CCL: Causal-aware In-context Learning for Out-of-Distribution Generalization

Hoyoon Byun, Gyeongdeok Seo, Joonseong Kang, Taero Kim, Jihee Kim, Kyungwoo Song[*]
Department of Statistics and Data Science, Yonsei University
{hoyun.byun, gd.seo, doongsae, taero.kim, jihee_sta, kyungwoo.song}@yonsei.ac.kr

Abstract

In-context learning (ICL), a nonparametric learning method based on the knowledge of demonstration sets, has become a de facto standard for large language models (LLMs). The primary goal of ICL is to select valuable demonstration sets to enhance the performance of LLMs. Traditional ICL methods choose demonstration sets that share similar features with a given query. However, our experiments reveal that these traditional ICL approaches perform poorly on out-of-distribution (OOD) datasets, where the demonstration set and the query originate from different distributions. To ensure robust performance in OOD datasets, it is essential to learn causal representations that remain invariant between the source and target datasets. Inspired by causal representation learning, we propose causal-aware in-context learning (CCL). CCL captures the causal representations of a given dataset and selects demonstration sets that share similar causal features with the query. To achieve this, CCL employs a novel VAE-based causal representation learning technique. We demonstrate that CCL improves the OOD generalization performance of LLMs both theoretically and empirically. Code is available at: https://github.com/MLAI-Yonsei/causal-context-learning

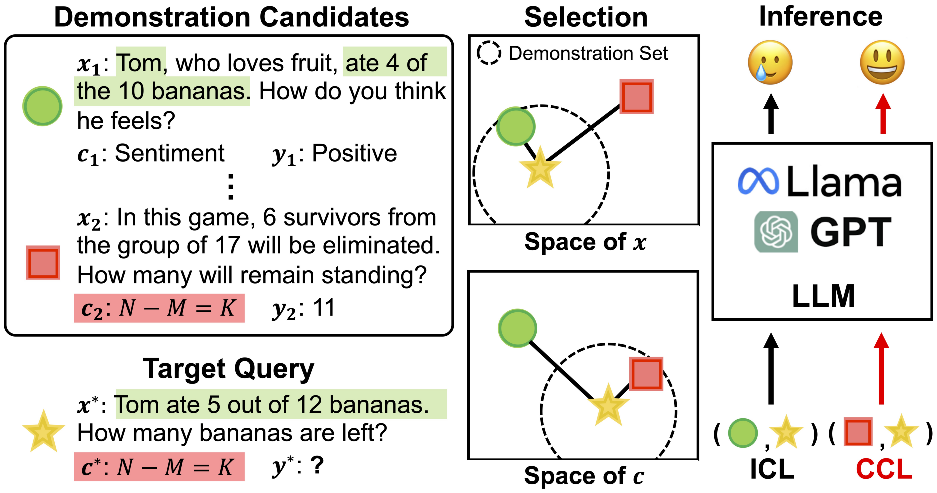
1 Introduction

While large language models (LLMs) excel as general-purpose pre-trained models, in-context learning (ICL) has become a key approach for aligning them to target tasks. ICL [1] enables LLMs to adapt to new tasks with a few demonstrations and without parameter updates, making it applicable in various fields. While ICL has shown significant promise, it still faces difficulties in achieving robust generalization [2]. A primary challenge is that LLMs rely on superficial patterns in demonstration sets, which restrict their capability in unseen environments [3]. Recent studies indicate that distribution shifts between demonstration sets and target queries in out-of-distribution (OOD) scenarios impede the ability of LLMs to generalize effectively [4, 5, 6]. To fully unlock the potential of LLMs and enable reliable deployment in real-world applications, ensuring robustness in OOD scenarios plays a pivotal role.

The pursuit of ensuring generalization beyond observed data naturally leads to the question of how the data was generated. Drawing on insights from causality [7], the structural knowledge of data is expressed using causal language. In causal representation learning (CRL) [8], observed data reflect underlying latent causal variables that drive the data-generating process (DGP). CRL aims to model the causal mechanisms among these variables [9]. For example, if two causal variables are independent, one remains invariant even when the other, acting as an environmental factor, changes [10, 11]. The assumption about causal mechanisms suggests that learning invariant causal variables is an effective approach for models robust to distribution shifts. Consequently, CRL lays the groundwork

[*] Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



x_1: Tom, who loves fruit, ate 4 of the 10 bananas. How do you think he feels?

c_1: Sentiment y_1: Positive

⋮

x_2: In this game, 6 survivors from the group of 17 will be eliminated. How many will remain standing?

c_2: $N-M=K$ y_2: 11

x^*: Tom ate 5 out of 12 bananas. How many bananas are left?

c^*: $N-M=K$ y^*: ?

Figure 1: To enhance the OOD performance of LLM, a causally-related demonstration set is important. Current ICL methods compare the non-causal representation x_i and x^*, and they might choose a worthless demonstration set (Candidate 1). However, our method, CCL, compares the causal representation c_i and c^* to construct a demonstration. Because CCL leverages the causal-related demonstration (Candidate 2), CCL shows superior performance on the OOD dataset.

for research on OOD generalization by learning invariant representations from training sets collected across multiple environments. [12, 13, 14].

Considering causal mechanisms allows for constructing a more suitable demonstration set from demonstration candidates when their environments differ from that of the target query. In Figure 1, the target contextual problem x^* is highly similar to x_1 (highlighted in green), making x_1 a strong candidate in ICL [15, 16]. But what exactly is the *problem* embedded in the target query? For LLMs to successfully generalize to the target query, the demonstration set should be constructed to reflect the fundamental context rather than relying on superficial patterns, such as frequently occurring words or characteristics of the data collection environment [17].

Therefore, even if the superficial context differs, a candidate x_2 that addresses the same problem (N-M=K) should be included in the demonstration set (highlighted in red). It ensures the demonstration set captures problem-level invariance even when generalizing to OOD targets from given candidates. Since causal variables c, which generate the contextual problem, are not observable objects, it is necessary to model c under the assumption of causal mechanisms that remain invariant across environments.

In this study, we focus on constructing a robust demonstration set to enhance the generalization of LLMs in OOD scenarios. Inspired by CRL, we propose a novel demonstration selection method, causal-aware in-context learning (CCL), which learns causal representations that remain invariant across environments and prioritizes candidates by assigning higher ranks to those with causal representations similar to the target query. Under the causal mechanism, we theoretically demonstrate that the demonstration set selected by CCL comprises candidates that are more closely related to the underlying problem addressed by the target query, rather than merely matching its context. The problem-level invariance of CCL ensures generalization performance for the target query even in unseen environments. We empirically validate that CCL operates robustly in OOD scenarios and demonstrates superior generalization performance on both synthetic and real datasets.

2 Related Works

2.1 In-context learning

ICL is a method where LLMs perform tasks by leveraging examples from the input context without updating model parameters [1]. This approach enhances computational efficiency and achieves competitive performance in various natural language tasks without the need for model fine-tuning

[2, 18]. However, the performance of ICL is sensitive to demonstration organization, including demonstration selection [19, 20]. Various approaches aim to optimize demonstration selection in ICL, including unsupervised methods that use similarity metrics like k-nearest neighbors [15], as well as supervised techniques that leverage task-specific retrievers [21] and reinforcement learning [22].

Despite these advancements, LLMs depend on surface-level patterns in the demonstration set, leading to a primary challenge with out-of-distribution (OOD) examples [3]. While larger models tend to reduce the performance gap between in-distribution (ID) and OOD scenarios, even transformers, which handle minor distribution shifts, face significant challenges when encountering major shifts [6, 4]. The BOSS benchmark evaluates OOD robustness in ICL, highlighting the importance of addressing OOD generalization [5]. An approach designed to improve OOD performance involves inferring latent variables from the context using the transformer architecture. However, this method struggles to apply those variables effectively in prediction, limiting OOD generalization [23]. We propose CCL, drawing on causal representation learning, to improve OOD performance in ICL by focusing on task-relevant causal features and enhancing robustness to distribution shifts.

2.2 Causal representation learning

Unlike statistical approaches, which describe the distributional characteristics of data, causality [7] focuses on the structural relationships between variables. The DGP is determined by the underlying causal relationships among variables, and a structural causal model (SCM) is a generative model that describes the DGP [10, 24]. The SCM expresses the uncertainty of exogenous factors in a probabilistic manner and defines functional relationships for the variables of interest (endogenous variables), thus structurally describing the causal mechanisms of the DGP. Observed data represent one of the realizations of these causal mechanisms. A causal graph visually represents the structural relationships between the variables, as induced by the SCM [7].

Recently, research in machine learning has increasingly focused on moving beyond models limited to statistical associations [25], aiming to model the underlying structural properties of the data by applying the causality framework to machine learning [26, 27]. CRL aims to construct latent variables that capture the underlying causal mechanisms, allowing for the discovery of causal representations within observed data [8]. It seeks to deploy robust models in OOD scenarios, ensuring reliable performance even when the data distribution shifts. For example, leveraging the stability of causal mechanisms across different environments, several studies have utilized the invariant properties of causal representations under distribution shifts to enhance model performance in OOD scenarios through invariant prediction [12, 28].

Furthermore, there has been ongoing research into utilizing deep generative models to explicitly represent causal variables. Notably, under the assumption of independent causal mechanisms [10], several studies have modeled these mechanisms as separate, independent modules or have focused on learning disentangled and interpretable representations [11, 29, 30]. Research has evolved toward learning causal representations that maintain stable mechanisms under distribution shifts, to improve OOD generalization [13]. Inspired by CRL, we construct a novel ICL framework using causal knowledge for OOD generalization. To build a robust demonstration set, we utilize the invariant causal representation constructed by a Variational Autoencoder (VAE) [31]–based model [13, 32].

3 Methodology

3.1 Generative model and inference model

In CCL, we consider several key variables: the task variable t represents the specific task being performed. The latent causal variable c represents the fundamental context of the query. It is generated from the task variable t and serves as a causal factor for both the input query x and the (ground truth) answer y. Additionally, we introduce the latent source variable s, which influences components of x that are unrelated to the task, such as the structure of the text. The environmental variable e acts as an observable proxy for the latent source variable s. It represents contextual attributes of the data, such as the dataset's origin or the language used.

Note that both latent variables, c and s, generate x, where c represents task-specific information, and s represents domain-specific information. That is, we assume that the domain shift in the observed data is induced by changes in s, while c remains invariant, as shown in Figure 2.

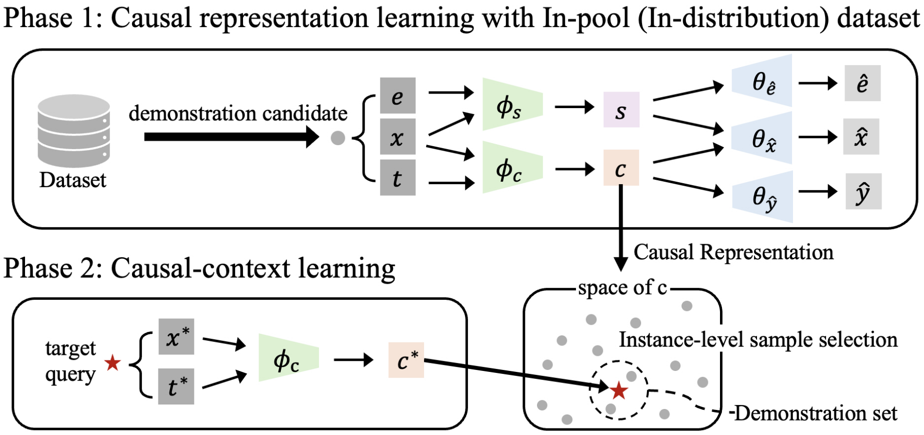
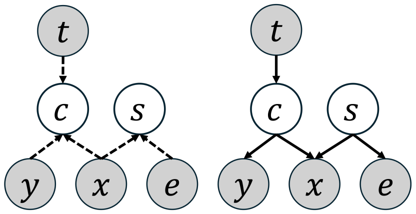


Figure 3: Our proposed method, Causal-aware In-Context Learning (CCL), utilizes causally related demonstration sets to enhance performance on out-of-distribution (OOD) datasets. (Phase 1) First, we optimize a novel VAE-based causal representation learning method to capture the causal representations of a given in-distribution dataset. After optimization, we store the causal representations, c, produced by the optimized model for the in-distribution dataset. (Phase 2) Second, CCL captures the causal representation, c^*, of the target query and selects the appropriate demonstration sets by comparing c and c^*.

We aim to model the joint distribution of observed variables $\{x, y, t, e\}$ along with latent variables $\{c, s\}$. We assume the generative model

$$p_\theta(x, y, t, e, c, s) \;=\; p_\theta(x, t, e, c, s)\; p_\theta(y \mid c),$$

where $p_\theta(y \mid c)$ is an invariant causal mechanism. We let θ denote all parameters of the generative model. We denote the unknown true source-domain distribution as $p_{\theta^*}(x, y, t, e)$, and we approximate it with $p_\theta(x, y, t, e)$.



(a) Inference model (b) Generative model

Figure 2: Graphical model of CCL. The generative model shows that t influences the latent causal variable c, which in turn directly affects both x and y.

Figure 3 illustrates the overall workflow of CCL in two phases. In Phase 1, we learn causal representations from an in-distribution (ID) dataset using our VAE-based model: the inference networks ϕ_s and ϕ_c infer the latent variables s (environment-related) and c (task-related), respectively, while the decoders $\theta_{\hat{e}}, \theta_{\hat{x}}, \theta_{\hat{y}}$ reconstruct the observed variables. This process yields the causal embeddings c for the ID data. In Phase 2, given a target query (x^*, t^*), we apply ϕ_c to obtain its causal embedding c^*. Comparing c^* with the stored causal embeddings c, CCL then selects the most relevant demonstration examples, those with similar causal factors, to construct the prompt context. This causal representation approach ensures that our examples align with the true causal structure of the query, thereby improving model performance even under distribution shifts.

3.2 Learning causal representations via variational inference

Since direct maximization of $\log p_\theta(x, y, t, e)$ is often intractable due to the latent variables, we employ variational inference. We introduce a tractable inference model $q_\phi(c, s \mid x, y, t, e)$, where ϕ are the variational parameters. The standard Evidence Lower BOund (ELBO) on $\log p_\theta(x, y, t, e)$ is:

$$\begin{aligned}\log p_\theta(x, y, t, e) &= \log \int p_\theta(x, y, t, e, c, s)\, dc\, ds = \log \mathbb{E}_{q_\phi(c,s|x,y,t,e)}\Big[\frac{p_\theta(x, y, t, e, c, s)}{q_\phi(c, s \mid x, y, t, e)}\Big] \\ &\geq \mathbb{E}_{q_\phi(c,s|x,y,t,e)}\left[\log \frac{p_\theta(x, y, t, e, c, s)}{q_\phi(c, s \mid x, y, t, e)}\right] := L_{\text{ELBO}}\end{aligned}$$

Maximizing this ELBO with respect to both θ and ϕ yields a tight approximation when $q_\phi(c, s \mid x, y, t, e) \approx p_\theta(c, s \mid x, y, t, e)$.

Since θ^* is unknown, we instead optimize the ELBO using the observed data distribution in the source domain, $p_D(x, y, t, e)$:

$$\max_{\theta,\phi} \mathbb{E}_{(x,y,t,e)\sim p_D(x,y,t,e)} \left[L_{\text{ELBO}}\right] \tag{1}$$

3.2.1 Reformulating variational inference for unobserved y

At test time, y is always unobserved, as it is the target variable we aim to infer. While one common approach, such as in CEVAE [33], is to introduce an auxiliary model to explicitly predict y, we instead modify the objective function to enable variational inference without conditioning on y. Specifically, we factorize the inference model:

$$q_\phi(c, s, y \mid x, t, e) = q_\phi(c, s \mid x, t, e)\, p_\theta(y \mid c),$$

which reflects the conditional independence $y \perp (x, t, e, s) \mid c$. This design is key, as it directly injects the generative model's causal assumption ($c \to y$) into the inference process. It serves to constrain the inference model q_ϕ to find a c that is consistent with $p_\theta(y \mid c)$, the actual causal mechanism from the generator. This formulation allows us to marginalize out y. By applying this factorization and Bayes' rule to the standard ELBO, we analytically marginalize out the unobserved y, reformulating the objective to depend only on $q_\phi(c, s \mid x, t, e)$ (see Appendix A for the full derivation). We define $\Phi_{y|x,t,e} = \mathbb{E}_{q_\phi(c,s|x,t,e)}[p_\theta(y|c)]$ as the implicit predictive distribution of y. The final objective of CCL is given by:

$$\begin{aligned}\max_{\theta,\phi} \mathbb{E}_{p_D(x,y,t,e)}[L_{\text{ELBO}}] = \mathbb{E}_{p_D(x,y,t,e)}\Big[\log \Phi_{y|x,t,e} \\ + \frac{1}{\Phi_{y|x,t,e}} \mathbb{E}_{q_\phi(c,s|x,t,e)}[p_\theta(y|c) \times \log \frac{p_\theta(x,t,e,c,s)}{q_\phi(c,s|x,t,e)}]\Big].\end{aligned} \tag{2}$$

We construct the reconstruction model p_θ following the generative structure outlined in Figure 2b. Implementing Equation 2 requires this model, which is composed of decoders (e.g., $p_\theta(x \mid c, s)$, $p_\theta(y \mid c)$, $p_\theta(e \mid s)$) that reconstruct the observed variables from the latent variables. This reconstruction process, particularly the $p_\theta(y \mid c)$ mechanism, ensures that the learned causal representation c effectively captures task-relevant information.

3.3 Regularization and conditional prior

In practice, to prevent unintended dependencies between c and s during training, we further employ Maximum Mean Discrepancy (MMD) [34] loss as a regularization term [9]. Additionally, the task variable t (the parent of c) is treated as an observed input, not a latent variable requiring posterior inference. Instead, following the iVAE [35] framework, we define a conditional prior $p_\theta(c \mid t)$ for the generative model based on this observed t. Our variational inference formulation follows the approach proposed in [32].

3.4 Theoretical analysis

Prioritizing demonstrations that are causally similar to the query yields provably better in-context learning (ICL) than prioritizing demonstrations that are merely input similar. We show that input nearest selection can induce large label discrepancies even when inputs are arbitrarily close in Theorem 3.3. Furthermore, Theorem 3.4 provides both a theoretical explanation and a practical guideline: prioritizing causally similar examples is key to robust ICL.

Our analysis begins by assuming the data-generating process is modeled using an SCM $\mathcal{M} := (\mathcal{S}, P_\varepsilon)$ and a collection $\mathcal{S}$ of assignment equations as follows [10]:

$$t := \varepsilon_t, \quad c := f_c(t, \varepsilon_c), \quad s := \varepsilon_s, \quad e := f_e(s, \varepsilon_e), \quad x := f_x(c, s, \varepsilon_x), \quad y := f_y(c, \varepsilon_y). \tag{3}$$

Here, $\varepsilon_t, \varepsilon_c, \varepsilon_s, \varepsilon_e, \varepsilon_x \in \mathbb{R}^d$ are random vectors with $d \geq 2$ and $\varepsilon_y \in \mathbb{R}$ is a random variable. We assume $\varepsilon = \{\varepsilon_t, \varepsilon_c, \varepsilon_s, \varepsilon_e, \varepsilon_x, \varepsilon_y\}$ satisfies joint independence. The parents of x are c and s, while y has only c as its parent. The causal graph is achieved by drawing edges from RHS variables of Equation (3) to LHS variables except the noise variables ε.

We adopt a linear setting in line with [36], who demonstrate that attention-based updates in LLMs can be approximated by steps of gradient descent with a convex loss on a linear parameter w with respect to $w^\top x$. Although real-world LLMs are more complex, the linear approximation provides a clear analytical framework.

Assumption 3.1 (Linear-causal assumption)**.** We formalize a simplified data-generating process via the following linear-causal assumption:

$$x_i := \mathcal{B}_1\, c_i + \mathcal{B}_2\, s_i + \varepsilon_{x,i}, \quad y_i := (w^*)^\top c_i + \varepsilon_{y,i}.$$

Each coordinate of $\varepsilon_{x,i}$ is σ_x^2-sub-Gaussian, and $\varepsilon_{y,i}$ is σ_y^2-sub-Gaussian. $\mathcal{B}_1$ and $\mathcal{B}_2$ denote coefficient matrices to c_i and s_i. x_i, c_i, and s_i are d-dimensional vectors and y_i is a scalar.

A prerequisite for ICL is to construct a demonstration set $\mathcal{D}_S = \{(x_i, y_i)\}_{i\in S}$ from the training dataset $\mathcal{D}_T = \{(x_i, y_i)\}_{i\in\mathcal{I}}$, where $S \subset \mathcal{I}$ is the selected index set. A common strategy, forming the set $\mathcal{D}_x$, selects pairs (x_i, y_i) by assessing how similar x_i is to the input query x^*, with the expectation that y^* will be similar to y_i [15]. Our method leverages latent causal variables: we associate the query x^* with a causal variable c^*, and each training pair (x_i, y_i) with its own causal variable c_i. We then select pairs whose c_i lie close to c^*, forming a causally similar set $\mathcal{D}_c$. All our theorems are based on Assumption 3.1.

Definition 3.2 (Demonstration sets by strategy)**.** Let $sim(\cdot,\cdot)$ be a similarity measure and N the size of the demonstration set. We define:

$$\mathcal{D}_c = \underset{S\subset\mathcal{I},|S|=N}{\operatorname{argmax}} \sum_{i\in S} sim(c_i, c^*), \quad \mathcal{D}_x = \underset{S\subset\mathcal{I},|S|=N}{\operatorname{argmax}} \sum_{i\in S} sim(x_i, x^*) \tag{4}$$

Theorem 3.3 (Input proximity can lead to prediction discrepancy)**.** *Let (x^*, y^*) and (x, y) be two samples potentially generated by different latent pairs (c^*, s^*) and (c, s). Under Assumption B.1, B.2 in Appendix B, for every $\epsilon > 0$, there exists a $\kappa > 0$ such that*

$$\|c^* - c\| \; > \; \frac{\kappa}{\|\mathcal{B}_1\|_{\text{op}}} \quad \Longrightarrow \quad \|y^* - y\| > \kappa\frac{\gamma}{\|\mathcal{B}_1\|_{\text{op}}} \quad \textit{where } \|\cdot\|_{\text{op}} \textit{ is the operator norm}$$

for some constant γ, if $\|x^* - x\| < \epsilon$. In other words, one can make $\|x^* - x\|$ arbitrarily small while allowing $\|y^* - y\|$ to remain arbitrarily large, due to the interplay between (c^*, s^*) and (c, s).

Theorem 3.3 shows that even when the distance between x^* and x is made arbitrarily small, the distance between the corresponding y^* and y can still be significant, as there is no upper bound on this gap. Consequently, the predicted value based on x^* may coincide with y, causing a discrepancy with the true y^*.

Picking demonstrations from $\mathcal{D}_c$ yields better in-context learning than picking from $\mathcal{D}_x$. The upper bound of the estimation error of the learned parameter is smaller compared to that of input-based selection. Furthermore, the upper bound on the test prediction error with CCL is also smaller. The parameter update in ICL, under a transformer architecture, is approximated by gradient descent on the demonstration set, following the formulation in [36]. Let $w_c^{(M)}$ be the weight updated via M steps of gradient descent using the empirical risk on $\mathcal{D}_c$, and let $w_x^{(M)}$ be the corresponding weight updated from $\mathcal{D}_x$.

Theorem 3.4 (Performance of the c-similarity)**.** *For sufficiently large N, M, with probability at least $1 - \delta_{tail}$, the following holds under Assumption C.1–C.4 in Appendix C:*

1. ***Tighter upper bound on estimation error.*** *The estimation errors admit upper bounds U^c_{param} and U^x_{param} such that*

   $$\|w_c^{(M)} - w^*\| \le U^c_{param}, \quad \|w_x^{(M)} - w^*\| \le U^x_{param}, \quad \textit{and} \quad U^c_{param} < U^x_{param}.$$

   $U^c_{param} = (1/\lambda_{min}(\Gamma_c)) \cdot C_u S_N$ and $U^x_{param} = (1/\lambda_{min}(\Gamma_x)) \cdot C_u S_N$. C_u is a some constant and $S_N = \sqrt{\log(1/\delta_{tail})/N}$. $\lambda_{min}(A)$ denotes the minimum eigenvalue of a matrix A. Γ_c and Γ_x are the empirical second moment matrices of $\mathcal{D}_c$ and $\mathcal{D}_x$.

2. ***Tighter upper bound on test error.*** *For the test query (x^*, y^*) with $y^* = (w^*)^\top c^* + \varepsilon^*_y$, the prediction errors admit upper bounds U^c_{test} and U^x_{test} such that*

   $$\left|(w_c^{(M)})^\top x^* - y^*\right| \le U^c_{test}, \quad \left|(w_x^{(M)})^\top x^* - y^*\right| \le U^x_{test}, \quad \textit{and} \quad U^c_{test} < U^x_{test}.$$

   $U^c_{test} = \|x^*\| U^c_{param} + |\mathcal{R}|$ and $U^x_{test} = \|x^*\| U^x_{param} + |\mathcal{R}|$. $\mathcal{R} = (w^*)^\top x^* - y^*$.

Theorem 3.4 shows that, with high probability, the parameter error $\|w_c^{(M)} - w^*\|$ and the test error $\|(w_c^{(M)})^\top x^* - y^*\|$ admit upper bounds that are tighter than the corresponding bounds obtained from $\mathcal{D}_x$. In essence, when $\mathcal{D}_c$ is used for demonstrations, the underlying design matrix becomes better conditioned with respect to c, mitigating the confounding effect of s and leading to tighter error bounds.

4 Experiments

We validate the effectiveness and validity of CCL by addressing three main points. First, in Section 4.2, we verify that the latent variables c and s inferred by CCL indeed capture domain-invariant and domain-variant features, respectively, for modeling the causal factors of x. In Section 4.3, we examine whether the samples characterized by c exhibit similarity to the test samples or convey the same underlying intent. In Section 4.4, we evaluate how the demonstration sets constructed using CCL enhance in-context learning performance under OOD scenarios. In Section 4.5, we qualitatively analyze how the latent features c and s capture distinct features. Lastly, in Section 4.6, we investigate the capability of CCL on new or more intricate reasoning tasks and perform a sensitivity analysis.

4.1 Experimental setup

We adopt a query-dependent demonstration strategy that dynamically selects the suitable examples for each test input. After embedding a test query, we compute its cosine distances to all candidates in the in-distribution training pool. In the K-nearest-neighbor (KNN) variant, the K closest instances, where K equals the predefined shot size (Ω), are selected directly. We also investigate a K-means-based selection method that is governed by two hyperparameters, R and P. A proportion R of the shot budget is allocated to the most similar instances, obtained exactly as in the KNN procedure. The remaining budget $K = \Omega - R$ is filled by clustering: among the next P (with $P \in \{50, 100, 300\}$) most similar candidates, we run K-means clustering and, from each cluster, select the sample whose embedding is closest to the centroid. This combined strategy yields prompts that simultaneously maintain high relevance to the query while covering a broader range of semantic regions.

4.2 Synthetic data

Method	ID Task Comparison			Env. Comparison			OOD Task Comparison		
	Acc.	NDCG	F1	Acc	NDCG	F1	Acc	NDCG	F1
x	57.7	71.5	58.6	85.7	91.3	86.0	45.0	57.9	45.4
CVAE (z)	33.3	60.2	32.5	33.1	43.2	32.2	32.4	53.6	33.9
Oracle (c)	**100.0**	**100.0**	**100.0**	33.5	48.7	33.7	**100.0**	**100.0**	**100.0**
CCL (c)	**100.0**	**100.0**	**100.0**	40.8	55.0	39.9	**100.0**	**100.0**	**100.0**
Oracle (s)	33.3	48.9	33.9	**100.0**	**100.0**	**100.0**	32.7	51.3	33.0
CCL (s)	36.2	51.4	36.2	**100.0**	**100.0**	**100.0**	33.2	48.3	31.9

Table 1: Retrieval experiments on synthetic data show that CCL consistently outperforms alternatives on both in-distribution and out-of-distribution task queries, confirming that c captures the underlying causal structure of the tasks. Conversely, when retrieval is conditioned on environment labels, s-based retrieval excels, highlighting their sensitivity to domain-specific factors. CCL's learned representation, CCL (c), tracks the ground-truth causal feature particularly closely.

We construct synthetic data with three tasks and five environments. Following Figure 2b, we first define the root nodes: the task variable t and the s variable. We enforce independence among task embeddings t by randomly initializing them with orthogonality constraints, applying the same approach to s. Then, we generate the c embedding using a three-layer fully connected neural network that takes t as input and add random noise to its output. Other variables follow a similar process. We train the neural networks, viewed as non-linear data-generating functions, using contrastive learning to ensure that c is similar within the same task and e is similar within the same s, while enforcing dissimilarity across different tasks or environments.

To better reflect realistic scenarios, we consider similar tasks or environments. Specifically, for the root nodes t and s, we set the cosine similarity between any two t or s embeddings to a value between 0 and 1 (in our experiment, we use 0.7). During contrastive training of the generating functions, we adjust the loss weights to reduce the penalty for similar tasks or environments, ensuring their embeddings are not pushed too far apart.

Table 1 presents the proportion of retrieved samples whose task or environment (Env.) matches that of the target input, across different embedding types, under both in-distribution (ID) and out-of-distribution (OOD) settings. Additional experimental results and discussions on the synthetic experiments are provided in Appendix D.

4.3 MGSM

Metric	x **embedding**	c **embedding**
Total Accuracy	81.03	**85.84**
ID Accuracy	97.05	**99.74**
OOD Accuracy	53.00	**61.52**
Total NDCG	86.00	**88.73**
ID NDCG	99.12	**99.89**
OOD NDCG	63.03	**69.21**

(a) Comparison of retrieval accuracy and NDCG for x and c embeddings on MGSM in the 5-shot setting.

Method	**Total**	**ID**	**OOD**
ZS	87.71	89.43	84.70
ICL (Fix.)	91.20	91.26	91.10
ICL (KNN)	94.07	95.83	91.00
CCL	**94.55**	**96.11**	**91.80**

(b) Comparison of performance. ZS denotes the zero-shot baseline, ICL (Fix.) uses a fixed demonstration set. ICL (KNN) and CCL utilize KNN retrieval

Table 2: (a) compares five-shot MGSM retrieval performance between embeddings derived from the original inputs x and from the causal features, c. (b) reports overall, in-distribution (ID), and out-of-distribution (OOD) accuracies for four prompting regimes—zero-shot (ZS), fixed demonstrations, KNN-based retrieval, and CCL.

As another dataset to evaluate the performance of our methodology, we employ the MGSM (Multilingual Grade School Math) dataset [37]. The MGSM dataset is a human-annotated translation of 250 problems from the GSM8K dataset [38] into ten different languages.

Utilizing the MGSM dataset, our goal is to evaluate the precision with which CCL deduces latent variables c, that represent the fundamental context of problems. For this purpose, we evaluate the retrieval performance by examining how correctly the model retrieves the same problem given a specific question.

First, we extract embeddings for each question using OpenAI's text-embedding-3-small model. Based on these embeddings, we split the data into an ID and an OOD dataset. We use Swahili, Thai, Telugu, and Bengali for the OOD dataset, while the remaining languages are designated as ID. We provide a detailed explanation of the classification criteria in Appendix D.

In this experiment, we define the problem category as the task t. The categories include six classes, such as "Arithmetic Operations" and "Geometry and Measurements". These categories are generated by labeling each question using OpenAI's o1, followed by human verification. During the labeling process, only English questions are labeled, and the same labels are directly applied to corresponding questions in other languages.

Table 2a presents the retrieval performance of the x embeddings and the c embeddings. We evaluate how accurately each method retrieves the same problem in a different language. The results demonstrate a significant improvement in accuracy and NDCG for both ID and OOD when using our approach instead of x embeddings.

Next, we perform ICL based on the retrieval results. In the MGSM dataset, we evaluate performance by measuring the model's prediction accuracy. Similarly to the retrieval process, we use a 5-shot setting to assess performance and compare zero-shot (ZS), ICL (Fixed sample, KNN) and CCL. Unlike ICL (KNN) and CCL, which can retrieve samples from different languages, ICL (Fix.) uses predefined samples specific to each language. We use GPT-4o-mini for in-context learning. We refer to Appendix D for a detailed explanation of the MGSM experiment.

Table 2b illustrates the experimental results. The results demonstrate that CCL-based retrieval for in-context samples achieves higher accuracy in both ID and OOD settings than other approaches. This aligns with the strong retrieval performance of c embedding indicated in Table 2a, demonstrating that selecting in-context samples based on the latent causal feature c is crucial for problem solving and improves in-context learning accuracy.

4.4 Generalization across tasks and domains

Language model	Retrieval method	QNLI	PIQA	WSC273	YELP	Avg.
Llama-3.2-3B-IT	ZS	43.36	**71.33**	55.31	*88.98*	64.75
	LLM-R	29.93	69.91	61.17	79.48	60.12
	ICL (K-means)	68.13	69.04	49.82	75.81	*65.70*
	CCL	**75.18**	*70.46*	**61.91**	**95.44**	**75.74**
Phi-4-mini-IT	ZS	**86.34**	**76.01**	64.10	95.76	80.55
	LLM-R	*85.21*	74.10	65.93	**96.37**	80.40
	ICL (K-means)	83.18	74.81	*71.06*	96.25	*81.33*
	CCL	82.26	*75.73*	**71.43**	*96.33*	**81.44**
GPT-4o	ZS	**91.30**	*94.07*	90.84	97.47	93.42
	LLM-R	90.32	**94.23**	*92.67*	*98.27*	93.87
	ICL (K-means)	88.28	93.04	87.55	98.17	91.76
	CCL	*90.77*	93.15	**93.77**	**98.36**	**94.01**

Table 3: Out-of-distribution accuracy on QNLI, PIQA, WSC273, and Yelp for three language models—Llama-3.2-3B-IT, Phi-4-mini-IT, and GPT-4o—under four prompting regimes: zero-shot (ZS), the learned-retriever baseline (LLM-R), and two K-means-based retrieval approaches, vanilla ICL and CCL. Bold numbers denote the highest score in each column, and italics denote the second highest. CCL attains the best average accuracy for every model, with particularly pronounced improvements for the smaller Llama-3.2-3B-IT.

We evaluate whether examples selected by CCL improve performance on OOD NLP tasks. Adopting the experimental protocol of LLM-R [39], we compare against their retrieval method but instead assess the generated outputs rather than relying on token probabilities. Our approach retrieves examples with similar c embeddings via KNN, clusters them using K-means, and selects the cluster centers as final candidates. As shown in Table 3, CCL consistently yields strong performance across diverse OOD tasks. We follow the same 8-shot setting used in LLM-R to ensure a fair comparison.

4.4.1 Sensitivity to the embedding models

Language model	Embedding model	QNLI	PIQA	WSC273	YELP	Avg.
Phi-4-mini-IT	text-embedding-3-small	**82.26**	**75.73**	*71.43*	**96.33**	*81.44*
	multilingual-e5-large-instruct	**82.26**	*75.25*	**73.99**	*95.72*	**81.81**

Table 4: CCL accuracy on four out-of-distribution benchmarks when the same language model (Phi-4-mini-IT) is paired with two embedding models (OpenAI's text-embedding-3-small and the multilingual-e5-large-instruct). Scores are given for each task and averaged; bold indicates the highest score per column, and italics the second-highest. The multilingual-e5 encoder attains the top overall score, yet the gap is small, indicating that CCL remains robust to the choice of embedding model.

To evaluate CCL's sensitivity to the encoder, we reran the entire pipeline across the NLP benchmarks using multilingual-e5-large-instruct [40], an open-source embedding model that ranks among the top performers on the MTEB text-embedding leaderboard [41]. Table 4 experimentally demonstrates that CCL maintains comparable performance despite changes in the embedding model, highlighting its robustness in inferring causal features.

4.5 Qualitative analysis

We provide a qualitative analysis of the learned latent features to better understand how c and s are interpreted in practice. To visualize the semantics encoded in these variables, we decode sentence embeddings while zeroing out one latent dimension. Specifically, we first infer c and s from an input embedding x. We then set $s = 0$ to generate $x'_{s=0}$, which highlights the domain-invariant features represented by c. Similarly, we set $c = 0$ to generate $x'_{c=0}$, which reveals the domain-variant information captured by s. Table 5 lists the top-5 nearest words to each decoded embedding.

x	$x'_{s=0}$	$x'_{c=0}$
horribleappetizers	unappetizing	review
pancakes	flavorless	reviewers
potatos	horribleappetizers	critiques
hadhorrible	inedible	soggy
bad	trashed	reviews

(a) Original negative sentence is "*the red velvet pancakes were horrible and brown, and potatos were over cooked and bland.. would not recommend*"

x	$x'_{s=0}$	$x'_{c=0}$
dvd	unusable	reverb
eject	expired	throw
disks	cancelled	film
unusable	crappy	review
purchased	trashed	trip

(b) Original negative sentence is "*Worked for about 4 months. DVD player will not eject or accept disks. Do not buy.*"

Table 5: Top-5 nearest words on Yelp and Amazon. The sentence embedding x captures both semantic and contextual tokens. In contrast, $x'_{s=0}$ clusters strongly around negative sentiment expressions, while $x'_{c=0}$ clusters tokens associated with contextual metadata.

4.6 Generalization and sensitivity analysis

4.6.1 Advanced tasks

Table 6 presents the generalization capability of CCL across advanced tasks. The unseen generation task involves sentiment reversal paraphrasing: the model rewrites a negative sentence to express the opposite sentiment, and we automatically assess its sentiment using GPT-4o-mini. Although CCL trains only on classification tasks, it generalizes well to this unseen generation setting. For MMLU [42], we retrieve five examples for each query without distinguishing among the 57 domains. For HotpotQA [43], we provide each query with its corresponding document and retrieve examples to form document-example pairs. This experiment provides evidence that CCL may help with hierarchical and composite language-understanding problems.

	Unseen & generation		Reasoning	Multi-hop QA
	Yelp	Amazon	MMLU	HotpotQA
ZS	86.26	86.73	60.48	82.43
ICL	87.68	85.80	61.37	84.14
CCL	**90.05**	**87.70**	**61.52**	**84.43**

Table 6: Performance comparison of ZS, ICL, and CCL across tasks using Phi-4-mini-IT.

4.6.2 Sensitivity analysis

	QNLI	PIQA	WSC273	YELP
ZS	43.4 (± 0.00)	71.3 (± 0.00)	55.3 (± 0.00)	89.0 (± 0.00)
LLM_R	29.9 (± 0.00)	69.9 (± 0.00)	61.2 (± 0.00)	79.5 (± 0.00)
ICL	68.1 (± 0.00)	69.0 (± 0.00)	49.8 (± 0.00)	75.8 (± 0.00)
CCL	**75.2 (± 0.45)**	**72.4 (± 1.12)**	**58.98 (± 2.78)**	**95.10 (± 0.25)**

(a) Mean accuracy and std over 5 random seeds.

dim(c)	QNLI	PIQA	WSC273	YELP	Avg.
128	69.5	**72.6**	56.1	93.5	72.9
256	**75.3**	71.6	60.4	94.9	75.6
1024 (ours)	75.2	70.5	**61.9**	**95.4**	**75.7**

(b) Accuracy variation w.r.t. dim(c).

Table 7: Performance of CCL under different training conditions using Llama-3.2-3B-IT. (a) OOD benchmark accuracy across five random seeds, showing stable results despite stochastic variation in VAE training. (b) Performance changes with respect to latent dimensions, indicating that smaller dimensions do not significantly degrade accuracy.

Table 7a shows the OOD benchmark results under different random seeds used for training the VAE within CCL. Since response generation is deterministic (non-sampling), other baselines exhibit zero variance. Table 7b reports the effect of varying the latent dimensions of c and s during VAE training. The results suggest that model performance remains stable even with smaller latent dimensions.

5 Conclusion and discussion

We propose CCL, the first framework to integrate causal representation learning into ICL, addressing a key limitation of conventional ICL in OOD settings. By selecting demonstrations based on causal representation rather than surface-level similarity, CCL improves robustness, and parameter estimation, with theoretical guarantees.

Limitation and Impact statement. Since CCL employs a VAE-based latent embedding, the inherent structural limitations of VAE may hinder its ability to fully capture the rich and nuanced representations of natural language. We leave the deeper integration of embedding-based retrieval with causal inference as future work.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00457216).

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[2] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[3] Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. In-context learning generalizes, but not always robustly: The case of syntax. *arXiv preprint arXiv:2311.07811*, 2023.

[4] Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, and Aaron Courville. On the compositional generalization gap of in-context learning. *arXiv preprint arXiv:2211.08473*, 2022.

[5] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

[6] Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts. *arXiv preprint arXiv:2305.16704*, 2023.

[7] J Pearl. *Causality*. Cambridge university press, 2009.

[8] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[9] Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*, 2022.

[10] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[11] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[12] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[13] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

[14] Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2024.

[15] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[16] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR, 2023.

[17] Sharut Gupta, Stefanie Jegelka, David Lopez-Paz, and Kartik Ahuja. Context is environment. In *The Twelfth International Conference on Learning Representations*, 2023.

[18] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR, 2022.

[19] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

[20] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[21] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.

[22] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*, 2022.

[23] Sarthak Mittal, Eric Elmoznino, Leo Gagnon, Sangnie Bhardwaj, Dhanya Sridhar, and Guillaume Lajoie. Does learning the right latent variables necessarily improve in-context learning? *arXiv preprint arXiv:2405.19162*, 2024.

[24] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[25] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[26] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[27] Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *arXiv preprint arXiv:2310.11011*, 2023.

[28] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

[29] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint arXiv:1812.03253*, 2018.

[30] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[31] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[32] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6155–6170. Curran Associates, Inc., 2021.

[33] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[34] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[35] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

[36] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.

[37] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

[38] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[39] Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, 2024.

[40] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[41] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

[42] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[43] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction of the paper, we compare and introduce our core contribution, which is the first study to alleviate the difficulties of existing ICL methods in OOD in-context learning from a causal representation perspective, and validate the significance of our methodology through theoretical validation and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss our limitations in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We proceed with the theoretical analysis of our methodology in Section 3. The necessary assumptions and definitions for the theoretical analysis are mentioned in the main text, and the proofs are given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: We mention specific experimental setups in the text and appendix, and ensure reproducibility by providing a code repository.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer:[Yes]

   Justification: We provide all the code and data we used in our experiments in a code repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe in the text how we selected examples from ICL and which linguistic models we experimented with. For the sake of brevity, the specific experimental setup can be found in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report the mean and standard deviation of performance metrics over five random seeds to assess the consistency of our method. However, we did not perform formal statistical significance tests or report confidence intervals.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We mention in the appendix the resources required for the experiments due to the constraints of the paper length.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We followed the EthicsGuidelines for our review. Reviewed according to the

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We do not discuss this paper due to the lack of relevant experiments.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We didn't do anything differently because we were already using a widely used benchmark.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all code and data we reference.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Not relevant.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Not relevant.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Does not describe

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We got some help with simple grammar corrections and LaTeX syntax from LLM.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.