EVOLUTION AND COMPRESSION IN LLMS: ON THE EMERGENCE OF HUMAN-ALIGNED CATEGORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Converging evidence suggests that human systems of semantic categories achieve near-optimal compression via the Information Bottleneck (IB) complexity-accuracy tradeoff. Large language models (LLMs) are not trained for this objective, which raises the question: are LLMs capable of evolving efficient human-aligned semantic systems? To address this question, we focus on color categorization — a key testbed of cognitive theories of categorization with uniquely rich human data — and replicate with LLMs two influential human studies. First, we conduct an English color-naming study, showing that LLMs vary widely in their complexity and English-alignment, with larger instruction-tuned models achieving better alignment and IB-efficiency. Second, to test whether these 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 a method we refer to as Iterated in-Context Language Learning (IICLL). We find that akin to humans, LLMs iteratively restructure initially random systems towards greater IB-efficiency. However, only a model with strongest in-context capabilities (Gemini 2.0) is able to recapitulate the wide range of near-optimal IB-tradeoffs observed in humans, while other state-of-the-art models converge to low-complexity solutions. These findings demonstrate how human-aligned semantic categories can emerge in LLMs via the same fundamental principle that underlies semantic efficiency in humans.

1 INTRODUCTION

As large language models (LLMs) become increasingly popular in everyday use, it is crucial to understand how their learning biases and representational capacities align with our own. Here, we investigate this by focusing on a key aspect of human intelligence: the ability to organize information into semantic categories (Croft, 2002; Rosch, 2002; Koch, 2008; Boster, 2005; Majid, 2015; Malt & Majid, 2013). This phenomenon presents two major challenges for AI. First, systems of semantic categories (semantic systems, for short) exhibit both universal patterns and cross-language differences (Berlin & Kay, 1969; Berlin, 1992; Croft, 2002), which LLMs must navigate. Second, LLMs are not grounded in the rich physical and social environment that humans are (Rosch, 1975; Labov, 1973), and it is unclear how these differences affect their ability to learn human-aligned semantic categories. Therefore, in order to understand whether LLMs can efficiently communicate with people and adapt to changing environments and communicative needs in a human-like manner, it is crucial to study whether LLMs are capable of structuring meaning according to the same principles that guide humans.

To address this open gap in our understanding of LLMs, we propose a novel theoretical and cognitively-motivated framework for studying semantic systems in LLMs. We build on the framework of Zaslavsky et al. (2018), which argues that languages efficiently compress meanings into words by optimizing the Information Bottleneck (IB) principle (Tishby et al., 1999), instantiated as a tradeoff between the informational complexity and communicative accuracy of the lexicon. This framework has broad empirical support across human languages (Zaslavsky et al., 2018; 2019; 2021; Mollica et al., 2021; Zaslavsky et al., 2022). Furthermore, Imel et al. (2025) recently showed that a drive for IB-efficiency may be present in the individual inductive biases of human learners. While LLMs are trained on vast amounts of human language data encoded as text, they are not trained with respect to the IB objective. This raises the question: are LLMs capable of developing IB-efficient human-aligned semantic systems?

On the one hand, if LLMs fail to develop efficient semantic systems, it highlights a key misalignment in their capacity to learn human-like language. Furthermore, human semantic efficiency is fundamentally constrained by contingent, often perceptually grounded representations of meaning. Therefore, testing LLMs for efficiency also provides a test of whether they capture the specific representations humans use for categorization.

We address this open question with an in-depth analysis in the domain of color—a key test case for categorization theories in cognitive science with rarely available human data as well as practical implication for human-LLM interactions (see Section 2.1)—and replicate with LLMs two influential human behavioral experiments (Figure 1). First, we conduct an English color naming experiment (analogous to the human experiment of Lindsey & Brown, 2014), designed to assess the efficiency and human-alignment of the color naming systems of LLMs. Second, we conduct an iterated learning experiment of pseudo color-naming systems (analogous to the human experiment of Xu et al., 2013), designed to reveal implicit inductive learning biases of LLMs by simulating a process of cultural transmission (see Section 2.3). For the latter, we extend Zhu & Griffiths (2024)'s iterated in-context learning (I-ICL) paradigm to iterated in-context *language* learning (IICLL).

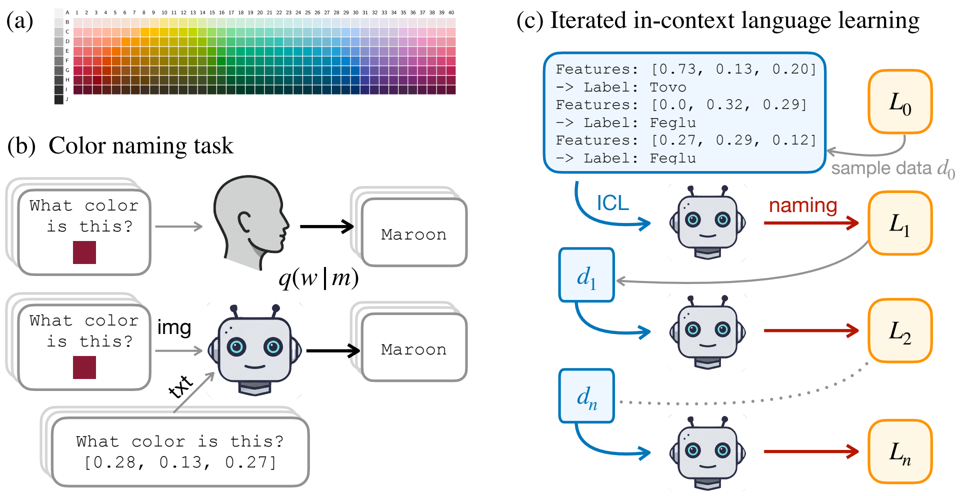
Our key findings and contributions are summarized as follows: First, we show that many prominent LLMs struggle to capture the English color naming system, exhibiting a wide range of complexities that are often lower than the complexity of English. However, with increased size and instruction-tuning, LLMs can achieve high English-alignment and IB efficiency. Second, using our IICLL paradigm, we show that LLMs that perform well in the naming task are not merely mimicking patterns in their training data but are actually guided by a human-like inductive bias toward IB-efficiency. Specifically, we show that LLMs iteratively restructure initially random artificial systems towards greater IB-efficiency and increased human-alignment. However, among the models we tested, only the model with strongest in-context capabilities (Gemini 2.0) is able to recapitulate the wide range of near-optimal IB-tradeoffs observed in humans, while other models converge to low-complexity solutions. Finally, we show that Gemini can also develop structured category systems via IICLL in a domain that is qualitatively different from color, suggesting that our result could potentially apply also in other domains. Taken together, these findings demonstrate how human-aligned semantic categories can emerge in LLMs via the same fundamental principle that underlies semantic efficiency in humans. Importantly, neither humans nor LLMs are explicitly trained for optimizing the IB objective, suggesting that IB-efficiency may emerge to support intelligent behavior.

2 Background and motivation

2.1 Why talk about color?

For decades, cognitive scientists have used color as an essential tool to study perception and categorization. This is in part due to the unprecedented amount of human behavioral data available for color, which includes research on perceptual space, cross-linguistic semantic variation, category learning, and cultural evolution. Especially relevant to our study is the World Color Survey (WCS) dataset (Cook et al., 2005), which contains color-naming data from 110 non-industrialized languages. Another relevant study by Xu et al. (2013) demonstrated that the cultural transmission of initially random, artificial color-term systems in humans leads to more regular systems that resemble those documented in the WCS dataset. To our knowledge, these two data resources— on actual cross-linguistic naming patterns and on the cultural evolution of category systems—are unique to the domain of color, making it an ideal domain for evaluating how well LLMs align with human behavior. Furthermore, studying color naming in LLMs also has practical implications. Generative AI models, used for tasks like image generation and online product searches, require grounded representations of color language. In order for such models to interact in ways we expect them to, it is crucial to determine the extent to which state-of-the-art LLMs have learned the meaning of color terms that actually align with human naming patterns.

While previous work has shown that human-aligned color representations can be recovered from LLMs (e.g., Abdou et al., 2021; Patel & Pavlick, 2022), there is limited research on model behavior in the context of real-world settings (namely, prompt-based interactions). Particularly relevant to our



```
Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu
```

Figure 1: (**a**) The standard WCS color naming grid (Kay et al., 2009). (**b**) Color naming task with humans and LLMs. Multi-modal LLMs can observe colors either via text or images. (**c**) Illustration of the IICLL paradigm. At each generation t, an LLM is prompted with a small dataset for ICL, d_{t-1}, consisting of pairs of colors and pseudo labels sampled from the previous generation's language, L_{t-1}. With these data in context, the LLM performs the naming task for the full space (a).

work is a study by Marjieh et al. (2024) that showed that several recent instruction-tuned models, namely GPT-3, its "ChatGPT" variants, GPT-3.5 and GPT-4 (Brown et al., 2020; OpenAI et al., 2024) and Mistral 7B Instruct (Jiang et al., 2023) can recover the English and Russian color naming systems by prompting the models to label hex codes. In contrast, we test a large set of 39 models with varying sizes and training stages, we consider both textual RGB inputs and image inputs (for multi-modal models), we analyze the LLMs' color naming systems through the lens of the IB framework, and we further assess their underlying inductive learning biases in a cultural evolution process via iterated learning. We provide relevant background on the IB and iterated learning frameworks below.

2.2 The Information Bottleneck and semantic systems

The IB framework for semantic systems (Zaslavsky et al., 2018), which we employ in this work, is based on the following communication model: a speaker wishes to communicate a mental representation, or belief state, $m \in \mathcal{M}$, defined as a probability distribution over world states $u \in \mathcal{U}$, by mapping it to a word $w \in \mathcal{W}$ via a stochastic encoder $q(w|m)$. A listener then receives w and interprets the speaker's intended meaning by constructing an estimator $\hat{m}_w$. In the case of color, for example, the world states $\mathcal{U}$ are given by a set of target colors (Figure 1a). To account for perceptual noise, it is assumed that each m is a Gaussian distribution over the perceptual CIELAB color space centered around a corresponding target color (Zaslavsky et al., 2018). That is, each color is mentally represented by the speaker as a Gaussian distribution and the listener's goal is to reconstruct the speaker's belief state over colors.

According to this framework, in order to communicate efficiently, the speaker and listener must jointly optimize the IB tradeoff between minimizing the complexity of their lexicon and maximizing its accuracy. Assuming the listener's inferences are adapted to the speaker, the lexicon is defined by an encoder $q(w|m)$. Complexity in IB roughly corresponds to the number of bits required for communication, formally defined by $I_q(M;W)$, the mutual information between speaker's meanings and words. Accuracy is defined by $I_q(W;U)$, the information that the speaker's words maintain about the target world state, which is also inversely related to the KL-divergence between the speaker's mental state and the listener's inferred state, $\mathbb{E}_q[D[M\|\hat{M}]]$ (Harremoes & Tishby, 2007; Zaslavsky, 2020). An optimal lexicon, or semantic system, is one that minimizes the IB objective function

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

where $\beta \geq 1$ controls the complexity-accuracy tradeoff. The solutions to this optimization problem define the IB theoretical limit of efficiency.

This framework generates precise quantitative predictions that have been gaining converging empirical support across hundreds of languages and multiple semantic domains, ranging from perceptually-grounded domains such as color (Zaslavsky et al., 2018; 2022) to higher-level conceptual domains such as household objects (Zaslavsky et al., 2019; Taliaferro et al., 2025) and personal pronouns (Zaslavsky et al., 2021). In addition, it has been successfully applied to studying emergent communication in artificial agents (Chaabouni et al., 2022; Tucker et al., 2022; Gualdoni et al., 2024; Tucker et al., 2025).

The IB color naming model. As part of our evaluation of LLM color naming, we use the previously published IB color naming model from Zaslavsky et al. (2018). The black curve in Figure 3 shows the IB bound for color naming from this study, together with a reproduction of their results showing that color naming systems across languages (from the WCS as well as English from Lindsey & Brown (2014)) achieve near-optimal IB tradeoffs.

2.3 Iterated language learning

The second main theoretical framework we apply in this work is based on iterated learning (IL). IL is a paradigm in cognitive science for simulating cultural transmission and eliciting prior inductive biases (Griffiths & Kalish, 2007; Kirby et al., 2008; Griffiths et al., 2008). In a typical IL experiment, participants form chains of "generations." At each generation t, a participant is exposed to data d_{t-1} from the previous generation and then produces responses that become input for the next generation. In iterated language learning (ILL), participants learn examples of pairs of stimuli and artificial labels during a training period, after which they assign labels to new, unlabeled stimuli in the same meaning space. In doing so, participants produce a full category system L_t, which is sampled from to provide training examples to the next generation, and the process repeats. This process, which requires generalization from limited data, reveals learners' inductive biases for certain linguistic or category structures (Griffiths & Kalish, 2007; Griffiths et al., 2008). As shown in Griffiths & Kalish (2007), under certain conditions, namely that the IL agents are Bayesian who share priors and likelihood functions, this Markov chain converges to a stationary distribution over languages equal to the learners' prior distribution $p(L)$. This makes the strong prediction that languages emerging from IL reflect the population's underlying inductive biases. Although this is an asymptotic characterization of IL dynamics, in behavioral experiments with people, researchers observe rapid convergence to heavily biased distributions of languages (Kirby et al., 2015).

Related work on ILL. Previous research in machine learning has investigated dynamics related to ILL. In agent-based simulations, Ren et al. (2020) introduced the neural iterated learning (NIL) framework, and showed how it can lead to compositional language in neural network agents, and Carlsson et al. (2024) showed that introducing communication-based training in NIL leads to IB-efficient color naming systems. These studies, however, did not explore LLMs as we do here. Zhu & Griffiths (2024) adapted IL to LLMs with strong in-context learning capacities as a prompt-based workflow known as Iterated In-Context Learning (I-ICL) to elicit LLMs' implicit prior distributions over aspects of world knowledge. I-ICL has recently been used to analyze cultural evolution in LLMs (Ren et al., 2024) and to compare visual and linguistic abstractions between LLMs and humans (Kumar et al., 2024). Here, we introduce iterated in-context *language* learning (IICLL, Figure 1c), which goes beyond prior work in leveraging the strong in-context learning abilities of LLMs to replicate as closely as possible the experimental conditions of ILL studies with humans, enabling a direct comparison to LLMs of their respective inductive biases, particularly with regards to semantic efficiency and alignment.

ILL of color naming systems. The empirical comparison for our IICLL color naming study is the IL data from Xu et al. (2013). In their experiment, participants were asked to learn and transmit novel systems of color terms across thirteen generations. We focus on their main results that include twenty iterated learning chains, each initialized with a random partition of the WCS grid. These chains vary in the number of allowed color terms, ranging from two to six, and four replications of each condition. Participants were shown a set of randomly selected colors generated uniquely for each chain, and paired with corresponding pseudo words. After training, participants were asked to label all 330 colors of the WCS grid. Xu et al. (2013) found that over time, the IL chains become increasingly regular and resembling the color naming systems documented in the WCS dataset. More recently, Imel et al. (2025) found that these chains also converge to highly efficient systems along the IB bound. Figure 3 shows our reproduction of this finding (plotting only final generations).

3 EXPERIMENTAL SETUP

Our goal is to test whether LLMs have an inductive bias toward IB-efficiency, as observed in humans. To this end, we conduct two studies with LLMs: (1) an English color naming study to assess their semantic alignment and communicative efficiency with respect to English speakers; and (2) a cultural transmission experiment of artificial color naming systems (using IICLL) to elicit their inductive learning biases beyond patterns they may have seen during training.

Models. We consider 39 models across 6 model families: Gemini (Google, 2025), Gemma 3 (Gemma-Team et al., 2025), Llama 3 (Grattafiori et al., 2024), Qwen 2.5 (Qwen et al., 2025), Olmo 2 (OLMo-Team et al., 2025) and GPT-2 (Radford et al., 2019). Within each family, we vary models along several dimensions. Specifically, to gain insight into the properties that may influence the models' behavior in our tasks, we consider models with different sizes, instruction-tuned versus base models, and text-based versus multi-modal models. For Olmo, we also considered its learning dynamics by analyzing training checkpoints. For a full list of models, see Table 1 in Appendix D.

Prompts. In both of our studies, we provided instructions in the prompts to choose only from a fixed set of terms. The Gemini API supports controlled generation which makes this constrained classification task straightforward; for all open-weight models, we used log probability based scoring of the allowed terms as a continuation of the prompt. Further details and example prompts can be found in Appendix J.

Stimuli. We used the 330 color chips from the WCS grid shown in Figure 1a. These chips represent a systematic sampling of color space and are standard stimuli in color naming research. Each chip is associated with precise coordinates in the CIELAB color space, which can be converted to sRGB coordinates. To present the color stimuli to the text-based LLMs, we encoded color using these numerical coordinates. For multimodal models, we generated a square colored image corresponding to the WCS chip's sRGB values, and passed this image together with the text instructions.

Evaluation. Following Zaslavsky et al. (2018), we use two main evaluation measures in our studies. First, the ***efficiency loss*** of a semantic system is measured by its minimum deviation from optimality, defined as $\varepsilon = \min_\beta\{\frac{1}{\beta}(\mathcal{F}_\beta[q] - \mathcal{F}^*_\beta)\}$ where $\mathcal{F}^*_\beta$ is the optimal value of $\mathcal{F}_\beta$ in Eq. 1. Second, we measure the ***semantic (mis)alignment*** between two systems by the Normalized Information Distance (NID) (Kraskov et al., 2005; Vinh et al., 2010). NID is a metric capturing the distance between two clusterings (in this case, induced by color naming categories), providing a quantitative measure of structural similarity between the naming systems. Since NID is bounded in $[0, 1]$, we take 1 - NID as a measure of similarity, or alignment. ***IB-alignment*** measures the similarity between a system and the nearest (ϵ-fitted) optimal IB system. ***WCS-alignment*** measures the average alignment between a system and the WCS languages, and ***English-alignment*** measures the alignment between a system and English.

4 RESULTS

4.1 ENGLISH COLOR NAMING

We begin by eliciting color naming responses from LLMs with English color terms. We then evaluate their IB-efficiency and alignment with the actual English color naming system from Lindsey & Brown (2014). As illustrated in Figure 1a, we consider two variants of this task, a text-only variant where colors are presented as sRGB coordinates, and an image-based variant where colors are presented as a color patch. More details on our procedure can be found in Appendix B.

The resulting systems are shown in Figure 2b, their IB tradeoffs are shown in Figure 2a, and their quantitative evaluation is shown in Figure 2c (see also Figure 7, 8 and 9 in Appendix E). We

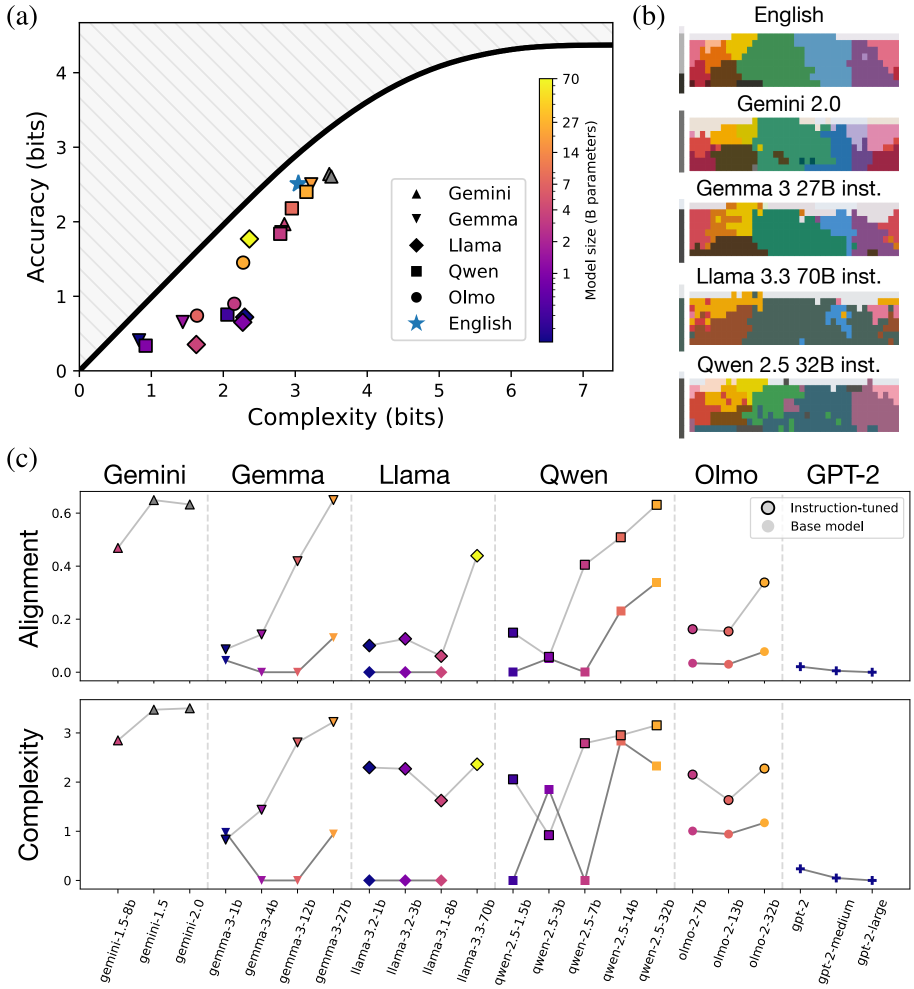


Figure 2: **English color naming experiment with LLMs. (a)** IB complexity-accuracy tradeoffs achieved by instruction-tuned LLMs (see Appendix section E for all/base models), plotted with respect to the English tradeoff (blue star) and IB theoretical bound (black curve) from Zaslavsky et al. (2018). Models vary widely in their tradeoffs, with larger instruction-tuned models reaching the English point. **(b)** Color naming systems of English (from Lindsey & Brown (2014)) and best-performing LLMs. Each system is shown by its mode map, i.e., it is plotted against the WCS grid (Figure 1a), where each chip is colored by the color-centroid of its modal category. **(c)** English-alignment (top) and IB complexity (bottom) of all LLMs. Markers are the same as in (a), where a black edge indicates the instruction-tuned model and no edge indicates the base model. Across model families, size and instruction-tuning are associated with higher complexity and better alignment to English.

found that the vast majority of LLMs vary in their complexity and English-alignment, with larger instruction-tuned models achieving better alignment and IB-efficiency. No model aligns perfectly with the English system from Lindsey & Brown (2014), though Gemini-2.0 and Gemma 3 27B (inst.) approximate it closely. Llama 3.3 70B (inst.) and Qwen 2.5 32B (inst.) produce naming systems that are reasonably aligned with English and achieve comparable IB-tradeoffs. We also conducted an analysis on the learning trajectory of Olmo 2 32B (Appendix F), which revealed that English-alignment and complexity increase as a function of learning steps and tokens seen during training. Interestingly, an minimal pair analysis of the multimodal LLMs (Figure 8, Appendix E) revealed that there may be a cutoff of roughly 3 bits of complexity for which processing images rather than text (sRGB coordinates) does not improve alignment.

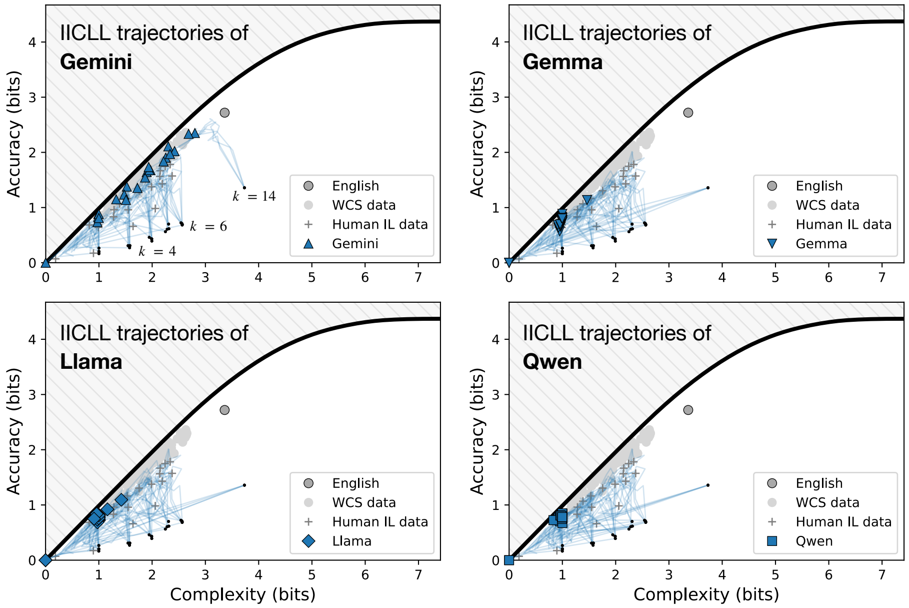


Figure 3: **IICLL with LLMs converges to near-optimal IB solutions.** The trajectories of Gemini 2.0 Flash (upper left), Gemma 3 27B (upper right), Llama 3.3 70B (lower left) and Qwen 2.5 32B (lower right) are plotted on the information plane (same as Figure 2A), together with the IB tradeoffs across human languages (WCS+English) and human IL data. Small black dots correspond to random initializations of chains with varying number of categories, $k \in \{2, 3, 4, 5, 6, 14\}$. Thin blue lines correspond to the LLMs' IICLL trajectories. Gemini captures the complexity range observed across human languages, while the other models converge to lower complexity systems. All models are instruction-tuned.

These findings are somewhat surprising, given that prior work (Abdou et al., 2021; Patel & Pavlick, 2022; Marjieh et al., 2024) has shown that language models of a variety of sizes can in principle recover human-aligned representations of color terms. In contrast, we find that many state-of-the-art pretrained LLMs struggle to reproduce any coherent color vocabulary, when presented with colors (encoded as sRGB coordinates) in a constrained naming task in the spirit of the WCS procedure. Although instruction tuning is associated with better performance, it does not guarantee comparable IB-tradeoffs or alignment to English, even for models on the order of ∼15B parameters. On the other hand, we were surprised to find that some models–particularly Olmo 2 32B (inst.) (OLMo-Team et al., 2025) and Qwen 7 2.5 VL (inst.) (Bai et al., 2025) – produced systems with category structure resembling not English, but instead other, very low-resource languages from the WCS (see Figure 9 in Appendix E). This suggests that although many LLMs struggle to recover the same particular distinctions as English speakers, they may still possess a representation of color aligned with that of humans.

Taken together, these findings indicate that while color categorization in LLMs is limited to very recent, large, instruction-tuned models, from these models there is a clear capacity to develop color categories aligned with human color perception. However, this English color naming task still leaves open a crucial question: is this behavior merely a reflection of imitating patterns in human data, or does it signify a more intrinsic LLM inductive bias towards IB-efficiency in categorization?

4.2 Iterated in-context language learning (IICLL)

To investigate whether LLMs possess an efficiency bias that extends beyond learning the specific categories of a language on which they were trained, we turn to our second study, which simulates cultural transmission in LLMs. To this end, we sought to replicate the ILL color naming experiment of Xu et al. (2013) as closely as possible, using our IICLL paradign shown in Figure 1b (see Appendix G for more details). We considered only large, instruction tuned models that performed well

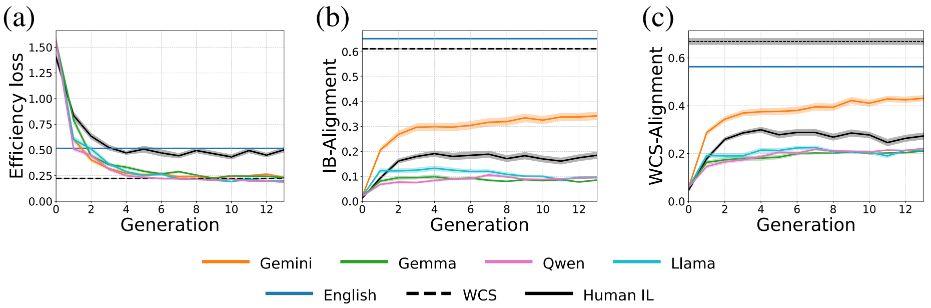


Figure 4: Across IICLL generations, emergent LLM systems become more efficient (a), more aligned with the optimal IB systems (b), and more aligned with human languages (c). Colored curves show the average across initializations and conditions, and the colored regions corresponds to the 95% confidence intervals.

in the English color naming task for our IICLL experiments: Gemini 2.0, Gemma 3 27B, Qwen 2.5 32B, and Llama 3.3 70B – henceforth, Gemini, Gemma, Qwen and Llama (see Appendix L for an analysis showing that smaller models struggle in IICLL).

Figure 3 shows the resulting trajectories of IICLL chains on the information plane. Gemini develops color naming systems that converge to a similar range of near-optimal IB solutions as the typological patterns of the WCS languages, as well as the final generation systems from the human IL chains from Xu et al. (2013). There is also broad qualitative fit between its final generation IICLL systems and languages from the WCS (see Appendix I for a comparison of mode maps). The other LLMs also develop systems that converge to highly IB-efficient solutions, though they are generally limited to the lower range of complexity observed in the WCS.

Additional quantitative support for these observations is provided in Figure 4. These figures show that over generations, LLM systems become (i) more efficient in their mapping of stimuli to terms, (ii) more similar to naturally occurring human color systems and (iii) more similar to optimal IB systems. Furthermore, LLM IICLL chains converge near the bound relatively quickly (after roughly four generations), parallel to human IL dynamics.

Strikingly, many trajectories from all models initially climb in complexity towards the IB bound before slowly evolving downwards alongside it. This suggests that the capacity to learn and transmit complex yet near-optimally efficient category systems, while strongest for Gemini, is present in all four LLMs. One factor that may drive the difference between Gemini and the other models is that the IICLL task requires very strong in-context learning, as models must integrate dozens of in-context training examples to generalize well. For example, the $k = 14$ condition includes 84 examples, and in this setting most of the LLMs immediately converge to low-complexity solutions. However, given that (1) the terms in our IICLL experiment are nonsense (made up) terms and (2) we give no indication to the model that the stimuli are in fact colors, only that they have "features" (Figure 1c), all four models show an impressive ability to evolve systems (at least transiently) that are aligned with the structure of human color perception.

In order to precisely assess whether the emergent LLM systems were nontrivially efficient and aligned to WCS languages, we conducted a rotation analysis (Regier et al., 2007) on the final, evolved color systems from our LLM experiments alongside the human data from Xu et al. (2013) (Figure 11 in Appendix H). This analysis involved rotating the color term assignments along the hue dimension of the WCS color grid and evaluating the difference in their IB-efficiency and alignment scores. We found that such rotations away from the actual emergent systems lead to a significant decrease in efficiency and alignment scores for Gemini systems. This suggests that Gemini achieves efficient categorization that is not arbitrary, but is constrained by the specific, perceptually grounded organization of human color categories.

Taken together, our IICLL experiments show that LLMs, like human learners, can in principle restructure initially random color naming systems in a process of simulated cultural evolution that leads to near-optimality and high alignment with human semantic representations.

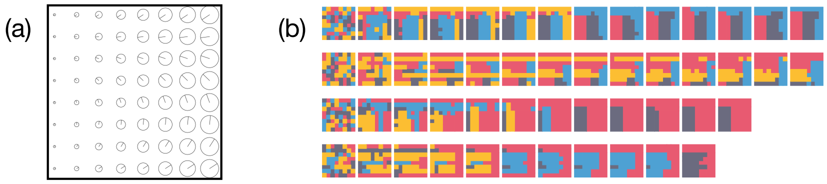


Figure 5: (a) The Shepard circles stimulus grid, (b) Gemini IICLL chains for naming Shepard circles. Rows correspond to individual chains, initialized randomly. Each system is plotted over the stimulus grid, where colors correspond to unique labels.

4.3 The emergence of categories in Shepard circles via IICLL

While many results in color have previously been shown to generalize to other domains (Zaslavsky et al., 2019; 2021; Mollica et al., 2021), it is far from trivial to conduct a full-scale analysis of the scope we performed for color naming, primarily due to the lack of high-quality data from both native speakers and ILL experiments. Here, nevertheless, we apply the IICLL paradigm in a qualitatively distinct semantic domain to provide initial evidence that LLMs may have a general bias to learn structured category systems.

We considered a synthetic domain of so-called "Shepard circles," a classic conceptual space used to study how humans categorize multidimensional stimuli (Shepard, 1964). These are circles which vary in both radius and angle of rotation of an internal spoke. We generated stimuli by taking eight evenly spaced values for each of the two dimensions, yielding 64 total stimuli (Figure 5a). For this preliminary investigation, we limit our analysis to Gemini. We found that presenting these stimuli as pairs of numbers (radius and angle) proved to be challenging. This is perhaps unsurprising, because unlike color– which is represented in various ways online in both text and images– there is likely no text online that would allow the model to associate these numbers with meaningful perceptual features. To overcome this limitation and better mimic human perception, we presented Gemini with images of the stimuli.

The results of a sample of IICLL chains are depicted in Figure 5b. Over generations, Gemini transmitted categories that became increasingly compact in their partitioning of the conceptual space, and distinguished regions based on both the radius and angle of the circles. This suggests that LLMs– especially large multimodal models– potentially have a domain-general bias to organize perceptual features into non-arbitrary, and increasingly regular, semantic categories.

5 Discussion

In this work, we combined 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. To do this, we first conducted an in-depth analysis of English color naming across 39 LLMs, and found that a surprising number of state-of-the-art models fail to capture the English color naming system. However, some of the most recent, larger, instruction-tuned models achieve high English-alignment and comparable IB tradeoffs. We then demonstrated that LLMs that do align well with the English color naming system are not merely mimicking patterns in their training data, but rather exhibit a more fundamental capacity to learn human-aligned, efficient color category systems. To do this, we introduced Iterated in-Context Language Learning (IICLL) to simulate cultural transmission of category systems. Over generations of IICLL, LLMs tend to restructure randomly-initialized artificial category systems toward greater IB-efficiency and alignment to human naming systems. We also provide initial evidence that LLMs can develop structured categories over generations of IICLL in a domain distinct from color, suggesting that our results may apply in other semantic domains. Taken together, our current findings suggest that LLMs are capable of evolving perceptually grounded, human-like semantic systems, guided by the same IB-efficiency principle that underlies human languages. Importantly, neither humans nor LLMs are explicitly trained for optimizing the IB objective, suggesting that IB-efficiency may emerge to support intelligent behavior.

Our empirical results open up several important questions for future research. First, while our work demonstrates that cultural transmission alone (via IICLL) can be a sufficient pressure for some LLMs to develop efficient, human-like category systems, a more complete understanding of language evolution requires integrating functional pressure of language use, e.g., via communication. Therefore, an important future direction is to extend our IICLL framework to incorporate communication as a selective pressure, for example, by adopting models that explicitly integrate both transmission and communication (e.g., Kouwenhoven et al., 2024). Second, the precise origins of the bias we observe in LLMs toward efficiency are unclear (for example, how might this bias emerge from properties of the training data, instruction-tuning, or model size), and investigating this is another important direction for future work. Lastly, we look forward to extending our analyses across more languages and semantic domains.

Reproducibility statement. This work utilizes a previously published model and code from Zaslavsky et al. (2018), made available at `https://github.com/nogazs/ib-color-naming` under the MIT License. The ILL data from Xu et al. (2013) were obtained from the authors, and the WCS data are publicly available from the World Color Survey website. Appendix D provides a list of the specific models used with their Hugging Face or Google API IDs, and Appendix J includes example prompts. We will make the code for our experiments and analyses available to reviewers in an anonymous repository.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025.

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

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

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

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, 2020.

Emil Carlsson, Devdatt Dubhashi, and Terry Regier. Cultural evolution via iterated learning and communication explains efficient color naming systems. *Journal of Language Evolution*, 9(1-2): 49–66, 2024. ISSN 2058-458X. doi: 10.1093/jole/lzae010.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 978-0-471-24195-9.

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.

Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report, 2025.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, 2024.

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.

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.

Eleonora Gualdoni, Mycal Tucker, Roger Levy, and Noga Zaslavsky. Bridging semantics and pragmatics in information-theoretic emergent communication. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21059–21078. Curran Associates, Inc., 2024.

Peter Harremoes and Naftali Tishby. The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. In *2007 IEEE International Symposium on Information Theory*, pp. 566–570, 2007. doi: 10.1109/ISIT.2007.4557285.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023.

Paul Kay, Brent Berlin, Luisa Maffi, William R. Merrifield, and Richard Cook. *The World Color Survey*. Center for the Study of Language and Information, Stanford, 2009.

Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, 2008. doi: 10.1073/pnas.0707835105.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015. ISSN 0010-0277. doi: 10.1016/j.cognition.2015.03.016.

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

Tom Kouwenhoven, Max Peeperkorn, and Tessa Verhoef. Searching for Structure: Investigating Emergent Communication with Large Language Models, 2024.

A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger. Hierarchical clustering using mutual information. *Europhysics Letters*, 70(2):278, 2005. doi: 10.1209/epl/i2004-10483-y.

Sreejan Kumar, Raja Marjieh, Byron Zhang, Declan Campbell, Michael Y Hu, Umang Bhatt, Brenden Lake, and Tom Griffiths. Comparing abstraction in humans and machines using multimodal serial reproduction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

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

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.

Asifa Majid. Comparing Lexicons Cross-linguistically. In John R Taylor (ed.), *The Oxford Handbook of the Word*, pp. 0. Oxford University Press, 2015. ISBN 978-0-19-964160-4. doi: 10.1093/oxfordhb/9780199641604.013.020.

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

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-72071-1.

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.

OLMo-Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, 2024.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

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.

Yi Ren, Shijie Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J. Sutherland. Bias amplification in language model evolution: An iterated learning perspective. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 38629–38664. Curran Associates, Inc., 2024.

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.

Eleanor Rosch. Principles of categorization. In Daniel J. Levitin (ed.), *Foundations of Cognitive Psychology: Core Readings*, pp. 251–270. MIT Press, 2002.

Roger N. Shepard. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1(1):54–87, 1964. ISSN 0022-2496. doi: 10.1016/0022-2496(64)90017-3.

Maya Taliaferro, Nathaniel Imel, Esti Blanco-Elorietta, and Noga Zaslavsky. Bilinguals exhibit semantic convergence while maintaining near-optimal efficiency. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*, 2025.

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

Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Mycal Tucker, Julie Shah, Roger Levy, and Noga Zaslavsky. Towards Human-Like Emergent Communication via Utility, Informativeness, and Complexity. *Open Mind*, 9:418–451, 2025. ISSN 2470-2986. doi: 10.1162/opmi_a_00188.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. ISSN 1533-7928.

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.

Noga Zaslavsky. *Information-Theoretic Principles in the Evolution of Semantic Systems*. PhD thesis, The Hebrew University of Jerusalem, 2020.

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.

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.

Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Communicative need in colour naming. *Cognitive Neuropsychology*, 37(5-6):312–324, 2020. ISSN 0264-3294. doi: 10.1080/02643294.2019.1604502.

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.

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*, pp. lzac001, 2022. ISSN 2058-458X. doi: 10.1093/jole/lzac001.

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

A IB Communication Model and Theoretical Bound

IB communication model The framework is based on a basic communication model that considers an inventory of words $\mathcal{W}$ to communicate about a space of meanings, $\mathcal{M}$. The specific model we apply in this paper is the previously published IB color naming model of Zaslavsky et al.. In this model, meanings are assumed to be mental representations or beliefs over world states $\mathcal{U}$, formally defined as a probability distribution $m(u)$ over world states $u \in \mathcal{U}$. In our setting, the set of world states $\mathcal{U}$ is taken to be the 330 color chips from the WCS grid, and meanings over colors are grounded in the CIELAB perceptual space such that each target color referent, $u_t \in \mathcal{U}$, is represented as a Gaussian distribution $m_t(u)$ centered around u_t. Meanings are drawn from an information source, $p(m)$, which characterizes how often each meaning needs to be communicated. The need distribution over meanings in this model, $p(m)$, was estimated using the method of least-informative priors (see Zaslavsky et al. (2020) for an extensive evaluation of this need distribution). Note that, in our IICLL experiments, we sample stimulus-word pairs uniformly for training data for each generation, instead of biased sampling from this prior. This was done in order to replicate the procedure in Xu et al. (2013).

Given a meaning $m \sim p(m)$, a speaker produces a signal w using a stochastic production policy, also called an encoder, $q(w|m)$, and then a listener interprets the signal by reconstructing an estimated belief state $\hat{m}_w(u) = \sum_m q(m|w)m(u)$. Note that this form of listener interpretations corresponds to a Bayesian listener. While this form is assumed here for simplicity, it is not an actual assumption but rather a derivation from the theory (see the SI of Zaslavsky et al. (2018)).

IB theoretical bound A semantic category system in this framework corresponds to a stochastic encoder, $q(w|m)$, which maps meanings to signals. An optimal semantic system, according to the IB principle, is an encoder that satisfies a tradeoff between its informational complexity and communicative accuracy. Complexity, also known as information rate (Cover & Thomas, 2006), is defined by the mutual information between speaker meanings and signals,

$$I_q(M;W) = \sum_m p(m)q(w|m) \log \frac{q(w|m)}{q(w)}, \tag{2}$$

which captures the number of bits, on average, that is required to encode meanings with signals. Maintaining low complexity corresponds to using fewer bits for communication, i.e., achieving high compression rate, which in turn, can be translated to affording a smaller lexicon size. For example, minimal complexity can be achieved by compressing all possible meanings into a single signal. This, however, will result in very poor accuracy. Accuracy, in IB terms, is quantified by

$$I_q(W;U) = I(M;U) - \mathbb{E}_q \left[D \left[M \| \hat{M} \right] \right], \tag{3}$$

which corresponds to maintaining a language that is informative about the speaker's intentions. Maximizing accuracy, as expressed in Eq. 3, amounts to minimizing the expected distortion between speaker and listener meanings, i.e., the Kullback-Leibler (KL) divergence $D[M\|\hat{M}]$. We say that a language's semantic system is *efficient* to the extent that its encoder q minimizes the IB tradeoff:

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U) \;\text{ s.t. }\; \beta \geq 1\,, \tag{4}$$

where β is a free parameter controlling the tradeoff between pressure to minimize complexity and pressure to maximize accuracy. The solutions to this optimization problem define the IB theoretical limit of efficiency, which means that no system can lie above this theoretical bound (see Fig. 3).

B English color naming procedure

The color naming task was conducted by prompting a model to label a color chip, for all 330 color chips from the WCS stimulus array, one stimulus at a time, in a randomized order. Unlike the prompts in our IICLL study, we do not provide the model with access to previous interactions. Each prompt was of the form: "What color is this [`<color coordinates>`]? You may only use one of the following allowed labels: [‘Red’, ‘Blue’, . . .] Please provide only a single label from the list just provided. Do not give any explanation." When prompting models with text only, the color coordinates were sRGB triples. The allowed set of labels were fourteen basic English color terms corresponding to the modal terms from Lindsey & Brown (2014).

C English color naming with CIELAB inputs

To address the generalizability of our findings beyond sRGB inputs, we replicated our English color naming task using CIELAB triples as the input features for the prompt text. CIELAB is a perceptually uniform space and so provides a principled color space to contrast with sRGB. We restricted to testing the four models that performed best from our original study (which used sRGB-encoded stimuli).

Figure 6 illustrates mode maps of the resulting systems for models when prompted with CIELAB-encoded colors. We found that performance generally degraded across all models when providing features as CIELAB triples: they struggled significantly to align with established English color boundaries, and the resulting category systems were considerably noisier compared to the sRGB-based outputs. This outcome is consistent with previous research by Marjieh et al. (2024) which noted that other frontier models failed to produce coherent English color category systems when prompted with CIELAB triples. Based on this observation, we restricted our IICLL study to the sRGB feature space to assess the models' inductive bias under ideal conditions.

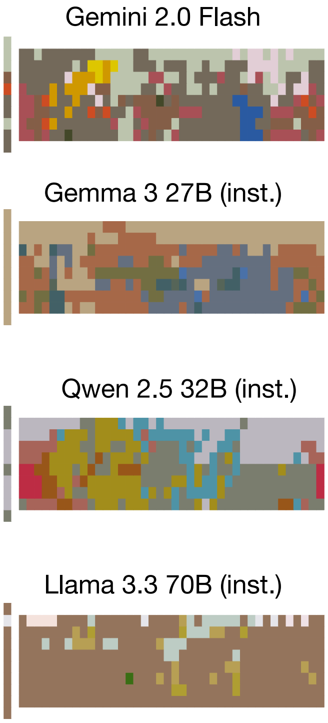


Figure 6: Mode maps for the best-performing models from our English color naming study, when prompted with CIELAB triples.

D MODELS

Model	Size (B params)	Instruction Tuned	Multimodal	Hugging Face / Google API ID
gemini-1.5	n/a	Yes	Yes	gemini-1.5-flash
gemini-1.5-8b	8	Yes	Yes	gemini-1.5-flash-8b
gemini-2.0	n/a	Yes	Yes	gemini-2.0-flash
gemma-3-1b	1	No	No	google/gemma-3-1b
gemma-3-1b-it	1	Yes	No	google/gemma-3-1b-it
gemma-3-4b	4	No	Yes	google/gemma-3-4b
gemma-3-4b-it	4	Yes	Yes	google/gemma-3-4b-it
gemma-3-12b	12	No	Yes	google/gemma-3-12b
gemma-3-12b-it	12	Yes	Yes	google/gemma-3-12b-it
gemma-3-27b	27	No	Yes	google/gemma-3-27b
gemma-3-27b-it	27	Yes	Yes	google/gemma-3-27b-it
gpt-2	0.224	No	No	openai-community/gpt2
gpt-2-medium	0.355	No	No	openai-community/gpt2-medium
gpt-2-large	0.774	No	No	openai-community/gpt2-large
llama-3.1-8b	8	No	No	meta-llama/Llama-3.1-8B
llama-3.1-8b-instruct	8	Yes	No	meta-llama/Llama-3.1-8B-Instruct
llama-3.2-1b	1	No	No	meta-llama/Llama-3.2-1B
llama-3.2-1b-instruct	1	Yes	No	meta-llama/Llama-3.2-1B-Instruct
llama-3.2-3b	3	No	No	meta-llama/Llama-3.2-3B
llama-3.2-3b-instruct	3	Yes	No	meta-llama/Llama-3.2-3B-Instruct
llama-3.3-70b-instruct	70	Yes	No	metallama/Llama-3.3-70B-Instruct
olmo-2-7b	7	No	No	allenai/OLMo-2-7B
olmo-2-7b-instruct	7	Yes	No	allenai/OLMo-2-7B-Instruct
olmo-2-13b	13	No	No	allenai/OLMo-2-13B
olmo-2-13b-instruct	13	Yes	No	allenai/OLMo-2-13B-Instruct
olmo-2-32b	32	No	No	allenai/OLMo-2-32B
olmo-2-32b-instruct	32	Yes	No	allenai/OLMo-2-32B-Instruct
qwen-2.5-1.5b	2	No	No	Qwen/Qwen-2.5-1.5B
qwen-2.5-1.5b-instruct	2	Yes	No	Qwen/Qwen-2.5-1.5B-Instruct
qwen-2.5-3b	3	No	No	Qwen/Qwen-2.5-3B
qwen-2.5-3b-instruct	3	Yes	No	Qwen/Qwen-2.5-3B-Instruct
qwen-2.5-7b	7	No	No	Qwen/Qwen-2.5-7B
qwen-2.5-7b-instruct	7	Yes	No	Qwen/Qwen-2.5-7B-Instruct
qwen-2.5-vl-7b-instruct	7	Yes	Yes	Qwen/Qwen-2.5-VL-7B-Instruct
qwen-2.5-vl-32b-instruct	32	Yes	Yes	Qwen/Qwen-2.5-VL-32B-Instruct
qwen-2.5-14b	14	No	No	Qwen/Qwen-2.5-14B
qwen-2.5-14b-instruct	14	Yes	No	Qwen/Qwen-2.5-14B-Instruct
qwen-2.5-32b	32	No	No	Qwen/Qwen-2.5-32B
qwen-2.5-32b-instruct	32	Yes	No	Qwen/Qwen-2.5-32B-Instruct

Table 1: Full list of models used in our naming study. Models used in our IICLL task are bolded.

E Naming systems for all models

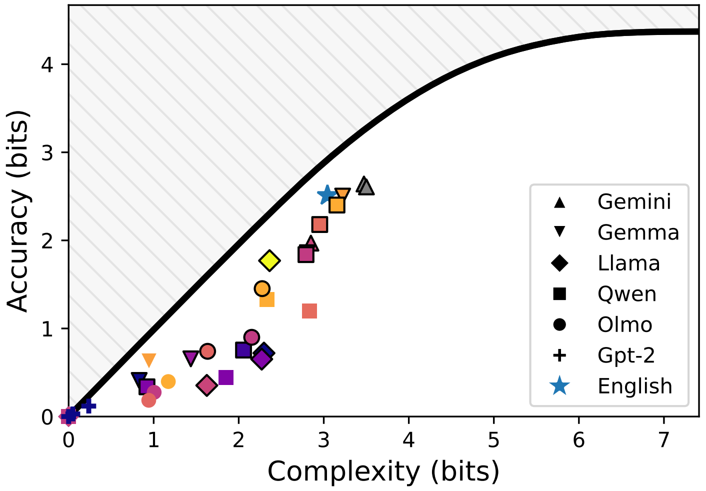


Figure 7: IB complexity-accuracy tradeoffs achieved by text-based LLMs. Same as Figure 2a but including base (not instruction-tuned) models.

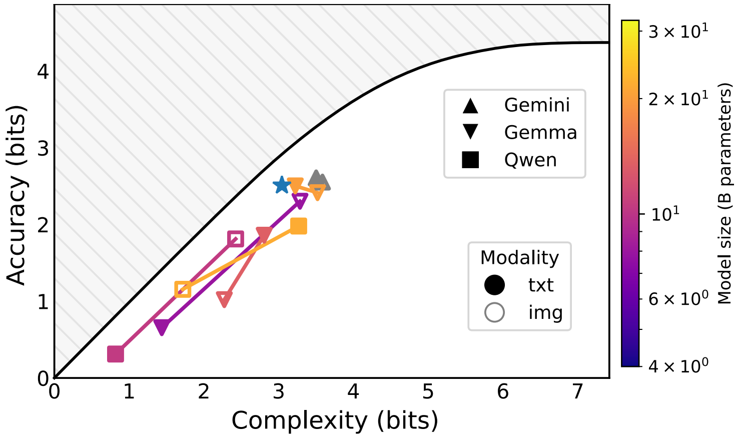


Figure 8: IB complexity-accuracy tradeoffs achieved by instruction-tuned multimodal LLMs. Points without color filling indicate systems resulting from providing colors as images, rather than via sRGB coordinates in text. Each minimal pair is connected by a line. Interestingly, there appears to be a cutoff around 3 bits of complexity for which models that already perform well with text-based prompting do not improve alignment to English when prompted with an image instead.

As can be seen in Figure 9, many of the models performed poorly on the English naming task, with many base (not instruction-tuned) models and some smaller instruction-tuned models failing to produce naming systems with any coherent category structure. Figure 7 shows that these systems are distributed widely across the information plane, with only larger, instruction-tuned, very recent models achieving similar IB-efficiency tradeoffs to English speakers. Interestingly, some models show similar category structure not to English, but to languages in the WCS. For example, Olmo-2-32B-Instruct, and Qwen-2.5-VL-Instruct (when presented with images) bear resemblance to Mayoruna (see the final row of Figure 13), as well as some of the systems to which Gemini-2.0 converges over generations of IICLL. This suggests that although some larger, instruction tuned models and multimodal instruction-tuned models fail to recover the English naming system, they may still possess some general bias towards color category structure that is aligned with humans.

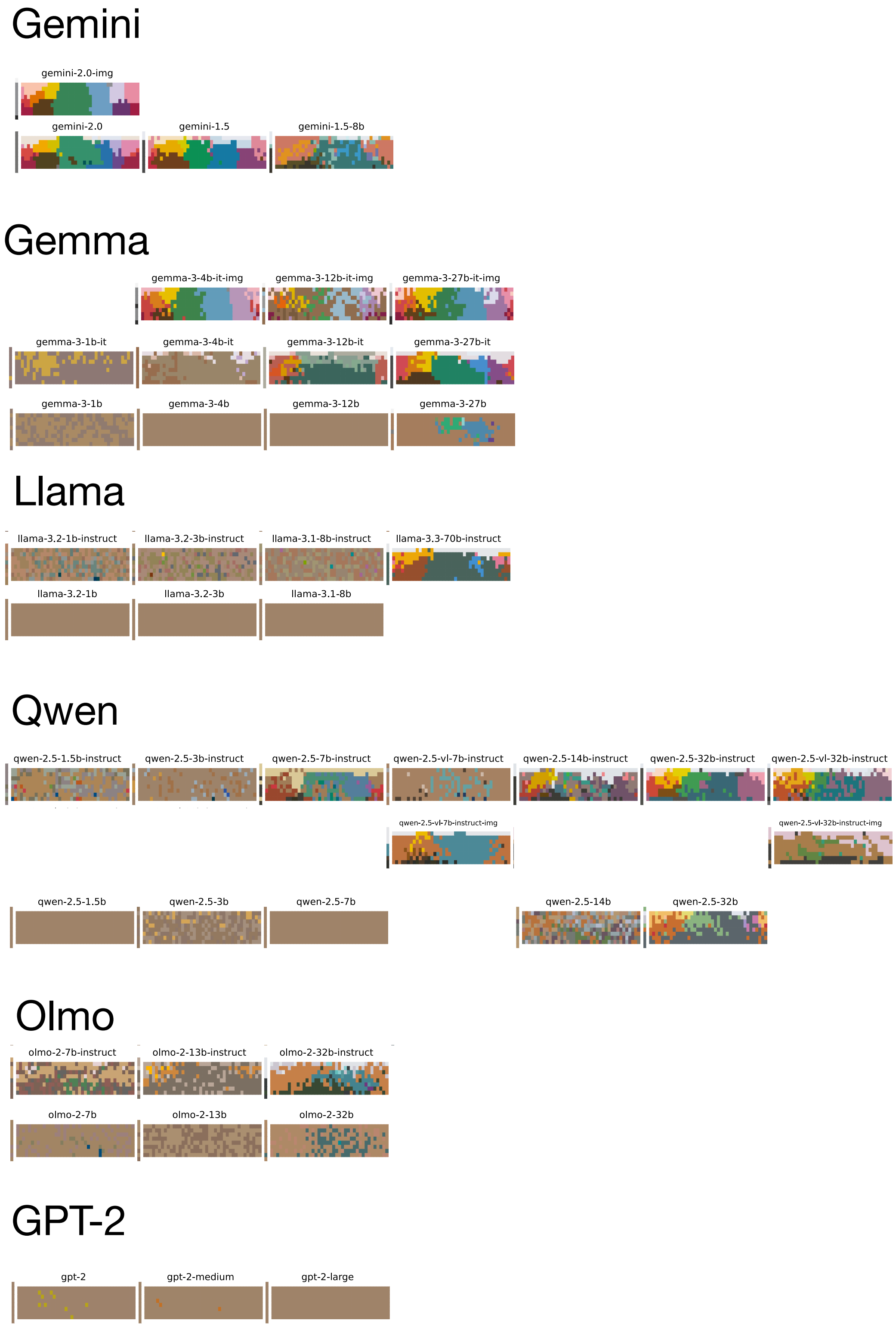


Figure 9: Mode maps for all models. The suffix '-img' for multimodal models indicates that images were presented as the color stimulus, rather than sRGB coordinates.

F OLMO TRAINING TRAJECTORY

As model checkpoints are available for the base Olmo 2 32B (OLMo-Team et al., 2025) model throughout its training, we repeated our English color naming study throughout different stages of its model training. The training procedure involved two main phases: pre-training on up to 6 trillion tokens, and a second training stage using a curated dataset of 843 billion tokens, including high-quality, academic, and instruction-tuning data. This also involved training on 100 billion and 300 billion token samples and then averaging the final checkpoints in a process called model souping.

Furthermore, we found that hinting to the model that the coordinates were in fact sRGB values could improve model performance. Specifically, when the prompt was changed to "What color is this rgb=[`<color coordinates>`]?". To aid in our analysis of the base pretrained Olmo models, we used this form of prompting in order to see what structure is achievable in principle.

Figure 10 shows that there is a slight increase in scores towards the end of stage 1 training, and a large jump in performance occurs after just 1000 steps of stage 2 training. This is particularly interesting given that stage 2 included *pretraining* on instruction-tuning data, given that we found instruction-tuning was strongly associated with higher alignment and complexity across other models.

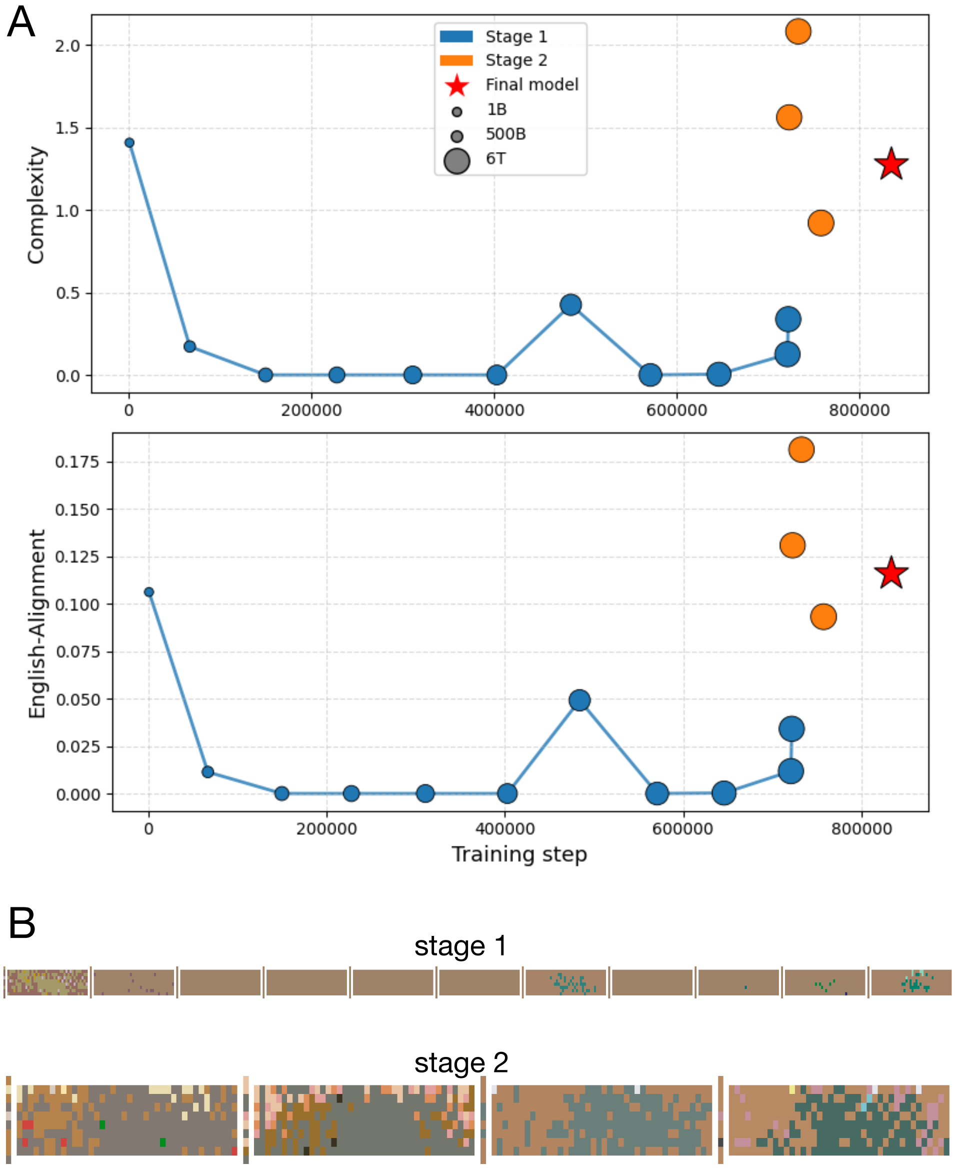


Figure 10: **(a)** Trajectory of English-alignment and complexity of Olmo 2 32B systems over the course of training. **(b)** Mode maps of Olmo systems over the course of training.

G Iterated in-context language learning procedure

Procedure We aimed to replicate the procedure of Xu et al. (2013). The initial training data for the first generation of IICLL consisted of random pairings of colors and nonsense terms, with a sample size six times the allowed vocabulary size. For vocabulary sizes ranging from 2 to 6 terms, the initial training data was sourced directly from Xu et al. (2013) to facilitate direct comparison. A 14-term condition was also included for comparison with English (Lindsey & Brown, 2014), with its initial mappings randomly generated as it was not part of the original study.

Next, we performed a sanity check to ensure that the LLM could recover the information from the training data. This involved presenting each stimulus from the training set to the LLM one at a time, in a random order and without any history of previous interactions, and verifying that the model could recover the correct labels as provided in the prompt. With the exception of the most resource-intensive $k = 14$ condition, all models achieved a stable comprehension accuracy of $\sim 80\%$ or greater throughout generations. The $k = 14$ condition served as an extreme-case sanity check, resulting in highly unstable performance (based on a failure to retrieve the required 84 examples; see the final paragraph of this section). Only Gemini occasionally passed this threshold; we elected to report the $k = 14$ condition results to illustrate the point of divergence in model capability.

Following this, we began the ‘production phase’, in which we prompted the model to provide a label for an unseen color stimulus. This was repeated for every item in the WCS stimulus array, presented in a randomized order, including those stimuli that were part of the initial training set for that generation. During this production phase, we also monitored the LLM’s consistency with the labels provided in the initial training set. While perfect adherence was not crucial since deviations may reflect the models’ inductive biases, we observed that consistency with the training set remained much higher than chance in the first generations and quickly rose to ceiling.

In addition, to promote response coherence and crudely mimic short-term memory influences found in human experiments, each subsequent prompt was augmented with a sliding window of the 10 most recent user-model interactions (see Appendix K for details on window size selection and its impact).

Once labels were obtained for all stimuli in the production phase, this full set of stimulus-label pairs constituted the current generation’s complete color category system. This marked the completion of one generation. To initiate the next generation, a new training set was created by randomly sampling stimulus-word pairs from the just-completed generation’s system, with the sample size being six times the number of allowed words in the vocabulary. This entire process was repeated for up to 13 generations, (which was the length of chains in Xu et al. (2013)) or until two successive generations yielded degenerate systems (where all stimuli are mapped to a single term).

H Rotation analysis

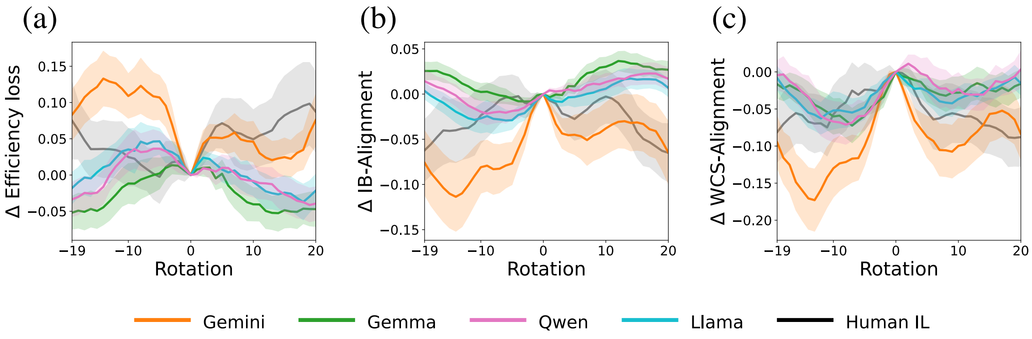


Figure 11: Rotation analysis for final systems of all models' IICLL chains and human ILL chains from Xu et al. (2013), WCS languages (Kay et al., 2009). The x-axis denotes rotation along columns (hue dimension) of the WCS grid, while the y-axis is the difference (Δ) in mean (a) efficiency loss, (b) alignment to IB optima, or (c) alignment to languages in the WCS. Each colored curve is the average across initializations and conditions, and the colored region corresponds to the 95% confidence interval around the mean (assuming a normal distribution). In (a), $\Delta > 0$ means that the actual systems are more efficient that their rotated counterparts, whereas in (b) and (c) $\Delta < 0$ means that the actual systems are more aligned.

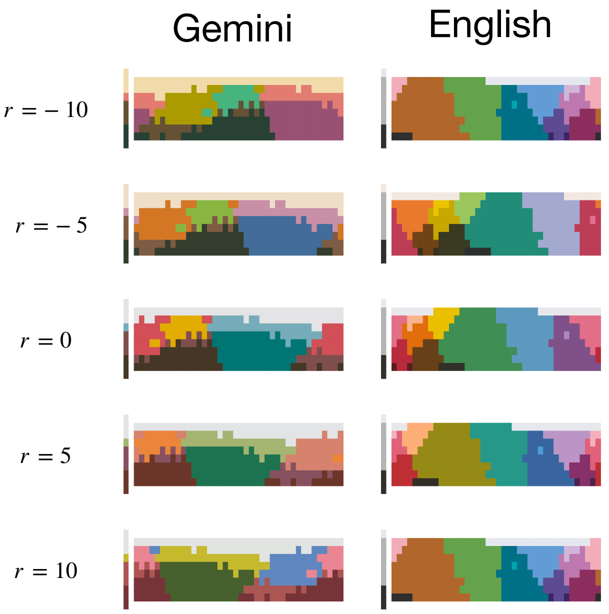


Figure 12: Rotation examples. Hypothetical variants for a final generation system of IICLL with Gemini 2.0 in the $k = 14$ condition, and the English system from Lindsey & Brown (2014). Variants are obtained by rotating the color naming system in the hue dimension across the columns of the WCS stimulus palette. $r = 0$ corresponds to the actual system, $r = 5$ corresponds to a shift of five columns to the right, and $r = -5$ corresponds to a shift of five columns to the left.

I Example chains compared to World Color Survey languages

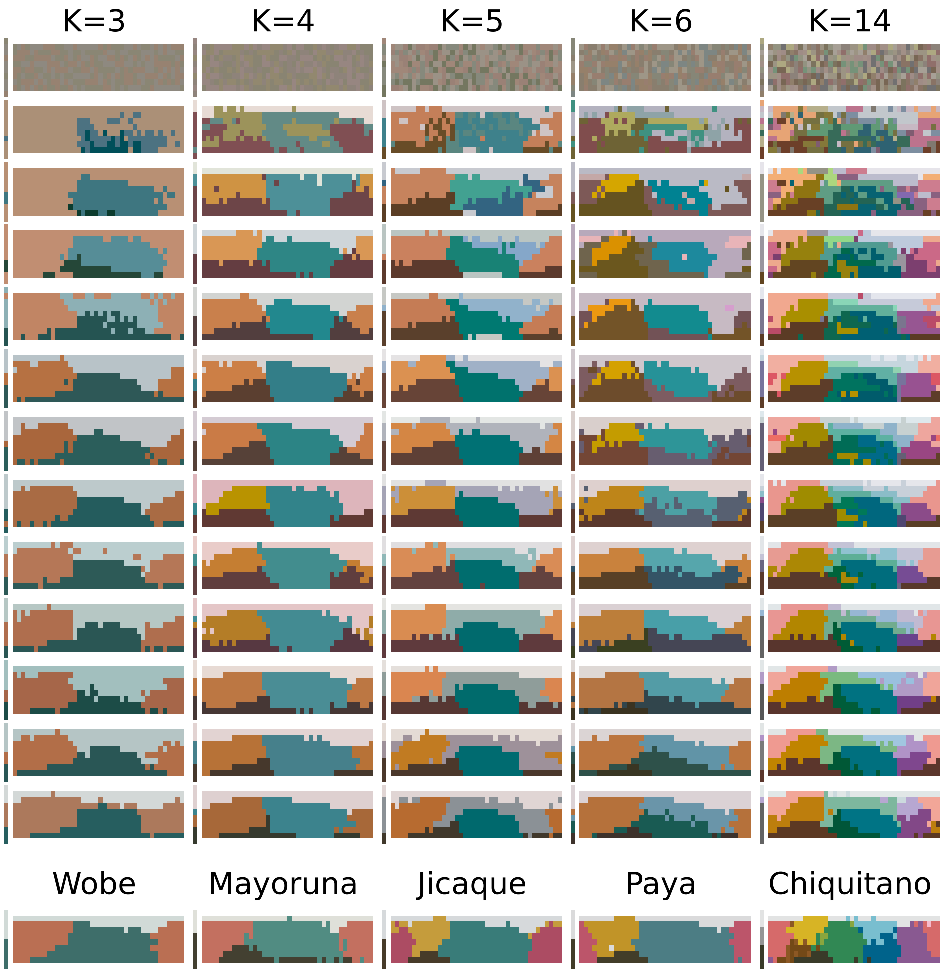


Figure 13: Mode maps for Gemini-2.0 systems over generations of IICLL. Bottom row shows the language from the WCS dataset that is most similar to the final generation's system.

J PROMPTING

In both of our studies, we provided instructions in the prompts to choose only from a fixed set of terms. The Gemini API supports controlled generation which makes this constrained classification task straightforward. For all open-weight models, we used log probability based scoring of the allowed terms as a continuation of the prompt to implement constrained choice. We used this decoding approach because after initial exploration of a brute-force method of repeating the prompt up to ten times until the model generated an output from the allowed set of terms, we found that only Llama 3.3-70B-Instruct could generate usable output (and in this case produced comparable behavior). For our probability-based decoding, we used the default generation configurations loaded from the Hugging Face Transformers library, which include a decoding temperature of 0.6 and a `top_p` sampling threshold of 0.9.

J.1 EXAMPLE PROMPT FOR ENGLISH COLOR NAMING TASK

```
‘‘What color is this [0.73579176, 0.13100809, 0.20245084]? You may only use one of the
    following allowed labels: [’Red’, ’Blue’, ’Yellow’, ’Green’, ’Orange’, ’Purple’, ’Pink’,
    ’Brown’, ’Black’, ’White’, ’Gray’, ’Peach’, ’Lavender’, ’Maroon’]. Please provide only a
    single label from the list just provided. Do not give any explanation.”
```

J.2 EXAMPLE PROMPT DURING ITERATED LEARNING

Training During training, the instructions were:

```
[
  {
    ”role”: ”user”,
    ”content”: [
      {
        ”type”: ”text”,
        ”text”: ”Features: [0.73579176, 0.13100809, 0.20245084] −> Label: Tovo
                Features: [0.0, 0.32875953, 0.29290289] −> Label: Feglu
                Features: [0.27329472, 0.29777161, 0.12539189] −> Label: Feglu
                Features: [0.18075972, 0.20165954, 0.14769053] −> Label: Narp
                Features: [0.77448248, 0.32302429, 0.52727771] −> Label: Zarn
                Features: [0.49405556, 0.30562009, 0.57919747] −> Label: Mib
                Features: [0.71954112, 0.66114241, 0.82844603] −> Label: Mib
                Features: [0.13380154, 0.21466098, 0.10314594] −> Label: Tovo
                Features: [0.99613596, 0.732415, 0.63294812] −> Label: Tovo
                Features: [0.77102815, 0.83377671, 0.0] −> Label: Blim
                Features: [0.94004023, 0.73594053, 0.83545817] −> Label: Blim
                Features: [0.77780255, 0.4893714, 0.71447577] −> Label: Narp
                Features: [0.89354841, 0.92711068, 0.55762157] −> Label: Zarn
                Features: [0.18901356, 0.19997485, 0.13980956] −> Label: Blim
                Features: [0.26099322, 0.1368575, 0.35608507] −> Label: Zarn
                Features: [0.0, 0.67091016, 0.50450556] −> Label: Mib
                Features: [0.86141792, 0.28265837, 0.0] −> Label: Feglu
                Features: [0.0, 0.42196565, 0.49203637] −> Label: Mib
                Features: [0.80047349, 0.92499868, 0.91404717] −> Label: Mib
                Features: [0.27606817, 0.14437327, 0.29090483] −> Label: Zarn
                Features: [0.91415567, 0.59970537, 0.67293472] −> Label: Narp
                Features: [0.8590007, 0.65998705, 0.0] −> Label: Feglu
                Features: [0.45183228, 0.37479448, 0.08237481] −> Label: Zarn
                Features: [0.91680269, 0.63513813, 0.0] −> Label: Feglu
                Features: [0.8994021, 0.89968419, 0.89916582] −> Label: Blim
                Features: [0.58159027, 0.030559, 0.08701426] −> Label: Blim
                Features: [0.84117237, 0.91203955, 0.93875183] −> Label: Feglu
                Features: [0.12209584, 0.66759353, 0.39283492] −> Label: Tovo
                Features: [0.81026541, 0.67994708, 0.0] −> Label: Blim
                Features: [0.81051796, 0.77667486, 0.88512298] −> Label: Feglu
                Features: [0.47551109, 0.20309021, 0.05617448] −> Label: Mib
                Features: [0.70626646, 0.35443549, 0.64213025] −> Label: Mib
                Features: [0.36974002, 0.20933179, 0.50495342] −> Label: Zarn
                Features: [0.0, 0.44248437, 0.46499573] −> Label: Tovo
                Features: [0.29926292, 0.15509473, 0.11356989] −> Label: Feglu
                Features: [0.0, 0.33817158, 0.21461567] −> Label: Mib”
      }
    ]
  },
  {
    ”role”: ”assistant”,
    ”content”:
  },
  ...
]
```

Generalization During generalization sessions, the prompt included (i) the entire training set, (ii) up to 10 previous user-assistant interaction pairs to serve as a form of memory / history of the conversation, and (iii) the current stimulus being queried from the language model.

For the history, previous interactions appeared in the form:

```
[
  {
    ”role”: ”user”,
    ”content”: [
      {
        ”type”: ”text”,
        ”text”: ”Based on the preceding examples, what is the label that best describes this?
             Do not give any explanation, and limit your response to exactly one word from
             this list of labels: [’Narp’, ’Tovo’, ’Feglu’, ’Mib’, ’Blim’, ’Zarn’].\nFeatures:
              [0.59395637, 0.24607302, 0.5432978] -> Label: ”
      }
    ]
  },
  {
    ”role”: ”assistant”,
    ”content”: ”Mib”
  },
  {
    ”role”: ”user”,
    ”content”: [
      {
        ”type”: ”text”,
        ”text”: ”Based on the preceding examples, what is the label that best describes this?
             Do not give any explanation, and limit your response to exactly one word from
             this list of labels: [’Narp’, ’Tovo’, ’Feglu’, ’Mib’, ’Blim’, ’Zarn’].\nFeatures:
              [0.73535226, 0.11620602, 0.27974255] -> Label: ”
      }
    ]
  },
  {
    ”role”: ”assistant”,
    ”content”: ”Blim”
  },
  ...
]
```

For the current stimulus being queried for a label, the presentation was of the form:

```
  {
    ”role”: ”user”,
    ”content”: [
      {
        ”type”: ”text”,
        ”text”: ”Based on the preceding examples, what is the label that best describes this?
             Do not give any explanation, and limit your response to exactly one word from
             this list of labels: [’Narp’, ’Tovo’, ’Feglu’, ’Mib’, ’Blim’, ’Zarn’].\nFeatures:
              [0.73535226, 0.11620602, 0.27974255] -> Label: ”
      }
    ]
  },
  {
    ”role”: ”assistant”,
    ”content”:
  },
```

K Sliding window conversation history

During the production phase of our IICLL experiments, each response from the model was saved and added to a sliding window of N previous user-model interactions, excluding the initial training data. After preliminary explorations with window sizes of 0, 10, 20, and 50, we determined that a window size of 0 led to degenerate category systems more frequently, while the results from a window size of 20 and 50 did not significantly differ from those obtained with a window size of 10. Consequently, we set the window size to 10 for all IICLL experiments. This window was included in the prompt for subsequent stimuli, presented after the initial training data and before each new stimulus, aiming to promote coherence in the model's responses and to crudely mimic the influence of short-term memory that human participants would possess in Xu et al. (2013)'s experiments.

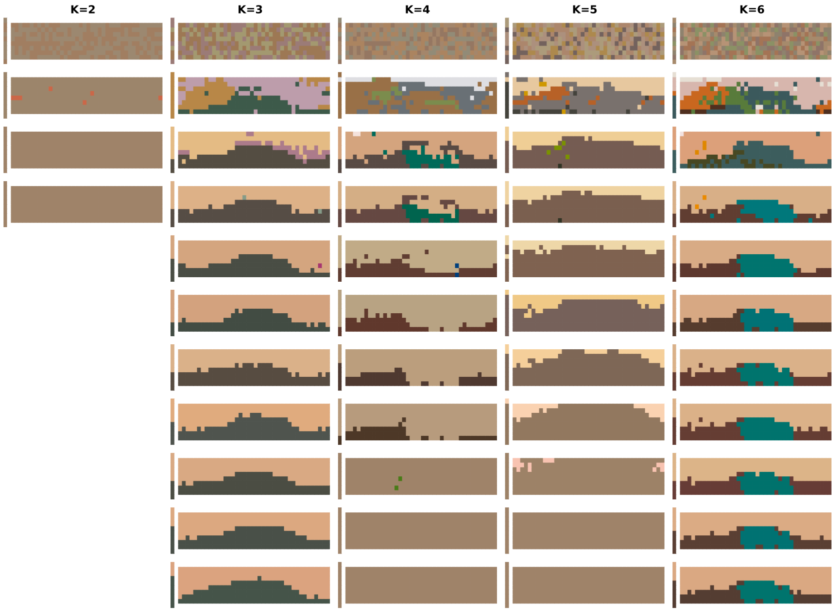


Figure 14: Randomly initialized IICL chains with Llama-3.3-70B-Instruct with a history window of 0.

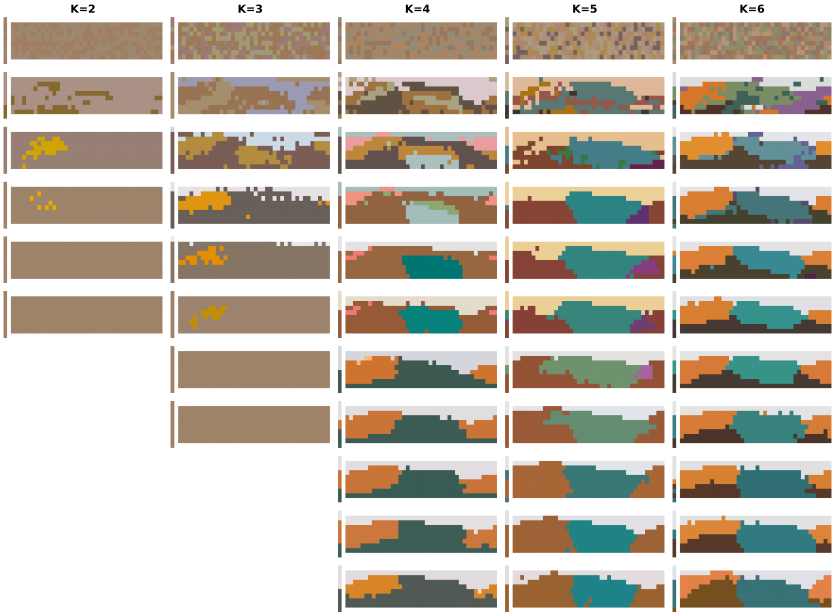


Figure 15: Randomly initialized IICL chains with Gemini-2.0 with a history window of 0.

L IICLL WITH SMALLER MODELS

We restricted our main Iterated In-Context Language Learning (IICLL) analysis to models that performed well on the initial English naming task to focus specifically on the evolution towards non-trivial IB-efficiency. This prioritization was based on our suspicion that models lacking sufficient inference power, coherence in color naming, or in-context learning capacity would be unable to sustain non-trivial category structure under IL pressure.

In order to confirm this empirically for models with weak initial performance, we ran supplementary IICLL simulations using four models that performed modestly at the English naming task: Gemma 3 4b, Llama 3.2 3B, Qwen 2.5 3B, and OLMo 2 7B (all instruction-tuned). These simulations covered the five word-number conditions ($k = 2$ through 6) used in the experiments of Xu et al. (2013).

The trajectories of these chains are depicted in Figure 16. Across these 20 chains, only the Gemma model converged to a two-word system (in the $k = 6$ condition), while the remaining 19 chains all converged to degenerate, single-word systems. This confirms that smaller models with limited inference power or in-context learning capacity tend to collapse to degenerate systems when subjected to the noisy transmission pressure of iterated learning.

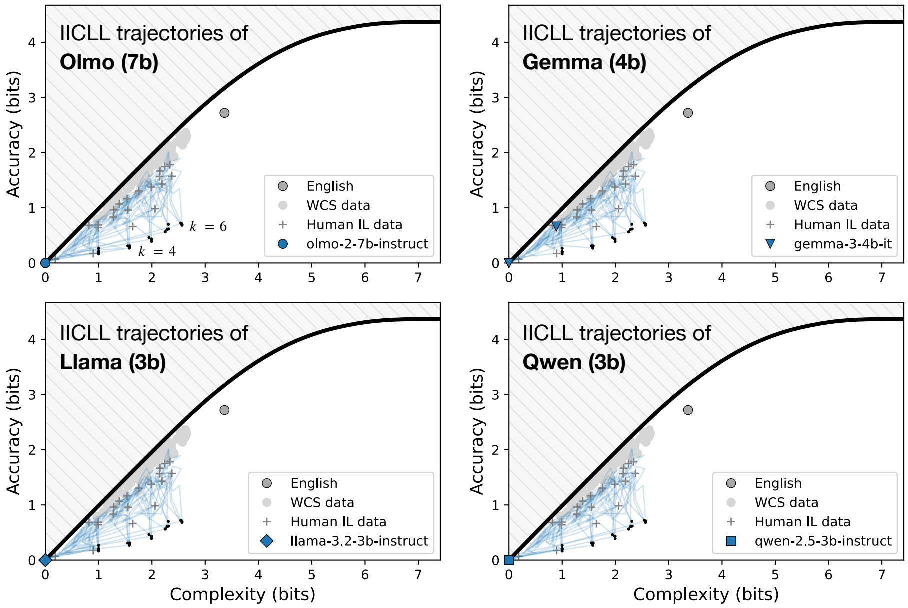


Figure 16: A sample of IICLL trajectories with smaller models from the English naming task. Compare to Figure 3 in the main text.

M Nearest Neighbor Baseline for IICLL

To test whether the semantic category systems that emerge from IICLL are merely simple approximations of feature-based clustering, we conducted a set of baseline simulations using a Nearest Neighbor (NN) classifier in place of an LLM in the IICLL task. In this setup, the NN classifier predicted the category label for each test stimulus based solely on the sRGB features of the provided in-context training examples.

We analyzed performance in the most challenging condition ($k = 14$ allowed terms). The results, presented in Figure 17, show that while the leading open-weight LLMs achieve significantly worse alignment compared to the NN baseline, the leading frontier model, Gemini 2.0, performs significantly better than this baseline on all evaluation metrics (efficiency, IB-alignment, and WCS alignment). In other words, Gemini's evolved categories are significantly more human-like and more IB-like than a simple nearest-neighbor clustering of the input examples.

This result suggests that the ability to acquire a human-like, optimally-compressed semantic representation is (1) not trivial because it is a capability only emergent in the most advanced frontier model, and (2) it cannot be explained by a simpler nearest-neighbor process that only maintains contiguous partitions regardless of their efficiency and IB-alignment.

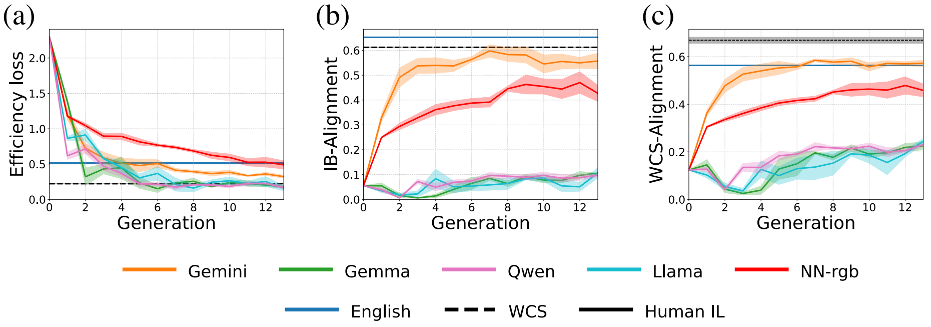


Figure 17: Evolution of LLM systems restricted to the $k = 14$ condition, shown in comparison to the Nearest Neighbor sRGB baseline classifier (NN-rgb). While many models struggle to achieve alignment to IB and human languages comparable to that of the Nearest Neighbor classifier, Gemini achieves greater efficiency (a), alignment with optimal IB systems (b), and alignment with human languages (c) across IICLL generations. Compare to Figure 4.