In-Context Exemplars as Clues to Retrieving from Large Associative Memory

Jiachen Zhao
Department of Computer Science
University of Massachusetts Amherst
jiachenzhao@umass.edu

Abstract

Recently, large language models (LLMs) have made remarkable progress in natural language processing (NLP). The most representative ability of LLMs is in-context learning (ICL), which enables LLMs to learn patterns from in-context exemplars without training. However, there remains limited intuition for how in-context learning works. In this paper, we present a novel perspective on prompting LLMs by conceptualizing it as contextual retrieval from a model of associative memory, which can be biologically plausible. We establish a theoretical interpretation of ICL based on an extension of the framework of Hopfield Networks. Based on our theory, we further analyze how in-context exemplars influence the performance of ICL. Our study sheds new light on the mechanism of ICL by connecting it to memory retrieval, with potential implications for advancing the understanding of LLMs.

1 Introduction

In recent years, large language models (LLMs) have garnered significant attention due to their ability to revolutionize natural language processing (NLP) by demonstrating impressive language understanding and reasoning capabilities (5; 4; 29; 37; 28). LLMs are first pretrained on extensive data using the language modeling technique where the model predicts the next token given a context. Without finetuning on task-specific data, LLMs leverage in-context learning (ICL), also referred to as few-shot prompting, to make predictions. Through ICL, LLMs can find underlying patterns of the input query through given in-context exemplars, such as a set of input/output pairs, and use them to complete the response.

The understanding of ICL currently remains intuitive and lacks theoretical foundation. Past works on ICL mainly focus on empirical investigation (23; 11; 19; 40) or data distribution to explain how ICL emerges (6; 38). In this work, we adopt a novel and distinct perspective by theoretically reframing ICL as contextual *retrieval* rather than a *learning* problem (8; 35; 9), as there is no actual weight update involved. We conceptualize LLM as a biologically plausible model of associative memory (13), also known as content-addressable memory. In the realm of machine learning, memory models (13; 17; 31; 16; 25; 18) have been widely studied for a long time. Two fundamental models are Hopfield Network (13) and its extension, sparse distributed memory (17). The retrieval process can also be viewed as pattern recognition (18).

Through the lens of memory models, we demonstrate that ICL with self-attention (34) in LLMs can be interpreted as retrieving patterns from associative memory of Hopfield Networks with context. Correspondingly, we establish a theoretical framework and analyze retrieval error. We further look into the influence of in-context exemplars on performance of ICL based on our contextual retrieval formulation. We also conduct extensive experiments as sources of evidence to our theoretical analysis.

Associative Memory & Hopfield Networks in 2023. NeurIPS 2023 workshop.

2 ICL as Contextual Retrieval

This section presents a formulation of ICL as pattern retrieval based on context from memories of modern Hopfield Networks (MHNs)(25; 18; 22). Brief overview of Hopfield Networks is provided in Appendix A. In this section, we first give a formal setup of ICL. For a pre-trained language model whose parameters are denoted as θ, given an input $\mathbf{x}$, the model will predict $\tilde{\mathbf{y}}$ for ground truth by conditioning on the query and a context sequence containing K exemplars that are drawn from an accessible labeled dataset $\mathcal{D}_{(x,y)}$ (each exemplar $e_i = (x_i, y_i)$). Formally, we denote the sequence of all K in-context exemplars $\mathbf{e}$, i.e., $\mathbf{e} = e_1, ..., e_K$. We can then have

$$\tilde{\mathbf{y}} = \text{argmax}_y P(y|\mathbf{e}, \mathbf{x}, \theta). \tag{1}$$

From the perspective of HNs, the input string [$\mathbf{e}$,$\mathbf{x}$] is a cue to the associative memory. The feed-forwarding process in the language model is to reconstruct the completion $\tilde{\mathbf{y}}$ of $\mathbf{x}$ that aligns with patterns of context $\mathbf{e}$ by recalling information stored into the model's memory during pretraining. For LLM, the pretraining is implemented as predicting masked/ next tokens of sentences, which is essentially teaching the model to reconstruct completion based on context like HNs.

To demonstrate the close relation between ICL and retrieving from HNs, we first extend the model definitions discussed by Ramsauer et al. (25); Millidge et al. (22) to construct a Hopfield Network with Context (HN-C). We then show that contextual retrieval from HN-C is equivalent to self-attention in LLMs. To incorporate context in HNs, we consider stored patterns in memory as applying a linear transformation to raw vectors with a memory matrix, which is different from past frameworks (25; 18; 22) that assume a static array of stored patterns. Thus, in our case, context patterns are dynamically defined depending on the input context. It is also important to note that retrieval does not necessarily mean extracting the exact stored patterns in memory without loss, but rather involves the induction of completion based on the input patterns that are typically not fully identical to the stored memories (18). Actually, contextual retrieval setting is indeed how the human brain retrieves episodic memory (26).

Formally, we denote some underlying query vector of input strings by $\sigma \in \mathbb{R}^{d_m}$. We define there are M context vectors $\lambda_i \in \mathbb{R}^{d_m}$ which are represented by a matrix $\Lambda \in \mathbb{R}^{d_m \times M}$. We define memory matrix $\xi_Q \in \mathbb{R}^{d_m \times d_q}$ and $\xi_K \in \mathbb{R}^{d_m \times d_q}$ respectively for query vector σ and context vector λ. We further define $\mathrm{Z} := \xi_K^T \Lambda$ and each column vector z is **context pattern**. Accordingly, we have $u := \sigma \xi_Q$ as **query pattern**. We then define the update rule for u of the model based on the Universal Hopfield Network (22) as follows:

$$u^{\text{new}} = sep(\gamma\, sim(u,\, \mathrm{Z}))\Lambda^T \xi_K, \tag{2}$$

where γ is a scalar value, sim is a similarity function and sep is a separation function. We set sim as dot production and sep as $softmax$ function. Then the update rule can be further specified as Eq. 4.

$$u^{\text{new}} = softmax(\gamma\, u\mathrm{Z})\Lambda^T \xi_K \tag{3}$$

$$= softmax(\gamma\, \sigma \xi_Q \xi_K^T \Lambda)\Lambda^T \xi_K \tag{4}$$

This formulation can be converted to self-attention by applying a linear transformation to u^{new}, i.e., $u^{\text{new}} W_v = softmax(\gamma \mathbf{Q}\mathbf{K}^T)\mathbf{V}$, where we write $\sigma \xi_Q = \mathbf{Q}$, $\xi_K^T \Lambda = \mathbf{K}^T$ and $\Lambda^T \xi_K W_v = \mathbf{V}$. Therefore, the update rule for contextual retrieval from HNs can be equivalent to self-attention through simple conversion. For self-attention, query pattern is $\mathbf{Q}$ and the context pattern is namely $\mathbf{K}$. We give further detailed interpretation of ICL as pattern retrieval in Appendix B.

Definition 1 (Query-Context Separation) *For z_i, $\delta := uz_i - uz_j$, where $z_j \neq \{z_i | i \in [1, M]\}$ and $i, j \in [1, M]$.*

We then establish the distinction in similarity scores between context patterns and query patterns as a metric for evaluating the degree of separation between two context patterns with respect to the query pattern. The larger δ is between z_i and some other patterns z_j, the easier it will be for z_i to be matched by the query pattern u. Moreover, we define the pattern retrieval error as $\|f(u) - u^\star\|$, where f is the update rule for the query pattern and $u^\star$ is the corresponding underlying ground-truth pattern of y in the same associative space to u. It is assumed that both the query vectors and context vectors follow some distribution within the pre-trained model, allowing the model to effectively capture and represent their patterns. Different from the defined error of HNs in (25), we consider a general case where u^T is not necessarily in $\{z_i | i \in [1, M]\}$.

Theorem 1 (Retrieval Error) *For some z_i that has t instances, i.e., $t = \sum_{j=1}^{M} \mathbb{1}\{z_j = z_i\}$. The ground-truth pattern $u^\star = (z_i + \Delta z)^T$. We define $c := exp(-\gamma(uz_i - max_{z_i \neq z_j} uz_j)) = exp(-\gamma\delta_{min})$, and $z_{max} = max(z_1, ..., z_M)$. The retrieval error $\epsilon := \|f(u) - u^\star\|$ is then bounded by $[0,\ \|\Delta z\| + \beta\|z_{max}\|]$,* where $\beta = \left(1 - \left(1 + \frac{c(M-t)}{t}\right)^{-1} + c(M-t)\right)$ and $\beta \propto c\frac{M}{t}$

The proof is displayed in Appendix D. We can see the upper bound consists of two parts, i.e., $\|\Delta_z\|$ and $\beta\|z_{max}\|$. We name $\|\Delta_z\|$ as **Instance Error** which directly reflects the match between a context pattern z_i and the target pattern $u^\star$. On the other hand, $\beta\|z_{max}\|$ is named as **Contextual Error** that mainly indicates the separation of z_i from other context patterns (remind that $\beta \propto c\frac{M}{t} = \exp(-\gamma\delta_{min})\frac{M}{t}$), i.e., how easy for the model to rely on z_i more for the retrieval. Additionally, when $t = 1$ and $||\Delta z|| = 0$, we are directly retrieving the pattern from context patterns stored in the HN. We next discuss two primary questions on ICL, utilizing our retrieval framework as a foundation, and offer some theoretical predictions. Additionally, we show some phenomena implying the biological plausibility of ICL in Appendix C.

How does the relation among context patterns influence retrieval error? We first assume the instance error is already acceptably small, otherwise the decrease of the contextual error in the upper bound can be trivial to the total error. Then from Theorem 1, given the $\|z_{max}\|$, the contextual error is proportionate to $c\frac{M}{t}$. Recall that $c = \exp(-\gamma(uz_i - \max_{z_i \neq z_j} uz_j)) = \exp(-\gamma\delta_{min})$. When δ_{min} is larger, the context pattern z_i is well separated from other distinct context patterns for the query pattern u. z_i will be more prominent when conducting softmax in Eq. 4 and the upper bound of ϵ will be lower, which indicates the potential of smaller error.

How does the number of context patterns influence retrieval error? With the increase of M, $\frac{M}{t}$ and c may change accordingly depending on newly introduced context patterns. Given the instance error, this fluctuation leads to the different tendencies of the upper bound, which means varied potential of the actual error. When context vectors are randomly sampled from the distribution, larger M is often observed to enable generally better performance of ICL (23; 4; 39). However, chances are that if one context pattern z_i already has minimum instance error, larger M may lead to declined performance due to the introduced contextual error from other context patterns (i.e., increased $\frac{M}{t}$ and c). We empirically show this in Sec. 3. Thus, the influence of M can be uncertain depending on chosen context patterns.

2.1 Exemplar Selection

This section analyzes the default random selection based on our retrieval framework. We first detail the definition of exemplar selection.

Definition 2 (Exemplar Selection) *For an input query $\mathbf{x}$ and output $\mathbf{y}$ sampled from distribution $p(\mathcal{D}^{te})$ of task $\mathcal{T}$, a set of K exemplars $\mathcal{S}_{\text{context}}$ is selected from training data $\mathcal{D}^{tr}_{(x,y)}$ to minimize $\ell(\mathbf{y}, \tilde{\mathbf{y}})$, where $\tilde{\mathbf{y}} = \text{argmax}_{\mathbf{y}} P(\mathbf{y}|e_1, ..., e_K, \mathbf{x})$, $e_i = (x_{e_i}, y_{e_i})$, $e_i \in \mathcal{S}_{\text{context}}$.*

We assume $p(\mathcal{D}^{te}) \approx p(\mathcal{D}^{tr}_{(x,y)}) \approx p^*$ that is the population distribution. We also regard patterns as latent variables that underlie string sequences.

Random Selection. Random selection is the default method (4) that can be considered as sampling exemplars from $p(\mathcal{D}^{tr}_{(x,y)})$. When K is large enough, we assume $p(\mathcal{S}_{\text{context}}) \approx p(\mathcal{D}^{tr}_{(x,y)}) \approx p^*$. Accordingly, the mode of context patterns, i.e., $\text{argmax}_z \sum_{j=1}^{M} \mathbb{1}\{z_j = z\}$ may approximate the mode (denoted by $\hat{z}$) of the pattern distribution of samples from p^*. Then for the upper bound of retrieval error ϵ with $z_i = \text{argmax}_z \sum_{j=1}^{M} \mathbb{1}\{z_j = z\}$, the instance error can be approximated to the error given by $\hat{z}$, i.e., $\|\Delta z\| \approx \|\hat{z} - (u^*)^T\|$. Because $\hat{z}$ may more or less be relevant to $(u^*)^T$ when they follow the same pattern distribution, with sufficiently large K, random selection may give a decent retrieval error. On the other hand, when K is small, random selection may perform poorly and have great variance depending on sampled exemplars. We provide empirical verification of our theoretical prediction in Fig. 1 of Sec. 3.

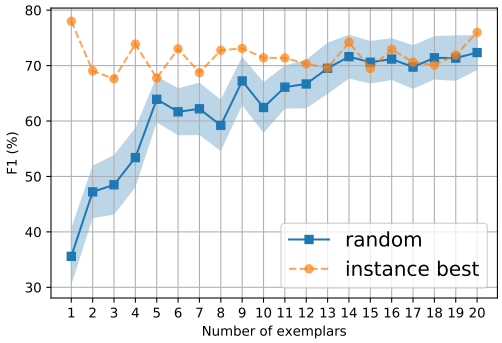

Figure 1: The relation between the performance of ICL with random exemplar selection and the number of exemplars (i.e., K). Apart from random selection, we assume the oracle access to ground truth so as to select the best exemplars for each query (denoted as "instance best").

3 Experiments

Experimental Setup. We consider inducing linguistic structures as our testbed that is a fundamental ability to downstream NLP tasks (20). We employ the commonly used Penn Treebank corpus (21) with the standard splits (2-21) for training containing 39832 sentences, 22 for validation, 23 for test). PTB is an often-used benchmark for constituency parsing in English. The corpus is collected from a variety of sources including stories, news, and scientific abstracts. We employ Code-Davinci-002, known as Codex (7). For the evaluation, we report sentence-level unlabeled parsing F1 that is computed separately for each sentence and then averaged across the dataset.

Effects of K. We conduct experiments with different number of in-context exemplars. The results are shown in Fig. 1. The performance of ICL increases with more exemplars for random selection. Recall in our theory in Sec. 2.1, we consider the drawback of random selection is it needs a large K to reach an optimal status where the mode of exemplars approximates the major patterns of the population. Fig. 1 verifies our reasoning and demonstrates the random selection will achieve a decent performance and then tend to be stabilized with the increase of K.

Comparison with oracle exemplars. For each query instance, we also assume the access to ground truth and rank training data that can give the best performance when used as the only exemplar for the query. We report the average result in Fig. 1. When $K = 1$, such oracle method gives the best result, while the performance drops immediately with additional exemplars and starts to stagnate. This can be caused by the increased contextual error as is discussed in our theoretical analysis of Sec. 2. Including more optimal exemplars turns out to give similar performance to random selection. Therefore, different from supervised tuning, for ICL more exemplars do not always guarantee a better performance, which depends on added exemplars. Additionally, the observation indicates when knowing the optimal exemplar for a query (which is likely to be impossible in practice), we do not need many-shot prompting. However, for cases with no access to such information, simply increasing K may actually be a good strategy for better performance.

4 Conclusion

In this work, we investigate in-context learning (ICL) of large language models through the lens of models of associative memory. We have shown that in-context learning can be theoretically equivalent to contextual retrieval from a Hopfield Network. Based on our theoretical framework, we further analyze the influence of in-context exemplars on ICL performance. All in all, our work interprets ICL as contextual retrieval from memory and links recent LLMs to biologically plausible Hopfield Networks, which may shed new light on understanding LLMs.

References

[1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*, 2022.

[2] Helen C Barron, Ryszard Auksztulewicz, and Karl Friston. Prediction and memory: A predictive coding account. *Progress in neurobiology*, 192:101821, 2020.

[3] Leonardo Bonetti, Elvira Brattico, Francesco Carlomagno, Giovanni Donati, Joana Cabral, Niels Trusbak Haumann, Gustavo Deco, Peter Vuust, and Morten L Kringelbach. Rapid encoding of musical tones discovered in whole-brain connectivity. *NeuroImage*, 245:118735, 2021.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[6] Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[8] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

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

[10] James Eric Eich. The cue-dependent nature of state-dependent retrieval. *Memory & Cognition*, 8:157–173, 1980.

[11] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[12] Duncan R Godden and Alan D Baddeley. Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, 66(3):325–331, 1975.

[13] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[14] Yong Sang Jo and June-Seek Choi. Memory retrieval in response to partial cues requires nmda receptor-dependent neurotransmission in the medial prefrontal cortex. *Neurobiology of learning and memory*, 109:20–26, 2014.

[15] Yong Sang Jo, Eun Hye Park, Il Hwan Kim, Soon Kwon Park, Hyun Kim, Hyun Taek Kim, and June-Seek Choi. The medial prefrontal cortex is involved in spatial memory retrieval under partial-cue conditions. *Journal of Neuroscience*, 27(49):13567–13578, 2007.

[16] Łukasz Kaiser and Samy Bengio. Can active memory replace attention? *Advances in Neural Information Processing Systems*, 29, 2016.

[17] Pentti Kanerva. *Sparse distributed memory*. MIT press, 1988.

[18] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

[19] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Association for Computational Linguistics, 2022.

[20] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

[21] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 1993.

[22] Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pages 15561–15583. PMLR, 2022.

[23] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics, December 2022.

[24] Douglas L Nelson, Cathy L McEvoy, and Martha A Friedrich. Extralist cuing and retrieval inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(2):89, 1982.

[25] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[26] Charan Ranganath and Robert T Knight. Prefrontal cortex and episodic memory: Integrating findings from neuropsychology and functional brain imaging. *The cognitive neuroscience of memory: Encoding and retrieval*, 1:83, 2002.

[27] Kirkpatrick Scott and Sherrington David. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384–4403, 1978.

[28] Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

[29] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[30] Steven M Smith. Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, 5(5):460, 1979.

[31] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

[32] Endel Tulving and Zena Pearlstone. Availability versus accessibility of information in memory for words. *Journal of verbal learning and verbal behavior*, 5(4):381–391, 1966.

[33] Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*, 2022.

[36] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023.

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[38] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

[39] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. *arXiv preprint arXiv:2302.05698*, 2023.

[40] Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. Association for Computational Linguistics, 2022.

A Brief Review of Hopfield Networks

Hopfield Networks (HNs) (13) were introduced to store and retrieve information. The standard HN (13) consists of a neural network of N neurons that can in total store M binary patterns of dimension D. Memory ξ is denoted as an array of stored pattern vectors, i.e., $\xi = [m_1, ..., m_M]$, where $\xi \in \mathbb{R}^{M \times D}$. During the retrieval process, the configuration of neurons is fixed to the query pattern (e.g., incomplete m_i), and an update rule f for σ is defined to retrieve the similar or the same pattern to the query. Each update lowers the energy function E of the network, which belongs to the Ising spin-glass model (27) in physics. The energy is expected to converge to an attractor state (local minimum) through repeated updates. Eventually, HNs will return the pattern from its memory that is the most similar to the input. Additionally, HNs are similar to humans' memory system. The neuron's state corresponds to the firing rate or activity level of biological neurons. The weights of the network correspond to the strength of the synaptic connections between neurons in the brain. Similar to HNs, memories in brains are stored in a distributed manner across many regions and neurons. There are associative areas storing relations between features. Complex memories can then be recalled to generate predictions based on partial cues or associations (3; 2) just like HNs.

B Pattern Retrieval

From the perspective of memory models, ICL can be reinterpreted as retrieving underlying patterns of input based on context vectors λ following the update rule. This interpretation is focused on the association among neurons in some middle layer of the model, where the hidden states at each token position may encode some unique information (1). Query and context are thus assumed to be encoded into separate vectors. The retrieval process consists of the following stages. **(1)** Query vector σ and context vectors λ are mapped to the associative space through linear transformation with ξ_Q and ξ_K to reveal underlying patterns. **(2)** Then a similarity score between u and z is computed to measure their mutual closeness in the associative space. Dot product is considered as the similarity function for self-attention. **(3)** An exponential separation function, i.e., $Softmax$ is computed to stress the prominent context patterns that have higher similarity scores. **(4)** After separation, u^{new} is computed as a weighted sum of context patterns. Remind that there can be repetition in context patterns (which means some $z_i = z_j$). So more frequent context patterns might thus have a larger contribution to the weighted summation.

C Biological Plausibility

The process of ICL in LLMs exhibits similarities to the memory retrieval process in the human brain, both of which involves the use of prompts or cues related to targeted information to retrieve. Similar to LLMs, human memory retrieval also heavily depends on contextual cues for successful recall (33; 10; 12; 30).

Human's memories can actually exist in a state of being *available* but *inaccessible* (32). When some information cannot be recalled with internal cues (i.e., without external hints), such as in free recall tasks, it is considered inaccessible. However, external cues, e.g., category cues related to the target items to recall, can greatly increase the accessibility of memory. Likewise, LLMs can provide answers to questions that they initially fail in zero-shot prompting scenarios when given related in-context exemplars. The query together with in-context exemplars can also be viewed as partial information cues for memory retrieval, providing incomplete or fragmented versions of the target (15; 14). Additionally, the cue-to-target similarity, also known as encoding specificity, is critical to the likelihood of successful recall for human brain (33; 24). Similarly, LLMs that are trained through language modeling may exhibit such requirements for in-context exemplars (36).

For humans, prompts are typically extralist cues, originating from a different list of stored memories to be retrieved. But extralist cues can still be effective if they are relevant to the target (33; 24). Similarly, in the case of ICL, it is uncommon to encounter context and target output that exactly match the training data. However, by providing relevant exemplars, LLMs may still capture underlying patterns of query with the guide of in-context demonstrations and generalize to unseen cases.

D Proof of Theorem 1

Proof:

$$\begin{aligned}\|f(u)-u^\star\| &= \|\Delta z + z_i - t[softmax(\gamma\, u, \mathrm{Z})]_i z_i - \sum_{j,\, z_j \neq z_i}^{M} [softmax(\gamma u \mathrm{Z})]_j z_j\| \\ &= \|\Delta z + \left(1 - t\frac{\exp(\gamma u z_i)}{\sum_j^M \exp(\gamma u z_j)}\right) z_i - \sum_{j,\, z_j \neq z_i}^{M} \frac{\exp(\gamma u z_j)}{\sum_k^M \exp(\gamma u z_k)} z_j\| \\ &= \|\Delta z + \left(1 - \frac{t}{1+\sum_{j,\, j\neq i}^{M} \exp(\gamma(u z_j - u z_i))}\right) z_i - \sum_{j,\, z_j \neq z_i}^{M} \frac{\exp(\gamma(u z_j - u z_i))}{1+\sum_{k,\, k\neq i}^{M} \exp(\gamma(u z_k - u z_i))} z_j\|\end{aligned}$$

For z_i, $\delta_{min} = uz_i - \max_{z_i \neq z_j}(uz_j)$ and recall $t = \sum_{j=1}^{M} \mathbb{1}\{z_j = z_i\}$, so we can get,

$$1 - \frac{t}{1+\sum_{j,\, j\neq i}^{M} \exp(\gamma(uz_j - uz_i))} \leq 1 - \frac{t}{t + (M-t)\exp(-\gamma\delta_{min})} = 1 - \frac{1}{1 + \frac{M-t}{t}\exp(-\gamma\delta_{min})}.$$

For z_j and $z_j \neq z_i$,

$$\frac{\exp(\gamma(uz_j - uz_i))}{1+\sum_{k,\, k\neq i}^{M} \exp(\gamma(uz_k - uz_i))} \leq \frac{1}{\exp(\gamma\delta_{min})} = \exp(-\gamma\delta_{min}).$$

Then, for the retrieval error, we can have,

$$\epsilon \leq \|\Delta z\| + \left(1 - \frac{1}{1 + \frac{M-t}{t}\exp(-\gamma\delta_{min})}\right)\|z_i\| + \exp(-\gamma\delta_{min}) \sum_{j,\, z_j \neq z_i}^{M} \|z_j\|.$$

Let $c := \exp(-\gamma\delta_{min})$ and $z_{max} = \max(z_1, ..., z_M)$. Then,

$$\epsilon \leq \|\Delta z\| + \left(1 - \frac{1}{1 + \frac{c(M-t)}{t}} + c(M-t)\right)\|z_{max}\|.$$

$$\text{Furthermore, } -\frac{1}{1 + \frac{c(M-t)}{t}} + c(M-t) \propto c\left(\frac{M}{t} + (M-t)\right),$$

$$\text{Therefore, } \epsilon \propto c\frac{M}{t}, \text{ given } \|\Delta z\| \text{ and } \|z_{max}\|.$$

Thus, we have proved the upper bound of the retrieval error. For the lower bound, if $u^\star$ is retrieved without loss, the error will be naturally zero.