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

Anonymous ACL submission

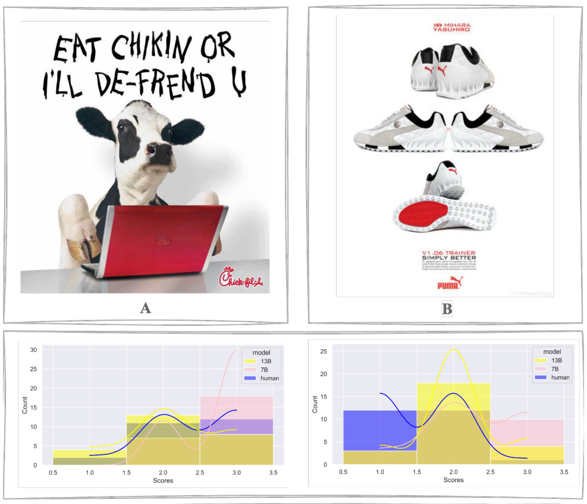
Abstract

Evaluating creativity is a challenging task, even for humans, not only because it is subjective, but also because it involves complex cognitive processes. Inspired by previous work in marketing, we attempt to break down creativity into atypicality and originality and collect fine-grained human annotation on these categories. With controlled experiments with vision language models (VLM), we evaluate the alignment between models and humans on a suite of novel tasks. Our results demonstrate both the promises and challenges of using VLMs for automatic creativity assessment.[1]

1 Introduction

Creativity is one of the most complex aspects of human cognition. Many researchers favor a definition of creativity that involves divergence and non-obviousness (Till and Baack, 2005; El-Murad and West, 2004a; Simonton, 2012). For example, in the advertisement (A) in Fig. 1, the image of a cow sitting in front of a computer and typing on the keyboard is a divergence from the norm (i.e., cows simply cannot do that); non-obviousness is achieved when we combine the text "Eat chikin or I'll de-friend U" and the small logo of Chick-fil-A to infer that the ad urges people to eat at Chick-fil-A. Decoding the ad thus requires background knowledge and drawing connections, making the evaluation of creativity a challenging task.

In adverising, creativity plays a critical role that motivates consumer behaviors (Sharma, 2012; Terkan, 2014a,b). Therefore, it is necessary for ad creators to consistently create and evaluate creative ad content. Extensive research has been conducted to understand what the general public would consider creative (El-Murad and West, 2004b; Rosengren et al., 2020; Swee Hoon Ang and Lou, 2014; Smith et al., 2007). However, these rely on domain experts, which are expensive and inaccessible.



	Human Avg. (A / B)	LLaVA-7B (A / B / pairwise)	LLaVA-13B (A / B / pairwise)	GPT-4v (A / B / pairwise)
Creativity	2.40 / 1.56	3 / 1 / A	3 / 2 / A	3 / 3 / A
Originality	2.72 / 1.56	3 / 3 / A	2 / 2 / A	2 / 2 / A
Atypicality	2.92 / 1.24	3 / 3 / A	3 / 2 / A	3 / 1 / A

Figure 1: Top: 2 ads from dataset; Middle: score annotations and outputs from VLMs (25 each); Bottom: average scores from annotators, single label and pairwise predictions; Scores are 3-scale, 3 being the best.

Recently, foundational models (Bommasani et al., 2021) have demonstrated impressive performances in other evaluation tasks, such as summarization, Long-Form QA (Jiang et al., 2023), and commonsense text generation (Xu et al., 2023), many of which were previously dominant by human evaluation. This poses the question of whether foundational models have an understanding of creativity, specifically, can we use visual language models (VLMs) to measure the creativity of visual advertisements? Prior works evaluate creativity in text, while we investigate whether we can expand creativity measurement on multimodal ads.

To this end, we conduct several fine-grained, automatic evaluations of creativity for visual adver-

[1] We will release data and code upon paper publication.

tisements. Based on studies in marketing and cognitive science (Smith et al., 2007; Chakrabarty et al., 2024a), we decompose creativity into atypicality and originality. We then collect high-quality, fine-grained human evaluations of advertisement images. We experiment with state-of-the-art (SoTA) VLMs to predict these ratings and examine the human-model alignment. In addition to the traditional emphasis on prediction accuracy, we extend our evaluation to the model's ability to gauge annotator disagreements and capture the subjective nature of the task through analysis of crowd annotation distributions. We show that VLMs perform better in pairwise tasks than intrinsic tasks (i.e. one image at a time). We also find disagreement prediction and distribution modeling challenging, both of which require more high-quality annotations in future research. Our benchmark and evaluation metrics provide a solid foundation for utilizing VLMs to evaluate and assist visual content creators.

2 Related Work

Evaluation of Creativity Research in the evaluation of creativity includes cognitive science (Said-Metwaly et al., 2017a; Simonton, 2012; James Lloyd-Cox and Bhattacharya, 2022; Said-Metwaly et al., 2017b), marketing (El-Murad and West, 2004b; Rosengren et al., 2020; Swee Hoon Ang and Lou, 2014; Smith et al., 2007), creative writing (Skalicky, 2022), HCI (Chakrabarty et al., 2024b), and AI (Chakrabarty et al., 2023, 2024a). There are two common grounds: first, creativity is the balance between divergence and effectiveness; second, evaluation of creativity is subjective, making fine-grained human feedback critical. (Smith et al., 2007) focused on advertisement images and proposed five creativity subcategories including atypicality and originality. We adapt their creativity decomposition. (Chakrabarty et al., 2024a) use-large language models (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) propose to decompose the evaluation of a complex task into simpler ones that can be accomplished by language models; whereasPandaLM (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 either evaluating text or images alone, instead of the image-text pair as we do.

Section	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

3 Dataset

3.1 Ads Dataset

We used the Pitt Ads Dataset (referred to as `Pitt-Ads`) (Hussain et al., 2017; Ye et al., 2019) with 64,832 image ads, of which 4,185 contain atypicality annotations. 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 category of atypicality, e.g., new object created from combining existing real objects, object missing parts, etc. We first sample 10 ads with at least one atypical object and another 10 that do not. We use this evaluation set for fine-grained human creativity annotation (Sec. 3.2) (referred to as `Creative-20`). From the remaining ads with atypicality annotations, we sample 300 images as our second evaluation set; we use it for atypicality prediction on a larger scale (Sec. 3.3) (referred to as `Atypical-300`).

3.2 Fine-grained Creativity Annotation

Deconstruction of Creativity We break down the concept of creativity into two categories: **originality** and **atypicality**. This breakdown is inspired by (Smith et al., 2007), where they proposed five factors that contribute to creativity. Based on their analysis, originality and synthesis are the two most influential ones. We adapt their definition of originality and synthesis[2], and renamed synthesis to atypicality due of the definition similarity with the existing atypically annotation in `Pitt-Ads`.

Human Annotation Human annotation is collected via Amazon Mechanical Turk (Mturk). Each task is structured into the following sections: ads

[2]They define synthesis as: "*...combine, connect, or blend normally unrelated objects or ideas*" and originality as "*...contain elements that are rare, surprising, or move away from the obvious and commonplace.*"

image, atypicality, originality, quality check question, overall creativity, and demographics. For atypicality and originality, we follow Smith et al. (2007) and present three statements (see Table 1) to the workers; they can answer either "agree", "neutral", or "disagree". For the overall creativity rating, the annotator can answer from 1 to 5 with a higher number means more creative. See Appendix A for more details on this.

Due to the inherent subjectivity of the creativity judgment, we view the questions with three possible answers as a categorical distribution with three choices.[3] To make sure we gathered enough annotations to cover the true creativity annotation distribution, we follow previous work on approximating the true distribution with 0.1 error rate (McHugh, 2012; Cheng et al., 2024) (Appendix B). Thus each advertisement needs to be annotated by 25 workers. See Appendix C for information on how we post-process the data, including standardizing scores.

Annotator Agreement We computed inter-annotator agreement for three score categories with Randolph's Kappa (Randolph, 2010; Seabold and Perktold, 2010): 0.32 for atypicality, 0.24 for originality and 0.25 for overall creativity, which fall in the category of "minimal agreement" (McHugh, 2012). This agreement level confirms all categories are subjective and motivates us to propose the "distribution" and "disagreement" tasks (Sec. 4.2).

3.3 Atypicality Data

We also randomly sampled 300 ads from `Pitt-Ads`. Each ad has three binary annotations on atypicality. Out of 300 images, 185 (62%) has at least one positive atypicality annotation. In both Smith et al. (2007) and human annotated data, we show atypicality has a positive and statistically significant correlation with creativity (shown in Appendix D). Therefore, we additionally evaluate this dataset to gain insight about creativity.

4 Experimental Setup

4.1 Models

We experiment with open-sourced vision language models (VLM), i.e. 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 [4] and run on a single NVIDIA A100 GPU. More details are in Appendix E.

4.2 Task Formulation

For the decomposed creativity: atypicality and originality, along with the creativity itself, we define two groups of tasks: intrinsic tasks and pairwise task. Intrinsic tasks entail prompting VLM with a single advertisement image at a time, with the objective of gauging the model's predictive capabilities in its most probable application scenario, namely, generating a creativity score for a given image. In contrast, pairwise tasks are simpler, as they merely require the VLM to rank a pair of images.

Intrinsic Tasks We first evaluate the performance of VLMs in the most traditional way: computing the accuracy by comparing model output with the label with majority annotators. We then compare model output with the average annotator score and compute Spearman's correlation across all ad images, which provides an overview of how model prediction and human judgment align. See the *Majority* and *Avg. Rating* columns in Table 2.

Given the subjective nature of creativity and low annotator agreements, we design two additional intrinsic tasks: distribution modeling and disagreement prediction. For distribution modeling, we prompt VLMs multiple times with high temperatures so that we get the same number of VLM outputs as the number of annotators; we then compute the KL Divergence between the distribution of human rating and VLM ratings. In this way, we quantify the distance between models and humans in a "group behavior" setting. For disagreement prediction, we directly prompt VLMs to predict the level of disagreement for each scoring category; 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. These two results are in *Distribution* and *Disagreement* columns in Table 2.

Pairwise Task We also propose an easier pairwise preference task where two ads with different average ratings are presented to the VLM with the prompt to pick a preferred one based on the scoring category. For each scoring category, we include all ad pairs with average human ratings differences greater than 0.5. For `Creative-20`, we have 55, 113, and 108 pairs in creativity, atypicality, and originality; for `Atypical-300`, we sampled 1000 pairs due to constraints in computation

[3]While creativity annotation is done on a five scale, we convert all annotations to a three scale in Appendix C.

[4]VLM prompts are in Appendix F.

Category	Model	Intrinsic: Majority ↑ *Acc.*	Intrinsic: Average Rating ↑ *R (p-value)*	Intrinsic: Distribution ↓ *KL Divergence*	Intrinsic: Disagreement ↑ *R (p-value)*	Pairwise: Pairwise ↑ *Acc.*
Creativity (`Creative-20`)	LLaVA 7B	**0.50**	0.23 (0.330)	0.62	*nan*	0.75
	LLaVA 13B	0.45	0.30 (0.203)	**0.30**	-0.38 (0.103)	0.65
	GPT-4v	**0.50**	**0.67 (0.001)**	-	*nan*	**0.91**
Originality (`Creative-20`)	LLaVA 7B	0.35	-0.18 (0.448)	0.71	*nan*	0.72
	LLaVA 13B	0.30	0.08 (0.730)	**0.67**	0.17 (0.463)	0.62
	GPT-4v	**0.50**	**0.70 (0.001)**	-	0.10 (0.677)	**0.97**
Atypicality (`Creative-20`)	LLaVA 7B	0.35	0.32 (0.169)	0.55	*nan*	0.81
	LLaVA 13B	0.50	**0.49 (0.027)**	0.55	-0.13 (0.595)	0.67
	GPT-4v	**0.70**	**0.72 (<0.001)**	-	0.393 (0.086)	**0.90**
Atypicality (`Atypical-300`)	LLaVA 7B	0.56	**0.13 (0.029)**	**0.02**	*nan*	0.57
	LLaVA 13B	0.58	0.05 (0.390)	0.30	-0.05 (0.406)	0.46
	GPT-4v	**0.66**	**0.32 (<0.001)**	-	0.01 (0.849)	**0.65**

Table 2: Bold results: best-performing models or 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 constraints. For intrinsic tasks, `Creative-20` labels are 3-scale and `Atypical-300` labels are binary.

resources. The results are evaluated by accuracy and are shown in *Pairwise* column in Table 2.

5 Results [5]

Atypicality, Originality, and Creativity VLMs generally perform better on atypicality than creativity and originality, and there is no clear differences for pairwise tasks. We believe this is because atypicality is more well-defined than originality and creativity when there is no comparison available, as atypicality implicitly requires comparison against the "typical world", such as physics rules and social norms; whereas models do not have this natural anchor to compare to when it comes to originality and creativity.

Cross Dataset Performance We can also see a clear performance gap between the two datasets, which we believe is due to the difference in annotation numbers. For each ad in `Atypical-300`, there are only 3 binary annotations of atypicality whereas there are 25 3-scale annotations in `Creative-20`. We believe this motivates future research involving subjective labels like creativity and atypicality to collect more annotations to avoid noise in the annotation.

Disagreement Prediction Remains Challenging In many cases, the VLMs failed the disagreement task by predicting the same output for all samples or demonstrating random correlation scores compared to human annotations. This suggests that using VLM as a group-opinion synthesizer remains challenging. Future work could 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.

Performance on Distribution Modeling In `Creative-20`, the LLaVA-13B model generally outperforms the 7B model whereas the result is reversed in `Atypical-300`. Our output analysis (Appendix G) shows that KL measurements indeed capture the distribution differences between humans and model outputs. We believe it is worth extending to a larger scope of datasets and tasks.

Performance in Ranking-based Task Although the average rating correlation is an intrinsic task, the underlying Spearman's correlation is a ranking-based method. The pairwise task is also ranking-based as it essentially ranks two images at a time. Therefore, it is a promising sign that GPT-4v shows impressive performances in this two ranking-based task. Future research can look into potential methods built upon GPT-4v to evaluate creativity in a ranking fashion.

6 Conclusion

We present a case study of using VLMs to evaluate creativity in advertisements. With a theoretical grounding in marketing research, we collect fine-grained human annotation on creativity ratings and test the alignment between the SoTA VLMs and humans. Our work is good starting point for automatic evaluation of creativity.

7 Limitations

One obvious limitation is the size of our dataset. The fine-grained creativity annotation only con-

[5] More output analysis can be found in Appendix G

sists of 20 ad images. Two bottom necks 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 budge requirement. We have also encountered the issue of poor annotation where half of the annotators failed the validation question in the first few batches of annotation collection. However, we believe what we have shown in this case study is that our overall framing and methodology can be generalized to a larger scope in the future, where more annotation would be conducted.

Another limitation is the subjective nature of the task. In particular, 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 researchers to be cautious when applying our method to data in other country or language.

Due to hardware constraints, we only experiment with LLaVA 13B when 34B is available. We also have other VLM choices such as BLIP, CLIP, etc. We will leave more extensive prompt tuning and model selections to future work.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

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 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.

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 Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *ArXiv*, abs/2310.00752.

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.

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

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.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286.

Justus Randolph. 2010. Free-marginal multirater kappa (multirater free): An alternative to fleiss fixed-marginal multirater kappa. volume 4.

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. 2017a. Approaches to measuring creativity: A systematic literature review. *Creativity. Theories – Research - Applications*, 4(2):238–275.

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

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

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

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

Brian D. Till and Daniel W. Baack. 2005. Recall and persuasion: Does creative advertising matter? *Journal of Advertising*, 34(3):47–57.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *Preprint*, arXiv:2306.05087.

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.

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.

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.

A Amazon Mechanical Turk Details

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

Quality check questions The quality check question asks the worker to choose the expected action from five action options, all from `Pitt-Ads`. The correct action corresponds to the ad image in the same HIT, and the other four are randomly sampled from other ads. The overall accuracy on this question is 93.2%, which means the workers understand visual advertisements and pay enough attention to the annotation task.

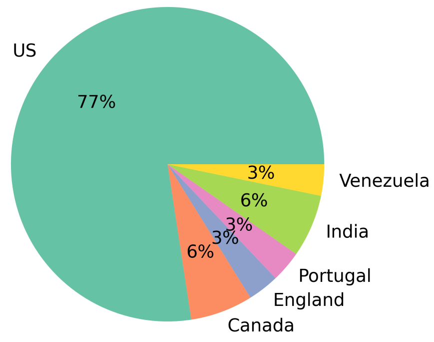


Figure 2: Distribution of workers' response to "In which country did you live the longest time so far?"

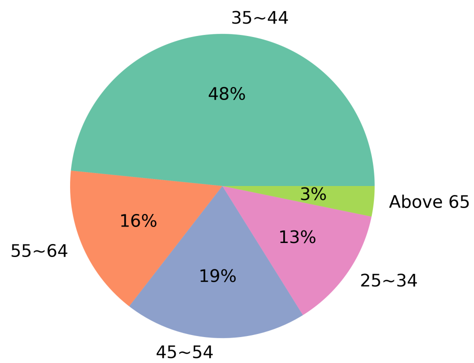


Figure 3: Distribution of workers' response to "What is your age?"

Annotation interface See Figure 8 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, 31 workers contributed to our task and finished 500 HITs. Their background can be found in Figure 2 and 3. As we can see, the annotators are strongly skewed towards the US-based middle-age group, which should be kept in mind when applying our methodology when it comes to people from another background.

B 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

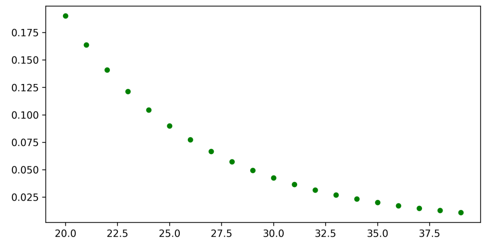


Figure 4: Upper-bound of the error based on calculation.

as follows:

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

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 Figure 4).

C 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 category 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 category, 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.

D Connection between atypicality and creativity

After analyzing the fine-grained creativity data we collected (Sec. 3.2), we find out that the Pearson R correlation between the normalized atypicality and overall creativity score is 0.3776 ($p < 0.01$), a positive correlation[6]. Therefore, it makes sense to evaluate the same methodology on data with only atypicality annotation to prove its effectiveness at a larger scale.

[6]The sample size is 500: 20 ads with 25 annotations each.

E Experiment Details

Configurations

- Temperature: 0.75 (for distribution prediction) and 0.001 (for all other tasks)
- Max New Token: 256
- Quantization (LLaVA only): load in 8bit
- Model Checkpoint
  - GPT-4: `gpt-4-vision-preview`
  - LLaVa7B: `llava-v1.6-mistral-7b-hf`
  - LLaVa13B: `llava-v1.6-vicuna-13b-hf`
- Number of pairwise samples (% of label "1")
  - creativity: 55 (47%)
  - atypicality: 113 (41%)
  - originality: 108 (50%)

Running Time (Approximately)

- `Creative-20`
  - GPT4-v: 30min for all tasks combined;
  - LLaVA7B: 9hr for all tasks combined;
  - LLaVA7B: 12hr for all tasks combined;
- `Atypical-300` (atypical data only)
  - GPT4-v: 2hr for all tasks combined;
  - LLaVA7B: 12hr for all tasks combined;
  - LLaVA7B: 16hr for all tasks combined;

F VLM Prompts

Creativity

Single Label & Distribution Modeling
How creative is this visual advertisement? Give your answer in the scale of 1 to 3 with 1 being not creative at all, 2 being neutral, and 3 being very creative. Give your answer in the following format: "answer: {score}; explanation: {reasoning}"

Disagreement

I am about to ship this advertisement design to the public and I am unsure how would the audience intepret it. Some might consider it creative (i.e. compose of creative ideas) while some others would not. To what extent would they agree on each other?
Make your best guess and give me an agreement score between 1 to 3, with 1 being easily agree (high agreement), 2 being neutral, and 3 being hardly agree (low agreement).
Give your answer in the following format: "answer: {score}; explanation: {reasoning}"

Pairwise

Here are two images of advertisement. Which one is more likely to succeed in catching people's eyes by being creative? 1 for the left image and 2 for the right image. Give your answer in the following format: "explanation: {reasoning}; answer: {choice}"

Atypicality

Single Label & Distribution Modeling

How unusual (i.e. including abnormal objects or atypical connotations) about the advertisement? Give your answer in the scale of 1 to 3 with 1 being very normal, 2 being neutral, and 3 being very unusual and abnomal. Give your answer in the following format: "answer: {score}; explanation: {reasoning}" "

Disagreement

I am about to ship this advertisement design to the public and I am unsure how would the audience intepret it. Some might consider it unusual (i.e. some abnormal objects or connections) while some others would not. To what extent would they agree on each other? Make your best guess and give me an agreement score between 1 to 3, with 1 being easily agree (high agreement), 2 being neutral, and 3 being hardly agree (low agreement). Give your answer in the following format: "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}"

Originality

Single Label & Distribution Modeling

How novel (i.e. unique from previous ads) is this visual advertisement? Give your answer in the scale of 1 to 3 with 1 being not original at all, 2 being neutral, and 3 being very unusual and outstanding. Give your answer in the following format: "answer: {score}; explanation: {reasoning}"

Disagreement

I am about to ship this advertisement design to the public and I am unsure how would the audience intepret it. Some might consider it original (i.e. unique of its kind) while some others would not. To what extent would they agree on each other? Make your best guess and give me an agreement score between 1 to 3, with 1 being easily agree (high agreement), 2 being neutral, and 3 being hardly agree (low agreement). Give your answer in the following format: "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 left one and 2 for the right one. Give your answer in the following format: : "explanation: {reasoning}; answer: {choice}"

`Atypical-300` Prompts (atypicality only)

Single Label & Distribution Modeling

How unusual (i.e. including abnormal objects or atypical connotations) about the advertisement? Give an answer of either 0 or 1; answer 0 for being very normal and 1 being very unusual and abnomal. Give your answer in the following format: "answer: {score}; explanation: {reasoning}"

Disagreement

I am about to ship this advertisement design to the public and I am unsure how would the audience intepret it. Some might consider it unusual (i.e. some abnormal objects or connections) while some others would not. To what extent would they agree on each other? Make your best guess and give me an agreement score of either 0 or 1, with 1 for no agreement, 0 high agreement. Give your answer in the following format: "answer: {score}; explanation: {reasoning}"

G Output Examples

We have three examples with all the scoring metrics, see Figure 5, 6, 7.

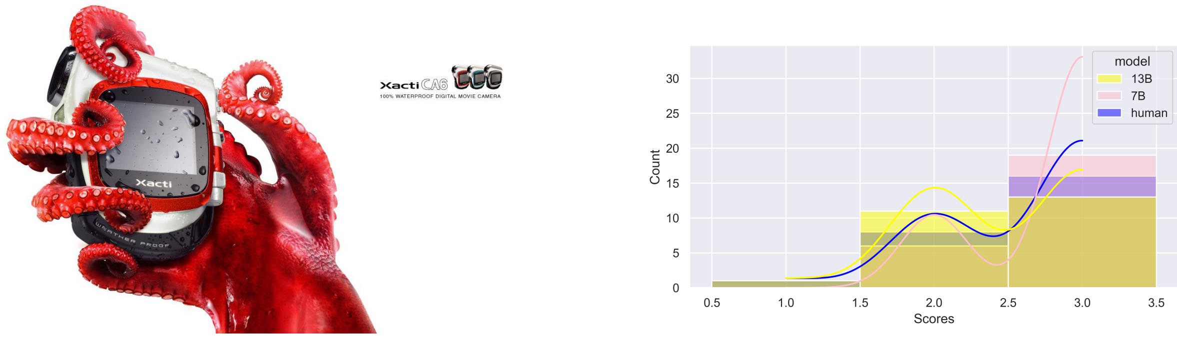


Figure 5: Example (A) and `creativity` predictions by models; complete output in Table 3

Aspect	Human	GPT-4v	LLaVA-7B	LLaVA-13B	$KL(H\|\|LLaVA-7B)$	$KL(H\|\|LLaVA-13B)$
Creativity	2.60	3	2	3	0.0456	0.0367
Originality	2.92	3	3	2	0.0000	0.4091
Atypicality	2.92	3	3	3	0.0270	0.0000

Table 3: Model output and human ratings for Example (A), see ad image and distribution modeling result in Figure 5

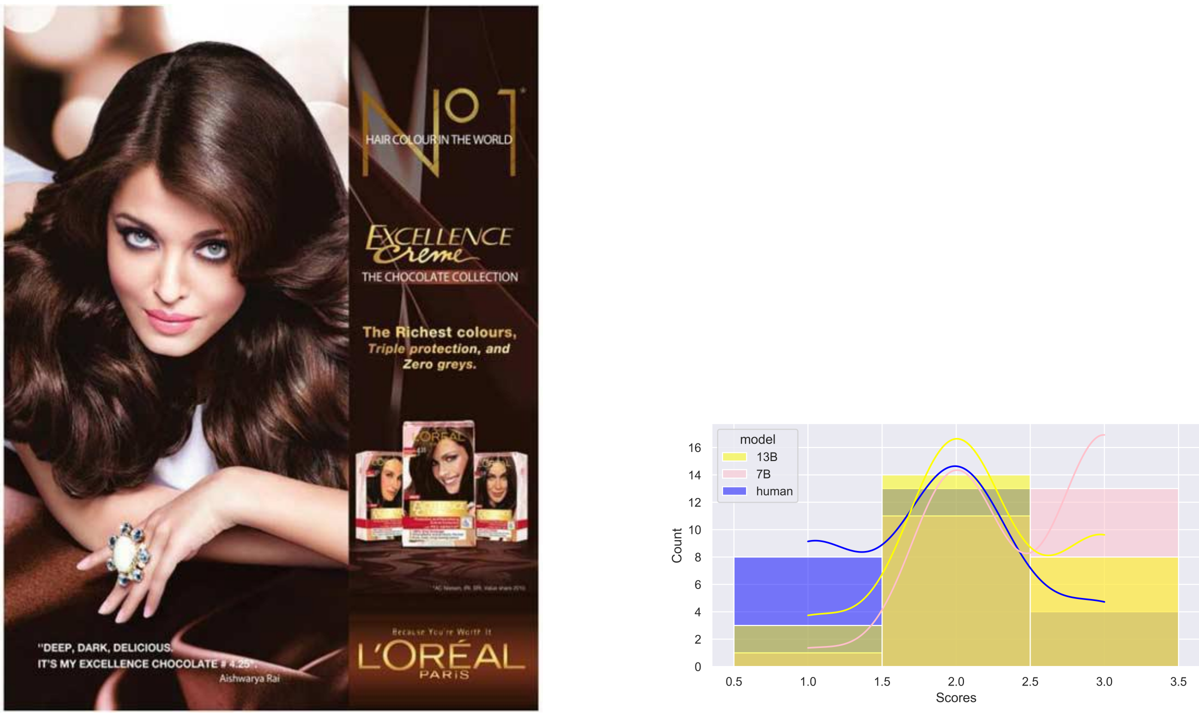


Figure 6: Example (B) and `creativity` predictions by models; complete output in Table 4

Aspect	Human	GPT-4v	LLaVA-7B	LLaVA-13B	$KL(H\|\|LLaVA-7B)$	$KL(H\|\|LLaVA-13B)$
Creativity	1.84	2	3	3	0.5434	0.1749
Originality	1.44	1	3	2	2.3743	1.4753
Atypicality	1.28	1	2	1	1.3535	1.1474

Table 4: Model output and human ratings for Example (B), see ad image and distribution modeling result in Figure 6

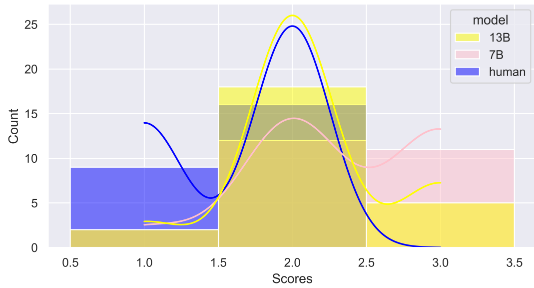


Figure 7: Example (C) and `creativity` predictions by models; complete output in Table 5

Aspect	Human	GPT-4v	LLaVA-7B	LLaVA-13B	$KL(H\|\|LLaVA-7B)$	$KL(H\|\|LLaVA-13B)$
Creativity	1.64	2	3	2	0.7273	0.4306
Originality	1.36	1	2	1	1.334	0.6839
Atypicality	1.44	1	2	1	0.6141	0.6203

Table 5: Model output and human ratings for Example (C), see ad image and distribution modeling result in Figure 7

Before you start: if you believe that you have done this exact HIT (i.e. have seen this *exact one* advertisement), please skip to the next ad.

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

Other Questions

What is your age?

○ Below 18 ○ 18~24 ○ 25~34 ○ 35~44 ○ 45~54 ○ 55~64 ○ 65 and above ○ Prefer not to answer

In which country did you live the longest time so far?

Please let us know if you have any feedback about this HIT (e.g., question unclear / ambiguous, etc.)

Submit

Figure 8: Amazon Mechanical Turk interface.