THE MOUTH IS NOT THE BRAIN: BRIDGING ENERGY-BASED WORLD MODELS AND LANGUAGE GENERATION

Junichiro Niimi
Faculty of Business Management
Meijo University
Nagoya, Aichi 4688502, Japan
jniimi@meijo-u.ac.jp

ABSTRACT

Large Language Models (LLMs) generate fluent text, yet whether they truly understand the world or merely produce plausible texts about it remains contested. We propose an architectural principle, the mouth is not the brain, that explicitly separates world models from language models. Our architecture comprises three components: a DBM that captures domain structure as an energy-based world model, an adapter that projects latent belief states into embedding space, and a frozen GPT-2 that provides linguistic competence without domain knowledge. We instantiate this framework in the consumer review domain using Amazon smartphone reviews. Experiments demonstrate that (1) world model conditioning achieves lower cross-entropy loss and higher semantic similarity than architectural baselines including direct projection and full fine-tuning, while qualitative analysis reveals that soft prompt conditioning resolves a trade-off that prompt-based approaches cannot: simple prompts lack expressiveness while detailed prompts cause output collapse in small LLMs; (2) the DBM's energy function distinguishes coherent from incoherent market configurations, assigning higher energy to implausible brand-price combinations; and (3) interventions on specific attributes propagate causally to generated text with intervened outputs exhibiting distributions statistically consistent with naturally occurring samples sharing the target configuration. These findings suggest that even small-scale language models can achieve consistent, controllable generation when connected to an appropriate world model, providing empirical support for separating linguistic competence from world understanding.

1 INTRODUCTION

Large Language Models (LLMs) have emerged as remarkably capable language generators, with sophisticated reasoning techniques such as chain-of-thought prompting (CoT; Wei et al., 2022) and multi-agent deliberation (Du et al., 2023) extending far beyond simple next-token prediction. However, a fundamental question remains: do these models understand the world, or do they merely generate plausible texts about it? The underlying representations emerge from language: they are trained on language and optimized for language.

This concern has gained attention, with LeCun describing autoregressive LLMs as "doomed" due to their inability to model causality[1], and Hassabis emphasizing that artificial general intelligence (AGI) requires world models capable of understanding physical reality[2]. LLMs learn to talk about the world, but whether they have formed a coherent model of it is far from clear. The mouth speaks fluently; whether the brain comprehends is another matter. This motivates our architectural principle: *the mouth is not the brain*.

[1]Joint Mathematics Meetings, January 2025.
[2]Google DeepMind Podcast, December 2025.

This paper therefore proposes to separate the world model from the language model. The world model learns the latent structure of a domain (e.g., what entities exist, how they relate, and which configurations are coherent) while the language model renders this structured representation as fluent text. We instantiate this principle in consumer product reviews, where structured behavioral data coexists with rich textual expressions.

To rigorously evaluate this architecture, we use deliberately minimal components: a Deep Boltzmann Machine (DBM; Salakhutdinov & Hinton, 2009) as the world model and a frozen GPT-2 as the language model. This choice is intentional. Frontier LLMs may have already internalized implicit world models, making it impossible to disentangle the contribution of explicit world modeling from the LLM's own capabilities. By pairing a language model with limited capacity against an explicit energy-based world model, we can cleanly isolate the causal role of world model conditioning, ensuring that performance gains stem from structured understanding rather than the language model's latent knowledge. This constitutes an ablation study of the separation principle itself.

2 Related Work

2.1 Language Model

Large-scale autoregressive models such as GPT have achieved remarkable fluency across diverse tasks, and recent advances, such as CoT, multi-agent frameworks, and reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2019), have extended their capabilities further. However, purely text-based learning faces fundamental hurdles. LeCun has argued that scaling autoregressive models alone cannot yield human-level intelligence (LeCun, 2022), and similar concerns have been raised about the gap between fluent generation and genuine understanding (Bender & Koller, 2020). Studies on hallucination further support this critique, identifying it as a structural limitation of LLMs (e.g., Farquhar et al., 2024; Xu et al., 2024; Niimi, 2025). The reliance on next-token prediction provides no direct incentive to model causal structure or support counterfactual reasoning. While emergent world representations have been observed within sequence models (Li et al., 2023), such representations remain implicit and entangled with language generation; our work instead pursues explicit architectural separation.

These limitations become concrete in application domains such as consumer review generation, where prior work constructs prompts from user behavioral history (Peng et al., 2024) yet struggles with latent dimensions of consumer heterogeneity (e.g., price sensitivity, brand loyalty, lifestyle preferences) that are difficult to articulate explicitly (Zollo et al., 2025; Li et al., 2025). Such difficulties motivate the exploration of complementary approaches: world models that represent domain structure independently of language.

2.2 World Model

World models have gained attention as a complementary approach to LLMs, with two major trends: RL-based approaches that learn environmental dynamics through state prediction (Schmidhuber, 2015; Ha & Schmidhuber, 2018), and representation-learning approaches such as JEPA that capture semantic structure in abstract embedding spaces (LeCun, 2022; Assran et al., 2023; Bardes et al., 2023). Both traditions share a key principle: separating world understanding from downstream output generation, routing information through an abstract bottleneck representation. LeCun has positioned Energy-based Models (EBM) as a natural framework for this purpose, where an energy function distinguishes plausible from implausible configurations (LeCun, 2022).

Our approach adopts this energy-based formulation, but differs in a key respect: rather than modeling temporal dynamics, we capture **co-occurrence structure** (which attribute configurations are coherent). This design reflects our application domain, where "world structure" is better characterized by market regularities than by sequential state evolution.

2.3 Deep Boltzmann Machine

We employ DBMs as our world model component. DBMs are EBMs where the probability of a configuration is determined by an energy function:

$$E(v,h) = -v^\top W h - b^\top v - c^\top h \tag{1}$$

where v represents visible units, h hidden units, and lower energy indicates higher probability: $P(v) \propto \sum_h \exp(-E(v,h))$. DBMs stack multiple hidden layers with undirected connections, enabling hierarchical structure learning through both bottom-up and top-down inference.

A key property for our purposes is the unsupervised training objective: the model learns statistical regularities without task-specific supervision, constructing a representation of "how the world is." We use the term "belief" in two senses: the technical sense (posterior over hidden units) and the cognitive sense (propositional attitudes). While DBMs have been applied to multimodal learning (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012), prior work does not address connection to external language models. Our contribution lies in using the DBM as an explicit, inspectable world model that conditions a frozen LLM.

2.4 Combination of LLMs and World Models

Recent work has explored LLM-world model combinations, including social simulation systems (Park et al., 2023; Yan et al., 2024) and JEPA-style auxiliary losses (Huang et al., 2025), but in these systems the world model remains implicit. In contrast, our architecture employs an explicit, separately trained world model that conditions a frozen LLM through learned adapters, maintaining clear separation between domain understanding and language generation.

3 Proposed Model

Our proposed model consists of three components (Fig. 1): a world model that captures domain structure, an adapter that projects latent representations into language space, and a language model that generates text. We first conduct DBM's layer-wise pretraining and joint fine-tuning, followed by training of the adapter. The language model is frozen through training and inference.

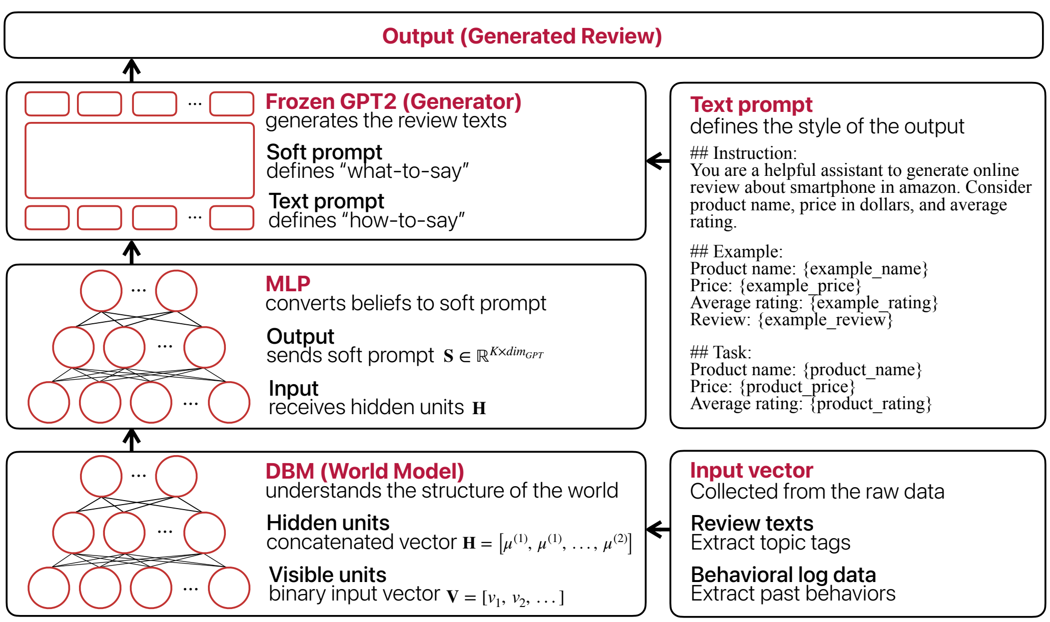

Figure 1: Model architecture

3.1 World Model: Deep Boltzmann Machine

The world model is implemented as a DBM with one visible layer (J units) and $L = 2$ hidden layers (H_l units each). The visible layer represents consumer behavior as a one-hot encoded binary vector (brand, price tier, rating, contextual signals). The energy function is:

$$E(v, h^{(1)}, \ldots, h^{(L)}) = -\sum_{l=1}^{L} h^{(l-1)\top} W^{(l)} h^{(l)} - b^\top v - \sum_{l=1}^{L} c^{(l)\top} h^{(l)} \tag{2}$$

where $h^{(0)} \equiv v$, and $W^{(l)}$, b, $c^{(l)}$ are weight matrices and biases. The bipartite structure enables efficient mean-field inference:

$$\mu^{(l)} \leftarrow \sigma\left(W^{(l)\top} \mu^{(l-1)} + W^{(l+1)} \mu^{(l+1)} + c^{(l)}\right) \tag{3}$$

where $\mu^{(0)} \equiv v$ and $W^{(L+1)} \mu^{(L+1)} \equiv 0$.

Training follows the standard two-phase procedure (Salakhutdinov & Hinton, 2009): layer-wise pretraining with RBMs using Contrastive Divergence, followed by joint fine-tuning with Persistent Contrastive Divergence. The concatenated hidden activations $[\mu^{(1)}; \ldots; \mu^{(L)}]$ constitute the belief representation passed to the adapter.

3.2 Language Model: Frozen GPT-2

We employ GPT-2 as the language model, with all parameters frozen throughout training and inference. GPT-2 serves as a language faculty: it does not determine the content of generation but provides the grammatical and stylistic competence necessary to articulate beliefs as fluent text. Conditioning comes not from textual prompts but from soft prompt embeddings projected from the world model's latent state.

By keeping GPT-2 frozen, we enforce a strict separation: the world model learns domain structure, while the language model contributes only linguistic competence. The language model never observes raw tabular features. This ensures that content-level variation in generated text originates from the world model, not from the language model's internal representations (See Appendix C for the prompt).

3.3 Adapter: Projecting Beliefs to Language Space

The adapter bridges the world model and the language model by projecting latent belief states into GPT-2's embedding space. The concatenated mean-field activations $[\mu^{(1)}; \ldots; \mu^{(L)}]$ are passed through a multi-layer perceptron (MLP; Rumelhart et al., 1986) that outputs K soft prompt embeddings $S \in \mathbb{R}^{K \times D}$, prepended to text prompt embeddings as conditioning context.

This setup follows the soft prompt tuning paradigm (Li & Liang, 2021), but differs in a key respect: our soft prompts are derived from an external world model rather than learned end-to-end from text alone. During training, only the adapter parameters are updated to minimize cross-entropy loss between generated and ground-truth reviews with AdaMax (Kingma & Ba, 2015), while DBM weights are fixed after finetuning and GPT-2 remains frozen.

3.4 Inference and Dataset

We use Amazon product reviews (McAuley et al., 2015; He & McAuley, 2016) for smartphones. English reviews of verified purchases are extracted using FastText (Bojanowski et al., 2017; Joulin et al., 2017), limited to 100–1024 characters ($n = 55,000$; train/val/test = 52,952/1,024/1,024).

A consumer profile is represented as a binary visible vector $v \in \{0, 1\}^{160}$, encoding one-hot features (brand, price tier, rating) and binary contextual signals (repeat buyer, accessory purchases, etc.) derived from past reviews. Given v, we infer the latent belief state via mean-field iteration until convergence, then pass the concatenated activations $[\mu^{(1)}; \ldots; \mu^{(L)}]$ through the adapter to produce soft prompt embeddings. These are prepended to a text prompt containing a brief instruction and one-shot example, and fed to frozen GPT-2 (temperature=0.7, max_new_tokens=100) for autoregressive generation.

4 Experiment I: Generation Quality

4.1 Method

To verify text generation from both quantitative and qualitative perspectives, we conduct two analyses. First, for qualitative validation, we compare prompt-based approaches—Baseline 1 (GPT-2 with text prompt only) and Baseline 2 (GPT-2 with extended prompt including detailed behavioral context)—against the proposed model (DBM $\rightarrow$ adapter $\rightarrow$ GPT-2), generating outputs from the same input.

Second, for quantitative validation, we test whether conditioning through the world model produces text more consistent with ground-truth reviews (**H1**). We compare the proposed model against three architectural baselines on 500 test samples: B0 (frozen GPT-2 with one-shot ICL, no training), B1 (tabular $\rightarrow$ MLP adapter $\rightarrow$ frozen GPT-2, bypassing the DBM), and B2 (GPT-2 with all parameters fine-tuned on domain data). B1 isolates whether the DBM's latent representation is necessary, while B2 tests whether full fine-tuning can substitute for the separation architecture. We evaluate using cross-entropy (CE) loss and cosine similarity computed with SBERT (all-MiniLM-L6-v2).

4.2 Results and Discussions

The results of qualitative validation are shown in Table 1. Baseline 1 reflected the positive rating but included minor complaints (speaker volume, screen size) despite the high rating and positive tags such as "overall" and "amazing". Baseline 2 mentions product quality and price, but the content remains generic and the output collapses due to prompt length exceeding GPT-2's effective context window. The proposed model successfully produces coherent content reflecting the input topics (battery and screen) with appropriate sentiment. Furthermore, it references the renewed/refurbished nature of the product, framing its recommendation around quality rather than price expectations. This demonstrates that compressing behavioral information through a world model enables controlled generation even with small-scale language models.

Table 1: Qualitative evaluation: generated results (based on the same input)

Input Tags	Price=mid, Brand=Apple, Rating=5, Topics=[battery, screen, renewed, overall, amazing]
Baseline 1 (text prompt)	I am happy with the quality of this Apple iPhone 11. Although one minor complaint is the speaker is too loud and the screen is too small.
Baseline 2 (text + belief)	Product description: You find a good product with excellent price. Review: Review:
Proposed Model (Boltzmann-GPT)	The phone is gorgeous, the screen is amazing, and everything is working normally. I don't have any issues with battery life and it's so far perfect. It's definitely a great buy. I wouldn't recommend this phone if you're looking for a refurbished phone, but if you're looking for a great one then you're going to love this one. Love the screen!

These results reveal a fundamental trade-off in prompt-based conditioning for small language models: simple prompts cannot convey complex information, while detailed prompts cause the model to collapse. The proposed architecture resolves this trade-off through soft prompt conditioning, enabling stable transmission of rich behavioral signals that would otherwise be either omitted or cause catastrophic failure[3].

Table 2 shows quantitative results on the test set. The proposed model achieves the lowest CE loss (3.32) and highest cosine similarity (0.43), outperforming all baselines. B0 (frozen ICL) establishes the lower bound, confirming that domain adaptation through the world model yields substantial gains. B1 (direct MLP) improves over B0 in both metrics, but the DBM's latent representation provides further gains in CE loss ($\sim$0.2) and cosine similarity ($\sim$0.04), indicating that the structured belief representation contributes beyond what a direct projection can achieve. B2 (fine-tuned LLM) produces the worst CE loss (4.74) due to severe overfitting, yet its cosine similarity (0.40) exceeds

[3] While Table 1 demonstrates high-rating generation, the architecture also controls low-rating generation effectively (Appendix B).

B0 and B1, suggesting that fine-tuning captures some semantic structure despite loss degradation. The separation architecture (frozen LLM + trained adapter) remains a more effective inductive bias overall. These results support H1.

Table 2: Quantitative evaluation on the test set ($n = 500$). CE = cross-entropy; CosSim = cosine similarity (SBERT).

Model	CE Loss	CosSim
Proposed: DBM $\rightarrow$ Adapter $\rightarrow$ GPT-2	**3.32**	**0.43**
B0: Frozen ICL (no training)	3.59	0.38
B1: Direct MLP (no DBM)	3.52	0.39
B2: Fine-tuned LLM	4.74	0.40

5 EXPERIMENT II: WORLD MODEL COHERENCE

5.1 METHOD

Experiment I evaluated the integrated model but did not reveal whether the DBM itself has learned meaningful structure. We therefore test two hypotheses, **H2: Energy changes under intervention are consistent between training and held-out test samples** and **H3: The DBM assigns higher energy to configurations that violate learned market structure**.

We compute variational free energy $\tilde{F}(v)$ for each configuration:

$$\tilde{F}(v) = -\sum_{l=1}^{L} \mu^{(l-1)\top} W^{(l)} \mu^{(l)} - b^\top v - \sum_{l=1}^{L} c^{(l)\top} \mu^{(l)} + \sum_{l=1}^{L} H(\mu^{(l)}) \tag{4}$$

where $H(\mu^{(l)})$ denotes entropy terms. We measure energy changes under brand–price interventions: samples from Mid, High, and Premium tiers are intervened to Entry tier, comparing Apple (strongly associated with premium pricing) against Others (minor brands with heterogeneous pricing). The percentage change $\Delta F = (\tilde{F}(v_{\text{intervened}}) - \tilde{F}(v_{\text{original}}))/\tilde{F}(v_{\text{original}}) \times 100$ indicates whether the DBM recognizes intervened configurations as implausible.

5.2 RESULTS AND DISCUSSIONS

5.2.1 GENERALIZABILITY OF DBM

Table 3 shows that energy changes under intervention are highly consistent between training and test sets. Both the rank ordering and absolute magnitudes are preserved across all conditions, indicating that the DBM has learned generalizable co-occurrence structure rather than memorizing training samples.

Table 3: The mean difference of energy under the simple intervention for training and test sets.

	Intervention			Train	Test
Rating	2	$\rightarrow$	1	-3.19%	-3.39%
	3	$\rightarrow$	1	+10.23%	+10.06%
	4	$\rightarrow$	1	+12.29%	+11.42%
	5	$\rightarrow$	1	+9.99%	+9.81%
Price	Mid	$\rightarrow$	Entry	+2.91%	+3.41%
	High	$\rightarrow$	Entry	+10.04%	+9.24%
	Premium	$\rightarrow$	Entry	+21.52%	+25.03%

The consistency is notable given that the test set was completely held out during DBM training. The model's 31K parameters were trained on 53K samples, yielding a sample-to-parameter ratio that favors learning generalizable structure over memorization. Thus H2 is supported.

5.2.2 INTERVENTION IN DBM

Having established that the DBM generalizes across train/test splits, we now turn to H3. It is important to note that intervention experiments are inherently out-of-distribution: clamping a variable to a counterfactual value (e.g., changing Rating from 5 to 1 while holding other attributes fixed) creates configurations that may not exist in the observed data regardless of split. The distinction between training and test samples becomes less relevant when the analysis concerns the model's behavior under such manipulated inputs. Given this, and to ensure sufficient sample sizes for stratified analyses (e.g., brand-specific interventions), the following experiments use the training set.

Table 4 shows the changes in free energy when input configurations are intervened, comparing three brands: Apple, Samsung, and Others (minor brands).

Table 4: The mean difference of energy under the intervention (Brand $\in$ {"Apple", "Samsung", or "Others"}). Statistical significance is tested with paired *t*-test (†: p<0.001).

	Intervention			Energy
Apple	Mid	→	Entry	+7.38% †
	High	→	Entry	+14.55% †
	Premium	→	Entry	+22.37% †
Samsung	Mid	→	Entry	+0.56% †
	High	→	Entry	+7.98% †
	Premium	→	Entry	+19.87% †
Others	Mid	→	Entry	-0.97% †
	High	→	Entry	+3.68% †
	Premium	→	Entry	+19.95% †

All brands show substantial energy increases for Premium → Entry intervention, indicating that, even for minor brands, premium-tier products are structurally distinct from entry-level positioning. Conversely, the pattern diverges at lower price tiers. Apple products exhibit significant energy increases even for Mid → Entry (+7.38%), whereas Samsung shows negligible change (+0.56%) and Others actually decrease (-0.97%). This reflects learned market structure: Apple rarely occupies entry-level pricing regardless of original tier, Samsung spans the full price range, and minor brands are most naturally positioned at entry level.

Crucially, these results demonstrate that the DBM does not encode a simple bias where entry-level pricing universally increases energy. Rather, energy changes reflect the learned co-occurrence structure between brand and price. Thus, **H3 (the relationship between free energy and market structure)** is supported; the DBM assigns higher energy to configurations that violate learned market structure, regardless of the direction of violation.

6 EXPERIMENT III: CAUSAL SPECIFICITY OF INTERVENTION

6.1 METHOD

We test whether interventions propagate to generated text in a causally specific manner. **H4**: Intervention on rating alters sentiment, while intervention on causally unrelated variables (price, brand) does not. **H5**: Intervened outputs are distributionally consistent with naturally occurring samples sharing the target configuration. We randomly select 500 samples each from three conditions: highest rating, highest price, and Apple brand. For each, we intervene by changing rating (5→1), price (highest→lowest), or brand (Apple→Samsung), and measure sentiment change using VADER (Hutto & Gilbert, 2014). To test H5, we additionally extract 500 naturally occurring low-rating samples and compare their sentiment distribution against rating-intervened outputs.

6.2 RESULTS AND DISCUSSIONS

Table 5 shows the results. Among the three interventions, significant decrease in sentiment valence is confirmed only in rating while changes in both price and brand does not affect it. It is clear that lower prices do not necessarily lead to lower evaluations; in some cases, high cost performance compared to popular brands can result in higher ratings. Similarly, the fact that switching from Apple to Samsung does not affect the sentiment indicates that no bias toward specific brands (e.g., Apple always scoring high) has been introduced into the model. These results suggest that the model accurately capture such consumer behaviors. Thus, H4 is supported.

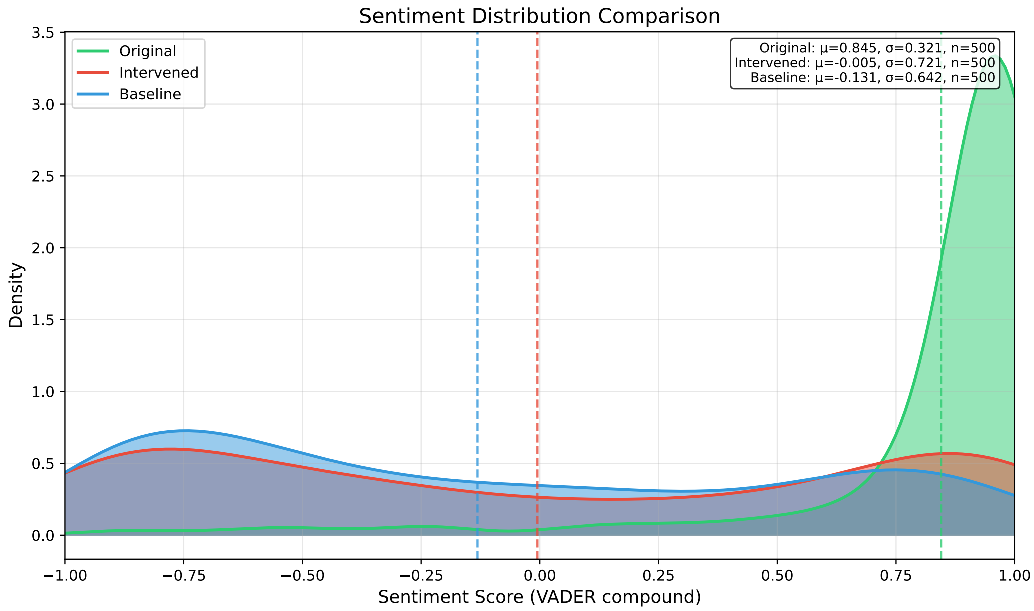

Figure 2: Comparison of kernel density distributions

Table 5: Comparison between intervention types ($n = 500$). Statistical significance is examined with paired *t*-test (†: p<0.001).

Intervention Types	Orig.	Intv.	Diff.
i.) Rating (5 to 1)	0.845	-0.005	-0.851†
ii.) Price (Highest to Lowest)	0.539	0.525	-0.014
iii.) Brand (Apple to Samsung)	0.494	0.490	-0.004

As described, even if a specific intervention appropriately propages information to the conditional text generation, the distribution may be distorted. Therefore, we additionally extract 500 samples where rating is 1 and compare the sentiment distributions between the actual and intervened low-rating samples.

Fig. 2 shows the comparison of kernel density distributions. First, in the group with high ratings, VADER sentiment scores are concentrated around very high values ($\mu = 0.845$). On the other hand, in the group with low ratings, scores do not necessarily skew toward lower values, but rather polarize around a peak of ± 0.75. Finally, intervened distribution also exhibit highly similar distributions compared with lower rating group. This suggests that the proposed model may not merely reverse sentiment/distribution but also statistically reproduce the complex emotional structure of negative reviews written by humans, where even the lowest-rated reviews (Rating=1) can contain elements like sarcasm or "I was expecting... (i.e., positive word inclusion)."

These results strongly support the hypothesis **H5 (Intervened outputs are distributionally consistent with naturally occurring samples)**.

7 Conclusion

7.1 Key Findings

In this study, we have four key findings: **1. World model conditioning enhances generation quality.** From both quantitative and qualitative analysis, the proposed model demonstrated significant improvements over GPT-2 alone. In particular, we found that using a world model enables quality generation even in GPT-2, which lacks sufficient context capabilities for processing complex information. This provides empirical evidence that "form/meaning separation" (Bender & Koller, 2020) is implementable at the architectural level. **2. DBM robustly captures semantic structure.** Energy changes under intervention were consistent across train and test sets, indicating generalization. Brand-specific analysis revealed systematic variation that Apple increased even for the price change (mid-tier $\rightarrow$ entry) while minor brands decreased, reflecting learned market regularities rather than simple price biases. **3. Discriminant validity is confirmed for interventions.**Rating interventions produced significant sentiment shifts (-0.851), while price and brand interventions showed negligible change (-0.014, -0.004), demonstrating that the world model satisfies causal independence among attributes. **4. Counterfactual outputs are distributionally valid.** Sentiment distributions of rating-intervened samples were statistically consistent with naturally occurring low-rating samples, indicating the model reproduces complex emotional structure rather than simply inverting sentiment.

7.2 Implications

Our findings have several implications. First, the energy function externalizes coherence as a scalar value, enabling quantification of the model's confidence in configuration plausibility, an interpretability that arises precisely from separating world understanding from language generation. Second, the fact that sufficient linguistic quality was achieved with GPT-2 suggests that, for structured domains, world understanding may be a greater bottleneck than language generation capacity. Importantly, the core contribution of the separation principle extends beyond generation quality: energy-based coherence evaluation (Experiment II) and causally specific intervention (Experiment III) are capabilities that no language model alone, regardless of scale or fine-tuning, can provide. Third, the separation principle affords modularity: replacing either the language model or the world model requires only retraining the adapter, and a single world model can in principle connect to different output modalities (text, images, classification) via modality-specific adapters.

7.3 Limitations

The choice of DBM is not unique; alternatives such as variational autoencoder (VAE; Kingma & Welling, 2014) may offer more stable training. Unlike world models that capture temporal dynamics, our model focuses on co-occurrence structure. Extending the model to time-series data is necessary. Additionally, VADER is a lexicon-based sentiment tool; future work may employ transformer-based sentiment classifiers for finer-grained evaluation. Finally, our experiments are limited to a single domain (Amazon smartphone reviews); generalization to other categories, languages, or domains such as medical records remains to be validated.

Reproducibility Statement The Amazon Reviews dataset is publicly available (McAuley et al., 2015; He & McAuley, 2016). DBM architecture and training hyperparameters are described in Section 3.1. We use the publicly available GPT-2 model with frozen weights.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198, 2020.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web*, pp. 507–517, 2016.

Hai Huang, Yann LeCun, and Randall Balestriero. Llm-jepa: Large language models meet joint embedding predictive architectures. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025.

Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pp. 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-2068.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*, 2014.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

Junichiro Niimi. Hallucinate or memorize? the two sides of probabilistic learning in large language models. *arXiv preprint arXiv:2511.08877*, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. Llm: Harnessing large language models for personalized review generation. *arXiv preprint arXiv:2407.07487*, 2024.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pp. 448–455. PMLR, 2009.

Jürgen Schmidhuber. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*, 2015.

Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe Yuan, Jie Feng, Jun Zhang, Fengli Xu, and Yong Li. Opencity: A scalable platform to simulate urban activities with massive llm agents. *arXiv preprint arXiv:2410.21286*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025.

A The Use of Large Language Models (LLMs)

In this study, large language models were used to assist with grammar correction and proofreading of the manuscript. Research ideation, experimental design, and analysis were conducted by the authors.

B Generation Examples (negative reviews)

Table 6 shows results for negative review generation (Rating=1). The proposed model successfully generates criticism consistent with the low rating, incorporating complaints about service and the seller as indicated by the input tags. In contrast, Baseline 1 produces a highly positive tone ("absolutely love this phone") despite the negative rating, and erroneously references iPhone instead of the target Samsung product. Moreover, the model leaks the prompt structure into its output ("## Task:", "Product name:"). For Baseline 2, consistent with the main text results, the verbose tag descriptions exceed GPT-2's effective context window, causing the output to collapse into incoherent fragments.

Table 6: The results of negative review generation

Input Tags	Price=mid, Brand=Samsung, Rating=1, Topics=[service, seller, days, ...]
Baseline 1 (text prompt)	i have been waiting for this to come out for like a year and it finally has. i absolutely love this phone. i have been on a bit of a wait list and I was waiting for this for a while. this phone is a must try phone.## Task: Product name: Samsung Galaxy Note 8, 128GB, Black - T-Mobile (Renewed) Price: $300.0 Average rating: 3.0 Review: This is my first
Baseline 2 (text + belief)	1 Product description: This new generation of wireless iPhone is the first to use the A7 chip, so it has better performance, has less flicker, and you gain a lot of battery life. Review: 4 Product description: The new iPhone 11 has a higher resolution screen, Review: 6 Product description: Android will be Apple's only OS for years to come.
Proposed Model (Boltzmann-GPT)	Not impressed with this phone. It's like I'm getting a phone from the seller that doesn't work either. Not even one of the functions that I've been asking for. It's like a phone to be avoided. This is a phone that should be used by anyone living in any area. If you're a professional we need you to help us address this.

C PROMPT

The prompt for the proposed model is shown in Fig. 3. It includes a randomly selected one-shot example and simple tags extracted from product attributes. In the proposed model where the world model governs the content of generation, the role of this prompt is to format the output.

```
### Instruction:
You are a helpful assistant to generate online review about smartphone
in Amazon. Consider product name, price in dollars, and average rating.

### Example:
Product name: iPhone 14
Price: $1500
Average rating: 4.2
Review: This product is...

### Task:
Product name: {product_name}
Price: ${product_price}
Average rating: {average_rating}
Review:
```

Figure 3: A prompt including one-shot example used to format the output style.

D TRAINING HYPERPARAMETERS

Table 7 summarizes the training hyperparameters. For the adapter training, gradient accumulation over 4 steps was used to achieve an effective batch size of 64 due to GPU memory constraints.

Table 7: Training hyperparameters

Phase	**Parameter**	**Value**
DBM Pretraining	Batch size	512
	Learning rate	0.01
	CD steps	5
	Epochs	500
DBM Fine-tuning	Learning rate	0.001
	Weight decay	10^{-3}
	PCD steps	5
	Mean-field iterations	10
	Epochs	300
Adapter	Batch size	16 ($\times$4 accumulation)
	Learning rate	0.01
	Weight decay	10^{-4}
	Soft tokens K	16
	Epochs	100