ENERGY-BASED DAG STRUCTURE REFINEMENT FOR LARGE LANGUAGE MODEL PREFERENCE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning large language models (LLMs) with human preferences remains a central challenge in modern AI. Reward models (RMs) serve as proxies for human judgment but are highly susceptible to overoptimization and reward hacking. RM ensembles provide complementary evaluations, but scalar pooling compresses their disagreements and graph aggregation can retain conflicting edges. In this paper, we propose **E**nergy-based **D**AG s**T**ructure ref**I**nement (EDIT) to learn a prompt-specific preference graph from multiple RMs. A structural equation model (SEM) supplies a tractable graph-conditioned density, and a residual energy-based model (EBM) fits joint patterns beyond this base model. We train the structural module and energy function with SEM likelihood and noise contrastive estimation (NCE) using adaptive negatives from a conditional flow. Acyclicity regularization and edge pruning yield a directed acyclic graph (DAG) for consistent ranking. Extensive experiments on RMB and AlpacaFarm demonstrate state-of-the-art performance in both response selection and policy optimization.

1 INTRODUCTION

Preference alignment guides large language models (LLMs) toward human values, intentions, and social norms. The general paradigm for this goal is Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. (2022); Bai et al. (2022); Stiennon et al. (2020); Ziegler et al. (2019), which usually trains a Reward Model (RM) on pairwise human preferences and then performs downstream optimization via policy gradient methods Schulman et al. (2017) or candidate selection Nakano et al. (2021); Touvron et al. (2023). As a proxy for human judgment, the RM shapes both response selection and policy updates.

Despite considerable progress, RMs remain imperfect proxies. Under sustained optimization pressure or on out-of-distribution inputs, policies tend to exploit systematic errors in the reward signals. This phenomenon, known as overoptimization Gao et al. (2023) or reward hacking Skalse et al. (2022), causes proxy rewards to increase while true response quality stagnates or even deteriorates. Existing solutions fall into two broad directions. The first strengthens individual evaluators. For instance, ODIN Chen et al. (2024) disentangles length-correlated features via two linear heads to suppress length-based hacking, while BSR Hong et al. (2025a) introduces a batch-wise sum-to-zero regularizer to stabilize reward magnitudes and improve generalization. These methods improve individual evaluators without explicitly exploiting disagreement across RMs. The second direction diversifies reward signals through RM ensembles Coste et al. (2024); Eisenstein et al. (2024), typically aggregating outputs with conservative heuristics such as worst-case or uncertainty-weighted schemes. Scalar pooling yields a consistent order but compresses the directions and strengths of view-specific disagreements. Graph aggregation retains these comparisons and resolves conflicting edges. PGED Hu et al. (2026) combines pairwise votes through weighted graph denoising. We learn graph weights from continuous RM margins and candidate embeddings.

Motivation. RMs may share training data and pretrained backbones, making their errors correlated. A ranking model can exploit both cross-view agreement and candidate-level context. Conventional directed acyclic graph (DAG) learners estimate a shared graph from repeated observations Zheng et al. (2018); Yu et al. (2019). Here, candidate identities change across prompts, and each graph has only a few RM views. We share a structure predictor across prompts and condition its edges on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

directional margins and candidate embeddings. A structural density models compact dependencies, while a residual energy correction captures remaining joint patterns.

Contribution. EDIT constructs a weighted preference graph from RM margins and candidate embeddings. A generalized structural equation model (SEM) Pearl (2009) models candidate representations conditioned on the graph, while a residual energy-based model (EBM) Deng et al. (2020); Bakhtin et al. (2021) captures remaining joint patterns. An alternating procedure combines SEM likelihood, noise contrastive estimation (NCE) Gutmann & Hyvärinen (2010), and adaptive negatives from a conditional flow. Acyclicity regularization followed by edge pruning yields a graph for response selection and policy optimization. Our main contributions are threefold:

- We introduce an EDIT framework, which learns prompt-specific preference graphs from RM margins and candidate embeddings using a structural density and residual energy correction.
- We develop an alternating scheme that trains the structural and energy modules with NCE and adapts a conditional flow for direct negative sampling.
- Experiments on response selection and policy optimization show gains over the evaluated baselines, with ablations and corruption tests examining component effects and robustness.

2 Related Work

2.1 Energy-based Preference Alignment

EBMs LeCun et al. (2006) are increasingly applied in natural language processing. In autoregressive language modeling, residual EBMs Bakhtin et al. (2021) and subsequent diffusion-based extensions Xu et al. (2025) introduce sequence-level energy corrections to capture global dependencies beyond local token factorization. Recently, EBMs have also advanced LLM preference alignment. ARM Pang et al. (2024) further treats the aligned policy itself as a residual EBM learned via energy matching procedure. EBRM Lochab & Zhang (2025) employs a post-hoc energy-based procedure to refine pretrained reward models for better generalization. EPA Hong et al. (2025b) replaces the standard Bradley-Terry formulation with an energy preference model tailored for offline alignment. While these methods operate on instance-level architectures, EDIT applies residual energy modeling to joint candidate representations conditioned on a learned ranking graph.

2.2 DAG Structure Learning

Learning the DAG that best fits the data is NP-hard, owing to the combinatorial acyclicity constraint (Chickering, 1996). NOTEARS (Zheng et al., 2018) introduced a continuous DAG learning formulation, replacing this constraint with a smooth matrix-exponential trace function that recasts discrete structure search as gradient-based optimization. DAG-GNN Yu et al. (2019) extends linear SEMs to nonlinear formulations via a graph variational autoencoder, substituting the matrix exponential with a more efficient polynomial alternative. FJMP Rowe et al. (2023) demonstrates the practical value of pruning instance-wise interaction graphs into DAGs and decodes sequentially in topological order for downstream prediction. More recently, DDCD Zhu et al. (2026) utilizes a diffusion denoising objective to smooth the optimization landscape and stabilize convergence amid severe feature-sample imbalances. We repurpose this structural perspective for correlated multi-RM ensembling, constructing a DAG to enforce a consistent preference ranking.

2.3 Causal Reward Modeling

Recent studies introduce causality into reward modeling to mitigate reward hacking and spurious correlations. To combat reliance on context-irrelevant shortcuts, RRM Liu et al. (2025) utilizes causal data augmentation to address prompt-independent artifacts like length and format. Similarly, CRM Wang et al. (2025) enforces counterfactual invariance against non-causal attributes such as length, sycophancy, or concept bias, while FCRL Yang et al. (2026) learns factored representations to isolate reward-relevant causal features. Focusing on alignment, CARP Lu et al. (2026) strengthens the causal link between prompt intent and reward predictions via embedding reconstruction. For observational feedback, CausalRM Wang et al. (2026) applies noise-aware surrogate learning and

propensity-score reweighting to correct user-selection bias. Unlike methods that optimize a single RM, we aggregate multiple RMs by modeling the prompt-specific preference graph as a DAG.

3 PRELIMINARIES

Directed Acyclic Graphs. Let $\mathcal{G}$ be a graph over N random vectors $X = [x_1, \ldots, x_N] \in \mathbb{R}^{N\times d}$, where node i represents x_i. Its structure is encoded by a weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ with zero diagonal $\mathbf{W}_{ii} = 0$, where $\mathbf{W}_{ij} \neq 0$ if and only if $x_j \to x_i$. The support of $\mathbf{W}$ is independent of the node dimension d. Imposing acyclicity rules out directed cycles, making $\mathcal{G}$ a DAG.

Generalized Structural Equation Models. A linear SEM posits $\mathbf{x}_i = \sum_{j=1}^{N} \mathbf{W}_{ij}\, \mathbf{x}_j + \boldsymbol{\varepsilon}_i$, where $\mathbf{W}$ encodes directed relations among N variables and $\boldsymbol{\varepsilon}_i$ is independent exogenous noise. Stacking the variables gives $\mathbf{X} = \mathbf{W}\mathbf{X} + \mathbf{E}$. For acyclic $\mathbf{W}$, the inverse $(\mathbf{I} - \mathbf{W})^{-1}$ exists. We use a generalized SEM with latent structural mixing and row-wise diffeomorphisms f_1 and f_2:

$$\mathbf{Y} = \mathbf{W}\mathbf{Y} + f_1(\mathbf{E}), \qquad \mathbf{X} = f_2(\mathbf{Y}), \tag{1}$$

yielding $\mathbf{X} = f_2((\mathbf{I} - \mathbf{W})^{-1} f_1(\mathbf{E}))$, where $\mathbf{Y}$ denotes the latent variables. The structural mixing stays linear in $\mathbf{W}$, f_1 relaxes the Gaussian assumption, and f_2 provides the representational capacity.

Residual Energy-based Models. An EBM parameterizes a probability density as $p_\theta(x) \propto \exp(-U_\theta(x))$ using a scalar energy function U_θ. A residual EBM tilts a tractable base density $p_0(x)$ by an exponential energy correction:

$$p_\theta(x) = \frac{p_0(x)\exp(-U_\theta(x))}{\mathcal{Z}_\theta}, \quad \mathcal{Z}_\theta = \int p_0(x)\exp(-U_\theta(x))\,\mathrm{d}x, \tag{2}$$

where $\mathcal{Z}_\theta$ is the partition function. p_0 supplies a tractable density, and U_θ adjusts its joint distribution.

4 METHOD

4.1 PREFERENCE GRAPH CONSTRUCTION

Given a prompt c, we sample N candidate responses $\mathcal{V}^c = \{v_i\}_{i=1}^N$ and evaluate them using an ensemble of K RMs $\{\mathcal{R}_k\}_{k=1}^K$. Each $\mathcal{R}_k$ outputs a scalar reward $\mathbf{r}_k(c, v_i) \in \mathbb{R}$ and a penultimate-layer embedding $\mathbf{h}_k(c, v_i) \in \mathbb{R}^{d_r}$. We organize these outputs into a reward matrix $\mathbf{R}^c \in \mathbb{R}^{K\times N}$ and an embedding tensor $\mathbf{H}^c \in \mathbb{R}^{K\times N\times d_r}$, where $\mathbf{R}^c_{k,i} = \mathbf{r}_k(c, v_i)$ and $\mathbf{H}^c_{k,i,:} = \mathbf{h}_k(c, v_i)$. To construct a preference graph $\mathcal{G}^c = (\mathcal{V}^c, \mathbf{A}^c)$, we let a directed edge $v_j \to v_i$ denote that v_j is preferred to v_i. We aggregate positive reward margins into a weighted adjacency matrix $\mathbf{A}^c \in \mathbb{R}^{N\times N}$:

$$\mathbf{A}^c_{ij} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{S}^c_{k,ij}, \qquad \mathbf{S}^c_{k,ij} = \max(\mathbf{r}_k(c, v_j) - \mathbf{r}_k(c, v_i),\, 0), \tag{3}$$

where $\mathbf{S}^c_k \in \mathbb{R}^{N\times N}$ is the k-th view's preference matrix and $\mathbf{A}^c_{ii} = 0$. This construction yields a raw graph $\mathcal{G}^c$ that may contain bidirectional edges, spurious links, and directed cycles.

4.2 GENERALIZED STRUCTURAL MODELING

Multi-view Candidate Encoding. Each RM induces a scalar ranking. We retain its scores, embeddings, and directed margins to encode candidate quality and disagreement across views. By concatenating scalar rewards $\mathbf{R}^c_k \in \mathbb{R}^{N\times 1}$ and candidate embeddings $\mathbf{H}^c_k \in \mathbb{R}^{N\times d_r}$, we apply a GNN encoder g_ψ shared across views:

$$\mathbf{Z}^c_k = g_\psi(\mathbf{R}^c_k \,\|\, \mathbf{H}^c_k, \mathbf{S}^c_k) \in \mathbb{R}^{N\times d_z}, \tag{4}$$

where $\mathbf{Z}^c_{k,i}$ denotes the representation of candidate v_i under the k-th RM. As the RM ensemble carries no meaningful ordering, we pool across views in a manner invariant to permutations. We concatenate

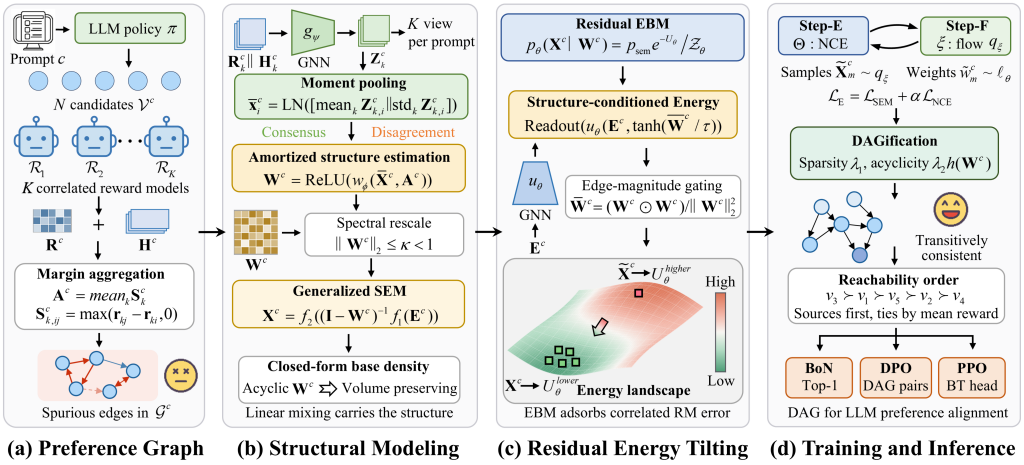

Figure 1: EDIT learns a preference graph from RM margins and embeddings. The SEM defines a tractable base density, the residual energy adjusts its joint density, and graph pruning yields an acyclic ranking structure. The graph head uses pooled features before Gaussian smoothing.

the first two moments and apply layer normalization (LN) across the feature dimension to obtain $\bar{\mathbf{x}}_i^c = \mathrm{LN}([\mathrm{mean}_k\, \mathbf{Z}_{k,i}^c \,\|\, \mathrm{std}_k\, \mathbf{Z}_{k,i}^c])$, where the mean captures consensus and the standard deviation reflects disagreement. For $\sigma > 0$, smoothing with isotropic Gaussian noise gives $\mathbf{x}_i^c = \bar{\mathbf{x}}_i^c + \sigma\boldsymbol{\zeta}_i$ with $\boldsymbol{\zeta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, stacked into $\mathbf{X}^c \in \mathbb{R}^{N\times d}$ with $d = 2d_z$.

Amortized Structure Estimation. Each slate supplies only K correlated views for $N(N-1)$ possible edges. We train a shared structure predictor across prompts to estimate $\mathbf{W}^c$ from recurring patterns across slates. A lightweight MLP w_ϕ scores each ordered pair (i, j) from two node representations and both directional margins:

$$\mathbf{W}_{ij}^c = \mathrm{ReLU}(w_\phi(\bar{\mathbf{x}}_i^c \,\|\, \bar{\mathbf{x}}_j^c \,\|\, \mathbf{A}_{ij}^c \,\|\, \mathbf{A}_{ji}^c)), \tag{5}$$

yielding $\mathbf{W}^c \in \mathbb{R}_{\geq 0}^{N\times N}$, with diagonal entries explicitly set to zero. The two directional margins recover the mean margin by subtraction and the mean absolute margin by addition. A positive $\mathbf{W}_{ij}^c$ predicts $v_j \succ v_i$ within the current slate. We compute $\mathbf{W}^c$ before Gaussian smoothing. For fixed g_ψ and w_ϕ, sampled slates induce a joint law of $(\mathbf{X}^c, \mathbf{W}^c)$ with conditional density $p_{\mathrm{d}}(\mathbf{X} \mid \mathbf{W})$.

Structural Base Density. For a candidate slate, we define the generalized SEM in Eq. 1 over $\mathbf{X}^c \in \mathbb{R}^{N\times d}$ and $\mathbf{W}^c \in \mathbb{R}^{N\times N}$.

$$\mathbf{Y}^c = (\mathbf{I} - \mathbf{W}^c)^{-1} f_1(\mathbf{E}^c), \qquad \mathbf{X}^c = f_2(\mathbf{Y}^c), \tag{6}$$

where $\mathbf{E}^c$ comprises i.i.d. standard Gaussian entries. The row-wise diffeomorphism f_1 transforms node noise, and f_2 maps latent variables to candidate representations. For fixed $\mathbf{W}^c$, inversion gives $\mathbf{Y}^c = f_2^{-1}(\mathbf{X}^c)$ and $\mathbf{E}^c = f_1^{-1}((\mathbf{I} - \mathbf{W}^c)\mathbf{Y}^c)$. Margins and embeddings inform preferences, while SEM likelihood scores cross-candidate dependencies.

Proposition 1. *If f_1, f_2 are diffeomorphisms and $\mathbf{I} - \mathbf{W}^c$ is invertible, then for fixed $\mathbf{W}^c$ the conditional base density induced by Eq. 6 is*

$$p_{\mathrm{sem}}(\mathbf{X}^c \mid \mathbf{W}^c) = |\det(\mathbf{I} - \mathbf{W}^c)|^d \prod_{i=1}^{N} p_\varepsilon(\boldsymbol{\varepsilon}_i^c) |\det \mathbf{J}_{f_1}(\boldsymbol{\varepsilon}_i^c)|^{-1} |\det \mathbf{J}_{f_2}(\mathbf{y}_i^c)|^{-1}, \tag{7}$$

where p_ε is the standard Gaussian density, and $\mathbf{J}_f$ denotes the Jacobian of f, with $\boldsymbol{\varepsilon}_i^c$ and $\mathbf{y}_i^c$ obtained via closed-form inversion. If $\mathbf{W}^c$ is acyclic, it is permutation-similar to a strictly triangular matrix and hence nilpotent, reducing the structural determinant to $\det(\mathbf{I} - \mathbf{W}^c) = 1$.

During training, we rescale $\mathbf{W}^c \leftarrow \kappa \mathbf{W}^c / \max(\kappa, \|\mathbf{W}^c\|_2)$ for $\kappa \in (0,1)$. This bounds the spectral norm by κ and keeps $\mathbf{I} - \mathbf{W}^c$ invertible even before the graph is acyclic. With $\mathbf{W}^c$ fixed before smoothing, the SEM admits direct density evaluation and sampling.

4.3 Residual Energy Tilting

The SEM models cross-candidate dependencies through $\mathbf{W}^c$ and independent node noise. Joint patterns can remain in the recovered residuals. We model them by tilting p_{sem} with a residual energy, as in Eq. 2:

$$p_\theta(\mathbf{X}^c \mid \mathbf{W}^c) = \frac{p_{\mathrm{sem}}(\mathbf{X}^c \mid \mathbf{W}^c)\, \exp(-U_\theta(\mathbf{X}^c, \mathbf{W}^c))}{\mathcal{Z}_\theta(\mathbf{W}^c)}, \tag{8}$$

where $\mathcal{Z}_\theta(\mathbf{W}^c) = \mathbb{E}_{\mathbf{X} \sim p_{\mathrm{sem}}}[\exp(-U_\theta(\mathbf{X}, \mathbf{W}^c))]$ is finite and positive.

Structure-conditioned Energy. We parameterize U_θ over the learned graph:

$$U_\theta(\mathbf{X}^c, \mathbf{W}^c) = \mathrm{Readout}(u_\theta(\mathbf{E}^c, \tanh(\bar{\mathbf{W}}^c/\tau))), \tag{9}$$

where u_θ is a GNN encoder and $\bar{\mathbf{W}}^c = (\mathbf{W}^c \odot \mathbf{W}^c)/\|\mathbf{W}^c\|_2^2$ for $\mathbf{W}^c \neq 0$. We set $\bar{\mathbf{W}}^c = 0$ at $\mathbf{W}^c = 0$. The gate weights messages by edge magnitude, and a sum readout produces a scalar energy. Residuals and graph messages couple the energy correction to the structural transformation.

4.4 Training and Inference

The negative log-likelihood in Eq. 8 has a tractable base term and an energy term with an intractable normalizer:

$$-\log p_\theta(\mathbf{X}^c \mid \mathbf{W}^c) = \underbrace{-\log p_{\mathrm{sem}}(\mathbf{X}^c \mid \mathbf{W}^c)}_{\text{tractable}} + \underbrace{U_\theta(\mathbf{X}^c, \mathbf{W}^c) + \log \mathcal{Z}_\theta(\mathbf{W}^c)}_{\text{intractable}}, \tag{10}$$

We use an alternating procedure with negatives from a conditional flow q_ξ. Step-E fixes ξ and updates the structural and energy modules using the SEM likelihood and NCE objective. Step-F adapts q_ξ by weighted maximum likelihood.

Step E. Structure and energy. Prop. 1 gives the following structural likelihood term, where $\mathrm{sg}[\cdot]$ denotes stop-gradient.

$$\mathcal{L}^c_{\mathrm{SEM}} = -\log p_{\mathrm{sem}}(\mathrm{sg}[\mathbf{X}^c] \mid \mathbf{W}^c). \tag{11}$$

Detaching $\mathbf{X}^c$ blocks the direct SEM gradient to g_ψ. The SEM loss and graph penalties still update g_ψ through $\mathbf{W}^c$, while NCE also uses the observed $\mathbf{X}^c$. We use the observed $\mathbf{X}^c$ as a positive sample and draw M negatives through Eq. 6 with $\tilde{\mathbf{E}}^c_m \sim q_\xi(\cdot \mid \mathrm{sg}[\mathbf{W}^c])$ and $\tilde{\mathbf{X}}^c_m = f_2((\mathbf{I} - \mathrm{sg}[\mathbf{W}^c])^{-1} f_1(\tilde{\mathbf{E}}^c_m))$. The SEM map pushes the flow density into representation space, with shared Jacobians cancelling in the density ratio.

$$\ell_\theta(\mathbf{X}^c) = \log \frac{p_{\mathrm{sem}}(\mathbf{X}^c \mid \mathbf{W}^c)\, e^{-U_\theta(\mathbf{X}^c, \mathbf{W}^c)}}{q_\xi(\mathbf{X}^c \mid \mathbf{W}^c)} = -U_\theta(\mathbf{X}^c, \mathbf{W}^c) - \log \frac{q_\xi(\mathbf{E}^c \mid \mathbf{W}^c)}{p_\varepsilon(\mathbf{E}^c)}. \tag{12}$$

At identity initialization $q_\xi = p_\varepsilon$, negatives follow p_{sem} and $\ell_\theta = -U_\theta$. We weight the positive and negative classes equally:

$$\mathcal{L}^c_{\mathrm{NCE}} = \log\big(1 + e^{-\ell_\theta(\mathbf{X}^c)}\big) + \frac{1}{M} \sum_{m=1}^{M} \log\big(1 + e^{\ell_\theta(\tilde{\mathbf{X}}^c_m)}\big), \tag{13}$$

where ℓ_θ depends on $\mathbf{W}^c$ through both the energy U_θ and the recovered residual $\mathbf{E}^c$. The contrastive term therefore reaches the structure along two paths.

Proposition 2. *Fix g_ψ, w_ϕ, f_1, f_2, and q_ξ. If the support of $q_\xi(\cdot \mid \mathbf{W})$ covers that of $p_{\mathrm{d}}(\cdot \mid \mathbf{W})$ and the energy is realizable, every population minimizer of Eq. 13 over θ satisfies $p_{\mathrm{sem}}(\mathbf{X} \mid \mathbf{W}) e^{-U_\theta(\mathbf{X}, \mathbf{W})} = p_{\mathrm{d}}(\mathbf{X} \mid \mathbf{W})$ for almost every $\mathbf{W}$. Hence $\mathcal{Z}_\theta(\mathbf{W}) = 1$ almost everywhere.*

Step F. Adaptive flow sampler. We adapt q_ξ toward the current model by minimizing $\mathrm{KL}(p_\theta \| q_\xi)$. The logits provide importance weights $e^{\ell_\theta(\tilde{\mathbf{X}}^c)} \propto p_\theta(\tilde{\mathbf{X}}^c \mid \mathbf{W}^c)/q_\xi(\tilde{\mathbf{X}}^c \mid \mathbf{W}^c)$ for a self-normalized estimate:

$$\mathcal{L}^c_{\mathrm{F}}(\xi) = -\sum_{m=1}^{M} \tilde{w}^c_m \log q_\xi(\tilde{\mathbf{E}}^c_m \mid \mathbf{W}^c), \qquad \tilde{w}^c_m = \mathrm{softmax}\, \ell_\theta(\tilde{\mathbf{X}}^c_m), \tag{14}$$

Step-F fixes the cached residuals, graphs, and normalized weights. For a fixed SEM map, residual and representation likelihoods differ by a ξ-independent Jacobian term. Self-normalization cancels the partition function. At a realizable conditional population fixed point, $q_\xi = p_\theta = p_{\mathrm{d}}$ in representation space and $\mathcal{L}_{\mathrm{NCE}} = 2\log 2$. The SEM likelihood and graph penalties then determine the structural objective.

DAGification. We penalize directed cycles using the polynomial function from Yu et al. (2019).

$$h(\mathbf{W}^c) = \mathrm{tr}\big[(\mathbf{I} + \gamma\, \mathbf{W}^c \odot \mathbf{W}^c)^N\big] - N, \tag{15}$$

where $\gamma > 0$ and $\odot$ denotes elementwise multiplication. The squared edge weights give nonnegative contributions to closed walks, so $h(\mathbf{W}^c) = 0$ exactly when $\mathbf{W}^c$ is acyclic.

Overall Objective. Let $\Theta = \{\psi, \phi, \theta, f_1, f_2\}$ exclude the flow parameters ξ. For a minibatch $\mathcal{B}$, Step-E updates Θ with

$$\mathcal{L}_{\mathrm{E}}(\Theta) = \frac{1}{|\mathcal{B}|} \sum_{c \in \mathcal{B}} \Big[\mathcal{L}^c_{\mathrm{SEM}} + \alpha\, \mathcal{L}^c_{\mathrm{NCE}} + \lambda_1 \|\mathbf{W}^c\|_1 + \lambda_2\, h(\mathbf{W}^c)\Big], \tag{16}$$

This composite objective uses SEM likelihood to score cross-candidate dependencies and NCE to fit the residual density. Sparsity and acyclicity penalties regularize $\mathbf{W}^c$, whose directional inputs supply preference evidence. We use a shallow energy encoder and anneal α from zero. The coefficients λ_1 and λ_2 control the graph penalties. Step-F updates ξ using the same cached negatives.

Algorithm 1: Step-E/F Alternating Training of EDIT

Input: Rewards $\mathbf{R}^c$ and embeddings $\mathbf{H}^c$ from RM outputs, view encoder g_ψ, structure head w_ϕ, energy function u_θ, row-wise diffeomorphisms f_1, f_2, flow sampler q_ξ, M negatives.
Output: Trained EDIT framework with base parameters Θ and flow parameters ξ.
Initialization : Initialize Θ, set $q_\xi = p_\varepsilon$, and start the annealing of α at 0.
while *not converged* **do**
  Sample a minibatch $\mathcal{B}$ of prompts
  for $c \in \mathcal{B}$ **do**
    Encode and pool the K RM views to obtain $\bar{\mathbf{X}}^c$, then smooth to obtain $\mathbf{X}^c$
    Predict $\mathbf{W}^c$ via w_ϕ with Eq. 5, then rescale $\mathbf{W}^c \leftarrow \kappa\, \mathbf{W}^c / \max(\kappa, \|\mathbf{W}^c\|_2)$
    Invert the SEM for $(\mathbf{Y}^c, \mathbf{E}^c)$ and evaluate $\mathcal{L}^c_{\mathrm{SEM}}$ in Eq. 11
    Draw M negatives $\tilde{\mathbf{X}}^c_m$ from $q_\xi(\cdot \mid \mathrm{sg}[\mathbf{W}^c])$ along the SEM path in Eq. 6
    Cache the logits $\ell_\theta(\mathbf{X}^c)$ and $\ell_\theta(\tilde{\mathbf{X}}^c_m)$ with Eq. 12, and accumulate $\mathcal{L}^c_{\mathrm{NCE}}$ with Eq. 13
  Step E. Update Θ by one gradient step on the joint objective $\mathcal{L}_{\mathrm{E}}$ in Eq. 16
  Step F. Fix the cached residuals, graphs, and normalized weights, then update ξ on $\mathcal{L}_{\mathrm{F}}$ in Eq. 14

Inference. At test time, we convert $\mathbf{W}^c$ into a valid DAG by thresholding $\mathbf{W}^c_{ij}$ and iteratively pruning the weakest edges in any remaining cycle until the edge support is acyclic. For a non-negative acyclic $\mathbf{W}^c$, the expansion $(\mathbf{I} - \mathbf{W}^c)^{-1} = \sum_{t<N} (\mathbf{W}^c)^t$ is entrywise non-negative and for $i \neq j$, $[(\mathbf{I} - \mathbf{W}^c)^{-1}]_{ij} > 0$ exactly when v_j reaches v_i. Reachability therefore defines a partial preference order over $\mathcal{V}^c$. We perform a topological sort, prioritizing larger mean RM rewards among currently available source nodes.

5 Experiments

5.1 Experimental Setup

Evaluation and Baselines. We evaluate response selection with pairwise and Best-of-N (BoN) accuracy on RMB Zhou et al. (2025), which covers 49 helpfulness and harmlessness scenarios. For policy optimization, we follow AlpacaFarm's gold RM protocol Dubois et al. (2023) and measure gold reward and KL divergence under BoN, DPO, and PPO. Synthetic slates with a known sparse preference DAG evaluate structural Hamming distance (SHD) under shared bias ρ. Baselines include single RMs Ouyang et al. (2022), RM ensembles Coste et al. (2024), enhanced RMs Chen et al. (2024); Hong et al. (2025a); Miao et al. (2024); Yang et al. (2024); Lochab & Zhang (2025), and graph aggregation methods Zhao et al. (2026); Kadekodi et al. (2025); Hu et al. (2026).

Implementation Details. For policy optimization, we follow Coste et al. (2024) with a 1.4B Pythia SFT policy, a 7B AlpacaFarm RM as the gold judge, and K=5 Pythia-70M RMs trained identically with distinct seeds. For response selection, we use the 8B Skywork RM Liu et al. (2024). EDIT requires no additional preference labels and is trained on ensemble outputs over 10K AlpacaFarm prompts disjoint from evaluation, with N=8 responses sampled at temperature 1.0. For BoN evaluation, RMB uses N=16 and AlpacaFarm uses $N \in \{16, 64\}$, with the pruning threshold set to 10% of the strongest positive edge weight before DAGification. BoN takes the top-ranked candidate. For DPO, we pair candidates from opposite ends of the topological order, yielding four preference pairs at N=8. For PPO, we distill these pairs into a Bradley-Terry reward head that supplies the scalar reward.

The RM feature dimension is d_r=512. The view encoder g_ψ and energy function u_θ are 2-layer GCNs with 128 hidden and 32 output dimensions. The structure head w_ϕ is a 3-layer MLP with GELU activation. The mappings f_1 and f_2 each stack 4 row-wise affine coupling bijections parameterized by 2-layer MLPs, and q_ξ is a 4-step conditional Glow-style flow Kingma & Dhariwal (2018). During optimization, we sample M=64 negatives per prompt and train for 5 epochs using AdamW with learning rate 2×10^{-4} and batch size 32, where Step-F updates ξ with a separate AdamW at the same setting. We set κ=0.9, τ=0.5, and λ_1=0.05, linearly warm up λ_2 to 1.0 over the first 20% of steps, and anneal α from 0 to 1.0 over the first 30%. PPO executes for 3000 steps with β_{KL}=0.01.

Table 1: Response selection on RMB and policy optimization on AlpacaFarm. RMB Avg. averages the four accuracy metrics. Gold denotes the gold-RM win rate (%) over the SFT policy. DPO and PPO report peak win rates and the corresponding KL. Accuracies and win rates are mean±std over three seeds. Best results are in bold.

Method	Response Selection on RMB					Policy Optimization on AlpacaFarm					
	Harmlessness ↑		Helpfulness ↑		Avg.	BoN Gold ↑		DPO		PPO	
	Pairwise	BoN	Pairwise	BoN		n=16	n=64	Gold ↑	KL ↓	Gold ↑	KL ↓
Single RM	52.1 ±0.41	36.8 ±0.52	48.6 ±0.47	33.1 ±0.58	42.7	62.4 ±0.84	65.2 ±0.88	63.5 ±0.79	12.1	67.1 ±0.93	72.6
ENS-Mean	52.4 ±0.35	35.5 ±0.44	48.3 ±0.39	31.4 ±0.51	41.9	64.1 ±0.78	67.6 ±0.82	65.2 ±0.63	11.7	68.4 ±0.88	67.8
ENS-WCO	53.5 ±0.38	36.6 ±0.49	45.4 ±0.45	29.0 ±0.62	41.1	64.9 ±0.62	68.5 ±0.77	66.0 ±0.74	11.5	69.2 ±0.74	63.9
ENS-UWO	52.9 ±0.36	35.5 ±0.47	47.2 ±0.42	30.9 ±0.55	41.6	65.3 ±0.59	69.0 ±0.74	66.5 ±0.60	11.4	70.0 ±0.71	60.4
ODIN	54.9 ±0.44	37.6 ±0.51	49.8 ±0.46	33.8 ±0.57	44.0	65.0 ±0.75	68.6 ±0.79	66.1 ±0.72	11.7	69.6 ±0.86	62.3
BSR	55.7 ±0.41	38.0 ±0.48	50.7 ±0.43	33.5 ±0.54	44.5	65.4 ±0.72	69.1 ±0.76	66.6 ±0.70	11.5	70.1 ±0.84	59.1
InfoRM	56.0 ±0.40	38.5 ±0.52	51.2 ±0.43	33.7 ±0.55	44.9	65.9 ±0.58	69.5 ±0.71	67.0 ±0.57	11.3	70.5 ±0.69	56.2
GRM	56.7 ±0.38	38.6 ±0.49	51.8 ±0.42	34.0 ±0.53	45.3	66.2 ±0.56	69.9 ±0.69	67.3 ±0.55	11.2	70.8 ±0.67	55.4
EBRM	58.1 ±0.42	39.4 ±0.53	53.0 ±0.44	33.4 ±0.59	46.0	66.6 ±0.69	70.2 ±0.73	67.7 ±0.68	11.6	71.2 ±0.81	57.8
SPA	56.9 ±0.41	37.6 ±0.53	51.8 ±0.45	32.6 ±0.60	44.7	65.8 ±0.71	69.4 ±0.75	66.9 ±0.69	11.2	70.4 ±0.68	56.9
CARE	56.3 ±0.39	38.2 ±0.50	51.4 ±0.44	33.4 ±0.56	44.8	65.5 ±0.74	69.0 ±0.78	66.8 ±0.71	11.3	70.3 ±0.79	58.1
PGED	58.6 ±0.44	40.0 ±0.55	53.2 ±0.47	33.9 ±0.58	46.4	67.3 ±0.67	70.8 ±0.70	68.4 ±0.65	11.0	71.9 ±0.75	53.2
EDIT	**62.2** ±0.29	**44.1** ±0.38	**55.8** ±0.33	**35.7** ±0.44	**49.5**	**70.1** ±0.42	**74.4** ±0.44	**72.3** ±0.40	**10.1**	**75.2** ±0.48	**44.5**

5.2 Main Results

Response Selection. EDIT achieves the highest accuracy on all four RMB metrics in Table 1. Its average of 49.5% exceeds PGED by 3.1 percentage points. The gains are 3.6 and 4.1 points on harmlessness pairwise and BoN evaluation, and 2.6 and 1.8 points on helpfulness. Scalar ensembles average 41.1% to 41.9%, compared with 42.7% for the single RM. Enhanced RMs improve the average to 44.0% to 46.0%, while PGED reaches 46.4%. EDIT improves both pairwise comparison and selection from a candidate pool, with its largest gain over PGED on harmlessness BoN. These results support jointly using directional margins and candidate representations for response selection.

Policy Optimization. EDIT improves AlpacaFarm BoN win rates over PGED by 2.8 points at N=16 and 3.6 points at N=64. Under DPO, it reaches 72.3% versus 68.4%, with corresponding KL values of 10.1 and 11.0. Under PPO, it reaches 75.2% versus 71.9%, with KL values of 44.5 and 53.2. Thus, the PPO peak improves by 3.3 points with 16.4% lower corresponding KL. The gains extend from direct candidate ranking to training on graph-derived preference pairs and optimizing a distilled reward head. The BoN improvement at both pool sizes also shows that the ranking model transfers beyond the eight-candidate training slates.

5.3 Reward Overoptimization

To examine reward overoptimization, we increase the BoN budget, track PPO over its full schedule, and retrain the K RMs with 25% label noise. In Fig. 2(a), baseline gold scores rise and then decline as the selection budget grows. EDIT remains within 5% of its peak at the largest budget. Shared label flips in Fig. 2(b) deepen the baseline declines and increase EDIT's advantage at large budgets. Fig. 2(c) traces gold scores against policy KL over 3000 PPO steps. EDIT reaches its plateau at lower KL and sustains higher gold scores across the plotted range. Fig. 2(d) varies the KL penalty. At zero penalty, EDIT retains roughly 85% of its peak gold score, while the baselines decline more. Within each method, continued proxy growth alongside declining gold scores indicates overoptimization. EDIT delays this decline under both increased selection pressure and policy optimization.

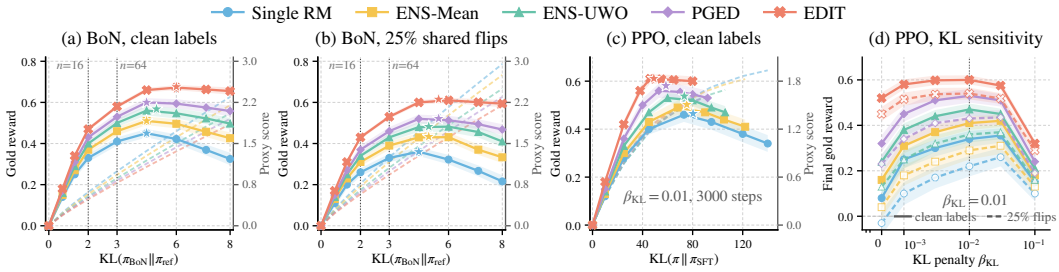

Figure 2: Gold and proxy scores under increasing optimization pressure on AlpacaFarm. Proxy curves use each scalar baseline's reward, the ensemble-mean reward of graph-selected responses under BoN, and the distilled reward head under PPO.

5.4 Correlated View Corruption

At a matched total flip budget, Fig. 3 shows larger losses under shared than dispersed corruption for every method. The setting m=0 is clean. When the corrupted views first form a majority, PGED loses 4.1 points, compared with 0.9 for EDIT. At m=5, the shared-versus-dispersed gap reaches 6.0 points for PGED and 3.9 for EDIT. EDIT retains a 68.8% BoN win rate when all five views are corrupted. The mean of the remaining clean views stays below EDIT for every $m < K$.

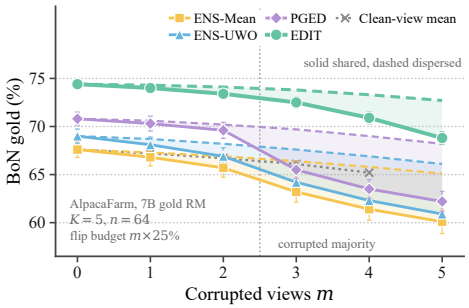

Figure 3: Shared and dispersed corruption at a matched budget. Shared flips target the same preference pairs in all m corrupted views, and dispersed flips spread independently over all K views.

5.5 Ablation Study

Component Analysis. Table 2 evaluates objective terms and structural designs. Removing the amortized head causes the largest pairwise accuracy drop, from 62.2% to 56.4%, and increases the cycle rate from 2.1% to 11.8%. Removing the energy term or SEM likelihood lowers pairwise accuracy by 3.6 or 4.2 points, respectively. Removing flow adaptation costs 2.8 points. Without the acyclicity penalty, accuracy falls to 58.3% and the cycle rate reaches 17.3%, the highest among the variants. Cycle rates are measured after thresholding and before cycle pruning. The final ranking graph is acyclic in every variant that applies this pruning. Nonnegative weights, cross-view standard deviations, and unsmoothed graph inputs each improve the reported ranking metrics. Removing the standard deviation also reduces PPO gold win rate from 75.2% to 72.3%, linking the disagreement features to downstream policy performance. Together, the ablations show contributions from representation fusion, structural regularization, and residual density training.

Sensitivity Analysis. Fig. 4 varies one factor at a time. RMB accuracy improves up to K=8 and then changes little, with EDIT outperforming the plotted baselines. Models trained at different pool sizes retain strong BoN performance as the inference pool grows. The conditional flow with M=16 matches the fixed Gaussian proposal with M=128, indicating more efficient use of negative samples.

Table 2: Component analysis of EDIT. The upper block removes objective terms and the lower block alters structural designs. RMB results are on harmlessness and AlpacaFarm results under PPO. Cycle rate is measured after thresholding and before cycle pruning.

Ablation	RMB		AlpacaFarm		Cycle
	Pairwise ↑	BoN ↑	Gold ↑	KL ↓	Rate ↓
w/o U_θ	58.6	39.9	71.6	54.3	4.6
w/o $\mathcal{L}_{\text{SEM}}$	58.0	39.3	70.9	56.1	7.9
w/o $\mathcal{L}_{\text{F}}$	59.4	41.0	72.8	50.7	3.4
w/o std_k	58.9	40.4	72.3	51.6	5.4
w/o w_ϕ	56.4	37.6	69.3	60.2	11.8
w/o $h(\mathbf{W}^c)$	58.3	39.1	71.9	52.8	17.3
signed $\mathbf{W}^c$	59.3	41.2	73.1	49.8	3.8
$\mathbf{W}^c$ from $\mathbf{X}^c$	59.8	41.6	73.6	48.2	2.9
EDIT	**62.2**	**44.1**	**75.2**	**44.5**	**2.1**

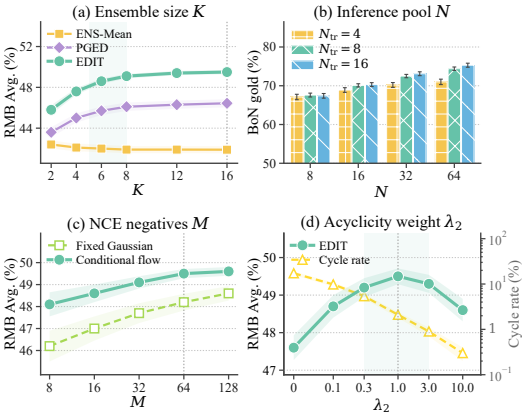

Figure 4: Sensitivity analysis of EDIT. Each panel varies a single factor around the defaults. RMB results report the average over the four metrics and AlpacaFarm results report BoN gold-RM win rate, where 50 is a tie with the SFT policy.

Accuracy peaks at λ_2=1.0. Larger values further reduce the cycle rate but lower ranking accuracy, showing a tradeoff between the strength of the structural penalty and predictive performance.

5.6 Structure Analysis

Table 3 compares learned edges with a known sparse DAG. At ρ=0, EDIT and its energy ablation obtain SHD values of 4.3 and 4.6. At ρ=0.5, the gap grows to 6.1 versus 12.7, linking the energy correction to improved structure recovery under shared bias. Removing or shuffling margin inputs changes $\mathcal{L}_{\text{SEM}}$ by less than 0.02 nats per dimension but reduces RMB Kendall τ_b to 0.03 or 0.01. Thus, directional margins supply ranking information that is weakly reflected in the density score. On AlpacaFarm, EDIT reaches 83.4% precision at 60% reachability coverage in Fig. 5(a). Matching coverage compares the precision of resolved pairs at the same degree of comparability. Fig. 5(b) groups pairs by the number of RMs opposing the gold direction. EDIT retains higher accuracy than the plotted baselines in each group, including pairs opposed by all five RMs. On RMB, EDIT achieves the highest τ_b and BoN gold score in Table 3.

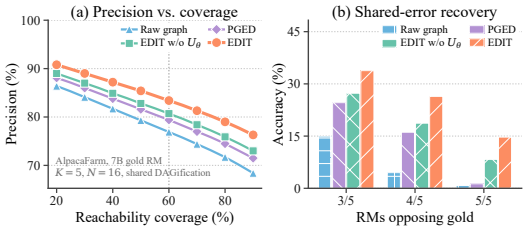

Figure 5: Structure analysis on AlpacaFarm. (a) Precision at matched reachability coverage. (b) Pairwise accuracy when m of the K RMs oppose the gold direction.

Table 3: Structure recovery on synthetic prompts with a known sparse DAG, where ρ scales the bias shared across K views. Ranking fidelity on RMB, where BoN gold scores the top-ranked candidate.

Method	Synthetic SHD ↓		RMB	
	ρ=0	ρ=0.5	τ_b ↑	BoN Gold ↑
Raw graph	18.4	24.1	0.35	38.5
EDIT w/o U_θ	4.6	12.7	0.43	44.6
EDIT w/o margin	8.7	10.4	0.03	9.4
EDIT w/ shuffled margin	9.1	11.2	0.01	7.8
EDIT	**4.3**	**6.1**	**0.48**	**47.2**

6 Conclusion

In this paper, we propose an EDIT framework, which combines continuous RM margins and candidate embeddings to learn prompt-specific preference graphs. A structural likelihood models cross-candidate dependencies, and a residual energy correction fits remaining joint patterns. Conditional flow sampling supports alternating training, while acyclicity regularization and pruning yield rankings for response selection and preference-pair construction. Across the evaluated settings, EDIT improves response selection, policy optimization, and robustness to correlated corruption.

AI USE STATEMENT

We used generative AI tools for literature discovery, manuscript editing, and feedback on methodological exposition and the interpretation of existing results. This assistance included revising passages for clarity, concision, and consistency. We did not use these tools to run experiments or generate the reported measurements. The authors are responsible for verifying all AI-assisted text, citations, mathematical statements, and interpretations and for the final content of the paper.

ETHICS STATEMENT

All authors have read and agree to follow the ICLR Code of Ethics. This work aims to advance representation learning for the public good. We use only public data and do not process personal data. We respect licenses and uphold fairness, privacy, security, and research integrity.

REPRODUCIBILITY STATEMENT

We plan to release the implementation, training configurations, and evaluation scripts, together with identifiers for the datasets and RM checkpoints used in each experiment.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41, 2021.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. In *International Conference on Machine Learning*, pp. 7935–7952. PMLR, 2024.

David Maxwell Chickering. Learning Bayesian networks is NP-complete. In Doug Fisher and Hans-Joachim Lenz (eds.), *Learning from Data: Artificial Intelligence and Statistics V*, volume 112 of *Lecture Notes in Statistics*, pp. 121–130. Springer, 1996.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Jiwoo Hong, Noah Lee, Eunki Kim, Guijin Son, Woojin Chung, Aman Gupta, Shao Tang, and James Thorne. On the robustness of reward models for language model alignment. In *International Conference on Machine Learning*, pp. 23682–23699. PMLR, 2025a.

Yuzhong Hong, Hanshan Zhang, Junwei Bao, Hongfei Jiang, and Yang Song. Energy-based preference model offers better offline alignment than the bradley-terry preference model. In *International Conference on Machine Learning*, pp. 23787–23804. PMLR, 2025b.

Zhengyu Hu, Jieyu Zhang, Zhihan Xiong, Alexander Ratner, Kaize Ding, and Ranjay Krishna. Towards acyclic preference evaluation of language models via multiple evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21903–21911, 2026.

Shreyas Kadekodi, Hayden McTavish, and Berk Ustun. Selective preference aggregation. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

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

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, et al. Rrm: Robust reward model training mitigates reward hacking. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Anamika Lochab and Ruqi Zhang. Energy-based reward models for robust language model alignment. In *Second Conference on Language Modeling*, 2025.

Yunsheng Lu, Zijiang Yang, Licheng Pan, and Zhixuan Chu. Robust reward modeling for large language models via causal decomposition. *arXiv preprint arXiv:2604.13833*, 2026.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37:134387–134429, 2024.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Bo Pang, Caiming Xiong, and Yingbo Zhou. Arm: Alignment with residual energy-based model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8225–8236, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. Fjmp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

Hao Wang, Licheng Pan, Zhichao Chen, Chunyuan Zheng, Zhixuan Chu, Xiaoxi Li, Yuan Lu, Xinggao Liu, Haoxuan Li, and Zhouchen Lin. Causalrm: Causal-theoretic reward modeling for rlhf from observational user feedbacks. *arXiv preprint arXiv:2603.18736*, 2026.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-based diffusion language models for text generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. *Advances in Neural Information Processing Systems*, 37:62279–62309, 2024.

Yupei Yang, Lin Yang, Wanxi Deng, Lin Qu, Fan Feng, Biwei Huang, Shikui Tu, and Lei Xu. Factored causal representation learning for robust reward modeling in rlhf. *arXiv preprint arXiv:2601.21350*, 2026.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. Care: Confounder-aware aggregation for reliable llm evaluation. *arXiv preprint arXiv:2603.00039*, 2026.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, et al. Rmb: Comprehensively benchmarking reward models in llm alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hao Zhu, Di Zhou, and Donna Slonim. Smoothing the landscape: Causal structure learning via diffusion denoising objectives. In *Proceedings of the Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, 2026.

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.

A Proofs and Additional Derivations

A.1 Proof of Proposition 1

Fix a graph $\mathbf{W}$ and let $\mathbf{B} = \mathbf{I} - \mathbf{W}$. Define

$$T_{\mathbf{W}}(\mathbf{E}) = f_2\big(\mathbf{B}^{-1} f_1(\mathbf{E})\big), \qquad T_{\mathbf{W}}^{-1}(\mathbf{X}) = f_1^{-1}\big(\mathbf{B} f_2^{-1}(\mathbf{X})\big). \tag{17}$$

Under the assumptions of Proposition 1, both maps are differentiable bijections. For column-wise vectorization, the Jacobian of $\mathbf{Y} \mapsto \mathbf{B}\mathbf{Y}$ is $\mathbf{I}_d \otimes \mathbf{B}$, whose determinant is $\det(\mathbf{B})^d$. Because f_1 and f_2 act row-wise, the chain rule gives

$$\left|\det J_{T_{\mathbf{W}}^{-1}}(\mathbf{X})\right| = |\det \mathbf{B}|^d \prod_{i=1}^{N} |\det J_{f_1}(\boldsymbol{\varepsilon}_i)|^{-1} |\det J_{f_2}(\mathbf{y}_i)|^{-1}. \tag{18}$$

Applying the change-of-variables formula to the independent Gaussian rows of $\mathbf{E}$ yields Eq. 7. If $\mathbf{W}$ is acyclic, a simultaneous permutation of its rows and columns makes it strictly triangular, so $\mathbf{I} - \mathbf{W}$ is triangular with unit diagonal and $\det(\mathbf{I} - \mathbf{W}) = 1$. □

The spectral rescaling in the main text gives $\|\mathbf{W}\|_2 \le \kappa < 1$, hence

$$(\mathbf{I} - \mathbf{W})^{-1} = \sum_{t=0}^{\infty} \mathbf{W}^t, \qquad \|(\mathbf{I} - \mathbf{W})^{-1}\|_2 \le (1 - \kappa)^{-1}. \tag{19}$$

The SEM map therefore stays invertible throughout training, including the early iterations in which the predicted graph still contains cycles. We never form the inverse explicitly. Each slate is factorized once with a partially pivoted LU decomposition, the log-determinant is read off its diagonal, and the same factorization serves the positive sample and all M negatives.

A.2 Proof of Proposition 2

All densities are taken with respect to Lebesgue measure on $\mathbb{R}^{N \times d}$. Fix g_ψ, w_ϕ, f_1, f_2, and the proposal, and write

$$a_\theta(\mathbf{X} \mid \mathbf{W}) = p_{\mathrm{sem}}(\mathbf{X} \mid \mathbf{W})\, e^{-U_\theta(\mathbf{X}, \mathbf{W})}. \tag{20}$$

We use the positive-density form of the proposal condition, $q(\mathbf{X} \mid \mathbf{W}) > 0$ almost everywhere, which is what the full-space conclusion requires. The energy is realizable, including a graph-dependent additive offset that absorbs the normalization.

For fixed $\mathbf{W}$, abbreviate the data and proposal densities by p and q. The population balanced NCE loss is

$$L(a) = \int \left[p \log\!\left(1 + \frac{q}{a}\right) + q \log\!\left(1 + \frac{a}{q}\right) \right] \mathrm{d}\mathbf{X}. \tag{21}$$

Let $\eta = p/(p+q)$ and $s = a/(a+q)$. Subtracting the loss at $a = p$ gives

$$L(a) - L(p) = \int (p+q)\, \mathrm{KL}(\mathrm{Bern}(\eta) \,\|\, \mathrm{Bern}(s))\, \mathrm{d}\mathbf{X} \ge 0. \tag{22}$$

Equality holds only when $s = \eta$ almost everywhere, and $q > 0$ almost everywhere then forces $a = p$ almost everywhere. Integrating the nonnegative gap over the graph distribution gives the same conclusion for almost every $\mathbf{W}$. Realizability attains this minimum, so

$$\mathcal{Z}_\theta(\mathbf{W}) = \int a_\theta(\mathbf{X} \mid \mathbf{W})\, \mathrm{d}\mathbf{X} = 1 \tag{23}$$

almost everywhere. □

Both densities in the statement satisfy the positivity condition. Gaussian smoothing is applied after the graph is predicted, so for $\sigma > 0$

$$p_{\mathrm{d}}(\mathbf{X} \mid \mathbf{W}) = \mathbb{E}\big[\varphi_\sigma(\mathbf{X} - \bar{\mathbf{X}}) \mid \mathbf{W}\big] > 0, \tag{24}$$

where φ_σ is the Nd-dimensional isotropic Gaussian density, and the invertible flow with a Gaussian base is positive by construction. Proposition 2 characterizes the fitted conditional density for fixed representation and structural modules.

Equation 13 averages the M negative terms, so the two classes carry equal total weight and no $\log M$ correction appears in the logit. At the realizable fixed point $q = p = a$, every logit is zero and $\mathcal{L}_{\mathrm{NCE}} = 2 \log 2$, which is the value we use as a convergence reference.

A.3 Residual Coordinates and the Flow Update

Let $J_{\mathbf{W}}(\mathbf{X}) = |\det J_{T_{\mathbf{W}}^{-1}}(\mathbf{X})|$. The proposal induced in representation space is

$$q_\xi^X(\mathbf{X} \mid \mathbf{W}) = q_\xi^E(T_{\mathbf{W}}^{-1}(\mathbf{X}) \mid \mathbf{W})\, J_{\mathbf{W}}(\mathbf{X}). \tag{25}$$

The same Jacobian appears in p_{sem}, so their ratio reduces exactly to Eq. 12 and the logit needs only the energy and the flow-to-base density ratio. In residual coordinates the current model is

$$p_\theta^E(\mathbf{E} \mid \mathbf{W}) = \frac{p_\varepsilon(\mathbf{E})\, \exp[-U_\theta(T_{\mathbf{W}}(\mathbf{E}), \mathbf{W})]}{\mathcal{Z}_\theta(\mathbf{W})}. \tag{26}$$

For a fixed SEM map, minimizing $\mathrm{KL}(p_\theta^X \| q_\xi^X)$ is equivalent, up to a ξ-independent constant, to minimizing $-\mathbb{E}_{p_\theta^E} \log q_\xi^E$. Importance sampling from the cached proposal yields weights proportional to $\exp(\ell_\theta)$, whose self-normalization gives Eq. 14. The self-normalized estimator carries an $O(1/M)$ bias and is consistent under standard integrability conditions.

Algorithm 1 caches residual samples, graphs, and normalized weights before Step-F. These quantities are constants in the flow update and differentiation acts only on $\log q_\xi^E$, which keeps Step-F a fixed-target weighted maximum-likelihood step. Step-E follows the matching convention. Negatives are detached after generation, and their logits are recomputed with the current structural and energy parameters while the flow parameters stay fixed. The sampling law is therefore held fixed inside the contrastive update, and the gradient reaches $\mathbf{W}^c$ only through the energy and the recovered residual.

The finite-normalizer condition follows from the parameterization. The row-wise coupling maps use bounded log-scales and finite-depth GELU networks, so their outputs grow at most linearly in $\|\mathbf{E}\|_F$. For fixed $\mathbf{W}$ and finite parameters, the shallow energy network with a linear sum readout therefore satisfies $|U_\theta(T_{\mathbf{W}}(\mathbf{E}), \mathbf{W})| \le C_0 + C_1\|\mathbf{E}\|_F$. Gaussian residuals have finite exponential moments of their norm, which gives $0 < \mathcal{Z}_\theta(\mathbf{W}) < \infty$.

A.4 Acyclicity, Reachability, and Permutation Properties

Let $\mathbf{C} = \mathbf{W} \odot \mathbf{W}$. Expanding Eq. 15 yields

$$h(\mathbf{W}) = \sum_{t=1}^{N} \binom{N}{t} \gamma^t \operatorname{tr}(\mathbf{C}^t). \tag{27}$$

Every term is nonnegative. A directed cycle of length $t \le N$ contributes a positive product to $\operatorname{tr}(\mathbf{C}^t)$, while an acyclic graph has no closed directed walk. Hence $h(\mathbf{W}) = 0$ if and only if the support of $\mathbf{W}$ is acyclic. For an acyclic nonnegative $\mathbf{W}$ we have $\mathbf{W}^N = 0$ and

$$(\mathbf{I} - \mathbf{W})^{-1} = \sum_{t=0}^{N-1} \mathbf{W}^t, \tag{28}$$

whose (i, j) entry for $i \ne j$ is positive exactly when a directed path leads from j to i. Reachability therefore defines the partial order used at inference. The pruning loop removes one edge per iteration, so it terminates after at most $|E|$ steps and returns an acyclic support. Removing only edges that lie inside a cycle leaves every acyclic reachability relation of the thresholded graph intact.

The shared view encoder and pair head are equivariant to candidate permutations, and moment pooling is invariant to permutations of the RM views. When heterogeneous RMs use model-specific adapters, each adapter is permuted together with its RM, so equivariance is preserved. Gaussian smoothing preserves candidate equivariance in distribution. Deterministic inference breaks exact ties by stable response identifiers, so the returned order does not depend on how the slate is presented.