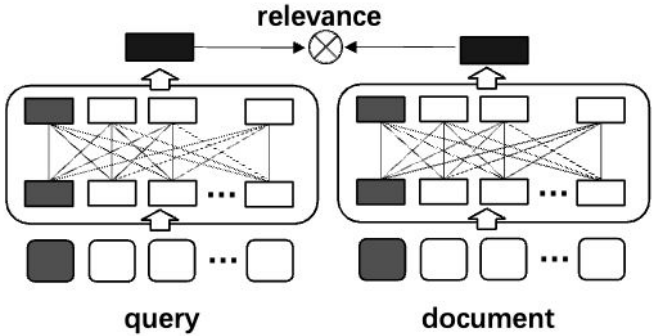
GRACE: GENERATIVE REPRESENTATION LEARNING VIA CONTRASTIVE POLICY OPTIMIZATION

Jiashuo Sun[1] **Shixuan Liu**[2] **Zhaochen Su**[3] **Xianrui Zhong**[1] **Pengcheng Jiang**[1]
Bowen Jin[1] **Peiran Li**[4] **Weijia Shi**[5] **Jiawei Han**[1]

[1]University of Illinois Urbana–Champaign [2]Australian National University
[3]Hong Kong University of Science and Technology [4]University of Wisconsin–Madison
[5]University of Washington

Under review as a conference paper at ICLR 2026

Traditional Training for Representation Model—Contrastive Learning

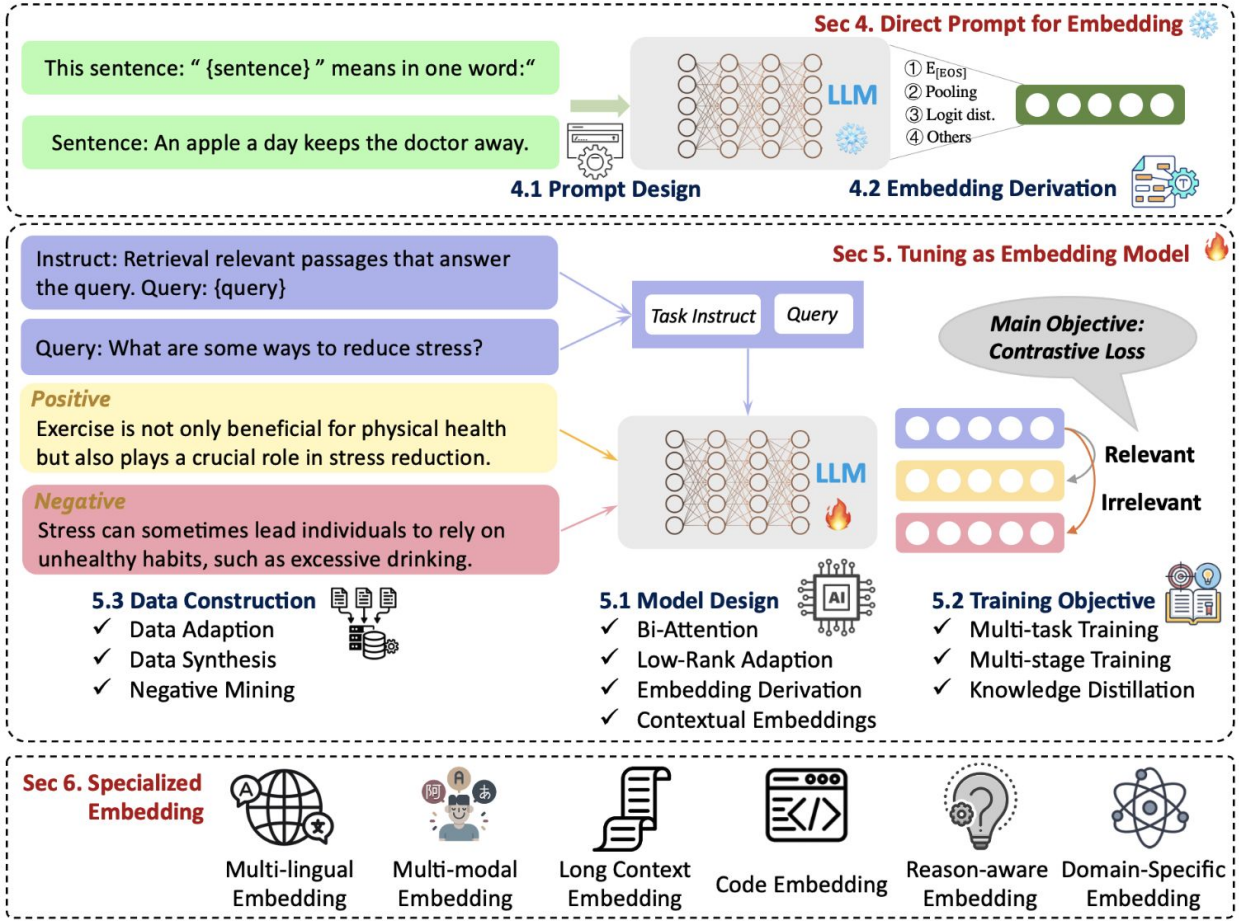


(a) Retriever

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}$$

Representation Model: BERT, RoBERTa, DeBERTa

Current Training for Representation Model—Contrastive Learning



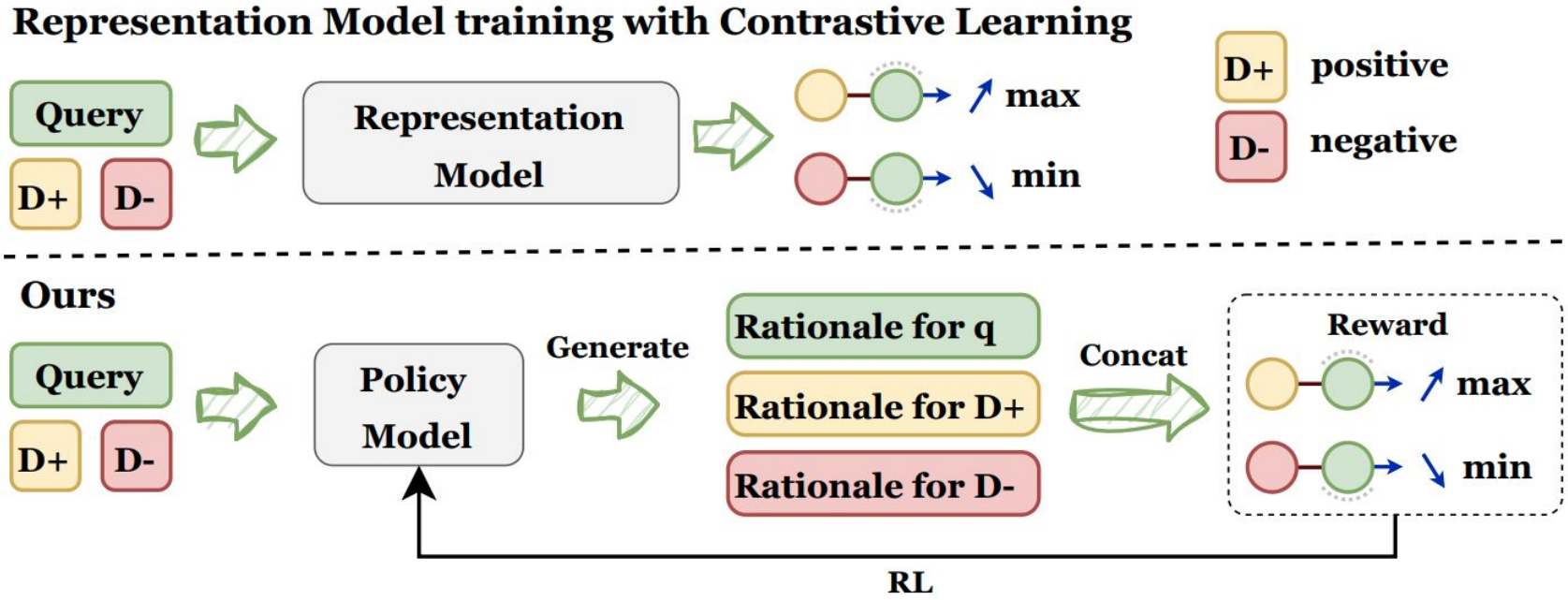
Issues and Chanllenges

1. LLM merely as a parameterized encoder function $f_\theta : \mathcal{X} \to \mathbb{R}^d$
Forces generative models to produce static embedding vectors through simple pooling mechanisms.

2. Suppressing their capacity for structured reasoning and natural language generation.

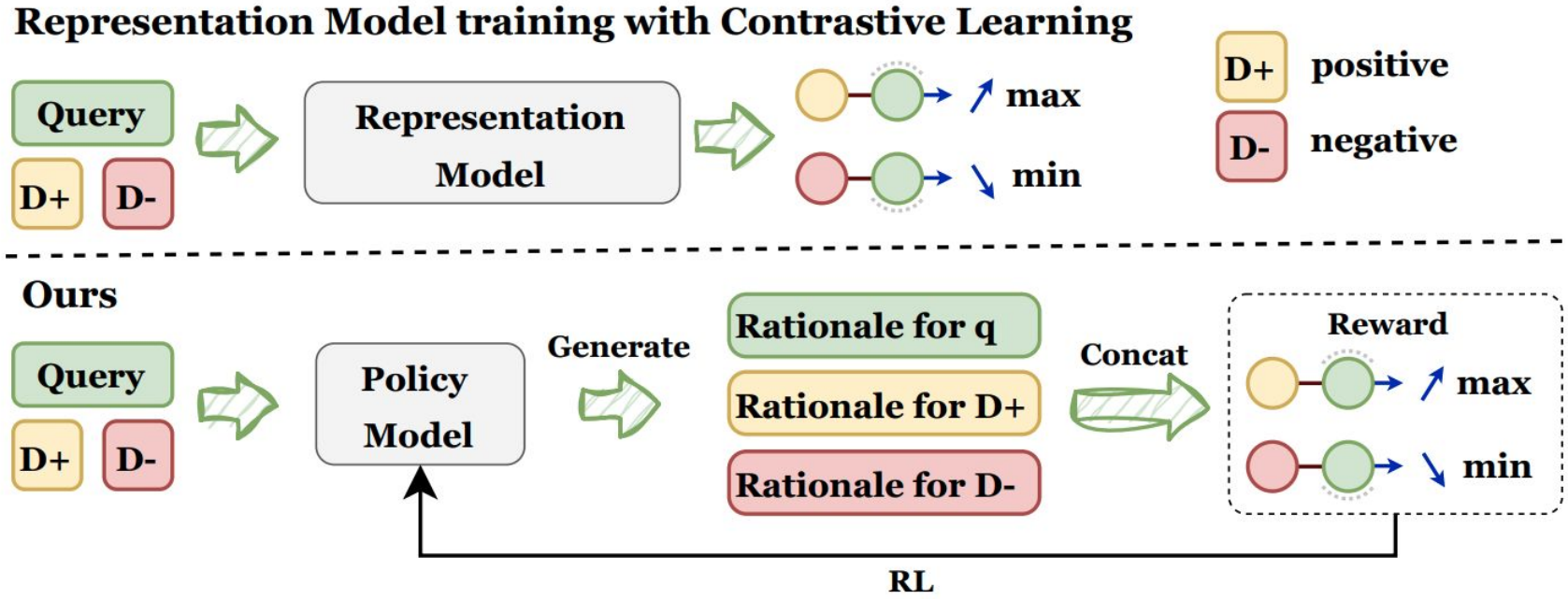
3. Lose the interpretability that makes LLMs valuable—the ability to understand and showing their reasoning process.

Motivation: Rethinking the contrastive signals in representation learning



1. We view them as reward signals that guide a generative policy.
2. LLM acts as a policy πθ that generates interpretable understandings of input texts.
   a. provide human-readable explanations of the model's semantic reasoning.
   b. encoded into high-quality representations of the inputs.

Method

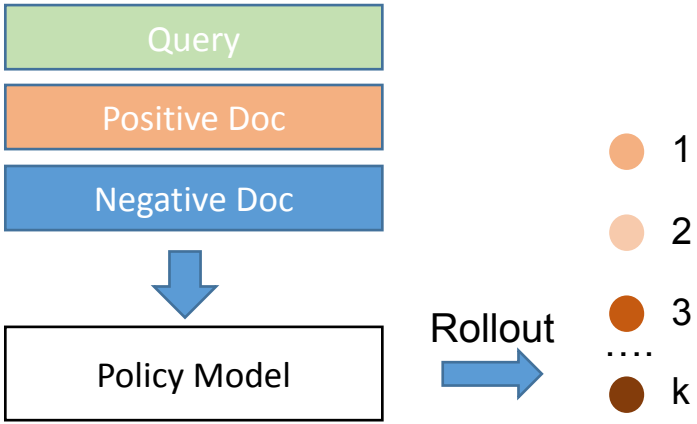


$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$

$$= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}$$

Our Base Reward: $\mathcal{R}_{\text{CL}}^{(i,k)} = \text{sim}(\mathbf{h}_{q_i}, \mathbf{h}_{d_i^+}^{(k)}) - \sum_{m=1}^{M_i} \text{sim}(\mathbf{h}_{q_i}, \mathbf{h}_{d_{i,m}^-})$.

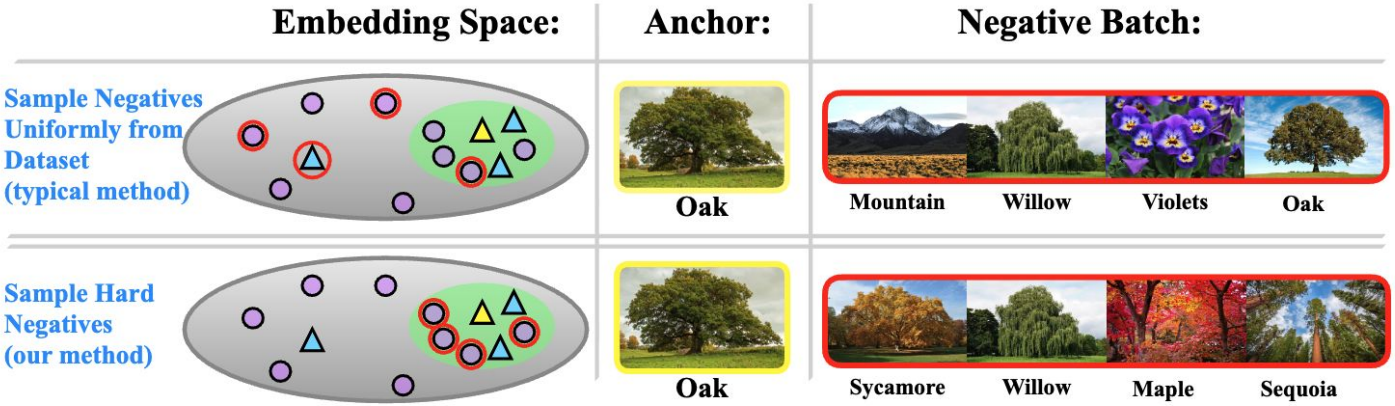
Method



Consistency Reward: $\mathcal{R}_{\text{consist}}^{(i,k)} = \frac{1}{K-1} \sum_{j \neq k}^{K} \text{sim}(\mathbf{h}_{d_i^+}^{(k)}, \mathbf{h}_{d_i^+}^{(j)})$

Encourage similar representations among concurrent rollouts to ensure semantic coherence across multiple interpretations of the same input.

Method

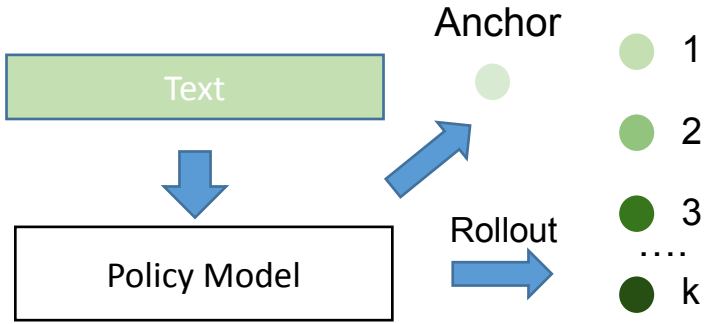


Hard-Negative Reward: $\mathcal{R}_{\text{hard}}^{(i)} = -\frac{1}{B-1}\sum_{\substack{j=1 \\ j\neq i}}^{B} \max_{1\leq l\leq K} \text{sim}\left(\mathbf{h}_{q_i}, \mathbf{h}_{d_j^+}^{(l)}\right).$

$$\mathcal{R}_{\text{total}}^{(i,k)} = \mathcal{R}_{\text{CL}}^{(i,k)} + \lambda_1 \mathcal{R}_{\text{consist}}^{(i,k)} + \lambda_2 \mathcal{R}_{\text{hard}}^{(i)}$$

Illustration from *Contrastive learning with hard negative samples* (ICLR 2021)

Unsupervised-Method



$$\mathcal{R}_{\text{total}}^{(i,k)} = \mathcal{R}_{\text{CL}}^{(i,k)} + \lambda_1 \mathcal{R}_{\text{consist}}^{(i,k)} + \lambda_2 \mathcal{R}_{\text{hard}}^{(i)}$$

$$R_{\text{self}}^{(i,k)} = \text{sim}\left(h_{x_i}^{\text{anchor}},\, h_{x_i}^{(k)}\right).$$

Inspired by *SimCSE: Simple Contrastive Learning of Sentence Embeddings* (EMNLP 2021)

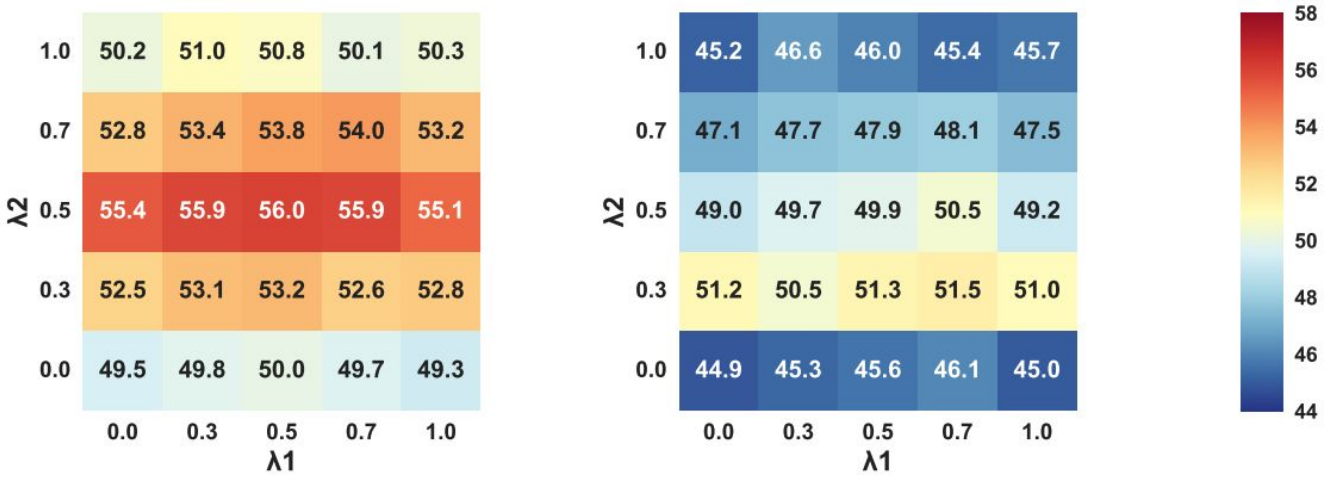
Experiment Massive Text Embedding Benchmark (MTEB)

Categories →	Retr.	Rerank.	Clust.	PairClass.	Class.	STS	Summ.	Avg.
# of datasets →	15	4	11	3	12	10	1	56
Qwen2.5-1.5B-Instruct								
Base	22.15	29.32	25.44	36.18	35.77	44.11	26.32	30.33
w/ reasoning	24.83	32.45	27.88	39.20	39.42	47.26	26.78	32.92
w/ CL training	38.95	43.88	36.21	52.02	53.87	56.39	28.43	43.21
GRACE	40.44	46.95	39.55	54.84	55.36	59.42	30.41	45.48
LLaMA-3.2-3B-Instruct								
Base	31.28	38.16	32.05	48.12	47.36	59.25	27.78	39.34
w/ reasoning	33.21	40.44	34.28	51.27	50.89	61.78	28.14	41.54
w/ CL training	42.42	47.35	39.92	58.66	58.15	65.55	28.63	47.39
GRACE	44.01	49.12	41.30	60.44	60.72	64.02	29.10	48.49
Qwen2.5-3B-Instruct								
Base	37.38	44.16	36.85	53.72	53.36	66.15	26.26	44.12
w/ reasoning	39.42	46.55	38.82	56.61	57.23	68.22	28.55	46.59
w/ CL training	45.90	52.87	43.26	74.08	65.94	70.05	29.68	52.10
GRACE	49.42	54.85	44.73	79.64	68.25	74.65	30.10	54.74
Qwen3-4B-Instruct-2507								
Base	37.42	48.16	38.55	55.33	54.87	66.02	29.44	45.49
w/ reasoning	38.91	49.72	40.76	57.20	55.41	68.35	29.62	46.87
w/ CL training	48.66	53.38	43.02	78.81	69.94	74.12	29.91	54.34
GRACE	52.11	55.85	45.24	82.94	71.02	77.38	30.46	56.64

Experiment Massive Text Embedding Benchmark (MTEB)

Categories →	Retr.	Rerank.	Clust.	PairClass.	Class.	STS	Summ.	Avg.
# of datasets →	15	4	11	3	12	10	1	56
Other Open Models								
BERT	10.59	43.44	30.12	56.33	61.66	54.36	29.82	38.33
RoBERTa	62.63	29.05	56.95	41.92	8.62	55.24	28.64	37.86
$\text{LLM2Vec}_{\text{LLaMA-3-8B}}$	24.75	49.20	39.74	65.91	69.00	67.85	25.59	48.84
$\text{Echo}_{\text{Mistral-7B}}$	71.63	33.51	72.31	47.43	22.85	73.64	31.02	49.02
`Qwen2.5-1.5B-Instruct`								
Base	22.15	29.32	25.44	36.18	35.77	44.11	26.32	30.33
w/ SimCSE	31.28	38.94	33.12	50.67	47.53	58.21	28.34	39.65
GRACE	34.57	41.86	35.49	51.12	50.78	56.44	29.07	41.45
`LLaMA-3.2-3B-Instruct`								
Base	31.28	38.16	32.05	48.12	47.36	59.25	27.78	39.34
w/ SimCSE	35.42	41.27	34.96	52.88	50.73	65.44	29.24	43.00
GRACE	36.55	43.15	36.88	55.27	53.62	63.18	29.52	44.04
`Qwen2.5-3B-Instruct`								
Base	37.38	44.16	36.85	53.72	53.36	66.15	26.26	44.12
w/ SimCSE	42.25	48.33	39.87	68.22	60.15	70.84	29.56	49.17
GRACE	43.15	49.72	41.58	70.44	62.91	69.38	29.63	50.15
`Qwen3-4B-Instruct-2507`								
Base	37.42	48.16	38.55	55.33	54.87	66.02	29.44	45.49
w/ SimCSE	42.18	50.72	41.63	69.14	61.25	72.48	29.62	50.11
GRACE	43.67	52.34	42.87	70.05	62.73	71.66	30.16	51.03

Reward Function Ablation Study



1. Removing all reward constraints (λ1 = 0, λ2 = 0) yields poor performance for supervised and unsupervised training.
2. The model exhibits significantly higher sensitivity to the hard negative mining weight (λ2) compared to the consistency weight (λ1), suggesting that hard negative discrimination plays a more critical role in determining overall performance.

Reward Algorithm Ablation Study

Algorithm # of datasets →	Retr. 3	Rerank. 2	Clust. 3	PairClass. 1	Class. 3	STS 3	Summ. 1	Avg. 16
GRACE-3B								
w/ ReMax Li et al. (2024)	42.63	61.24	35.61	68.3	62.3	70.8	29.04	53.36
w/ REINFORCE++ Hu et al. (2025)	43.10	64.9	37.67	68.0	63.18	71.60	29.91	54.64
w/ DAPO Yu et al. (2025)	44.58	65.62	39.24	70.67	63.58	72.8	30.04	55.78
w/ GRPO Shao et al. (2024)	44.72	65.71	38.99	70.87	64.35	72.53	30.09	55.89

Table 3: Comparison of our supervised method with different RL algorithms on subsets of MTEB.

1. Our method consistently improves performance across all four algorithms, demonstrating both its portability and generalizability.
2. GRPO is most effective in our setting, other algorithms focuses on issues that are less critical in our tasks.

Generalization to General Domain Tasks

Table 4: Performance on General Domain Tasks

Dataset → / Metric →	GSM8K EM	MMLU EM	TriviaQA EM	FEVER Acc	BBH EM	HumanEval Pass@1	Avg.	Δ
Qwen-2.5-1.5B-Instruct								
Base	32.06	54.94	18.35	66.91	25.25	46.95	40.74	–
Bsse w/ CL training	0.0	0.0	0.0	50.28	0.0	0.0	8.38	**-32.36**
GRACE (Supervised)	32.54	55.12	18.10	67.43	25.01	47.30	41.08	**+0.34**
GRACE (Unsupervised)	32.21	54.81	18.29	66.75	25.34	47.05	40.88	**+0.14**
LLaMA-3.2-3B-Instruct								
Base	16.75	14.74	29.21	64.52	9.72	38.41	28.89	–
Bsse w/ CL training	0.0	0.0	0.0	50.01	0.0	0.0	8.33	**-20.56**
GRACE (Supervised)	17.02	15.01	29.05	65.10	9.61	38.90	29.27	**+0.38**
GRACE (Unsupervised)	16.81	14.69	29.18	64.40	9.80	38.55	28.91	**+0.02**
Qwen-2.5-3B-Instruct								
Base	57.90	62.60	28.80	71.50	35.00	52.80	51.40	–
Bsse w/ CL training	0.0	0.0	0.0	50.13	0.0	0.0	8.35	**-43.05**
GRACE (Supervised)	59.10	61.20	27.60	72.80	34.30	53.10	51.50	**+0.10**
GRACE (Unsupervised)	58.00	61.50	28.30	71.20	34.80	52.90	51.10	**-0.30**
Qwen-3-4B-Instruct-2507								
Base	75.96	69.45	31.55	83.53	35.01	68.90	60.73	–
Bsse w/ CL training	0.0	0.0	0.0	51.07	0.0	0.0	8.51	**-52.22**
GRACE (Supervised)	76.42	69.71	31.40	84.02	34.88	69.35	61.13	**+0.40**
GRACE (Unsupervised)	76.05	69.38	31.52	83.40	35.09	69.01	60.74	**+0.01**

1. The CL fine-tuning baseline suffers severe performance drop, indicating that naive contrastive objectives can substantially erode general-domain capability.
2. Ours keep reasoning generation so the policy continues to practice skills needed for general tasks.

Case Study

Original Doc: BBC - History - British History in depth: Prime Ministers and Politics Timeline On This Day Prime Ministers and Politics Timeline Do you know which prime minister brought 'fallen women' to 10 Downing Street? Or which one fought a duel? Or who was known as 'the Goat'? Take a political journey through nearly 300 years of high ideals and low cunning, from Gordon Brown to the first man to hold prime minister...

Base Model (480 tokens)	**Training Step 200 (562 tokens)**	**Training Step 600 (824 tokens)**	**Training Step 1000 (1051 tokens)**
"The text is comprised of an interactive timeline on the history of **British Prime Ministers**, a concise biography of several past **Prime Ministers**, and a broader overview of the office." ... "It begins with timed anecdotes highlighting key figures and events from **1681** to the present day."	"Provides an overview of several key **British prime ministers** and the history of the office." ... "Starts with **Margaret Thatcher**, Britain's first **female PM**, and also mentions **Robert Walpole** and **Henry Campbell-Bannerman**." ... "Walpole is seen as the first PM, Campbell-Bannerman the first formally called so."	"Informative passage about key British Prime Ministers, such as **Margaret Thatcher**, **James Callaghan**, and **Harold Wilson**." ... "Traces the evolution of the role from **18th century cabinet system**, with Walpole as early figure." ... "Highlights crises like Harold Wilson's **industrial disputes** and **economic instability**." ... "Explains **parliamentary accountability** and **collective responsibility**."	"Provides a comprehensive understanding of the office of the **Prime Minister** and its historical context." ... "Roots traced back to **Robert Walpole**, later evolution shaped by constitutional changes." ... "Mentions modern figures like **Thatcher** and **Tony Blair**, situating them in **UK's political-economic**." ... "Discusses powers and constraints of **PM**, from cabinet appointments to parliamentary limits." ... "References anecdotes (e.g., '**fallen women**') as outdated, while contextualizing gender role shifts."

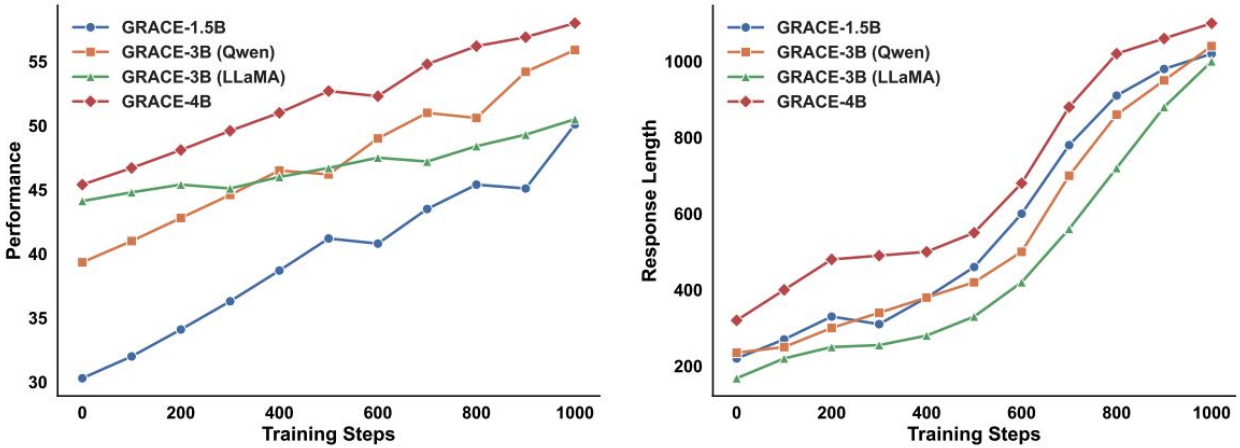
Table 6: Case study for output evolution. As training progresses, outputs grow from brief topic recognition to structured lists, then detailed coverage, and finally full contextualization.

Case Study

Doc Case (Belt and Road Initiative)	**Query Case (Machu Picchu)**
Original Document: "policy coordination is an important guarantee for implementing the Initiative... We should carry forward the spirit of friendly cooperation of the Silk Road by promoting cultural and academic ..."	**Original Query:** "In which country would you find the ruins of the ancient Inca city known as Machu Picchu?"
Model Response: "The Belt and Road Initiative is an extensive cooperation project proposed by China... *Policy Coordination* involves inter-governmental cooperation, expansion of shared interests, and enhancing mutual trust... *People-to-People Bonds* promote cultural exchange, volunteer services, and media cooperation... The initiative faces challenges such as debt sustainability and geopolitical risks..."	**Model Response:** "Machu Picchu is located in **Peru**... situated in the Andes Mountains, Cusco Region, at about 2,430m elevation... rediscovered in 1911 by Hiram Bingham... a UNESCO World Heritage Site and major tourist destination..."

Table 7: Representative cases of model responses: left column shows document understanding (doc case), right column shows query expansion (query case).

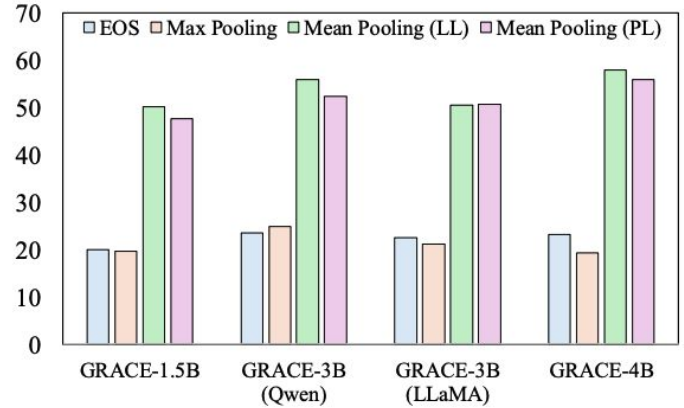
Ablation Study for Training Progression



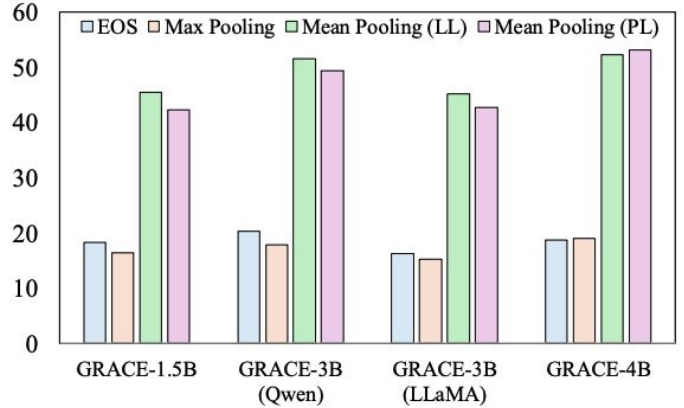
(a) Accuracy progression across training steps. (b) Response length progression across training steps.

Figure 5: Training progression of GRACE models. Left: accuracy on subtasks steadily improves with more training steps. Right: response length also increases, reflecting enhanced information density and richer reasoning chains.

Ablation Study for various representation approaches.



(a) Performance of various representation approaches in supervised fine-tuning.



(b) Performance of various representation approaches in unsupervised training.

Takeways

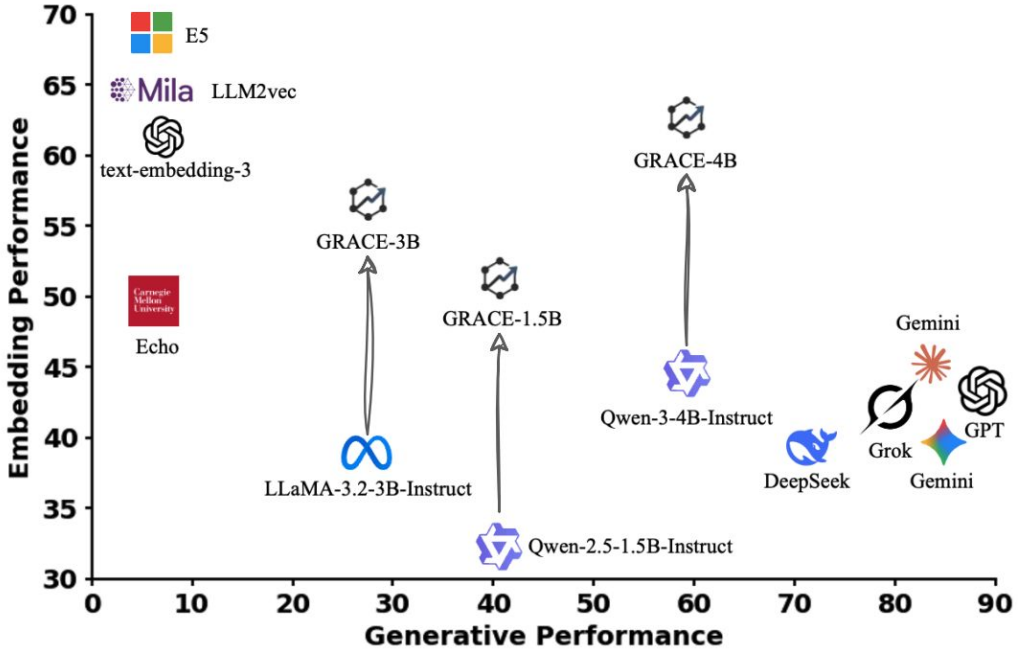
1. We present the first empirical evidence that rewards derived from contrastive learning can be leveraged to train policy models, resulting in improved representational capabilities.

2. We propose a novel methodology that enables the transformation of existing LLMs into powerful representation models while preserving their general capabilities without performance degradation.

3. This work represents a substantial advancement in text representation interpretability.

4. Our method yields overall score by 11.5% over base models, and the unsupervised adds 6.9% on the MTEB benchmark.

5. We make all models, datasets, and code publicly available to facilitate reproducibility and advance future research in this domain.



1. Retrieval -> Generate (Traditional RAG)

2. Retrieval -> thinking -> (retrieval) -> thinking ->generate (Agentic RAG)

3. Retrieval (thinking)

Thanks!