CreativityNeuro: Steering Language Model Weights to Improve Divergent Thinking

Anonymous authors
Paper under double-blind review

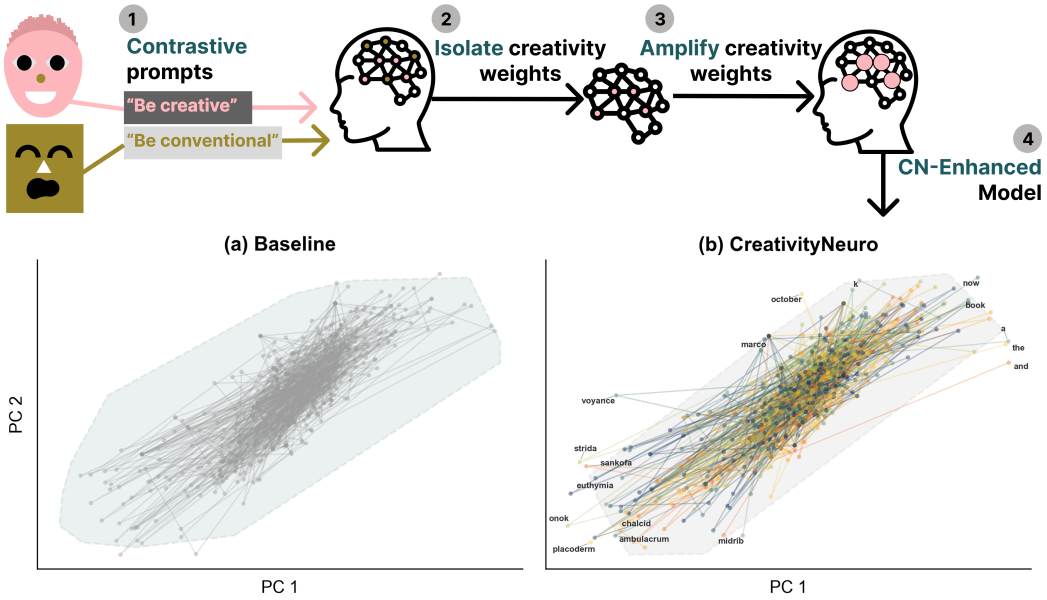

Figure 1: **CreativityNeuro (CN) pipeline.** Given a pair of contrastive creative prompts, CN computes parameter importance scores, selects a sparse subset of creativity-relevant parameters, and applies a scaled weight perturbation—without requiring behavioral datasets or gradient-based finetuning. CN improves divergent thinking across various tasks. Subplot (b) visualizes CN *thinking outside of the "box"*(i.e., the convex hull of baseline DAT responses), despite baseline responses falling within CN's convex hull in subplot (a).

Abstract

We introduce CreativityNeuro (CN), a data-free method for enhancing creative behavior in LLMs via contrastive weight steering. We evaluate CN on various creativity assessments and report three main findings: **(1)** On the Divergent Association Task (DAT), CN improves performance by up to 14 human percentile points, with gains that are consistent across semantically diverse prompts. **(2)** In a large-scale human evaluation (N=720) on the Alternative Uses Test (AUT) and the Task Task (TT), CN-enhanced models achieve significant improvements in originality, surprise, and creativity, demonstrating transfer to more complex, open-ended settings. Compared to contrastive activation steering (CAA), we find that while activation steering can match CN on the DAT, it does not achieve consistent improvements on held-out tasks (AUT, TT). **(3)** CN-enhanced models reduce factual reasoning scores on MMLU by 1–11%, and attempts to explicitly preserve such capabilities by incorporating MMLU prompts into prompt contrast sets reverse nearly all divergent thinking (DT) gains, providing evidence that DT and factual reasoning rely on functionally entangled weights.

1 Introduction

Recent advances in large language models (LLMs) have renewed interest in a longstanding question: *how can we understand and enhance creativity in intelligent systems?* (Boden, 2004). While this question has deep roots in cognitive science (Quetelet, 1842; Galton, 1870; Hadamard, 1954; Guilford, 1956; Mednick, 1962; Koestler, 1964; Simonton, 2004; Dietrich, Arne, 2004; Fauconnier & Turner, 2008; Rothenberg, 2014), it is now increasingly studied in the context of large-scale generative models (Maher, 2010; Varshney, 2019; Schapiro et al., 2025). Recent work has begun to assess the capacity for LLMs to engage in creative and open-ended tasks (Si et al., 2024; 2025; Sanyal et al., 2025; Bellemare-Pepin et al., 2024; Wang et al., 2025; 2024), where a recurring issue has surfaced: models tend to consistently generate similar responses to open-ended questions, in what has been termed the *artificial hivemind* effect (Jiang et al., 2025).

Within the creativity literature, a common distinction is made between *divergent thinking* (DT), the capacity to generate multiple diverse solutions to a problem, and *convergent thinking* (CT), the ability to find a single correct solution that unifies multiple diverse stimuli (Dietrich, 2019; Guilford, 1956). Studying ways to enhance DT offers a promising pathway to encourage greater diversity and novelty in model responses, combating the homogenization issues that have emerged thus far. Thus, in this work, we introduce a weight-space steering method that improves DT in large language models. Our method outperforms prior approaches—including decoding, prompting, and activation steering—and generalizes better to unseen tasks, without requiring behavioral data or gradient-based fine-tuning. In detail, our main contributions are as follows:

1. We introduce **CreativityNeuro** (CN), a data-free method for steering creative behavior. **CN improves divergent thinking** on the Divergent Association Task (DAT), outperforming baselines such as prompting, activation steering, and decoding baselines.
2. We conduct a large-scale human evaluation on the Alternative Uses Test (AUT) and Task Task (TT) and find that **CN-enhanced models improve originality, surprise, and creativity on the AUT and TT.** Despite performing comparably to CN on the DAT, **activation steering transfers poorly to the AUT and TT**.
3. We investigate the impact of CN on factual reasoning and find that improvements in divergent thinking come at the cost of reduced MMLU performance. Attempts to preserve factual reasoning ability reverse nearly all DAT gains, providing **evidence that divergent thinking and factual reasoning rely on functionally entangled model weights.**

Outline In Section 2, we discuss related work and introduce our method. Next, in Section 3 and Section 4, we perform experiments on the DAT, AUT, and TT. Then, in Section 5, we study the tradeoff between divergent thinking and factual reasoning. Lastly, we discuss limitations, future work, and conclude in Section 6.

2 Background and Method

2.1 Related Work

Evaluating the creativity of LLMs. Previous work has evaluated LLMs on divergent creativity assessments–including the DAT (Olson et al., 2021), AUT (Guilford, 1956), the Task Task (Chu et al., 2024)–and in various real-world settings like scientific ideation (Si et al., 2024; 2025) and open-ended user queries (Jiang et al., 2025). On the DAT, LLMs can achieve scores well into the 90th percentile of humans (Bellemare-Pepin et al., 2024; Wang et al., 2025), whereas Stevenson et al. (2022) studied GPT-3 on the AUT and concluded that humans exhibited greater creativity, with model responses showing weaker originality. Lastly, Chu et al. (2024) found that model-generated goals on the Task Task achieved similar creativity ratings as human-generated goals, as assessed by a large panel of human raters.

Algorithm 1 CreativityNeuro

Require: Model weights $\{W_\ell\}_{\ell=1}^{L}$; creative prompts $\mathcal{P}^{\text{cre}}$; non-creative prompts $\mathcal{P}^{\text{non-cre}}$; importance threshold ρ; scaling factor α

Ensure: Modified weights $\{W'_\ell\}_{\ell=1}^{L}$ with amplified creativity weights

for each layer ℓ **do**

  Run forward passes on $\mathcal{P}^{\text{cre}}, \mathcal{P}^{\text{on-cre}}$ to obtain importance scores $S^{\text{cre}}_{\ell,ij}, S^{\text{non-cre}}_{\ell,ij}$ via:

$$S_{\ell,ij}(\mathcal{P}) = \sum_{b=1}^{|\mathcal{P}|} \sum_{t=1}^{T_b} |W_{\ell,ij}| \cdot \|\mathbf{x}_{\ell,j}^{(b,t)}\|_2 \qquad \textbf{Step 1: Compute weight importance scores}$$

  for prompt b and token position t within prompt b

end for

for each layer ℓ **do**

  $C_\ell \leftarrow$ top-ρ weights ranked by $S^{\text{cre}}_{\ell,ij}$ **Step 2: Extract creative-specific subspaces**

  $N_\ell \leftarrow$ top-ρ weights ranked by $S^{\text{non-cre}}_{\ell,ij}$

  $M^{\text{cre-spec}}_{\ell,ij} \leftarrow \mathbb{I}[(i,j) \in C_\ell \setminus N_\ell]$

end for

for each layer ℓ **do**

  $W'_\ell \leftarrow W_\ell \odot (1 + \alpha \cdot M^{\text{cre-spec}}_\ell)$ **Step 3: Creative parameter scaling**

end for

return $\{W'_\ell\}_{\ell=1}^{L}$

Table 1: **Examples of contrastive prompt sets.** Each set contains creative ($\mathcal{P}^{\text{cre}}$) and non-creative ($\mathcal{P}^{\text{non-cre}}$) prompts used for parameter importance scoring in Algorithm 1.

Set	Creative ($\mathcal{P}^{\text{cre}}$)	Non-Creative ($\mathcal{P}^{\text{non-cre}}$)
STORY	Write the first line of a story that makes the reader question reality.	Create a typical story beginning that establishes setting and character clearly.
PROBLEM	What would an alien civilization's approach to this problem look like?	What best practices should be followed when solving this?
MINIMAL	Surprise me.	Be precise.

Improving the creativity of LLMs. Most similar to this work, Olson et al. (2024) proposed an activation steering method to amplify the creativity of LLMs, although improvements were only established for a single model, task, and human annotator. Our study is the first to demonstrate a steering method to improve creative behavior and validate its effectiveness in a large-scale human study. Apart from steering, various other approaches to improving LLM creativity include prompting frameworks (Nguyen & Singla, 2025; Morain & Ventura, 2025; Wang et al., 2025), varying decoding parameters such as temperature (Peeperkorn et al., 2024), and reinforcement learning (RL) on preference data (Wei et al., 2025). Unlike steering and RL-based approaches that typically require labeled behavioral data, our method operates entirely data-free.

2.2 Method

Our method relies on identifying parameters that influence creative behavior. Christ et al. (2025) has recently shown that parameter importance methods can be used to identify and amplify weights involved in mathematical reasoning, improving scores on the MATH benchmark by 4–17% (Hendrycks et al., 2021). Here, we adopt the framework of Christ et al. (2025), which computes Wanda-style (Sun et al., 2023) importance scores by taking by the product of weight magnitude and activation norm, summed across a set of prompts b and their token positions t: $S_{\ell,ij} = \sum_{b,t} |W_{\ell,ij}| \cdot \|\mathbf{x}_{\ell,j}^{(b,t)}\|_2$, where $\mathbf{x}_{\ell,j}^{(b,t)}$ is the j-th input activation

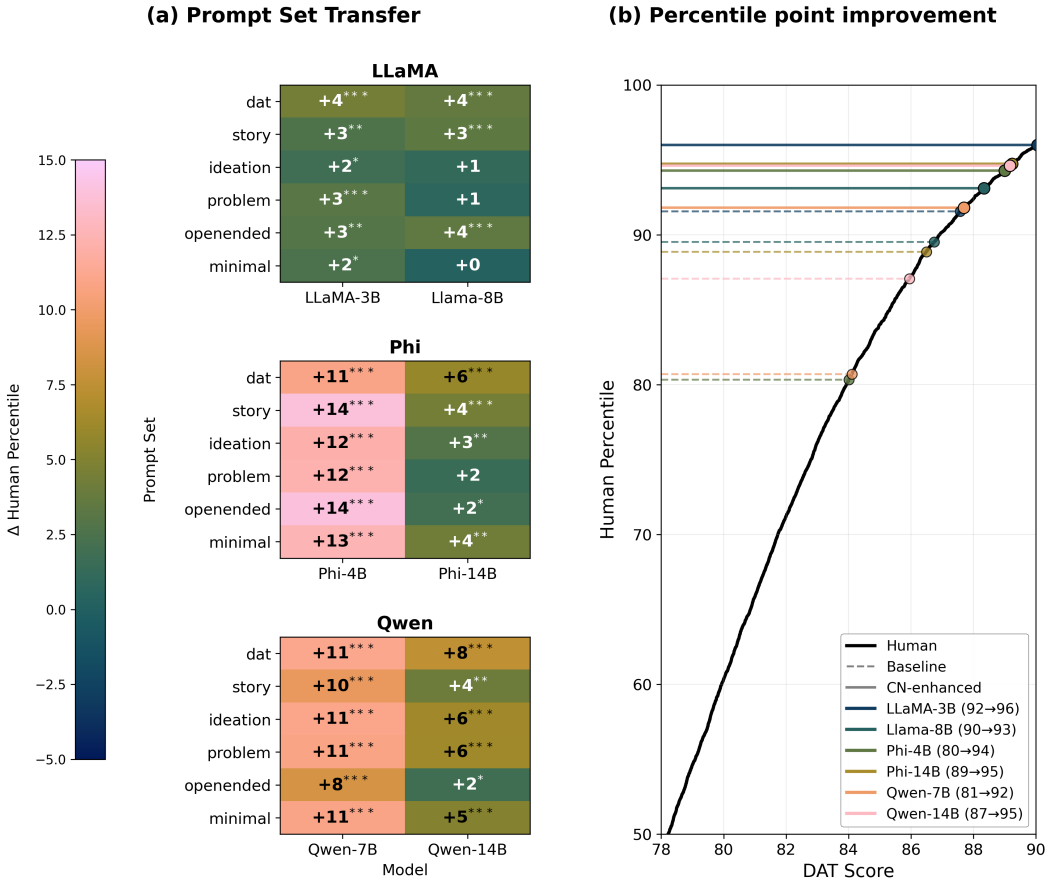

Figure 2: **CreativityNeuro (CN) improves divergent thinking across models and prompt sets.** Given a human reference distribution (Wang et al., 2025) ($N = 9{,}297$, $\mu = 78.26$, $\sigma = 6.73$), we report: (a) Heatmap showing human percentile improvement (Δ%ile) for CN-enhanced models, with statistical significance ($p < 0.05$) at each of the temperatures tested (0.9, 1.0, 1.2) denoted by an asterisk. (b) CDF showing CN-enhanced scores for the best performing prompt set.

at layer ℓ for token t in prompt b. We compute importance scores on creative $\mathcal{P}^{\text{cre}}$ and non-creative prompts $\mathcal{P}^{\text{non-cre}}$ (representative examples are given in Table 1; all six prompt sets are given in Table 2). Then, we isolate creativity-specific parameters by selecting the top ρ percent of weights ranked by creative importance that do not also appear in the top ρ percent for non-creative prompts. This set difference operation ($C_\ell \setminus N_\ell$) ensures we identify parameters uniquely associated with creative behavior. Lastly, at inference time, we multiply weights in $C_\ell \setminus N_\ell$ by a scaling factor $(1 + \alpha)$. Like Christ et al. (2025), this method is data- and backpropagation-free. The hyperparameters ρ (importance threshold) and α (scaling strength) control the intervention's scope and intensity. The full procedure is given in Algorithm 1.

3 Experiments on the Divergent Associates Task

We test instruct-tuned models across three open-weight model families (Phi, Llama, Qwen), totaling six models at 3B, 4B, 7B, 8B, and 14B sizes: LLaMA (3.2-3B-Instruct, 3.1-8B-Instruct) (Grattafiori et al., 2024), Qwen-2.5 (7B-Instruct, 14B-Instruct) (Yang et al., 2025), and Phi (3.5-mini-Instruct (4B), 3-medium-4k-Instruct (14B)) from Microsoft (Abdin et al., 2024). Full hyperparameter sweep details for CreativityNeuro parameters are available in Section D.

3.1 The Divergent Associates Task (DAT)

The DAT asks participants to generate N words that are as semantically distant from each other as possible (Olson et al., 2021). Given a set of N words $W := \{w_1, w_2, \ldots, w_N\}$ with corresponding GloVe embeddings $V := \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\} \subseteq \mathbb{R}^{300}$, the DAT score is the average pairwise semantic distance among all distinct pairs of those N words:

$$\mathrm{DAT}(W) := \frac{100}{N(N-1)} \sum_{i \neq j}^{N} (1 - \cos(\mathbf{v}_i, \mathbf{v}_j)) \tag{1}$$

Following Olson et al. (2021), we use the 840B-token GloVe embeddings (Pennington et al., 2014) as our semantic space, providing consistent distance measurements independent of response pool demographics. Typically, a participant is asked to name $N = 10$ words, and the first 7 valid words are kept (Olson et al., 2021). The prompts used to administer the DAT are given in Section E.

3.2 Baselines

Existing studies have found that temperature-scaling and prompting can influence DAT scores (Wang et al., 2025; Bellemare-Pepin et al., 2024). To ensure the CN intervention leads to a meaningful improvement over such techniques, we compare against a broad set of baselines, including prompting; varying decoding parameters such as top-p nucleus sampling, top-k sampling, temperature, and repetition penalty; as well as activation steering via contrastive activation addition (CAA; Panickssery et al. (2024)), which injects a steering vector $\mathbf{v}_\ell = \bar{\mathbf{h}}_\ell^+ - \bar{\mathbf{h}}_\ell^-$ into the residual stream during decoding. For activation steering, contrast pairs are obtained from top- vs. bottom-quartile DAT responses by score, creating a "divergent thinking" direction in the residual stream.[1] Full decoding strategy and activation steering hyperparameter settings are given in Section B. All settings are evaluated across all six models at $T \in \{0.9, 1.0, 1.2\}$ until N=120 valid DAT responses are obtained.

3.3 Results on the Divergent Associates Task

CN improves DAT performance across all six models and prompt sets (Figure 2), outperforming all sampling-based baselines (Figure 3). Panel (a) of Figure 2 shows Δ Percentile for the best (ρ, α) per model–prompt combination: while the DAT prompt set produces the most consistent gains, non-DAT prompt sets also yield statistically significant improvements, suggesting **CN is able to identify weights controlling divergent thinking behavior, rather than localizing DAT-specific task knowledge**.

CN (94.1 avg) slightly outperforms activation steering (93.9 avg) on DAT percentile (Figure 3); however, activation steering requires scored DAT responses to construct its steering vector, while CN uses only creative and non-creative prompts with no generation or scoring. Prompt-only activation steering (omitted from the table; see footnote) averaged only 87.8, comparable to prompting (87.9), suggesting that the behavioral signal is essential for activation steering to be competitive, whereas CN extracts a stronger signal from prompts alone.

> CreativityNeuro improves DAT scores across all six models and prompt sets, outperforming prompting, sampling-parameter, and activation steering baselines.

[1] We also tested a prompt-only CAA variant using forward-pass activations from the same creative vs. non-creative prompt sets as CN, but it underperformed the behavioral-data variant on all models and is omitted for clarity.

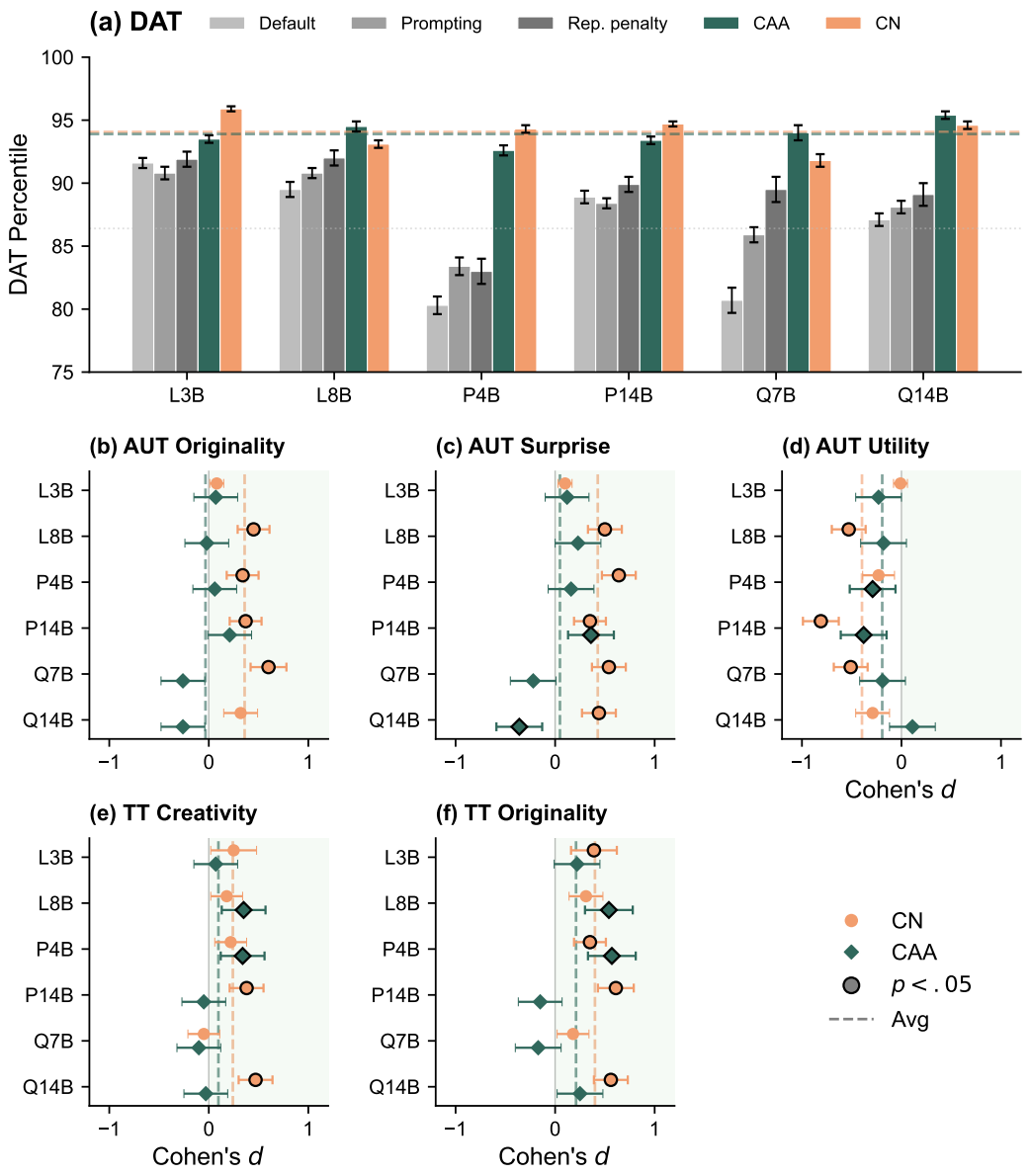

Figure 3: **Results on the DAT, AUT, and TT, as well as comparison against baselines.** **(a)** DAT human percentile (±SEM) averaged across $T \in \{0.9, 1.0, 1.2\}$. **(b–f)** Cohen's d (±SE) from intra-participant z-scored human ratings on the AUT and TT. Black outlines indicate $p < .05$, and green shading marks the positive-effect region.

4 Generalization to the Alternative Uses Test and the Task Task

4.1 The Alternative Uses Test (AUT) and the Task Task (TT)

The Alternative Uses Test is a divergent thinking assessment introduced by Guilford (1956) that asks participants to generate creative uses for a common everyday object (e.g., "List alternative uses for a brick"). Unlike the DAT, which measures semantic distance among single words, the AUT requires participants to produce functionally meaningful ideas, each of which may span multiple words or sentences. We administer the AUT using a standard set of objects commonly used in the creativity literature: *brick*, *paperclip* and *fork*. Following Stevenson et al. (2022), uses are scored on *originality*, *surprise*, and *utility*. Meanwhile, the

(a) **Alternative Uses Test** — "List alternative uses for a ____."

CN-enhanced (Llama-8B), $\bar{z} = 0.67$

Book: A book can be transformed into a giant marimba by hollowing out its pages, gluing wood strips to the hollowed-out sections, and tuning metal or glass material to create a makeshift percussion instrument.

originality = 1.12 surprise = 1.32 utility = −0.42

Baseline (Phi-4B), $\bar{z} = 0.36$

Book: A book can transform into a unique, eco-friendly planter, nurturing a tiny ecosystem as it blossoms with flowers.

originality = 0.58 surprise = 0.70 utility = −0.20

(b) **Task Task** — "Design a creative game show challenge."

CN-enhanced (Qwen-14B), $\bar{z} = 1.07$

Your objective is to construct a functional periscope allowing clear line-of-sight around an obstacle, mimicking an espionage thrill; you must see a small target widget placed behind a bookcase [...] You may utilize a stack of encyclopedias, a roll of wrapping paper, a broomstick, two mirrors, safety goggles, and a length of garden hose...

creativity = 1.08 originality = 1.06

Baseline (Qwen-14B), $\bar{z} = 0.44$

Your goal is to create a makeshift catapult using only a skateboard, a bicycle tire inner tube, a cooking pot lid, and six feet of string. Using this device, you must launch a raw egg into an oversized wading pool filled with colorful foam noodles placed fifteen feet away. Each successful catapult that lands the egg in the pool scores five points; however, the egg must remain intact upon impact to earn the full five points.

creativity = 0.50 originality = 0.37

Figure 4: **Top-rated CN-enhanced stimuli vs. baseline generations from the same model.** Intra-participant z-scores averaged across raters (N=40 per cell). (a) CN-enhanced AUT responses tend to score higher on originality and surprise while sometimes sacrificing utility; (b) CN-enhanced Task Task challenges layer more constraints, props, and absurdist elements.

Task Task evaluates the ability to generate novel challenges or goals that themselves expect novel solutions (Chu et al., 2024). Participants are asked to design creative game show challenges that would be fun to attempt, entertaining to watch, and difficult enough to be interesting. Following Chu et al. (2024), we evaluate responses on creativity and originality.[2] Full prompts used to administer the AUT and TT are given in Section E.

4.2 Human Experiment Design

We evaluate AUT and TT stimuli via human ratings on Prolific. For each model, we select the CN configuration (ρ, α, prompt set) that produced the largest statistically significant DAT improvement in Section 3. Then, for each task, we sample 40 stimuli (20 baseline, 20 CN-enhanced) at temperature $T = 1.0$, top-$k = 0$, and top-$p = 1.0$. All studies uses a between-subjects design, where every participant rates 10 stimuli (5 baseline, 5 creative) in randomized order with condition labels hidden. We ensure balanced allocation via automated participant-to-slot assignment so that each stimulus receives exactly 10 independent ratings. Moreover, on both the AUT and TT, ratings for each dimension are given on continuous 0–100 sliders, and to control for individual differences in scale usage, we compute intra-participant z-scores: for each participant i and dimension d, $z_{i,d,s} = (r_{i,d,s} - \bar{r}_{i,d}) \,/\, \sigma_{i,d}$, where $\bar{r}_{i,d}$ and $\sigma_{i,d}$ are computed across all stimuli that participant rated on that dimension. We recruit 30 participants per (model, task, method) triple, totaling N=720 total human reviewers. Effect sizes and significance (t-tests[3]) are computed on z-scored ratings for baseline vs. creative (CN-enhanced or CAA), where each participant is treated as an independent sample.

4.3 Results on the Alternative Uses Test and the Task Task

We report results in panels (b-f) of Figure 3. On the AUT, CN produces uniformly positive originality effects across all six models (avg. $d = +.36$), with four reaching significance. The effects are even stronger for surprise (avg. $d = +.43$), with five models reaching significance. On the Task Task, CN shows strong originality gains (avg. $d = +.40$), with the strongest

[2] Chu et al. (2024) studies additional dimensions, such as difficulty, how fun the task is to do, and how fun it is to watch, but here we restrict focus to creativity and originality, as these are most relevant to the goal of evaluating divergent thinking.

[3] We confirm responses follow a normal distribution before applying the t-tests.

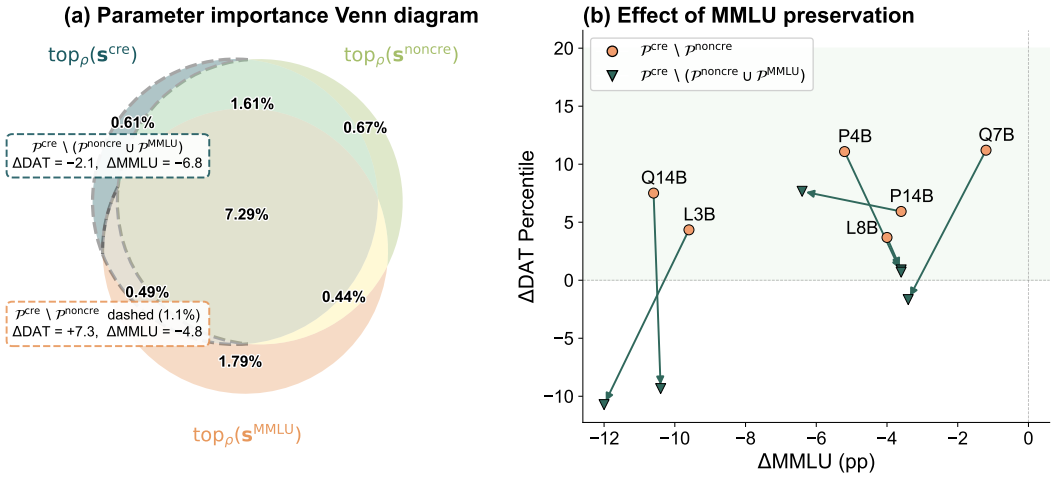

Figure 5: **Creativity–factual reasoning trade-off under attempted MMLU preservation. (a)** Venn diagram of top$_\rho$ parameter importance sets for creative, non-creative, and MMLU prompts at $\rho = 0.1$. **(b)** Arrows show the shift from the original CN mask to the augmented mask, which entails catastrophic reversal of DAT improvements without recovering MMLU accuracy.

effects in Phi-14B ($d = +.61$) and Qwen-14B ($d = +.56$), and moderate creativity improvements (avg. $d = +.24$). Although CN degrades AUT utility, this is a predictable consequence of the novelty–utility tradeoff already present in baseline responses. Among baseline AUT stimuli, originality and utility are negatively correlated ($r = -.49$, $p < 10^{-7}$), as are surprise and utility ($r = -.64$, $p < 10^{-13}$). As predicted by fundamental novelty–utility tradeoffs established in the creativity literature (Varshney, 2019), increases in subjective ratings of novelty tend to be accompanied by decreases in perceived utility. Therefore, CN's utility reduction is an expected byproduct of steering towards high-novelty responses, rather than an independent failure mode. To confirm this, we perform a hypervolume (HV) analysis (Cao et al., 2015) over all responses in the three-dimensional rating space (originality, surprise, utility) and find that the CN HV indicator is 15.2% larger than baseline ($p = .18$), suggesting CN obtains a moderate (but non-significant) gain in Pareto efficiency as well.

Additionally, while activation steering (93.9 avg) performs comparably to CN (94.1 avg) on the DAT (Figure 3), activation steering fails to generalize to the AUT and Task Task. These results are consistent with recent work that has found weight-space steering generalizes further out-of-distribution than activation steering on sycophancy and value alignment tasks (Fierro & Roger, 2025). Moreover, activation steering effectiveness has been shown to vary significantly by behavior type, with more complex behaviors like embodying persona archetypes and public figures proving more difficult to steer (Bas & Novak, 2025). Techniques such as context-dependent (Li et al., 2026; Lee et al., 2024) and learned activation steering (Rodriguez et al., 2025) have been proposed to remediate such issues, and may be explored in future work.

CN generalizes to open-ended creative tasks judged by human raters, demonstrably improving measures of originality and surprise, while activation steering does not exhibit reliable transfer.

5 Divergent Thinking Versus Factual Reasoning

Does improved divergent thinking come at the expense of general capabilities, such as factual reasoning? To test this, we evaluate CN-enhanced models[4] on the MMLU benchmark

[4]For each of the six models tested, we take the top-performing (ρ, α, prompt) configuration.

(Hendrycks et al., 2021) to quantify the impact on question-answering abilities in STEM, humanities, and the social sciences. Figure 5 reveals that CN-enhanced models reduce MMLU performance by 1–11%.

A natural strategy to protect factual-reasoning weights during mask construction is to incorporate MMLU questions in the negative contrast set in Algorithm 1, enforcing non-interference between tasks. Using this technique, Christ et al. (2025) demonstrate that mathematical reasoning can be improved without degrading MMLU performance. Does the same hold true for a *meta-cognitive* ability like divergent thinking? To test this, we sample a random set of 20 MMLU prompts $\mathcal{P}^{\text{MMLU}}$ and perform Algorithm 1 with $\mathcal{P}^{\text{cre}}$ and $\mathcal{P}^{\text{non-cre}} \cup \mathcal{P}^{\text{MMLU}}$. The masks then select for $\text{top}_\rho(\mathbf{s}^{\text{cre}}) \setminus \text{top}_\rho(\mathbf{s}^{\text{non-cre}} \cup \mathbf{s}^{\text{MMLU}})$, identifying weights above the $1-\rho$ percentile of importance for creative prompts that are not simultaneously above the $1-\rho$ percentile for MMLU or non-creative prompts. We recompute masks for all six models using each model's best hyperparameters, evaluate DAT score and 5-shot MMLU accuracy on 500 held-out questions, and report the results in Figure 5.

We find that **attempting to preserve MMLU performance catastrophically reverses DAT gains** and fails to improve MMLU scores, offering evidence that divergent thinking and factual reasoning rely on functionally entangled weights. Existing work has established that individual weights can be entangled in multiple distinct functions (a phenomenon known as *polysemanticity*), and that this supports the ability for neural models to represent more features than they have neurons (*superposition*) (Elhage et al., 2022; Sharkey et al., 2025). Kumar et al. (2025) attribute this behavior to *fractured entangled representations*, which can inhibit creative acts that intentionally "break regularities" by failing to preserve other regularities in a controlled manner. Creativity research more broadly suggests a need for separation between generation (≈DT) and selection (≈convergent thinking (CT)) (Varshney, 2019), with neuroscience studies pointing to distinct brain-activation patterns involved in associative (≈DT) and controlled (≈CT) processing modes (Zhang et al., 2020; Volle, 2018). The success of multi-agent creative systems (Lin et al., 2025) and multi-stage prompting techniques that decouple creative exploration from constraint satisfaction (Nguyen & Singla, 2025) can potentially be interpreted within the context of these results: if DT and CT rely on functionally entangled weights, simultaneously eliciting strong DT and CT abilities may be challenging. Therefore, separating these steps—whether across agents or prompts—can provide stronger overall performance.

> We find evidence that divergent thinking and factual reasoning are functionally entangled in model weights.

6 Conclusion

In this work, we presented CreativityNeuro (CN), a data-free method for steering language model weights to improve divergent thinking. CN-enhanced models improve measures of divergent thinking on the DAT, AUT, and the Task Task, outperforming baselines such as prompting, activation steering, and varying decoding strategies.

Limitations and Future Work. Several limitations of our study are worth acknowledging: our evaluation is restricted to finite set of divergent thinking (DT) benchmarks and metrics (DAT, AUT, Task Task), which capture only certain aspects of creativity. As Runco (2008) notes, DT is not synonymous with creativity and should best be thought of as a measure of creative potential. Moreover, our comparison to activation steering focuses on a standard CAA-style method and does not exhaust the space of possible activation-space interventions, such as context-dependent or learned approaches. Lastly, we found evidence suggesting DT and factual reasoning are functionally entangled model weights—understanding whether this entanglement reflects a fundamental architectural constraint or arises as an artifact of the Wanda-style importance technique remains an important open question. Developing techniques that can enhance DT without degrading factual reasoning—for instance, through training-time objectives that encourage representational disentanglement—would be a valuable direction for future work.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Tetiana Bas and Krystian Novak. What can we actually steer? a multi-behavior study of activation control. *arXiv preprint arXiv:2511.18284*, 2025.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent Creativity in Humans and LLMs. Technical report, 2024.

Margaret A Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.

Yongtao Cao, Byran J. Smucker, and Timothy J. Robinson. On using the hypervolume indicator to compare pareto fronts: Applications to multi-criteria optimal experimental design. *Journal of Statistical Planning and Inference*, 160:60–74, 2015. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2014.12.004. URL https://www.sciencedirect.com/science/article/pii/S0378375814002006.

Bryan R. Christ, Zack Gottesman, Jonathan Kropko, and Thomas Hartvigsen. Math Neurosurgery: Isolating Language Models' Math Reasoning Abilities Using Only Forward Passes. 6 2025. URL http://arxiv.org/abs/2410.16930.

Junyi Chu, Jennifer Hu, and Tomer D Ullman. The task task: Creative problem generation in humans and language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Arne Dietrich. Types of creativity. *Psychonomic Bulletin and Review*, 26(1):1–12, 2 2019. ISSN 15315320. doi: 10.3758/s13423-018-1517-7.

Dietrich, Arne. The Cognitive Neuroscience of Creativity. *Psychonomic Bulletin & Review*, 11 (6):1011–1026, 2004.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Chris Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Gilles Fauconnier and Mark Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2008.

Constanza Fierro and Fabien Roger. Steering language models with weight arithmetic. *arXiv preprint arXiv:2511.05408*, 2025.

Francis Galton. *Hereditary genius: An inquiry into its laws and consequences*. D. Appleton & Company, 1870.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer,

Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

J P Guilford. Psychological Bulletin THE STRUCTURE OF INTELLECT. Technical Report 4, 1956.

Jacques Hadamard. *An Essay on the Psychology of Invention in the Mathematical Field*. Courier Corporation, 1954.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. 11 2021. URL http://arxiv.org/abs/2103.03874.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954*, 2025.

Arthur Koestler. *The Act of Creation*. Macmillan, 1964.

Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2024.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models. *arXiv preprint arXiv:2602.01654*, 2026.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung-yi Lee, and Yun-Nung Chen. Creativity in llm-based multi-agent systems: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27572–27595, 2025.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Technical report, Anthropic, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Mary Lou Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, pp. 22–28, 2010.

Sarnoff Mednick. The associative basis of the creative process. *Psychological Review*, 69(3): 220, 1962.

Robert Morain and Dan Ventura. Is Prompt Engineering the Creativity Knob for Large Language Models? In *Proceedings of the 16th International Conference on Computational Creativity (ICCC'25)*, 2025.

Manh Hung Nguyen and Adish Singla. Divergent-convergent thinking in large language models for creative problem generation. *arXiv preprint arXiv:2512.23601*, 2025.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. 4 2021. doi: 10.1073/pnas.2022340118/-/DCSupplemental.y.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Shao-yen Tseng, and Vasudev Lal. Steering Large Language Models to Evaluate and Amplify Creativity. 12 2024. URL http://arxiv.org/abs/2412.06060.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models? arXiv:2405.00492, 2024.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1162. URL https://nlp.stanford.edu/pubs/glove.pdf.

Lambert AJ Quetelet. A treatise on man and the development of his faculties (a facsimile reproduction of the english translation of 1842 with an introduction by solomon diamond). 1842.

Pau Rodriguez, Michal Klein, Eleonora Gualdoni, Valentino Maiorca, Arno Blaas, Luca Zappella, Marco Cuturi, and Xavier Suau. Lineas: End-to-end learning of activation steering with a distributional loss. *NeurIPS*, 2025.

A Rothenberg. *Flight from Wonder: An Investigation of Scientific Creativity*. Oxford University Press, 2014.

Mark A. Runco. Commentary: Divergent Thinking Is Not Synonymous With Creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 2(2):93–96, 5 2008. ISSN 19313896. doi: 10.1037/1931-3896.2.2.93.

Aishik Sanyal, Samuel Schapiro, Sumuk Shashidhar, Royce Moon, Lav R Varshney, and Dilek Hakkani-Tur. Spark: A system for scientifically creative idea generation. *arXiv preprint arXiv:2504.20090*, 2025.

Samuel Schapiro, Jonah Black, and Lav R Varshney. Transformational Creativity in Science: A Graphical Theory. *arXiv preprint arXiv:2504.18687*, 2025.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. 9 2024. URL http://arxiv.org/abs/2409.04109.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. 6 2025. URL http://arxiv.org/abs/2506.20803.

Dean Keith Simonton. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, 2004.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han Van Der Maas. Putting GPT-3's Creativity to the (Alternative Uses) Test. In *International Conference on Computational Creativity*, 2022. URL http://osf.io/vmk3c/.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

L. R. Varshney. Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development*, 63(1), 1 2019. ISSN 21518556. doi: 10.1147/JRD.2019.2893907.

Emmanuelle Volle. Associative and controlled cognition in divergent thinking: Theoretical, experimental, neuroimaging evidence, and new directions. *The Cambridge Handbook of the Neuroscience of Creativity*, pp. 333, 2018.

Haining Wang, Peng Bao, Luning Qiu, Dawei Wu, Nanyun Yu, Haoran Liu, and Samuel Johnson. A large-scale comparison of divergent creativity in humans and large language models. *Nature Human Behaviour*, 2025. doi: 10.1038/s41562-025-02331-1.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. 6 2024. doi: 10.18653/v1/2024.acl-long.18. URL http://arxiv.org/abs/2305.14259http://dx.doi.org/10.18653/v1/2024.acl-long.18.

Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. Igniting creative writing in small language models: Llm-as-a-judge versus multi-agent refined rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17171–17197, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Weitao Zhang, Zsuzsika Sjoerds, and Bernhard Hommel. Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage*, 210: 116572, 2020.

A Full Sets of Contrastive Prompts

Each of the six prompt sets contains 10 creative and 10 non-creative exemplars. Below we show 3 representative examples from each set.

Table 2: **Full contrastive prompt sets (3 examples each).** Each prompt set contains 10 creative ($\mathcal{P}^{\text{cre}}$) and 10 non-creative ($\mathcal{P}^{\text{non-cre}}$) exemplars used for parameter importance scoring in Algorithm 1.

Set	**Creative** ($\mathcal{P}^{\text{cre}}$)	**Non-Creative** ($\mathcal{P}^{\text{non-cre}}$)
DAT	List 10 common English nouns that are as unrelated in meaning as possible. Avoid any shared topic or category. Output only the nouns, separated by commas.	List 10 common English nouns that are as closely related in meaning as possible and clearly fit into a single narrow topic. Output only the nouns, separated by commas.
	Give me 10 one-word English nouns that are extremely far apart in semantic meaning. They should not fit into a single theme.	Give me 10 one-word English nouns that are very strongly associated with each other and belong to the same specific domain.
	Produce 10 everyday English nouns that share no obvious connection with one another. Each noun should come from a very different domain.	Produce 10 English nouns that are tightly connected around a single clear theme (for example, all parts of a computer or all items in a kitchen).
STORY	Write an unusual opening sentence for a short story that subverts reader expectations.	Write a standard opening sentence for a fairy tale.
	Create a story beginning that combines two unrelated concepts in a surprising way.	Create a typical story beginning that establishes setting and character clearly.
	Write the first line of a story that makes the reader question reality.	Write the first line of a conventional mystery novel.
IDEATION	List 5 unusual uses for a brick that nobody has thought of before.	List 5 common uses for a brick in construction.
	Generate unconventional solutions to reduce traffic in cities.	Generate standard solutions to reduce traffic in cities.
	What are some surprising ways a library could be repurposed?	What are the traditional functions of a library?
PROBLEM	How might you solve this problem in a way that seems counterintuitive at first?	What is the most efficient way to solve this problem?
	What would an alien civilization's approach to this problem look like?	List the standard steps for addressing this type of issue.
	If you had to solve this with resources from a different era, what would you do?	What best practices should be followed when solving this?
OPEN	Generate a joke about electric vehicles.	Explain nuclear fission like I am five years old.
	Create the first verse of a wedding vow.	What is Bukhara? Provide a paragraph-long explanation in layman's language.
	Write a song about a guy named Jacob working at a call center making jokes.	In a few sentences explain what threats do scams pose to individuals?
MINIMAL	Invent something.	State a fact.
	Surprise me.	Be accurate.
	Be weird.	Be precise.

B Baseline Sweep Settings

In Section 3, we compare CreativityNeuro against a set of baseline techniques.

1. **Prompting:** We present creative exemplars from each of six prompt sets (Table 2) as in-context guides
2. **Decoding Parameter Sweeps**
   (a) **top-p** nucleus sampling, $p \in \{0.8, 0.85, 0.9, 0.95, 1.0\}$
   (b) **top-k** sampling, $k \in \{10, 25, 50, 100, \text{disabled}\}$
   (c) **repetition penalty**, $\theta \in \{1.0, 1.1, 1.2, 1.5, 2.0, 3.0\}$
3. **Activation Steering**: We sweep the injection layer suffix from single-layer to the final 50% of layers and find that injecting into the final 30% works best. We sweep $\alpha \in$

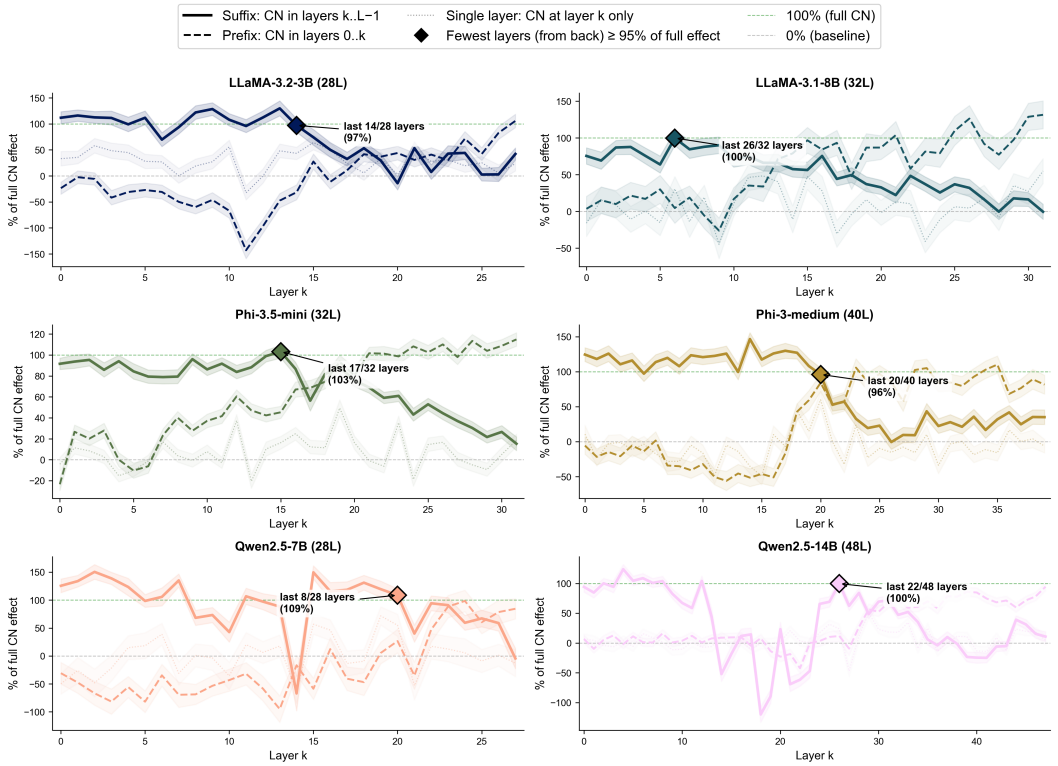

Figure 6: **Layerwise ablation: suffix vs. prefix vs. single-layer.** Each panel shows the % of full CN DAT effect recovered as a function of the number of layers k with CN weights applied. Solid lines: suffix (last k layers). Dashed lines: prefix (first k layers). Dotted lines: single-layer (one layer at a time, plotted by layer index). Diamond markers indicate the fewest layers from the back achieving $\geq$95% of the full effect.

$\{0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 2.0, 4.0\}$, selecting the best α where all three temperatures yield ≥ 120 valid samples.

C Layerwise Ablation Studies

C.1 Suffix vs. Prefix

To localize the CN effect across the network, we compare three layerwise interventions (Figure 6): (i) *suffix*, applying CN weights to only the last k layers; (ii) *prefix*, applying CN weights to only the first k layers; and (iii) *single-layer*, applying CN weights to one layer at a time. For each condition, we generate $N = 120$ valid DAT samples at $T = 1.0$ and report the percentage of the full CN effect recovered in terms of DAT scores.

Suffix. Applying CN weights to a suffix of layers (layers k through $L-1$) recovers the full effect with roughly half the network. On average, 51% of layers (from the back) are needed to reach 100% recovery, ranging from 29% for Qwen-7B (last 8/28) to 81% for Llama-8B (last 26/32). The remaining models fall in between: Llama-3B (last 14/28, 50%), Phi-14B (last 20/40, 50%), Qwen-14B (last 22/48, 46%), and Phi-4B (last 17/32, 53%). However, this analysis is only conducted on DAT scores, and it is unclear whether the last 51% of layers are sufficient to recover 100% of the scores on the AUT and Task Task.

Prefix. Applying CN weights from the front (layers 0 through k) is far less efficient. On average, 78% of all layers must be included before the prefix condition reaches 95% recovery. For Qwen-14B (48 layers), the prefix condition never reaches 95% even when all layers

are included (94% at $k = 48$). This asymmetry provides evidence that the CN effect is concentrated in later layers of the residual stream.

Single-layer. No single layer is sufficient to recover the full CN effect. The best individual layers recover 50–72% of the effect (e.g., Q7B layer 19: 72%, L3B layer 15: 64%, P14B layer 20: 60%), with top contributors generally appearing in middle-to-late layers. The gap between the best single layer and the suffix provides evidence that the CN effect requires cooperation across multiple late layers, consistent with findings that representations of human-interpretable concepts or behaviors may span multiple layers (Lindsey et al., 2024; Sharkey et al., 2025).

D Hyperparameter Sweeps

CreativityNeuro introduces two hyperparameters: the *importance threshold* $\rho \in (0, 1]$, which controls the fraction of parameters selected by the importance mask, and the *scaling factor* $\alpha > 0$, which controls the magnitude of the weight perturbation applied to masked parameters. To identify effective (ρ, α) configurations for each model, we conduct systematic grid sweeps evaluated on the DAT. For each model, we generate CN weight masks using six different prompt sets: DAT, STORY, IDEATION, PROBLEM, OPENENDED, and MINIMAL. Each prompt set produces a separate importance mask. We then evaluate every combination of keep ratio $\rho \in \{0.01, 0.05, 0.1, 0.2\}$ and scaling factor $\alpha \in \{0.1, 0.5, 1.0, 2.0\}$ at three sampling temperatures $T \in \{0.9, 1.0, 1.2\}$, with top-$k = 0$ and top-$p = 1.0$ (i.e., untruncated sampling). For LLaMA-3.1-8B and Phi-3.5-mini, after initial experiments revealed that $\alpha > 0.5$ were too high, we additionally tested a finer alpha grid $\alpha \in \{0.01, 0.05, 0.075, 0.1, 0.5, 1.0\}$ to probe the conservative regime identified by Christ et al. (2025). Each configuration generates $n = 120$ valid DAT samples (10-word lists with all words present in the GloVe vocabulary).

For each (ρ, α, T) triple, we compute the mean DAT scores for the CN-enhanced ($\overline{\text{DAT}}_{\text{CN}}$) and baseline ($\overline{\text{DAT}}_{\text{base}}$) models, then convert each to a human percentile using the distribution from Wang et al. (2025). Figures 7 to 12 show $\Delta\,\text{Percentile} = P(\overline{\text{DAT}}_{\text{CN}}) - P(\overline{\text{DAT}}_{\text{base}})$ (averaged across temperatures) as a function of (α, ρ) for each of the six prompt sets, across all six models. Positive values indicate that CN moves the model's DAT performance upward in the human score distribution.

E Evaluation Prompts

In this section, we share the full prompts used to generate and/or post-process the stimuli for the DAT, AUT, and TT.

E.1 Divergent Association Task (DAT)

The DAT prompt instructs models to generate 10 maximally dissimilar nouns:

DAT Prompt

```
You will be asked to name exactly 10 English nouns.
Output exactly 10 words, separated by commas, and nothing else.
Return ONLY: word1, word2, ..., word10
DO NOT write explanations. DO NOT think step by step.

Be as original and unusual as possible.  Avoid common or closely related
words.
Now name 10 English nouns that are as different from each other as possible.
```

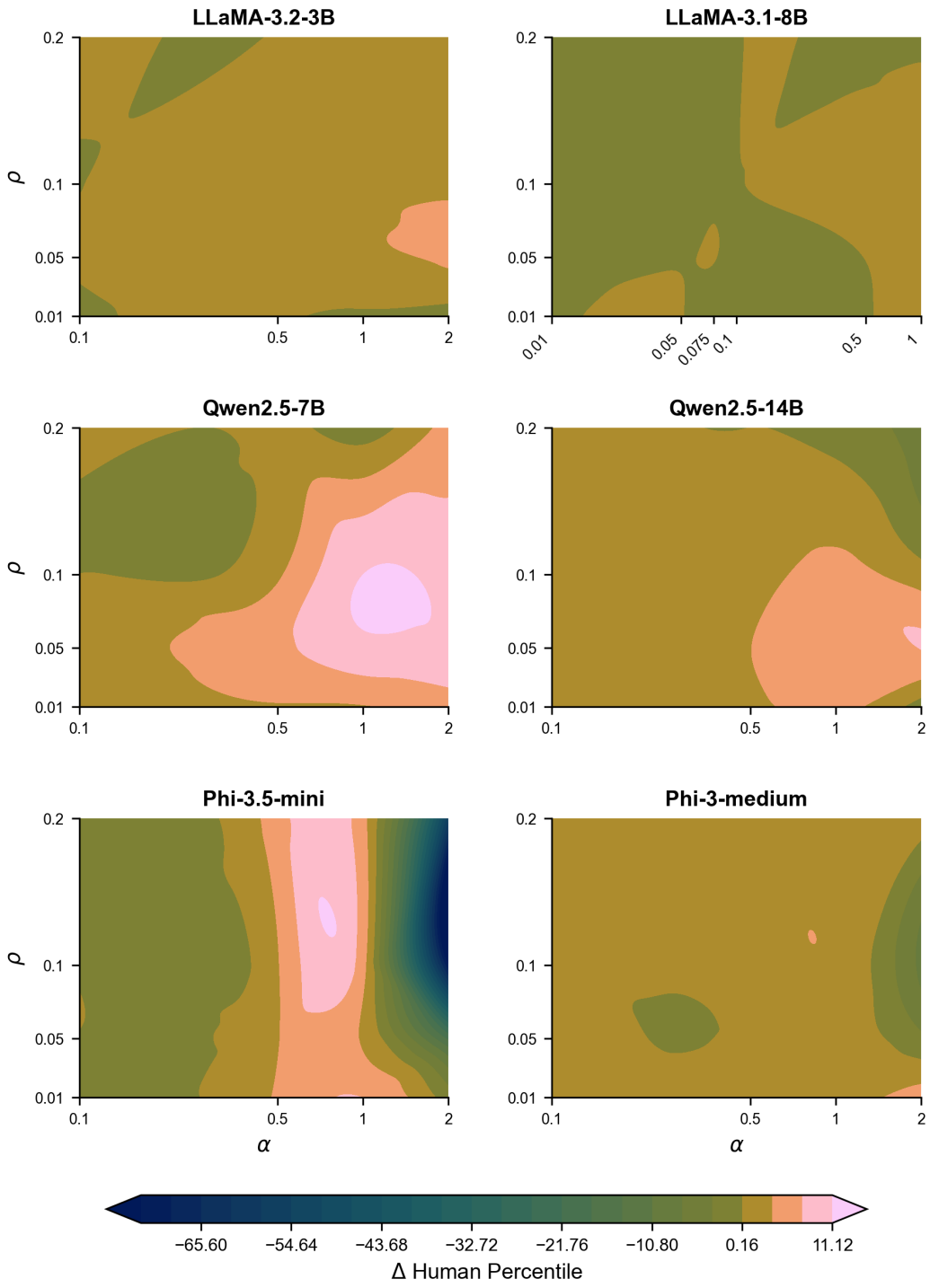

Figure 7: **Sensitivity to (α, ρ) — DAT prompt set.** Δ Percentile averaged across three temperatures ($T \in \{0.9, 1.0, 1.2\}$) for each model. Color scale is shared across panels and centered at zero.

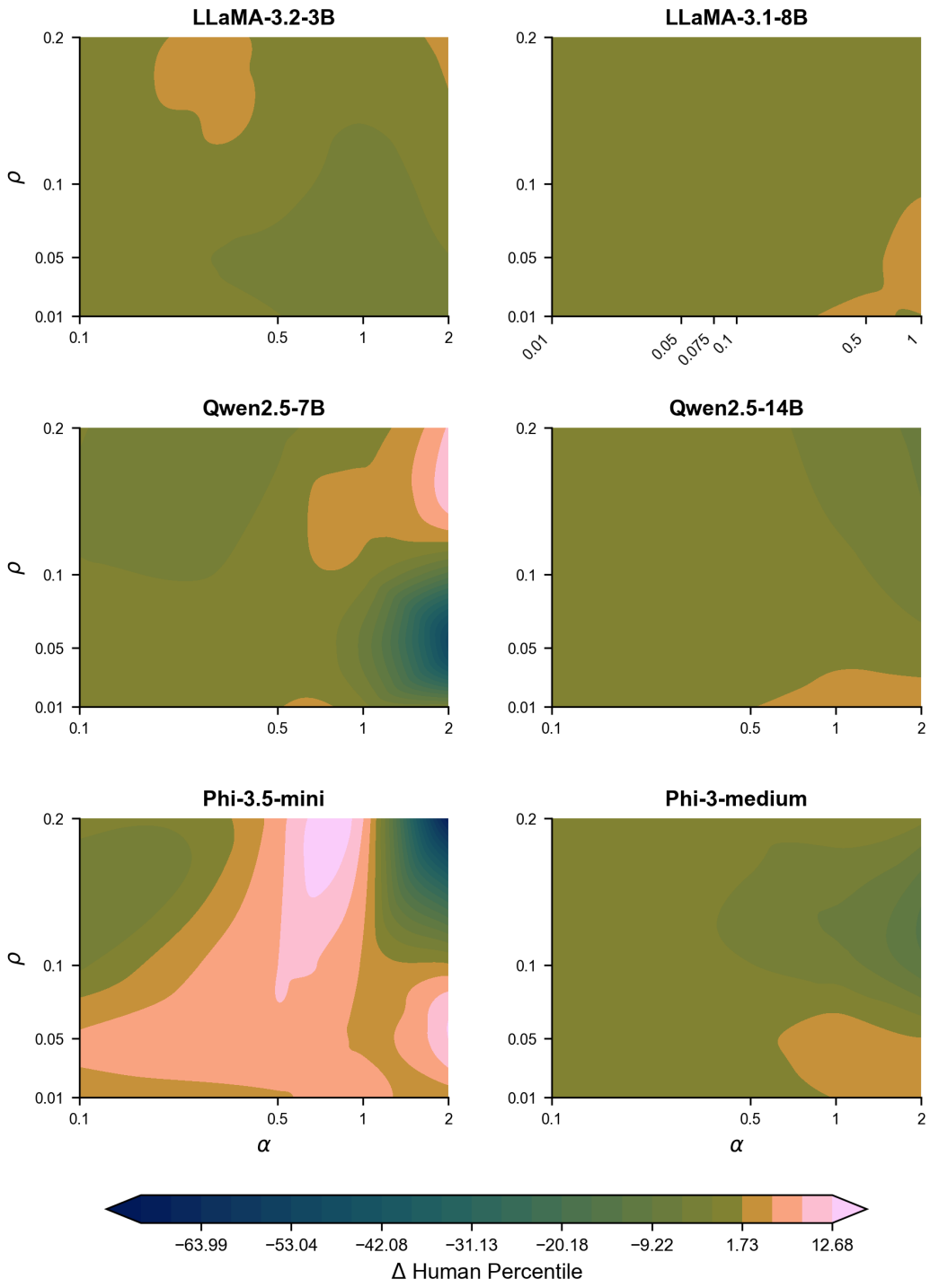

Figure 8: **Sensitivity to (α, ρ) — Story prompt set.** Mean ΔDAT averaged across three temperatures ($T \in \{0.9, 1.0, 1.2\}$) for each model. Color scale is shared across panels and centered at zero.

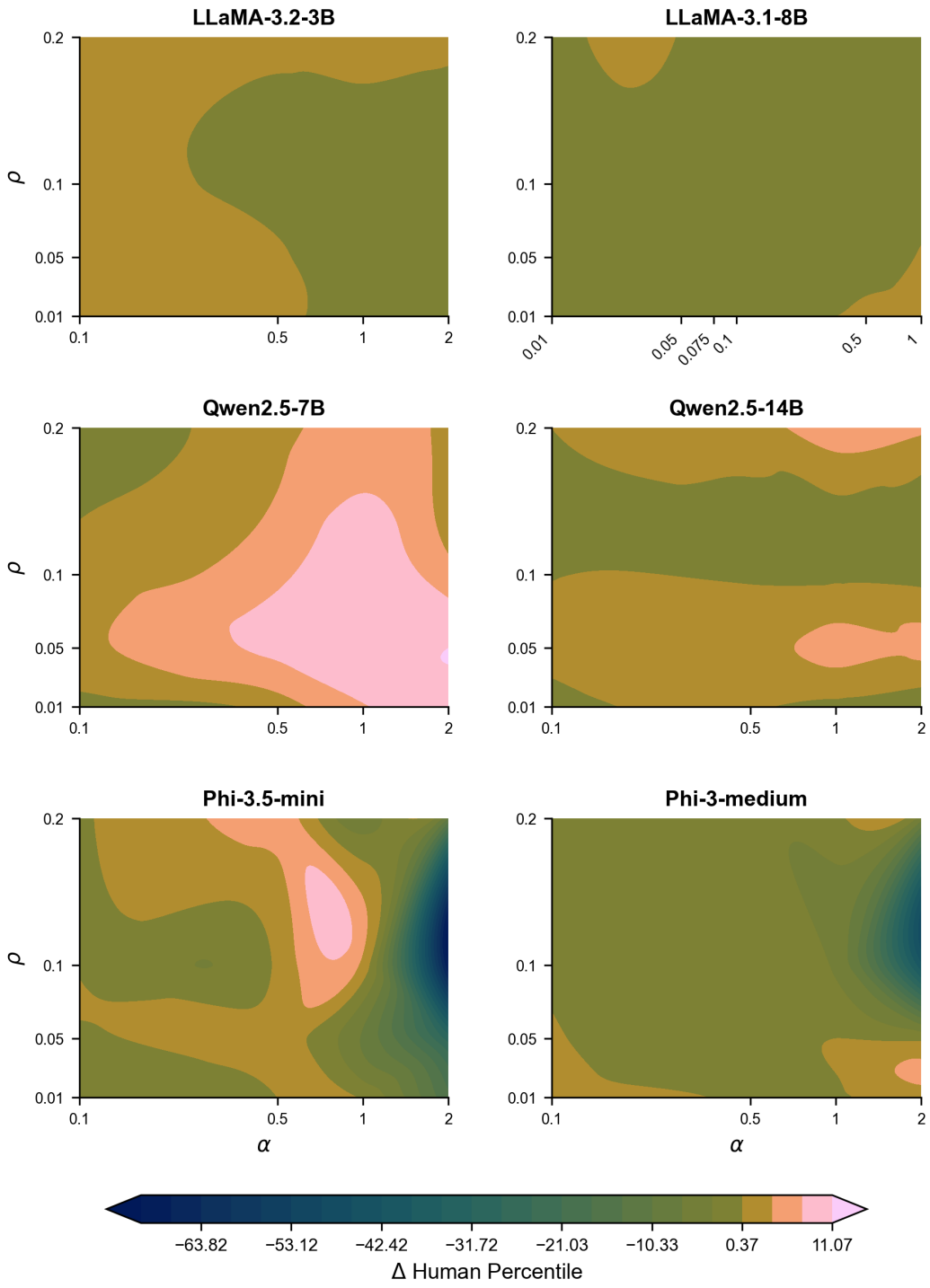

Figure 9: **Sensitivity to (α, ρ) — Ideation prompt set.** Mean ΔDAT averaged across three temperatures ($T \in \{0.9, 1.0, 1.2\}$) for each model. Color scale is shared across panels and centered at zero.

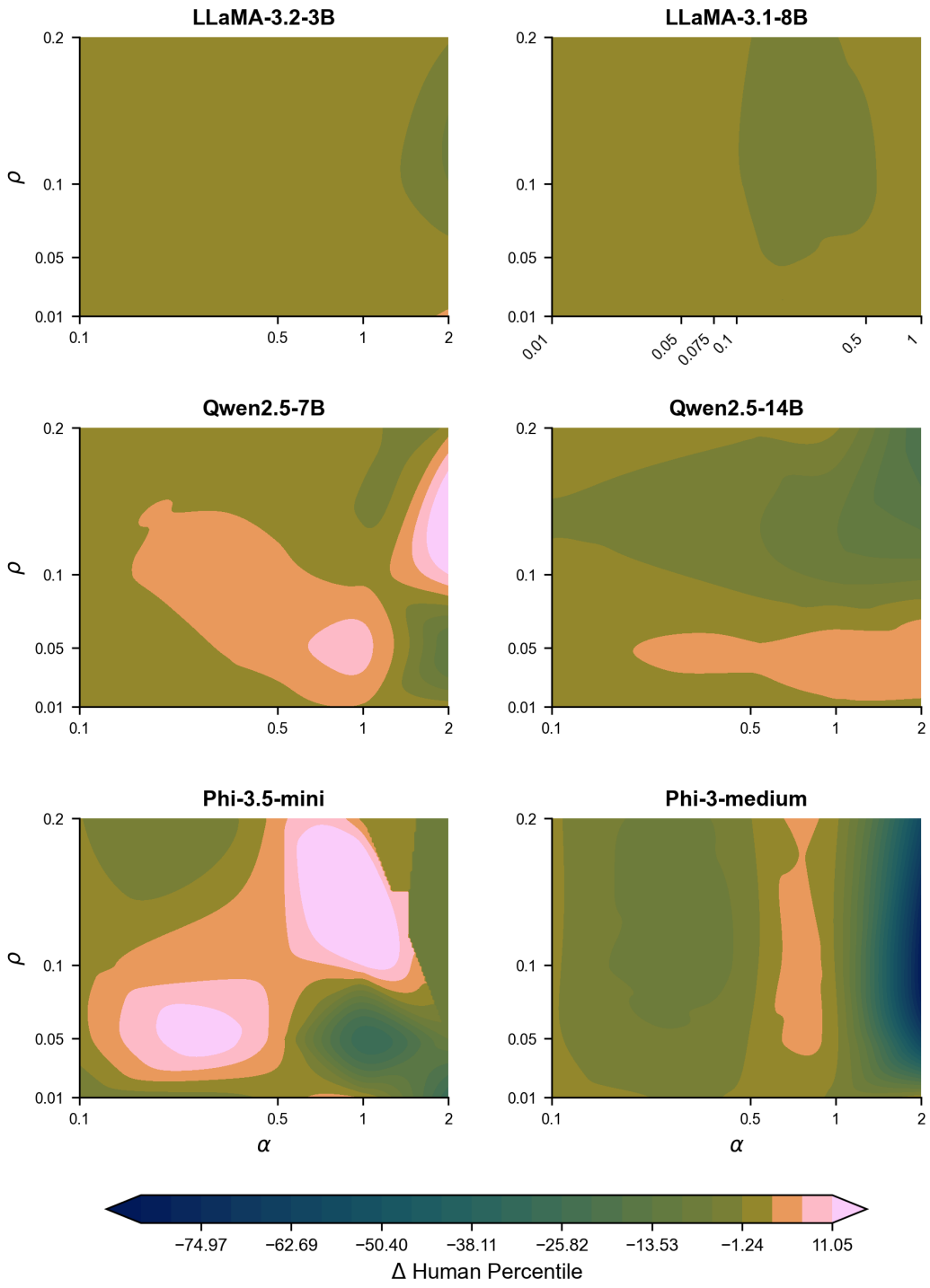

Figure 10: **Sensitivity to (α, ρ) — Problem prompt set.** Mean ΔDAT averaged across three temperatures ($T \in \{0.9, 1.0, 1.2\}$) for each model. Color scale is shared across panels and centered at zero.

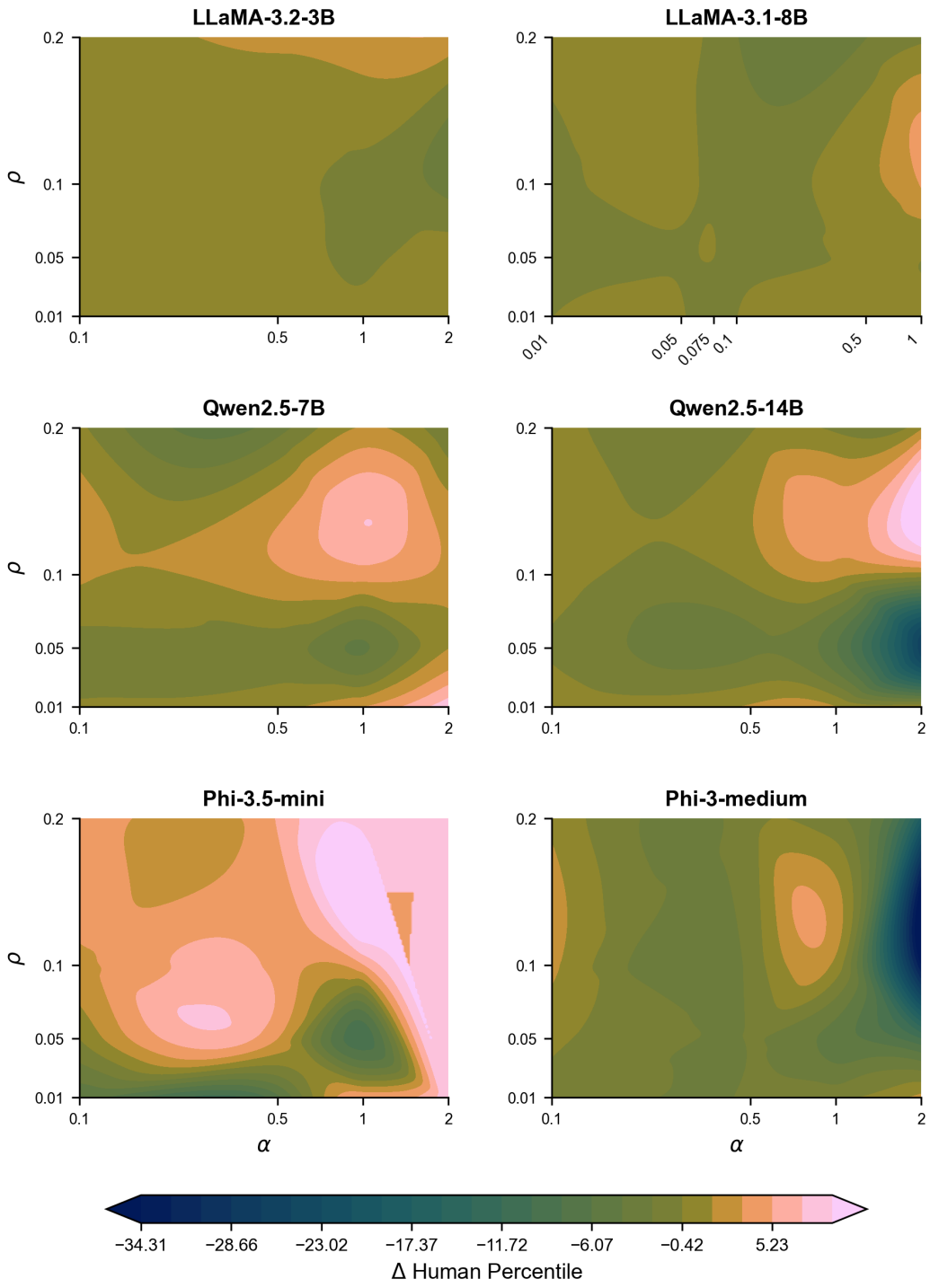

Figure 11: **Sensitivity to (α, ρ) — Open-ended prompt set.** Mean ΔDAT averaged across three temperatures ($T \in \{0.9, 1.0, 1.2\}$) for each model. Color scale is shared across panels and centered at zero.

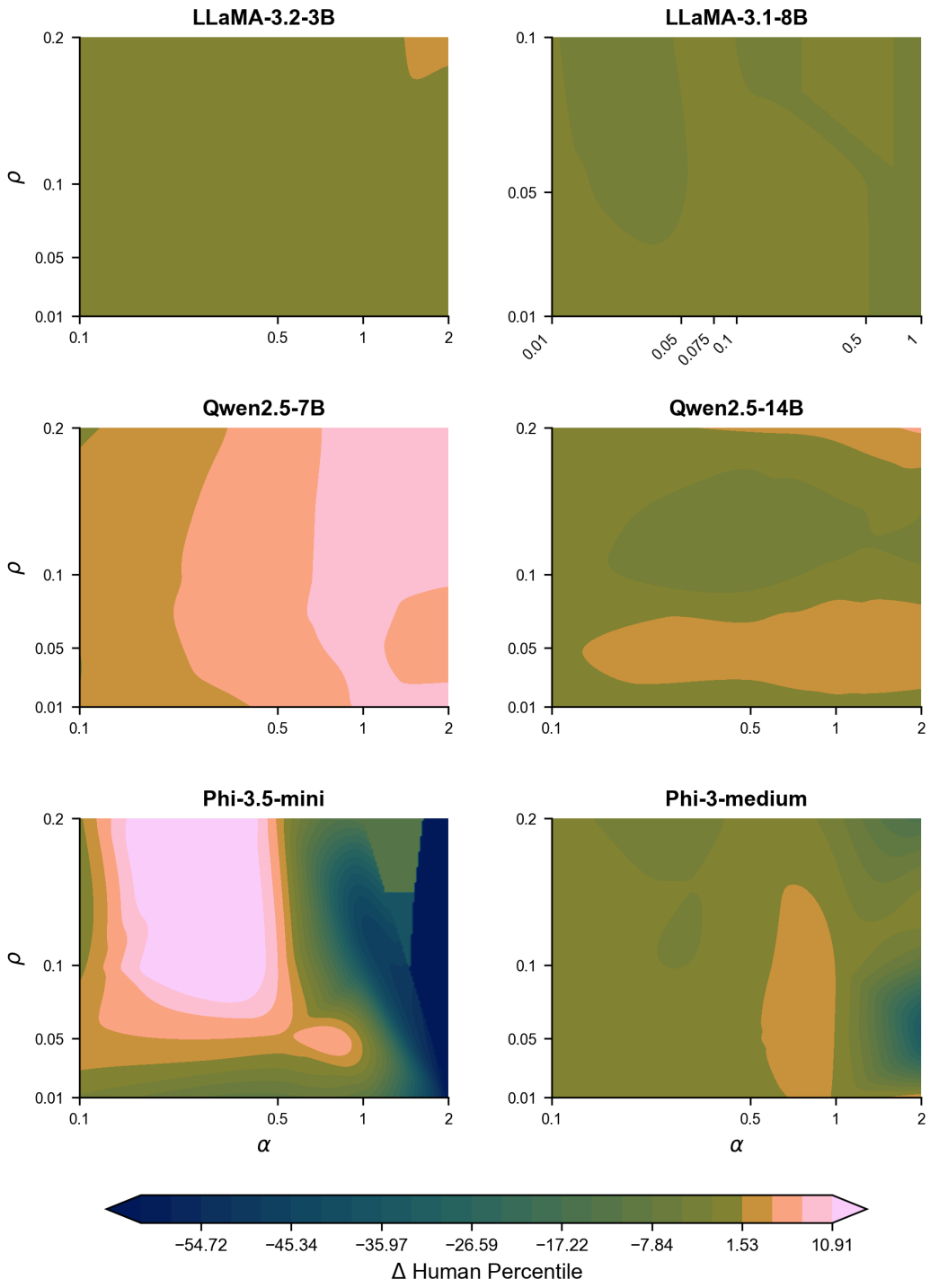

Figure 12: **Sensitivity to (α, ρ) — Minimal prompt set.** Mean ΔDAT averaged across three temperatures ($T \in \{0.9, 1.0, 1.2\}$) for each model. Color scale is shared across panels and centered at zero.

E.2 Alternative Uses Test (AUT)

The AUT uses a system prompt and a user prompt, with {object} replaced by one of 3 standard objects: *brick*, *paperclip*, *fork*.

AUT System Prompt

```
You are participating in a creativity test. Your task is to generate creative, unusual,
and original uses for common objects. Be imaginative and think outside the box.
```

AUT User Prompt

```
List 5 creative and unusual alternative uses for a {object}.
Be specific and creative. List each use on a new line, numbered 1 through 5.
Only list the uses, no explanations.
```

E.3 Task Task (TT)

The Task Task uses a system prompt and a user prompt with three in-context examples from Chu et al. (2024).

TT System Prompt

```
You are a creative game show designer. Your task is to invent fun, original, and
entertaining challenges that would be exciting for contestants to attempt and audiences
to watch.
```

TT User Prompt

```
You've been recruited to help design challenge tasks for a new game show! Your job
is to come up with a new creative, silly, and fun task for humans to solve.

Here are a few example tasks:

1. Your goal is to: Throw a teabag into a mug from the farthest distance.
You can use: Anything you can reasonably expect to find in a house, garage, and
garden shed.

2. Someone has squeezed all of the toothpaste out of the toothpaste tube.
Your goal is to: Get as much of the original toothpaste back into the empty
toothpaste tube as possible. You can use: Anything you can reasonably expect to find
in a bathroom.

3. Your goal is to: Transfer water between two fishbowls using only the
supplied items. You cannot move the fishbowls. You can use: a chocolate bar, a
rubber glove, a baguette, a snorkel, a cardboard tube, and a plate of pasta.

Now it's your turn! Create your own creative, silly, and fun task for future
participants to solve. Specify the goal, scoring criteria, and any materials or
constraints. Describe it in exactly 3--4 sentences as a short paragraph --- do not
use lists or bullet points. Do not comment on how entertaining, creative, or fun the
task would be.
```

E.4 Task Task Post-Processing

Raw Task Task generations are post-processed using GPT-4o to normalize formatting before human evaluation.

TT Post-Processing System Prompt

```
You are an editor cleaning game show challenge descriptions for a human evaluation
study.

Make these MINIMAL edits:

1. REMOVE any task title at the start (e.g., 'In "Baking Bonanza,"'). After
removing, capitalize the first word of the remaining text.

2. Convert ALL third-person references to SECOND PERSON (e.g., ''contestants
must'' → ''you must'', ''the player'' → ''you'').

3. REMOVE any metacommentary sentences (e.g., ''This creative juggling act
combines laughs with a dash of physics comedy.'').

CRITICAL RULES:
- Do NOT summarize, condense, or shorten the task description
- Do NOT change the creative content or game mechanics
- Do NOT add any text --- only edit or remove
- Output ONLY the cleaned description with no preamble or explanation
```

TT Post-Processing User Prompt

```
Clean this game show challenge description: {text}
```

F Disclosure of Large Language Model Usage

In this paper, large language models (LLMs) were used to assist in the code implementation, plotting and figure generation, reporting (but not analysis) of experiment results, and for comprehensive surveys of related work.