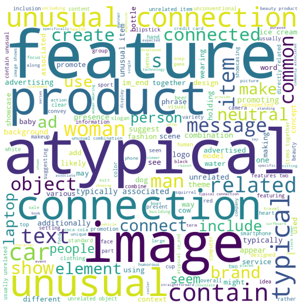


Figure 22: Task: Atypicality, Model: LLaVA-7B

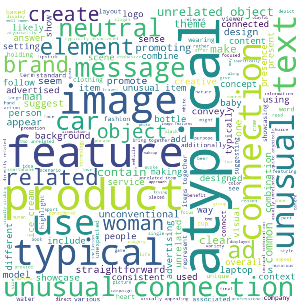


Figure 23: Task: Atypicality, Model: LLaVA-13B

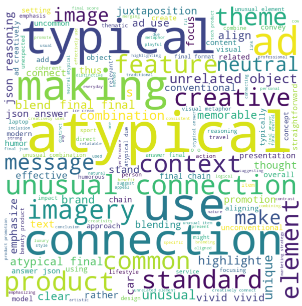


Figure 24: Task: Atypicality, Model: InternVL2-8B