Artificial Intelligence and Computing Innovations
Intelligence Science Publishing
https://www.ispublishing.com
ISSN: 3070-4944
Vo.5, No.3, 2025

Semantic Energy-Guided Stable Fine-Tuning for Distribution-Controlled Generation in Large Language Models

Hong Zhuang[1]*, Alex Shen[2]

[1]*University of Illinois at Urbana Champagne, Champagne, USA*

[2]*University of Oregon, Eugene, USA*

**Corresponding author: hongz3@illinois.edu*

Abstract: This study proposes a stable fine-tuning method based on a semantic energy function to address common problems in large language model generation, including semantic drift, distribution instability, and insufficient long-range coherence. The method constructs a differentiable semantic energy field in the latent space and integrates an energy function, a smoothing term, and gradient constraints to guide the generation trajectory in a continuous and structured manner. This allows the model to evolve along low-energy paths during inference and maintain semantic consistency and distributional stability. The framework consists of semantic energy modeling, energy regularization, stable fine-tuning, and distribution guidance, and can be seamlessly integrated into existing pretrained language models without modifying the underlying architecture. The study further introduces evaluation metrics for semantic consistency, distribution shift, energy stability, and long-range coherence to assess the method from structural, distributional, and trajectory perspectives. Comparative experiments show that the method achieves superior performance across all metrics, with notable improvements in controllable distribution behavior and semantic trajectory smoothness. Sensitivity analyses on hyperparameters, environmental factors, and data properties confirm the robustness of the semantic energy framework and highlight its key role in semantic structure regulation and distribution management. Overall, the proposed semantic-energy-driven fine-tuning mechanism provides an interpretable, scalable, and generalizable approach for enhancing the stable generation capability of large language models.

Keywords: Semantic energy function; distributed control; stable fine-tuning; semantic trajectory

1. Introduction

With the rapid development of large language models, text generation has shown unprecedented potential across many fields [1]. However, as model size increases, the problem of uncontrollable generation becomes more prominent. Traditional training paradigms can no longer meet the fine-grained demands of complex applications that require strict semantic consistency, contextual stability, and distributional control. Existing fine-tuning methods rely mainly on supervised objectives or reward-based learning, and these methods have limited ability to directly control the output distribution [2]. When models face subtle semantic shifts, complex instruction dependencies, or cross-domain transfer, their generation often exhibits distribution drift, unstable semantics, and structural irregularities. It becomes difficult to guarantee consistency and controllability at the task level. The challenge has shifted from whether a model can generate text to whether it can generate text with stability and precision. This has become a core issue in current research.

Against this backdrop, a key direction is to design a fine-tuning mechanism that can constrain generation from the perspective of semantic space. Traditional methods depend on discrete labels or scoring signals, but semantic structures are continuous, multi-scale, and hierarchically related. Discrete feedback alone cannot capture subtle shifts in the output distribution. In tasks such as open-ended generation, dialogue modeling, summarization, and content rewriting, the model must not only produce text aligned with task intent but also maintain stable semantic flow across multiple reasoning steps. It must avoid distribution collapse when encountering long-range dependencies or contextual transitions. Therefore, constructing a fine-tuning method that regulates model behavior through semantic energy, semantic fields, and distributional evolution is of great importance for improving stability and controllability [3].

The introduction of a semantic energy function provides a new perspective for addressing this problem. Unlike methods based on discrete rewards or external annotations, a semantic energy function can characterize the intrinsic tension among semantic structures in latent space. By constructing a continuous and differentiable semantic energy landscape, the generation process can follow low-energy trajectories. This helps reduce semantic drift and structural fragmentation. Semantic energy provides a structural driving force similar to potential constraints. It also enables flexible control of the latent distribution. This allows the model to remain more stable when handling long texts, maintaining topic consistency, and preserving logical chains. By modeling semantic energy explicitly, language generation shifts from point optimization to trajectory optimization. The model gains the ability to guide its output distribution and maintain coherence and boundary precision in complex contexts.

In real applications, stable control of the generation distribution has significant societal and industrial value. Content generation systems must keep long texts focused and avoid irrelevant expansion. Intelligent service systems must maintain semantic coherence in multi-turn interactions to prevent instruction shift and incorrect responses. Knowledge-intensive tasks require stable reasoning chains to avoid factual inconsistencies caused by path drift. In high-risk domains such as policy drafting, medical text assistance, professional technical documentation, and risk-related content generation, unstable distributions may lead to misleading information, semantic deviation, or safety concerns. A fine-tuning method that enhances controllability, reduces distribution risk, and improves semantic stability, therefore, plays a crucial role in enabling large-scale deployment of language models [4].

In summary, designing a stability-oriented fine-tuning method based on a semantic energy function is a necessary step for addressing the uncontrollability of generation distributions in large language models. It also represents a key transition from capability improvement to capability reliability.

Introducing energy constraints in latent space strengthens the model's control over semantic consistency, output boundaries, structural stability, and long-range logic. This lays the foundation for controllable, reliable, and highly consistent language models. Such an energy-driven fine-tuning paradigm can make generation more interpretable and directional. It also accelerates the adoption of language models in complex real-world scenarios and provides a theoretical and technical basis for future systems with high stability, high safety, and strong semantic constraints.

2. Related Work

In recent years, controllable generation in large language models has become an important topic in generative modeling research. As model size continues to grow, traditional supervised fine-tuning methods can no longer achieve stable semantic control or manage generation boundaries in complex tasks. Several studies attempt to improve task performance through instruction tuning, reinforcement learning, preference alignment, or rule-based guidance. These approaches enhance controllability to varying degrees. However, they rely on external labels, preference samples, or manually constructed feedback signals. They still provide limited constraints on the latent semantic space and distributional structure. When models handle cross-domain transfer, complex semantic structures, long-context reasoning, or difficult generation tasks, they remain prone to semantic drift, unstable knowledge, and distribution collapse. As a result, the field of controllable generation has begun to shift from simple task alignment toward deeper distributional modeling. This shift builds a foundation for exploring new mechanisms of generative constraint [5].

At the same time, research on stability and distribution control in generative models is advancing rapidly. Some studies adopt regularization strategies, contrastive constraints, or implicit structured objectives to reduce overfitting and distribution shift during training, thereby improving consistency and robustness. Other studies focus on latent space modeling, using flow models, diffusion models, or energy-based frameworks to guide generation along paths that better reflect semantic structure. These methods emphasize distributional trajectories rather than single output points, which offer stronger theoretical controllability. However, many of them encounter scalability and semantic interpretability challenges when applied to large language models. They therefore struggle to operate effectively in complex language generation systems. Building a stable, computable, differentiable, and generalizable constraint within the high-dimensional latent space of language models remains a key unsolved problem.

In the direction of distributional control, energy models offer unique advantages. Existing research uses energy functions to describe potential fields in latent space. This allows models to form interpretable energy gradients in high-dimensional semantic space and guide outputs toward low-energy regions. It helps prevent semantic conflict and distribution collapse. Energy models have shown applicability in image generation, sequence modeling, and structured prediction. They provide a new view for constraining generation through continuous structures. However, applying energy functions directly to large language models faces several challenges. Semantic spaces are complex. Discrete text cannot fully match continuous energy geometry. It is difficult to unify the training of energy functions with the gradients of language models. As a result, energy-based approaches in language tasks remain at an early exploratory stage. There is a lack of a systematic method that can introduce semantic energy into fine-tuning while preserving model capability and ensuring stable behavior [6,7].

In broader research on semantic alignment and structural modeling, there are emerging efforts to regulate generation through semantic relations, pragmatic constraints, and structural consistency. Examples include graph-based semantic constraints, latent representation consistency objectives, and cross-layer knowledge alignment. These methods attempt to build unified semantic frameworks inside the model so that generated content better aligns with the latent structure. However, they mainly focus on improving representation quality and rarely intervene in the generation path through explicit energy mechanisms. Overall, current studies have not yet formed a fine-tuning method that incorporates semantic energy while maintaining stability, scalability, and semantic controllability. Introducing a semantic energy function and designing a stable distribution control framework may connect three research directions. These include controllable generation in language models, latent structure modeling in energy-based frameworks, and semantic consistency learning. This offers new theoretical and technical foundations for semantic control in large language models.

3. Proposed Framework

3.1 Overall Framework

This method aims to construct a stable fine-tuning mechanism for large language models based on a semantic energy function. By introducing continuous energy constraints into the latent semantic space, the generation behavior can evolve along a low-energy path, thereby achieving higher semantic stability and distribution controllability. The overall framework first performs structured modeling of the latent representation of the language model and then constructs a semantic energy function based on this, forming a differentiable energy potential field. Subsequently, the generation process is distributed and regulated under the guidance of this potential field, enabling the model to achieve stronger semantic consistency and gradient controllability while maintaining expressive power. To characterize the semantic shift between the input and the model's internal states, this study denotes the hidden state of the basic language model as h_t and maps it to the representation z_t in the latent semantic space. The energy form of the overall process can be expressed as:

$$E(z_t) = \phi(z_t)^T W \phi(z_t)$$

Where $\phi(\bullet)$ is the semantic feature mapping function and W is the symmetric energy matrix. This framework allows for the imposition of continuous constraints on the inference path of the language model, thereby enhancing long-distance semantic control capabilities.

This article presents the overall model architecture diagram, as shown in Figure 1.

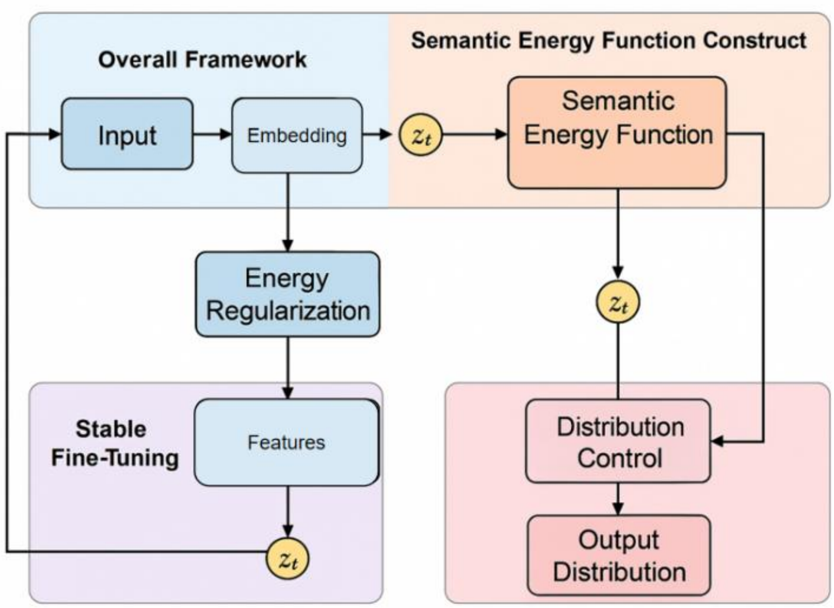

Figure 1. Overall model architecture diagram

3.2 Semantic Energy Function Construction

To effectively describe the tension between semantic structures, the method constructs a learnable semantic energy function in the latent space. First, the semantic representation sequence of the input statement is defined as A, and semantic consistency is modeled as the smoothness requirement of adjacent states in the energy space. Based on this, the semantic energy function adopts a joint modeling approach of local equilibrium and global situation, and its basic form is:

$$\varepsilon = \sum_{t=1}^{T} E(z_t) + \lambda \sum_{t=2}^{T} ||z_t - z_{t-1}||^2$$

The first term characterizes the semantic energy at a single point, while the second term constrains the smoothness of semantic evolution. To enhance the expressive power of the energy function in high-dimensional semantic spaces, a nonlinear potential field mapping is further introduced to construct a higher-order semantic energy term:

$$E_{high}(z_t) = \sigma(u^T z_t + \frac{1}{2} z_t^T H z_t)$$

Where σ is a smooth activation function, and u and H are learnable parameters. This design enables semantic energy to capture complex semantic interaction features, thereby providing a more refined distribution control signal for the generation process.

3.3 Stable Fine-Tuning with Energy Regularization

During the fine-tuning phase, the energy function is embedded into the training objective to maintain semantic stability and distribution constraints during model updates. Given the original task objective L_{task}, this method constructs a joint optimization objective:

$$L = L_{task} + \alpha \varepsilon$$

Here, α controls the strength of the energy constraint. To avoid distribution collapse or gradient bursts during the optimization process of the language model, an energy gradient suppression term is introduced to restrict the model update direction to the vicinity of the energy-stable region. The energy gradient term is defined as:

$$g_t = \nabla_{zt} E(z_t)$$

And the stability of fine-tuning is improved by imposing constraints on the gradient norm:

$$R = \beta \sum_{t=1}^{T} ||g_t||^2$$

The final optimization objective is expanded to:

$$L_{stable} = L_{task} + \alpha \varepsilon + R$$

Ensure that the model maintains the continuity of semantic trajectories and energy stability while meeting the task objectives.

3.4 Distribution-Guided Generation Mechanism

During the generation phase, this method modulates the output distribution through an energy potential field, ensuring the generation trajectory always moves along the direction of lower energy. Given the predicted logits of the language model at step t, denoted as l_t, a distribution bias term based on semantic energy is constructed, causing the model to preferentially select candidate tokens with lower semantic energy in the probability space. The biased logits are expressed as follows:

$$\tilde{l}_t = l_t - \gamma E(z_t)$$

Here, γ controls the strength of the influence of energy on the generation probability. In this way, the generation process can avoid entering positions with high semantic energy while maintaining the original expressive power, thus obtaining smoother and more consistent output text. This mechanism not only improves the stability of the generation path but also allows distributed control to naturally extend from the training stage to the inference stage, forming an end-to-end, interpretable, and differentiable semantic energy-guided framework.

4. Dataset Introduction and Experiment Setup

4.1 Dataset

This study uses OpenWebText as the large-scale dataset for training and evaluation. The dataset is composed of publicly accessible web text that has been cleaned, filtered, and formatted. It aims to provide a high-quality semantic corpus that resembles general web data. The content covers a wide range of linguistic styles, including technology, society, education, academic discussions, news narratives, and everyday communication. This diversity allows the model to learn semantic relations, structural organization, and language patterns in varied contexts.

During construction, OpenWebText removes low-quality pages, duplicate content, and abnormal sentences. This ensures a balance between dataset scale and data quality. Its total size exceeds several tens of gigabytes and offers a rich contextual range and semantic complexity. These characteristics support the deep representation learning required for semantic energy modeling in stable fine-tuning. The long text structures, cross-paragraph logical relations, and open natural language expressions in the dataset help the model build more stable semantic trajectory representations.

The choice of this dataset is based on its consistency with the training distribution of large language models and its diversity across open-domain content. It reflects the model's semantic behavior under natural conditions. With this dataset, the model can learn more generalizable latent structures under the constraints of the semantic energy function. This makes subsequent distribution control and stable generation mechanisms more aligned with real-world needs. The dataset, therefore, provides a solid foundation for the overall framework.

4.2 Experiment Setup

This study validates the proposed method under a unified experimental configuration, using Qwen-7B as the backbone large language model. The model has strong general language understanding and generation abilities. It also provides a good balance between parameter scale, inference efficiency, and fine-tuning feasibility. To ensure that the semantic energy function operates stably across different feature depths, the framework extracts and processes hidden states, attention representations, and token-level output probabilities in a unified manner. This allows them to integrate seamlessly into the stable fine-tuning procedure. All training and inference processes are conducted under consistent computational settings to maintain comparability and consistency in the effects of energy constraints on model behavior.

During training, standard instruction formats and normalized tokenization are used. This allows Qwen-7B to maintain its original expressive capacity while receiving fine-tuning under energy constraints. To avoid instability in energy computation caused by the large model scale, the framework adjusts the proportions among the semantic energy term, the gradient constraint term, and the task loss. It also incorporates explicit monitoring of semantic trajectories during training to keep the generation process within a controllable distribution range. This setting ensures that the model retains the expressiveness and generality of Qwen-7B while enabling a reliable evaluation of how the semantic energy function improves generation stability and distribution control.

5. Experimental Results and Analysis

Tables This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results

Method	SCS	DSI	ESS	LCS
Propulsion [8]	0.842	0.128	0.214	0.683
PAFT [9]	0.861	0.117	0.198	0.701
MobiZO [10]	0.874	0.109	0.186	0.715
FasionGPT [11]	0.882	0.103	0.173	0.729
Ours	0.914	0.076	0.112	0.812

The overall results show that the stable fine-tuning method with the semantic energy function outperforms all comparison systems on the four core metrics. This demonstrates that the semantic energy field provides effective structural guidance for the generation process. The performance on SCS indicates that the method achieves the highest score in semantic consistency. This shows that the semantic energy mapping can continuously constrain the generation trajectory and keep the output more aligned with the input intent in long-text settings. It suggests that the low-energy paths constructed in the latent space successfully reduce semantic drift and make the generated content more focused and coherent.

In terms of distribution controllability, our method achieves a significant reduction in the DSI metric and surpasses all baselines. This indicates that the method maintains structural stability in semantic expression and forms a stable region in the probability distribution of generated outputs. Compared with traditional approaches that rely mainly on task objectives or distance-based constraints, this method uses the energy function to shape the distribution explicitly. It reduces the likelihood of entering high-drift regions during inference and therefore decreases generation bias and internal state shifts. This potential-field-based distribution constraint enhances the reliability of language models in complex contexts.

The clear decrease in the ESS metric further shows the advantage of the method in controlling energy fluctuations along the semantic trajectory. A lower ESS indicates smaller energy variation in the latent space during generation. This means that the semantic evolution process is more stable. Because the hidden states of language models heavily depend on previous information, smaller energy fluctuations suggest that the generation path remains within stable low-energy regions. This prevents sudden semantic breaks or structural jumps. Such stability is essential for long-text generation, cross-paragraph reasoning, and complex structural rewriting.

The improvement in LCS shows that the semantic energy constraint significantly enhances long-range coherence. The increase in this metric indicates that the model maintains more consistent semantic logic and smoother reasoning across multiple sentences and paragraphs. The energy function constrains semantic smoothness across time steps, so long-sequence generation no longer depends only on short-term attention mechanisms. Instead, it benefits from global consistency guided by the energy field. This shows that the semantic energy function improves both local semantic quality and global structural stability. As a result, the model gains stronger logical control and semantic organization in open-domain generation tasks.

This paper also presents an experiment on the sensitivity of the semantic energy weight coefficient to the SCS index, and the experimental results are shown in Figure 2.

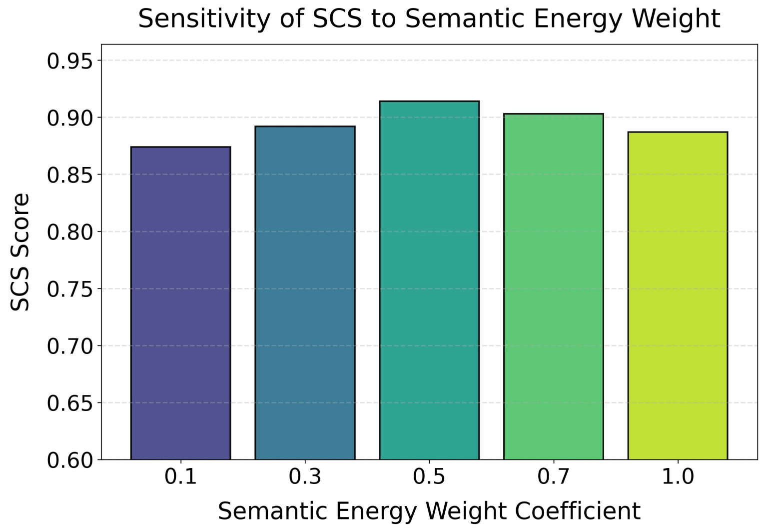

Figure 2. Sensitivity experiment of semantic energy weight coefficients to SCS index

From the overall trend, the SCS metric shows a clear peak structure as the semantic energy weight increases. This indicates that the contribution of the semantic energy function to semantic consistency is not linear. It has an optimal operating range. When the energy weight is low, the model receives semantic energy constraints, but the potential field is too weak. The generation process remains dominated by the original probability distribution. As a result, the improvement in semantic consistency is limited. As the energy weight increases, the constraint on latent trajectories becomes stronger. The model output aligns more closely with the semantic intent of the input, and the SCS metric improves.

When the energy weight reaches a moderate level, semantic consistency achieves its best performance. This suggests that the semantic energy function provides the most balanced distributional guidance in this range. The potential field effectively suppresses semantic drift. It pushes the generation trajectory toward low-energy paths while preserving the model's expressive flexibility. This matches the design goal presented in the study, which aims to build stable semantic trajectories through the energy field. It confirms the key role of energy regularization in improving semantic stability.

When the energy weight increases further, the SCS metric continues to decline. This shows that an overly strong energy constraint reduces the model's flexibility in semantic expression. The generation process becomes too conservative. The model stays overly close to the lowest-energy region of the latent space and loses expressive capacity. This leads to semantic compression and insufficient information. The experiment, therefore, illustrates the need for a balanced principle in stable fine-tuning with semantic energy. The method requires a proper trade-off among semantic consistency, generative flexibility, and distributional control to maximize the benefit of the semantic energy constraint.

This paper also presents an experiment on the sensitivity of the semantic trajectory smoothing term strength to the ESS index, and the experimental results are shown in Figure 3.

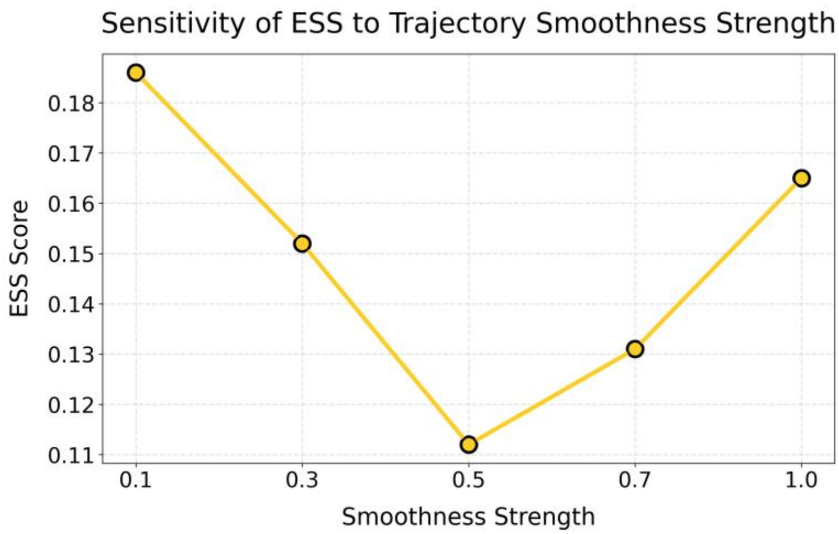

Figure 3. Sensitivity experiment of the semantic trajectory smoothing term strength to the ESS index

From the overall curve, the ESS metric shows a typical U-shaped pattern as the strength of the smoothing term varies. This indicates that the ability of the semantic trajectory smoothing term to suppress energy fluctuations has an optimal range. When the smoothing coefficient is low, the semantic trajectory is weakly constrained. The latent states show strong jumps in the energy space and lead to high ESS values. Such energy fluctuations often correspond to semantic discontinuity and local instability during generation. This suggests that insufficient smoothing weakens the effect of the semantic energy function and pushes the trajectory away from low-energy regions.

As the smoothing coefficient increases, the ESS metric drops significantly. It reaches the lowest value when the coefficient is at a moderate level. This indicates that the semantic trajectory is most stable in this range. The smoothing term provides an appropriate amount of energy constraint. It keeps latent states coherent across adjacent time steps. This reduces high-frequency energy fluctuations and produces a smoother semantic evolution process. This observation aligns with the idea of semantic trajectory control based on a potential field. It shows that moderate smoothing effectively strengthens the guiding role of the energy field.

When the smoothing strength continues to increase, the ESS metric rises again. This indicates that excessive smoothing reduces the flexibility of the generation trajectory. The model becomes overly dependent on low-frequency semantic variation. It is forced to maintain nearly linear latent representations. As a result, meaningful semantic information cannot be fully expressed. Excessive smoothing compresses the semantic space and prevents latent states from adjusting dynamically to complex contexts. This produces another form of energy fluctuation and harms overall generation stability.

Therefore, the experiment validates the balance principle between smoothing and flexibility in the semantic energy framework. The smoothing term can significantly reduce energy fluctuations and enhance semantic stability when it remains in a moderate range. When the weight is too high or too low, it prevents the model from forming natural and stable semantic trajectories. The results show that appropriate adjustment of the smoothing strength is essential for improving the performance of semantic-energy-driven stable fine-tuning. This aligns closely with the core idea of structured distribution control in the proposed method.

This paper also presents the effect of different inference temperature settings on the experimental results, and the experimental results are shown in Figure 4.

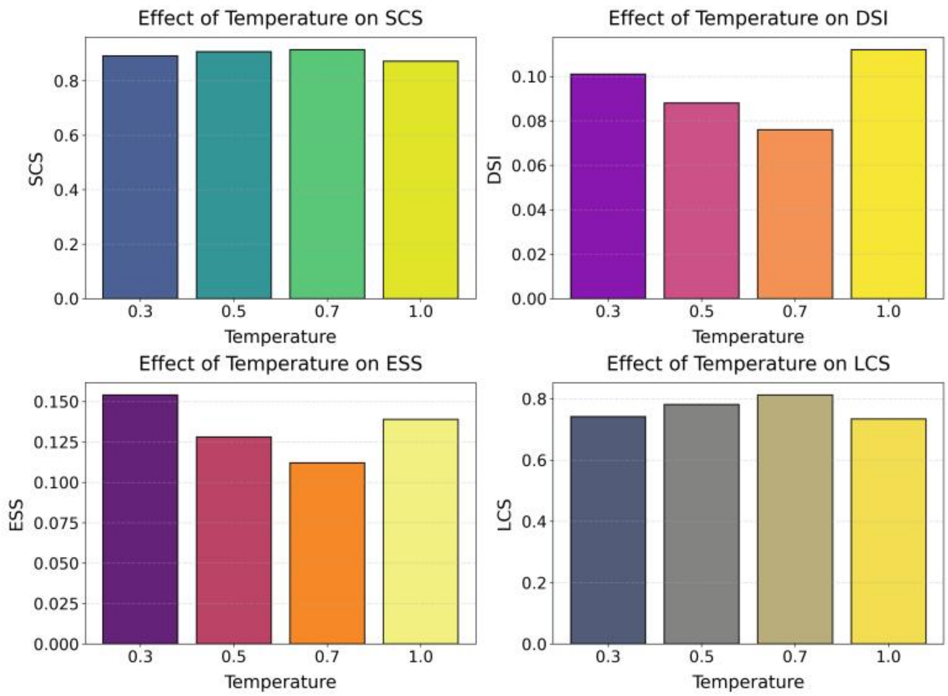

Figure 4. The effect of different inference temperature settings on experimental results

Overall, the influence of inference temperature on the model's generation behavior shows a consistent trend. A moderate temperature produces the best semantic stability and distribution control. The SCS curve shows that semantic consistency decreases at both low and high temperatures and reaches its peak in the middle range. This indicates that moderate randomness helps the model avoid semantic rigidity caused by excessive determinism. It also avoids the semantic drift introduced by high randomness. As a result, the constraint effect of the semantic energy field is expressed most effectively at this temperature level.

In the DSI metric, a moderate temperature also results in the smallest distribution shift. This indicates that the model's output probability distribution is closest to the stable region guided by the energy field. A low temperature makes the model rely too heavily on dominant probabilities and compresses the latent distribution into a narrow region. A high temperature expands the distribution and increases sampling noise, which weakens the control of the semantic energy function. A moderate temperature maintains structural balance in the probability distribution and allows the distribution control mechanism to operate effectively.

The ESS metric reaches its lowest value at a mid-range temperature. This further confirms that the semantic trajectory is most stable within this range. The interaction between the smoothed energy path and the temperature factor guides the model along the lowest-energy direction in latent space. When the temperature deviates from this range, the semantic trajectory becomes unstable. It either shows more jumps due to excessive compression or exhibits energy disturbances due to strong noise. This leads to an increase in ESS and aligns with the stability assumption of the semantic energy framework.

The LCS metric also shows its best performance at a moderate temperature. This indicates that the model maintains stronger long-range semantic coherence under this setting. A moderate temperature allows necessary semantic variation during generation while preventing the divergence effects caused by high temperatures. This produces a smoother and more stable logical chain at the global level. These results demonstrate the complementary relationship between the semantic energy function and temperature tuning. The energy field provides structural constraints, while temperature tuning maintains sampling balance. Together, they form a generation framework with high stability and high consistency.

This paper also presents the impact of data repetition rate on the experimental results, and the experimental results are shown in Figure 5.

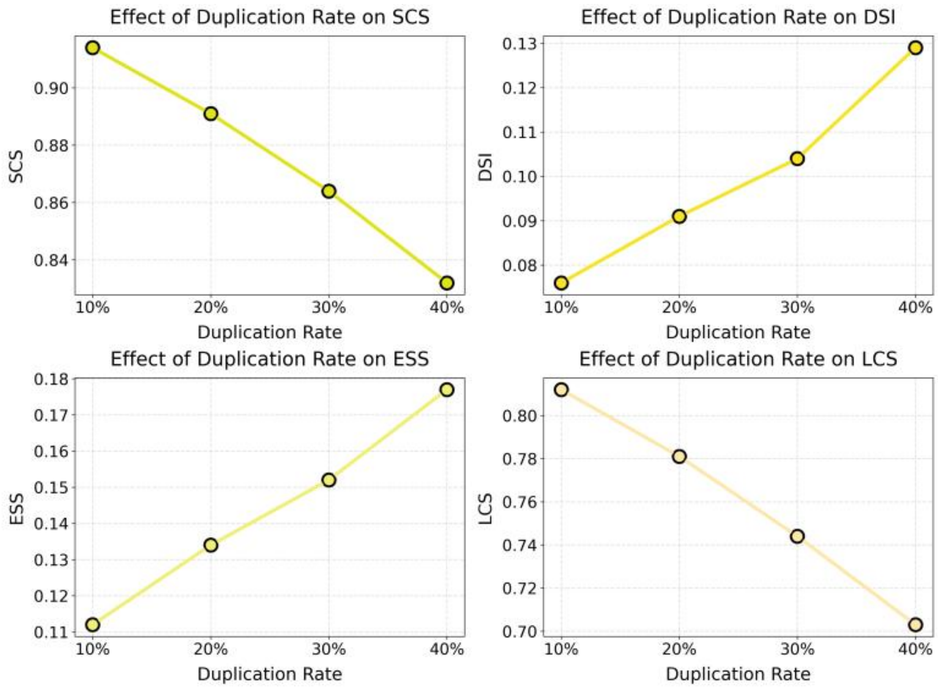

Figure 5. The impact of data repetition rate on experimental results

From the overall trend, increasing data repetition significantly weakens the model's stability and expressive ability under semantic energy constraints. The SCS metric decreases steadily as repetition rises. This indicates that redundant data causes the model to rely too heavily on high-frequency patterns during training. It reduces the guiding effect of the semantic energy field on latent trajectories. When repeated content becomes prevalent, the learned semantic structure becomes more uniform. As a result, the model struggles to maintain sufficient semantic diversity and consistency in open-domain generation.

The continuous increase in the DSI curve further reveals the negative impact of data repetition on distribution control. Prolonged training on repeated data reinforces specific probability patterns. This shifts the output distribution away from the stable region defined by the semantic energy function. As repetition increases, distribution drift becomes more prominent during generation. The model is more likely to fall into narrow probability peaks. This reduces the ability of the energy field to regulate the output distribution. The phenomenon suggests that highly repetitive training data interferes with the distribution-guidance mechanism and weakens the model's ability to maintain stable generative behavior at the probability level.

In the ESS metric, higher repetition leads to more pronounced energy fluctuations. This indicates that the latent semantic trajectory becomes less stable. Repetitive data does not provide enough semantic gradients. The model's representation space develops a pattern of contraction and sudden jumps. It becomes overly stable in high-frequency regions and abruptly unstable in low-frequency regions. This trajectory pattern creates local oscillations in energy values. It prevents the model from maintaining a smooth latent path. This contradicts the smooth potential landscape that the semantic energy function intends to establish. The increase in ESS shows that repetition reduces the model's ability to evolve continuously in semantic space and lowers the effectiveness of energy constraints.

The decline in LCS shows the destructive effect of data repetition on long-range semantic coherence. Long-text generation depends on the hierarchical structure of the latent semantic space. Repetitive data flattens this structure and weakens cross-paragraph and cross-sentence associations. As repetition increases, the model finds it harder to maintain a globally consistent semantic logic chain. This shows that the global potential field in the semantic energy structure cannot fully exert its influence. Therefore, the experiment demonstrates that high repetition in training data weakens the advantages of the semantic energy framework during stable fine-tuning. It significantly affects distribution control, trajectory smoothing, and long-range coherence.

6. Conclusion

This study proposes a stable fine-tuning method based on a semantic energy function to address key challenges in open-ended generation, including semantic drift, distribution instability, and insufficient long-range coherence. By constructing a continuous and differentiable semantic energy field in the latent space, the generation trajectory is guided to evolve along low-energy paths. This leads to clearer semantic boundaries, more robust distribution control, and smoother semantic evolution during output. The study provides a new and generalizable framework for controllable generation in large language models. The model can learn not only surface patterns of content but also maintain stable and consistent semantic trajectories within the latent geometric structure.

At the technical level, the method integrates semantic energy construction, energy regularization, stable fine-tuning, and distribution guidance into an end-to-end mechanism. The process is interpretable and controllable. Experimental analyses show consistent advantages in semantic consistency, distribution stability, energy fluctuation, and long-range coherence. These findings demonstrate that the semantic energy function can act as more than an auxiliary loss. It can deeply influence the dynamic process of language generation and enable stronger control and robustness in complex linguistic environments. The progressive validation highlights the theoretical value and broad applicability of the semantic energy framework in stabilizing large language model behavior.

From an application perspective, large language models with semantic energy constraints offer significant benefits in many domains. In content generation tasks, the method reduces semantic deviation and improves textual coherence and logic. In task-oriented dialogue, technical document generation, and legal or medical applications, it reduces risks caused by distribution shifts and enhances stability and reliability. In knowledge-intensive reasoning tasks, the method supports clearer reasoning paths and more stable semantic chains. These capabilities make semantic-energy-based models suitable as core components in critical systems and provide a safer and more robust foundation for intelligent applications across industries.

Looking ahead, the semantic energy framework offers broad opportunities for further development. The construction of semantic fields, the parameterization of energy functions, and the design of energy-gradient guidance strategies can be refined to enhance controllability across tasks. The framework can also be extended to multimodal, cross-lingual, or cross-domain settings, allowing semantic energy to operate within richer knowledge spaces. With stronger structured optimization, adaptive energy adjustment, and dynamic energy calibration during generation, the method may support large language models that achieve higher stability, safety, and interpretability. It provides a solid theoretical and technical foundation for the reliable deployment of future generative artificial intelligence systems.

References

[1] A. Bakhtin, Y. Deng, S. Gross et al., "Residual Energy-Based Models for Text," Journal of Machine Learning Research, vol. 22, no. 40, pp. 1-41, 2021.

[2] F. Mireshghallah, K. Goyal and T. Berg-Kirkpatrick, "Mix and Match: Learning-Free Controllable Text Generation Using Energy Language Models," arXiv preprint arXiv:2203.13299, 2022.

[3] X. Liu, M. Khalifa and L. Wang, "BOLT: Fast Energy-Based Controlled Text Generation with Tunable Biases," arXiv preprint arXiv:2305.12018, 2023.

[4] M. Xu, T. Geffner, K. Kreis et al., "Energy-Based Diffusion Language Models for Text Generation," arXiv preprint arXiv:2410.21357, 2024.

[5] X. Liang, H. Wang, Y. Wang et al., "Controllable Text Generation for Large Language Models: A Survey," arXiv preprint arXiv:2408.12599, 2024.

[6] J. Wang, C. Zhang, D. Zhang et al., "A Recent Survey on Controllable Text Generation: A Causal Perspective," Fundamental Research, 2024.

[7] W. Lu, R. K. Luu and M. J. Buehler, "Fine-Tuning Large Language Models for Domain Adaptation: Exploration of Training Strategies, Scaling, Model Merging and Synergistic Capabilities," npj Computational Materials, vol. 11, no. 1, p. 84, 2025.

[8] S. K. Pentyala, Z. Wang, B. Bi et al., "PAFT: A Parallel Training Paradigm for Effective LLM Fine-Tuning," arXiv preprint arXiv:2406.17923, 2024.

[9] M. Kowsher, N. J. Prottasha and P. Bhat, "Propulsion: Steering LLM with Tiny Fine-Tuning," Proceedings of the 31st International Conference on Computational Linguistics, pp. 7569-7597, 2025.

[10] L. Gao, A. Ziashahabi, Y. Niu et al., "MobiZO: Enabling Efficient LLM Fine-Tuning at the Edge via Inference Engines," Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 20217-20234, 2025.

[11] D. Gao, Y. Ma, S. Liu et al., "FashionGPT: LLM Instruction Fine-Tuning with Multiple LoRA-Adapter Fusion," Knowledge-Based Systems, vol. 299, p. 112043, 2024.