From Complex to Clear: Energy-Guided Latent Simplification in Language Models

Dheeraj Malle Rudrappa, Arjunaditya
Department of Computer Science
University of Illinois Chicago

December 4, 2025

Abstract

Large Language Models (LLMs) often generate text that is technically accurate but inaccessible to non-expert audiences. Standard approaches to text simplification, such as supervised fine-tuning, require expensive parallel corpora and lack inference-time flexibility. In this work, we propose a **Latent-Layer Energy-Based Framework** for controlled text simplification. By defining a simple energy function based on semantic fidelity and cosine similarity to simple reference embeddings, we steer the generation process of GPT-2 using Langevin Dynamics. We intervene at the latent space (Layer 4) to balance text complexity with factual preservation. Our experiments on the ASSET dataset demonstrate that our method improves Flesch Reading Ease scores by over 30 points compared to baselines while maintaining high semantic fidelity (BERTScore > 0.90), offering a modular solution for accessible AI communication.

1 Introduction

Communicating complex information in an accessible manner is a persistent challenge in domains such as medicine, law, and finance. While Large Language Models (LLMs) have demonstrated remarkable generative capabilities, their outputs often remain inaccessible due to technical jargon and convoluted sentence structures. This linguistic gap exacerbates informational inequality; for instance, patients may struggle to interpret diagnostic reports, leading to suboptimal health decisions.

Traditional approaches to text simplification rely on supervised fine-tuning on parallel corpora (e.g., WikiAuto). However, such corpora are scarce and domain-specific. Furthermore, fine-tuning modifies model parameters globally, preventing adaptive control of complexity at inference time.

To address these limitations, we propose an energy-based guidance framework using **Contrastive Optimized Latent Decoding (COLD)**. Unlike token-level steering, we intervene at intermediate layers of the LLM. We define an energy landscape where "low energy" corresponds to text that is both semantically faithful to the input and structurally simple. By optimizing the hidden states via Langevin Dynamics prior to decoding, we achieve dynamic control without retraining the backbone model.

2 Methodology

We utilize GPT-2 Small as our backbone. Our approach combines a semantic anchor (preservation) with a simplicity attractor (simplification) within the latent space.

2.1 Architecture

We intervene at Layer $L = 4$ of GPT-2. Let h_t be the hidden state of the t-th token at layer L. We aim to find an optimized embedding h_t^* that minimizes an energy function $E(h_t)$.

2.2 Energy Function

We propose a dual-term energy function to guide the simplification process. For a candidate latent vector e and the original latent vector e_p, the energy is defined as:

$$E(e) = \lambda_{\text{preserve}} \underbrace{\|e - e_p\|^2}_{\text{Semantic Anchor}} + \lambda_{\text{simplify}} \sum_{j=1}^{K} \underbrace{(1 - \cos(e, e_j))}_{\text{Simplicity Attractor}} \tag{1}$$

where:

- e_p is the original hidden state from the model forward pass.
- $\{e_j\}_{j=1}^{K}$ are reference embeddings derived from $K = 3$ simple sentences (e.g., "The heart is weak"), extracted via average pooling at Layer 4.
- $\lambda_{\text{preserve}}$ and $\lambda_{\text{simplify}}$ are hyperparameters controlling the trade-off between fidelity and simplification.

2.3 Langevin Dynamics Optimization

To minimize $E(e)$, we employ Langevin Dynamics, a gradient-based sampling method. At each step i, the latent vector is updated as:

$$z_{i+1} = z_i - \eta \nabla_z E(z_i) + \epsilon\sqrt{2\eta} \tag{2}$$

where η is the step size and $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise. This process allows the model to traverse the energy landscape toward "simpler" regions while the preservation term prevents the representation from drifting off the semantic manifold.

3 Experiments

3.1 Setup

We evaluated our method on the ASSET dataset using GPT-2 Small.

- **Baselines:** Zero-shot GPT-2 prompting ("Simplify: [Input]").
- **Parameters:** We used Layer 4, $N = 5$ Langevin steps, step size $\eta = 0.001$, $\lambda_{\text{preserve}} = 50.0$, and $\lambda_{\text{simplify}} = 1.0$.
- **Metrics:** Flesch Reading Ease (Readability), BERTScore (Semantic Fidelity).

3.2 Quantitative Results

Table 1 summarizes the performance. The Latent COLD approach significantly outperforms the baseline in readability while maintaining high semantic similarity.

Table 1: Comparison of Baseline vs. Latent COLD (N=20 subset).

Method	Flesch Score ↑	BERTScore (F1) ↑	Win Rate
GPT-2 Baseline	28.4	**0.970**	-
Latent COLD (Ours)	**62.1**	0.942	+33.7 pts

3.3 Qualitative Analysis

We highlight two distinct test cases to demonstrate the model's capabilities (Table 2).

Case 1 (General Complexity): The model successfully transformed "proliferation" and "exacerbated" into "spread" and "made worse," significantly increasing accessibility (Flesch 28.4 $\rightarrow$ 62.1).

Case 2 (Factual Preservation): A major risk in generative simplification is hallucination. In the "Sudan" example, our model successfully simplified "armed conflicts" to "war" but correctly retained the named entities ("Sudanese military", "Janjaweed"). This validates the effectiveness of our high $\lambda_{\text{preserve}}$ weight.

Table 2: Qualitative Examples of Simplification.

Case	Baseline Output	Latent COLD Output
Social Isolation	The proliferation of digital technology has exacerbated the issue...	The spread of digital technology has made the issue of social isolation worse.
Sudan Conflict	The armed conflicts are not a single conflict. They are...	One side of the war is the Sudanese military and the Janjaweed militia.

3.4 Ablation Study

We analyzed the sensitivity of the simplification weight $\lambda_{\text{simplify}}$.

- $\lambda = 0.5$: Output remained similar to baseline (Flesch ≈ 30).
- $\lambda = 2.0$: Output achieved optimal simplification (Flesch ≈ 68).
- $\lambda > 5.0$: The model began to suffer from "manifold departure," generating fluent but hallucinated text (e.g., drifting to unrelated topics), highlighting the need for strong preservation constraints.

4 Discussion and Ethics

4.1 Manifold Departure Risks

During development, we observed that unconstrained energy optimization leads to hallucinations (e.g., the model generating text about "social media users" instead of "digital technology"). This phenomenon, known as manifold departure, occurs when the latent vector is pushed into low-probability regions of the embedding space. We mitigated this by introducing "Safe Mode" parameters: reducing the step size to $\eta = 0.001$ and increasing the preservation weight to $\lambda = 50.0$.

4.2 Ethical Considerations

Simplification promotes accessibility but risks over-simplification or factual distortion. Our semantic fidelity checks (BERTScore > 0.90) ensure that critical details—such as proper nouns in geopolitical contexts—are

preserved. Future work should incorporate Entity Matching (NER) directly into the energy function to strictly enforce factual consistency.

5 Conclusion

This project demonstrates that **Latent COLD** is a viable, inference-time method for controlled text simplification. By intervening at Layer 4 with a dual-term energy function, we achieved a substantial improvement in readability (+33 points Flesch) without fine-tuning. The approach offers a modular, data-efficient alternative to supervised learning, making AI-generated text more accessible to diverse audiences.

References