Energy-Based World Models for Controllable Text Generation with Frozen LLMs

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) generate fluent text but, as monolithic systems, lack native mechanisms for fine-grained clamps on individual concepts (e.g., clamping a single attribute such as brand preference while holding everything else fixed). We propose an architectural separation grounded in form/meaning distinction: a frozen LLM provides linguistic competence (form), while a learned energy-based world model captures the latent belief structure (meaning) on which clamps act. Concretely, a Deep Boltzmann Machine (DBM) models the domain's latent co-occurrence structure, a lightweight adapter projects its mean-field beliefs into soft prompts, and a frozen Qwen2.5-0.5B-Instruct renders them as text. This split yields a falsifiable prediction: a form-optimized direct-MLP (No DBM) baseline should match or marginally beat the proposed model on form-level metrics (CE, SBERT), while meaning-level capabilities, including energy-based coherence scoring, predictively-valid clamping, and independent two-axis attribute control, should be exclusive to the DBM-based pipeline. We evaluate this on two Amazon review domains (Smartphones, Beauty) under a rigorous 3-seed protocol. On form metrics, the proposed model statistically dominates LoRA, full fine-tuning, CTRL, PPLM, and one-shot ICL in both domains ($p < 10^{-5}$) and ties direct-MLP as predicted. On meaning-level metrics, it outperforms direct-MLP on every cross-domain $\times$ cross-metric cell: energy-based scoring recovers each domain's market structure while returning near-null scores on control pairs; coupling-aware clamping cuts Wasserstein distance to natural low-rating distributions by 23% on Beauty; and two-axis clamps yield 1.6–3.4 times larger content shifts than direct-MLP with sentiment held fixed. A generously tuned activation-steering baseline (CAA) matches the raw sentiment shift but is rejected on distributional match to natural reviews under RoBERTa, composes destructively under multi-attribute clamps, and entangles sentiment when steered on the content axis. Architectural separation thus offers a practical, theoretically grounded path toward controllable, interpretable, and model-consistent counterfactual text generation in structured domains.

1 Introduction

Large Language Models (LLMs) have emerged as remarkably capable language generators. Yet a fundamental question remains: do they truly understand the world, or do they merely generate plausible texts about it? Their representations are shaped exclusively by linguistic data, and as monolithic systems, they do not natively expose mechanisms for fine-grained clamping of individual latent concepts (e.g., what the model would generate if a single attribute such as brand preference were clamped to a different value while holding everything else fixed). All knowledge remains entangled within a single set of parameters (Cloud et al., 2025). Existing conditional generation methods, such as control codes (Keskar et al., 2019) and gradient-based steering (Dathathri et al., 2020), rely on correlational signals rather than structured, inspectable representations of the underlying domain.

The broader question of whether LLMs can genuinely model the world has gained renewed attention. Leading researchers have argued that autoregressive LLMs are structurally limited in their ability to model causality and world structure, and that capable AI systems require explicit world models that predict in abstract representation space rather than token space (LeCun, 2022; Ha & Schmidhuber, 2018). The mouth speaks

fluently; whether the brain comprehends is another matter. This motivates our architectural principle: *the mouth is not the brain.*

We therefore propose a modular architecture that explicitly separates the two: (i) a Deep Boltzmann Machine (DBM; Salakhutdinov & Hinton, 2009) serves as an energy-based world model that captures the domain's latent co-occurrence structure; (ii) a lightweight MLP adapter projects the DBM's mean-field beliefs into soft prompt embeddings; (iii) a frozen modern LLM provides linguistic competence, receiving only the minimal textual prompt as all baselines. By externalizing domain knowledge into a dedicated, inspectable world model, this separation naturally yields two critical properties absent from standard monolithic adaptation: **interpretability** (the system's beliefs can be directly examined via the energy-based model) and **intervenability** (specific latent factors can be selectively clamped to generate model-consistent counterfactuals).

The position we advance. For domains that exhibit structured latent regularities, **world understanding and language generation should be architecturally separated**: an inspectable energy-based world model should condition a frozen language model through a lightweight adapter. We frame this separation explicitly in Bender & Koller (2020)'s *form/meaning distinction*: **form** (linguistic surface) is the competence of the LLM, while **meaning** (the latent belief structure over domain attributes, on which clamps act) is captured by the world model.

This architectural split yields a clear, testable prediction. A strong form-only baseline that optimizes linguistic quality end-to-end with direct access to visible features (B1: direct-MLP adapter, see §4) is expected, by construction, to match or marginally outperform the proposed model on surface-level (form) metrics such as cross-entropy and SBERT similarity. In contrast, meaning-level capabilities (energy-based coherence scoring, predictively valid attribute clamps, and independent two-axis counterfactual-style control) should be exclusive to the DBM-based pipeline.

We test and confirm this prediction using a deliberately compact 31K-parameter DBM together with an open-weight LLM (Qwen2.5-0.5B-Instruct). Four lines of evidence support the framework. **1. Strong linguistic competence:** under a 3-seed $\times$ n=1024 protocol the proposed model statistically dominates LoRA, full fine-tuning, CTRL (Keskar et al., 2019), PPLM (Dathathri et al., 2020), and one-shot ICL on CE and SBERT in both domains ($p < 10^{-5}$), and ties B1 as the form/meaning split predicts (§5). **2. Coupling discovery validates the world model:** the DBM recovers the strongest empirical couplings per domain while returning near-null scores on control pairs (§6). **3. Independent two-axis attribute control:** cross-domain invariant couplings (e.g., Rating–Service) give a sentiment handle and domain-specific couplings (e.g., Brand–Topic specialty) give a content handle; coupling-aware combined clamps cut Wasserstein distance to natural low-rating distributions on Beauty by 23%, while a generously tuned activation-steering baseline (CAA; Rimsky et al. 2024) is rejected on distributional match under RoBERTa, composes destructively, and breaks two-axis independence on the content axis (§7–§7.4). **4. Meaning-level superiority on every cell:** the DBM-based model beats B1 across all cross-domain meaning-level comparisons; the architectural claim rests on B1's structural inability to expose meaning-level capabilities, not on form-side failure (§8, App. C).

2 Related Work

Modern approaches to model the world structure fall into two complementary lines. RL-inspired methods learn environmental dynamics via state prediction (Schmidhuber, 2015; Ha & Schmidhuber, 2018), typically using an asymmetric architecture in which a small controller operates on abstract representations produced by a larger perception module. Representation-learning approaches such as JEPA (LeCun, 2022; Assran et al., 2023; Bardes et al., 2023) instead predict in abstract embedding space rather than raw inputs. LeCun (2022) explicitly advocates energy-based models (EBM; LeCun et al., 2006) as a natural framework, where low energy corresponds to plausible configurations and high energy to implausible ones. We adopt the energy-based formulation but focus on modeling **static co-occurrence market structure**, i.e., which attribute configurations are coherent within a given market rather than temporal dynamics.

Among EBMs, Boltzmann machines (BMs; Ackley et al., 1985) learn distributions over binary variables via an energy function. Restricted Boltzmann Machines (RBMs; Smolensky, 1986; Hinton, 2012) remove intra-layer connections to enable efficient mean-field inference and Contrastive Divergence (CD) training. Deep Boltzmann Machines (DBMs; Salakhutdinov & Hinton, 2009) further stack multiple RBMs with undirected connections, supporting bidirectional inference. While DBMs have been applied to multimodal learning (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012), to our knowledge, they have not been used as an explicit, inspectable world model that conditions a frozen LLM. Compared with variational autoencoders (VAEs; Kingma & Welling, 2014), which learn continuous latents via a reconstruction-plus-KL objective, DBMs offer a distinct advantage: they provide a natural energy function that directly quantifies the *coherence* of arbitrary input configurations. This combination of energy-based evaluation, generative capability, and structured latent representations makes DBMs particularly well-suited as inspectable world models.

Recent work combines LLMs with world models implicitly: social simulacra (Park et al., 2023; Yan et al., 2024; Bougie & Watanabe, 2025) use LLMs simultaneously as agents and world model. LLM-JEPA (Huang et al., 2025) adds a JEPA-style auxiliary loss to LLM fine-tuning, treating the world model as an implicit regularizer. We instead employ an explicit, separately trained world model that conditions a frozen LLM through learned soft prompts.

For controllable generation and PEFT, classic methods such as CTRL (Keskar et al., 2019) prepend control codes, while PPLM (Dathathri et al., 2020) steers a frozen LM via attribute-classifier gradients. Both operate on individual attributes without joint modeling of correlational structure. Among parameter-efficient fine-tuning methods, LoRA (Hu et al., 2022) has become the standard modern baseline. In addition, soft prompt tuning (Li & Liang, 2021) learns prompts end-to-end from text; we differ in deriving prompts from an external world model rather than from text alone.

A complementary line controls frozen LLMs by editing their internal representations. Activation steering methods — activation addition (Turner et al., 2023), contrastive activation addition (Rimsky et al., 2024), inference-time intervention (Li et al., 2023), and function/in-context vectors (Todd et al., 2024; Hendel et al., 2023; Liu et al.) — extract a direction in the LLM's activation space (typically a contrastive difference-of-means or linear probe) and add it back at inference time; representation engineering (Zou et al., 2023) systematizes the recipe on the longstanding linear-direction view (Kim et al., 2018; Park et al., 2024). Sparse-autoencoder steering clamps interpretable SAE features (Templeton et al., 2026; Gao et al., 2025; Chalnev et al., 2024), and DExperts (Liu et al., 2021) reshapes logits with auxiliary LMs at decoding time. In all cases the steering signal is a single correlational direction (or feature) per attribute applied inside the frozen LLM, with no separate jointly-trained model of how attributes co-occur. Our controllable object is instead an external energy-based world model whose binary latents encode attribute co-occurrence; clamping visible units and re-running mean-field inference propagates the clamp through an inspectable joint distribution while the LLM's weights and activations remain untouched.

Finally, regarding consumer review generation, early work used RNNs with attention mechanisms (Dong et al., 2017; Ni & McAuley, 2018). More recent approaches construct prompts from user behavioral history and fine-tune LLMs (Peng et al., 2024). A central challenge is *consumer heterogeneity* along latent difference such as price sensitivity and brand loyalty (Smith, 1956; DeSarbo et al., 1997; Wedel & Kamakura, 2000) which cannot be observed (Tseng et al., 2024; Zollo et al., 2025). LLM-based persona generation often reproduces systematic biases relative to true population distributions (Li et al., 2025). Our approach instead learns latent representations of heterogeneity as a market structure directly from behavioral data through an energy-based world model.

3 Architecture

The proposed architecture consists of three components (Fig. 1): (i) a DBM that serves as an energy-based world model, learning the latent co-occurrence structure of a domain from binary visible features, (ii) a lightweight MLP adapter that projects the DBM's mean-field beliefs into soft-prompt embeddings, and (iii) a frozen LLM that renders those beliefs into fluent natural language.

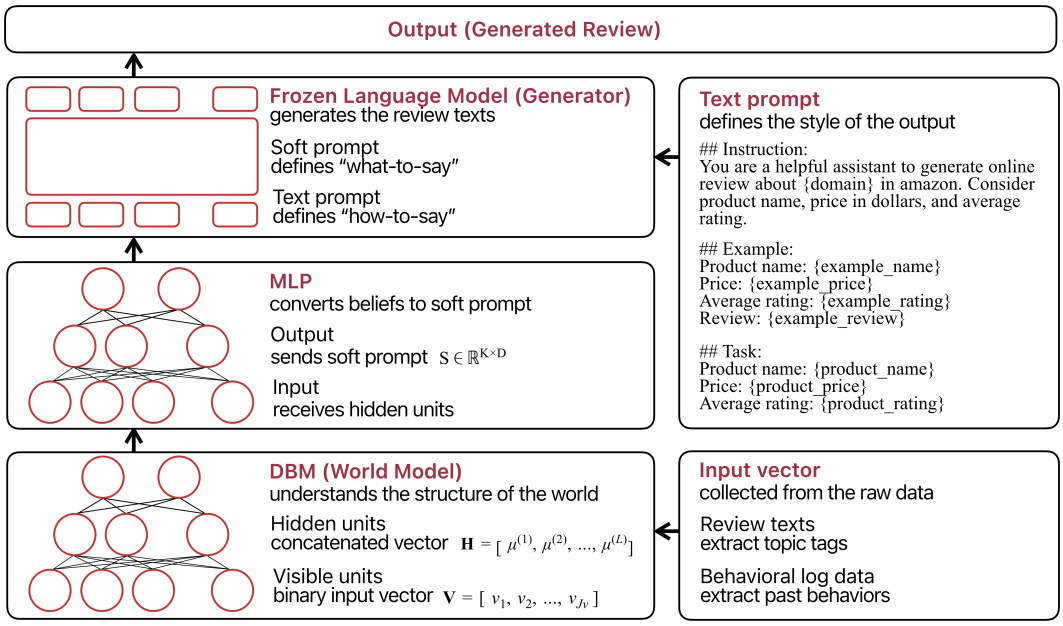

Figure 1: Overview of the three-component architecture. The DBM is trained unsupervised on binary visible features extracted from reviews. The adapter and LLM never observe the raw tabular features directly. During the final training phase, only the adapter is trained while both the DBM and LLM remain frozen.

World model: Deep Boltzmann Machine The DBM comprises one visible layer $\mathbf{v}$ with J_v binary units and L hidden layers $\mathbf{h}^{(l)}$ with J_l units each ($l = 1, \ldots, L$). The energy function is defined as

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

where $\mathbf{h}^{(0)} \equiv \mathbf{v}$. Thanks to the bipartite structure between consecutive layers, mean-field inference for layer l can be performed efficiently via the fixed-point iteration

$$\mu^{(l)} \leftarrow \sigma\Big(W^{(l)\top} \mu^{(l-1)} + W^{(l+1)\top} \mu^{(l+1)} + \mathbf{c}^{(l)}\Big), \tag{2}$$

with $\mu^{(0)} \equiv \mathbf{v}$ and $W^{(L+1)}\mu^{(L+1)} \equiv \mathbf{0}$. At inference time, we use the deterministic mean-field activations rather than stochastic samples for adapter conditioning. The concatenated vector $\mathbf{H} = [\mu^{(1)}; \ldots; \mu^{(L)}] \in \mathbb{R}^{J_1 + \cdots + J_L}$ serves as the abstract *belief* representation that is passed to the adapter. The free-energy-based coherence score $\tilde{F}(\mathbf{v})$ is defined in App. A.

Adapter: belief → soft prompt. A two-layer MLP maps the concatenated mean-field activations $\mathbf{H}$ to a sequence of K soft-prompt embeddings $\mathbf{S} \in \mathbb{R}^{K \times D}$, prepended to the LLM's input sequence. Only the adapter parameters are updated during the final training phase, using the AdaMax optimizer (Kingma & Ba, 2015) and cross-entropy loss with respect to ground-truth reviews.

Language model: frozen LLM. We employ Qwen2.5-0.5B-Instruct as the frozen generator, with all parameters kept frozen throughout training and inference. The LLM receives only a minimal textual product context (product name, price, and average rating) shared with every baseline. All additional structured domain knowledge is supplied exclusively through the DBM-derived soft prompt. The exact prompt template is provided in App. B.

4 Experimental Design

Dataset. We conduct experiments on Amazon product reviews (McAuley et al., 2015; He & McAuley, 2016) from two distinct domains: Smartphones (electronics category with oligopolistic brand structure) and Beauty products (long-tail category with brand–topic specialization). We use English reviews from verified purchases with lengths between 100 and 1024 characters (detected via FastText (Bojanowski et al., 2017; Joulin et al., 2017)). The dataset is split into train/val/test = 52,952/1,024/1,024 per domain. Each consumer profile is represented as a binary one-hot vector of dimension $J_v = 160$ (Smartphones) or $J_v = 170$ (Beauty), encoding brand, price tier, rating, and additional contextual signals (e.g., repeat buyer, accessory purchases) derived from the user's prior review history.

Inference. Given visible $\mathbf{v}$, we run mean-field iteration to convergence, project beliefs through the adapter to soft prompts, prepend to a one-shot text prompt, and generate text autoregressively with the frozen LLM (temperature 0.7, max_new_tokens = 100).

Baselines. Six baselines, all built on Qwen2.5-0.5B-Instruct as the base language model: **B0** Frozen ICL with a one-shot example (generator weights frozen, no adaptation); **B1** Direct-MLP adapter that bypasses the DBM but otherwise matches the proposed setup (frozen generator, identical text prompt), serving as a leakage-controlled comparator that shares the proposed model's visible-vector conditioning; **B2** Full fine-tuning of all LLM parameters on the in-domain training corpus; **B3** LoRA (Hu et al., 2022) (rank 8) applied to the LLM; **CTRL** (Keskar et al., 2019) controlling codes prepended at decoding time (frozen LLM); **PPLM** (Dathathri et al., 2020) attribute-classifier-guided steering (frozen LLM). The proposed model, B0, B1, CTRL, and PPLM keep the generator frozen. B1's MLP is trained end-to-end on the LLM's CE loss with direct access to visible features, giving it a structural advantage on form metrics that we analyze in §8.

Metrics. We evaluate form-level quality using test-set cross-entropy (CE; lower is better) and SBERT cosine similarity (CosSim; higher is better) on $n = 1{,}024$ held-out reviews. Statistical significance is assessed via paired t-tests on item-level differences (and Wilcoxon signed-rank tests where noted); see the multi-seed protocol below. Sentiment is measured with a RoBERTa sentiment classifier (`cardiffnlp/twitter-roberta-base-sentiment-latest`; scored as $P(\text{positive}) - P(\text{negative}) \in [-1, 1]$).

Multi-seed protocol. All quantitative results are reported under a rigorous 3-seed protocol designed to separate model variance from data variance. (i) **Fixed test split**: The data split seed is fixed at 0, so the same 1,024 test reviews (and matched clamped samples) are used across all seeds. (ii) **Training/generation seed**: For trainable models (Proposed, B1, B3), we vary the random seed in $\{0, 1, 2\}$, affecting DBM training (PCD sampling and initialization), DataLoader shuffling, and decoding. For non-trainable baselines, only the generation seed varies. (iii) **Statistical inference**: We report 3-seed mean ± std. Because the test split is fixed, each test item yields one paired difference per seed; we average the three seed-replicates of each item and run paired tests on the resulting n *item-level* differences (paired t; Wilcoxon signed-rank as a non-parametric cross-check). Items—rather than seed-replicates—are thus the unit of inference, which avoids the pseudo-replication of pooling $3 \times n$ correlated differences and, for this balanced design, is equivalent to a cluster-robust analysis with items as clusters. Pooling all $3 \times n$ per-sample differences instead yields the same significance pattern on every reported claim.

Conditioning and feature provenance. Visible features consist of *history-derived* signals (e.g., purchase frequency, repeat-buyer status) and *auto-extracted topical* features from the target review. This setup is best interpreted as *controlled rendering of text given desired attributes* rather than pure next-review prediction. All baselines receive the same per-sample attribute information as the proposed model, but through different channels: text fields appended to the prompt (B0/B2/B3/CTRL/PPLM) and the visible vector passed through the adapter (Proposed and B1).

Hyperparameters. Detailed hyperparameters for the three training phases (LPT, JFT, and adapter training) are provided in App. D.

5 Experiment I: Generation Quality

Before evaluating the meaning-side capabilities that motivate the architecture (§6–§7), we verify that the proposed pipeline does not sacrifice form-side generation quality relative to standard adaptation methods. The proposed model is compared against the six baselines defined in §4 (B0 One-shot ICL, B1 Direct-MLP, B2 Full FT, B3 LoRA, CTRL, PPLM).

Table 1 reports the main comparison. Across 3 seeds the proposed model dominates B0/B2/B3/CTRL/PPLM on every cell in both domains ($p < 10^{-10}$ on every cell except B2 Smartphone CE, p=5.1×10^{-6}); B1 (direct MLP, no DBM) wins one form metric per domain (Smartphone CosSim and Beauty CE) by margins within seed-to-seed std, as expected in §1. The dominance over Full FT and LoRA in particular establishes that DBM-mediated conditioning does not compromise generation quality.

Table 1: Generation quality on Smartphone and Beauty review generation. CE and CosSim on n=1024 test reviews per seed (PPLM rows drop 1–4% of failed generations); 3-seed protocol per §4, mean ± std; p-values are item-level paired t-tests against the proposed model (per-item differences averaged across the 3 seeds; §4). Since PPLM's dropped generations break the item-level pairing for CosSim, CosSim p is the per-seed maximum (every seed $< 10^{-10}$).

	Smartphone				**Beauty**			
Model	CE	p	CosSim	p	CE	p	CosSim	p
Proposed (DBM + adapter)	**3.135 ± 0.005**	—	0.534 ± 0.007	—	3.181 ± 0.006	—	**0.455 ± 0.005**	—
B1: Direct MLP (no DBM)	3.137 ± 0.003	0.383	**0.541 ± 0.002**	1.3×10^{-3}	**3.175 ± 0.003**	2.9×10^{-3}	0.453 ± 0.003	0.354
B2: Full fine-tuning	3.161 ± 0.003	5.1×10^{-6}	0.438 ± 0.003	$< 10^{-10}$	3.261 ± 0.009	$< 10^{-10}$	0.407 ± 0.007	$< 10^{-10}$
B3: LoRA (Hu et al., 2022)	3.197 ± 0.001	$< 10^{-10}$	0.440 ± 0.002	$< 10^{-10}$	3.240 ± 0.001	$< 10^{-10}$	0.405 ± 0.002	$< 10^{-10}$
CTRL (Keskar et al., 2019)	3.473 ± 0.000	$< 10^{-10}$	0.371 ± 0.002	$< 10^{-10}$	3.418 ± 0.000	$< 10^{-10}$	0.364 ± 0.001	$< 10^{-10}$
PPLM (Dathathri et al., 2020)	3.583 ± 0.000	$< 10^{-10}$	0.275 ± 0.002	$< 10^{-10}$	3.583 ± 0.000	$< 10^{-10}$	0.233 ± 0.006	$< 10^{-10}$
B0: One-shot ICL	3.674 ± 0.005	$< 10^{-10}$	0.421 ± 0.002	$< 10^{-10}$	3.619 ± 0.004	$< 10^{-10}$	0.401 ± 0.000	$< 10^{-10}$

6 Experiment II: Coupling Discovery as World-Model Validation

Experiment I evaluates the integrated model; however, it does not isolate what the DBM has learnt. We therefore conduct Experiment II for world-model validation. In Phase 1 (discovery), for all $\binom{7}{2}$=21 cross-prefix visible-attribute pairs in each domain, we rank empirical Pearson correlations and select the strongest cross-prefix couplings. In Phase 2 (validation), we clamp the discovered visible units in the DBM and measure the marginal flip of the conditionally clamped variable; we additionally include *control pairs* with empirical $r\approx0$ to test for false positives.

Phase 1: empirical coupling discovery. Table 2 shows that the strongest cross-prefix couplings differ markedly between domains (full rankings in App. F). Smartphone exhibits an oligopolistic Brand–Price/Purchase coupling, consistent with a market where a few brands occupy distinct price tiers. Beauty shows weaker price coupling but distinctive brand–category specialization (e.g., Sally Hansen–Color, Bath & Body Works–Fragrance), consistent with a more fragmented market where brand identity is anchored in product specialty rather than price tier. Across both domains, Rating–Topic(Service) is consistently coupled, suggesting a cross-domain invariant rooted in consumer psychology.

Phase 2: DBM clamp validation. Table 3 reports the DBM's response to clamping each discovered (and control) coupling. The model recovers all positively-coupled pairs with measurable marginal flips and correctly returns near-null on control pairs (empirical $r\approx0$): predictive validity rather than confirmation bias. For invariant pairs, the flip is asymmetric: clamping Topic(Service)= 1 raises $P(\text{Rating} = 1)$ substantially (+0.83–+1.24%), while clamping Rating= 1 produces a near-null or slightly negative change in $P(\text{Service})$; this reflects that the coupling is stronger in the consumer-psychology direction (service complaints predict low ratings) than in reverse.

Three-tier latent structure. Combining Phases 1 and 2 yields a three-tier organization of the learnt belief space: **(i) cross-domain invariant** (Rating–Service in both domains, $r\sim$0.21–0.25), reflecting consumer-

Table 2: Top empirical couplings (max $|r|$ across all unit pairs within prefix combinations). Brand–Price coupling differs by 3.6× between domains; Rating–Topic coupling is invariant.

Coupling	**Type**	**Smartphone** $\lvert r\rvert$	**Beauty** $\lvert r\rvert$
PriceRange × Purchase	oligopoly (domain-specific)	0.908	0.289
Brand × Purchase	oligopoly (domain-specific)	0.636	0.027
Brand × Price	oligopoly (domain-specific)	0.277	0.076
Rating × Topic(SERVICE)	cross-domain invariant	0.254	0.212
Brand × Topic specialty †	long-tail (domain-specific)	0.249	0.161

† Apple–Damage (Smartphone), Sally Hansen–Color (Beauty).

Table 3: DBM clamp response on discovered couplings. *Specialty* pairs are domain-specific positive couplings; *control* pairs have empirical $r \approx 0$; *invariant* pairs hold across domains. "Flip $b \mid a{=}1$" is the percentage change in the marginal probability of b when a is clamped to 1.

Domain	**Pair**	**Type**	r	flip $b\|a{=}1$ (%)	flip $a\|b{=}1$ (%)
Smartphone	Apple × Topic(DAMAGE)	specialty	+0.249	−0.093	+1.693
	Google × Topic(SCREEN)	specialty	+0.117	+0.187	+1.162
	Apple × Topic(GAMING)	control	−0.064	+0.413	+1.377
	Google × Topic(BATTERY)	control	+0.001	+0.354	+0.882
	Rating=1 × Topic(SERVICE)	invariant	+0.254	−0.039	+1.237
Beauty	Sally Hansen × Nails	specialty	+0.158	+0.861	+0.690
	Sally Hansen × Color	specialty	+0.161	+1.236	+0.702
	Bath & Body Works × Fragrance	specialty	+0.135	+1.337	+0.668
	Essie × Color	specialty	+0.103	+1.406	+0.760
	Sally Hansen × Fragrance	control	−0.034	+0.152	+1.295
	Bath & Body Works × Nails	control	−0.014	−0.217	+0.956
	Rating=1 × Topic(SERVICE)	invariant	+0.212	−0.346	+0.831

psychology regularities; **(ii) domain-specific** (Smartphone Brand–Price oligopoly, Beauty Brand–Topic specialty), reflecting market structure; **(iii) common-null** (control pairs in both domains), where the DBM correctly refrains from imposing structure. The DBM does not impose rigid priors; it learns whichever coupling structure each domain exhibits. App. G shows the full Pearson correlation matrices of visible variables under both the empirical distribution and DBM samples, confirming that the DBM tracks each domain's second-order co-occurrence structure rather than imposing a shared prior.

Energy ranks natural over swap-perturbed configurations. Beyond per-pair flips, the DBM's free energy globally ranks naturally observed configurations as more plausible than swap-perturbed ones (LeCun et al., 2006). For each of the six single-attribute clamps used in §7, we compare $\tilde{F}(v_{\text{original}})$ to $\tilde{F}(v_{\text{clamped}})$ where one attribute is swapped to a less-coupled value (Table 4). $\Delta\bar{F}$ is positive in all six conditions and the natural configuration receives the lower energy on 84–99% of paired samples; the gap scales with empirical coupling strength (largest for Smartphone Price Premium→Entry, +11.9; smallest for Smartphone Brand swap, +3.9).

7 Experiment III: Coupling-Aware Targeted Clamp

The latent structure uncovered in Experiment II provides two distinct clamp handles: *sentiment-axis* clamps exploit cross-domain invariant couplings (e.g., Rating–Service), while *content-axis* clamps exploit domain-specific couplings (e.g., Brand–Topic specialty). We therefore verify their independence, using several baselines: (i) B1 Direct-MLP which shares the proposed model's visible-vector conditioning but has no clampable

Table 4: Free-energy ranking of natural vs. swap-perturbed configurations under the six single-attribute clamps used in §7. $\Delta\bar{F} = \bar{F}(v_{\text{clamped}}) - \bar{F}(v_{\text{original}})$; "% lower" is the fraction of paired samples for which the natural configuration has lower energy (chance is 50%).

	Smartphone			**Beauty**		
Clamp	n	$\Delta\bar{F}$	% lower	n	$\Delta\bar{F}$	% lower
Rating 5→1	500	+4.37	83.8%	500	+4.90	96.4%
Price (Premium/Luxuary) → (Entry/Budget)	162	+11.88	99.4%	500	+5.81	94.6%
Brand swap	500	+3.90	89.6%	112	+6.18	96.4%

belief space, as the leakage-controlled comparator throughout and (ii) a generously tuned activation-steering baseline (CAA; Rimsky et al. 2024) on the sentiment axis (§7.4)[1].

7.1 Single-attribute selectivity

For each domain we draw n=200 samples for Rating and n=100 for Price/Brand and apply three single-attribute clamps: Rating 5 → 1, Price (Premium/Luxury → Entry/Budget), Brand (Apple→Samsung, MAC→L'Oréal). The 3-seed evaluation applies; sentiment is measured with RoBERTa (§4).

The Rating clamp produces a strong sentiment shift in both domains ($p < 10^{-10}$; Table 5). Most Price/Brand cells remain null. The one Price/Brand cell with a detectable signal is Beauty Price (Δ=−0.088, paired-t p=1.5×10^{-3}, Wilcoxon p=0.039), consistent with the empirical Brand–Price coupling r=0.076 being weak but non-zero. Selectivity in our framework therefore reflects fidelity to the data: where the empirical coupling is appreciable, the clamp leaves a detectable signature; where it is essentially absent, no significant shift is induced.

Table 5: Single-attribute clamp effects on sentiment under 3-seed evaluation. The paired-t test is computed against zero on item-level differences (each item's per-seed differences averaged across the 3 seeds; §4). †: paired-t $p < 10^{-3}$; *: $p < 0.05$; ∘: $p<0.1$; ns: $p\geq0.1$ with Wilcoxon signed-rank test.

Domain	**Clamp**	$n_{\text{per-seed}}$	Δ RoBERTa	paired-t p	Wilcoxon p
Smartphone	Rating 5→1	200	-1.388 ± 0.045†	$< 10^{-10}$	$< 10^{-10}$
	Price Premium→Entry	100	$+0.038\pm0.028^{ns}$	0.250	0.415
	Brand Apple→Samsung	100	$+0.058\pm0.029^{\circ}$	0.056	0.183
Beauty	Rating 5→1	200	-1.489 ± 0.065†	$< 10^{-10}$	$< 10^{-10}$
	Price Luxury→Budget	100	$-0.088\pm0.090^{*}$	1.5×10^{-3}	0.039
	Brand MAC→L'Oréal	100	$+0.017\pm0.036^{ns}$	0.545	0.382

7.2 Coupling-aware combined clamp sharpens counterfactuals

We further test whether jointly clamping a coupled pair (Rating = 1 *and* Topic(SERVICE)= 1) brings the generated distribution closer to natural low-rating samples than clamping Rating alone. We report Kolmogorov–Smirnov p and Wasserstein distance against two natural baselines: all reviews with Rating= 1, and the conditional subset Rating= 1∩Service.

On Beauty the combined clamp sharpens the distributional match: Wasserstein drops from 0.147±0.030 (rating-only ↔ all R=1) to 0.113±0.016 (combined, −23%). On Smartphone the improvement is modest: the combined clamp improves Wasserstein from 0.235 to 0.210 (−11%). The effect is therefore Beauty-dominant. Figure 2 visualizes the distributions: the natural low-rating mass concentrates near −1, the

[1]CTRL and PPLM are omitted here: their form-side outputs lie outside the valid review distribution (Table 1) so pre/post-clamp Δs would measure OOD shifts, and they implement attribute control via marginal channels rather than the joint-distribution exploitation the meaning-level tests require.

rating-only clamp reproduces this mode while retaining residual positive mass, and the combined clamp moves the residual toward the natural Rating= 1∩Service profile.

Table 6: Coupling-aware combined clamp vs. single-attribute Rating clamp under 3-seed evaluation (mean ± std), RoBERTa sentiment. Distance to the model's naturally-conditioned low-rating generations (the *model-consistent* reference: generation conditioned on natural Rating=1 / Rating=1∩Service samples without clamp); smaller is better.

Domain	**Comparison**			**KS** p	**Wasserstein**
Smartphone	rating-only	↔	all R=1	0.006 ± 0.008	0.235 ± 0.054
	combined	↔	all R=1	0.000 ± 0.000	0.210 ± 0.066
	combined	↔	R=1∩Service	0.017 ± 0.023	0.192 ± 0.064
Beauty	rating-only	↔	all R=1	0.013 ± 0.012	0.147 ± 0.030
	combined	↔	all R=1	0.078 ± 0.100	0.113 ± 0.016
	combined	↔	R=1∩Service	0.028 ± 0.036	0.091 ± 0.020

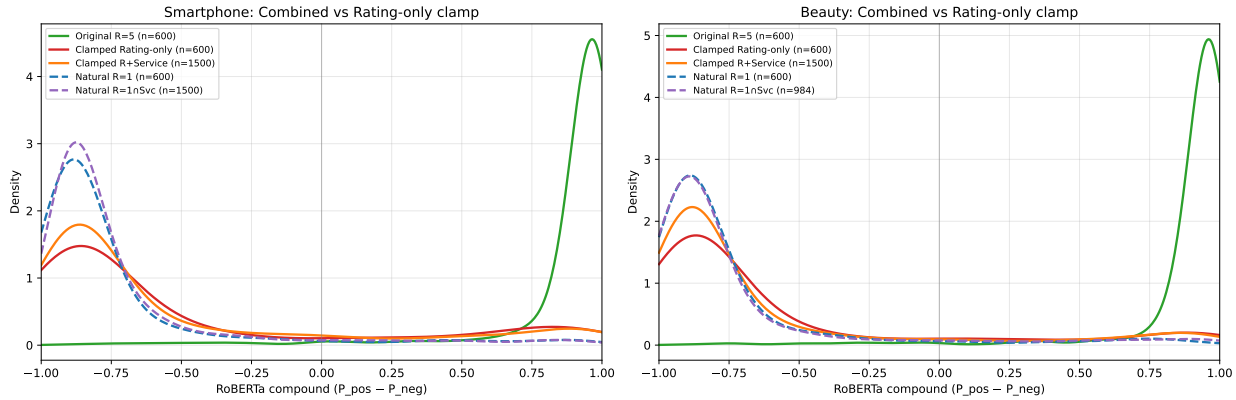

Figure 2: Combined clamp KDE (RoBERTa). The single-attribute Rating clamp (red) approximates the natural Rating= 1 distribution (blue, dashed) but retains residual positive mass; the coupling-aware combined clamp (orange) tightens the match toward the natural Rating= 1∩Service mode (purple, dashed).

7.3 Content-axis clamp via brand–topic specialty

Utilizing Beauty's domain-specific Brand–Topic specialty couplings (Sally Hansen–Nails, Bath & Body Works–Fragrance), we perform content-level counterfactuals. For each pair, we clamp the brand and target topic visible units and generate n=500 samples per seed under the 3-seed evaluation, then measure: keyword frequency (average per-generation count of a hand-curated topic lexicon under case-insensitive whole-word match[2]), SBERT similarity to natural baseline (mean-best-match cosine), and sentiment (to verify content/sentiment independence).

Keyword frequency rises sharply in both pairs (Table 7): Sally Hansen → Nails moves from 0.29 to 0.86 (∼ 3× over baseline, natural kw 4.08), and Bath & Body Works → Fragrance moves from 0.03 to 0.63 (∼ 21×, natural kw 2.93); B1 reaches only +0.17 and +0.31 respectively, so the DBM-mediated content shift is 1.9–3.4× that of the direct-MLP comparator under the same clamp. SBERT mean-best-match against natural Brand∩Topic samples supports this: the proposed model's clamp moves content similarity by +0.059±0.020 (Sally Hansen) and +0.077±0.013 (Fragrance), versus B1's +0.020±0.005 and +0.047±0.014, a 1.6–3.0× stronger pull toward the target Brand∩Topic mode (Table 7). The sentiment-axis control test is

[2]Nails: {nail, polish, color, manicure, pedicure, lacquer, enamel}. Fragrance: {fragrance, scent, smell, perfume, aroma, cologne}.

largely respected. Sally Hansen × Nails remains null (p=0.28); Bath & Body Works × Fragrance resolves a small positive shift (Δ=+0.046±0.012, p=6.1×10^{-4}). The effect appears in B1 as well (Δ=+0.048±0.026, p=5.6×10^{-4}), so it is not specific to the DBM-mediated handle: Fragrance reviews are intrinsically positive-leaning and the content shift drags sentiment marginally with it. The two-axis decomposition therefore holds in the strong sense for Sally Hansen and in a qualified sense for Bath & Body Works (orthogonality is ∼95%; the residual is detectable but the effect size is an order of magnitude smaller than Rating's $|\Delta|\geq 0.9$). App. H provides representative verbatim generation samples illustrating each axis.

Table 7: Brand–Topic specialty clamp (Beauty) under 3-seed evaluation (mean ± std). ΔRoBERTa: sentiment shift (item-level paired-t p, §4); kw orig→int: topic-keyword frequency on the clamped generations vs. the original-attribute generations from the same prompts; kw natural: average kw count on the model's naturally-conditioned generations over the natural Brand∩Topic samples (n=51/70 for the two pairs; the natural review texts themselves average 1.94/1.30); SBERT$_{o\to b}$/SBERT$_{i\to b}$: mean-best-match cosine similarity of original/clamped generations to the same naturally-conditioned generations (model-consistent reference, cf. Table 6; higher = content closer to the target Brand∩Topic mode).
∗: $p < 0.05$; ∘: $0.05<p<0.1$; otherwise $n.s.$ ($p \geq 0.1$). See §7.3 for the paired-t p-values.

Pair	ΔRoBERTa	kw orig→int	kw nat.	SBERT$_{o\to b}$	SBERT$_{i\to b}$
Sally Hansen × Nails					
Proposed (DBM + adapter)	−0.015±0.012	0.29 → 0.86	4.08	0.398±0.024	0.458±0.017
B1: Direct MLP (no DBM)	−0.002±0.021	0.26 → 0.44	4.05	0.412±0.010	0.431±0.011
B&B Works × Fragrance					
Proposed (DBM + adapter)	+0.046±0.012 *	0.03 → 0.63	2.93	0.430±0.012	0.506±0.004
B1: Direct MLP (no DBM)	+0.048±0.026 *	0.03 → 0.35	2.37	0.448±0.008	0.496±0.012

7.4 Comparison with activation steering

We finally implement contrastive activation addition (CAA; Rimsky et al., 2024) on the same frozen Qwen2.5-0.5B-Instruct generator. The rating steering vector is the unit-normalized difference of mean residual-stream activations between natural Rating=1 and Rating=5 reviews (256 per side, disjoint from the generation samples), added to one decoder block during decoding with coefficient α; original and steered generations share an identical text prompt, so the activation edit is the only modification channel. The hyperparameters (block 18 of 24, α=16, per domain) are selected *in CAA's favor* on held-out dev prompts by maximizing the sentiment shift over a 4 × 3 (layer, α) grid, subject to a fluency guard (median steered perplexity ≤ 3× unsteered; the selected configurations show no degradation). The combined condition adds the sum of the rating and a Topic(Service) steering vector, which is CAA's native vector-arithmetic composition. Since extraction is deterministic, only the generation seed varies for the 3-seed evaluation, as for the other non-trainable baselines.

Raw sentiment shift: steering is a competent handle. On Rating 5→1 CAA produces ΔRoBERTa=−1.011/−1.290 (Smartphone/Beauty) against the proposed model's −1.388/−1.489 and B1's −1.279/−1.335 (Table 5). A single correlational direction therefore suffices to move the *mean* of the sentiment distribution by an amount comparable to clamping, as the steering literature predicts.

Distribution-level match: only the world model reproduces the natural distribution. The model-consistent reference of Table 6 is undefined for CAA: unsteered, the system has no conditioning channel and its generations on natural low-rating prompts remain strongly positive, leaving no naturally-conditioned generation to compare against. We therefore score every system against the same model-independent reference: the sentiment distribution of the natural review *texts* (Rating=1 for rating-only; Rating=1∩Service for combined). Table 8 shows that the proposed model attains the best match in three of four cells and is the only system whose match is not rejected by KS anywhere; CAA is rejected in every cell (p≈0), including Beauty rating-only where its Wasserstein is nominally smallest—the first moment matches, the shape does

not. Diagnostically, the combined condition *degrades* CAA on Smartphone (W: 0.347 → 0.448): summed steering vectors interfere, while coupling-aware clamping improves the match in both domains.

Table 8: Distribution match against the natural review-text reference (RoBERTa), 3-seed mean ± std. Wasserstein distance (smaller is better) and KS p (higher = harder to distinguish from the natural texts; 0.000 = rejected).

		Proposed		**B1**		**CAA**	
Domain	**Condition**	W	KS p	W	KS p	W	KS p
Smartphone	rating-only	**0.216 ± 0.063**	0.098 ± 0.122	0.317 ± 0.075	0.002	0.347 ± 0.017	0.000
Smartphone	combined	**0.193 ± 0.046**	0.004 ± 0.006	0.283 ± 0.059	0.000	0.448 ± 0.003	0.000
Beauty	rating-only	0.124 ± 0.063	**0.374 ± 0.325**	0.285 ± 0.097	0.013	**0.101 ± 0.013**	0.000
Beauty	combined	**0.138 ± 0.078**	0.096 ± 0.069	0.248 ± 0.039	0.000	0.247 ± 0.008	0.000

Content axis: steering moves keywords, not controlled content. We construct brand and topic steering vectors for the two Brand–Topic specialty pairs and add their sum, with (layer, α) re-tuned in CAA's favor on a keyword-shift objective under the same fluency guard (block 14, α=16). Against the model-independent reference (Table 9), steering moves target keywords far more aggressively than clamping (0.29 → 2.12, 0.09 → 4.31, the latter overshooting the natural rate of 1.30 by 3.3×), but the shift is not controlled content. Two-axis independence fails: the Sally Hansen content edit drags sentiment by Δ_{RoBERTa}=−0.605 (proposed: −0.015, n.s.), and Bath & Body Works leaks +0.127 (2.8× the proposed model's residual). And on Sally Hansen the steered generations move *away* from the natural target mode (SBERT 0.384 → 0.278) where clamping moves toward it (+0.060). The mechanism is the contrast sets' incidental valence skews (−0.02/−0.03 Nails/Sally Hansen; +0.13/+0.06 Fragrance/Bath & Body Works), whose signs match the induced sentiment shifts: a difference-of-means direction carries every co-varying attribute, and the coefficient required for a content-sized effect amplifies the leak. The DBM clamp does not inherit this entanglement, because mean-field inference propagates the clamp through a joint distribution in which Rating and Topic are distinct variables.

Table 9: Content-axis clamp (Brand×Topic clamp vs. brand+topic steering vector) against the model-independent reference: the natural Brand∩Topic review texts (n=51/70; keyword rates 1.94/1.30). 3-seed mean ± std. ΔRoBERTa: sentiment side-effect (should be ≈0 for a controlled content edit); $\text{SBERT}_{o\to n}$/$\text{SBERT}_{i\to n}$: mean-best-match cosine of original/clamped generations to the natural texts. For Proposed/B1 the deltas are nearly identical to the model-consistent ones of Table 7.

Pair	**Model**	ΔRoBERTa	kw orig→int	$\text{SBERT}_{o\to n}$	$\text{SBERT}_{i\to n}$
Sally Hansen × Nails	Proposed	-0.015 ± 0.012^{ns}	0.29 → 0.86	0.373±0.005	**0.433 ± 0.005**
	B1	-0.002 ± 0.021^{ns}	0.26 → 0.44	0.370±0.002	0.394±0.006
	CAA	−0.605±0.041 †	0.29 → 2.12	0.384±0.001	0.278±0.003
B&B Works × Fragrance	Proposed	+0.046±0.012 *	0.03 → 0.63	0.412±0.004	0.484±0.010
	B1	+0.048±0.026 *	0.03 → 0.35	0.408±0.003	0.462±0.009
	CAA	+0.127±0.031 †	0.09 → 4.31	0.438±0.002	**0.644 ± 0.001**

†: item-level paired-t $p < 10^{-3}$; *: $p < 0.05$; ns: $p \geq 0.1$.

What the comparison establishes. CAA is the strongest competing handle we evaluate, and it differs from the proposed architecture in kind, not degree: it shifts sentiment means but is everywhere rejected on distributional match to natural reviews under RoBERTa, composes destructively where coupling-aware clamps sharpen, entangles sentiment on the content axis, and requires a hand-built contrast set plus a per-task (layer, α) sweep—whereas both axes of §7.1–§7.3 are clamps on a single learnt belief space that also exposes a coherence score (§6) and discovered coupling structure to exploit. Steering edits the mouth; it does not consult a brain.

8 Discussion and Conclusion

Form vs. meaning: the prediction and how it resolves. The architectural separation (§1) predicts that B1's direct MLP—which receives the LLM's CE gradient all the way back to the inputs while our DBM is trained on the joint visible likelihood with that gradient blocked at the adapter boundary—should match or marginally beat the proposed model on form, while meaning-level capabilities (coherence scoring, predictively-valid clamps, two-axis counterfactual-style control) remain exclusive to the DBM-based pipeline. Both halves resolve as predicted: B1 ties or marginally edges the proposed model on form (Table 1, App. C), while the proposed model statistically dominates LoRA, full fine-tuning, CTRL, PPLM, and one-shot ICL ($p < 10^{-5}$). The shared form-side advantage over LoRA/Full FT is parsimoniously attributed to the soft-prompt channel itself (Lester et al., 2021); the DBM's contribution is the joint distribution structure that enables the meaning-side capabilities of §7.

Meaning-side observations strictly favor the DBM-based architecture. On meaning (§7) the proposed model wins B1 across all cross-domain comparisons (Rating Δ amplified $\sim$9–12%, combined-clamp W reduced $\sim$11–23%, content-axis shifts 1.6–3.4$\times$ larger). A generously tuned CAA baseline matches the raw sentiment shift but is rejected on distributional consistency under RoBERTa, composes destructively under multi-attribute clamps, and entangles sentiment on the content axis (§7.4). The DBM's belief space carries empirically near-orthogonal sentiment and content handles—a sentiment clamp leaves content unchanged, and a content clamp leaves sentiment unchanged in three of four pair$\times$metric cells (the residual Δ=+0.046 on Bath & Body Works $\times$ Fragrance also appears in B1, marking it as a fragrance-domain artifact). B1 has no structured belief space to factor over, and can be re-prompted but cannot be *clamped* along a chosen axis with the other held fixed. The framework's central claim is therefore established not by denying B1 its form-level advantage but by the structural argument that meaning-level capabilities are a separate axis of system design.

What we argue should change. Two practices should change. *First*, when methods such as LoRA already approach the ceiling on cross-entropy and embedding similarity, further marginal gains on the same metrics no longer discriminate among architectures; the discriminative axis is whether the system supports coherence scoring and predictively-valid attribute clamps. *Second*, interpretability and intervenability should be first-class architectural requirements rather than post-hoc probes on monolithic models. The present architecture is one concrete realization, not a unique one. The mouth speaks; the brain, separately and inspectably, decides what to say.

Scope and limitations. Causal interpretation. "Clamping" here denotes fixing visible units in the learnt DBM and re-running mean-field inference before decoding; the selectivity reflects *model-internal distributional consistency*, not identification of real-world causal effects. We do not specify a structural causal model nor address latent confounding. **Form/meaning framing.** We adopt Bender & Koller (2020)'s form/meaning distinction as an *organizing frame*, not as a commitment to the stronger position that LLMs trained on language alone cannot acquire meaning. Our weaker working claim suffices: form-level metrics are insufficient to evaluate meaning-level capabilities, which require structural clamping rather than re-prompting (§7). **Sample sizes.** Exp I uses n=1024/seed; Exp III uses n=200/seed for Rating and n=100/seed for Price/Brand. Seed-to-seed std is $\leq$ 0.009 on CE/CosSim and $\leq$ 0.09 on clamp Δ. **Cross-model and cross-domain scope.** Transfer to Llama-3.2-1B-Instruct and OLMo-2-1B-Instruct is reported single-seed in App. E as directional evidence. Broader generalization (other categories, languages, non-review text, larger generators) remains to be verified.

Reproducibility. Code and pre-trained model checkpoints will be made publicly available upon acceptance.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

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.

Nicolas Bougie and Narimawa Watanabe. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 215–229, 2025.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*, 2024.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1edEyBKDS`.

Wayne DeSarbo, Asim Ansari, Pradeep Chintagunta, Charles Himmelberg, Kamel Jedidi, Richard Johnson, Wagner Kamakura, Peter Lenk, Kannan Srinivasan, and Michel Wedel. Representing heterogeneity in consumer response models 1996 choice conference participants. *Marketing letters*, 8(3):335–348, 1997.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 623–632, 2017.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

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.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade: Second Edition*, pp. 599–619. Springer, 2012.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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.

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.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

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.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

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, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 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.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706, 2021.

Sheng Liu, Haotian Ye, Lei Xing, and James Y Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Forty-first International Conference on Machine Learning*.

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.

Jianmo Ni and Julian McAuley. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 706–711, 2018.

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.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39643–39666, 2024.

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.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

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.

Wendell R. Smith. Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, 21(1):3–8, 1956. ISSN 00222429. URL `http://www.jstor.org/stable/1247695`.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, 1986.

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

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Michel Wedel and Wagner A Kamakura. *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media, 2000.

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.

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.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.