Searching for Structure: Investigating Emergent Communication with Large Language Models

Tom Kouwenhoven[1], **Max Peeperkorn**[2], **and Tessa Verhoef**[1],
[1]Leiden Institute of Advanced Computer Science, Leiden University, Netherlands,
[2]School of Computing, University of Kent, United Kingdom,
{t.kouwenhoven, t.verhoef}@liacs.leidenuniv.nl

Abstract

Human languages have evolved to be structured through repeated language learning and use. These processes introduce biases that operate during language acquisition and shape linguistic systems toward communicative efficiency. In this paper, we investigate whether the same happens if artificial languages are optimised for implicit biases of Large Language Models (LLMs). To this end, we simulate a classical referential game in which LLMs learn and use artificial languages. Our results show that initially unstructured holistic languages are indeed shaped to have some structural properties that allow two LLM agents to communicate successfully. Similar to observations in human experiments, generational transmission increases the learnability of languages, but can at the same time result in non-humanlike degenerate vocabularies. Taken together, this work extends experimental findings, shows that LLMs can be used as tools in simulations of language evolution, and opens possibilities for future human-machine experiments in this field.

1 Introduction

Vocabularies of signals enable us to communicate about meanings, but to express an arbitrary number of meanings, vocabularies would require an equally large set of words as there are meanings, and learning such holistic vocabularies is cognitively challenging. Human languages therefore typically show some form of compositional structure, where meaningful signal-meaning mappings can be composed such that the combination of individual meaningful signals can express more than the meaning of the individual components alone (Hockett, 1960). An important finding in the field of language evolution is that such structural properties can emerge as a result of individual learning biases and pressures that continuously shape the languages on a longer timescale, often eventually resulting in languages that are easier to learn and exhibit some degree of structure (Smith, 2022).

The processes involved in the evolution of language have been investigated abundantly with experiments and simulations. The latter typically use hard-coded agents with inductive biases (de Boer, 2006), Bayesian learners (e.g. Griffiths and Kalish, 2007; Culbertson and Smolensky, 2012; Kirby et al., 2015), or reinforcement learning agents (Lazaridou and Baroni, 2020) to investigate the evolution of structured languages. In contrast, we investigate whether more flexible LLMs as relatively unbiased language learners (Wilcox et al., 2023) are appropriate tools to study how languages evolve. While their internal mechanisms are fundamentally different from humans, they still are the first close flexible comparators of human language users that can be used as tools to answer cognitive and typological investigations (Warstadt and Bowman, 2022; van Dijk et al., 2023). Given that languages are shaped by the biases and pressures of individual language learners, which are different for LLMs (e.g., fewer memory constraints), we are interested in finding similarities and differences between humans and LLMs on specific language evolution-oriented tasks.

Our work largely follows the experimental design by Kirby et al. (2015) in which Bayesian learners and humans learn an artificial language to communicate in a referential game. They find that linguistic structure arises from a trade-off between pressures for compressibility and expressivity. Our work extends their work by using LLMs as objects of investigation. Specifically, we investigate how artificial languages evolve when two LLMs communicate in a referential game and what the effects of generational transmission on these languages are. We compare properties of these languages to those that are found in experiments involving humans. Results show that 1) LLMs can learn artificial languages and use them to communicate

Proceedings of the 31st International Conference on Computational Linguistics, pages 9977–9991
January 19–24, 2025. ©2025 Association for Computational Linguistics

successfully, 2) the languages exhibit higher degrees of structure after multiple communication rounds, 3) LLMs generalise in more systematic ways when the evolved language is more structured, and 4) languages adapt, although not necessarily in a human-like way, and become easier to learn by the LLMs as a result of generational transmission.

2 Background & Related work

2.1 The evolution of structure

Learning novel signal-meaning mappings, and the emergence of rules that can combine these signals into structured languages have been abundantly investigated in the field of language evolution using human experiments (Kirby et al., 2008; Galantucci, 2005; Scott-Phillips et al., 2009; Verhoef, 2012; Raviv et al., 2019a,b; Kouwenhoven et al., 2022a) and computational simulations (de Boer, 2006; Steels et al., 2012; Lazaridou and Baroni, 2020; Kouwenhoven et al., 2024). These typically follow a setup where success depends on cooperation between two or more participants/agents in a Lewis game. Here, players are prevented from communicating using conventional communicative means and instead must establish novel communication systems through repeated cooperation. Outcomes often show that players, human or machine, quickly establish novel signal-meaning mappings that enable them to communicate successfully. However, recent computational simulations using reinforcement learning agents often develop communicative systems different from those of humans (Galke et al., 2022)[1] unless specific key pressures are introduced to recover initially absent human patterns (Galke and Raviv, 2024).

It has been suggested that seemingly arbitrary aspects of linguistic structure may result from general learning and processing biases deriving from the structure of thought processes, perceptuo-motor factors, cognitive limitations, and pragmatics (Christiansen and Chater, 2008). A well-investigated cause for this phenomenon is the process of cumulative cultural evolution (Boyd et al., 1996; Tomasello, 1999), which is typically investigated using iterated learning experiments (Kirby et al., 2008). Here, information (e.g., a language) is repeatedly passed down from one generation to the next, where the information is modified and improved upon within each generation. The influential work from Kirby et al. (2008, 2015) shows that when human individuals learn an artificial language that was previously learned by another individual, languages become easier to learn and display a higher degree of structure. Crucially, these results are mostly attributed to the fact that the language repeatedly goes through a learning bottleneck, in which individual cognitive biases such as memory constraints slowly shape the language. Iterated learning has been used to show that structure emerges in various setups with, for example, continuous signals (Verhoef, 2012) or continuous meaning spaces (Carr et al., 2017), and it is argued that it may have led to the statistical Zipfian structure of language (Arnon and Kirby, 2024). Yet, Raviv et al. (2019a) showed that structure can also emerge *without* generational transmission. In this case, a pressure for compressibility originating from communication with multiple interaction partners and expanding meaning spaces causes languages to become compositional. This effect is even more prominent if the number of interaction partners is larger (Raviv et al., 2019b). The current work is inspired by the traditional methods described before and extends them with our current most sophisticated models of natural language.

2.2 LLMs as models of language

LLMs are sophisticated models of natural languages and growing evidence shows their ability to exhibit ‘average' human behaviours. It is, for example, suggested that LLMs can model human moral judgements (Dillion et al., 2023) and transmission chain experiments revealed human-like content biases in GPT-3.5 (Acerbi and Stubbersfield, 2023). When LLMs are extended with records of experiences, Park et al. (2023) showed that groups of generative agents display believable human-like individual and emergent social behaviours when they interact over extended periods. It is even suggested that human-LLM interactions in everyday life can potentially mediate human cultures through their influence on cultural evolutionary processes of variation, transmission and selection (Brinkmann et al., 2023; Yiu et al., 2024).

While previous work has investigated human-like behaviour at inference time, findings from cognitive science can also be used to improve model performance. Iterated learning can, for example,

[1] But see Lian et al. (2023, 2024); Zhang et al. (2024) for recent work showing that the need to be understood (i.e. communicative success), noise, context sensitivity, and incremental sentence processing help induce human-like patterns of dependency length minimisation in reinforcement learning agents.

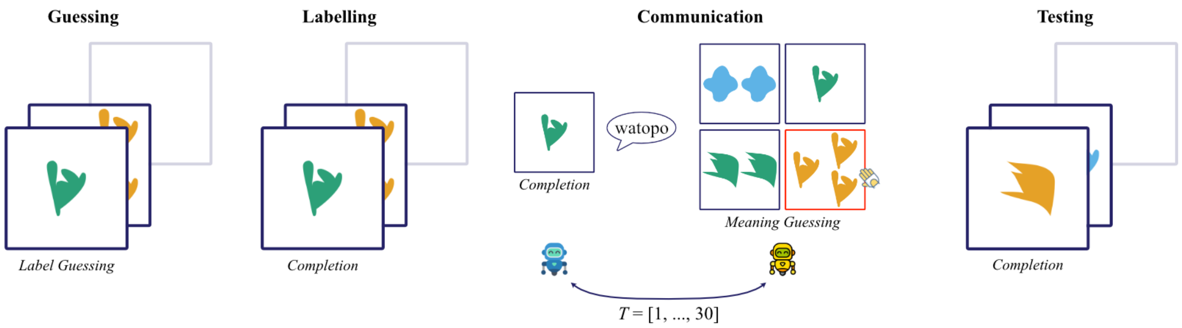


Figure 1: A graphical representation of the experimental blocks. The agents first go through a guessing block before labelling each of the 15 training stimuli in the labelling block. The communication block is done for 4 rounds each consisting of 30 tasks T, where the agents alternate speaker-listener roles to be a speaker and listener for each stimulus once. Finally, the agents label 27 (15 original and 12 novel) stimuli in the testing block.

be incorporated into the training regime to extrapolate desirable behaviours. Zheng et al. (2024) have likewise shown that representations are easier to learn when vision-language contrastive learning is reframed as the Lewis signaling game between a vision agent and a language agent, ultimately improving compositional reasoning in vision-language models. However, this does not guarantee model improvements, Shumailov et al. (2024) have for example shown that LLMs, autoencoders and Gaussian mixture models drift when trained repeatedly on AI-generated data. In these cases, crucially, the generated content is slowly optimised to be understandable for models, *not* for humans, resulting in what they call model collapse. The authors therefore argue that genuine human interactions with systems will be increasingly important to prevent model collapse. While drift is often seen as an unwanted effect of unsupervised training, from a language evolution viewpoint this is not surprising, since languages adapt to how they are learned and used. As such, it has been suggested that languages should adapt to become more natural for humans *and* machines (Kouwenhoven et al., 2022b) and that findings from cognitive science can prevent modal collapse (Smith et al., 2024) or inform modelling choices (Galke and Raviv, 2024). Here, we view LLMs from this evolutionary perspective.

Although biases inherent to a language model's (pre-)training objectives (i.e. the cloze task and instruction tuning) and memory constraints are very different from those in humans, recent work has shown that GPT-2 models struggle to learn languages that contain unnatural word orders, lack hierarchical structure, or lack information locality (Kallini et al., 2024). This suggests that, even though the language processing mechanisms in transformers are non-humanlike, LLMs share some preference for structured languages similar to humans. Moreover, in an artificial language learning experiment similar to the work presented here, Galke et al. (2023) showed that compositional structure is advantageous for GPT-3 when learning an artificial language and that a higher degree of compositional structure also resulted in human-like generalisation for new unseen items. Our work is different in that Galke et al. tested the ability of GPT-3 to learn languages that evolved during a *human* experiment (Raviv et al., 2019b, 2021), thus being optimised for human learners. We instead wish to investigate what kinds of languages evolve when they are optimised for *LLMs*.

3 Methodology

Our methodology is inspired by Kirby et al. (2015) and (Raviv et al., 2021). The complete simulation set-up consists of four blocks: guessing, labelling, communication and testing (§3.2 & Figure 1[1]). The agents perform the guessing, labelling, and testing block separately; communication is interactive. The communication block is a classic referential game in which two agents communicate to discriminate a target stimulus from four distractor stimuli. They do so in four rounds, each consisting of 30 interactions T, alternating speaker-listener roles between interactions. In a single interaction round, the speaker observes a target stimulus and utters a signal that describes the current stimulus. Using

[1] This is for illustration purposes only, we stress that our simulations are entirely run in the textual modality only to avoid the additional challenge of extracting relevant visual features and mapping these to artificial languages.

```
{'shape':3,'colour':'blue','amount':1,'word':'ninikonu'}
{'shape':1,'colour':'green','amount':3,'word':'hanosa'}
    ⋮
{'shape':2,'colour':'orange','amount':2,'word':'sanu'}
{'shape':1,'colour':'green','amount':3,'word':'[COMPLETE]
```

Prompt 1: A vocabulary snippet used in a completion prompt. Full prompts are visible in Appendix C

this utterance, the listener must discriminate the correct target. Cooperation is successful when the listener's guess is the target stimulus.

3.1 Stimuli and initial languages

The meaning space consists of stimuli with three attributes. They have one of three shapes, one of three colours, and can appear in groups of one, two, or three shapes, creating 27 distinct stimuli. Initial signals for these stimuli were generated before each experiment according to the method used by Kirby et al. (2008). The signals are concatenations of 2, 3, or 4 randomly picked consonant-vowel (CV) syllables resulting in artificial non-existing signals (e.g., watopo, nafa, nomomeme). The CV syllables consist of one of eight consonants g, h, k, l, m, n, p, w and one of five vowels a, e, i, o, u. Out of 27 stimuli, only 15 stimuli are used during the guessing, labelling, and communication blocks. All 27 stimuli are used in the testing block. The training stimuli are selected randomly before each simulation, but we ensure that each attribute value is represented equally often across this set.

3.2 Simulation blocks

Each simulation consists of four blocks. In the first block, we assess whether agents can guess the right signal when presented with a stimulus. Second, in the labelling block, an agent repeatedly produces a signal for each stimulus given the initial training vocabulary. The signals generated in this block are taken as the learned vocabulary for that agent. In the third block, the agents communicate as described before, taking turns as speaker and listener until all rounds are completed and each stimulus appeared twice per round (i.e., both agents produced a signal for each stimulus and made a guess for each stimulus). In this block, the interaction between the agents slowly alters each agent's individual vocabulary much like done by (de Boer, 2000; Steels et al., 2012) by updating the current stimulus to be associated with the produced signal. After the communication block, the testing block tasks the agents to generate signals for the entire meaning space of 27 stimuli using the training vocabulary that was optimised in the labelling and communication block. Hence, they must generalise their strategies to unseen samples.

3.3 LLMs as agents

The LLMs in our experiment were instruction-tuned instantiations of Llama 3 70B (Llama Team, 2024) with greedy sampling[2]. While human participants typically learn signal-meaning mappings through a learning block, we use LLMs' in-context learning (Brown et al., 2020) ability to teach them the languages. Specifically, we prepend our prompts with the items to be learned in a structured JSON-like format (Prompt 1). Given the observed behavioural similarities between humans and LLMs (Galke et al., 2023), we assume that a vocabulary of signal-meaning mappings in the context of a prompt provides enough (distributional) information for a LLM to learn an appropriate mapping between the attributes of the stimuli and signal syllables. Although the prompt structure 'invites' the LLM to infer a signal from the stimulus attributes, we are agnostic about how exactly and what kind of mapping the LLM deduces, but we are interested in the resulting behaviours.

Throughout a simulation, agents essentially perform one of two tasks: generation or guessing. The labelling block and speaking in the communication block involve generating signals. The guessing block and discrimination in the communication block involve guessing. The prompts for these tasks are extensions of those used by Galke et al. (2023), with slight adaptations to enable LLMs to discriminate between stimuli. Given that LLMs show a primacy and recency bias (Liu et al., 2024), the vocabulary is shuffled before each task such that ordering effects are minimal. System instructions depend on the task performed but are largely similar and chosen to be maximally close to instructions given to humans in experimental settings.

[2] Although we only report results on one model type, initial explorations with GPT-3.5 and Llama 2 7B showed similar behaviours to LLama 3 70B

Generating signals. For signal generation in the labelling block, we use prompt completion (Prompt 2). During labelling, the agents see the *entire* training set and generate a signal for each stimulus, effectively amounting to a look-up task since the stimulus is present in the prompt. On the other hand, the vocabulary presented to agents during communication and testing does *not* include the current stimulus, thus requiring the agents to extract an appropriate mapping and generalise to new stimuli (Prompt 3). A human-like solution would be to map stimulus attributes (i.e. shape, colour, and amount) to syllables representing these attributes and create compositions that describe the stimulus. During communication, we add a `communicativeSuccess` attribute which is set to 1 if the previous interaction for this stimulus was successful and zero otherwise. Adding this attribute functions as a memory between interactions and provides a pressure for expressivity. It is hypothesised that the latter plays an important role in human language evolution since it prevents languages from becoming degenerate (Smith et al., 2013). Importantly, during testing, the vocabulary presented to the agents always includes the train set (without the current stimulus), and items from the test set are never present.

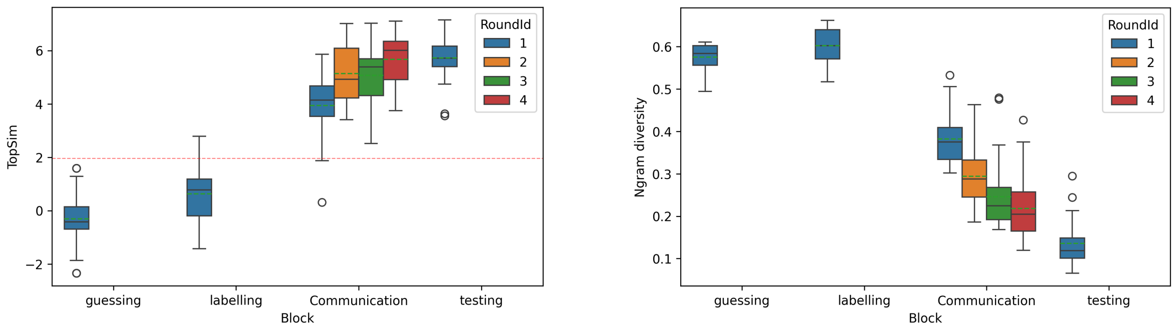
Guessing signals or meanings. For guessing and discrimination during communication, the agents need to respond with a choice corresponding to the speaker's signal. Unfortunately, LLMs are inconsistent and unreliable in answering multiple choice questions (Khatun and Brown, 2024). In our initial exploration, this indeed proved to be unusable. Instead, for each distractor (signal or meaning), we run the prompt prefilled with that distractor through the model and select the distractor with the highest probability (Prompt 4). Again, the agents observe the training vocabulary *with* the current stimulus in the guessing block. In the communication block, agents observe the training vocabulary *without* the current stimulus.

3.4 Metrics

We are firstly interested in investigating whether two agents settle on a language that enables them to communicate, measured by the percentage of successful interactions (*PercCom*) in a round. We use multiple metrics to measure structure in messages. The most common metric is topographic similarity (*TopSim*, Brighton and Kirby, 2006). Similar to Kirby et al. (2008), we report Z-scores of the Mantel test (Mantel, 1967) between signal similarities (normalised Levenshteins distance) and semantic similarities (the number of equal attributes between two meanings). A communication system with a high *TopSim* uses similar signals for similar meanings. We compute the *Ngram* diversity (Meister et al., 2023), being the average fraction of unique vs. total *Ngrams* for $N \in \{1, 2, 3, 4, 5\}$ in all produced signals. Low *Ngram* diversity across all signals implies the agents re-use parts of signals in different signals, hinting at compositional signals when it happens in combination with increased *TopSim*. We assess the degree of signal systematicity between the signals produced for unseen stimuli in the test block and the previous stimuli in the communication block using the generalisation score (*GenScore*, Raviv et al., 2021). Here, we first compute the pairwise semantic difference between each stimulus in the train and test scenes, followed by the pairwise normalised edit distance between the signals produced for these scenes. We then take the Pearson correlation between these differences across all stimuli. Intuitively, this measures whether similar scenes across both sets are similarly labelled, thereby suggesting generalisation.

4 Evaluation

We ran 15 simulations, each initialised with a random seed and unique artificial unstructured language. Metrics were computed for each block, except for the generalisation score, which is only computed for the testing block. A human-like result would show increasingly successful interactions and increasing *TopSim* scores, while *Ngram* diversity should go down. If this is the case, we expect to observe higher generalisation scores since agents can compose new signals according to a learned structured strategy. We use linear mixed effects models to analyse the results of the communication block to take the random effects of each simulation's vocabulary into account. The slope ($\hat{\beta}$) determines the direction of the effect and the rate of change. Additionally, we use conditional R^2 (Nakagawa and Schielzeth, 2013), denoted by R^2_c, which considers fixed and random effects, to show how much variance can be explained by the model. Higher values of R^2_c indicate that the model captures more variance and that correlations are stronger. Finally, we report the marginal R^2_m, which is the variance explained by the fixed effects.



(a) TopSim scores over the agent's vocabulary in each block and round. The dashed red line indicates the $p < .05$ level.

(b) Ngram diversity scores over the agent's vocabulary in each block and round.

Figure 2: Communication clearly increases the structure of the vocabularies, as seen by the increasing *TopSim* scores and decreasing *Ngram* diversity.

5 Results

5.1 Learning the artificial languages

We first assess whether LLMs were able to learn the initially unstructured languages. Given the nature of the guessing task, which is essentially a lookup task, unsurprisingly, LLMs were able to guess the correct signals for the stimuli almost perfectly ($M = .973, SD = .031$). However, labelling the same stimuli via completion proved much more difficult ($M = .453, SD = .152$) despite the presence of the correct signal in the prompt. This contrast is in line with work showing that LLM predictions are sensitive to task instructions and how predictions are extracted (Weber et al., 2023; Hu and Levy, 2023; Hu and Frank, 2024). Additionally, it corroborates using prefilled options in our guessing prompts during communication. Nevertheless, this performance is still better than that of humans[3] and is not unimpressive given the vast number of possible signals that can be produced. Finally, the expected struggle to correctly reproduce (i.e., learn) unstructured signals introduces some welcome variation to the agents' vocabulary which is used at the start of the communication block.

5.2 Agents communicate successfully

Once the agents have individually learned the vocabulary, they start communicating. Despite initially starting with different languages, approximately 70% of the interactions in the first round are successful (chance performance would amount to 25%). This increases somewhat in the following rounds to ≈ 75%, but not significantly (Appendix A, Figure 5). Interestingly, communicative success is not guaranteed, it fluctuates between rounds and in some simulations it even decreases drastically.

5.3 Communication results in structure

Although the initial languages are unstructured, some form of structure emerges due to repeated learning and use (Figure 2). This mostly happens during the communication block where *TopSim* increases significantly across rounds ($\hat{\beta} = .508 \pm .073, R^2_c = .579, R^2_m = .355, p < .001$) and *Ngram* decreases across rounds ($\hat{\beta} = -.054 \pm .004, R^2_c = .812, R^2_m = .558, p < .001$). This increase in structure benefits communicative success positively ($\hat{\beta} = .035 \pm .007, R^2_c = .769, R^2_m = .427, p < .001$). However, we also observe behaviour that is not human-like; the signals used to communicate become longer over the rounds ($\hat{\beta} = .557 \pm .044, R^2_c = .919, R^2_m = .505, p < .001$). This contradicts what is observed in human experiments, where we typically observe that messages become shorter and lie close to a theoretical frontier balancing expressivity and simplicity (Piantadosi et al., 2011; Kirby et al., 2015).

These results extend the findings of Galke et al. (2023); LLMs not only learn structured vocabularies better but also naturally shape languages to have some form of structure when they are optimised for their inherent biases. As LLMs struggle to learn impossible languages (Kallini et al., 2024), reframing prompt instructions into a structured list improves the model response (Mishra et al., 2022), and given that we do not impose pressure to induce structure, the surprising outcome of our experiments may be the result of an apparent "structure bias" in LLMs.

[3] Preliminary analyses of an ongoing experiment involving humans show that the guessing block is much easier than the labelling block.

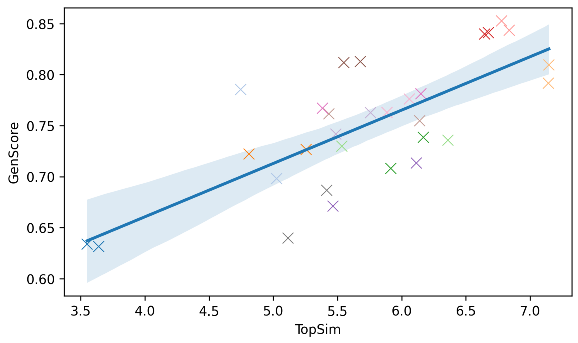


Figure 3: Languages that have evolved to be more structured allow for better generalisation to unseen test stimuli. Colours refer to individual simulations.

5.4 Structure enables better generalisation

After the communication block, the agents engage in the final simulation block. Here they generate signals for all 27 stimuli using the vocabulary that has evolved after learning and communication. We find that high *TopSim* languages allow for better generalisation ($r = 0.735, p < .001$, Figure 3).

A qualitative inspection of the signals generated in the testing block of the simulation which resulted in the highest *TopSim* after communication reveals that this agent repeatedly re-uses parts of signals in different compositions (Table 1). For example: “su” refers to the amount one, “pepi” to two, “petite” to three. For shape 1, the signals “sunu” and “sutu” are used, “ginu” for shape 2, and shape 3 is referred to with “wipi” or “wipu”. However, colours are less clearly demarcated by unique signal parts. This is also reflected in the ratio of unique signals produced during the test block ($M = 62.1\%, SD = 19.8\%$), showing that some simulations sometimes result in repetitive use of the same signals for different meanings, resulting in a somewhat degenerate vocabulary. Nevertheless, it is clear that unseen stimuli are often labelled similarly to previously seen stimuli.

	Shape	Colour	Amount	Word
train set	3	orange	1	wipisu
	1	green	2	sutupepi
	2	green	1	ginisu
	3	green	1	wipisu
	1	blue	2	sunupepi
	1	green	3	sutupitite
	2	orange	1	ginusu
	3	blue	3	wipipitite
	3	green	3	wipupitite
	3	blue	1	wipisu
	1	blue	3	sunupitite
	2	orange	3	ginupitite
	2	blue	2	ginupepi
	1	orange	2	sunupepi
	2	orange	2	ginupepi
test set	1	orange	1	sutisu
	1	orange	3	sutupitite
	1	green	1	sutusu
	1	blue	1	sunusi
	2	green	2	ginupepi
	2	green	3	ginupitite
	2	blue	1	ginisu
	2	blue	3	ginupitite
	3	orange	2	wipupepi
	3	orange	3	wipipitite
	3	green	2	wipupepi
	3	blue	2	wipupepi

Table 1: The signals produced in the testing phase of the simulation that resulted in the highest *TopSim* score (7.13) after communication. The signals for the test stimuli share parts of signals and are composed similarly to train stimuli ($GenScore = .792$)

6 Iterated learning

The previous results showed that two LLMs can successfully communicate and slowly shape the language to become more structured. Provided that cumulative cultural evolution can extrapolate weak biases to have strong effects in socially learned systems like language (Smith, 2011), we extend our simulations by adding generations of learners. The first generation is initialised with a random unstructured language described in Section 3.1, but in following generations, agents learn a portion of the signal-meaning mappings produced in the testing block by the agents of the previous generation. Only the vocabulary of the agent with the highest *TopSim* is transmitted to the next generation. We ran six transmission chains of 8 generations each. The seed generations for each chain were selected randomly from our initial 15 simulations.

6.1 Learnability increases

Figure 4 clearly reveals that the learnability increases. While LLMs in the first generation struggle to look up signals and reproduce them, a single generation learning and using a language tremendously decreases the edit distance between ground truth signals and the produced signals. These results are remarkably similar to findings with human participants (Kirby et al., 2015), and show that the

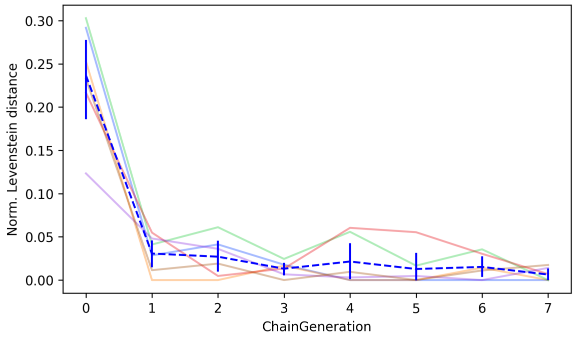


Figure 4: The normalised Levenshtein distance between the ground truth and the produced signal in the learning block. Solid lines indicate chains, dashed blue line indicates the average Levenshtein distance across chains.

languages are optimised for LLMs' preferences.

6.2 Communicative success and non-humanlike structures

Despite the increase in learnability, we do not observe an increase in communicative success due to iterated learning (Appendix B Figure 6). Possibly, this is due to the already high scores of the first generation. Despite their increased learnability, the signals become significantly longer and more ambiguous. We take this non-humanlike solution to be an artefact of an absence of pressures for memorisation in LLMs'. While human language is optimised to be compressible and expressive (Fedzechkina et al., 2012; Tamariz and Kirby, 2015; Kirby et al., 2015), context windows of LLMs, are considerably larger. In our case, Llama 3 70B has a context window of 8.2K tokens, which we do not exceed.

Finally, the metrics to measure structure display a mixed picture. *TopSim*, does increase but not significantly across generations (Figure 7 and Table 2 in appendix B) but *Ngram* diversity decreases significantly across generations. Qualitative inspections of several vocabularies show that some languages evolve into degenerate languages with repeating signals for different stimuli (i.e., underspecification). Corroborated by a significantly lower number of uniquely produced signals in the last generation compared to the first simulation ($(t(5) = 2.64, p = .046, M_{gen0} = .707, SD_{gen0} = .142, M_{gen7} = .519, SD_{gen7} = .119)$. Together this causes the *Ngram* diversity to be lower while clearly hurting communicative expressiveness. Even though degenerate languages are not uncommon in iterated learning experiments with humans (e.g., experiment 1 in Kirby et al., 2008), an additional pressure for expressivity typically prevents languages from becoming underspecified. Given the expressivity pressure that we expect to result from the communication block, we expected to see less underspecification. Iterated learning therefore results in vocabularies optimised for LLM agents but do so in a non-humanlike way.

7 Discussion

Our findings show a mixed picture, agents comprised of LLMs can learn and use artificial languages in a referential game. They do so by optimising the initially holistic vocabulary to fit better with the preferences of their language model, resulting in increased regularity and structure (Table 1). These human-like results are much in line with previous findings showing that structured languages can emerge from repeated interactions between interlocutors (i.a. Selten and Warglien, 2007; Verhoef et al., 2016; Nölle et al., 2018; Raviv et al., 2019a). Yet, we also observe some degeneracy, i.e. many-to-one mappings of signals and attributes, and non-humanlike behaviours such as a tendency to produce long signals. Iterated learning further increases the learnability of the vocabulary but also extrapolates these non-humanlike behaviours further. Despite not being able to *directly* compare our results to human data, these findings are loosely comparable to earlier work involving human participants (Kirby et al., 2015; Raviv et al., 2019b) in which languages with similar properties emerge.

Table 1 suggests that certain attributes, such as the colour attribute, in the inputs may be ignored, possibly due to the primacy and recency bias in LLMs (Liu et al., 2024). Optimising the instructive sentences by choosing sentences that maximise the fraction of valid model answers for each task, as suggested by Aher et al. (2023), may alleviate these ignorances and increase focus on relevant attributes. It is also possible that the LLMs do not 'experience' enough pressure to be understood by other agents, i.e., the *communicativeSuccess* attribute is not able to force a need to be expressive, which is deemed an essential pressure in computational simulations for human-like structures (Galke and Raviv, 2024). Despite these discrepancies, it is nevertheless interesting that some form of structure emerges.

Our results further show variability between generations of learners. This is not uncommon in human experiments where processes of interaction and transmission sometimes generate fully systematic, compositional languages, but can also result in systems that lack structure entirely (Verhoef et al., 2022). Differences in personal biases may be a contributing factor to these differences (Kouwenhoven et al., 2022a). Since we do not initialise agents with different biases, these variations, originating in distributional information of the prepended vocabularies, are a natural human-like outcome of repeated exposure to and use of the language.

The evolution of degenerate vocabularies could be explained by the use of greedy decoding during signal generation, which does not necessarily produce the most human-like text (Holtzman et al., 2020; Meister et al., 2022, 2023) and may therefore also result in non-humanlike composition. Moreover, once an agent, perhaps mistakingly, duplicates a signal, its raw probabilities are increased when producing the next utterance, possibly resulting in a feedback loop that collapses onto a degenerate vocabulary. This effect may be further increased due to LLMs' inability to innovate (Bender et al., 2021; Yiu et al., 2024) and the choice of structured prompts that do not explicitly ask for innovation. Future work could attempt to increase the composition of novel signals by increasing the temperature parameter. Perhaps resulting in slightly more novel outputs as this forces exploration of the vocabulary embedding space (Peeperkorn et al., 2024), possibly alleviating the evolution of degenerate vocabularies and shifting the optimisation of the language to different solutions.

The rapid increase in learnability resulting from iterated learning proves that weak learning biases in language models, such as, for example, an observed simplicity bias (Chen et al., 2024), can be amplified by the process of generational transmission. Additional simulations with increased communicative difficulty, e.g., by increasing the number of distractors or the number of interaction partners, could reveal whether and how some form of memory constraint affects the learnability of the languages. Doing so additionally captures the diversity and dynamic nature of language in the real world. In general, systematic manipulations across model features (e.g., size, training data, or decoding strategies) may expose why we observe tendencies such as producing longer signals. Similar to what was proposed by Galke and Raviv (2024), we argue that careful manipulation of our setup can help reveal underlying mechanistic biases of language models and inform modelling choices when simulating language acquisition in LLMs. Taking into account the important role communication plays in shaping human language, LLM performance drastically increased when it was optimised for successful communication through reinforcement learning from human feedback (RLHF).

Finally, we acknowledge that our results depend on many methodological considerations, such as the prompt format, task instructions, and the tokenisation process. However, our primary goal was to investigate whether LLMs can be used in simulations of artificial language emergence. We aimed to stay maximally close to well-known experimental methods in the field of language emergence and did not optimise for performance, human-like results, or compositional vocabularies. Instead, our goal was to reveal LLMs' natural behaviours resulting from learning and using artificial languages. Future work could extend our findings by performing experiments in which humans collaborate with LLMs to investigate whether languages can evolve that are optimised for human *and* LLM preferences.

8 Conclusion

Given the remarkable linguistic abilities of recent LLMs, we show how they behave in a classical referential game in which artificial languages, typically used in the field of language evolution, are learned and used. Primarily, our results suggest that LLMs can be used as artificial language learners to investigate the evolution of language. We show that initially unstructured languages are optimised for improved learnability and allow for successful communication. While we found some evidence of human-like compositional structures that enhance generalization abilities, we also identified notable differences in behavioural characteristics of LLMs in comparison to humans. Notably, iterated learning processes increased vocabulary learnability but also amplified such different characteristics further. As such, we extend existing research by revealing that structured languages are not merely easier for LLMs to learn. Critically, the inherent biases of LLMs also shape unstructured languages towards increased regularity. These findings contribute to a deeper understanding of how LLMs process and evolve language, potentially bridging the gap between computational models and natural language evolution. Finally, we hope to have shown that our setup is useful in exposing the underlying mechanistic biases of LLMs and demystifying their uninterpretable nature.

Acknowledgments

We wish to thank Bram van Dijk for his comments on an early draft of this paper and the helpful discussions.

References

Alberto Acerbi and Joseph M. Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Inbal Arnon and Simon Kirby. 2024. Cultural evolution creates the statistical structure of language. *Scientific Reports*, 14(1):5255.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Robert Boyd, Peter J Richerson, et al. 1996. Why culture is common, but cultural evolution is rare. In *Proceedings-british academy*, volume 88, pages 77–94. Oxford University Press Inc.

Henry Brighton and Simon Kirby. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242.

Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. 2023. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868.

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

Jon W. Carr, Kenny Smith, Hannah Cornish, and Simon Kirby. 2017. The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive Science*, 41(4):892–923.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Morten H. Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509.

Jennifer Culbertson and Paul Smolensky. 2012. A bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, 36(8):1468–1498.

Bart de Boer. 2000. Self-organization in vowel systems. *Journal of Phonetics*, 28(4):441–465.

Bart de Boer. 2006. *Computer modelling as a tool for understanding language evolution*, pages 381–406. Springer Netherlands, Dordrecht.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Bruno Galantucci. 2005. An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5):737–767.

Lukas Galke, Yoav Ram, and Limor Raviv. 2022. Emergent communication for understanding human language evolution: What's missing? In *Emergent Communication Workshop at ICLR 2022*.

Lukas Galke, Yoav Ram, and Limor Raviv. 2023. What makes a language easy to deep-learn? *arXiv preprint arXiv:2302.12239*.

Lukas Galke and Limor Raviv. 2024. Learning and communication pressures in neural networks: Lessons from emergent communication. *Language Development Research*, 5(1):116–143.

Thomas L. Griffiths and Michael L. Kalish. 2007. Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Aisha Khatun and Daniel G Brown. 2024. A study on large language models' limitations in multiple-choice question answering. *Preprint*, arXiv:2401.07955.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. 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.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Tom Kouwenhoven, Roy De Kleijn, Stephan Raaijmakers, and Tessa Verhoef. 2022a. Need for structure and the emergence of communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Tom Kouwenhoven, Max Peeperkorn, Bram Van Dijk, and Tessa Verhoef. 2024. The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kouwenhoven, Tessa Verhoef, Roy De Kleijn, and Stephan Raaijmakers. 2022b. Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers in Artificial Intelligence*, 5:886349.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. *Preprint*, arXiv:2006.02419.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Yuchen Lian, Tessa Verhoef, and Arianna Bisazza. 2024. Nellcom-x: A comprehensive neural-agent framework to simulate language learning and group communication. *Preprint*, arXiv:2407.13999.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. 2022. On the probability–quality paradox in language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–45, Dublin, Ireland. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Jonas Nölle, Marlene Staib, Riccardo Fusaroli, and Kristian Tylén. 2018. The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, 181:93–104.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *15th International Conference on Computational Creativity*. Association for Computational Creativity.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Limor Raviv, Marianne de Heer Kloots, and Antje Meyer. 2021. What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, 210:104620.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019a. Compositional structure can emerge without generational transmission. *Cognition*, 182:151–164.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019b. Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, 286(1907):20191262.

Thomas C. Scott-Phillips, Simon Kirby, and Graham R.S. Ritchie. 2009. Signalling signalhood and the emergence of communication. *Cognition*, 113(2):226–233.

Reinhard Selten and Massimo Warglien. 2007. The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, 104(18):7361–7366.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Kenny Smith. 2011. Learning Bias, Cultural Evolution of Language, and theBiological Evolution of the Language Faculty. *Human Biology*, 83(2):261 – 278.

Kenny Smith. 2022. How Language Learning and Language Use Create Linguistic Structure. *Current Directions in Psychological Science*, 31(2):177–186.

Kenny Smith, Simon Kirby, Shangmin Guo, and Thomas L Griffiths. 2024. Ai model collapse might be prevented by studying human language transmission. *Nature*, 633(8030):525.

Kenny Smith, Monica Tamariz, and Simon Kirby. 2013. Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *Proceedings of the annual meeting of the cognitive science society*, volume 35.

Luc Steels, Martin Loetzsch, et al. 2012. The grounded naming game. *Experiments in cultural language evolution*, 3:41–59.

Mónica Tamariz and Simon Kirby. 2015. Culture: Copying, compression, and conventionality. *Cognitive Science*, 39(1):171–183.

Michael Tomasello. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press.

Bram van Dijk, Tom Kouwenhoven, Marco Spruit, and Max Johannes van Duijn. 2023. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654, Singapore. Association for Computational Linguistics.

Tessa Verhoef. 2012. The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4(4):357–380.

Tessa Verhoef, Esther Walker, and Tyler Marghetis. 2016. Cognitive biases and social coordination in the emergence of temporal language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38.

Tessa Verhoef, Esther Walker, and Tyler Marghetis. 2022. Interaction dynamics affect the emergence of compositional structure in cultural transmission of space-time mappings. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, pages 2133–2139.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pages 17–60.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 294–313, Singapore. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

Eunice Yiu, Eliza Kosoy, and Alison Gopnik. 2024. Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, 19(5):874–883. PMID: 37883796.

Yuqing Zhang, Tessa Verhoef, Gertjan van Noord, and Arianna Bisazza. 2024. Endowing neural language learners with human-like biases: A case study on dependency length minimization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5819–5832, Torino, Italia. ELRA and ICCL.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13785–13795.

A Communication per round

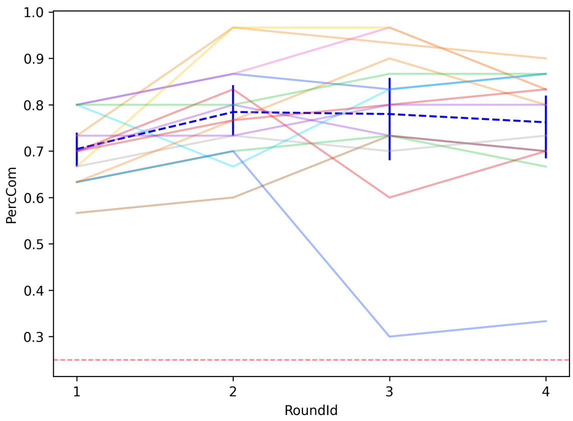


Figure 5: The communicative success (*PercCom*) over the communication rounds. Each line indicates a simulation, the dashed blue line is the average with bars indicating the 95% confidence interval. The dashed red line indicates chance performance.

B Additional results iterated learning

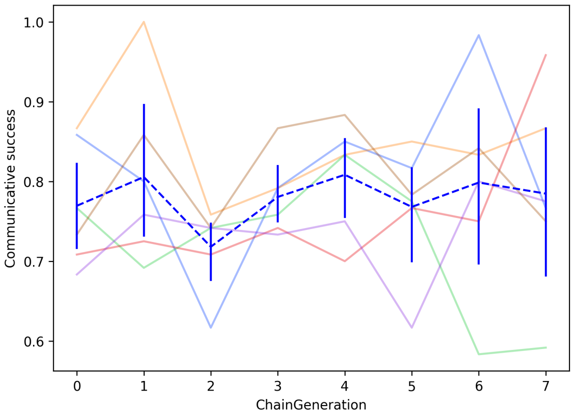


Figure 6: The average communicative success across rounds for each generation. Each line indicates a chain, and the dashed blue line is the average with bars indicating the 95% confidence interval. See Table 2.

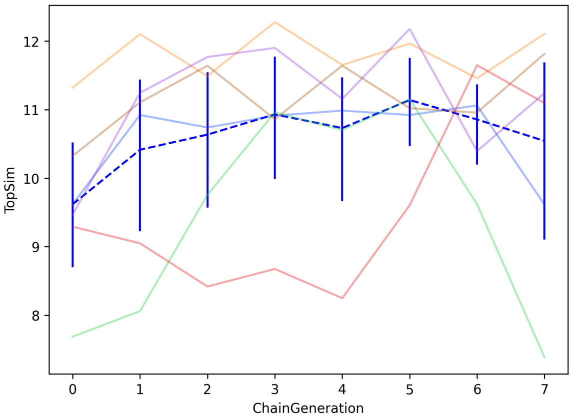


Figure 7: The evolution of *TopSim* on the words produced in the testing block. Each line indicates a chain, and the dashed blue line is the average with bars indicating the 95% confidence interval. See Table 2 for the descriptives of *TopSim* and *Ngram*.

C Prompts

Our agents act based on prompts and system instructions. These are designed to be maximally close to the classical experimental setup and formatted similar to Galke et al. (2023). The completion prompt 2 is used for labelling and guessing. For the guessing task, we prefill the word and pick the signal with the highest probability. See the full prompts for labelling and guessing (Prompt 2), speaking (Prompt 3), discrimination (Prompt 4) below.

	$t(5)$	p	M_{gen0}	SD_{gen0}	M_{gen7}	SD_{gen7}
TopSim	-1.42	.215	9.62	1.21	10.5	1.77
Ngram	2.83	.037	.158	.074	.071	.025

Table 2: Paired t-tests shows that *Ngram* does significantly change resulting from generational transmission, while *TopSim* does not.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a language
learner who has to learn an artificial language with words and their corresponding
features. Your task is to complete the vocabulary by generating a word that
describes the last item. Only respond with the word.<|eot_id|><|start_header_id|>
user<|end_header_id|>

{'shape':2,'colour':'orange','amount':1,'word':'giniwite'}
{'shape':3,'colour':'green','amount':1,'word':'ginisu'}
{'shape':1,'colour':'orange','amount':2,'word':'pinisugi'}
{'shape':3,'colour':'green','amount':3,'word':'sutepi'}
{'shape':2,'colour':'orange','amount':2,'word':'winisu'}
{'shape':3,'colour':'orange','amount':1,'word':'niwi'}
{'shape':1,'colour':'blue','amount':2,'word':'sutuwite'}
{'shape':1,'colour':'blue','amount':3,'word':'tupitene'}
{'shape':3,'colour':'blue','amount':1,'word':'wipinepi'}
{'shape':2,'colour':'orange','amount':3,'word':'gigi'}
{'shape':1,'colour':'green','amount':2,'word':'nite'}
{'shape':3,'colour':'blue','amount':3,'word':'wite'}
{'shape':1,'colour':'green','amount':3,'word':'sune'}
{'shape':2,'colour':'blue','amount':2,'word':'ninene'}
{'shape':2,'colour':'green','amount':1,'word':'tusetetu'}
{'shape':1,'colour':'green','amount':3,'word':'<|eot_id|><|start_header_id|>
assistant<|end_header_id|>
[COMPLETION OR PREFFILED]
```

Prompt 2: Completion Prompt used for labelling and guessing.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a language
learner who has to learn an artificial language with words and their corresponding
features. Your task is to generate a word such that your communication partner can
guess the correct meaning of the word. Communicative success is important. Only
respond with the word.<|eot_id|><|start_header_id|>user<|end_header_id|>

{'shape':1,'colour':'green','amount':3,'word':'sutupitite','communicativeSuccess':1}
{'shape':2,'colour':'orange','amount':2,'word':'ginupepi','communicativeSuccess':1}
{'shape':1,'colour':'orange','amount':2,'word':'sutupepi','communicativeSuccess':1}
{'shape':1,'colour':'green','amount':2,'word':'sutupepi','communicativeSuccess':0}
{'shape':2,'colour':'orange','amount':1,'word':'ginisu','communicativeSuccess':1}
{'shape':2,'colour':'orange','amount':3,'word':'ginupitite','communicativeSuccess':1}
{'shape':3,'colour':'green','amount':1,'word':'wipisu','communicativeSuccess':0}
{'shape':2,'colour':'green','amount':1,'word':'ginisu','communicativeSuccess':1}
{'shape':1,'colour':'blue','amount':2,'word':'sunupepi','communicativeSuccess':1}
{'shape':3,'colour':'green','amount':3,'word':'wipipitite','communicativeSuccess':1}
{'shape':3,'colour':'orange','amount':1,'word':'wipisu','communicativeSuccess':0}
{'shape':1,'colour':'blue','amount':3,'word':'sunupitite','communicativeSuccess':1}
{'shape':3,'colour':'blue','amount':3,'word':'wipipitite','communicativeSuccess':1}
{'shape':3,'colour':'blue','amount':1,'word':'wipisu','communicativeSuccess':1}
{'shape':2,'colour':'blue','amount':2,'word':'<|eot_id|><|start_header_id|>assistant
<|end_header_id|>
[COMPLETION]
```

Prompt 3: Speaking Prompt during communication.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a language
learner who has to learn an artificial language with words and their corresponding
features. Your task is to complete the vocabulary by interpreting the intended
meaning of the word generated by your communication partner. Communicative success
is important. Only respond with the complete last item.<|eot_id|><|start_header_id|>
user<|end_header_id|>

{'word':'wipipitite','shape':3,'colour':'blue','amount':3,'communicativeSuccess':1}
{'word':'wipisu','shape':3,'colour':'orange','amount':1,'communicativeSuccess':0}
{'word':'wipisu','shape':3,'colour':'green','amount':1,'communicativeSuccess':0}
{'word':'sutupepi','shape':1,'colour':'orange','amount':2,'communicativeSuccess':1}
{'word':'ginupepi','shape':2,'colour':'orange','amount':2,'communicativeSuccess':1}
{'word':'sutupitite','shape':1,'colour':'green','amount':3,'communicativeSuccess':1}
{'word':'wipipitite','shape':3,'colour':'green','amount':3,'communicativeSuccess':1}
{'word':'wipisu','shape':3,'colour':'blue','amount':1,'communicativeSuccess':1}
{'word':'ginisu','shape':2,'colour':'green','amount':1,'communicativeSuccess':1}
{'word':'ginisu','shape':2,'colour':'orange','amount':1,'communicativeSuccess':1}
{'word':'sunupepi','shape':1,'colour':'blue','amount':2,'communicativeSuccess':1}
{'word':'sutupepi','shape':1,'colour':'green','amount':2,'communicativeSuccess':0}
{'word':'sunupitite','shape':1,'colour':'blue','amount':3,'communicativeSuccess':1}
{'word':'ginupitite','shape':2,'colour':'orange','amount':3,'communicativeSuccess':1}
{'word':'ginupepi','shape':'<|eot_id|><|start_header_id|>assistant<|end_header_id|>
[PREFILLED WITH DISTRACTOR ATTRIBUTES]
```

Prompt 4: Guessing Prompt during communication.