

Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity

Anonymous ACL submission

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 suit of tasks specifically for such a subjective problem. We also evaluate the alignment between state-of-the-art (SoTA) vision language models (VLM) and humans on our proposed benchmark, demonstrating both the promises and challenges of using VLMs for automatic creativity assessment.¹

1 Introduction

Evaluation of creativity has been a lasting effort where researchers have taken diverse sets of perspectives (Said-Metwaly et al., 2017). Among those perspectives, evaluating the created product is the only one that focuses purely on the created artifact, no matter who authored that artifact, greatly reducing the complexity of problem framing. Examples from this perspective include evaluation of creative writing (Lu et al., 2024; Chakrabarty et al., 2024a), creative tool use (Tian et al., 2024), and creative advertising (Smith et al., 2007).

In advertising, creativity plays a critical role in driving consumer behaviors, where multiple studies have shown positive effects of creative ads, including increased purchase intent and positive brand impression (Sharma, 2012; Terkan, 2014). Therefore, ad creators are motivated to consistently develop and evaluate creative ad content. Extensive research has been conducted to understand what the general public would consider creative (El-Murad and West, 2004; Rosengren et al., 2020; Swee Hoon Ang and Lou, 2014; Smith et al., 2007), many of which consider ads creativity to be a combination of “atypicality” (or, synthesis, abnormalness,

¹Data and code will be released upon paper publication.

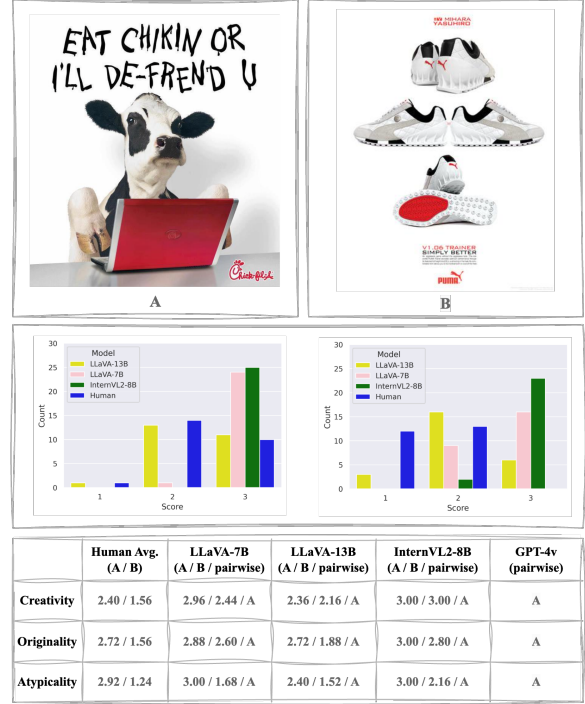


Figure 1: Top: two ads from dataset; Middle: human rating and VLM outputs (25 each); Bottom: average human rating, VLM outputs, and VLM pairwise predictions; Scores are 3-scale, 3 being the best. Ad A receives a higher average rating in all three categories compared to Ad B, also reflected in the model-predicted ratings and pairwise preference prediction.

etc.) and “originality” (or, novelty, uniqueness, etc.). To be specific: in the advertisement (A) in Fig. 1, the image of a cow typing on a computer is highly atypical, as cows do not do that; the text “Eat chikin or I’ll de-friend U”, which consists of multiple slangs, the cow, the computer, and the small logo of Chick-fil-A is a very unique combination given that it is a fast-food advertisement. Furthermore, decoding ads creativity under such a framework requires extensive reasoning, making the evaluation of creativity a challenging task. Unsurprisingly, previous work heavily relies on domain experts, who are expensive and inaccessible.

Recently, foundational models demonstrate im-

pressive performances in other evaluation tasks, such as summarization (Zhong et al., 2022), Long-Form QA (Jiang et al., 2023), and commonsense text generation (Xu et al., 2023), many of which were previously dominant by human evaluation. For creativity evaluation, prior work (Chakrabarty et al., 2024a) has also shown the capability of large language models (LLMs). 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 Vision Language Models (VLM).

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 Fig. 1 as “human rating”. We experiment with state-of-the-art (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 as the objective, instead of majority label prediction, in future research on 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, 2004; 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). There are two common grounds: first, creativity is the balance between divergence and effectiveness; we also include atypicality as part of our evaluation and ensure effectiveness through the quality check question in human annotation. Second,

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

evaluation of creativity is subjective, making fine-grained human feedback critical. This motivates our distribution modeling task instead of a traditional, majority-label prediction task. Two most related works are: (Smith et al., 2007) focused on advertisement images and proposed five creativity dimensions, including atypicality and originality. We adapt their creativity decomposition. (Chakrabarty et al., 2024a) use LLMs to evaluate short stories; in contrast, we analyzed the alignment between VLM outputs and human ratings.

Automatic Evaluation with Foundation Models GPTScore (Fu et al., 2023) and UniEval (Zhong et al., 2022) decompose the evaluation of a complex task into simpler ones that can be accomplished by language models; whereas PandaLM (Wang et al., 2024) focuses on pairwise evaluation for free-form text quality. In the vision domain, (Jayasumana et al., 2024; Otani et al., 2023) explore evaluating generated image content using CLIP embeddings. These prior works focus on single modality, instead of the image-text pair as we do.

3 Dataset

We use the Pitt Ads Dataset (referred to as Pitt-Ads) as our basis, where each ad image is annotated with its topic, expected actions from viewers after seeing the ad, binary labels of atypical objects in it (when applicable), and the topic of the ad (10 topic groups in total) (Hussain et al., 2017; Ye et al., 2019). We sample 100 ads and collect fine-grained human creativity annotations (Creative-100) (Sec. 3.1); we also sample 300 ads from the remaining ads for atypicality prediction on a larger scale (Atypical-300) (Sec. 3.2).

3.1 Creative-100

Creative-100 consists of 100 ads, with 10 from each topic group: food, pet, drinks, automobile, electronics, service, education, beauty, healthcare, clothing, home, leisure, shopping, and non-commercial. To do quality creativity evaluation, we

break down creativity into two dimensions: originality and atypicality, the two most influential dimensions for ads creativity according to (Smith et al., 2007). Human annotations are then collected via Amazon Mechanical Turk (Mturk) to collect fine-grained ratings in all three dimensions: **originality**, **atypicality**, and **creativity**. Due to the inherent subjectivity of the creativity judgment, we formulate the measurement of creativity as several multiple-choice questions with possible answers as a categorical distribution of those choices. In other words, the predictive target is not a single label (e.g., “creative”) but a distribution of human ratings. This motivates us to collect 25 annotations per ad image to cover the true rating distribution with at most 0.1 error rate (McHugh, 2012).

For atypicality and originality, we follow Smith et al. (2007) and record responses about various statements (Tab. 1). For creativity, we record an overall score from 1 to 5 and convert it to a 3-scale, aligning with other dimensions. We also include a quality check question by asking annotators to choose the action after seeing a given ad (e.g., “I should go to Chick-fil-A” for Ad A in Fig. 1). Five actions are given, with one correct action and four randomly sampled from Pitt-Ads. Annotators get 96.88% accuracy in this question, highlighting their understanding of visual advertisements. More dataset construction details are in Appx. B.

3.2 Atypical-300

We also randomly sampled 300 ads (Atypical-300) from Pitt-Ads, where 185(62%) include atypical object(s). Different from Creative-100, each ad in this set only has three binary annotations on atypicality. Both Smith et al. (2007) and human-annotated data (Sec. 3.1) show that atypicality has a positive correlation with creativity (Appx. B.4). Thus, we include this dataset to gain further insight into VLM’s ability to evaluate ad creativity.

4 Experimental Setup

4.1 Models

We experiment with open-sourced VLMs, i.e. 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 vllm inference framework (Kwon et al., 2023). More details about the experiment configurations are in Appx. 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 behavior 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. In this way, we simulate a “group behavior” instead of a single-point judgment about the level of creativity in the ad. 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 human rating distribution and that from VLMs. These two results are in *Rating Correlation* and *Distribution Divergence* columns in Tab. 2.

Disagreement Prediction tries to evaluate the model’s ability to predict people’s disagreement about the content, i.e., the ad image. 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 standard deviation of human ratings. This metric studies the ambiguity level of the ads. In reality, a very creative ad will have a low disagreement rate with a high creativity score. The results are in *Disagreement* column in Tab. 2.

Pairwise Preference aims at evaluating the model’s ability to correctly pick the more creative ad out of two ads, given that an absolute rating of creativity can be hard when there is no reference. For each scoring dimension, we include all ad 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 Tab. 2.

5 Results²

Promising Results in Rating Correlation 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, although they are statistically

²More output analyses are in Appx. D

Dimension	Model	Distribution Modeling		Disagreement	Pairwise		
		Rating Correlation ↑ R (p -value)	Distribution Divergence ↓ KL Divergence	↑ R (p -value)	All ↑ $F1$	Easy ↑ $F1$	Hard ↑ $F1$
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	<i>nan</i>	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	<i>nan</i>	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	<i>nan</i>	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.

significant. We believe this is due to the small annotation size (3 per ad) in Atypical-300, which can easily be biased by one annotation data point, leading to an overall more noisy distribution of ratings. 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. Output examples and reasoning text word cloud in Appx. 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*. This suggests that using VLM as a group-opinion synthesizer remains challenging³.

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” subsets: compared to median human rating differences if a pair of ad images has a higher absolute difference, it is an easy image pair, and

vice versa. As shown in Tab. 2, all VLMs perform better in easy tasks than hard ones, confirming the alignment with human preference differences.

Smaller Models’ Superior Performance Counterintuitively, 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 (more details in Appx. 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 LeaderBoard⁴ 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. We show decent alignment between VLMs and humans in Distribution Modeling and Pairwise Preference, whereas Disagreement Prediction remains challenging. Our work opens the opportunity for automatic evaluation of advertisement creativity by providing a benchmark and metrics.

³We also calculated the correlation between standard deviations of model predictions and that of human ratings, but the correlations are all near zero.

⁴HuggingFace LLM Leaderboard

7 Limitations

One obvious limitation is the size of our dataset. The fine-grained creativity annotation only consists of 100 ad images. Two bottlenecks that lead to such a limited number is budget and annotation quality. Since we want to explore distribution modeling, we need more annotation than typical machine learning tasks, leading to a huge budget requirement.

Another limitation is the natural biases contained in our annotation as a majority of our annotators are located in the U.S. We have plans to expand the annotation to other platforms (e.g., LabInTheWild) where a more diverse set of annotators is available. We would also suggest that researchers be cautious when applying our method to data from other countries or languages. 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 comparable to those of particular groups.

Also, 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.

References

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 emerging writers](#). *Preprint*, arXiv:2309.12570.

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.

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](#). *Preprint*, arXiv:2404.16821.

Qi Cheng, Michael Boratko, Pranay Kumar Yelugam, Tim O’Gorman, Nalini Singh, Andrew McCallum, and Xiang Lorraine Li. 2024. Every answer matters: Evaluating commonsense with probabilistic measures. *arXiv preprint arXiv:2406.04145*.

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

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. [Gptscore: Evaluate as you desire](#). *Preprint*, arXiv:2302.04166.

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.

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.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. 2024. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhui Chen. 2023. [Tigerscore: Towards building explainable metric for all text generation tasks](#). *ArXiv*, abs/2310.00752.

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](#)

403	systematic attribution of machine text against web	
404	text . <i>Preprint</i> , arXiv:2410.04265.	
405	Mary L. McHugh. 2012. Interrater reliability: the kappa	
406	statistic. <i>Biochemia medica</i> , 22(3):276–282.	
407	OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal,	
408	Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-	
409	man, Diogo Almeida, Janko Altschmidt, Sam Alt-	
410	man, Shyamal Anadkat, Red Avila, Igor Babuschkin,	
411	Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim-	
412	ing Bao, Mohammad Bavarian, Jeff Belgum, Ir-	
413	wan Bello, Jake Berdine, Gabriel Bernadett-Shapiro,	
414	Christopher Berner, Lenny Bogdonoff, Oleg Boiko,	
415	Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-	
416	man, Tim Brooks, Miles Brundage, Kevin Button,	
417	Trevor Cai, Rosie Campbell, Andrew Cann, Brittany	
418	Carey, Chelsea Carlson, Rory Carmichael, Brooke	
419	Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully	
420	Chen, Ruby Chen, Jason Chen, Mark Chen, Ben	
421	Chess, Chester Cho, Casey Chu, Hyung Won Chung,	
422	Dave Cummings, Jeremiah Currier, Yunxing Dai,	
423	Cory Decareaux, Thomas Degry, Noah Deutsch,	
424	Damien Deville, Arka Dhar, David Dohan, Steve	
425	Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti,	
426	Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,	
427	Simón Posada Fishman, Juston Forte, Isabella Ful-	
428	ford, Leo Gao, Elie Georges, Christian Gibson, Vik	
429	Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-	
430	Lopes, Jonathan Gordon, Morgan Grafstein, Scott	
431	Gray, Ryan Greene, Joshua Gross, Shixiang Shane	
432	Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris,	
433	Yuchen He, Mike Heaton, Johannes Heidecke, Chris	
434	Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele,	
435	Brandon Houghton, Kenny Hsu, Shengli Hu, Xin	
436	Hu, Joost Huizinga, Shantanu Jain, Shawn Jain,	
437	Joanne Jang, Angela Jiang, Roger Jiang, Haozhun	
438	Jin, Denny Jin, Shino Jomoto, Billie Jonn, Hee-	
439	woo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-	
440	mali, Ingmar Kanitscheider, Nitish Shirish Keskar,	
441	Tabarak Khan, Logan Kilpatrick, Jong Wook Kim,	
442	Christina Kim, Yongjik Kim, Jan Hendrik Kirchner,	
443	Jamie Kiros, Matt Knight, Daniel Kokotajlo,	
444	Łukasz Kondraciuk, Andrew Kondrich, Aris Kon-	
445	stantinidis, Kyle Kopic, Gretchen Krueger, Vishal	
446	Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan	
447	Leike, Jade Leung, Daniel Levy, Chak Ming Li,	
448	Rachel Lim, Molly Lin, Stephanie Lin, Mateusz	
449	Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue,	
450	Anna Makanju, Kim Malfacini, Sam Manning, Todor	
451	Markov, Yaniv Markovski, Bianca Martin, Katie	
452	Mayer, Andrew Mayne, Bob McGrew, Scott Mayer	
453	McKinney, Christine McLeavey, Paul McMillan,	
454	Jake McNeil, David Medina, Aalok Mehta, Jacob	
455	Menick, Luke Metz, Andrey Mishchenko, Pamela	
456	Mishkin, Vinnie Monaco, Evan Morikawa, Daniel	
457	Mossing, Tong Mu, Mira Murati, Oleg Murk, David	
458	Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak,	
459	Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh,	
460	Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex	
461	Paino, Joe Palermo, Ashley Pantuliano, Giambat-	
462	tista Parascandolo, Joel Parish, Emy Parparita, Alex	
463	Passos, Mikhail Pavlov, Andrew Peng, Adam Perel-	
464	man, Filipe de Avila Belbute Peres, Michael Petrov,	
	Henrique Ponde de Oliveira Pinto, Michael, Poko-	465
	rny, Michelle Pokrass, Vitchyr H. Pong, Tolly Pow-	466
	ell, Alethea Power, Boris Power, Elizabeth Proehl,	467
	Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh,	468
	Cameron Raymond, Francis Real, Kendra Rimbach,	469
	Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-	470
	der, Mario Saltarelli, Ted Sanders, Shibani Santurkar,	471
	Girish Sastry, Heather Schmidt, David Schnurr, John	472
	Schulman, Daniel Selsam, Kyla Sheppard, Toki	473
	Sherbakov, Jessica Shieh, Sarah Shoker, Pranav	474
	Shyam, Szymon Sidor, Eric Sigler, Maddie Simens,	475
	Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin	476
	Sokolowsky, Yang Song, Natalie Staudacher, Fe-	477
	lipe Petroski Such, Natalie Summers, Ilya Sutskever,	478
	Jie Tang, Nikolas Tezak, Madeleine B. Thompson,	479
	Phil Tillet, Amin Tootoonchian, Elizabeth Tseng,	480
	Preston Tuggle, Nick Turley, Jerry Tworek, Juan Fe-	481
	lipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya,	482
	Chelsea Voss, Carroll Wainwright, Justin Jay Wang,	483
	Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei,	484
	CJ Weinmann, Akila Welihinda, Peter Welinder, Ji-	485
	ayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner,	486
	Clemens Winter, Samuel Wolrich, Hannah Wong,	487
	Lauren Workman, Sherwin Wu, Jeff Wu, Michael	488
	Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim-	489
	ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong	490
	Zhang, Marvin Zhang, Shengjia Zhao, Tianhao	491
	Zheng, Juntang Zhuang, William Zhuk, and Bar-	492
	ret Zoph. 2024. Gpt-4 technical report . <i>Preprint</i> ,	493
	arXiv:2303.08774.	494
	Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke	495
	Ishigami, Yuta Nakashima, Esa Rahtu, Janne	496
	Heikkilä, and Shin’ichi Satoh. 2023. Toward ver-	497
	ifiable and reproducible human evaluation for text-to-	498
	image generation. In <i>Proceedings of the IEEE/CVF</i>	499
	<i>Conference on Computer Vision and Pattern Recog-</i>	500
	<i>nition (CVPR)</i> , pages 14277–14286.	501
	Sara Rosengren, Martin Eisend, Scott Koslow, and Mi-	502
	cael Dahlen. 2020. A meta-analysis of when and how	503
	advertising creativity works . <i>Journal of Marketing</i> ,	504
	84(6):39–56.	505
	Sameh Said-Metwaly, Wim Van den Noortgate, and Eva	506
	Kyndt. 2017. Approaches to measuring creativity: A	507
	systematic literature review . <i>Creativity. Theories –</i>	508
	<i>Research - Applications</i> , 4(2):238–275.	509
	Pooja Sharma. 2012. Advertising effectiveness: " un-	510
	derstanding the value of creativity in advertising", a	511
	review study in india. <i>Online Journal of Communi-</i>	512
	<i>cation and Media Technologies</i> , 2(3):1.	513
	Dean Keith Simonton. 2012. Quantifying creativity:	514
	can measures span the spectrum? <i>Dialogues in Clinical</i>	515
	<i>Neuroscience</i> , 14(1):100–104. PMID: 22577309.	516
	Stephen Skalicky. 2022. Liquid gold down the drain:	517
	Measuring perceptions of creativity associated with	518
	figurative language and play . <i>Cognitive Semantics</i> ,	519
	8(1):79 – 108.	520
	Robert E. Smith, Scott B. MacKenzie, Xiaojing Yang,	521
	Laura M. Buchholz, and William K. Darley. 2007.	522

523	Modeling the determinants and effects of creativity	Pitt-Ads Dataset and want to keep the terminol-	577
524	in advertising. <i>Marketing Science</i> , 26(6):819–833.	ogy aligned.	578
525	Yih Hwai Lee Swee Hoon Ang, Siew Meng Leong and	A.2 Sampling Process	579
526	Seng Lee Lou. 2014. <i>Necessary but not sufficient:</i>	The original Pitt Ads dataset consists of 38 top-	580
527	<i>Beyond novelty in advertising creativity. Journal of</i>	ics in total (Hussain et al., 2017). We group those	581
528	<i>Marketing Communications</i> , 20(3):214–230.	topics into 10 “topic groups”: clothing, food, auto-	582
529	Remziye Terkan. 2014. Importance of creative advertis-	mobile, beauty, leisure, electronics, drinks, service,	583
530	ing and marketing according to university students’	non-commercial, and healthcare. From each topic	584
531	perspective. <i>International Review of Management</i>	group, we sampled 10 ads, which makes up the 100	585
532	<i>and Marketing</i> , 4(3):239–246.	images in Creativity-100.	586
533	Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan	Details about “topic groups”	587
534	Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi,	• food: “Restaurants, cafe, fast food”, “Choco-	588
535	Thomas Griffiths, and Faeze Brahman. 2024. <i>Mac-</i>	late, cookies, candy, ice cream”, “Chips,	589
536	<i>Gyver: Are large language models creative problem</i>	snacks, nuts, fruit, gum, cereal, yogurt,	590
537	<i>solvers? In Proceedings of the 2024 Conference of</i>	soups”, “Seasoning, condiments, ketchup”	591
538	<i>the North American Chapter of the Association for</i>	• pet: “Pet food”	592
539	<i>Computational Linguistics: Human Language Tech-</i>	• drinks: “Alcohol”, “Coffee, tea”, “Soda,	593
540	<i>nologies (Volume 1: Long Papers)</i> , pages 5303–5324,	juice, milk, energy drinks, water”	594
541	Mexico City, Mexico. Association for Computational	• automobile: “Cars, automobiles”	595
542	Linguistics.	• electronics: “Electronics”	596
543	Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang,	• service: “Phone, TV and internet service	597
544	Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie,	providers”, “Financial services”, “Security	598
545	Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and	and safety services”, “Software”, “Other ser-	599
546	Yue Zhang. 2024. <i>Pandalm: An automatic evaluation</i>	VICES”	600
547	<i>benchmark for llm instruction tuning optimization.</i>	• education: “Education”	601
548	<i>Preprint</i> , arXiv:2306.05087.	• beauty: “Beauty products and cosmetics”	602
549	Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao	• healthcare: “Healthcare and medications”	603
550	Song, Markus Freitag, William Wang, and Lei Li.	• clothing: “Clothing and accessories”	604
551	2023. Instructscore: Towards explainable text genera-	• home: “Baby products”, “Cleaning products”	605
552	tion evaluation with automatic feedback. In <i>Proceed-</i>	, “Home improvements and repairs”, “Home	606
553	<i>ings of the 2023 Conference on Empirical Methods</i>	appliances”]	607
554	<i>in Natural Language Processing</i> , pages 5967–5994.	• leisure: “Games and toys”, “Vacation and	608
555	Keren Ye, Narges Honarvar Nazari, James Hahn, Za-	travel”, “Media and arts”, “Sports equipment	609
556	eem Hussain, Mingda Zhang, and Adriana Kovashka.	and activities”, “Gambling”	610
557	2019. Interpreting the rhetoric of visual advertise-	• shopping: “Shopping”	611
558	ments. <i>IEEE transactions on pattern analysis and</i>	• non-commercial: “Environment, nature, pollu-	612
559	<i>machine intelligence</i> , 43(4):1308–1323.	tion, wildlife”, “Animal rights, animal abuse”,	613
560	Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu	“Human rights”, “Safety, safe driving, fire	614
561	Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and	safety”, “Smoking, alcohol abuse”, “Domes-	615
562	Jiawei Han. 2022. <i>Towards a unified multi-</i>	tastic violence”, “Self esteem, bullying, cyber	616
563	<i>dimensional evaluator for text generation. In Pro-</i>	bullying”, “Political candidates”, “Charities”	617
564	<i>ceedings of the 2022 Conference on Empirical Meth-</i>		
565	<i>ods in Natural Language Processing</i> , pages 2023–		
566	2038, Abu Dhabi, United Arab Emirates. Association		
567	for Computational Linguistics.		
568	A Ads Dataset		
569	A.1 Terminology		
570	In the original study (Smith et al., 2007) that pro-		
571	posed the breakdown of creativity, they use the		
572	term “synthesis” instead of atypicality. However,		
573	they defined “synthesis” as “...combine, connect, or		
574	blend normally unrelated objects or ideas” which		
575	is very similar to “atypicality”. We use the term		
576	“atypicality” because that is the term used in the		

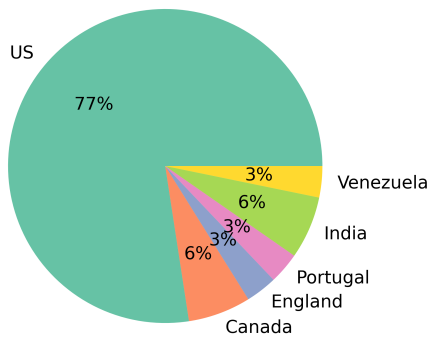


Figure 2: Distribution of workers' response to "In which country did you live the longest time so far?"

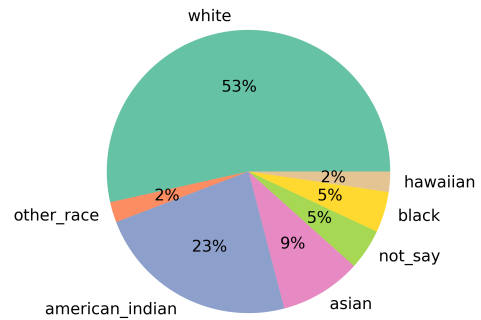


Figure 4: Distribution of workers' response to "Race Identification (select one or more)"

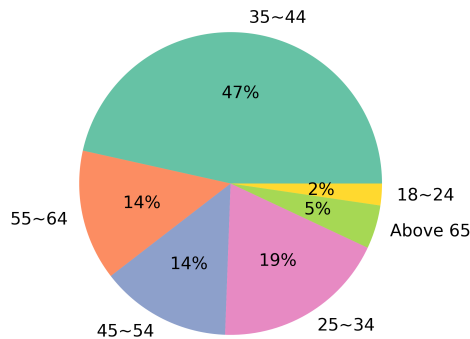


Figure 3: Distribution of workers' response to "What is your age?"

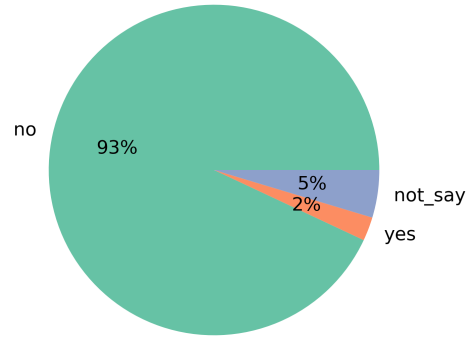


Figure 5: Distribution of workers' response to "Are you of Hispanic or Latino/a origin?"

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 Fig. 22 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 Fig. 2, 3, 4, 5, and 6. 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.

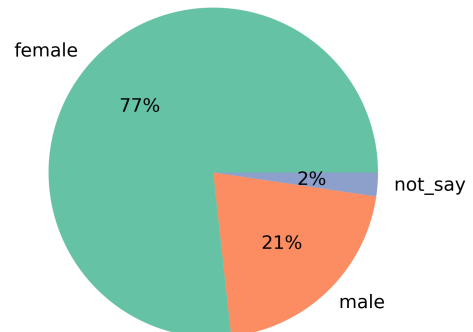


Figure 6: Distribution of workers' response to "What is your gender?"

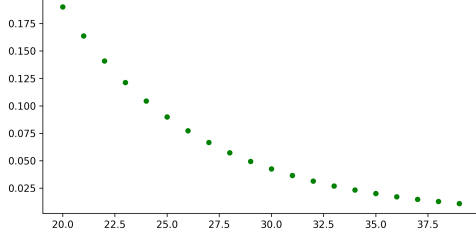


Figure 7: Upper-bound of the error based on calculation.

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:

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

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 Fig. 7).

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 Tab. 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 (Fig. 8, 9).

D.2 Distribution Modeling Examples

We have three examples with all the scoring metrics; see Fig. 10, 11, 12. We have also plotted WordClouds for the reasoning part of the output (Fig. 15, 18, 21). 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).

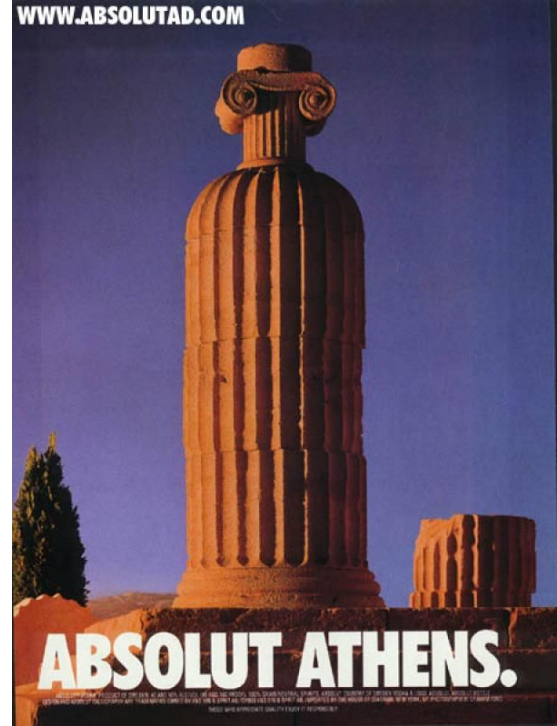
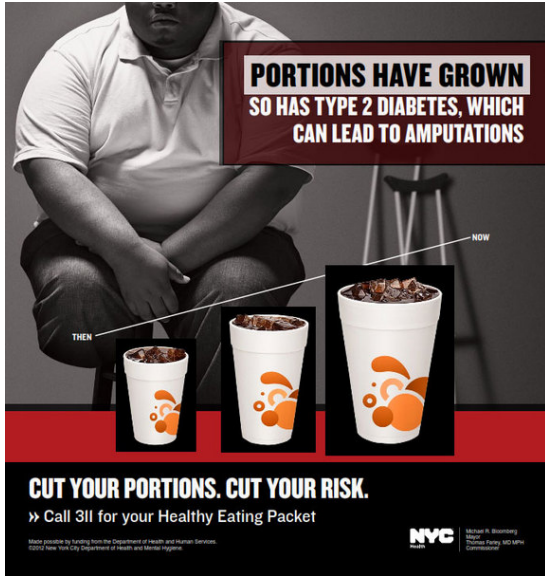


Figure 8: Task: **atypicality**; model outputs are in Tab. 4

Model	Model Output	Prediction
LLaVA-7B	<p>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</p>	right (2)
LLaVA-13B	<p>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.</p>	left (1)
InternVL2-8B	<p>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}</p>	right (2)
GPT-4v	<p>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</p>	right (2)

Table 4: Model output for ads pair in Fig. 8. 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 9: Task: **originality**; model outputs are in Tab. 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 Fig. 9. 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.

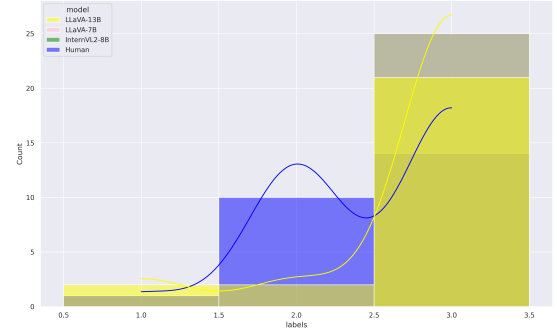


Figure 10: Example (A) and creativity predictions by models; complete output in Tab. 6

Aspect	Human	LLaVA-7B (<i>KL</i>)	LLaVA-13B (<i>KL</i>)	InternVL (<i>KL</i>)
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 Fig. 10; *KL* refers to $KL(Human||Model)$

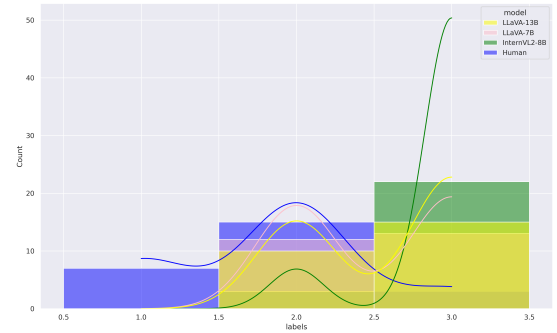


Figure 11: Example (B) and creativity predictions by models; complete output in Tab. 7

Aspect	Human	LLaVA-7B (<i>KL</i>)	LLaVA-13B (<i>KL</i>)	InternVL (<i>KL</i>)
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 Fig. 11; *KL* refers to $KL(Human||Model)$

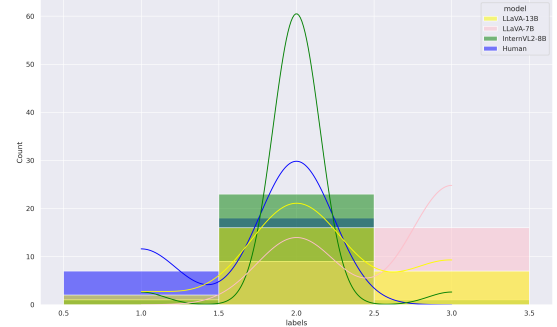


Figure 12: Example (C) and creativity predictions by models; complete output in Tab. 8

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 Fig. 12; KL refers to $KL(Human||Model)$



Figure 13: Task: Creativity,
Model: LLaVA-7B



Figure 14: Task: Creativity,
Model: LLaVA-13B



Figure 15: Task: Creativity,
Model: InternVL2-8B



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

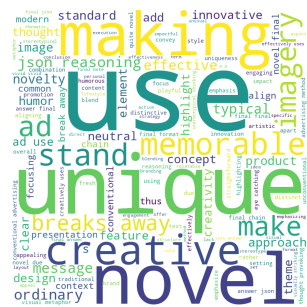
Figure 17: Task: Originality,
Model: LLaVA-13B

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

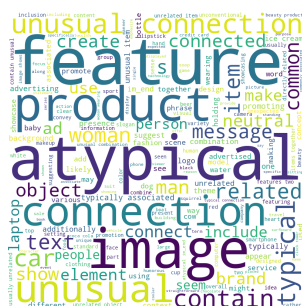


Figure 19: Task: Atypicality,
Model: LLaVA-7B

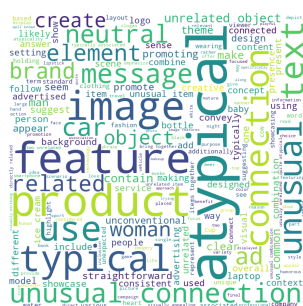


Figure 20: Task: Atypicality, Model: LLaVA-13B

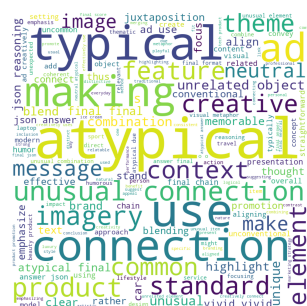


Figure 21: Task: Atypicality,
Model: InternVL2-8B

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

Ad image



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

Figure 22: Amazon Mechanical Turk interface.