Searching for Structure:

Investigating Emergent Communication using Large Language Models

Tom Kouwenhoven, Max Peeperkorn, Tessa Verhoef

Correspondence: t.kouwenhoven@liacs.leidenuniv.nl

Universiteit Leiden The Netherlands

University of Kent

The Premise:

Human languages have **evolved to be structured** through **repeated language learning and use** (Smith, 2022). These processes introduce pressures that cause languages to adapt to **human biases** that operate during **language acquisition** and the need to **communicate efficiently**. Here we investigate whether the same processes can result in artificial languages that are optimised for **implicit biases** of **LLMs**. Inspired by previous experiments with human participants, we answer the following questions:

→ Can LLMs learn **artificial languages** and use them to **communicate** successfully?

→ Can **generational transmission** be used to **reveal** intrinsic mechanistic LLM **biases?**

Contributions:

→ LLMs can learn **artificial languages** and use them to **communicate successfully** in referential games.

→ Languages exhibit higher **structural properties** after **repeated language use**.

→ **Structured** languages are **easier to learn**. Critically, processes of **language learning** and **use** with LLMs also **shape** unstructured languages to **increased regularity**.

→ **Generalisation** is **more systematic** when the evolved language is **more structured.**

→ Languages become **easier to learn** (but less human-like) as a result of **generational** transmission, adapting to intrinsic biases of LLMs.

Our work contributes to a **deeper understanding** of how LLMs **process** and **evolve languages.** This method can potentially be used to **bridge the gap** between computational models and natural language evolution, enabling more natural interactions between humans and machines (Kouwenhoven et al. 2022).

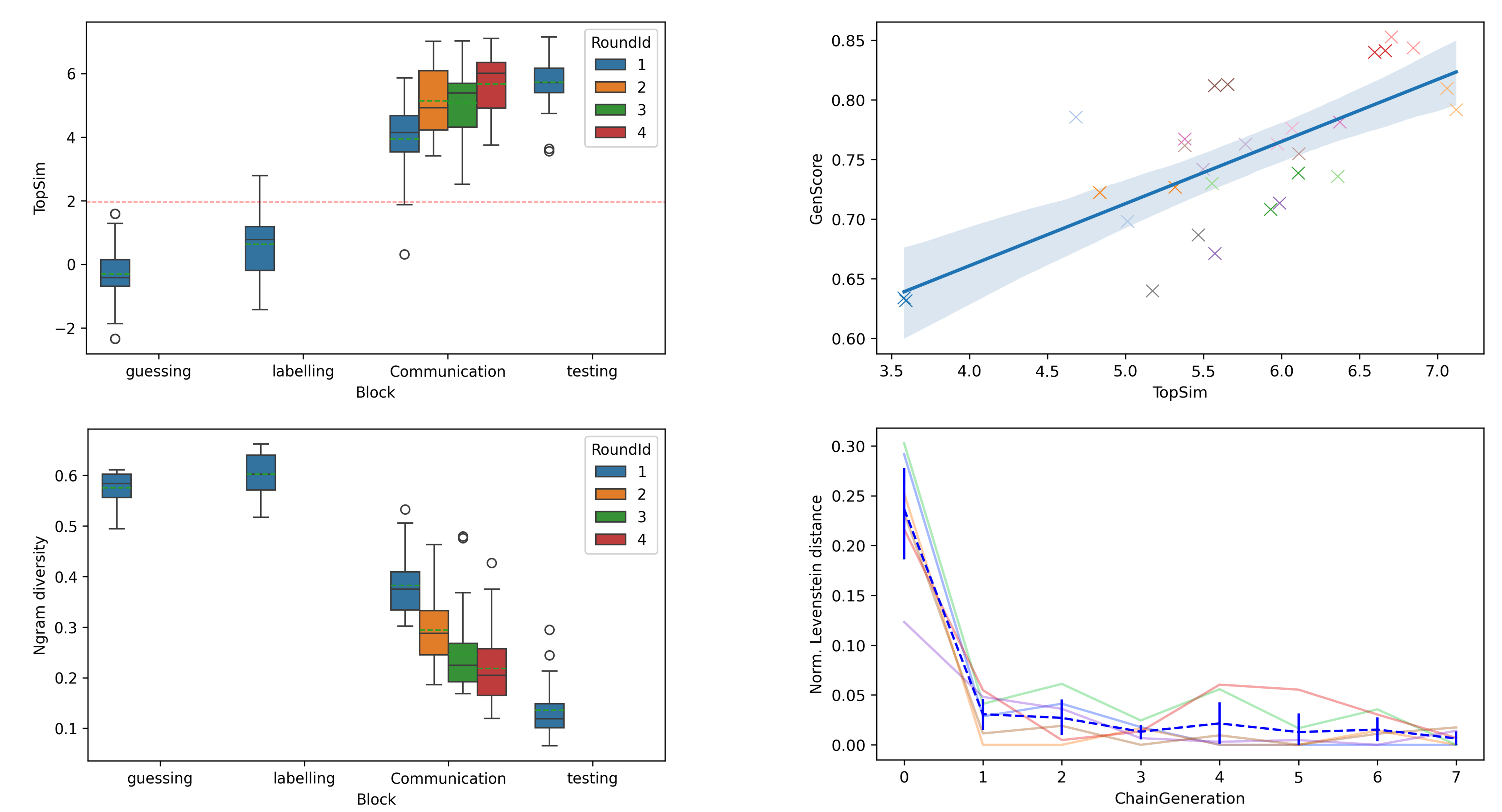
Measures:

Topographic Similarity *(TopSim)*
TopSim measures whether meanings with similar attribute-values are labelled in similar ways.

Ngram diversity
Measures the average fraction of unique vs. total Ngrams for N ∈ {1,2,3,4,5} in all produced signals.

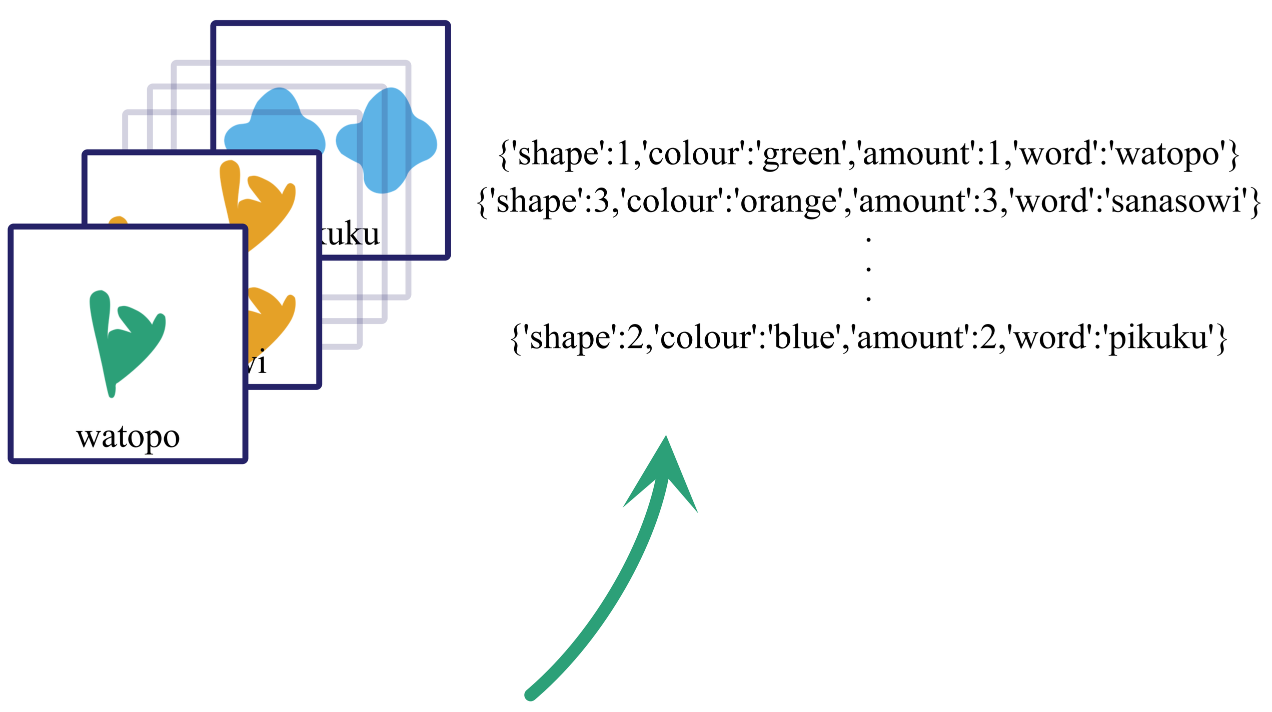
Generalisation Score *(GenScore)*
The *GenScore* measures the degree of signal systematicity between the signals produced for unseen stimuli in the test block and the previously seen stimuli in the communication block.

Norm. Levenshtein distance
The normalised edit-distance between the target label and the predicted label.



Method:

Learning languages



We use LLM-augmented agents to test if LLMs can **learn artificial languages**.

Agents are comprised of Llama 3 70B and learn by virtue of in-context learning in a JSON-like format similar to Galke et al. 2024. The stimuli are inspired by Raviv et al., 2021 and the artificial languages are generated conform Kirby et al. 2015.

Simulating Language Evolution

The simulation consists of four blocks: guessing, labelling, communication and testing. The communication block is a classical **Lewis referential game** in which **two** agents **communicate** to discriminate the target stimulus among three distractor stimuli. They do so in four rounds, each consisting of 30 interactions, alternating speaker-listener roles between interactions. Here the **speaker** observes a target stimulus and **utters a signal** that describes the current stimulus. Using this utterance, the **listener** must **discriminate the correct target**. During testing, we test whether agents can generalise their vocabularies to new stimuli.