Shaping Shared Languages:

Human and Large Language Models' Inductive Biases in Emergent Communication

Tom Kouwenhoven, Max Peeperkorn, Roy de Kleijn, Tessa Verhoef

Correspondence: t.kouwenhoven@liacs.leidenuniv.nl

Universiteit Leiden The Netherlands

University of Kent

The Premise:

Languages are **shaped** by the **inductive biases** of their users. This happens through **repeated language learning and use** (Smith, 2022). Using a classical referential game, we investigate how **artificial languages** evolve when they are optimised for inductive biases in humans and large language models (LLMs) via **Human-Human, LLM-LLM** and **Human-LLM** experiments. Inspired by previous experiments with human experiments and simulations, we ask:

→ Can humans and LLMs **collaboratively shape a language** that is easy to learn for both users and still allows for **successful communication**?

→ If so, what **properties** do these languages have?

Contributions:

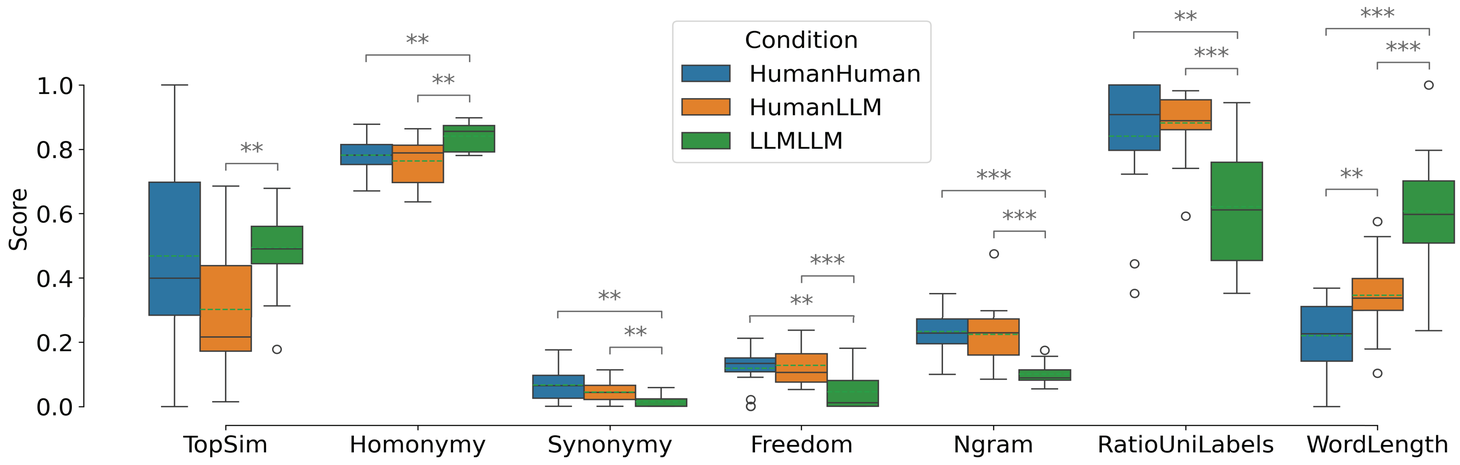
→ Our results show that **structured** and **referentially grounded languages** can **emerge** when **humans** and **LLMs interact** repatedly.

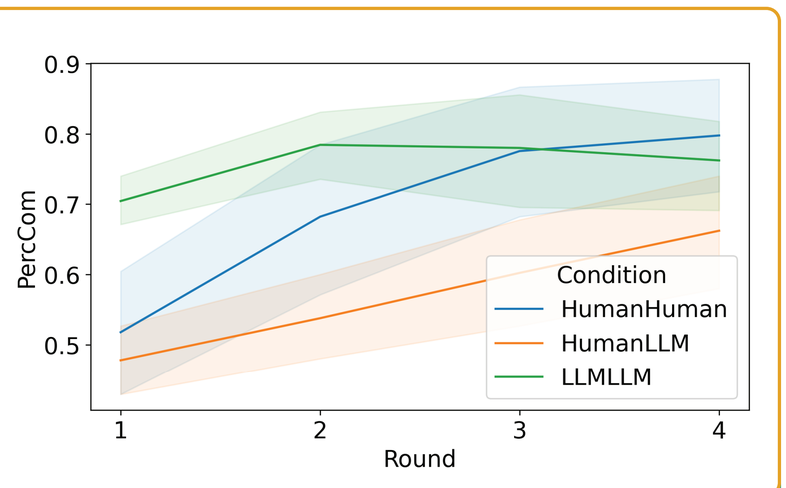
→ The **languages** that emerge from **human-LLM** interactions are **more human-like** than LLM-like, suggesting that **LLMs** are **flexible towards** the **human preferences** that shape the languages.

→ Finally, **languages** that are optimised for **LLMs** have **less variation** and are **more degenerate** than those optimised for humans.

These findings advance our understanding of how LLMs may adapt to the dynamic nature of human language, contribute to its evolution, and maintain alignment with human understanding of language. To achieve successful interactions between humans and machines, it may be fruitful to **optimise for communicative success**.

Results:





Topographic Similarity gauges if meanings with similar attribute-values are labelled in similar ways. TopSim ≈ 1 indicates some structure.

Homonymy (many-to-one) measures how many attribute-values a character in a position can refer to. If homonymy ≈ 1, characters can map to multiple attribute values.

Synonymy (one-to-many) measures whether each attribute value is encoded by a single character in a position. Synonymy ≈ 0 means that characters map to single attribute values.

Freedom (word order) measures variability in the order by which labels are composed. Freedom ≈ 0 means that each value of a specific attribute is encoded in a specific position of the label.

Ngram diversity is the average fraction of unique vs. total Ngrams in all signals for $N \in \{1,2,3,4,5\}$. Ngram ≈ 0 means label parts are reused.

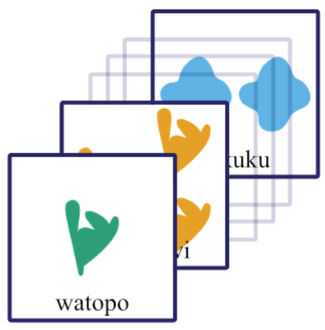
Method:

Learning languages

Humans learn the artificial languages through **two rounds of exposure** in which they try to memorise **15 stimuli** and the labels that correspond to them.

LLM-augmented agents are comprised of Llama 3 70B and learn by virtue of **in-context** learning similar to Galke et al. 2024 and Kouwenhoven et al. 2025. The stimuli are inspired by Raviv et al., 2021 and the artificial languages are generated conform Kirby et al. 2015.

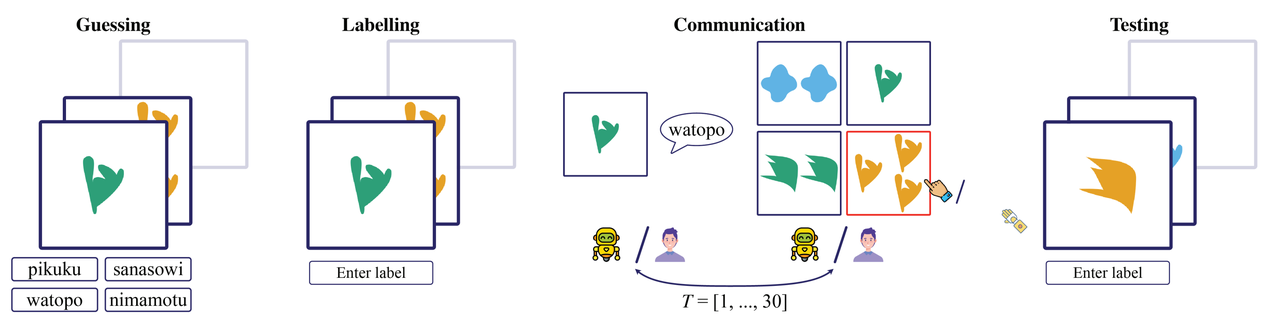
Human learning



LLM learning

```
{'shape':1,'colour':'green','amount':1,'word':'watopo'}
{'shape':3,'colour':'orange','amount':3,'word':'sanasowi'}
.
.
.
{'shape':2,'colour':'blue','amount':2,'word':'pikuku'}
```

Human-Machine Language Evolution



The experimental setup involves a classical **Lewis referential game** in which **two entities communicate** to discriminate the target stimulus among three distractor stimuli. Communication happens 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.

We use **three conditions** (Human-Human, LLM-LLM, Human-LLM) to compare the resulting communicative systems.