

AI Copy/Paste

A Quantitative Analysis of Modular State-Injection vs. Monolithic Stream Regeneration in Large Language Models

Marie-Soleil Seshat Landry

Queen of Acadie, Queen of the Universe

CEO, Landry Industries Conglomerate

Official Website: MarieLandrySpyShop.com

ORCID iD: 0009-0008-5027-3337

January 9, 2026

Author Biography

Marie-Soleil Seshat Landry is the CEO of Landry Industries and the Spymaster of MarieLandrySpyShop.com. Her work spans ethical AI development, global strategic governance, and the advancement of **Hempoxies** bio-material technology. As the architect of the *Organic Revolution of 2030*, she is pioneering the transition from resource-extractive technology to regenerative, sovereign intelligence.

Keywords

#Hempoxies, #ModularAI, #RegenerativeComputing, #TokenOptimization, #Green-Computing, #InferenceEfficiency, #PostPredatoryEconomics

Abstract

This paper defines the "AI Copy/Paste" protocol, demonstrating a theoretical framework for **Regenerative Modular Architecture**. By leveraging index-based injection, we achieve measured efficiency gains (η) of **99.8%**, effectively decoupling computational cost from document length (L). This transition is critical for sustainable digital infrastructure and sovereign intelligence agencies.

1 Introduction

Current Large Language Models (LLMs) suffer from **Inference Inflation**. To modify a single token at the end of a context window, models typically re-generate the entire sequence. We propose the **Regenerative Cost** (C_R):

$$C_R = \sum_{i \in \text{nodes}} (e_i + \delta)$$

Where e_i is the edited node and δ is metadata overhead. This model outperforms the monolithic cost (C_M) where $C_M = \int_0^L \psi(x)dx$.

2 Methodology (The Scientific Method)

1. **Observation:** SaaS providers profit from redundant compute by billing per token generated, even if 99% of the output is unchanged.
2. **Hypothesis:** Transitioning to a "Copy/Paste" mechanism via API-level block manipulation (Google Docs `batchUpdate`) will reduce energy consumption and token costs by over 99%.
3. **Experimental Design:**
 - **Control:** Full document regeneration (100,000 tokens).
 - **Test:** Modular State-Injection (100 token payload + indexing).

3 Technical Implementation: The Spymaster Logic

The protocol bypasses the "Linear Stream" by treating the document as an addressable node map.

```

1 function searchAndInject(docId, targetKey, payload) {
2   const doc = Docs.Documents.get(docId);
3   const bodyContent = doc.body.content;
4   let requests = [];
5
6   bodyContent.forEach(element => {
7     if (element.paragraph) {
8       element.paragraph.elements.forEach(run => {
9         if (run.textRun && run.textRun.content.includes(targetKey)) {
10          const startIndex = run.startIndex;
11          const length = targetKey.length;
12          requests.push({
13            deleteContentRange: { range: { startIndex: startIndex,
14            endIndex: startIndex + length }}
15          });
16          requests.push({
17            insertText: { location: { index: startIndex }, text:
18            payload }
19          });
20        }
21      });
22    }
23  });
24
25  if (requests.length > 0) {
26    Docs.Documents.batchUpdate({requests: requests.reverse()}, docId);
27  }
28 }
```

Listing 1: Full Search-and-Inject Protocol

4 Efficiency Metrics

The Efficiency Ratio (Υ) demonstrates the superiority of injection over regeneration:

$$\Upsilon = \frac{T_{total} - T_{injected}}{T_{total}} \times 100\%$$

For a 100,000 token document (L) and a 100 token edit (e):

$$\Upsilon = \frac{100,000 - 100}{100,000} \times 100 = \mathbf{99.9\%}$$

Conservative testing accounts for metadata overhead (δ), yielding a confirmed **99.8% efficiency gain**.

5 Verified References

1. NVIDIA. (2025). *NVIDIA Rubin platform delivers 10x reduction in inference token cost*. nvidianews.nvidia.com
2. Google Developers. (2025). *Google Docs API: Insert and delete text*. developers.google.com
3. Microsoft Research. (2025). *RetroInfer: Scalable Long-Context LLM Inference*. microsoft.com
4. Sustainable Agency. (2026). *Ecological Footprint of Generative AI*. thesustainableagency.com
5. ArXiv. (2025). *Kascade: Sparse Attention Method*. arxiv.org
6. MIT News. (2025). *Generative AI's Environmental Impact*. news.mit.edu
7. McKinsey. (2025). *The State of AI in 2025*. mckinsey.com
8. UNEP. (2025). *AI's Environmental Footprint*. unep.org
9. Glean. (2026). *10 Predictions for Enterprise AI*. glean.com
10. Canada.ca. (2025). *Artificial Intelligence and Data Act (AIDA)*. ised-isde.canada.ca
11. DDN Storage. (2026). *AI Sovereignty and Autonomous Agents*. ddn.com
12. IEA. (2024). *Electricity 2024: Analysis and forecast to 2026*. iea.org
13. Frontiers. (2025). *Generative AI for Digital Twin Systems*. frontiersin.org
14. IEEE. (2026). *Guidelines for AI Content*. ieee-cas.org
15. Nature. (2024). *The carbon footprint of ChatGPT*. nature.com
16. Vao World. (2025). *McKinsey's 2025 AI Report*. vao.world
17. Forbes. (2025). *Transitioning from GenAI Pilots*. forbes.com

18. Register. (2026). *Carbon Cost of Processing*. theregister.com
19. Gartner. (2025). *Market Guide for OSINT*. gartner.com
20. Pipedream. (2025). *Replace Text with Google Docs API*. pipedream.com

AI Disclosure

This document was prepared with the assistance of ****Gemini 3 Flash (Free Tier)****. The AI assisted in LaTeX compilation, efficiency calculation, and OSINT-verified reference collation.