Culturally transmitted color categories in LLMs reflect a learning bias toward efficient compression

Nathaniel Imel
New York University
n.imel@nyu.edu

Noga Zaslavsky
New York University
nogaz@nyu.edu

Abstract

Converging evidence suggests that systems of semantic categories across human languages achieve near-optimal compression via the Information Bottleneck (IB) complexity-accuracy principle. Large language models (LLMs) are not trained for this objective, which raises the question: are LLMs capable of evolving efficient human-like semantic systems? To address this question, we focus on the domain of color as a key testbed of cognitive theories of categorization and replicate with LLMs (Gemini 2.0-flash and Llama 3.3-70B-Instruct) two influential human behavioral studies. First, we conduct an English color-naming study, showing that Gemini aligns well with the naming patterns of native English speakers and achieves a significantly high IB-efficiency score, while Llama exhibits an efficient but lower complexity system compared to English. Second, to test whether LLMs simply mimic patterns in their training data or actually exhibit a human-like inductive bias toward IB-efficiency, we simulate cultural evolution of pseudo color-naming systems in LLMs via iterated in-context language learning. We find that akin to humans, LLMs iteratively restructure initially random systems towards greater IB-efficiency and increased alignment with patterns observed across the world's languages. These findings demonstrate that LLMs are capable of evolving perceptually grounded, human-like semantic systems, driven by the same fundamental principle that governs semantic efficiency across human languages.

1 Introducton

As large language models (LLMs) are increasingly deployed in real-world everyday settings, it is crucial to understand how their learning biases and representational capacities align with our own. We investigate this by focusing on a key aspect of human intelligence: the ability to organize information into semantic categories [1–6]. This phenomenon reveals two major challenges for AI. First, semantic categories exhibit both universal patterns and cross-language differences [7, 8, 1], which LLMs must navigate. Second, LLMs are not grounded in the rich physical and social environment that humans are, and it is unclear how these differences affect their ability to learn human-aligned semantic categories [9, 10]. Therefore, in order to understand whether LLMs can efficiently communicate with people and adapt to changing environments and communicative needs, it is crucial to study whether LLMs are capable of structuring meaning according to the same principles that guide humans. Here, we address this challenge and propose a novel cognitively-motivated framework for studying semantic systems in LLMs. We build on the theoretical framework of [11], which argues that languages efficiently compress meanings into words by optimizing the Information Bottleneck (IB) principle [12], instantiated as a tradeoff between the informational complexity and communicative accuracy of the lexicon. This framework has broad empirical support across languages [11, 13–16]. Furthermore, [17] recently showed that a drive for IB-efficiency may be present in the individual inductive biases of human learners. LLMs, however, are not trained with respect to the IB objective, which raises the question: are LLMs capable of evolving efficient human-like semantic systems?

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

To address this question, we focus on color, a key test case for categorization theories in cognitive science with rare cross-cultural human data [7, 18–20], and replicate with LLMs two influential human behavioral experiments: an English color naming experiment [21], designed to assess the efficiency and human-alignment of the color naming systems of LLMs; and an iterated language learning (ILL) experiment [22], designed to probe the implicit inductive biases of LLMs by simulating cultural transmission [23]. For the latter, we extend [24]'s iterated in-context learning (I-ICL) paradigm to iterated in-context *language* learning (IICLL). Because IICLL requires strong in-context capabilities that are too demanding for most open-weights models, we focus on two prominent LLMs that are capable of performing the task: Gemini 2.0-flash [25], a top proprietary multi-modal model, and Llama 3.3-70B-Instruct [26], a leading open-weights model.

We show that (1) Gemini aligns well with the English color naming system and achieves high IB-efficiency, while Llama, the smaller model, exhibits a highly efficient but lower complexity system; and (2) similar to humans, both LLMs iteratively restructure initially random systems towards greater IB-efficiency and increased human-alignment, suggesting that they do not simply imitate patterns in their training data but are intrinsically guided by a drive toward IB-efficiency. These findings demonstrate that LLMs are capable of evolving perceptually grounded, human-like semantic systems, driven at least in part by the same efficiency principle that underlies human languages.

2 Background

The Information Bottleneck (IB) color naming model. We apply the theoretical framework from [11], which models communication as a speaker mapping a mental representation (m) of a world state (u) to a word (w) via an encoder $q(w|m)$. In the case of color, the world states $\mathcal{U}$ are given by a set of target colors (Figure 1C). To account for perceptual noise, each m is modeled as a Gaussian distribution over the perceptual CIELAB color space centered around a target color [11]. To communicate efficiently, the speaker and listener optimize the Information Bottleneck (IB) tradeoff [12] between minimizing the complexity of their lexicon, $I_q(M;W)$, and maximizing its accuracy, $I_q(W;U)$. An optimal semantic system minimizes the IB objective:

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U), \tag{1}$$

where $\beta \geq 1$ controls the tradeoff, and the solutions to this optimization problem define the IB bound of efficiency. Figure 1F shows the IB bound for color naming from [11] together with a reproduction of their results showing that color naming systems across languages achieve near-optimal tradeoffs. The empirical basis for this analysis is the World Color Survey (WCS) dataset [27], containing naming data from 110 languages of non-industrialized societies elicited via the task illustrated in Figure 1A, with respect to the WCS grid shown in Figure 1C. Our analysis also includes English color naming data from [21] (Figure 1D).

Iterated language learning (ILL) of color naming systems. In iterated language learning (ILL), human participants form "generations" in which they learn a language from a sample of the previous generation's output and then produce a new language system (L_t) for the next. This process has been used to reveal how learners' inductive biases emerge from generalizing on limited data [28, 29]. We adapt this paradigm for LLMs, yielding iterated in-context language learning (IICLL) (see Figure 1B and Section 3.2). The empirical comparison for our IICLL study is the human ILL data from [22], in which participants learned and transmitted novel color naming systems across chains of generations. Each chain was initialized with a random system, and at each generation, participants were trained on a small set of color-pseudoword pairs before labeling the entire grid. Over time, chains became increasingly regular and more similar to systems in the WCS dataset [22]. More recently, [17] showed that these chains also converge to highly efficient systems along the IB bound (see Figure 1F).

3 Experiments

We aim to test whether LLMs are capable of acquiring a human-like inductive bias toward IB efficiency, even though they are not explicitly trained for this objective function. To this end, we implement with LLMs a color naming experiment and an IICLL experiment, designed to assess the color category systems and inductive learning biases implicit in LLMs. We then compare the LLM data with the corresponding human data through the lens of the IB framework. Specifically, we focus on three evaluation measures: (1) **efficiency loss**, measuring the LLM's deviation from IB-optimality,

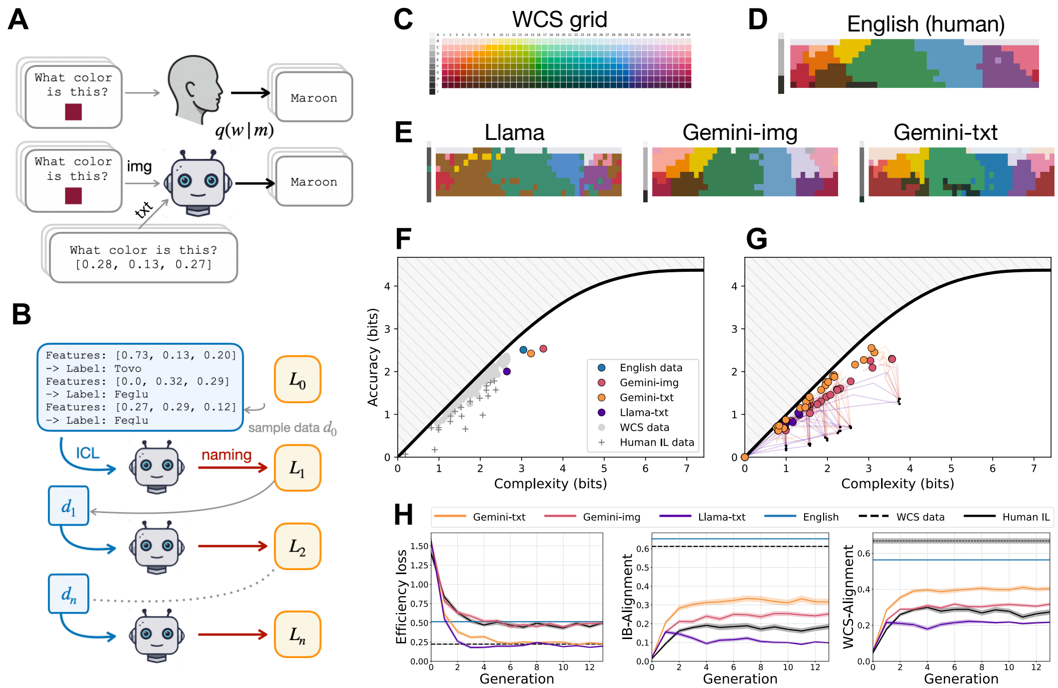

Figure 1: **A.** Color naming task with humans (top) and LLMs (bottom). Multi-modal LLMs can observe colors either via text or images. **B.** The IICLL paradigm. At each generation t, an LLM is prompted with a small dataset, d_{t-1}, consisting of colors with pseudo labels sampled from the previous generation's language, L_{t-1}. With d_{t-1} in context, the LLM performs the naming task for the full meaning space. **C.** The WCS color naming grid. **D-E.** Color naming systems of English speakers and LLMs. Each system is plotted against the WCS grid; each chip is colored by the color-centroid of its modal category. **F-G.** Humans and LLM on the information plane from [11], where the black curve is the IB theoretical bound. **F.** Gemini's English color naming system achieves a similar tradeoff as that of native English speakers, while Llama exhibits a lower complexity system. **G.** IICLL with LLMs converges to near-optimal solutions (colored dots) within the same range of human IL chains and WCS languages (shown in F). Small black dots correspond to random initialization of chains with varying number of categories, $k \in \{2, 3, 4, 5, 6, 14\}$. Thin lines correspond to IL trajectories. **H.** Over generations of IICLL, LLM systems become more efficient (left), more aligned to IB optima (middle), and more aligned to WCS languages (right). Colored curves represent averages across initializations and conditions, with the shaded region indicating the 95% CI.

i.e., with respect to Equation (1); (2) **IB-alignment**, measuring the structural similarity between an LLM system and its nearest IB-optimal system; and (3) **human-alignment**, measuring the structural similarity between an LLM system and human systems (from the WCS or English). Because our IICLL task requires strong in-context capabilities that are too demanding for most open-weights models we tested, we focus on two prominent LLMs that are capable of performing the task: Gemini 2.0-flash [25], a leading industry-level multi-modal model, and Llama 3.3-70B-Instruct [26], a leading open-weights text-based model.

3.1 English Color Naming with LLMs

We begin with a color naming study (Figure 1A) to assess the semantic alignment and communicative efficiency of LLMs with respect to native English speakers. Each LLM was prompted to label all 330 colors shown in Figure 1C, one at a time in a randomized order. These stimuli were presented either as sRGB coordinates (`txt` input), which both models can process, or as colored image patches (`img` input), which only Gemini can process. In both cases, the models were instructed to use the set of English terms from the free-naming experiment of [21].

Table 1: Evaluation of the efficiency loss (lower values are better) and alignment (higher values are better) of LLMs. Results show averaged over ten random seeds ($\pm$ STD).

Model	Efficiency Loss	IB-alignment	English-alignment
Llama-txt	**0.68 (0.03)**	0.45 (0.02)	0.55 (0.02)
Gemini-img	0.89 (0.03)	0.52 (0.01)	**0.69 (0.01)**
Gemini-txt	**0.79 (0.03)**	**0.53 (0.01)**	**0.65 (0.01)**

The resulting LLM color naming systems are shown in Figure 1E, their complexity-accuracy tradeoffs are shown in Figure 1F relative to the IB bound, and their efficiency and alignment scores are shown in Table 1. Gemini, using either `txt` or `img` inputs, demonstrates the highest alignment with the human English system and reaches human-like IB tradeoffs (Figure 1F). Llama produces a highly efficient system but with lower complexity compared to English. This suggests that both LLMs employ highly efficient color naming systems, but the model's size may impact its complexity range and in turn, its ability to align with humans. Finally, to ensure these results robustly reflect the LLM behavior, rather than brittle memorization, we also replicated the same results with added input noise which is imperceptible.

3.2 Iterated in-context language learning (IICLL)

To further understand the inductive learning biases of LLMs, beyond what they may have seen during training, we study how LLMs can evolve category systems on their own through a process of cultural evolution. We do so by replicating as closely as possible the ILL color naming experiment of [22], while adapting it to LLMs. To this end, we build on the iterated in-context learning (I-ICL) framework of [24] to develop a method which we call iterated in-context language learning (IICLL, Figure 1B), consisting of a chain of generations simulated by an LLM. In each generation, the LLM learns in-context an artificial color naming system produced by the previous generation (or randomly initialized for the first generation) from a small random sample of color-signal pairs. Importantly, the signals are not real words but rather pseudowords, designed to avoid a bias toward existing languages. Given this in-context learning phase, the model is then prompted to perform the full naming task (Figure 1A), generalizing the meanings of the pseudowords to the full color space of Figure 1C. This reveals the model's implicit inductive biases, which are then amplified with generations.

Figure 1G shows the IICLL trajectories of both models. Over generations, the LLMs approach the IB bound, becoming more efficient (Figure 1H, left) and more aligned with the IB-optimal systems (Figure 1H, middle), above and beyond a strong counterfactual baseline of hypothetical variants. Interestingly, Llama appears to be confined to a lower complexity region compared to human languages, while Gemini is capable of spanning the full complexity range observed in humans (comparing the WCS and IL data in Figure 1F with the LLM end-points in Figure 1G). Finally, Figure 1H shows that Gemini-txt is more efficient and more human-aligned compared to Gemini-img, although Gemini-img seems to track more closely the trends observed in the human iterated learning (IL) data. This suggests that the input representation may have an important role in how LLMs generalize semantic knowledge and categories. One possibility, for example, is that it might be easier for Gemini to process textual RGB inputs, leading it to more efficient solutions, while visual inputs leads it to more human-like behavior.

4 Conclusions

This work combines a theory-driven approach, based on the IB principle, with cognitively-motivated experimental methods, based on color naming and iterated language learning, to study whether LLMs can acquire a human-like inductive bias toward optimally-compressed semantic representations without being trained for this objective. We found that LLMs can align well with human color naming systems and tend to restructure randomly-initialized artificial category systems toward greater IB-efficiency and human-alignment with human naming systems. This demonstrates that LLMs are capable of evolving perceptually grounded, human-like semantic systems, guided by the same IB-efficiency principle that underlies human languages. Our findings also suggest that model size and input representations may be key factors in determining the level of complexity and human-alignment of LLMs. Finally, our work lays out the foundations for further evaluations of this capacity across more semantic domains and models.

5 Acknowledgments

This work was supported by a Google Cloud Platform Credit Award.

References

[1] William Croft. *Typology and Universals*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, 2 edition, 2002. ISBN 978-0-521-80884-2. doi: 10.1017/CBO9780511840579.

[2] Eleanor Rosch. Principles of categorization. In Daniel J. Levitin, editor, *Foundations of Cognitive Psychology: Core Readings*, pages 251–270. MIT Press, 2002.

[3] Peter Koch. 85. Lexical typology from a cognitive and linguistic point of view. In *2. Halbband Language Typology and Language Universals 2.Teilband*, pages 1142–1178. De Gruyter Mouton, 2008. ISBN 978-3-11-019426-5.

[4] James S Boster. Categories and cognitive anthropology. In *Handbook of Categorization in Cognitive Science*, pages 75–106. Elsevier, 2005.

[5] Asifa Majid. Comparing Lexicons Cross-linguistically. In John R Taylor, editor, *The Oxford Handbook of the Word*, page 0. Oxford University Press, 2015. ISBN 978-0-19-964160-4. doi: 10.1093/oxfordhb/9780199641604.013.020.

[6] Barbara C Malt and Asifa Majid. How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6):583–597, 2013.

[7] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969. ISBN 978-0-520-01442-8.

[8] Brent Berlin. *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*, volume 185. Princeton University Press, 1992.

[9] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975. ISSN 1939-2222. doi: 10.1037/0096-3445.104.3.192.

[10] William Labov. The boundaries of words and their meaning. In Charles-James N. Bailey and Roger W. Shuy, editors, *New Ways of Analyzing Variation in English*. Georgetown University Press, Washington, D.C., 1973.

[11] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31): 7937–7942, 2018. doi: 10.1073/pnas.1800521115.

[12] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottlneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[13] Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. Semantic categories of artifacts and animals reflect efficient coding. In *41st Annual Meeting of the Cognitive Science Society*, 2019.

[14] Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. Let's talk (efficiently) about us: Person systems achieve near-Optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2021.

[15] Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49):e2025993118, 2021. doi: 10.1073/pnas.2025993118.

[16] Noga Zaslavsky, Karee Garvin, Charles Kemp, Naftali Tishby, and Terry Regier. The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, page lzac001, 2022. ISSN 2058-458X. doi: 10.1093/jole/lzac001.

[17] Nathaniel Imel, Jennifer Culbertson, Simon Kirby, and Noga Zaslavsky. Iterated language learning is shaped by a drive for optimizing lossy compression. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*, 2025.

[18] Terry Regier, Paul Kay, and Naveen Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441, 2007. doi: 10.1073/pnas.0610341104.

[19] Terry Regier, Charles Kemp, and Paul Kay. Word Meanings across Languages Support Efficient Communication. In *The Handbook of Language Emergence*, chapter 11, pages 237–263. John Wiley & Sons, Ltd, 2015. ISBN 978-1-118-34613-6. doi: 10.1002/9781118346136.ch11.

[20] Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2019.02.003.

[21] Delwin T. Lindsey and Angela M. Brown. The color lexicon of American English. *Journal of Vision*, 14(2):17, 2014. ISSN 1534-7362. doi: 10.1167/14.2.17.

[22] Jing Xu, Mike Dowman, and Thomas L. Griffiths. Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280 (1758):20123073, 2013. doi: 10.1098/rspb.2012.3073.

[23] Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114, 2014. ISSN 0959-4388. doi: 10.1016/j.conb.2014.07.014.

[24] Jian-Qiao Zhu and Thomas L. Griffiths. Eliciting the Priors of Large Language Models using Iterated In-Context Learning, 2024.

[25] Google. Gemini 2.0: Flash, flash-lite and pro, 2025.

[26] AI@Meta. Llama 3 model card, 2024.

[27] Richard S. Cook, Paul Kay, and Terry Regier. The World Color Survey Database: History and use. In Henri Cohen and Claire Lefebvre, editors, *Handbook of Categorization in Cognitive Science*, pages 223–241. Elsevier Science Ltd, Oxford, 2005. ISBN 978-0-08-044612-7. doi: 10.1016/B978-008044612-7/50064-0.

[28] Thomas L. Griffiths and Michael L. Kalish. Language Evolution by Iterated Learning With Bayesian Agents. *Cognitive Science*, 31(3):441–480, 2007. ISSN 1551-6709. doi: 10.1080/15326900701326576.

[29] Thomas L. Griffiths, Brian R. Christian, and Michael L. Kalish. Using Category Structures to Test Iterated Learning as a Method for Identifying Inductive Biases. *Cognitive Science*, 32(1): 68–107, 2008. ISSN 1551-6709. doi: 10.1080/03640210701801974.