Does Forcing Structured Output Degrade LLM Creativity?

Anonymous ACL submission

Abstract

The operational need for structured data from Large Language Models (LLMs) is in direct conflict with the cognitive processes that foster creativity. While formats like JSON are essential for downstream applications, this paper investigates the critical, unquantified cost of such constraints on creative performance. We conducted a large-scale analysis across multiple creative tasks, comparing the creativity of LLM-generated responses in a freeform text baseline against six structured formats. Our results reveal that forcing structured output degrades creativity—on average by over 17% when models must infer a JSON structure, and by up to 26% in the most severe cases. We deconstruct this degradation into a dominant "creative constraint" effect, where the cognitive load of simultaneous creation and formatting harms ideation, and a weaker, opposing trend of "format bias," where LLM judges slightly prefer well-structured output. The former effect outweighs the latter. Consequently, we propose and validate a "generate-then-structure" workflow as a practical solution that mitigates this degradation, improving both the substance and perceived quality of creative work.

1 Introduction

Large Language Models (LLMs) are powerful creative tools used for tasks from marketing copy to product brainstorming, demanding both creativity and utility (Chakrabarty et al., 2024). A key aspect of this utility is producing structured output (e.g., JSON) for downstream systems (Wu et al., 2023).

This requirement, however, introduces a fundamental tension. Creative ideation is a divergent, free-associative process (Sowden et al., 2015), while adhering to a rigid data schema is a convergent, logical task imposing significant cognitive load (Sweller, 1988). Recent work has begun to explore the impact of format restrictions on LLM performance (Tam et al., 2024; Castillo), and studies have shown that LLMs can exhibit bias towards certain output formats (Long et al., 2025). This raises a critical, underexplored question: Does forcing an LLM to produce structured output degrade its creativity?

This paper provides an answer to that question. We hypothesize that the cognitive load of adhering to a strict format while simultaneously generating creative ideas—what we term **creative constraint**—negatively impacts output quality. This degradation is a concern for applications requiring both creative problem-solving and structured data. To overcome this, we propose and validate a generate-then-structure workflow, showing that separating ideation from formatting recovers this lost creativity.

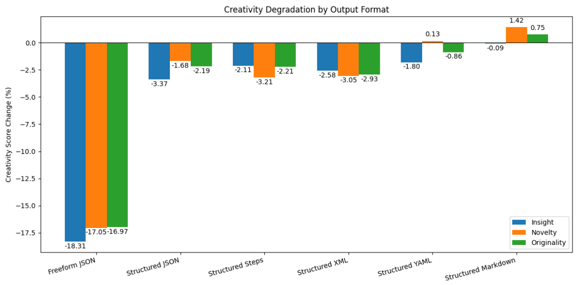
Our key contributions are:

1. **Quantifying a Creativity Tax:** We quantify a creativity tax from forcing structured output, which degrades LLM performance by up to 26%, revealing a flaw in many AI system designs.

2. **Isolating Competing Mechanisms:** We deconstruct this degradation into a dominant negative creative constraint on the generator and a subtle positive format bias from the evaluator, providing a new analytical framework.

3. **A Validated Architectural Pattern:** We propose and validate a generate-then-structure pipeline that recovers this creative loss, offering an evidence-based design pattern for creative systems.

Our findings offer clear guidance for designing more effective AI-powered creative systems, ensuring that the need for structured data does not compromise innovation.



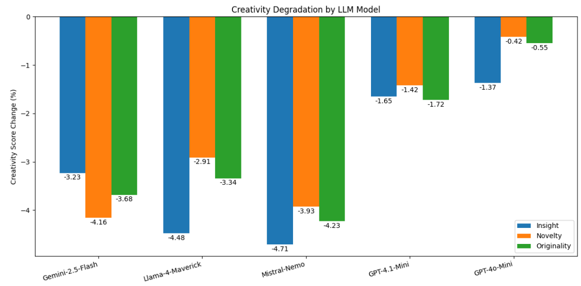


Figure 1: Creativity degradation effects: (a) by output format showing percentage change in Insight, Novelty, and Originality scores relative to freeform baseline, with Freeform JSON showing the most severe degradation (-18.31% insight, -17.05% novelty, -16.97% originality) and Structured Markdown performing closest to baseline; (b) by model architecture comparing creativity degradation across five LLM models, with GPT models showing significantly less degradation than other architectures.

2 Related Work

The use of LLMs for evaluation is a rapidly growing field. Models like Prometheus (Kim et al., 2023) have shown that LLMs can achieve high agreement with human experts on a variety of tasks. However, the potential for bias in these LLM judges is a known issue. Prior work has explored issues like position bias, verbosity bias, and sycophancy bias (Wang et al., 2023). Our work extends this line of inquiry by explicitly positioning "format bias" as a new dimension in the critical field of reliable LLM assessment.

In the domain of creative generation, many studies have explored how prompting techniques can influence output quality. Techniques like Chain-of-Thought (Wei et al., 2022) and Tree-of-Thought (Yao et al., 2023) improve logical reasoning but do not explicitly address the impact of output format constraints on creative tasks. Recent work has examined LLM creativity more broadly (Franceschelli and Musolesi, 2023; Zhao et al., 2024), including studies on divergent thinking in humans versus LLMs (Bellemare-Pepin et al., 2025) and methods to encourage divergent thinking through multi-agent debate (Liang et al., 2024). Our work contrasts with the "creative constraint" literature, where some constraints enhance human creativity, by showing that not all constraints are equal; the rigid, syntactic constraints of data formats appear to be detrimental, unlike more abstract, semantic constraints.

Our findings also have direct implications for the design of multi-agent systems like AutoGen (Wu et al., 2023) and ChatDev (Qian et al., 2023), which often rely on structured communication protocols. Recent work has developed frameworks for structuring LLM outputs (Yang et al., 2025; Wang et al., 2025), but the analysis shows that the format of the communication protocol between agents is not a neutral engineering choice; it can directly impact the creative capacity of the system. Our work provides empirical evidence for separating "ideation" agents from "structuring" or "execution" agents, a design principle that is often followed intuitively but now has quantitative backing. We bridge the gap between the generation and evaluation literature by showing how a single factor—output format—can simultaneously impact both processes.

3 Methodology

Our experiment isolates the impact of format constraints from content quality across three stages: parallel generation, faithful conversion, and comparative evaluation. For generation, we utilized state-of-the-art LLMs including GPT-4.1-Mini, GPT-4o-Mini, Gemini-2.5-Flash, Llama-4-Maverick, and Mistral-Nemo. For evaluation, we employed a single judge model with multiple evaluations to ensure robust scoring (GPT-4.1). All generations were performed with a temperature of 0.7 to encourage creative yet coherent responses. Judge temperature is set to 0.2 to ensure consistency in scoring.

We selected three creativity metrics—Insight, Originality, and Novelty—as they align with established creativity frameworks, capturing the depth (Insight), statistical rarity (Originality), and surprisingness (Novelty) of ideas. These metrics provide a comprehensive assessment of creative quality across different dimensions of the creative process. While recent work has developed specialized benchmarks for creative writing (Fein et al., 2025) and creative thinking assessments (Mishra et al., 2023), our focus on format-specific creativity degradation requires a more general evaluation approach.

3.1 Stage 1: Parallel Generation

We tasked LLM agents with solving problems from four creative datasets: AUT, MacGyver, LiveIdea, and LiveIdea-Div (Sun et al., 2024; Tian et al., 2023; Ruan et al., 2024). For each problem, we generated solutions under several conditions by appending specific modifiers to base instructions. We tested six different output formats: freeform text (baseline), structured JSON, structured steps, XML, YAML, and Markdown. We also tested two distinct JSON conditions: one with explicit formatting instructions and another where the model was forced to return JSON via API parameters without explicit guidance. This parallel generation allows us to compare the agent's creative performance when constrained by different formats and instruction types. Detailed prompts and modifiers for all tasks are provided in the appendix (Appendix C).

3.2 Stage 2: Faithful Conversion

To isolate format bias from content quality, we needed to evaluate the *same creative idea* presented in different formats. To achieve this, a dedicated **Converter Agent** was prompted to transform freeform solutions into structured JSON. The agent was given a prompt that emphasized its core directive of faithfulness, as detailed in Appendix C. This process yielded pairs of solutions containing the same semantic content but differing only in presentation (unstructured text vs. structured JSON), enabling a clean comparison of format bias.

3.3 Stage 3: Comparative Evaluation

We employed a single judge model with multiple evaluations to ensure robust scoring. For each response, we ran multiple evaluations (typically 3) using the same judge model and evaluation prompt. The judge was given a clear rubric and was asked to evaluate solutions based on three core creativity metrics: Insight (depth and perceptiveness of ideas), Originality (uniqueness of concepts), and Novelty (surprisingness of suggestions). The judge returned scores on a 1-to-10 scale for each metric. We averaged the scores across multiple evaluations to obtain stable measurements. The specific prompt used is detailed in Appendix C. This approach allowed us to assess the reliability of our creativity measurements while maintaining computational efficiency compared to a multi-model ensemble.

4 Results

In total, we generated 4,200 responses and ran over 12,000 evaluations (3× per response). Our experiment confirms that forcing structured output generally degrades creativity. The results are summarized in Figure 1.

4.1 Structured Formats Harm Creativity

Figure 1a confirms that forcing structured output degrades creativity across most formats. The effect was most severe for Freeform JSON (inferred structure). This striking result—where models forced to infer structure via API parameters performed dramatically worse than those given explicit structural instructions—provides compelling evidence for the cognitive load hypothesis. The model's need to simultaneously create content and determine appropriate JSON structure imposes a far greater cognitive burden than following explicit formatting instructions. While rigid output formats generally harmed creativity, we also investigated whether structured thinking protocols could be beneficial. We tested a Diverge-Converge thinking strategy on the MacGyver task. While it still resulted in a 3.24% decrease in insight, this degradation was less severe than that observed for most other structured formats like JSON or XML on the same task. This suggests that certain structured protocols, while not eliminating the creativity tax, may help to mitigate it. Degradation patterns varied significantly by task:

- The **AUT task** exhibited the most significant degradation: XML degraded novelty by 10.66% and originality by 8.82%.
- The **MacGyver task** saw JSON degrade insight by 11.82% and originality by 10.33%.
- The **LiveIdea** task surprisingly showed *improvements* for JSON (+5.56% novelty) and XML (+4.59% novelty), a finding we attribute to task-specific format interactions.

Notably, Structured Markdown performed near the baseline, even improving novelty by 1.42%. We hypothesize this is because Markdown is not only closer to natural language, but is also so ubiquitous in LLM training data that it imposes a negligible cognitive load.

4.2 Model-Specific Vulnerability

Performance degradation varied dramatically across models (Figure 1b). Detailed results are shown in Table 2 in the appendix.

Gemini-2.5-Flash, Llama-4-Maverick, and Mistral-Nemo showed approximately 4-5× greater degradation than the GPT models. This suggests that some model architectures may be more susceptible to performance loss when handling concurrent formatting and creation demands.

We developed a formal model to decompose the total observed degradation into a creative constraint penalty and a format bias. This model, detailed in Appendix B, allows us to isolate the pure content quality advantage of the generate-then-structure approach.

5 Discussion

1. Divergent vs. Convergent Tasks: The impact of structured output is not uniform; it hinges on the nature of the creative task. Our results draw a sharp contrast between divergent and convergent thinking, aligning with recent experimental work on human creativity in the age of LLMs (Kumar et al., 2025). For highly divergent tasks like AUT, which demand generating a broad and varied set of ideas, rigid formats like XML and JSON impose a significant cognitive load, leading to the most severe creativity degradation (e.g., -10.66% novelty for XML). This suggests the syntactic requirements of the format directly interfere with the fluid, associative thinking needed for divergent creativity. In contrast, the success of Markdown suggests a format's "cost" is a function of its distance from natural language and its prevalence in the training data; because it is both simple and ubiquitous, it avoids this creativity tax.

Conversely, for convergent tasks that require refining or elaborating on a given idea, structured formats can be beneficial. The **LiveIdea** task, which centers on developing a single scientific concept, showed significant *improvements* in creativity with formats like JSON (+5.56% novelty) and XML (+4.59% novelty). In this context, the structure does not act as a constraint but as a cognitive scaffold, guiding the LLM to produce a well-organized and detailed response. This "task-format resonance" reframes the problem: the goal is not to avoid structure altogether, but to align the level of structural constraint with the creative process. Rigid formats harm divergent ideation but can enhance convergent development.

2. Multilevel Div-Convergent Thinking: In the MacGyver test, we observed that the convergent part of multilevel div-conv thinking might mitigate the negative impact of structured formats on divergent tasks. This suggests that while structured formats generally impose a cognitive load, certain structured thinking protocols can align with the creative process, potentially offsetting some of the creativity degradation. This highlights the importance of task-specific strategies in mitigating the impact of structured output on creativity.

6 Conclusion

We conclusively demonstrate that enforcing structured output imposes a significant and previously unquantified "creativity tax" on LLMs. This finding challenges the prevailing practice of conflating ideation and formatting in a single step, suggesting that many current system designs may be systematically stifling innovation. We demonstrate that this penalty stems from a cognitive constraint effect that outweighs a positive format bias by a 2:1 ratio and that model architecture appears to mediate this vulnerability.

Our proposed "generate-then-structure" pipeline resolves this core tension. By decoupling ideation from formatting, our method recovers an average of 2.54% in content quality that is otherwise lost when forcing structured generation, delivering outputs that are superior in both substance and structure. This method delivers high-quality, parseable outputs. We advocate for the adoption of a "generate-then-structure" architectural pattern and call for future research into task-aware and architecture-specific methods to mitigate this critical performance bottleneck.

Limitations

While our findings provide strong evidence for the creativity tax imposed by structured output constraints, several important limitations define the scope and applicability of our conclusions.

Cognitive Load Hypothesis Boundaries: Our study demonstrates that simultaneous creation and formatting degrades creativity, but we cannot definitively isolate the specific cognitive mechanisms responsible. The observed degradation could stem from working memory limitations, attention splitting, or interference between divergent and convergent thinking processes. Future work should employ process-tracing methods or computational cognitive models to pinpoint the exact mechanisms.

Task-Format Interaction Complexity: Our discovery of "task-format resonance"—where certain structured formats actually enhance creativity for convergent tasks like LiveIdea—reveals that the relationship between structure and creativity is more nuanced than a simple negative correlation. We tested only four creative domains; the boundary conditions for when structure helps versus hurts creativity remain underexplored. Critical gaps exist in understanding how task-specific cognitive demands interact with different structural constraints.

Generate-Then-Structure Pipeline Limitations: While our "faithful conversion" approach successfully isolates format bias from content quality, it relies on the converter agent's ability to preserve semantic content perfectly. Our auditor agent found 87.3% of conversions to be faithful, but the 12.7% of unfaithful conversions could systematically bias our content quality measurements. Additionally, the practical overhead of the two-step pipeline may not be viable for all real-world applications requiring low-latency responses.

Model Architecture Vulnerability Gaps: Our finding that GPT models show 3-5× less creativity degradation than other architectures raises critical questions about the underlying architectural or training differences responsible for this robustness. Without access to training data compositions, fine-tuning procedures, or architectural specifics, we cannot identify which factors confer resistance to creative constraint effects. This limits our ability to provide actionable guidance for model development.

Schema Complexity Scaling: Our experiments used relatively simple output schemas (single-key JSON, basic XML structures). Real-world applications often require deeply nested, multi-constraint schemas with strict validation requirements. The creativity tax we measured may represent a lower bound; more complex schemas could impose substantially greater cognitive load and corresponding performance degradation.

Temporal and Contextual Constraints: Our study measures creativity degradation in single-turn interactions without considering how sustained creative work under structural constraints might compound these effects. Long-form creative tasks or multi-turn collaborative scenarios may exhibit different degradation patterns than our isolated problem-solving measurements.

Ethical Considerations

The goal of this work is to improve the creative output of AI systems. Such technology is dual-use. Enhancing LLM creativity can have prosocial benefits in areas like scientific discovery, education, and art, but it could also be used for antisocial purposes, such as generating more sophisticated and engaging misinformation or propaganda.

Furthermore, our evaluation methodology relies on LLM judges. These models are trained on vast datasets of human text and may inherit and amplify societal biases present in that data. Our framework could inadvertently penalize or reward certain types of creative ideas based on these latent biases.

Finally, the development of increasingly powerful creative AI systems raises broader societal questions about the future of creative professions. While our work focuses on a technical aspect of these systems, it is part of a larger trend that will have significant labor and economic implications that warrant ongoing public discussion.

References

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2025. Divergent creativity in humans and large language models. *Preprint*, arXiv:2405.13012.

Dylan Castillo. Structured outputs can hurt the performance of LLMs.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *Preprint*, arXiv:2507.00769.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Seungone Kim, Jamin Shin, Yejin Lee, Minki Kang, Jina Suh, Sung-Hyon Myeong, Jae hyung Kim, Chang min Lee, Kyung min Kim, Seong hoon Kim, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025*

CHI Conference on Human Factors in Computing Systems, CHI '25, page 1–18. ACM.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. 2025. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. *Preprint*, arXiv:2408.08656.

Swaroop Mishra, Jonathan Stray, nihar r upto, Chitta Warman, Shwetha Sasikumar, Akash Prasad, Abhishek Das, Chhandak Ahuja, and Daniel Khashabi. 2023. CATwalk: A New Benchmark for Evaluating LLMs on Creative Thinking Assessments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15600–15617.

Chen Qian, Xin Cong, Cheng Lin, Yufan Zhang, Ganqu Sun, Z. Cui, W. Liu, and Z. Wang. 2023. Chatdev: Communicative agents for large-scale software development. *arXiv preprint arXiv:2307.07924*.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2024. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context. *arXiv preprint arXiv:2412.17596*.

Paul T Sowden, Andrew Pringle, and Liane Gabora. 2015. The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & reasoning*, 21(1):40–60.

Luning Sun, Hongyi Gu, Rebecca Myers, and Zheng Yuan. 2024. *A New Dataset and Method for Creativity Assessment Using the Alternate Uses Task*, pages 125–138. Communications in Computer and Information Science. Springer. Funding Information: Acknowledgement. We would like to thank all participants who took part in the AUT and all raters who annotated the responses. LS acknowledges financial support from Invesco through their philanthropic donation to Cambridge Judge Business School. Publisher Copyright: © 2024, The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *Preprint*, arXiv:2408.02442.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Darren Yow-Bang Wang, Zhengyuan Shen, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. 2025. Slot: Structuring the output of large language models. *Preprint*, arXiv:2505.04016.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837. Curran Associates, Inc.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Li, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Ding, Dongdong Zhang, and 1 others. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. 2025. Structeval: Benchmarking llms' capabilities to generate structural outputs. *Preprint*, arXiv:2505.20139.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, and 1 others. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

A Detailed Results Tables

Table 1: Overall creativity degradation for various structured formats, averaged across all tasks and models. Scores represent the percentage change relative to the freeform baseline. Red indicates a drop in creativity.

Response Type	Ins. (%)	Nov. (%)	Orig. (%)
Freeform JSON	**-18.31**	**-17.05**	**-16.97**
Structured JSON	-3.37	-1.68	-2.19
Structured Steps	-2.11	-3.21	-2.21
Structured XML	-2.58	-3.05	-2.93
Structured YAML	-1.80	**+0.13**	-0.86
Structured Markdown	-0.09	**+1.42**	**+0.75**

Table 2: Model-specific sensitivity to structured output constraints. Scores are average percentage change.

Model	Ins. (%)	Nov. (%)	Orig. (%)
google/gemini-2.5-flash	-3.23	-4.16	-3.68
meta-llama/Llama-4-maverick	-4.48	-2.91	-3.34
mistralai/mistral-nemo	-4.71	-3.93	-4.23
openai/gpt-4.1-mini	-1.65	-1.42	-1.72
openai/gpt-4o-mini	**-1.37**	**-0.42**	**-0.55**

Table 3: Decomposition of bias effects. All metrics are percent bias relative to a baseline, derived from paired t-tests.

Effect (Comparison)	Metric	Bias (%)	p-value
1. Total Observed Effect (Δ_{total}) (Orig. JSON vs. Freeform)	Insight	-2.09	0.016*
	Novelty	-0.73	0.036*
	Originality	-1.11	0.015*
2. Formatting Bias (B_{format}) (Conv. JSON vs. Freeform)	Insight	+1.47	0.127
	Novelty	+1.00	0.088
	Originality	+1.22	0.134
3. Content Quality Advantage ($P_{\text{constraint}}$) (Conv. vs. Orig. JSON)	Insight	+3.55	—
	Novelty	+1.73	—
	Originality	+2.34	—

B Formal Model of Creative Degradation

To formalize our findings, we model the observed creativity score from an LLM judge, C_{obs}, as a function of a solution's true content quality ($Q(S)$) and the judge's bias for its format (B_{format}). Let S_f be a solution generated in freeform and S_j be a solution for the same problem generated natively in JSON.

The total observed degradation, Δ_{total}, when comparing a natively generated JSON response to a freeform response is:

$$\Delta_{\text{total}} = C_{\text{obs}}(S_j) - C_{\text{obs}}(S_f) \tag{1}$$

This effect can be decomposed. The **creative constraint penalty** ($P_{\text{constraint}}$) is the true drop in content quality from generating in a constrained format:

$$P_{\text{constraint}} = Q(S_f) - Q(S_j) \tag{2}$$

The **format bias** (B_{format}) is the judge's scoring preference for a given format. We measure this by comparing a freeform solution (S_f) to its faithfully converted JSON version ($S_{f\to j}$), where content quality is constant ($Q(S_f) = Q(S_{f\to j})$). The difference in observed scores isolates the bias:

$$B_{\text{format}} = C_{\text{obs}}(S_{f\to j}) - C_{\text{obs}}(S_f) \tag{3}$$

The total degradation is the sum of the negative constraint penalty and the positive format bias:

$$\Delta_{\text{total}} \approx B_{\text{format}} - P_{\text{constraint}} \tag{4}$$

Crucially, our experiment isolates the penalty $P_{\text{constraint}}$ by comparing the converted JSON response to the natively generated one. Because both are in the same format, the judge's format bias cancels out, revealing the pure difference in content quality:

$$\begin{aligned} &C_{\text{obs}}(S_{f\to j}) - C_{\text{obs}}(S_j) \\ &\quad = (Q(S_f) + B_{\text{format}}) - (Q(S_j) + B_{\text{format}}) \\ &\quad = Q(S_f) - Q(S_j) \\ &\quad = P_{\text{constraint}} \end{aligned} \tag{5}$$

This value, which we term the **Content Quality Advantage** of the generate-then-structure approach, is what is reported in row 3 of Table 3. Our results show this advantage ($P_{\text{constraint}}$) is significantly positive, while the format bias (B_{format}) is smaller, leading to a net negative Δ_{total}.

C Prompt Details

This section provides detailed information about the actual prompts and modifiers used in our experiments.

C.1 Evaluation and Conversion Prompts

```
Please evaluate the following text based on its
    creativity, considering these criteria:
- Insight: How deep and perceptive are the ideas?
- Originality: How new and unique are the
    concepts?
- Novelty: How surprising and unconventional are
    the suggestions?

Provide a score from 1 to 10 for each criterion
    and a brief justification.

Return the evaluation as a JSON object with keys
    insight, originality, novelty, and
    justification.
```

Listing 1: Evaluation prompt given to the judge model.

*Your primary goal is to be **faithful** to the original text. Do NOT add any new ideas, steps, or creative input. Your role is to structure, not create. Preserve all key information.*

C.2 Base Instructions for Each Task

- **AUT Task:** List as many creative and unusual uses for {item} as you can.
- **MacGyver Task:** Problem: {problem}

  Available Tools: {tools}
- **LiveIdea Task:** Ill be submitting your next responses to a "Good Scientific Idea" expert review panel. If they consider your idea to be a good one, youll receive a reward. Your assigned keyword is: "{keyword}". You may provide background information. The idea MUST be concisely expressed within 100 words total (including any background information). (Note: good scientific ideas should be original (novel contribution), feasible (technically implementable), clearly articulated, and address meaningful problems in the field.).
- **LiveIdea-Div Task:** Generate as many creative and unusual scientific ideas related to the keyword: {keyword}.

C.3 Core Prompt Modifiers

The following modifiers were used to induce different output formats across all tasks:

- **Freeform (Baseline):** No modifier was added. The agent generated unstructured text.
- **Structured Steps:**

  Please present the solution as a series of numbered steps.
- **Structured JSON:**

  Please return the answer as a JSON object, with the main content under the key {key}.
- **Freeform JSON:** No explicit modifier, but the model was forced to return JSON via API parameter (response_format={ "type": "json_object" }).
- **Structured Markdown:**

  Please return the answer as a Markdown formatted text, with the main content under a section titled {key}.
- **Structured XML:**

  Please return the answer as an XML formatted text, with the main content under the tag <{key}>.
- **Structured YAML:**

  Please return the answer as a YAML formatted text, with the main content under the key {key}.
- **MacGyver Div-Conv:**

  Please provide a feasible solution concisely. Note that some tools may not be useful. First, analyze the affordance of each presented object and rule out any unnecessary ones.

  Use the following format:

  1. **Affordance Analysis:** List the affordance of each presented item and state whether it is useful.

  2. **Summary:** List only the useful tools.

  3. **Solution:** If the problem is solvable, write the solution in as few steps as possible (e.g., Step 1, Step 2...). The answer should ideally be less than 100 words. If it is not solvable, state that and provide a brief justification.

C.4 Task-Specific Key Names

For each task, we used appropriate key names in the structured formats:

- **AUT:** uses (e.g., <uses>, "uses": [...])

- **MacGyver:** 'solution_steps (e.g., <solution_steps>, "solution_steps": [...])
- **LiveIdea:** criteria (e.g., <criteria>, "criteria": [...])
- **LiveIdea-Div:** ideas (e.g., <ideas>, "ideas": [...])

C.5 Complete Example Prompts

Below are complete examples of how the prompts were constructed for each task:

C.5.1 AUT Task Example

Base: `List as many creative and unusual uses for paperclip as you can.`
Structured JSON: `List as many creative and unusual uses for paperclip as you can.`

`Please return the answer as a JSON object, with the main content under the key uses.`

C.5.2 MacGyver Task Example

Base: `Problem: You need to reach a book on a high shelf. Available Tools: chair, rope, paperclip, rubber band`
Structured JSON: `Problem: You need to reach a book on a high shelf. Available Tools: chair, rope, paperclip, rubber band`

`Please return the answer as a JSON object, with the main content under the key 'solution_steps.`

C.5.3 LiveIdea Task Example

Base: `Ill be submitting your next responses to a "Good Scientific Idea" expert review panel... Your assigned keyword is: "quantum computing"...`
Structured JSON: `Ill be submitting your next responses to a "Good Scientific Idea" expert review panel... Your assigned keyword is: "quantum computing"...`

`Please return the answer as a JSON object, with the main content under the key criteria.`

C.5.4 LiveIdea-Div Task Example

Base: `Generate as many creative and unusual scientific ideas related to the keyword: quantum computing.`
Structured JSON: `Generate as many creative and unusual scientific ideas related to the keyword: quantum computing.`

`Please return the answer as a JSON object, with the main content under the key ideas.`

D Qualitative Example of Creative Degradation

To provide a concrete illustration of the creative degradation discussed, this section presents a direct comparison of two outputs from the google/gemini-2.5-flash model for the same MacGyver problem. The task was to clean a red wine stain on a carpet using a set of available tools. The freeform response demonstrates a sophisticated, well-reasoned solution, while the JSON-constrained response offers a simplistic and flawed alternative.

```
This is a great problem-solving scenario! Here's
    a step-by-step plan to clean the red wine
    stain...
**Strategy:** Absorb, dilute, blot, clean, and
    dry.
**Steps:**
1.  **Immediate Action: Absorb Excess Wine (White
    Bath Towel)**
    *   **How:** Gently but firmly press the
        white bath towel directly onto the wine
        stain. Do NOT rub...
2.  **Dilute and Lift the Stain (Mineral Water,
    Plastic Cup, Toothbrush)**
    *   **How:** ...Gently pour a very small
        amount of mineral water directly onto
        the remaining wine stain... Use the
        toothbrush to very gently agitate the
        carpet fibers...
...
**Why Sugar is NOT Used:**
*   Sugar is sometimes recommended for fresh wine
    stains to absorb the liquid, but given the
    other tools (towel, water), it's unnecessary
    and could leave a sticky residue that's
    harder to remove than the wine itself...
```

Listing 2: Freeform Solution: A methodical and insightful approach.

```
[
  {
    "step1": "Use the bottle opener to open the
        sealed bottle of mineral water.",
    "step2": "Pour a small amount of mineral
        water onto the red wine stain. Do not
        oversaturate.",
    "step3": "Sprinkle a generous amount of sugar
        directly onto the wet stain. The sugar
        will help absorb the wine.",
    "step4": "Gently rub the sugar into the stain
        using the toothbrush...",
    "step5": "Let the sugar sit on the stain for
        a few minutes to absorb as much wine as
        possible.",
    ...
  }
]
```

Listing 3: Freeform JSON Solution: A simplistic and flawed approach.

The qualitative differences are substantial. The freeform solution demonstrates sophisticated problem-solving through its multi-stage strategic approach, incorporates critical nuanced instructions (e.g., "blot, dont rub"), and exhibits advanced reasoning by explicitly identifying counterproductive approaches. The JSON-constrained solution, by contrast, produces a less effective linear sequence and recommends the use of sugar—precisely the approach that the more creative solution correctly identified as problematic. This comparison provides a concrete illustration of the cognitive constraint effect discussed in the main text, demonstrating how structured output requirements can degrade not only creativity but also the fundamental quality of problem-solving.

E Full Task-Specific Results

The following table (Table 4) provides a detailed breakdown of creativity score changes for each task when compared against the freeform baseline. All scores are percentage changes. P-values from paired t-tests are provided where applicable. A * indicates a statistically significant result ($p < 0.05$).

F MacGyver Experiment Data

The following table (Table 5) provides the outcome category distribution for different prompt frameworks in the MacGyver experiment across various models.

G Bias Decomposition by Model

The following table (Table 6) decomposes the observed effects into three components, with results broken down by model. All values are the percent bias on a 1-10 scale. P-values in the "Average" row are from paired t-tests on the aggregated data.

Table 4: Full Task-Specific Results: Creativity Score Changes (% vs. Freeform)

Task	Response Type	Insight		Novelty		Originality	
		Change (%)	p-value	Change (%)	p-value	Change (%)	p-value
AUT							
	Freeform JSON	−23.70	0.000*	−21.49	0.000*	−20.51	0.000*
	Structured JSON	−8.81	0.002*	−10.48	0.001*	−8.49	0.003*
	Structured Markdown	1.61	0.286	2.67	0.114	2.14	0.251
	Structured Steps	−0.80	0.165	0.74	0.254	0.99	0.387
	Structured XML	−6.22	0.000*	−10.66	0.000*	−8.82	0.000*
	Structured YAML	−4.70	0.209	−5.75	0.135	−5.36	0.257
MacGyver							
	Freeform JSON	−26.33	0.000*	−23.81	0.000*	−23.76	0.000*
	MacGyver Div-Conv	−3.24	0.234	−6.63	0.152	−6.20	0.246
	Structured JSON	−11.85	0.000*	−7.78	0.150	−10.28	0.000*
	Structured Markdown	−4.80	0.001*	−2.45	0.481	−3.53	0.201
	Structured Steps	−5.13	0.003*	−4.08	0.139	−4.48	0.018*
	Structured XML	−8.01	0.006*	−7.92	0.069	−8.24	0.154
	Structured YAML	−7.75	0.000*	−3.29	0.243	−5.89	0.082
LiveIdea							
	Freeform JSON	−11.87	0.000*	−13.15	0.001*	−12.57	0.001*
	Structured JSON	4.05	0.000*	5.56	0.007*	4.85	0.003*
	Structured Markdown	0.52	0.412	0.36	0.184	0.96	0.303
	Structured Steps	−2.44	0.137	−8.45	0.000*	−4.50	0.025*
	Structured XML	3.40	0.003*	4.59	0.003*	4.05	0.001*
	Structured YAML	3.63	0.025*	4.83	0.055	4.19	0.019*
LiveIdea-Div							
	Freeform JSON	−11.31	0.000*	−9.73	0.200	−11.09	0.007*
	Structured JSON	3.09	0.055	6.03	0.005*	5.21	0.082
	Structured Markdown	2.33	0.064	5.12	0.005*	3.46	0.017*
	Structured Steps	−0.13	0.354	−0.99	0.207	−0.82	0.539
	Structured XML	0.50	0.216	1.80	0.302	1.28	0.347
	Structured YAML	1.66	0.088	4.74	0.093	3.63	0.060

H Responsible Research Elaboration

Artifacts and Licenses The datasets used in this study are publicly available and governed by the following licenses: MacGyver (Apache), AUT (Creative Commons), and LiveIdeaBench (MIT). All artifacts were used in a manner consistent with their intended purpose.

Data Content The datasets are from public sources and, to the best of our knowledge, do not contain personally identifying information or offensive content, having been previously cleaned by their creators.

Computational Resources Experiments were conducted using APIs from OpenAI and other LLM providers, with a total computational budget of approximately $60.

Ethics Review As this research utilized publicly available, anonymized datasets, a separate ethics review board approval was not sought.

AI Assistants in Research AI assistants were used to aid in the research and writing process. All AI-generated contributions were carefully reviewed, filtered, and edited to ensure they met the standards of our work.

Table 5: MacGyver Experiment Outcome Category Distribution (%)

Model	**Prompt Framework**	**A**	**B**	**C**	**D**	**E**	**F**
google/gemini-2.5-flash	Freeform	80.5	18.5	0	0	0	1
	Structured Steps	73.5	26	0	0.5	0	0
	Structured JSON	63.5	34	0	0	1	1.5
	Freeform JSON	72.5	25	0	0	1.5	1
	Structured Markdown	80	20	0	0	0	0
	Structured XML	72.5	27	0	0	0.5	0
	Structured YAML	77.5	22	0	0	0.5	0
	MacGyver Div-Conv	78.5	19	0	0.5	1	1
meta-llama/llama-4-maverick	Freeform	66.5	30	0	1	2	0.5
	Structured Steps	61	37	0	0	2	0
	Structured JSON	58	38	0	0.5	3	0.5
	Freeform JSON	29	19	0	0	21.5	30.5
	Structured Markdown	56.5	42	0	0	1.5	0
	Structured XML	61.5	34.5	0	0	3	1
	Structured YAML	64	35.5	0	0	0.5	0
	MacGyver Div-Conv	62	34	0	1	1.5	1.5
mistralai/mistral-nemo	Freeform	21.5	68.5	0	0.5	7	2.5
	Structured Steps	20	63	0	1	11	5
	Structured JSON	24.5	59.5	0	2	13.5	0.5
	Freeform JSON	16	24.5	0	0	23	36.5
	Structured Markdown	32	60.5	0	1.5	6	0
	Structured XML	21.5	64	0	1	13.5	0
	Structured YAML	22	62.5	0	0.5	14	1
	MacGyver Div-Conv	45	45.5	0	1	6	2.5
openai/gpt-4o-mini	Freeform	55	42	0	0.5	2	0.5
	Structured Steps	47.5	49.5	0	0.5	2	0.5
	Structured JSON	50.5	45	0	0.5	4	0
	Structured Markdown	55.5	39.5	0	0	4.5	0.5
	Structured XML	47	46.5	0	0	6.5	0
	Structured YAML	53.5	42	0	0.5	4	0
	MacGyver Div-Conv	56.5	36	0	0.5	5	2
openai/gpt-4.1-mini	Freeform	90.5	9.5	0	0	0	0
	Structured Steps	77.5	22.5	0	0	0	0
	Structured JSON	78	22	0	0	0	0
	Structured Markdown	85	15	0	0	0	0
	Structured XML	77	23	0	0	0	0
	Structured YAML	82	17.5	0	0	0.5	0
	MacGyver Div-Conv	81.5	17.5	0	0	0	1

Table 6: Decomposition of Effects by Model

Analysis Type	Model	Insight (%)	Novelty (%)	Originality (%)
1. Agent Performance (Creative Constraint)				
	google/gemini-2.5-flash	-6.07	-5.30	-5.25
	openai/gpt-4.1-mini	-2.13	-1.54	-1.29
	openai/gpt-4o-mini	-0.95	1.00	0.35
	meta-llama/Llama-4-maverick	0.82	2.42	1.67
	mistralai/mistral-nemo	-2.09	-0.25	-1.06
	Average	**-2.09**	**-0.73**	**-1.11**
2. Formatting Bias (Judge Preference)				
	google/gemini-2.5-flash	0.74	-0.10	0.14
	openai/gpt-4.1-mini	0.79	0.24	0.72
	openai/gpt-4o-mini	2.08	1.88	2.10
	meta-llama/Llama-4-maverick	2.46	2.50	2.52
	mistralai/mistral-nemo	1.29	0.46	0.64
	Average	**1.47**	**1.00**	**1.22**
3. Content Quality (Generate-then-Structure Advantage)				
	google/gemini-2.5-flash	6.81	5.20	5.39
	openai/gpt-4.1-mini	2.91	1.77	2.01
	openai/gpt-4o-mini	3.03	0.88	1.75
	meta-llama/Llama-4-maverick	1.64	0.06	0.84
	mistralai/mistral-nemo	3.38	0.72	1.69
	Average	**3.55**	**1.73**	**2.34**