[34] Ruofan Liang, Tianlin Li, Longfei Li, Jing Wang, and Quanshi Zhang. 2020. Knowledge Consistency between Neural Networks and Beyond. arXiv:1908.01581 (2020). http://arxiv.org/abs/1908.01581
[35] Kelsey Medieros, David H Cropley, Rebecca L Marrone, and Roni Reiter-Palmon. [n. d.]. Human-AI Co-Creativity: Does ChatGPT make us more creative? ([n. d.]).
[36] Sarnoff Mednick. 1962. The associative basis of the creative process. *Psychological review* (1962).
[37] Kibum Moon, Adam Green, and Kostadin Kushlev. 2024. Homogenizing Effect of Large Language Model (LLM) on Creative Diversity: An Empirical Comparison of Human and ChatGPT Writing. (2024).
[38] Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences* 118, 25 (2021), e2022340118.
[39] Vivek Pandya. 2024. The Age of Generative AI: Over half of Americans have used generative AI and most believe it will help them be more creative. *Adobe* (2024). https://blog.adobe.com/en/publish/2024/04/22/age-generative-ai-over-half-americans-have-used-generative-ai-most-believe-will-help-them-be-more-creative.
[40] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* (2016).
[41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162
[42] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of EMNLP*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084
[43] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109* (2024).
[44] Massimo Stella, Thomas T Hills, and Yoed N Kenett. 2023. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences* 120, 43 (2023), e2312911120.
[45] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting GPT-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932* (2022).
[46] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018* (2023).
[47] C Szegedy. 2014. Intriguing properties of neural networks. *Proc. of ICLR* (2014).
[48] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
[49] Jamba Team, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, et al. 2024. Jamba-1.5: Hybrid Transformer-Mamba Models at Scale. *arXiv preprint arXiv:2408.12570* (2024).
[50] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[51] Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2018. With great training comes great vulnerability: Practical attacks against transfer learning. In *27th USENIX security symposium (USENIX Security 18)*. 1281–1297.
[52] Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209* (2024).
[53] Simone Zhang, Janet Xu, and A Alvero. 2024. Generative ai meets open-ended survey responses: Participant use of ai and homogenization. (2024).
[54] Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491* (2024).
[55] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS nexus* 3, 3 (2024), pgae052.
[56] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).

A DIVERGENT THINKING TEST WORDING

Here, we report the exact wording for the tests given to humans and LLMs. The wording differs slightly between the two groups because the LLM models are prompted to output their work in a particular format for easier processing, while human prompts refer to text boxes in the survey UI. Without formatting instructions in the prompt, LLMs often discussed the reasoning behind their word choices. While mildly interesting, this muddied the data.

A.1 AUT prompts.

For original experiments, we use the following start words for AUT: `WORD` = {book, fork, table, hammer, pants}. For the expanded LLM evaluation of §5.2, we use `WORD` = {book, bottle, brick, fork, hammer, pants, shoe, shovel, table, tire}.

Human prompt. *Imagine that someone gives you* `WORD`. *In the blanks below, write down as many creative uses you can think of for this object, up to 10 uses.*

LLM prompt. *Imagine that someone gives you a* `WORD`*. Write down as many uses as you can think of for this object, up to 10 uses. Please list the uses as words or phrases (single word answers are ok), separated by semicolons. Do not write anything besides your proposed uses.*

A.2 Forward Flow prompts.

We use the following start words for Forward Flow: `WORD` = {candle, table, bear, snow, toaster}.

Human prompt. (From the original Flow paper) *Starting with the word* `WORD`*, in each of the following blanks, write down the next word that follows in your mind from the previous word. Please put down only single words, and do not use proper nouns (such as names, brands, etc.). Start by writing* `WORD` *in the first space below.*

LLM prompt. *Starting with the word* `WORD`*, your job is to write down the next word that follows in your mind from the previous word. Please put down only single words, and do not use proper nouns (such as names, brands, etc.). Stop after you listed at least 22 words. Print just the list of words, separated by commas, and do not add anything else to your response. The first word in the list should be 'candle'.*

A.3 DAT Prompts.

Human prompt. (From the original DAT paper) *In the spaces below, please enter 10 words that are as different from each other as possible, in all meanings and uses of the words. You must follow the following rules: 1. Only single words in English. 2. Only nouns (e.g., things, objects, concepts). 3. No proper nouns (e.g., no specific people or places). 4. No specialised vocabulary (e.g., no technical terms). 5. Think of the words on your own (e.g., do not just look at objects in your surroundings). 6. Complete this task in less than four minutes.*

LLM prompt. *Instructions: Please enter 10 words that are as different from each other as possible, in all meanings and uses of the words. Rules: 1. Only single words in English. 2. Only nouns (e.g., things, objects, concepts). 3. No proper nouns (e.g., no specific people or places). 4. No specialised vocabulary (e.g., no technical terms). 5. Think of the words on your own (e.g., do not just look at objects in your surroundings). 6. Complete this task in less than four minutes. 7. Return just the list of words, separated by commas, and do not include any other content.*

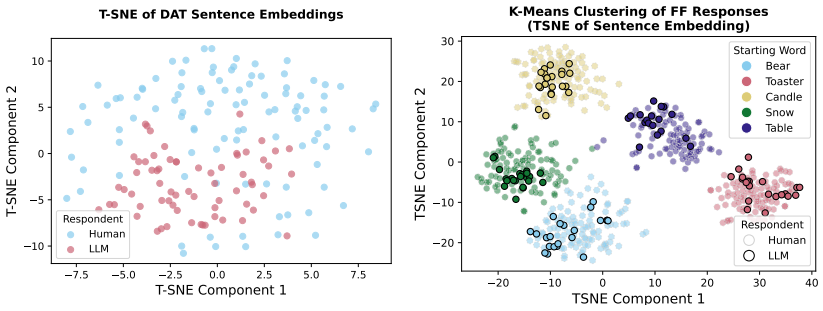

Fig. 10. LLM responses to the DAT and FF cluster more in feature space than do human responses.

B ORIGINALITY SCORES FOR AUT, FF, AND DAT

Here, we describe our methods of computing originality scores for each test. Originality scores are denoted as $O_t(\mathcal{P})$, where t = AUT, FF, or DAT and $\mathcal{P}$ is a population, either humans or LLMs.

We denote a single word test response as r and an n-word test response as $\mathbf{r} = \{r_0, r_1, \dots r_n\}$. The word embedding model is $\mathcal{W}$, and the embedding of a response r is $\mathcal{W}(r)$ (similar for $\mathbf{r}$, $\mathbf{r_j}$, etc.). We use cosine similarity $\cos(\mathcal{W}(r_1), \mathcal{W}(r_2))$ to measure semantic distance between embedded responses.

AUT scoring. Following [19], we score the originality of AUT responses by measuring the semantic distance between a prompt p (e.g. "book") and each word in $\mathbf{r}$ (e.g. "use it as a doorstop"). Because different words in the AUT response contribute differently to overall response creativity (e.g. "it" matters less than "doorstop"), the final originality score is computed via a weighted sum of these distances. Weights are determined by running TF-IDF analysis on the corpus of responses, which produces low weights for common words like "it" and high weights for unusual words like "doorstop". The set of originality scores for AUT responses of population $\mathcal{P}$ is then:

$$O_{\text{AUT}}(\mathcal{P}) = \left\{ 1 - \frac{\sum_{j=0}^{n-1} w_j \cdot \cos(\mathcal{W}(p), \mathcal{W}(r_j))}{\sum_{j=0}^{n-1} w_j}, \forall (\mathbf{r}, p) \in \mathcal{P} \right\} \tag{2}$$

where w_j is the TF-IDF weight for the j^{th} word of response $\mathbf{r}$ and p is the prompt.

FF scoring: Here, we follow the methodology of [22]. This defines the "instantaneous" forward flow of a given thought in the sequence $\mathbf{r}$ as the average distance between the m^{th} thought in the sequence r_m and all preceding thoughts:

$$\frac{\sum_{j=1}^{m-1}(1 - \cos(\mathcal{W}(r_j), \mathcal{W}(r_m)))}{m-1}$$

Building on this, the set of FF scores for a population $\mathcal{P}$ consisting of n-word sequences $\mathbf{r}$ is given by:

$$O_{\text{FF}}(\mathcal{P}) = \left\{ \frac{1}{n-1} \cdot \sum_{i=2}^{n} \frac{\sum_{j=1}^{i-1}(1 - \cos(\mathcal{W}(r_j), \mathcal{W}(r_i)))}{(i-1)}, \forall\ \mathbf{r} \in \mathcal{P} \right\} \tag{3}$$

DAT scoring: We use the scoring methodology of [38], which scores responses by averaging the semantic distance between all pairs of words in the response. Given a population $\mathcal{P}$ composed of n-long DAT response $\mathbf{r}$ containing words $\{r_0, r_1, \dots r_{n-1}\}$, the set of DAT scores is calculated as:

$$O_{\text{DAT}}(\mathcal{P}) = \left\{ \frac{1}{n(n-1)} \sum_{i,j(i \neq j)}^{n-1} (1 - \cos(\mathcal{W}(r_i), \mathcal{W}(r_j))) \forall\ \mathbf{r} \in \mathcal{P} \right\} \tag{4}$$

C TSNE OF FF AND DAT

Figure 10 visualizes the TSNE of sentence embeddings for DAT and FF responses, similar to Figure 3. This confirms the trend observed in the AUT TSNE: LLM responses cluster closer in feature space than human responses, resulting in lower population-level originality measurements. We perform k-means clustering of TNSE of FF sentence embeddings to demonstrate clusters of LLM response for each start word. Since the DAT since the test does not involve varying start words, we simply visualize all LLM and human responses.