IN-CONTEXT ADAPTATION

Yongqiang Chen[1,2], Chenxi Liu[3], Qingyi Guo[3], Bo han[3], Kun Zhang[1,2]

[1]MBZUAI, [2]Carnegie Mellon University, [3]TMLR Group, Hong Kong Baptist University
yqchen24@gmail.com

ABSTRACT

A defining characteristic of intelligence is reasoning – a capability of adapting learned knowledge to unfamiliar contexts. Although large language models (LLMs) exhibit strong reasoning capabilities in-context, it remains unclear whether they can perform human-like in-context reasoning under out-of-distribution (OOD) contexts. In this work, we introduce `In-Context Adaptation` (`ICA`), a paradigm that formalizes reasoning as the adaptive use of learned knowledge through a few demonstrations from new environments. Using a benchmark adapted from invariant learning, we show that transformers trained via next-token prediction are prone to spurious correlations and fail to reason effectively in OOD settings. To address this limitation, we propose `Adaptive Context Engineering` (`ACE`), a simple context reconstruction strategy that promotes adaptive exploitation of learned knowledge. Empirical results demonstrate significant improvements in in-context reasoning under distribution shifts, highlighting a path toward more human-like adaptive generalization in transformers.

1 INTRODUCTION

A definite capability of intelligence is *reasoning* – a process to recognize, adapt, and extrapolate the learned knowledge to solve tasks under unfamiliar contexts (Lombrozo, 2024). Recently, large language models (LLMs) have demonstrated similar capabilities as humans in adapting the knowledge learned from a massive scale of training to solve complex problems (Wei et al., 2022; OpenAI, 2024; Guo et al., 2025; Li et al., 2025b). At the heart of the success of LLMs is the capability of performing in-context learning (ICL) based on the architecture of transformers (Brown et al., 2020), where the model is able to learn and reason for the answer of a new task given a few demonstrations in-context (Dong et al., 2024).

Nevertheless, LLMs also demonstrate *trivial failures* in tasks that are intuitive and simple for humans. For example, LLMs can not solve simple logical reasoning tasks (Nezhurina et al., 2024), nor extract the desired information from an reversed order (Berglund et al., 2023), or can easily overlook critical information in the contexts (Shaikh et al., 2023; Li et al., 2024). Even with the state-of-the-art reinforcement learning training, the frontier reasoning LLMs can easily get lost in multi-turn conversations and hallucinate (Li et al., 2025a; Zou et al., 2026; Fan et al., 2026). The behaviors of LLMs *deviating from human reasoning* show the limitations of the existing LLM training and in-context learning paradigm in imitating human intelligence (Kargupta et al., 2025; Liu et al., 2026). Hence, it raises a challenging research question:

How can we train models to adapt and reason in-context like humans?

To answer the question, we resort to the literature of out-of-distribution (OOD) generalization that seeks to find the *invariance* for generalizing across different environments (Arjovsky et al., 2019) and avoid learning *spurious correlations* or shortcuts that hold only at training environments (Geirhos et al., 2020). As it is relatively challenging to establish the notion of OOD generalization for LLMs that are trained on arguably

all available data, we present a new paradigm to study the generalization capabilities of transformers with large-scale pretraining, termed as `In-Context Adaptation` (`ICA`). `ICA` requires the model not only to learn the desired knowledge for generalization, but also utilize the learned knowledge adaptively. As shown in Fig. 1, different from previous paradigms that focus on identifying the sample-level characteristics (Gupta et al., 2024) or the sample-label correlations (Brown et al., 2020) in order to generalize to samples from the same environment, `ICA` focuses on the adaptive utilization for generalization under OOD contexts that may not appear during training.

To examine the `ICA` capabilities of existing paradigms of next-token prediction training and in-context learning, we adapt the COLOREDMNIST dataset from Arjovsky et al. (2019). COLOREDMNIST requires the model to classify hand-written digits in the colored MNIST data. In the training environments, COLOREDMNIST constructs spurious correlations based on the colors and the labels. Therefore, the learned knowledge includes the correlations between the color and the label, as well as the correlations between the digit and the label. We then evaluate whether transformers trained through next-token prediction (Gupta et al., 2024) can generalize to samples from *OOD environments* where the correlations between the color and the label are *invalid*. We find that in-context learning with either GPT2 (Radford et al., 2019) architecture, or MEMORYMOSAICS, which shows the state-of-the-art ICL capabilities (Zhang et al., 2025), can not solve `ICA` under different OOD contexts.

To mitigate the issue, we develop a simple strategy called `Adaptive Context Engineering` (`ACE`) that strategically reconstructs the context to improve the adaptation capabilities of transformers. We show that `ACE` brings significant improvements based on both GPT2 and MEMORYMOSAICS architectures. Our contributions can be summarized as follows:

- We propose a new paradigm of generalization called `In-Context Adaptation` to study OOD generalization capabilities of transformers.
- We propose a simple strategy called `Adaptive Context Engineering` that significantly improves the `ICA` capabilities of transformers.

2 IN-CONTEXT ADAPTATION

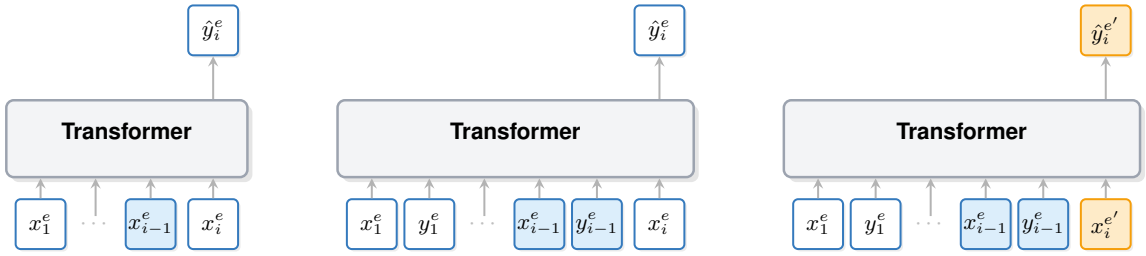
2.1 PARADIGMS OF GENERALIZATION

We begin by discussing and comparing different paradigms of generalization in the literature. A detailed comparison of existing paradigms is given in Table 1.

OOD Generalization. The OOD generalization problem is commonly formulated within a supervised learning framework, where the dataset $\mathcal{D} = {\mathcal{D}^e}_{e\in\mathcal{E}_{\text{all}}}$ is collected from multiple causally related environments $\mathcal{E}_{\text{all}}$. In each environment $e \in \mathcal{E}_{\text{all}}$, a subset of samples $\mathcal{D}^e = X_i^e, Y_i^e$ is independently and identically drawn from the distribution $\mathbb{P}^e$ (Peters et al., 2016).

Given data from a set of training environments ${\mathcal{D}^e}_{e\in\mathcal{E}_{\text{tr}}}$, the objective of OOD generalization is to learn a predictor $f : \mathcal{X} \to \mathcal{Y}$ that performs robustly across all environments, including unseen ones. Formally, this corresponds to minimizing the worst-case empirical risk across environments, i.e., $\max_{e\in\mathcal{E}_{\text{all}}} \mathcal{L}_e(f)$, where $\mathcal{L}_e$ denotes the empirical risk under environment e. The predictor is typically parameterized as a composition $f = w \circ \varphi$, where the feature extractor $\varphi : \mathcal{X} \to \mathcal{Z}$ maps inputs to a latent representation space, and the classifier $w : \mathcal{Z} \to \mathcal{Y}$ produces predictions based on the learned features.

A predominant body of solutions to OOD generalization is to seek the *invariance* between training and test environments (Ganin et al., 2016; Arjovsky et al., 2019; Sagawa* et al., 2020). However, it has been shown that seeking invariance can significantly affect the learning of useful features and pose an optimization



(a) Standard (e) (b) ICL (e) (c) Adapt ($e \to e'$)

Figure 1: Overview of the different in-context learning paradigms at test-time: (a) Standard Generalization (Gupta et al., 2024) generalizes empirical risk minimization (ERM) (Vapnik, 1991) to in-context and aims to train the transformer to recognize the environment of the given context from only samples $\{x_1^e, \ldots, x_{i-1}^e\}$ and to retrieve the learned knowledge from the closest training environment to e; (b) In-Context Learning (Brown et al., 2020) aims to infer the mapping strategy from the given sample – label pairs at the context $\{(x_1^e, y_1^e), \ldots, (x_{i-1}^e, y_{i-1}^e)\}$ and predict the label of the query sample x_i^e from the same environment e; and (c) In-Context Adaptation aims to infer what learned knowledge or correlations can be adapted based to the test context on the given sample – label pairs at the context $\{(x_1^e, y_1^e), \ldots, (x_{i-1}^e, y_{i-1}^e)\}$ and predict the label of the query sample $x_i^{e'}$ that may come from a different environment $e' \neq e$.

Table 1: **Comparison of paradigms.** At test time (step i), the model receives a context sequence (indices $1 \ldots i-1$) and a query x_i, enabling amortized inference.

Paradigm	Training data	Testing data	Estimates / objective
ERM	Labeled data (pooled) $\{(x_j, y_j)\} \sim \cup_{e \in \mathcal{E}_{tr}} P_e(X, Y)$	Query only $x_i^{e'}$	Global predictor $f(x_i) \approx P(Y \mid X)$
IRM / invariance DG Arjovsky et al. (2019)	Multi-environment labeled data $\{(x_j^e, y_j^e, e)\}_{e \in \mathcal{E}_{tr}}$	Query only $x_i^{e'}$	Invariant predictor $f(x_i) = g(\phi_{\text{inv}}(x_i)) \approx P(Y \mid \phi_{\text{inv}}(X))$
LLM next-token LM Brown et al. (2020)	Token sequences $z = (z_1, \ldots, z_T)$	Prefix / context $z_{1:i-1}$	Next-token predictor $f(z_{1:i-1}) \approx P(z_i \mid z_{1:i-1})$
ICRM Gupta et al. (2024)	Multi-env episodic sequences $\{(x_j^e, y_j^e)\}, e \in \mathcal{E}_{tr}$	Query $x_i^{e'}$ and context $C_i^{e'} = (x_j^{e'})_{j=1}^{i-1}$	In-context DG predictor $f(x_i, C_i) \approx P(Y_i^{e'} \mid x_i^{e'}, C_i^{e'})$ (amortized adaptation to e')
ICA (ours)	Same as above (or any multi-env training) + optionally label noise	Query $x_i^{e'}$ and labeled context $S_i^{e'} = \{(x_j^{e'}, y_j^{e'})\}_{j=1}^{i-1}$	In-context adaptation predictor $f(x_i, S_i) \approx P(Y_i^{e'} \mid x_i^{e'}, S_i^{e'})$ *Goal:* approximate $P_{e'}(Y \mid X)$ without gradient updates

dilemma due to the intrinsic conflicts between empirical risk minimization and OOD generalization (Chen et al., 2023b). Zhang et al. (2022); Chen et al. (2023a) present rich feature learning and show that learning all the underlying useful features is essential for smooth OOD generalization.

In-Context Learning with Next-Token Predictors. Brown et al. (2020) demonstrate the promise of transformers trained with next-token prediction and learn to reason for the answers of the sample x_i^e given the demonstration contexts $\{(x_1^e, y_1^e), \ldots, (x_{i-1}^e, y_{i-1}^e)\}$. Usually, the query sample and the context samples are from the sample environment.

Formally, the next-token predictor is trained to predict the next token z_i given the prefix token sequence $(z_1, \ldots, z_{i-1})$:

$$P(Z_i = z_i | z_1, \ldots, z_{i-1}). \tag{1}$$

The success of LLMs is built upon this paradigm and demonstrates emerging and impressive capabilities of learning in-context (Brown et al., 2020). The in-context generalization capability can be further manifested, conditioning on proper instructions (Wei et al., 2022) and reinforcement learning (Guo et al., 2025).

Following the success of the ICL, Gupta et al. (2024) propose the paradigm of In-Context Risk Minimization (ICRM) that replaces the labeled context of $\{(x_1^e, y_1^e), \ldots, (x_{i-1}^e, y_{i-1}^e)\}$ with unlabeled ones $\{x_1^e, \ldots, x_{i-1}^e\}$. They demonstrated that by constructing a context sequence based on samples from the training environment

$$P(y_i^e | \varphi(x_i^e), \varphi(x_1^e), \cdots, \varphi(x_{i-1}^e)), \tag{2}$$

to train transformer-based next-token predictors for OOD generalization. Nevertheless, ICRM differs from human reasoning where humans are given some sample labels or feedback to infer which learned knowledge is useful for generalization under the test context (Lombrozo, 2024).

In-Context Adaptation. Therefore, we propose a new paradigm of generalization called `In-Context Adaptation` (`ICA`) to fully model the realistic generalization of human reasoning. During training, the next-token predictor is given with the labeled pairs as ICL:

$$P(y_i^{e'} | \varphi(x_i^{e'}), \varphi(x_1^e), \varepsilon(y_1^e), \cdots, \varphi(x_{i-1}^e), \varepsilon(y_{i-1}^e)), \tag{3}$$

where the $\varepsilon(\cdot)$ is the encoder to encode label information for next-token prediction training. Different from previous paradigms, the primary objective of `ICA` is to evaluate the generalization capability in adapting the learned knowledge to OOD contexts of the next-token predictor. The environments of the context and the query sample are not necessarily the same. To emulate the realistic context scenarios in `ICA`, we also construct different scenarios, such as label noises, and a lack of labels in Sec. 3.

Evaluation of ICA. To evaluate the generalization capabilities of different training paradigms in `ICA`, we construct the evaluation protocol based on ColoredMNIST data from Arjovsky et al. (2019).

The data generation process of ColoredMNIST can mainly be parameterized by $\alpha_e, \beta_e \in [0, 1]$. For each environment e, the dataset $\mathcal{D}_e = \{X^e, Y^e\}$ is generated as

$$Y^e := \text{Rad}(0.5), \quad X^e = (X_1^e, X_2^e), \quad X_1^e := Y^e \cdot \text{Rad}(\alpha_e), \quad X_2^e := Y^e \cdot \text{Rad}(\beta_e), \tag{4}$$

where $\text{Rad}(\sigma)$ denotes a random variable that takes the value -1 with probability σ and $+1$ with probability $1 - \sigma$. The environment in ColoredMNIST can be denoted as $\mathcal{E}_\alpha = \{(\alpha, \beta_e) : 0 < \beta_e < 1\}$. In this setup, X_1^e functions as the invariant feature since α is fixed across environments, whereas X_2^e serves as a spurious feature because β_e varies with e. The training environments in ColoredMNIST contain $\{(0.25, 0.10), (0.25, 0.20)\}$ and the test environment is $\{(0.25, 0.90)\}$ where the spurious correlations between X_2^e and Y is inversed. Empirically, as we show in Sec. 3, the next-token predictor trained via previous paradigms can not generalize to OOD contexts.

Adaptive Context Engineering. We propose a simple strategy to mitigate the drawbacks of previous training methods of next-token predictors called `Adaptive Context Engineering` (`ACE`). `ACE` is built upon the intuition to strategically construct contexts for enabling adaptation following the definition of

Table 2: **Main Results**: Detailed comparison of Backbone and Configurations.

Config	Backbone	Train Acc	Test Accuracy				
			0	25	50	75	100
ACE	MEMORYMOSAICS	**92.5**	18.8	**62.1**	**62.2**	**62.1**	**62.1**
ICA	MEMORYMOSAICS	85.0	18.0	58.7	58.5	58.2	58.4
ICRM	MEMORYMOSAICS	86.5	24.0	23.0	24.0	32.0	30.0
ACE	GPT2	91.5	**23.0**	24.8	25.2	25.0	25.8
ICA	GPT2	84.0	17.8	18.7	18.5	18.3	18.7
ICRM	GPT2	83.5	14.0	7.0	10.0	20.0	15.0
ERM	MLP	88.0	11.9	11.9	11.9	11.9	11.9

ICA as in Eq. 3. As $(x_i^{e'}, y_i^{e'})$ are harder samples for the next-token predictors to predict during training, we construct a buffer to collect all the hard samples following the strategy of Chen et al. (2023a), and construct contexts based on the hard samples in the buffer and the contexts from all the training environments. In Sec. 3, we show that this simple strategy can already brings significant improvements over previous methods.

3 EXPERIMENTS

Evaluation protocol. To emulate realistic OOD context scenarios, we consider a number of cases inspired by human cognitive capabilities and the realistic challenges of LLM reasoning:

- Matched Context: the vanilla ICA setting, where the test context and the query sample are from the same OOD environment with respect to the training environment. It evaluates whether the model is able to recognize useful features learned from training environments (Lombrozo, 2024), as well as the relations given by the test context (Brown et al., 2020).
- Mismatched Context: the test context and the query sample are from different environments. Hence, the context does not provide information and even strengthens the spurious correlation, where the next-token predictor needs to calibrate the contextual clues (Zhao et al., 2021).
- Mixed Context: the test context are sampled from mixture of training and test environments.
- Few-show Context: the test context contains only few samples and the adaptation needs to happen fast (Evans & Stanovich, 2013).
- GT label and pred label: The ground-truth (GT) labels are not given in Pred label setting, where the next-token predictor needs to predict the label themselves as humans encounter some new environments without labels (Evans & Stanovich, 2013).

Baselines. We mainly consider ICRM and vanilla ICA as our baselines. We also implement all the methods based on GPT2 backbone and MEMORYMOSAICS backbone, where the latter is the state-of-the-art transformer architecture for in-context learning.

Empirical results. Table 2 presents the main OOD generalization results. We can find that when without the label information in the context, as in ICRM, the generalization results are poor. Under both GPT2 and MEMORYMOSAICS backbones, ACE and ICA significantly improve the performances under OOD contexts.

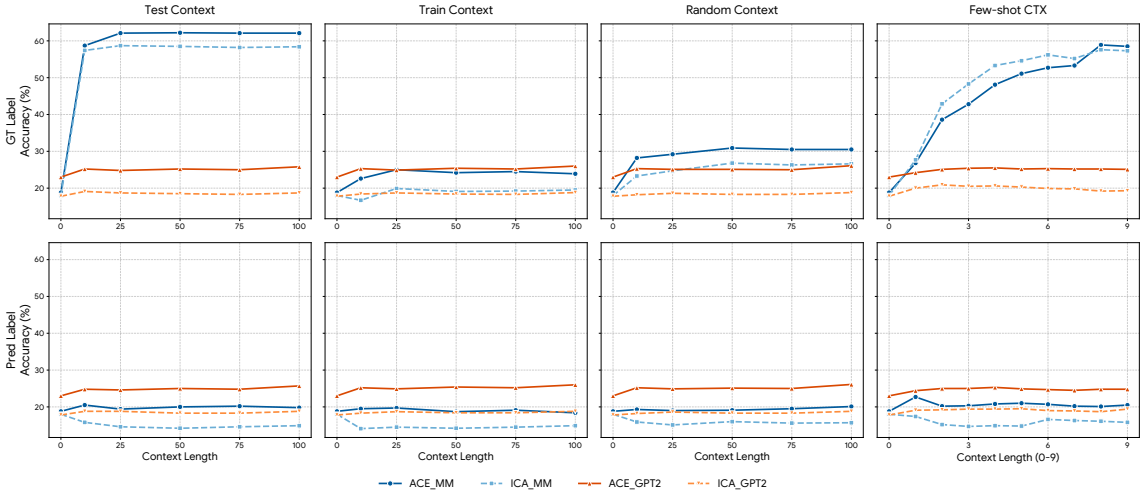


Figure 2: Performances of ICA and ACE under different ICA scenarios.

Fig. 2 shows the performances of ICA and ACE based on GPT2 and MEMORYMOSAICS backbones under different adaptation context settings. It can be found that ACE generically achieves better results than the vanilla ICA across different setttings. Although when given few samples with GT labels, ACE obtains similar performance as ICA based on MEMORYMOSAICS, which can be attributed to the strong ICL capability of MEMORYMOSAICS, notably, when given sufficient context information, ICA achieves better results under both GPT2 and MEMORYMOSAICS backbones.

When without GT label, the strong ICL capabilities of MEMORYMOSAICS also suffer from a severe performance decrease due to the noise in the predicted labels in the context, when with ICA. In contrast, ACE achieves significantly better results under the Pred Label setting given both GPT2 and MEMORYMOSAICS backbones.

4 CONCLUSIONS

In this work, we presented a new paradigm of generalization called In-Context Adaptation (ICA) to emulate how humans adapt the learned knowledge for generalizing to OOD contexts during reasoning. As we showed that previous training methods for next-token predictors are incapable of realizing ICA, we proposed a simple strategy called ACE that obtained significant improvements under a number of OOD contexts. ICA and ACE highlight a new path towards training more human-like reasoning models based on transformers.

ACKNOWLEDGMENTS

We thank the reviewers for their constructive comments and suggestions.

LLM USE STATEMENT

From the research side, this work presents a new simulation and study environment for human-like reasoning of next-token predictors. From the paper writing side, we use LLMs to assist with improving the writing of this work.

ETHICS STATEMENT

We study implementing human-like reasoning via next-token predictors that will benefit the whole humanity and society. This work does not involve human subjects or personally identifiable information beyond public benchmarks used under their licenses.

REFERENCES

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. (Cited on pages 1, 2, 3 and 4)

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint*, arXiv:2309.12288, 2023. (Cited on page 1)

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 *Advances in Neural Information Processing Systems*, 2020. (Cited on pages 1, 2, 3, 4 and 5)

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2023a. (Cited on pages 3 and 5)

Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, MA KAILI, Han Yang, Peilin Zhao, Bo Han, and James Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023b. (Cited on page 3)

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128. Association for Computational Linguistics, 2024. (Cited on page 1)

Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013. (Cited on page 5)

Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, and Maksym Andriushchenko. Halluhard: A hard multi-turn hallucination benchmark. 2026. (Cited on page 1)

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Mache Learning Research*, 17:59:1–59:35, 2016. (Cited on page 2)

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. (Cited on page 1)

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. (Cited on pages 1 and 4)

Sharut Gupta, Stefanie Jegelka, David Lopez-Paz, and Kartik Ahuja. Context is environment. In *International Conference on Learning Representations*, 2024. (Cited on pages 2, 3 and 4)

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L Griffiths, Max Kleiman-Weiner, Jiawei Han, Asli Celikyilmaz, and Yulia Tsvetkov. Cognitive foundations for reasoning and their manifestation in llms. *arXiv preprint arXiv:2511.16660*, 2025. (Cited on page 1)

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Steering llms towards unbiased responses: A causality-guided debiasing framework. *arXiv preprint*, arXiv:2403.08743, 2024. (Cited on page 1)

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025a. (Cited on page 1)

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025b. (Cited on page 1)

Chenxi Liu, Yongqiang Chen, Tongliang Liu, James Cheng, Bo Han, and Kun Zhang. On the thinking-language modeling gap in large language models. In *International Conference on Learning Representations*, 2026. (Cited on page 1)

Tania Lombrozo. Learning by thinking in natural and artificial minds. *Trends in Cognitive Sciences*, 2024. (Cited on pages 1, 4 and 5)

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*, 2024. (Cited on page 1)

OpenAI. Introducing openai o1-preview. https://openai.com/index/introducing-openai-o1-preview/, 2024. Accessed: 2024-09-12. (Cited on page 1)

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. (Cited on page 2)

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533. (Cited on page 2)

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. (Cited on page 2)

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4454–4470. Association for Computational Linguistics, July 2023. (Cited on page 1)

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pp. 831–838, 1991. (Cited on page 3)

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 1 and 4)

Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. *arXiv preprint arXiv:2203.15516*, 2022. (Cited on page 3)

Jianyu Zhang, Niklas Nolte, Ranajoy Sadhukhan, Beidi Chen, and Leon Bottou. Memory mosaics. In *International Conference on Learning Representations*, 2025. (Cited on page 2)

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*, pp. 12697–12706, 2021. (Cited on page 5)

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Garry Yang, Mufei Li, Qing Da, James Cheng, Pan Li, and Yu Gong. Reducing belief deviation in reinforcement learning for active reasoning. In *International Conference on Learning Representations*, 2026. (Cited on page 1)